การค้นหาเครื่องหมายชีวภาพของโรคไตอักเสบลูปัสด้วยวิธีการศึกษาแบบชีววิทยาเชิงระบบ

นายภูมิพัฒน์ ทองอยู่

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาชีวเวชศาสตร์ (สหสาขาวิชา)
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2558

IDENTIFICATION OF BIOMARKERS OF LUPUS NEPHRITIS BY SYSTEMS BIOLOGY APPROACH

Mr. Pumipat Tongyoo

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Biomedical Sciences
(Interdisciplinary Program)
Graduate School
Chulalongkorn University
Academic Year 2015

Thesis Title	IDENTIFICATION OF BIOMARKERS OF LUPUS NEPHRITIS BY SYSTEMS BIOLOGY APPROACH
By	Mr. Pumipat Tongyoo
Field of Study	Biomedical Sciences
Thesis Advisor	Professor Yingyos Avihingsanon, M.D.
Thesis Co-Advisor	Professor Nattiya Hirankarn, M.D.,Ph.D.
	Assistant Professor Santitham Prom-on, Ph.D.

Accepted by the Graduate School, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctoral Degree

.....Dean of the Graduate School
(Associate Professor Sunait Chutintaranond, Ph.D.)

THESIS COMMITTEE

.....Chairman
(Professor Apiwat Mutirangura, M.D.,Ph.D.)

.....Thesis Advisor
(Professor Yingyos Avihingsanon, M.D.)

.....Thesis Co-Advisor
(Professor Nattiya Hirankarn, M.D.,Ph.D.)

.....Thesis Co-Advisor
(Assistant Professor Santitham Prom-on, Ph.D.)

.....Examiner
(Trairak Pisitkun, M.D.)

.....Examiner
(Associate Professor Chatchawit Aporntewan, Ph.D.)

.....External Examiner
(Assistant Professor Treenut Saithong, Ph.D.)

ภูมิพัฒน์ ทองอยู่ : การค้นหาเครื่องหมายชีวภาพของโรคไตอักเสบลูปัสด้วยวิธีการศึกษาแบบชีววิทยาเชิงระบบ (IDENTIFICATION OF BIOMARKERS OF LUPUS NEPHRITIS BY SYSTEMS BIOLOGY APPROACH) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ศ. นพ. ยิ่งยศ อวิหิงสานนท์, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ศ. พญ. ดร. ณัฏฐิยา หิรัญกาญจน์, ผศ. ดร. สันติธรรม พรหมอ่อน, 153 หน้า.

Lupus Nephritis (LN) เป็นโรคภูมิต้านเนื้อเยื่อตนเองที่มีความสำคัญในประเทศไทย ผู้ป่วยเกิดความผิดปกติในหลายระบบของร่างกายเช่นเดียวกับโรค Systemic lupus erythematosus (SLE) โดย LN จะแสดงอาการเฉพาะที่ไตเป็นหลัก กลไกการเกิดโรค SLE เกิดจากปัจจัยทางพันธุกรรมร่วมกับปัจจัยทางสิ่งแวดล้อม คณะผู้วิจัยที่คณะแพทยศาสตร์จุฬาฯ ได้รายงานปัจจัยทางพันธุกรรมของโรคนี้และได้รายงานยีนที่สำคัญในผู้ป่วยเอเชียซึ่งต่างจากชาวคอร์เคเซียนไว้จำนวนหนึ่ง งานวิจัยนี้ได้ใช้เทคนิคทางด้าน integrative systems biology เปรียบเทียบระหว่างข้อมูลการแสดงออกของยีนในระดับ mRNA และโปรตีนของผู้ป่วย LN ที่ตอบสนองต่อการรักษาและไม่ตอบสนองต่อการรักษาเพื่อที่จะค้นหา biomarker ชนิดใหม่ๆเพื่อใช้ในการตรวจวิเคราะห์ ติดตามการรักษาโรค SLE/LN นอกจากนี้ผลการวิเคราะห์ Meta analysis ของร่วมกับผล GWAS มีการยืนยันถึงการเปลี่ยนแปลงของระดับของหมู่เมทิลในดีเอ็นเอ (DNA methylation) เป็นกลไกหนึ่งที่เกี่ยวข้องกับการเกิดโรค SLE กลไกนี้เป็นส่วนหนึ่งที่ช่วยอธิบายว่าสิ่งแวดล้อมจากแสง UV จากยา หรือจากอาหารบางอย่างอาจมีผลต่อการเกิดโรคได้โดยผ่านกระบวนการเปลี่ยนแปลงของระดับ methylation นั่นเอง ความเข้าใจถึงความผิดปกตินี้จึงมีความสำคัญมากต่อการพัฒนาการรักษาโรคเนื่องจากเราสามารถเปลี่ยนแปลงระดับ methylation ได้ง่ายกว่าการแก้ไขยีน ปัจจุบันมีผู้สนใจศึกษาความผิดปกติของ methylation ในส่วนโปรโมเตอร์ที่ควบคุมการแสดงออกของยีนเป็นหลัก อย่างไรก็ตามบริเวณที่เป็นเบสซ้ำเป็นอีกส่วนหนึ่งที่มีการควบคุมด้วย methylation เป็นหลักและน่าสนใจมากในโรค SLE ซึ่งการวิเคราะห์ระดับ methylation ใน IRS ของผู้ป่วย active SLE พบว่าในเซลล์ที่สำคัญต่อการเกิดโรค CD3+CD4+ T lymphocytes มี ระดับ hypomethylation ของเบสซ้ำชนิด HERV-E LTR2C ผู้วจัยจึงได้ทำการวิเคราะห์ข้อมูล HERV (human Endogenous retrovius) ซึ่งเป็นหนึ่งในเบสซ้ำที่เป็น transposable element และพบได้มากถึง 8% ใน genome intersperse repetitive sequences (IRS) ซึ่งกลุ่มของ HERV เป็นเบสซ้ำที่สามารถ transcribe เป็น mRNA ได้ เคลื่อนที่ในจีโนมได้ และมีรายงานว่าสามารถควบคุมยีนข้างเคียงได้ ดังนั้น ผู้วิจัยจึงได้ทำการวิเคราะห์การกระจายตัวของ HERV ใน genome มนุษย์และสร้างเป็นฐาน database ชื่อ EnHERV ขึ้นมารวมทั้งสร้างฟังก์ชันวิเคราะห์การแสดงออกของ HERV ร่วมกับการแสดงออกของยีน โดยสามารถเข้าใช้งานได้ที่ http://sysbio.chula.ac.th/enherv นอกจากนี้ คณะผู้วิจัยจึงตั้งสมมติฐานว่าในเซลล์ของผู้ป่วยลูปัสน่าจะมีเปลี่ยนแปลงการแสดงออกของยีนข้างเคียงโดยเกิดการสร้าง chimeric transcripts อันประกอบไปด้วยส่วนหนึ่งของ LTR และส่วนของยีนข้างเคียง ซึ่งจะเป็นการเพิ่มหรือลดการแสดงออกของยีนข้างเคียงนั้น ทั้งนี้ คณะผู้วิจัยมุ่งหวังที่จะค้นพบ chimeric transcripts ที่มีความสำคัญกับการดำเนินของโรค SLE รวมทั้งสามารถใช้เป็น biomarker ที่ใช้บอก prognosis หรืออาจใช้ในเป็นเป้าหมายสำหรับการพัฒนายาหรือการรักษาใหม่สำหรับโรค SLE ได้

สาขาวิชา ชีวเวชศาสตร์	ลายมือชื่อนิสิต .....
ปีการศึกษา 2558	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
	ลายมือชื่อ อ.ที่ปรึกษาร่วม .....
	ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

5387849620 : MAJOR BIOMEDICAL SCIENCES
KEYWORDS: SYSTEMS BIOLOGY / LUPUS NEPHRITIS / HUMAN ENDOGENOUS RETROVIRUS / ENHERV / RNA-SEQ / SLE

PUMIPAT TONGYOO: IDENTIFICATION OF BIOMARKERS OF LUPUS NEPHRITIS BY SYSTEMS BIOLOGY APPROACH. ADVISOR: PROF. YINGYOS AVIHINGSANON, M.D., CO-ADVISOR: PROF. NATTIYA HIRANKARN, M.D.,Ph.D., ASST. PROF. SANTITHAM PROM-ON, Ph.D., 153 pp.

Systemic lupus erythematosus (SLE) is a systemic autoimmune disease which is important in Thailand. Inflammation and tissue damage occurs at many organs of SLE patients. Lupus Nephritis (LN) is a major complication of SLE that cause inflammation of the kidney. The mechanisms underlying SLE pathogenesis including genetics and environmental factors. The integrative systems biology approach by transcriptome and proteomic data were used in this study to identify new biomarker. Asian Lupus Consortium has reported some important SLE-related genes which are unique in Asian population. Furthermore, our recent result from Meta analysis combining with GWAS data has identified SLE-related genes in Asian SLE patients which are involved in demethylation processes. This confirms that the dynamics of DNA methylation level is a mechanism related with SLE pathogenesis. Therefore, the environmental factors including UV light, drugs or some kind of food may contribute to SLE pathogenesis through the DNA methylation dynamics. Understanding of this mechanism is important for improving SLE treatment because DNA methylation can be easier manipulated than the genes themselves. Most DNA methylation researches have focused on the promoter which controls gene expression; however, the repetitive sequences in genome are also controlled by DNA methylation are very interesting in SLE. We recently reported the DNA methylation dynamics of intersperse repetitive sequences (IRS) in CD4+, CD8+ T lymphocytes, B lymphocytes and neutrophils. The results show that HERV (human Endogenous retrovius), a transposable element which occupies 8% of the genome, is hypomethylated in CD3+CD4+ T lymphocytes. Since HERV can transcribe into mRNA, retrotranspose in the genome, as well as control expression of the neighbor genes, we hypothesized that HERV hypomethylation in the SLE cells can alter expression of the neighbor genes by transcribing chimeric transcripts. We analyzed HERV distribution in human genome and constructed EnHERV, which is the database that allows researchers to search HERV in human genome based on their pattern or interested genes. EnHERV also provides enrichment analysis function which can identify specific HERV pattern in published expression data. EnHERV is available at http://sysbio.chula.ac.th/enherv. We also attempt to identify the chimeric transcripts using publish RNA-Seq data in SLE patients. We hypothesized that HERV-LTR can alter the expression of their neighbor genes by using their regulatory mechanism. We aimed to discover novel chimeric transcripts, which are important in SLE pathogenesis and can be used as biomarkers for predict the prognosis, develop drugs and improve the treatment of SLE.

Field of Study: Biomedical Sciences
Academic Year: 2015

Student's Signature .....
Advisor's Signature .....
Co-Advisor's Signature .....
Co-Advisor's Signature .....

ACKNOWLEDGEMENTS

First, I would like to thank you my thesis advisor, Prof Yingyos Avihingsanon for giving me a chance to join the Lupus research unit team under a funding support from Thailand Research Fund (TRF) under the Royal Golden Jubilee (RGJ) PhD program. I have to thank you my thesis co-advisors, Prof.Dr.Nattiya Hirankarn for a valuable suggestion and she always teach me how to improve a research skills. I also would like to thanks Asst.Prof.Dr.Santitham Prom-On, my thesis co-advisors as well. He also give me a lot of valuable suggestion in computational and bioinformatics work. I won't be able to complete this study without their support. I am also sincerely thanks RGJ for the support as well. It is very nice to work with lupus team especially Dr.Thitima Benjachart and Dr.Poorichaya Somparn who always give me valuable resources and precise advice to improve my work.

I would like to special thanks to Prof.Dr.Apiwat Mutirangura for always giving me guideline on my research direction. I would like to thank Dr. Sissades Tongsima, a head of Biostatistics and Informatics Laboratory, Genome Institute, BIOTEC who allows me to use the HPC server at BIOTEC. It was impossible to finish my analysis without that facility. I am really appreciate that.

I would like to thanks everyone that I didn't mention here who encouraged and supported me to complete this dissertation as well.

Finally, I would like to thanks my parents for their unconditional support during my entire study life and helped me to be the person I am today.

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BaEV	Baboon Endogenous Virus
CCDS	Consensus Coding Sequence Database
DLE	Discoid lupus erythematosus
DAVID	Database for Annotation, Visualization and Integrated Discovery
DNA	Deoxyribonucleic Acid
ERV	Endogenous Retroviruses
GEO	Gene Expression Omnibus database
GIRI	Genetic Information Research Institute
GO	Gene Ontology
HERV	Human Endogenous Retroviruse
LTR	Long Terminal Repeat
MMTV	Mouse Mammary Tumor Virus
MaLR	Mammalian apparent LTR-retrotransposon
mRNA	messenger Ribonucleic Acid
MuLV	Murine Leukemia Virus
NCBI	National Center for Biotechnology Information
ORF	Open Reading Frame
RepSeq	Reference Sequence Database
RNA	Ribonucleic Acid
RNP	Ribonucleoprotein
RU	Repbase Update
SLE	Systemic Lupus Erythematosus
TE	Transposable Element
tRNA	transfer Ribonucleic Acid
UCSC	University of California, Santa Cruz
NGS	Next Generation Sequencing
RNA-Seq	RNA Sequencing

CHAPTER I

INTRODUCTION

Lupus nephritis (LN) is an autoimmune disease which is the second leading cause of glomerular diseases in Thailand. It is the most common and serious complication of systemic lupus erythematous (SLE). One-third of patients died or reached end-stage kidney disease within seven years [1]. Most patients with LN had a renal relapse (flare) within five years after initial diagnosis of nephritis. The leading causes of death included infection, uremia, and cardio-pulmonary failure. There was a report showing that Asian patients have higher rates of LN and more active glomerulonephritis than the Caucasian patients [2]. Up to date, Kidney biopsy is still necessary for the diagnosis and confirmation of relapse. There is still need for a non-invasive tool to monitor relapse as well as to guide the treatment decision. Therefore, a seeking for other molecular biomarkers of kidney diseases are now the highlight research in nephrology. Moreover, the exact etiology of SLE/LN is still unclear nowadays. Many studies have reported the contribution of multi-factors such as genetic factor, environmental factor and also abnormalities of immune system. Furthermore, there are increasing evidences supporting role of epigenetics in SLE/LN pathogenesis, which can explain the link between environment and genetic regulation [3, 4].

The availability of high-throughput technology make the molecular research growth very fast in term of data generation which allow for identifying of various candidate molecular targets, such as mRNA and protein for specific question. However, using microarray technology, gene expression profiles only measure transcripts at the cells expression level for specific conditions. Most biocellular processes are affected by protein-protein or other protein-substrate interactions. At the same time, the transcriptomic analysis is able to track the regulation process to feedback regulations by the expressed proteins in bio cellular mechanism. In other words, gene expression is rather interconnected with protein profile at a certain time and condition. Therefore, the analysis of gene expression profiling along with proteome level could provide a snapshot of controlled biosynthesis, which might be regulated by the transcriptomic

profile level. Understanding the causal regulatory interactions of both mRNA and protein will improve the understanding at the certain conditions.

With the availability of LN Biobank samples includes kidney tissues, urine and blood at lupus research unit, faculty of medicine, Chulalongkorn university, we have recently reported the approach to identify novel biomarker using global gene expression by Illumina microarray platform. We found that 442 and 374 probe sets were upregulated and downregulated in the non-responder kidney tissues. The interesting gene sets that could predict a non-responder including tight junction gene (claudin), B-lymphocyte stimulating factors (BAFF, APRIL). Moreover, a loss of kidney function might be predicted by set of genes such as complement pathway (SERPINA) or ANXA13 [5]. Since, urine is a logical resource as non-invasive marker of kidney diseases as it is secreted from diseased kidneys. Thus, our team also analyzed protein profiles by 2 dimension gel electrophoresis for urine biomarker discovery [6]. We have validated two proteins by ELISA which are PGDS and ZAG that increased in active LN while PGDS was specific to lupus disease only. Moreover, urine protein profiles of non-responder were characterized as well. Interestingly, APRIL was discovered by microarrays as a biomarker for lupus nephritis and was validated in serum and kidney tissues [7]. APRIL and BAFF play an important role in the pathogenesis of lupus disease. Anti-BAFF pathway is now the most interesting molecule among the pharmaceutical target therapy.

There have been a number of biomedical studies that investigated the integration between transcriptomic and proteomic data. Among them, Ou and colleagues conducted a cancer biomarker study that integrates proteomic and gene expression mapping together [8]. In their study, the proteomic data were mapped to mRNA transcript database of cancer cell lines. They found novel proteomic biomarkers half of them were successfully validated. Interestingly, it should be noted that event the proteomic data and transcript database of cancer cell lines are from different sources, yet the integration still yield promising candidate biomarkers. With the data from the same cell culture, researcher can investigate the underlying molecular mechanism in human cells. Shibuya and colleagues conducted the integrative study to determine the genes and proteins of human RPE cells that are altered by exposure to TFPI-2 [9]. The transcriptomic and proteomic data were integrated using data mining. By integrating

both transcriptomic and proteomic data together, they found the potential mechanism of gene-protein association with the growth-promoting effect of TFPI-2 on the human RPE cells. Based on the assumption that the essential proteins tend to cluster together as a connected protein-networks for a particular biological process. The connections between single interactions that make up a whole biological network are proposed to directly affect the phenotype [10]. With the available of high throughput data in both mRNA expressions in the kidney and proteomics profile in the urine of lupus nephritis patients, our first objective is to reveal more comprehensive view on the molecular mechanism level in LN by using integrative approach.

The growing evidence of epigenetics as the science of changes to gene function not explained by structural changes to the genome indicates that aberrant in DNA methylation which is one of the most highly studied topics in epigenetics, seems to plays essential roles in the pathogenesis of SLE.[11, 12]. This phenomenon might help explain the complexity and emphasize how our genes are continually interacting with the environment around us. Since DNA methylation is not occurred only in promoter of genes but as a complex composition of our genome, it also occurs at the interspersed repetitive sequences (IRS) in human genome which found approximately 45% of the human genome. The IRS also known as transportable element (TE) due to their ability to copy or cut and then place in other location in human genome. They can be divided into DNA transposons and retroelements, encompass about 2.8% and 42.2% of the human genome, respective [13]. Even more surprisingly, our genomes carry both hosts and viruses's genetic content and hence, we are all part virus. Retroelements can be divided into 2 groups based on the presence or absence of long terminal repeats (LTRs). There are 2 types of a high copy number of non-LTR retroelements; short interspersed nuclear elements (SINEs e.g. ALU) and long interspersed nuclear elements (LINEs). While, the majority of LTR retroelements is Human endogenous retroviruses (HERVs). In most cases, HERVs contain in the human genome is solitary LTRs due to recombination of the two LTRs [14]. They were considered as junk DNA for a long time but currently studies reported the functional role of many IRSs in human genome, suggested that they can affect the human genome from generating insertion and instability effect to genome by serves as alternative promoter, enhancer, exon, or

polyadenylation signal to their neighbor genes [15]. As a result, it can change gene expression and might contribute to the disease etiologies.

Several studies have reported an association of HERVs and autoimmune disease. The expression of HERV-E gag level is increasing in peripheral blood mononuclear cells (PBMCs) of SLE and there was report of the increasing of HERV-K gag gene in rheumatoid arthritis (RA) patients as well. [16]. Furthermore, HERV-E gag transcription was reported to correlate with blood plasma concentrations of anti-U1 ribonucleoprotein (RNP) and anti-Sm antibodies in SLE patients. It was also reported that HERV element participated in splicing event of pre-mRNA to mRNA in SLE [17]. Nakkuntod and colleagues [18] reported results based on the examination of methylation status of two HERV-E and HERV-K in lymphocytes from patients with SLE. They found that hypomethylation of specific HERV was a feature for SLE patients. The implication is that lower methylation levels will allow for expression of HERV genes, which may then have some biological consequences. For example, 1) increased aberrant HERV transcripts might lead to the production of autoantibodies due to molecular mimicry, 2) HERV mRNA might serve as foreign nucleic acids and stimulate abnormal immune response via endogenous immune receptor, 3) regulatory region in the HERV such as LTR can affect neighboring gene expression. Therefore, the second objective in this study aim to create a bioinformatics tool to facilitate the analysis of HERV in the genome e.g., the association studies between HERV characters and expression level of their neighboring genes that might help discover the role of certain HERV in diseases.

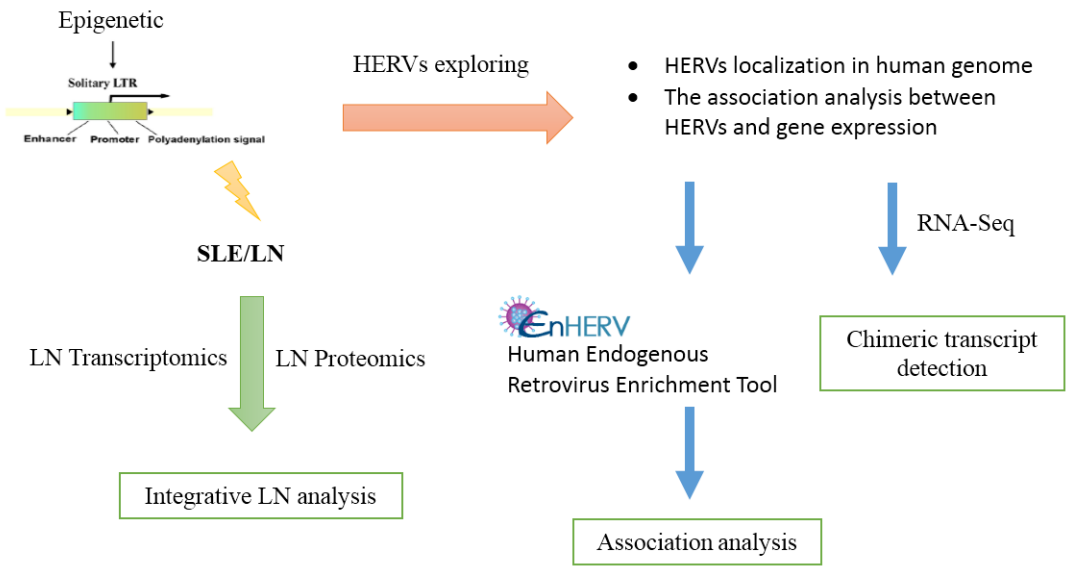
With the emergence of next generation sequencing (NGS) technology, high-throughput sequencing of the transcriptome, also known as RNA-Sequencing (RNA-Seq) provides a capture of an entire range of expression levels, with the advantage in the detection of novel transcripts and alternative splicing event from the generated data. Several RNA-Seq studies have proved that HERVs become one of the direct regulation on human genes by using their enhancer and promoter motifs present in their LTR. Moreover, these chimeric transcriptions in novel gene isoforms containing retroviral and human transcript sequence are transcribed from HERV promoters were report to associate with diseases. Therefore, by taking the advantage of the public RNA-Seq data, our third objective aims to identify chimeric transcripts in SLE. The finding of new

biomarkers using different available approaches nowadays will not only use for measuring the SLE/LN progression or early diagnosis but it might help to reveal the underlying knowledge in SLE/LN pathogenesis.

Objectives

1. To find new biomarker by using systems biology approach for SLE/LN.
2. To construct the HERVs database and analysis tool.
3. To find the association of HERV elements and their neighboring genes by using available high throughput human genomics and transcriptomics data.

Conceptual framework



CHAPTER II

LITERATURE REVIEW

Systemic lupus erythematosus (SLE) and lupus nephritis (LN)

Renal disease is a common and serious manifestation of systemic lupus erythematosus (SLE). Lupus nephritis (LN) is one of the most severe complications in SLE. Pathology of LN, including glomerular, tubulointerstitial, and vascular lesions, occurs in up to 60% of patients with SLE. LN is one of the most leading causes of morbidity and mortality in SLE [19]. The presentations can range from asymptomatic urinary abnormalities to rapidly progressive renal failure leading to end-stage renal disease. SLE is progressing to LN thought the dependent on the loss of self-tolerance and lead to the autoantibodies forming then deposit in the kidney to induce nephritis. Dysregulated apoptosis and inadequate removal of apoptotic cells and nuclear remnants are purpose to contribute in autoimmunity by causing prolonged exposure of the immune system to nuclear and cell membrane components [20]. The Recent studies have described specific genetic linkage to the development of renal disease in SLE among certain ethnic groups, including European American and African American populations, some of which may determine the severity of the glomerular disease. The immune dysregulation in lupus nephritis is characterized by polyclonal B-cell activation, which induced by cognate autoreactive helper T-cells, and the formation of autoreactive antibodies directed against nuclear antigen and other self-antigen, so-call autoantigen [21]. In general, LN is associate with high titers of circulating high-affinity, IgG anti-double-stranded DNA (anti-dsDNA) antibodies and glomerular immunoglobulin deposits. Elution of immunoglobulin from glomeruli revealed enrichment for anti-dsDNA antibodies. Therefore, it has been postulated that anti-dsDNA autoantibodies are nephritogenic in lupus nephritis. The binding of anti-double-stranded DNA (anti-dsDNA) autoantibodies to the glomerular basement membrane (GBM) in lupus nephritis can be explained by two mechanisms: i) direct cross-reactive binding to intrinsic glomerular antigens; ii) nucleosome-mediated binding to heparin sulfate in the GBM [22].

The dominant feature of renal in lupus is proteinuria, present in almost every patient and commonly leading to the nephrotic syndrome. Microscopic hematuria is almost always present, but never in isolation while macroscopic hematuria is rare. Surprisingly, hypertension is not overall more common in those with nephritis than in those without; but, as expected, those with more severe nephritis are more commonly hypertensive. About half will show a reduced GFR, and occasional patients present with acute renal failure. Renal tubular function is disturbed, which is not surprising in view of the finding of both immune aggregates in tubular basement membranes and the presence of interstitial nephritis. In a high proportion of patients, urinary excretion of light chains and β2-microglobulin are both increased. Recently, hyperkalemic renal tubular acidosis has been emphasized as a manifestation of lupus. However, there are also some lupus nephritis patients with no clinical evidence of renal involvement (no proteinuria, normal urine microscopy, normal renal function) nevertheless showed active histological change on renal biopsy specimens. This has become known as "silent lupus nephritis". Recently, the investigators found significant renal involvement (Class III, IV, or V LN) in SLE patients with < 1000 mg proteinuria with or without hematuria. These findings suggest that kidney biopsy is strongly considered in this patient population [23].

There are quite a number of studies on Genome wide association (GWA) in SLE patients (6-8). The results listed a number of candidate genes including *HLA*, *FCGR*, *PTPN22*, *STAT4* and *IRF5*. This information helps elucidating novel pathogenesis of SLE. For instance, the discovery of *BLK* and *BANK1* genes emphasize the crucial role of B cell in pathogenesis of SLE. Another novel gene namely, *ITGAM* [24] was also identified which is an adhesion molecule that regulates leukocyte adhesion to endothelial cells and may contribute to vasculitis in patients with SLE. Other study has also found the association such as *TNFAIP3* [25]. This data highlight the inflammatory role of TNF pathway in the pathogenesis of SLE including *ITPR3* [26] and *TNXB* [27].

Gene expression profiling in lupus nephritis

Several studies have analyzed global gene expression profiles of SLE patients. Interferon (IFN)-related genes are a dominant signature in SLE patients. There are still limited numbers of gene expression studies in kidney tissues of LN as shows in table 1. Nevertheless, the transcriptional profiles of lupus glomeruli was available. Similarly to lupus's peripheral blood mononuclear cells (PBMC) study, IFN inducible genes were over-expressed in lupus glomeruli. A large gene cluster with decreased expression found in all samples included ion channels and transcription factors, indicating a loss-of-function response to the glomerular injury [28]. In 2008, Kamatani *et al* [29] showed 161 genes of leukocyte that different expression between LN and healthy group. The differential expressed genes were associated with antiviral, immunity, helicase and hydrolase activity. Recent study in LN biopsy showed the 20 commonly key pathologic processes of immune cell infiltration/activation, endothelial cell activation/injury, and tissue remodeling/fibrosis with macrophage/dendritic cell activation as a dominant cross-species shared transcriptional pathway [30].

Table 1 Gene expression studies in lupus Nephritis

Year (reference)	Results
2004 [28]	A large gene cluster with decreased expression found in all samples included ion channels and transcription factors, indicating a loss-of-function response to the glomerular injury.
2007 [29]	161 genes were identified as differential expression. These gene were uniquely overrepresented in antiviral, immunity, helicase, hydrolase activity
2012 [30]	The 20 commonly key pathologic processes of immune cell infiltration/activation, endothelial cell activation/injury, and tissue remodeling/fibrosis, with macrophage/dendritic cell activation as a dominant cross species shared transcriptional pathway.
2015 [5]	Tight junction proteins were purposed as promising biomarker for refractory LN analysis.

Proteomic experiments in lupus nephritis

In this recent years, there has been increased interest in exploring the human urinary proteome and particularly in the establishment of reference maps to assist in biomarker discovery. Because urine is an easily accessible, noninvasive body fluid that carries proteins, peptides, and amino acids related to kidney disease, the analysis of the urinary proteome is a potential source of information regarding the kidney's physiopathology. A clear knowledge of the protein composition of normal urine is an essential prerequisite to look at its pathology. Technological evolution in the field of proteomics (2-dimensional (2D) electrophoresis, equalization, mass spectrometry and exosomes) has greatly expanded the power of analysis, allowing the detection of almost 2,000 spots in normal urine at one time [31]. There are several reports on the use of urine analysis for diagnosis and/or prognosis in the following diseases: bladder cancer, acute inflammation due to urogenital diseases [32, 33]. Some of the urine protein biomarker studies in lupus were list in table 2. For example, Oates *et al.* [16] analyzed urine proteins by using two-dimensional electrophoresis (2 DE) (at pI 4-7) followed by mass spectrometry (MS). They can identify α -1 acid glycoprotein, α 1 microglobulin, zinc α -2 glycoprotein, zinc α -2 glycoprotein, IgGκ light chain and α 1 microglobulin which they proposed to use those biomolecules to develop a clinical assay to predict ISN/RPS class and chronicity for patients with lupus nephritis. In other study, prostaglandin-H2-isomerase and hepcidin have been identified as activity biomarkers in LN [34], along with several overexpressed or underexpressed proteins that have also been found in urine IgA nephropathy. Decreased levels of aquaporin 2 and of inter-a-trypsin-inhibitor heavy chain 4 have also been reported as an associated marker to LN. By using an array-based proteomic, about 280 molecules were recently screened in lupus nephritis patients [35]. HMGB1 was introduced as a novel candidate biomarker in lupus nephritis [36]. The study showed that the interaction of HMGB1 with a variety of receptors, including receptor for advanced glycation end products (RAGE) and Toll-like receptors, might play a role in the pathogenesis of lupus nephritis. However, most of spot identified, about 80%, still need to be characterized. There was several reports on the use of urine analysis for diagnosis and/or prognosis in LN disease status. Importantly, urine proteins that are differentially expressed during flare resolution can be used as biomarkers of prognosis or response to therapy.

Table 2 Urine proteomic study in lupus

Year/ reference	Results (purpose biomarker)
2005 [37]	i.e. α-1 acid glycoprotein, α -1 microglobulin, zinc α -2 glycoprotein, zinc α -2 glycoprotein, IgG light chain and 1 microglobulin
2008 [38]	- Hepcidin 20 increased 4 months before renal flare - Hepcidin 25 decreased at renal flare - alpha 1-antitrypsin increased at renal flare
2012 [34]	Prostaglandin-H2-isomerase as activity biomarkers.
2013 [36]	Urine angiostatin was significantly increased in active SLE compared to inactive SLE.

A link between urine and renal tissue: a logical approach for systems biology

Since, podocytes or glomerular epithelial cells play a pivotal role in many glomerulopathies. The link between tissues in kidney and urinary sediments was demonstrated in the study of glomerulonephritis model [39]. The podocytes prevent a loss of protein through glomerular basement membrane. Another study revealed the association between vascular endothelial growth factor (VEGF) and podocyte marker (WT-1) in lupus nephritis [40]. The result showed that VEGF expression could protect podocyte cell loss and prevent proteinuria. The reduced amount of VEGF mRNA in the patients with LN was positively associated with increased numbers of urinary podocytes. They suggest that VEGF might play a crucial role in the preservation of renal function and may also serve as a useful biomarker in monitoring the progression of LN.

Biological network (Graph in molecular biology)

Since bioinformatics has increasingly shifted its focus from individual genes, proteins, into large scale, so-called omic, level such as genome, proteome, and metabolome. The inference of biological network provides a framework to model the complex events that will help to enhance the biological meaning from the interactions among these parts. Understanding these complex biological systems has become an

important task that will lead to the intensive research in disease gene identification and prediction. A biological network is any network (also called graphs) that applies to biological systems by applying the graph theory concept. Normally, network is any system which contains many sub-units that are linked together into a whole system. Biological networks come in a variety of forms. Commonly, nodes in biological networks represent biomolecules such as genes, proteins or metabolites, and edges connecting these nodes indicate functional, physical or chemical interactions between the corresponding biomolecules. Graphs play roles in three complementary areas. First, graphs provide a data structure for knowledge representation. Many types of biological knowledge representation networks exist, (e.g., protein–protein interaction (PPI) network; gene regulatory network (GRN); metabolic network (MN); gene co-expression network (GCEN)) [41]. Not many networks are characterized their complete structure content [42]. A second application of graphs is to measure relationships between biological molecules. For example, in a Yeast-Two-Hybrid screen which used to explore pair of proteins that worked together or finding a protein-DNA complex model in a chromatin immuno-precipitation experiment. The last application is in statistical modeling. For example, using graph to fit a model that describes which sets of proteins can assemble together to form a protein complex by given some data consisting of observations of pairwise interactions or of the co-precipitation of proteins. The example of biological network, protein-protein interaction, is shown in Figure 1, which illustrates the interaction of S100A8, the calcium binding protein family. Each node represents protein and edge represents interaction of protein.

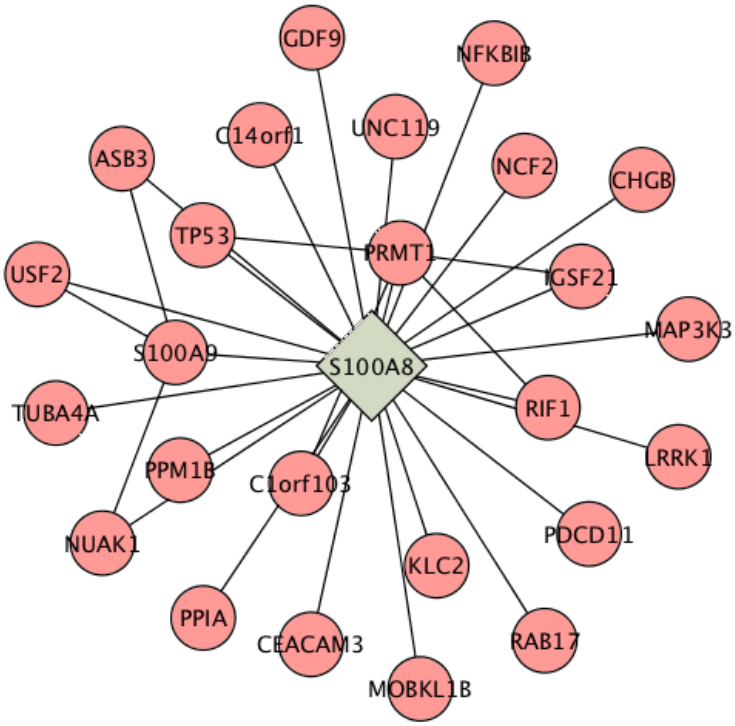


Figure 1 The protein-protein interaction network of S100A8 protein

Biologically significant sub-networks: Motifs and modules

There are many components of bio-cellular networks including genes, proteins, and other molecules are acting in collaboration to each other's to carry out specific biological processes and also with biochemical activities, by forming relatively isolated functional units called modules in molecular networks. A network motif is a significant recurring unit that become a subunit of biological modules. Elucidating the essential roles of motifs and identifying modules in molecular networks are the interests in both theoretically and biologically.

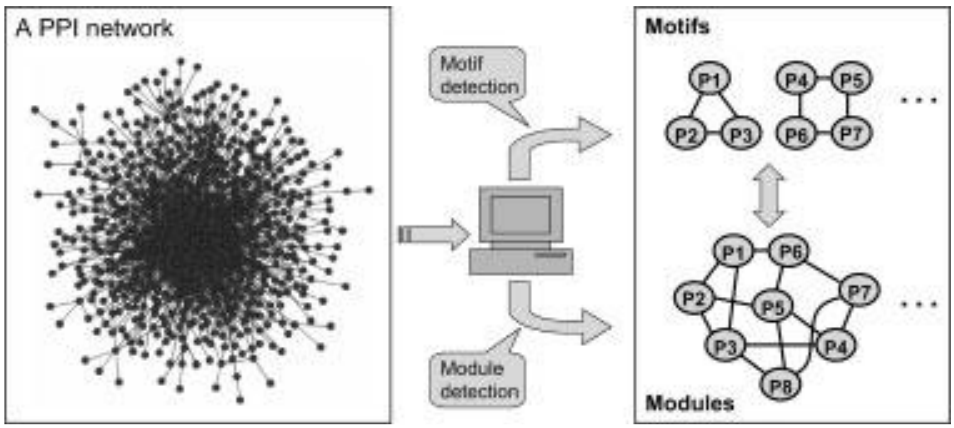


Figure 2 Motifs and modules in a PPI network (Figure is taken from [41]).

Databases of biological pathway and molecular interaction

Up to date there are several available molecular interaction and biological pathway including both commercial and academic/free of use. Table 3 shows some example of those databases. There are many successful reports that used these data sources as fundamental data for their studies.

Integrative analysis of transcriptomic and proteomic data

There have been a number of biomedical studies that integrate between transcriptomics and proteomic data. Among them, Ou and colleagues conducted a cancer biomarker study that integrates proteomic and gene expression mapping together [8]. In their study, the proteomic data were mapped to mRNA transcript database of cancer cell lines. They found novel proteomic biomarkers. Half of them were successfully validated. Interestingly, it should be noted that the proteomic data and the data in mRNA transcript database of cancer cell lines are from different sources, yet the integration still yield promising candidate biomarkers. With the data from the same cell culture, researcher can investigate the underlying molecular mechanism in human cells. Shibuya and colleagues conducted the integrative study to determine the genes and proteins of human RPE cells that are altered by exposure to TFPI-2 [9]. The transcriptomic and proteomic data were integrated using data mining. By integrating both transcriptomic and proteomic data together, they found the potential mechanism of gene-protein association with the growth-promoting effect of TFPI-2 on the human RPE cells. Mc Redmond *et al.* performed a qualitative correlation of the platelet

transcription profile with proteomic data. The result showed a set of 82 proteins secreted by thrombin-activated platelets in there analysis [43]. They found the presence of transcript still appears to be associated with presence of protein. Around 70% of these proteins were represented on the microarray. There is a clear discrepancy between the numbers of expressed genes identified by transcriptomic and confirmation of the protein found in different platforms.

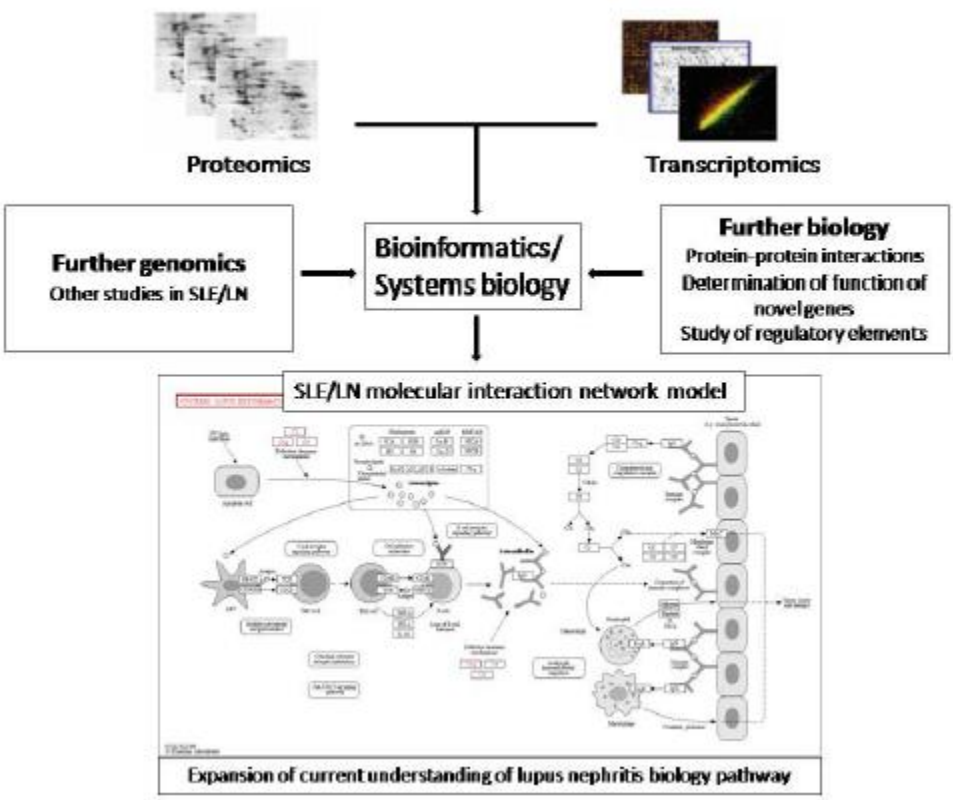


Figure 3 Integration of proteome and transcriptome data. Including with other genomic and biological information, this will lead to the expansion of the SLE/LN understanding based on the current disease model database.

Table 3 List of some biological pathway and molecular interaction related resources.

Type of database	Availability	Reference
Protein-Protein Interactions		
BIND - Biomolecular Interaction Network Database	Free	[44]
BioGRID - Biological General Repository for Interaction Datasets	Academic	[45]
DIP - Database of Interacting Proteins	Academic	[46]
GO - Gene Ontology	Free	[47]
HAPPI - Human Annotated and Predicted Protein Interaction Database	Free	[48]
IntAct–IntAct	Free	[49]
MiMI - Michigan Molecular Interactions	Free	[50]
MINT - Molecular Interaction Database	Free	[51]
MIPS-MPPI - MIPS Mammalian Protein-Protein Interaction Database	Free	[52]
Pathways Knowledge Base - Ingenuity Pathways Knowledge Base	Commercial	[53]
Signaling Gateway - UCSD-Nature Signaling Gateway Molecule Pages	Free	[54]
Metabolic Pathways		
BioCyc - BioCyc Knowledge Library	Academic	[55]
KEGG - Kyoto Encyclopedia of Genes and Genomes	Academic	[56]
MetaCore - MetaCore pathway database	Commercial	[57]
MetaCyc - Metabolic Pathway Database	Academic	[55]
NCBI BioSystems - NCBI BioSystems	Free	
Pathways Knowledge Base - Ingenuity Pathways Knowledge Base	Commercial	[53]
Reactome - ReactomeKnowledgeBase	Free	[58]
WikiPathways - WikiPathways	Free	[59]
Signaling Pathways		
BioModels - BioModels Database	Free	[60]
GeneNet - Genetic Networks	Free	[61]
GO - Gene Ontology	Free	[62]
iPath - Invitrogen iPath	Free	[63]
PANTHER - Protein ANalysisTHrough Evolutionary Relationships	Free	[64]
Reactome - ReactomeKnowledgeBase	Free	[58]
Pathway Diagrams		
BioCarta - BioCarta Pathway Diagrams	Free	
KEGG - Kyoto Encyclopedia of Genes and Genomes	Academic	[65]
MiMI - Michigan Molecular Interactions	Free	[50]
INOH - Integrating Network Objects with Hierarchies	Free	[66]
PANTHER - Protein ANalysisTHrough Evolutionary Relationships	Free	[64]
WikiPathways - WikiPathways	Free	[67]

Table 3. cont.

Type of database	Availibility	Reference
Transcription Factors / Gene Regulatory Networks		
CPDB - ConsensusPathDB	Academic	[68]
miRBase - microRNA Database	Free	[69]
JASPAR - JASPAR Transcription Factor Binding Profile Database	Free	[70]
MAPPER – MAPPER	Academic	[71]
TRANSFAC - Transcription Factor Database	Commercial	[72]
Genetic Interaction Networks		
BIND - Biomolecular Interaction Network Database	Free	[44]
BioGRID - Biological General Repository for Interaction Datasets	Academic	[45]
GeneNet - Genetic Networks	Free	[61]
Other		
iHOP - Information Hyperlinked Over Proteins	Free	[73]
PRID - Protein-RNA Interaction Database	Free	
MedGene–MedGene	Academic	[74]

A combined analysis of published transcriptomic and proteomic datasets could lead to the identification of additional novel proteins present in disease pathology. However, such an analysis is not trivial, as data can be in different formats and access to raw data can be restricted. In addition, there was not many large-scale analyses including the integrative studies in the LN especially in the identification of new biomarker for monitoring the LN therapy manner. Since, the transcriptomic and proteomic study in the treatment response in LN were perform recently in Thai cohort (in-house data). Based on the assumption, as the essential proteins tend to cluster in densely connected sub-networks with other proteins that are involved in the same biological process. The connections between the properties of single interactions and those of the whole biological network, which more directly gives rise to the phenotype [10]. These makes possible to reveal more comprehensive view on the molecular mechanism level in LN by using integration approach. Differentially regulated transcripts and proteins were mapped to their respective NCBI Gene Symbols for aligning the transcriptomic and proteomic name spaces. In the first analysis step, those features present in both the transcriptomic and proteomic lists were identified. In successive analyses, the overlap of lists was interpreted on the level of functional

annotation, molecular pathways and protein dependency networks. The results of network analysis provide novel hypotheses for functional pathway involved in disease pathogenesis and the candidate network can be effectively used to identify plausible underlying cellular mechanisms of given candidates biomarkers from a genomics study.

Epigenetics and SLE

Even the pathogenesis of systemic lupus erythematosus (SLE) is incompletely understood. Many studies indicate a clear role for epigenetic defects in the pathogenesis of lupus and other autoimmune diseases, particularly DNA methylation [12, 75, 76]. The on-going research findings in the relationship of abnormal DNA methylation and SLE is still consider interesting worldwide in both global and gene-specific analysis. Study roles of aberrant DNA methylation in the initiation and development of SLE will provide an insight into the related diagnosis biomarkers and therapeutic options in SLE. The imbalance of DNA methylation are commonly found in the Interspersed Repetitive Sequences (IRSs) region [77] which Human genome contains approximately 45% of IRSs [78].

Role of DNA methylation in SLE

From independent studies, it found that leukocytes, PBMCs, T cell, CD4+ T lymphocyte of SLE patient is globally hypomethylation comparing to normal group [79-82]. The evidence of the role of demethylation in development of SLE comes from studies with demethylating agents was published. They showed that treating T cell with demethylating agents (procainamide, hydralazine, or 5-azacytidine) induces major histocompatibility complex4 specific T cell autoreactivity [83, 84]. Furthermore, demethylation and overexpression of LFA-1, PRF1, CD70 (TNFSF7), and CD40 ligand (TNFSF5) were also observed in SLE patients [85-87]. It was suggested that these genes are standing out as importance genes affected by hypomethylation.

Human endogenous retroviruses (HERVs)

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 line [88]. 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 [89]. 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 [90, 91]. 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.

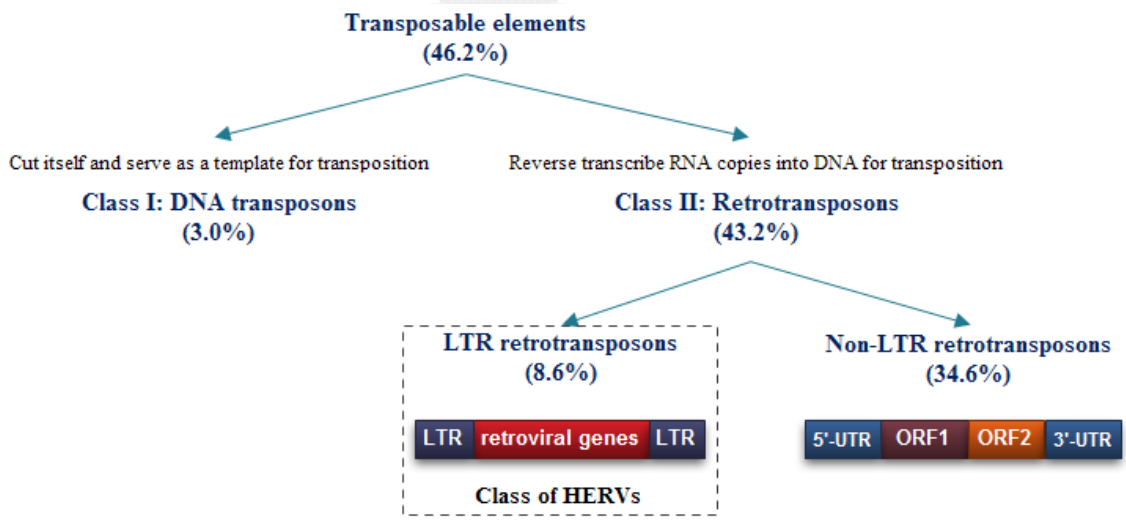


Figure 4 Classification of transposable elements (adopted from [92])

Generally, the transposable elements are DNA sequences that are able to move around and integrate into new sites within the genome [93]. As shown in Figure 4, 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.

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 [90, 94].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 5. 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 [90, 94, 95].

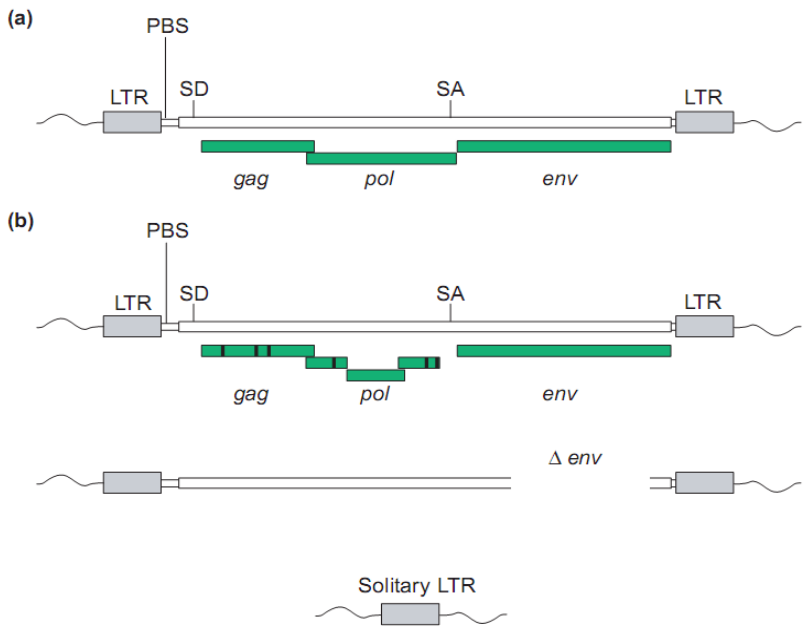


Figure 5 Genomic structures of retroviral proviruses (a) and HERVs (b) [90]

Classification of HERVs

HERVs are 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 designation 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 [91]. 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 [96]. 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 [94].

Another classification system is Repbase [97], 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 [96]. 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 [98]. This method seems to be more useful for the classification of the HERVs [94]. 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[98]. 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 [90, 92]. 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 mammy tumor virus (MMTV). Lastly, Class III HERVs, also termed ERVL superfamily, are distantly related to spumaretroviruses [90, 91]. 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 [99]. 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 [100]. 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 [90].

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.

1. Cellular functions: HERVs have been accumulated a number of mutations along the time resulting in the fact 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 [101]. The retroviral products can be beneficial for the humans. The best example is the physiological roles of some HERVs in the host. HERV-W and HERV-FRD envelope (*env*) proteins, 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 [95, 102]. 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 [94]. 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. 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 6). 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 6). 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 [95]. 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 [103].

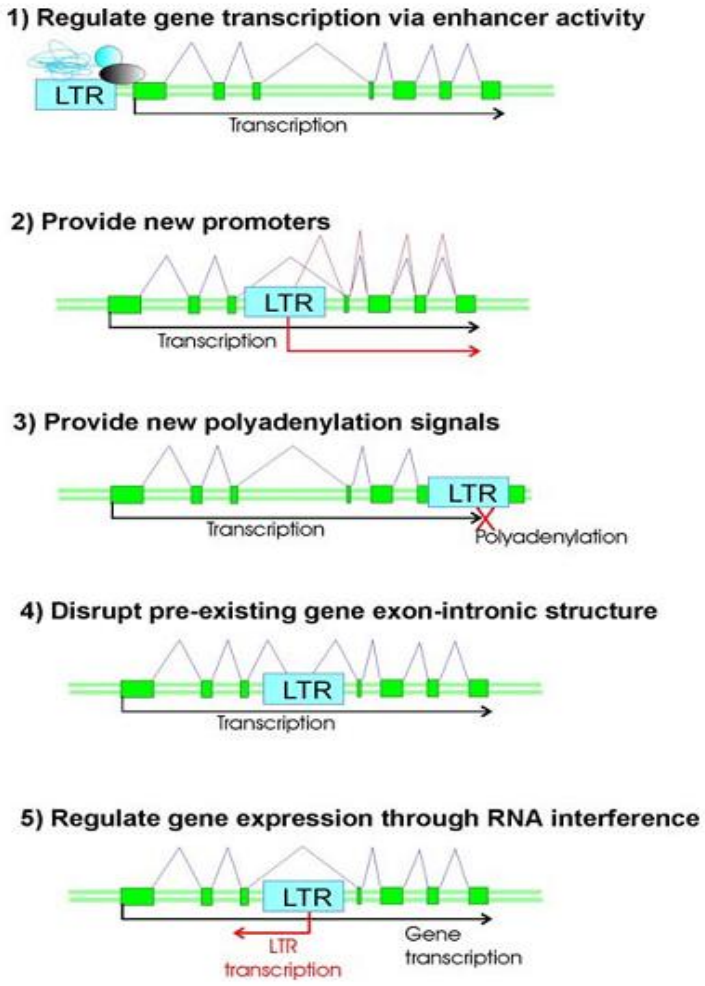


Figure 6 Five potential mechanisms of the HERVs modulating the expression of the neighboring genes [103]. An HERV element is indicated as LTR box in the figure.

Similar to 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 4.

Table 4 List of example human genes affected by HERV regulatory sequences

HERV regulator	HERV name	Gene name	Reference
1. Enhancer	ERV9 LTR	β-globin locus	[104]
	HERV-E	Amy1 (salivary amylase)	[94]
	HERV-L LTR	β1,3-galactosyltransferase 5	[105]
2. Promoter	HERV-E	APOCI (apolipoprotein CI)	[106]
	HERV-H LTR	DSCR4 and DSCR8 (Down syndrome critical region)	[107]
	HERV-E	EDNRB (endothelin receptor B)	[106]
	HERV-H	NAIP (neuronal apoptosis inhibitory protein)	[94]
	ERV9 LTR	ZNF80 (zinc finger protein)	[108]
3.Polyadenylation signals	HERV-K (KML2) LTR	LEPR (human leptin receptor)	[94]
4. Splice sites	HERV-H	PLA2L (phospholipase A2-like)	[109]
5. Antisense regulators	LTR91	CEBZ	[103]

Evidences linking HERVs to SLE

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 [110]. 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 [111]. One important 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 [112]. 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 [111]. 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[113]. 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 [114]. 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 [115]. 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 [116]. 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 [112].

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 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[113]. 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[117].

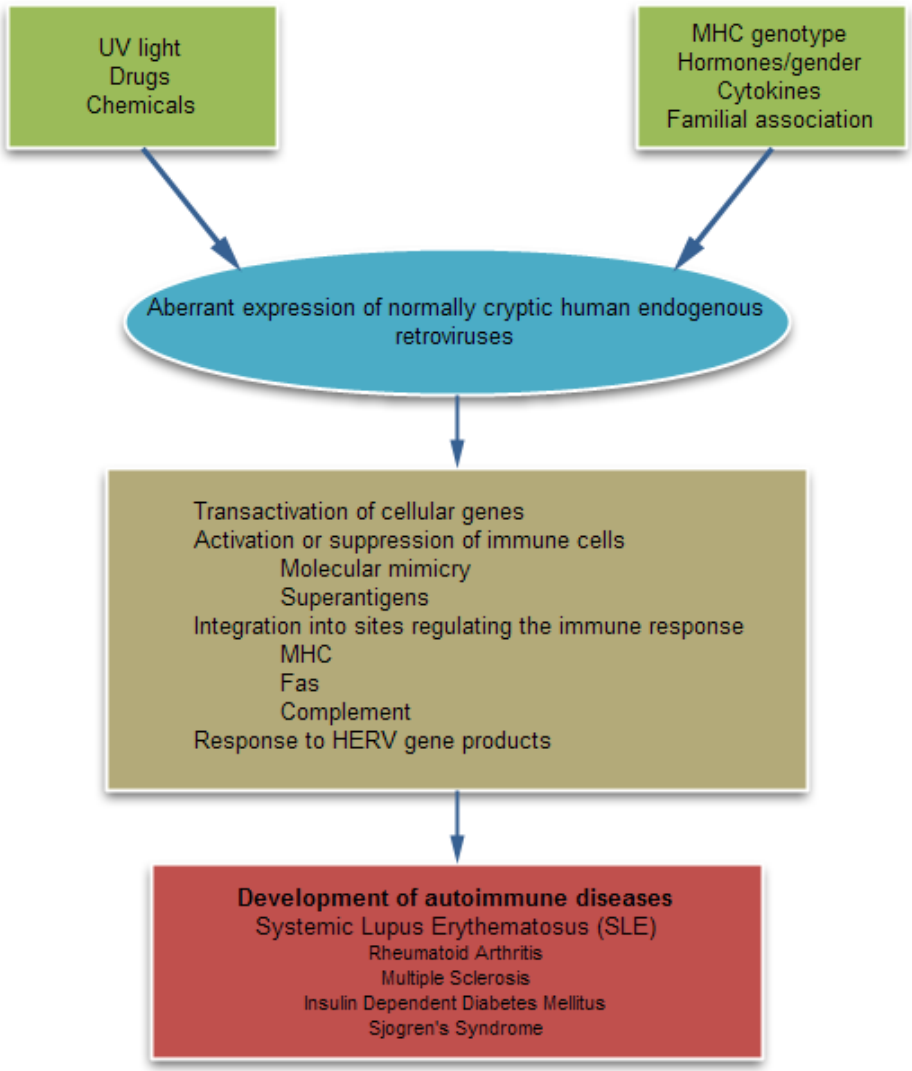


Figure 7 Summarization of purposed mechanisms used by human endogenous retroviruses in the etiopathogenesis of SLE and other autoimmune diseases[112]

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. 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)[97].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[118]. RU is available online for searching and downloading at http://www.girinst.org/ Repbase_Update.html (last viewed on December 19, 2011). 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[119]. 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 [120]. 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 [120]. There are several tools and scripts provided for this purpose, such as Process Repeats, LTR_MINER, Transposon Cluster Finder (TCF), MATCHER, and REannotate [121].

Enrichment analysis

One of a common problem in functional genomic studies is to detect significant enrichments of functional annotation in biological meaning category groups such as Gene Ontology (GO) in set of interesting genes which mostly are the group of significantly differentially expressed genes (DEG) [122]. The correct statement of enrichment testing problem leads to a unique exact null distribution of the number of DEG belonging to the GO category of interest, given the total gene number and the total number of genes belonging to the GO category. This concept was applied to calculate the enrichment of HERV in interested DE genes comparing to the number of HERV in the whole genome level.

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 [123]. Therefore, the Fisher's test 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 contingency tables (Table 5) [124].

Table 5 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 5, 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

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

Next Generation Sequencing technology (NGS)

Deep sequencing or Next Generation Sequencing (NGS) became the main application for complex biological research instead of Sanger sequencing (SS) today. NGS deliver manifold increases in sequencing throughput due to a parallel sequencing design. Millions of amplicons are generated simultaneously. Innovative NGS sample preparation and data analysis options enable a broad range of applications. Such as the huge number of whole genome sequences, deep target sequencing, RNA sequencing (RNA-Seq) to discover novel transcript forms, or precisely analysis in mRNAs quantity measurement, the analysis of whole genome methylation or DNA-protein interactions (ChIP-seq), and metagenomic which allows us to study microbial diversity in humans or in the environment in one single experiment.

Current NGS sequencers produce hundreds of gigabytes of raw sequence information that is subject to direct analysis. Considering, for example, Illumina® HiSeq 2500 sequencer, it outputs more than 4 billion 125-base reads per lane (www.illumina.com). With this sequencing performance, it leads to a major challenge for biologist to handle the massive amount of information. Data analysis is commonly handled by freely available software under the Linux environment such as the Burrows-Wheeler Aligner (BWA) [125], SOAP/SOAP2 [126, 127] alignment tools, Bowtie/TopHat [128] that allows mapping of splice junctions in RNA-Seq or Trinity [129] that allows user to reconstruct a full-length transcriptome without a genome from RNA-Seq data. Including with, the Galaxy platform [130] provides an easy accessible way to handle and analyze NGS data. Various genome alignment visualization tools also available, for example, the Integrative Genomics Viewer (IGV) [131] or through the UCSC Genome Browser [132].

CHAPTER III

MATERIALS AND METHODS

Integrative approach

Data sources

The transcriptome data of refractory lupus nephritis kidney biopsies which considered as the source of biological process of refractory lupus nephritis were integrated with the in house LN urine proteomics data since urine is considered as a logical approach of kidney diseases detection as it is secreted from diseased kidneys. Then the graph clustering method were applied to identify the integrated LN markers. The concept figure is shown in Figure 8. By using the bioinformatics functional analysis, the integrative approach will help to reveal the underlying refractory LN biological mechanism as well.

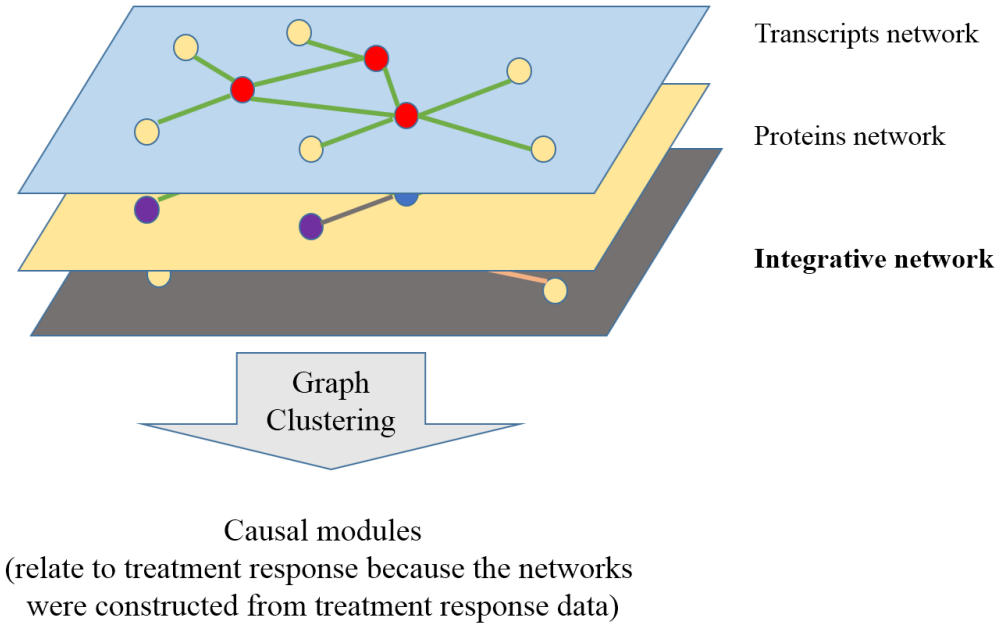


Figure 8 The integrated approach concept.

Transcript data

The candidate biomarkers of refractory lupus nephritis transcripts were discovered by using the Illumina® HumanHT-12 v4 expression BeadChip [5]. The list of differentially expressed genes were used in this study. Patients who was recruited in the experiment fulfilled the revised American College of Rheumatology criteria for SLE and criteria for lupus nephritis according to [5]. Renal involvement was documented by having one of the following criteria: 1) a total urinary protein level of more than 0.5 g/day, 2) an increment of serum creatinine levels of more than 0.5 mg/dl during 1 month period of follow-up, or 3) presence of pyuria, hematuria, or urinary cast by microscopic examination. Before medication treated, active LN patients were performed renal biopsies as routine protocol. All of renal tissue section were collected and divided to two parts. Firstly, frozen tissues were kept on ice and immediately transfer to pathology laboratory for histological diagnosis. Secondly, the section were transferred into RNA*later*® (Ambion Inc., Austin, TX, USA) solution and stored at -80°c until performed RNA extraction. All patients were treated with standard therapy, mycophenolate mofetil (MMF) or intravenous cyclophosphamide plus prednisone, by 6-month course, and were also followed and collected their clinical data. The therapeutic response was defined either by the improvement of pathological scores of activity and chronicity based on repeated kidney biopsies, or by the following clinical criteria, including: 1) stabilization or improvement in renal function; 2) X50% decrease in hematuria to less than 10 RBC per high-power field; and 3) significant reduction in proteinuria (X50% decrease to less than 3 g/day if baseline nephrotic range, p1 g/day if baseline non-nephrotic) for at least 3 months.

Proteomics data sources

Biomarker discovery utilized protein profiles by 1 dimension gel electrophoresis were perform on urine sample of response and non-response patients which collected at the day of kidney biopsy. Patients were treated with standard lupus nephritis treatment. Then the patients were classified as responder and non-responder at the sixth month of follow up as described in previous study [5]. The protein profiles were also performed for patients at sixth month of follow up as well. Only patients that

had biopsy-proven class III/IV LN were included in this experiment. Response to therapy was defined as the follows. 1) stabilization or improvement in the renal function; 2) X50% decrease in hematuria to less than 10 RBC per high-power field; and 3) significant drop of proteinuria (X50% decreased to less than 3 g/day if baseline was nephrotic range or <1 g/day if the baseline was non-nephrotic) for at least 3 months.

Urine samples were collected from normal healthy individuals (who had no medication and no menstrual blood) and LN patients (inactive, active disease). The urine sample was centrifuged at 1000g, 4 ˚C for 5 min to clear debris. The supernatant was precipitated by 75% ethanol and mixed. Samples were incubated on ice for 10 min following by centrifugation at 12,000g for five minute. The pellet was resuspended in a sample buffer containing 7 M urea, 2 M thiourea, 4% CHAPS (3[(3-cholamidopropyl) dimethylammonio]-1-propanesulfonate), and 30 mM Tris-base. The suspension samples were kept at -80 ºC until use. Mid-stream urine (≈25 mL) was precipitated by 75% ethanol, and then resuspended with 30 mM Tris-HCl (pH 8.5) buffer containing 2 M Thio urea, 7 M Urea and 4% CHAPS on ice. The protein suspension were subsequently desalted and concentrated by 0.3 μM VivaSpin (GE Healthcare Life Sciences). In order to separate protein by its molecular weight, the protein were loaded into 12% polyacrylamide gel electrophoresis (PAGE) and run at 100 V for 2 hours. The gel were fixed and stained with Coomassie Brilliant Blue R-250. In consequent, protein were prepared for step In-Gel digestion. Next, the gel were destained with 25% Sodium Bicarbonate with 50% acetonitrite (Merck), in consequent, the gel were digested with 12.5 μg/mL Gold Trypsin (Promega) at 37℃. The peptide were solubilized in the solution while residual peptide in gel were sonicated to ensure that there were entire peptide in suspension. In the following, peptides were dried by speed vacuum and stored at -80℃. Then the peptide were resuspended with 0.1% formic acid, then injected into Mass-spectrometer.

Integration method

We applied two integrative approaches for discovering refractor LN biomarkers in our study. For the first analysis approach, transcriptome and proteome data were directly mapped with their gene/protein names to reveal the direct overlapping relationship on these two different levels of expression in refractory LN. Unfortunately, it seems like the overlap of transcriptomic and proteomic features is low and ambivalent as described in [133]. Therefore in the second approach, we used PPI properties to connect transcripts and proteins for finding underlying relationship on those two difference expression levels as the indirect overlapping network. For transcription data, if there are multiple probe sets corresponding to the same gene were occurred we selected only one probe that give the highest statistical t-test significant value as a representative probe of that genes.

The first approach explores the mapping of differentially expressed genes and proteins to the well-characterized biological pathways. The GenMAPP-CS [134], an open-source software for pathway visualization and analysis was used to integrate the transcriptome and proteome data. At first, transcriptome and proteome expression data were imported into GenMAPP-CS then mapped these data to the existing biological pathways, including KEGG [65] (Kyoto Encyclopedia of Genes and Genomes, available at http://www.genome.jp/kegg/) and WikiPathways [59] (available at http://wikipathways.org/). Candidate biomarkers were then identified based on the obtained functional annotations and their biological relations to lupus nephritis.

In the second approach, gene-gene, gene-protein, and protein-protein interactions information were used for constructing the integrative network then identify the sub-network (module) based on the topology of the networks. Cytoscape, an open-source software for integration, visualization and analysis of biological networks were used to integrate these different layers of interaction networks together [135]. First, the transcriptional and protein networks were generated based on the protein-protein interaction (PPI) information which PPI was retrieved from BIND [136] and IntAct [137] databases through MiMi plugin [138]. Transcriptional and protein network were then integrated together based on common Entrez ID. After the network was constructed, graph-clustering algorithm was used to extract the sub-networks based on network topology. In this study, 6 clustering methods were used for comparing and

evaluating according to the review protocol that different network clustering methods can yield quite different network modules from the same data. The plugin that were used in the analysis are MCODE, MINE, NeMo, ClusterONE, APCluster, and ClusterExplorer [139, 140]. All clustering algorithms are online available at http://apps.cytoscape.org/ apps/. Then the candidate biomarkers were identified based on the highly connected molecules in sub-networks. Additionally, these highly connected molecules were also investigated by setting a minimum cut-off connected molecules to 5 molecules which at least 5 connected. Finally, the functions of each sub-network module can be inferred by identifying the enriched functions of its gene

Gene ontology and functional analysis

Differentially expressed genes and proteins will identify their enriched biological process to reveal common biological process for using in the integration procedure. GO [62] provides three structured, controlled vocabulary (ontology) to describe gene and gene product attributes in any organism, in terms of their associated biological processes, cellular components and molecular functions. Enriched biological processes will identify using the PANTHER (Protein Analysis Through Evolutionary Relationships) Classification System [141]. The two main categories ‘‘biological process’’ and “molecular function” will carry on for the analysis procedure. Other functional analysis will also analyze for mere comprehensive view in the disease mechanism by using DAVID (Database for Annotation, Visualization, and Integrated Discovery) tool [142], which provides gene-specific functional data mining tools and methods for functional category analysis.

Integration and clustering analysis software

Cytoscape [143] an open source bioinformatics software platform for visualizing molecular interaction networks and integrating with gene expression profiles and other state data. Cytoscape can be used to visualize and analyze network graphs of any kind involving nodes and edges. A key aspect of the software architecture of Cytoscape is the use of plugins for specialized features which are developed by many research groups.

Human Endogenous Retrovirus analysis procedure

There is no database facilitating the HERV neighboring genes available up to date. By utilized the genome information from Repbase [144], it will help in the investigation of effect of HERV to their nearby genes. Moreover, this study also facilitates the enrichment analysis for investigate the association between certain type of HERV and the certain expression pattern of their neighboring genes. Not only the association between HERV and gene expression were investigated in this study, but we also examined the RNA-Seq data for discovering HERV-gene chimeric transcripts in this study as well.

Analysis and programming tools

REannotate [121] is a computational tool that performs post-processing of repeat annotation results generated by RepeatMasker, a program that computationally detects interspersed repeats and low complexity DNA sequences [145]. The post-processing is required to improve the biological interpretation of the RepeatMasker annotations because the annotated sequences often correspond to fragments of the repetitive elements resulting from accumulation of insertions and deletions [120]. REannotatecan automatically perform main three tasks, including defragmentation of the dispersed repetitive elements, resolution of the temporal order of the elements' insertions in the nested clusters, and forecasting the age of the elements after the insertion time [121]. In this work, REannotate was employed to defragment the HERV annotations. Furthermore, there are some beneficial measurements additionally calculated by REannotate. These measurements would be kept as additional characteristics of HERVs.

Python is a high-level programming language that can be easily applied to many different problems and integrated to a system more efficiently [146]. It can run on a wide variety of operating systems: UNIX, Windows, Mac, and so on. Python is one programming language that is being used more and more in bioinformatics works. Also, there is an available tool that can facilitate the biological computation that is written in Python called Biopython.

R is a programming language and software environment for statistical computing and graphics. It is an open source under the terms of the Free Software

Foundation's GNU General Public License. There are a wide variety of statistical and graphical techniques provided in R, such as, classification, clustering, time-series analysis, linear and nonlinear modeling, and so on. It can compile and run on a various UNIX platforms, Windows, and MacOS [147].

MySQL database is a rational database management system or RDBMS that is the most popular today. It is freely downloadable and available as an open source software [148]. This software will use as the main database management system in this work.

Data resources

This is an annotation of all repeats, including short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), long terminal repeat elements (LTRs, which contains HERVs), DNA repeats, simple repeats, low complexity repeats, and satellite repeats, of the human reference sequence version hg19/GRCh37 (Feb., 2009). The annotation was created by using RepeatMasker [145], along with the Repbase repeat library (Release 20090120). It is a UCSC data table, named rmsk, and freely downloadable on the UCSC table browser [149]. This is also a UCSC track that shows the gene annotation resulting from the UCSCs' predictions [150]. This UCSC gene annotation table was named as knownGene table. Data from RepSeq, GenBank, CCDS and UniProt were used in the predictions. The track contains both protein-coding genes and putative non-protein coding genes. To be consistent with the repeat annotation data, the UCSC gene annotation of the hg19 human reference sequence would be used in this work. This is the central cross-reference table for the UCSC known genes [150]. The data has been collected in the table named kgXref and can be downloaded from the UCSC table browser as well. In the table, there are several cross-reference IDs for a UCSC known gene, including UCSC known gene ID, GenBank accession number, SWISS-PROT protein accession number, SWISS-PROT display ID, gene symbol, NCBI RefSeq ID, and NCBI protein accession number.

Data collection

The human repeat annotation, human gene annotation, and also the cross-reference IDs of the UCSC genes were all downloaded from the UCSC table browser [149]. In the UCSC table browser, these three data are in the UCSC table named rmsk, knownGene, and kgXref, respectively. They are all based on the February 2009 human reference sequence (hg19/GRCh37), which includes sequences of 24 main chromosomes (chromosome 1-22, X, and Y), 59 unplaced contigs, and 9 haplotype chromosomes. The haplotype chromosomes are a collection of haplotype segments, additionally generated during the genome assembly [151].

Data selection

As mentioned before, the annotation results and the cross-reference gene IDs contain not only the records based on the main placed chromosomes, but also based on the unplaced contigs and the haplotype chromosomes, which were separately generated from the main placed contigs. In this work, only the data of the main placed chromosomes would be included in further utilization, due to easier and tidier in referring to the chromosome on which a gene or a repeat is located. This step was thus performed to remove all of the unwanted records in the data resources before performing any manipulation of the data.

REannotate and determination of truncation patterns

Generally, the complete HERVs are composed of two flanking LTRs, 5’-LTRs and 3’-LTRs, as well as a set of internal sequences for the retroviral genes. However, due to the accumulations of insertions and deletions on their sequences, most of the HERVs currently present in the human are hardly found as the complete elements, but usually found as the truncated elements. Therefore, assigning the types of truncation patterns to each HERV element would facilitate us more in categorizing and referring to HERVs according to their structural truncation. In this work, the types of truncation patterns will define based on truncated parts of an element, including a 5’-LTR, an internal sequence, and a 3’-LTR. Thus, there are five types for the classification of the truncation patterns: complete, 5’-truncated, 3’-truncated, both 5’- and 3’-truncated

elements, as well as solitary or solo LTRs. The classification design is shown in Figure 9. To classify each HERV element, the result obtained from the step of the HERV defragmentation will be used for the purpose.

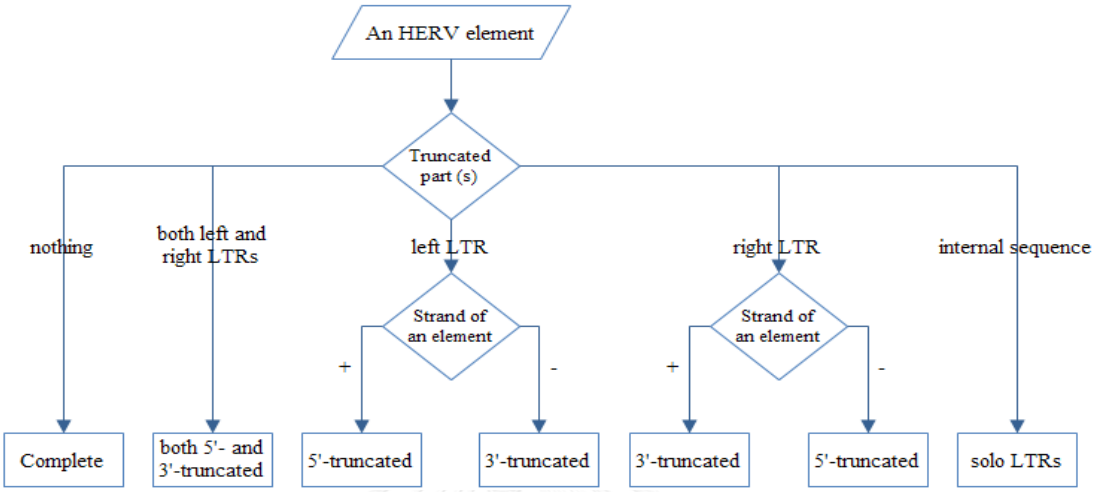


Figure 9 Truncation pattern determination process.

Mapping HERVs on the human genes

The edited REannoate output and the selected gene annotation were used. Each gene isoform were collected with their neighboring HERV elements which are located not far from the transcription start or transcription termination site more than 100,000 bps were considered as intergenic HERVs.

HERV distribution analysis

The HERV expression were detected in gene bank mRNA using blast. Five solo LTR characteristics were tested in this study which are list below,

1.) sense direction in up-stream region,

2.) sense direction in intragenic,

3.) sense direction in down-stream region,

4.) anti-sense direction in intragenic, and

5.) anti-sense direction in down-stream region.

EnHERV

As the available of CU-DREAMX [152] that allows researchers to investigate the association of specific list of genes with microarray data in GEO database. But according to limitation of the tool that are only available for Windows OS environmental. Therefore we constructed a web-based tool, EnHERV, which facilitates in the investigations of neighboring HERVs with any interested set of gene names. The System flow design shown in Figure 10. Then Fisher's exact were tested on all solo-LTR type and characteristics in various diseases in both up- and down-regulated expression pattern.

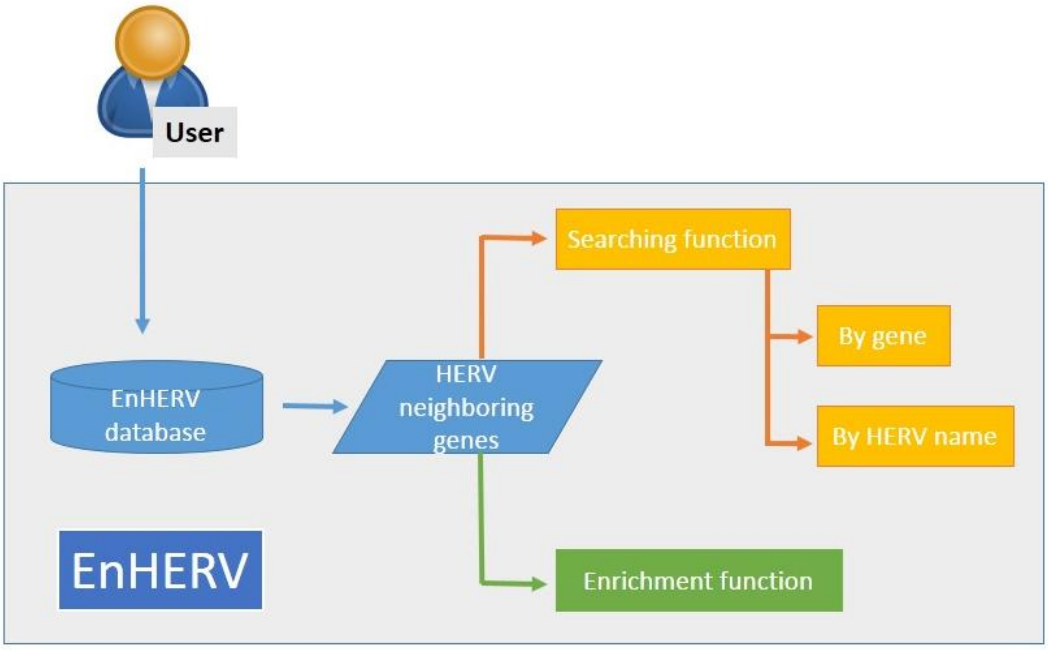


Figure 10 EnHERV diagram of system flow design

EnHERV construction

EnHERV was constructed as web database tool for user easily access. EnHERV provides 2 functions; the first function is the search function that allows user to connect to pre-built HERV profile database as described above. There are 2 different input types for accessing database which are a.) Searching by gene(s) and b.) Searching by HERV characteristics. There are total 7 searching characteristics; 1.) HERV superfamily, 2.) HERV family, 3.) HERV name, 4.) HERV orientation, 5.) HERV location in genome, 6.) HERV location in gene, and 7.) HERV completeness type. Another function is

Enrichment analysis function. This function implement Fisher's exact test for enrichment analysis test on user defined genes list and gene that contains specific HERV characteristics. Because of HERVs are mainly reports to involve in cancer and autoimmune disease, 10 different cancer and autoimmune experiments were retrieved from gene expression omnibus (GEO) [153, 154] for using as built-in pre-set gene lists in EnHERV.

Solo-LTR enrichment analysis in various disease conditions.

HERVs have been reports to be active due to hypomethylation in cancer and autoimmune diseases [18, 155, 156]. Therefore, distinct isoforms or gene silencing due to HERV were reported in various cancers [157-160] Alternative transcript of CD5 by HERV-E was also reported in SLE B cells [161]. However, there are still limited comprehensive data and tool available to analyze HERV in relation to other genes. Since we hypothesize that HERV can control neighboring genes by either up-regulation or down-regulation including with it might associate in specific direction and location. We performed the enrichment analysis in EnHERV to reveal the associated between specific HERV properties in various disease conditions. Forty nine GEO accessions were retrieved from NCBI GEO database. Then we classified these gene expression in to 56 conditions as shown in table 6 including autoimmune and other disease conditions. Differentially expressed genes were identified by using GEO2R function in GEO, available at http://www.ncbi.nlm.nih.gov/geo/geo2r/. Genes were considered as differentially expressed genes with criteria of p-value less than or equal to 0.05 and fold-change greater than 1 fold.

Table 6 List of gene expression conditions

GEO accession	Condition
GSE27011	Asthma white blood cells
GSE31773	Asthma CD4
GSE31773	Asthma CD8
GSE43696	Asthma bronchial epithelial cells
GSE71957	Graves' disease CD4
GSE71957	Graves' disease CD8
GSE1299	Breast cancer cells
GSE13911	Microsatellite instable gastric cancer
GSE2171	HIV infected PBMC
GSE2171	HIV infected PBMC cd4dec
GSE2171	HIV infected PBMC cd4inc
GSE3167	Bladder carcinoma situ
GSE5764	Ductal and lobular breast cancer cells
GSE5816	Lung adenocarcinoma
GSE6631	Head and neck cancer cells
GSE6740	HIV infected cd4 acute
GSE6740	HIV infected cd4 chronic
GSE6740	HIV infected cd4 non-pregressive
GSE6740	HIV infected cd8 acute
GSE6740	HIV infected cd8 chronic
GSE6740	HIV infected cd8 non-pregressive
GSE6919	Metastasis prostate cancer
GSE9750	Cervical cancer
GSE9764	5aza-Human Mesenchymal Stem Cells
GSE59695	H3K4me1, HepG2
GSE22859	H3K4me2, HeLa
GSE44084	H3K9, pre-iPSC
GSE25282	H3K9me3, HeLa
GSE41040	H3K9me3, primary fibroblasts
GSE32591	LN glomerular
GSE32591	LN tubular
GSE36474	Myeloma bone marrow
GSE13355	Psoriasis skin
GSE14905	Psoriasis skin
GSE32407	Psoriasis skin
GSE52471	Psoriasis skin
GSE10500	Rheumatoid arthritis macophage
GSE15573	Rheumatoid arthritis PBMC
GSE1919	Rheumatoid arthritis synovial tissues
GSE4588	Rheumatoid arthritis B
GSE4588	Rheumatoid arthritis CD4
GSE45175	SETDB1 knockdown, lung cancer cell lines

GSE73231	SETDB1 knockdown, mouse mesenchymal stem cells
GSE10325	SLE B cells
GSE10325	SLE myeloid cells
GSE10325	SLE T cells
GSE13887	SLE T cells
GSE20864	SLE PBMC
GSE24706	SLE PBMC ANA
GSE27427	SLE neutrophil
GSE4588	SLE B cells
GSE4588	SLE CD4
GSE52471	SLE/DLE, skin
GSE61635	SLE PBMC RNP+
GSE30153	SLE inactive condition, B cells
GSE61639	TRIM28 knockdown, breast cancer cells

Since, HERVs were reported to disease or condition specific association, there are 25 individual HERVs from all 4 superfamilies were used in this study which shows in Table 7. Most of them were found as expressed HERV elements in previous reported. We also tested the association between HERVs and various disease condition at superfamily and entire HERV neighboring gene as well. HERVs and disease conditions were considered as associated event using criteria as Fisher's exact p-value less than 0.001 and OR ratio more than 1.

Table 7 List of solo-LTR used in enrichment analysis

Super-family	family	Name/group
	ERV9	LTR12, LTR12C
	HERVH	LTR7
	HERVW	LTR2, LTR2B, LTR2C
	HUERSP1	LTR8
ERV1/ERVE	LOR1	LOR1a
	MER39	MER39
	MER4	MER4C
	MER52	MER52A
	MER57	MER57B1

Super-family	family	Name/group
ERV2/ERVK	HERVK10/HERVK (HML-2)	LTR5_Hs
	HERVK11/HERVK (HML-8)	MER11C
ERV3/ERVL	HERV16	LTR16A1
	HERVL33	LTR33
	HERVL52	LTR52
	HERVL66	LTR66
ERVL-MaLR	MLT	MLT1D, MLT2B3
	MST	MSTD
	THE1	THE1A, THE1B, THE1C, THE1D

Chimeric detection using RNA-Seq data

Hung T et al. conducted RNA-Seq of SLE whole blood and healthy controls to determine the gene expression changes in these patients [162]. By using the advantage of this public RNA-Seq data, this make possibility of exploring of chimeric sequences in their data. In this study 10 SLE and 3 control samples were retrieve from their accession GSE72509. Not only SLE PBMC were used in this study, but K562 cell line were also included in this study since it is hypomethylation model study. So, 3 of K562 RNA-Seq samples from GSE34740 accession [163] were included in this study.

The RNA-Seq experiments information were descript in NCBI GEO portal [154] while all raw NGS data were stored as SRA file format in SRA portal (http://www.ncbi.nlm.nih.gov/sra). The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®. For the first step in this analysis, sra files of GSE72509 and GSE34740 were retrieved from SRA portal and converted to fastq file format using SRA toolkit as descript in SRA HandBook in NBCI Bookself (accession NBK47528). Then the quality of raw NGS sequences were checked by FastQC which is one of the most popular tools for checking

the quality of raw NGS data. FastQC is available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

According to there is no single analysis pipeline can be used in the RNA-Seq analysis yet, this project used 2 different methods to analyze the raw RNA-Seq data.

1. The first approach was modified method of Trapnell C et al. which is the famous RNA-Seq analysis method for differential gene and transcript expression analysis by using TopHat and Cufflinks [164]. TopHat aligns RNA-Seq reads by using Bowtie [165] as an aligner then discovering RNA splice site by mapping the reference transcripts. The *H. sapiens* hg19 (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/) were used as references for mapping of reads. UCSC [166], RefSeq [167] and Ensembl [168] transcripts information were used in this study. Then assembled transcript were constructed using gffread function. Then assembled transcripts were used for detection of chimeric transcript in the next step. Figure 11 illustrates flow of this first approach.
2. *De novo* transcript assembly using Trinity [169] to avoid the repetitive problem during mapping process to retrieve the full range transcripts which beneficial in possible to discover any new transcripts from RNA-Seq data. Figure 12 shows the *De novo* analysis approach. The assembled transcripts were also used for detection of chimeric transcript in the next step.

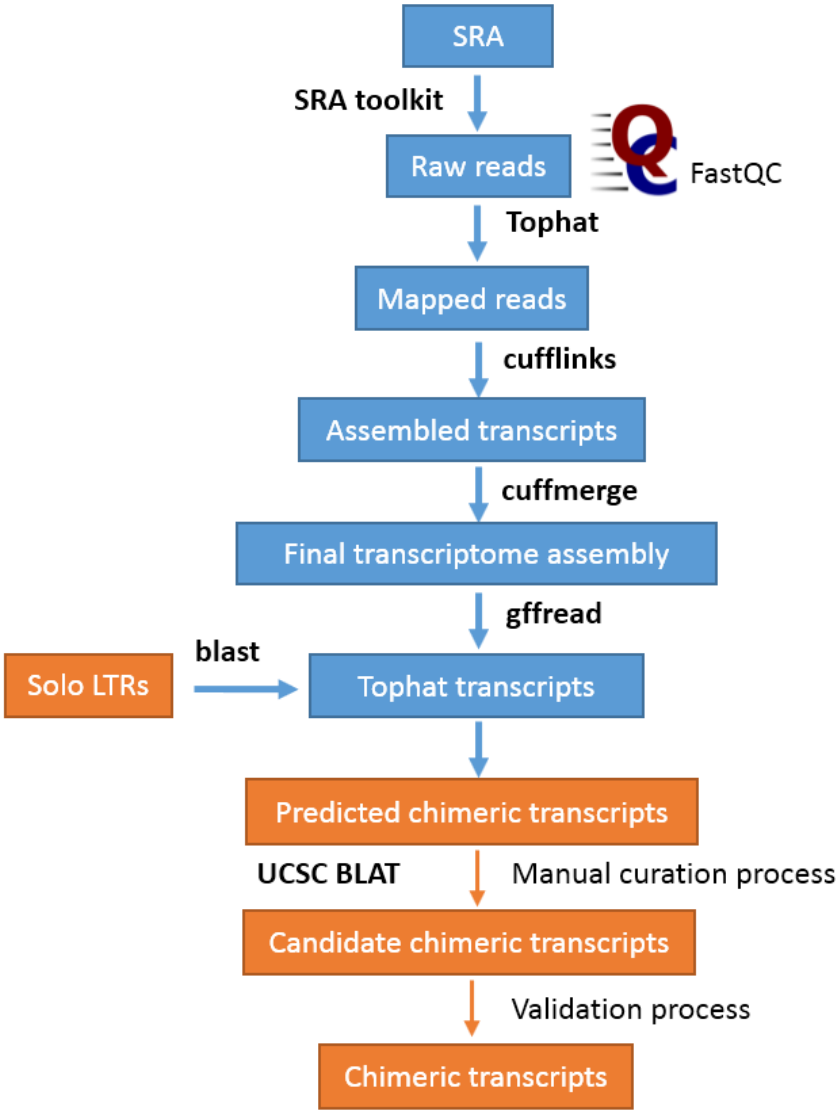


Figure 11 An overview of RNA-Seq reads mapping approach

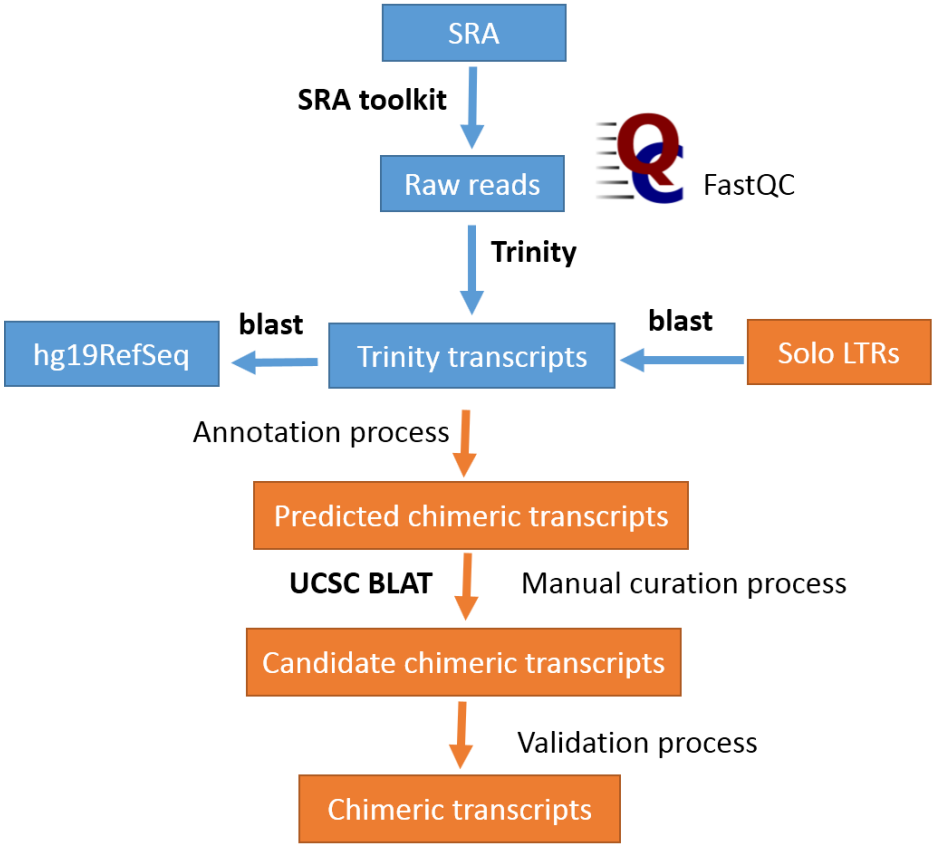


Figure 12 An overview of de novo RNA-Seq assembly approach

Identifying of chimeric sequences

NCBI-blast was used to detect chimeric sequences in assembled transcripts from both analysis approaches. List of solo-LTR that used for discovering chimeric sequences in RNA-Seq transcripts were listed in Table 8. Transcripts were consider as chimeric sequences by using following criteria; 1) gap free alignment sequence, 2) sequence identity is more than or equal to 98%, 3) less than 3 mismatch in the alignment, 4) the hit length is longer than or equal to 200 nucleotides, and 5) hit length/sequence length ratio is more than 20%. Transcripts that pass these criteria were consider as predicted chimeric transcripts which will manually occur the transcript structure and localization by BLAT [170] and visualizing on UCSC genome browser. Only transcript that chimeric transcript were aligned correctly between human exon and LTR fragment were consider as candidate chimeric transcripts.

Table 8 List of solo-LTR using as query for finding chimeric transcripts

Super-family	family	Name/group
ERV1/ERVE	ERV9	LTR12, LTR12CB, LTR12C, LTR12D, LTR12E, LTR12F
	HERVW	LTR2, LTR2B, LTR2C
	HUERSP1	LTR8
	LOR1	LOR1a
	MER39	MER39
	MER4	MER4C
	MER52	MER52A
	MER57	MER57B2
ERV2/ERVK	HERVK10/HERVK (HML-2)	LTR5_Hs
	HERVK11/HERVK (HML-8)	MER11C
	HERVK22/HERVK (HML-5)	LTR22, LTR22A, LTR22B, LTR22C
ERV3/ERVL	HERV16	LTR16A1
	HERVL52	LTR52
	HERVL66	LTR66
ERVL-MaLR	MLT	MLT1L, MLT2D
	MST	MSTD
	THE1	THE1A, THE1B, THE1C, THE1D, THE1I

Primer3 [171] was used to design PCR primer for detecting chimeric transcripts in 4 selected chimeric transcripts which are MER52A-CLEC4E, THE1C-CLEC2D, LTR5B-TOP3A, and THE1C-IFI44. Primers were also tested with UCSC *In-Silico* PCR before used in this analysis. We then confirming these chimeric that meet the expected amplicons size with Sanger sequencing platform (1st BASE Pte Ltd, Singapore).

CHAPTER IV

RESULTS AND DISCUSSION

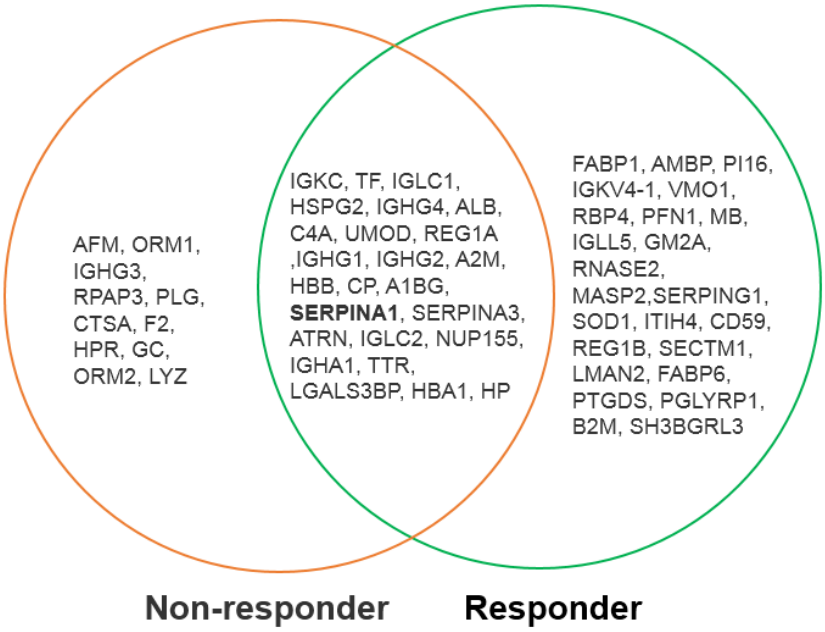
Integrative approach for LN biomarker discovery

LN Transcript data; the number of differentially expressed genes were list in Table 9. There were interesting gene proposed as potential biomarker for predicting a non-responder in our reported which are tight junction gene (claudin), B-lymphocyte stimulating factors (BAFF, APRIL). Moreover, a loss of kidney function might be predicted by set of genes such as complement pathway (SERPINA) or ANXA13.

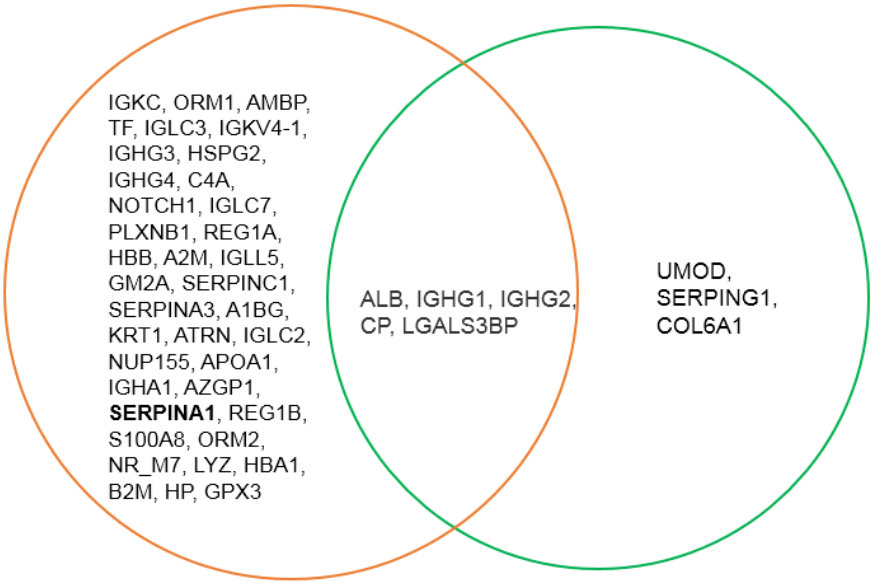
Table 9 Number of differentially expressed probes and genes

	Up-regulated	Down-regulated
Significant analysis		
Differentially expressed probes	442	374
Differentially expressed genes	396	353

LN proteomics data; the list of differentially expressed proteins between lupus nephritis treatment responder and non-responder were illustrated in Figure 13. We have to note that there are just only one sample for responder and non-responder were investigated in this study. So we purpose this analysis result as the discovering profile procedure, there is no statistical calculation applied in this study. The result show that there are 68 proteins were identified in responder while only 43 proteins were discovered in non-responders. There were interesting gene sets that could predict a non-responder including tight junction gene (claudin), B-lymphocyte stimulating factors (BAFF, APRIL). Moreover, a loss of kidney function might be predicted by set of genes such as complement pathway (SERPINA) or ANXA13. Figure 13 illustrated the list of different refractory urine proteins between responder and non-responder. Figure 13a displayed list of urine protein at biopsies time (first month of follow up and Figure 13b listed the urine proteins at sixth month of follow up. Interestingly, the number of responder urine protein, including with SERPINA1, was dramatically decrease at the sixth month of follow up. It might result of the good response to the LN treatment.



13a.



13b.

Figure 13 List of differential expressed urine protein in refractory LN 13a.) List of different proteins at first month and 13b.) List of different proteins at 6 month follow up

All differentially expressed transcripts and proteins were mapped to the characterized molecular pathways. The KEGG was used as the background pathway for this approach. In total, 815 genes and 35 proteins were identified as differentially expressed molecules in transcription and present proteins in refractory LN, respectively. The analysis result indicates 5 over expressed transcripts which 3 were present in refractory LN responders while 2 were found in non-responders. Simultaneously 2 under expressed transcripts were found in urine proteins which 1 common protein between responder and non-responder as listed in Figure 14.

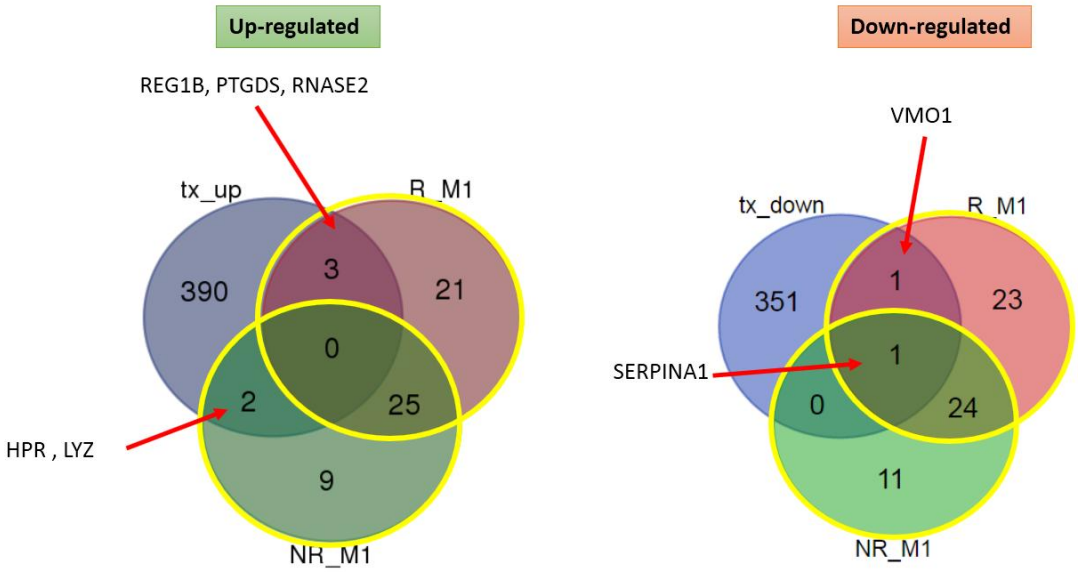


Figure 14 List of integrated kidney biopsy transcripts and urine protein at different expression occurred at the first month in both up- and down-regulated transcripts. The yellow circles indicate proteomic analysis results. (tx_up; up-regulated transcript, R_M1; protein profile of responder at biopsy time, NR_M1; protein profile of non-responder at biopsy time, tx_down; down-regulated transcript.)

REG1B, The related REG1 protein is associated with islet cell regeneration and diabetogenesis, and may be involved in pancreatic lithogenesis. This gene encodes a protein that is secreted by the exocrine pancreas. RNASE2, The protein encoded by this gene is a non-secretory ribonuclease that belongs to the pancreatic ribonuclease family. GO annotations related to this gene include nucleic acid binding and ribonuclease activity. The protein antimicrobial activity against viruses. PTGDS (prostaglandin-H2 D-isomerase) was illustrate as part of the arachidonic acid metabolism [PATH:ko00590], which illustrated in Figure 15, under Lipid metabolism pathway in KEGG. In the meanwhile, there is no arachidonic acid metabolism pathway reported in WikiPathways. Anyway there are two PTGDS involved pathways identified in WikiPathways. There are prostaglandin synthesis and regulation, and eicosanoid synthesis. Since we have reported that PTGDS as a biomarker for active lupus nephritis in proteomic aim, this molecule also involves in the eicosanoid synthesis in peripheral blood monocytes that also was reported as a marker of disease activity in lupus nephritis [24]. The involvement of PTGDS in these two molecular pathways which are prostaglandin synthesis and regulation and eicosanoid synthesis pathways as illustrated by WikiPathways in Figure 17 and 18, respectively. These finding might note the important role of PTGDS in refractory LN mechanism based on the consistency of its expression in term of both mRNA and protein expression level. While SERPINA1 and VMO1 were found as corresponding molecules with under expressed transcripts. According to the follow up urine proteomic screening at first and sixth mounts, SERPINA1 found in non-responder urine in both first month and at the sixth month while it absent in the sixth month follow of responder. This might help in tracking of protein leaking of LN patient as well. VMO1, Vitelline Membrane Outer Layer 1 Homolog is a Protein Coding gene. It found as extracellular exosome proteins. SERPINA1; serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1 was illustrated in Figure 16 as part of the complement and coagulation cascades (PATH:ko04610) in both the immune system pathway of KEGG and WikiPathways. This finding is support that complement is implicated in the pathogenesis of systemic lupus erythematosus (SLE) [23] and might be used as one of the candidate biomarker network for therapeutic response of LN.

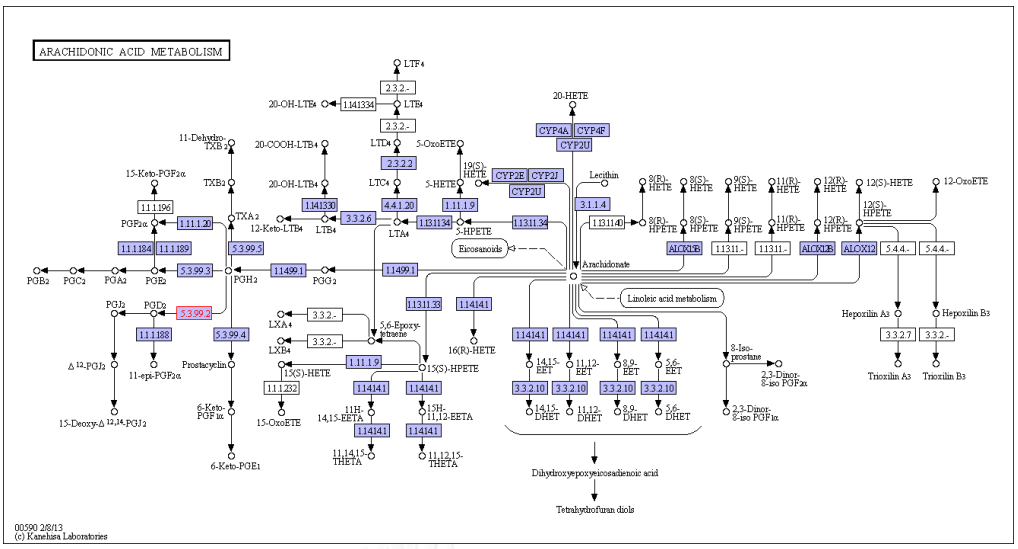


Figure 15 The KEGG arachidonic acid metabolism [PATH:ko00590] pathway. The PTGDS is illustrated using enzyme number EC 5.3.99.2 in red box.

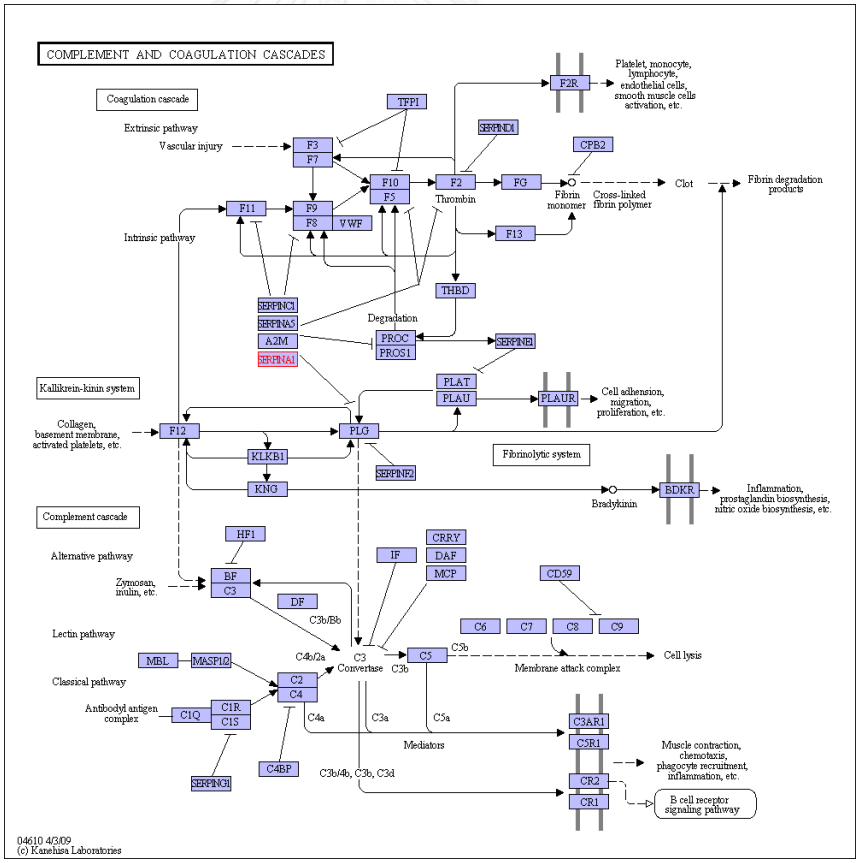


Figure 16 The KEGG complement and coagulation cascades (PATH:ko04610). The SERPINA1 is illustrated in the red box.

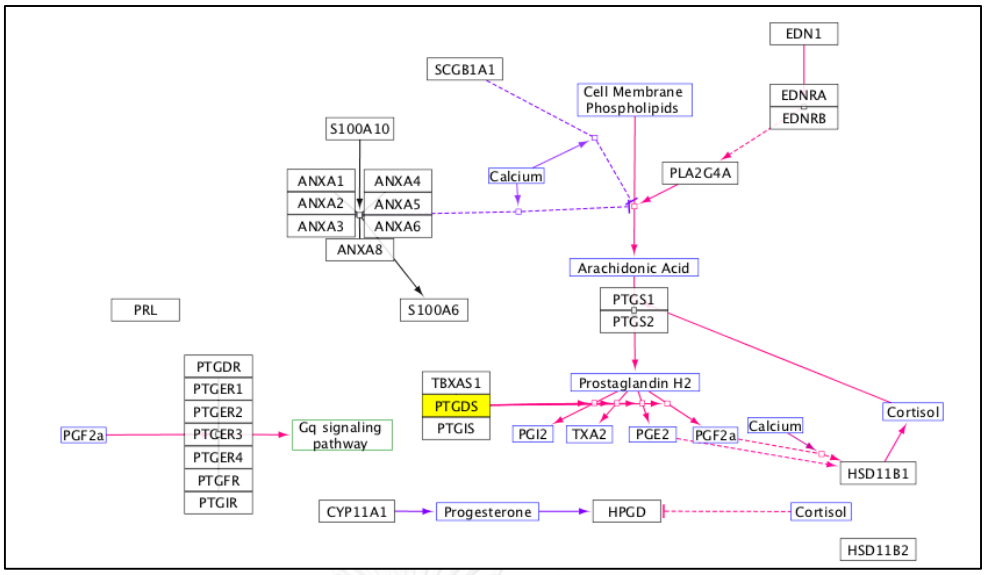


Figure 17 Prostaglandin synthesis and regulation pathway from WikiPathways. The PTGDS was highlight in yellow box.

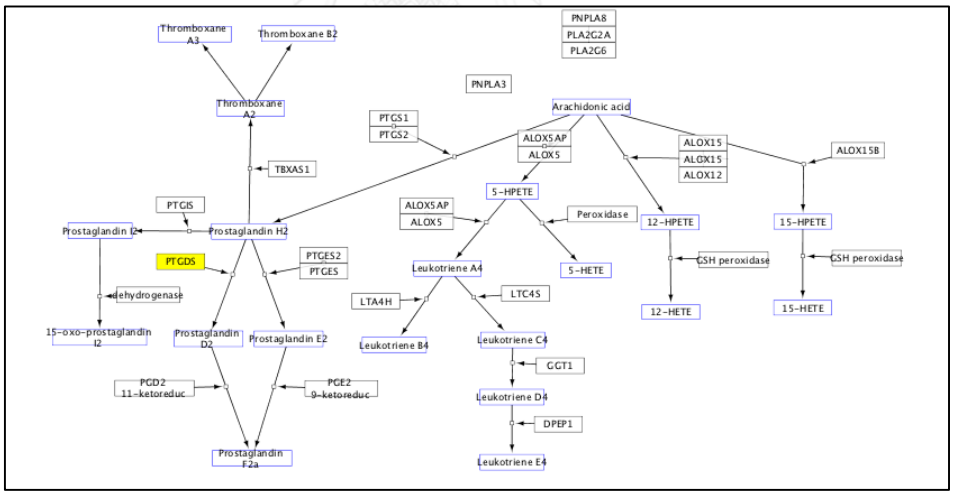


Figure 18 Eicosanoid synthesis pathway from WikiPathways. The PTGDS was highlight in yellow box.

As described in method section for the in-direct integration analysis, the molecular networks were constructed based on the protein-protein interaction networks. In order to map the mRNA expression data onto gene interaction network, we used Entrez Gene ID as the unique identifier for genes. When there are multiple probe sets corresponding to the same gene, we used the one with the maximum t-statistic as a representative. The integrated LN network was illustrated as organic interaction network shape in Figure 19. The gray square node represents the seed molecules, red node represents protein interaction molecule occurred from BIND and IntAct.

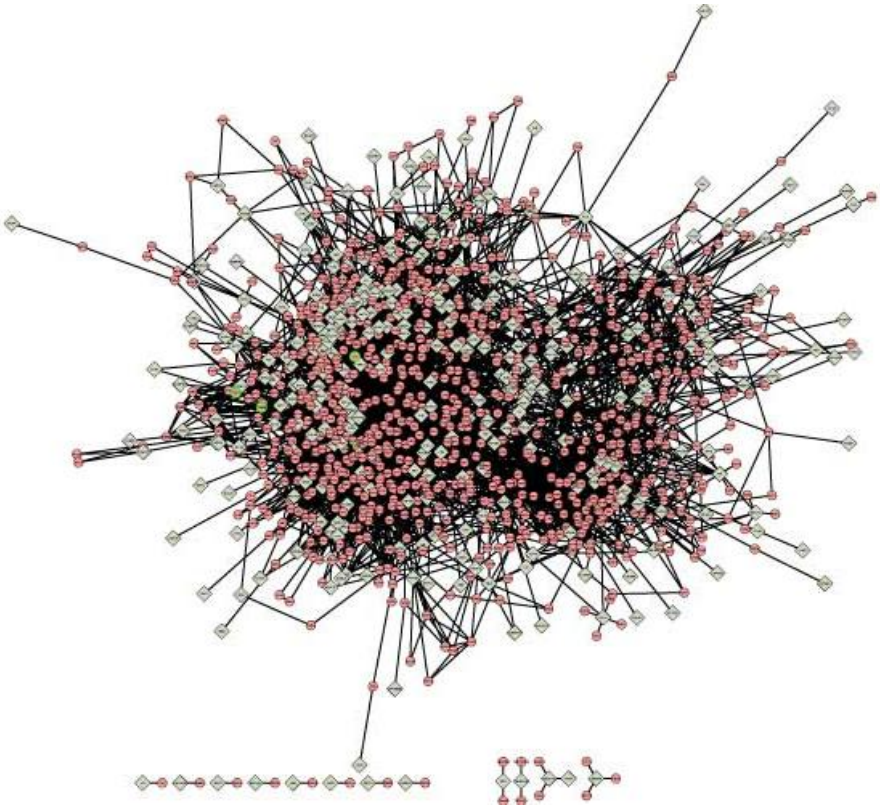


Figure 19 Integrated LN network based on BIND and IntAct PPI.

There are total 5 sub-networks were identified by MCODE. The score, number of node and edge are showed in Table 10. The top 3 sub-networks were illustrated in Figure 20. SERPINA1 was also identified in cluster 2, which shows a consistency to the direct map approach. This finding reveals the genes/proteins in complement system might play a major role in the LN treatment therapeutic response as the highly connected proteins were displayed as the core sub-network of refractory LN. The full list of each cluster members is listed in Appendix Table A1.

Table 10 List of sub-networks of integrated proteomics and transcriptome network

Cluster	Score (Density*#Nodes)	Nodes	Edges
1	4.461	89	397
2	4.368	144	629
3	2.286	7	16
4	1.4	15	21
5	0.5	2	1

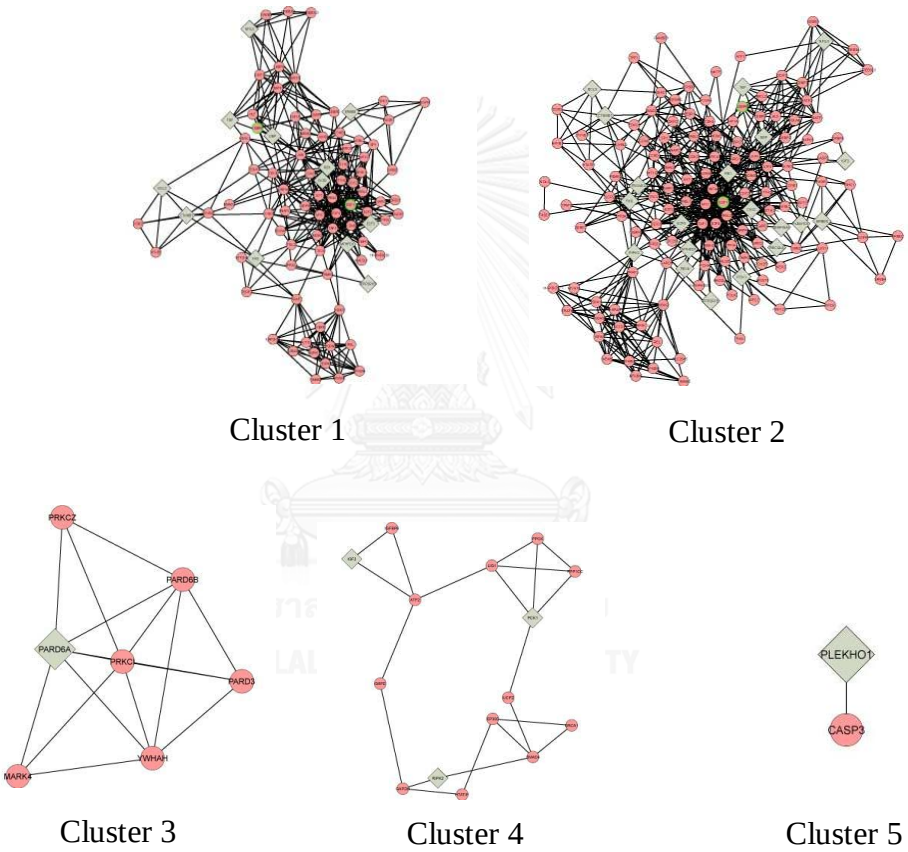


Figure 20 Sub-networks of integrated LN proteomics and transcriptome.

The molecules in cluster 1 involved in DNA/RNA biosynthetic process including with apoptosis and programmed cell death. Cluster 2 involved in a response to estrogen and steroid hormone stimulus. Cluster 3 revels tight junction and also with cell-cell junction. Cluster 4 seems to play a role in monosaccharide metabolic process and cell growth regulation process. While there are only 2 members in cluster 5.

Although there are some gene/protein that not significantly expressed in the transcript or protein level, but those candidates were also identified by this approach based on their highly connected interaction with other proteins. Thus, these gene/protein candidates might be useful as the candidate refractory LN biomarker. List of biological process and gene in those process were list in Table 11.

Table 11 Biological process of refractory LN sub-networks

cluster	Biological process	p-value	Gene
cluster 1	GO:0032774~RNA biosynthetic process	2.89E-09	BATF3, E2F2, TAF1, HLF, E2F3, CEBPB, CEBPD, SNAPC3, ESR1, GTF2H3, NFKB1, DDIT3, HIF1A, NFIL3, MYC
	GO:0006275~regulation of DNA replication	2.08E-06	CDC6, JUN, PPP2CA, PCNA, SHC1, TERF2, CDK2
	GO:0006915~apoptosis	0.00108438	E2F1, E2F2, TRAF2, ERBB3, MSH2, JUN, NFKBIA, NFKB1, BIRC5, FAS, MYC, CTNNB1
	GO:0012501~programmed cell death	0.001222535	E2F1, E2F2, TRAF2, ERBB3, MSH2, JUN, NFKBIA, NFKB1, BIRC5, FAS, MYC, CTNNB1
cluster 2	GO:0043627~response to estrogen stimulus	0.001249228	TRAF2, CEBPB, JUN, PPP2CA, CEBPG, IKBKG, BCL3, FAS, MYC, DDIT3
	GO:0048545~response to steroid hormone stimulus	2.96E-06	CCND1, EP300, ERBB4, BCL2, ERBB2, TGFBR2, ESR1, SERPINA1, TFF1, FAS, BRCA1, CTNNB1
	GO:0010628~positive regulation of gene expression	1.12E-11	E2F1, E2F3, MITF, NFKBIA, FOXO1, NFKB1, CTNNB1, REL, POU2F1, BCL3, MYC, TAF1, CEBPB, SMAD4, ESR1, SMAD2, RB1, BRCA1, CDK2, DDIT3, HIF1A, EP300, HNF4A, HDAC1, SP1, JUN, NCOA6, MAPK9
cluster 3	GO:0005923~tight junction	2.01E-04	PARD6A, PRKCZ, PARD3
	GO:0005911~cell-cell junction	0.001039732	PARD6A, PRKCZ, PARD3
	GO:0006468~protein amino acid phosphorylation	0.007521696	PRKCZ, PRKCI, MARK4

cluster 4	GO:0005996~monosaccharide metabolic process	4.49E-05	G6PD, IGF2, PPP1CC, GAPDH, PCK1
	GO:0001558~regulation of cell growth	7.47E-04	EP300, IGFBP6, SMAD4, IGF2
	GO:0032268~regulation of cellular protein metabolic process	8.27E-04	G6PD, EP300, SMAD4, IGF2, BRCA1

Human Endogenous Retrovisus analysis

Data collection

First step of the analysis is data preparation. All data were store in structured database. The number of records in all data collections was shown in Table 12. In the UCSC table browser, these three data are in the UCSC table named rmsk, knownGene, and kgXref, respectively. They are all based on the February 2009 human reference sequence (hg19/GRCh37), which includes sequences of 24 main chromosomes (chromosome 1-22, X, and Y), 59 unplaced contigs, and 9 haplotype chromosomes.

Table 12 the number of records in the original downloaded data according to the categories of assembled sequences

Data	Categories of assembled sequences	The number of records (%)
Human repeat annotation	Main chromosomes	5,232,237 (98.76%)
	Haplotype chromosomes	55,980 (1.05%)
	Unplaced contigs	9,913 (0.19%)
	Total	**5,298,130**
Human gene annotation/ Cross-reference gene IDs	Main chromosomes	73,660 (94.91%)
	Haplotype chromosomes	3,835 (4.94%)
	Unplaced contigs	119 (0.15%)
	Total	**77,614**

The annotation results and the cross-reference gene IDs contain not only the records based on the main placed chromosomes, but also based on the unplaced contigs and the haplotype chromosomes, which were separately generated from the main placed contigs. In this work, only the data of the main placed chromosomes were included in this study. This step was thus performed to remove all of the unwanted records in the data resources before performing any manipulation of the data. First, genome information from unplaced contigs and the haplotype chromosomes were removed.

Second selected only human repeat annotation which belong to HERV superfamilies. Finally, the annotation of HERV fragments, the gene annotation and the cross-reference IDs, which are based on only 24 main placed chromosomes were obtained from the data selection. At this step, the annotation records based on the unplaced contigs and the haplotype chromosomes were removed from downloaded data. Next, selected informative human repeat annotation by removing all of the repeats, which do not belong to HERV superfamilies. Finally, HERV fragments annotation, the gene annotation and the cross-reference IDs which are based on only 24 main assembled chromosomes, including chromosome 1-22, X and Y. The numbers of records resulting from the data selection are shown in Table 13.

Table 13 The number of selected records from the data resources

Data	The number of records
Annotation of HERV fragments	687,420
Selected human gene annotation	73,660
Selected cross-reference gene IDs	73,660

There are six HERV superfamilies found in the annotation of the HERV fragments, including ERV1, ERVK, ERVL, ERVL-MaLR, ERVL?, and ERV. For the last two, they are fragments that cannot be certainly determined for its superfamily. Thus, all of the fragments annotated belonging to ERVL? and ERV would have been considered as the unclassified fragments in this work. The detailed proportions in the annotation data of the HERV fragments, obtained from the removing unwanted records from the raw human repeat annotation, is shown in Table 14. The full list of HERV superfamily, family and name is list in Appendix Table B1.

Table 14 Detailed proportions of each HERV in the HERV annotation data

Superfamily	The number of records	Percentage
ERVL-MaLR	343,666	49.99%
ERV1	172,863	25.15%
ERVL	157,992	22.98%
ERVK	10,490	1.53%
Unclassified ERVs	2,379	0.35%
Total	**687,420**	100%

HERV defragmentation using REannotate

Generally, a complete HERV element is composed of two flanking LTRs and a set of internal retroviral genes in the central. However, the computational annotation of the HERVs is usually incompletely performed unlike the picture of that complete element, because the LTRs and the internal sequences would be separately detected in the annotation. There are two reasons supporting why the separate detection is required. First, to avoid the missing discoveries of the solitary LTRs. Secondly, due to massive accumulation of insertions and deletions on the HERV sequences, the annotation results are often displayed as several fragments of an element instead of one with a long gap [120]. Therefore, to enable observing the HERVs in a view of being the elements, the HERV defragmentation is required to join fragments of the same element into a single element. The numbers and the proportions of the HERV fragments in the annotation data are shown in Figure 21.

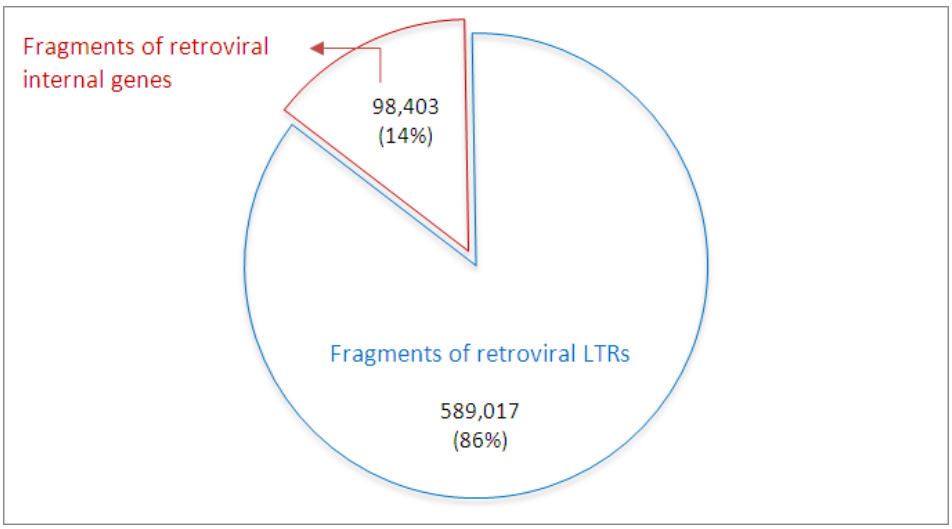


Figure 21 The number and proportions of each HERV fragment type

HERV superfamilies that were not reported in REannotate list were still obtained but their name were changed according to REannotate result manipulation to include with data clean-up. They were named as HERV name/group. In summary, there are 133 HERV families with 413 names or groups found in the human genome in total. The full list of HERVs are showed in Appendix Table B1. The summary of the numbers of the types of truncation patterns found is shown in Table 15.

Table 15 The numbers and percentages of HERV elements according to each type of truncation patterns

Type of truncation patterns	**The number of elements (elements)**	**Percentage**
complete element	7,975	1.48%
5’-truncated element	10,796	2.01%
3’-trunctated element	9,856	1.84%
both 5’- and 3’-truncated element	39,724	7.40%
solo LTRs	468,710	87.27%
Total	537,061	100.00%

Mapping HERVs on the human genes

The edited REannoate output and the selected gene annotation were used in this step. Each gene isoform were collected with their neighboring HERV elements which are located not far from the transcription start or transcription termination site more than 100,000 bp. This mapping was done using python program. The summary of this mapping result is shown in Table 16. More than a million copy number of HERV were found in human genome. The proportion distribution of each HERV superfamily in human chromosomes was shown in Figure 22. Their location in human gene were illustrated in Figure 22b. There is a slightly in anti-sense bias of HERV orientation in the genome. Total 899 expressed HERV were identified in genbank human transcripts. More than 50% of them are located in gene which slightly more abundant in exon region as shown in Figure 23b. There are 233 genes that contain HERV as part of their transcripts which associated to 1,086 MeSH disease terms. Most of expressed HERV type found in disease-associated genes are solo-LTR. The analysis results show that ERV3-MaLR is the most active HERV superfamily in term of disease association. These HERV neighboring genes were principally related to coenzyme metabolic process, protein kinase activity, and immunoglobulin function.

Table 16 Summary of both the numbers of genes and HERV elements resulting from mapping HERVs on the human genes

HERV elements			**Gene isoforms**		
gene-neighboring	**non-gene-neighboring**	**Total**	**HERV-hosting**	**non-HERV-hosting**	**Total**
382,622 (71.24%)	154,439 (28.76%)	537,061	73,645 (99.98%)	15 (0.02%)	73,660

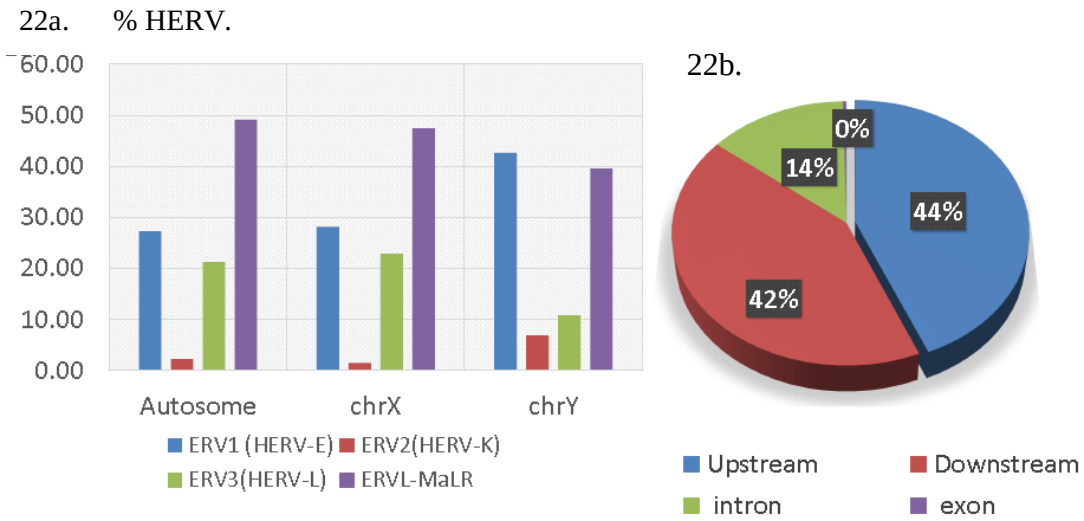


Figure 22 HERV distribution in human chromosomes, b) HERV's location distribution in human genome

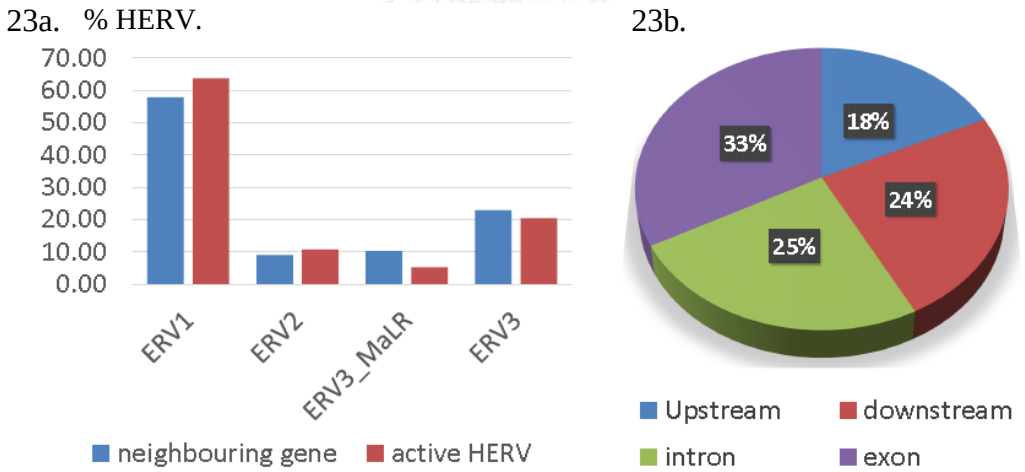


Figure 23 Comparing neighboring HERV and expressed HERV in gen bank mRNA. b) Expressed HERV in gen bank mRNA database

EnHERV construction

EnHERV can be access at http://sysbio.chula.ac.th/enherv/. The main page is shown in Figure 24. EnHERV provide two searching function 1.) Search by gene(s), which provide auto-complete gene input for user and 2.) Search by HERV characteristics including HERV superfamily, family, name, their location in genome, their location in gene, their orientation, and their structure completeness as shown in Figure 25, which illustrated in red box. The searched result is displayed in the table format. EnHERV provides UCSC genome browser link for visualizing the genomic location of the searched result in specific regions. The database allows user to download the result for downstream analysis. The search by HERV name option was organized into drop down function for more precisely selected the user interested HERV member and their characteristics.

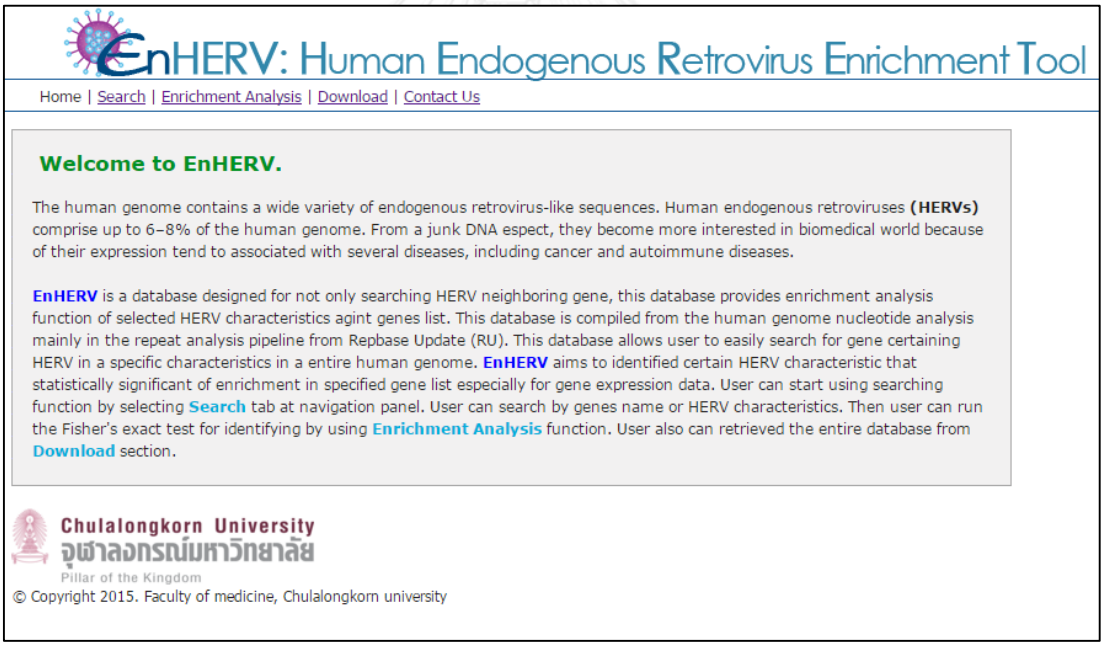


Figure 24 EnHERV homepage

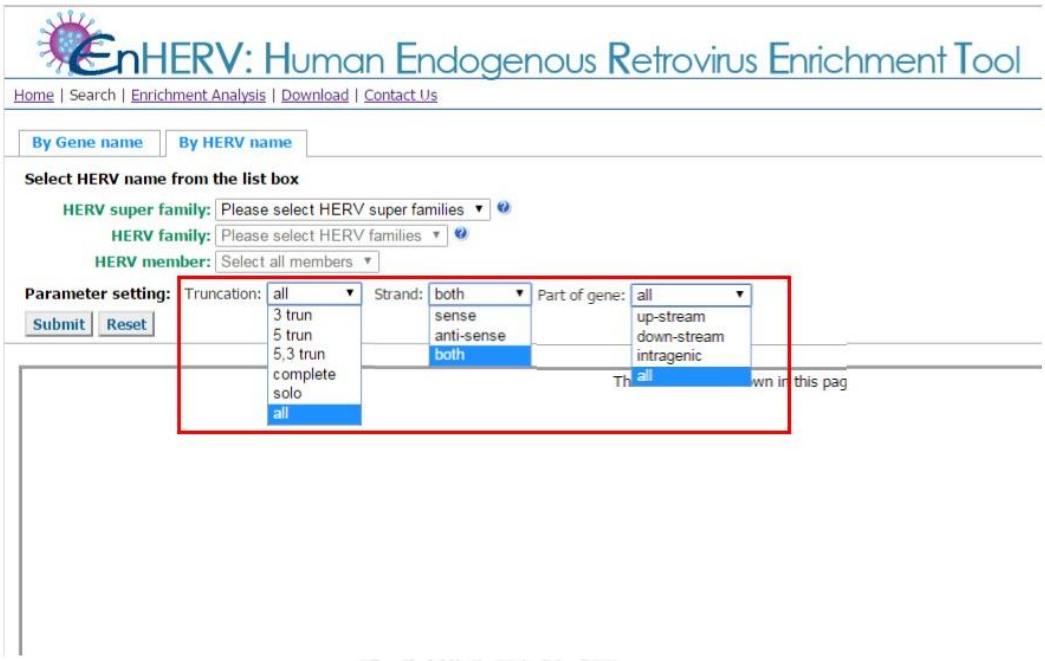


Figure 25 HERV characteristic parameter

The enrichment analysis function provides an enrichment analysis between genes with specific HERV characteristics and user-defined gene list as shown in Figure 25. EnHERV calculates Fisher's p-value including with OR ratio for the selected lists. Genes containing specified HERV characteristics will be shown in the result table. Then user can download the result table for further investigation.

Since solo-LTRs are the most abundance HERV structure and according to their properties that they contain the regulatory region of HERV which might effect to their neighboring gene expression. The next analysis step were focused on solo-LTR that located in both inter- and intra-genic regions for looking for any specific pattern in differences gene expression pattern in SLE including with some cancer types.

Table 17 List of GSE experiments used as pre-set gene lists in EnHERV

GSE experiment	Number of DEG
GSE50101:CD4+ Seasonal allergic rhinitis (SAR) - down regulated genes	460
GSE50101:CD4+ Seasonal allergic rhinitis (SAR) - up regulated genes	560
GSE32323:Colorectal cancer cells - down regulated genes	39
GSE32323:Colorectal cancer cells - up regulated genes	38
GSE52471:Discoid lupus erythematosus (DLE) PBMC - down regulated genes	247
GSE52471:Discoid lupus erythematosus (DLE) PBMC - up regulated genes	259
GSE32591:LN glomeruli - down regulated genes	223
GSE32591:LN glomeruli - up regulated genes	456
GSE32591:LN tubulointerstitial - down regulated genes	80
GSE32591:LN tubulointerstitial - up regulated genes	268
GSE1793:Melanoma cells - down regulated genes	174
GSE1793:Melanoma cells - up regulated genes	173
GSE10325:SLE Myeloid - down regulated genes	236
GSE10325:SLE Myeloid - up regulated genes	662
GSE61635:RNP autoantibody+ SLE PBMC - down regulated genes	41
GSE61635:RNP autoantibody+ SLE PBMC - up regulated genes	307
GSE4588:SLE B cells - down regulated genes	1138
GSE4588:SLE B cells - up regulated genes	634
GSE4588:SLE CD4 T cells - down regulated genes	988
GSE4588:SLE CD4 T cells - up regulated genes	900
GSE27427:SLE neutrophils - down regulated genes	458
GSE27427:SLE neutrophils - up regulated genes	1682
GSE20864:SLE PBMC - down regulated genes	1114
GSE20864:SLE PBMC - up regulated genes	1526

The proportion of the number of solo-LTR in differences gene locations was illustrated as pie chart in Figure 26. The majority of LTR are located in up-stream and down-stream region of genes. The pattern of HERV neighboring genes showed similar expression patterns as most of LTRs were located in intergenic region of the genes. While focusing on LTR intragenic, we found that the majority of LTRs are located in intron. Moreover, intragenic LTRs, especially the intron LTRs were located in anti-sense strand in all expression conditions. We further analyzed the number of solo-LTR in all conditions and plotted as line graph as some graph were illustrated in Figure 27. By comparing genes containing solo-LTR between down-regulated and up-regulated

genes, the analysis results show that number of solo-LTR in up-regulated genes were clearly higher than the down-regulated genes in SLE myeloid, RNP+ SLE PBMC, and both glomerular and tubular LN (green lines) in which the number of solo-LTR in anti-sense strand is slightly higher than sense direction (dark green line) comparing to expressed genes containing solo-LTR in colorectal cancer and DLE CD3+ T cells and other conditions. In summary, the intragenic solo-LTR in anti-sense direction pattern seems to associate with up-regulated genes in SLE myeloid, RNP+ SLE PBMC, and both glomerular and tubular LN.

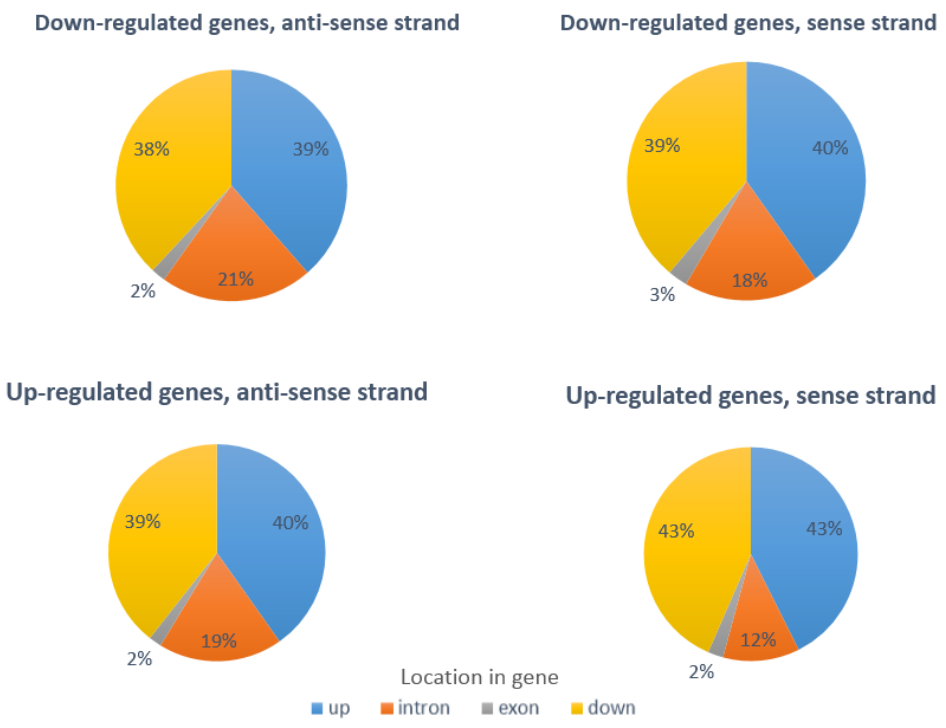
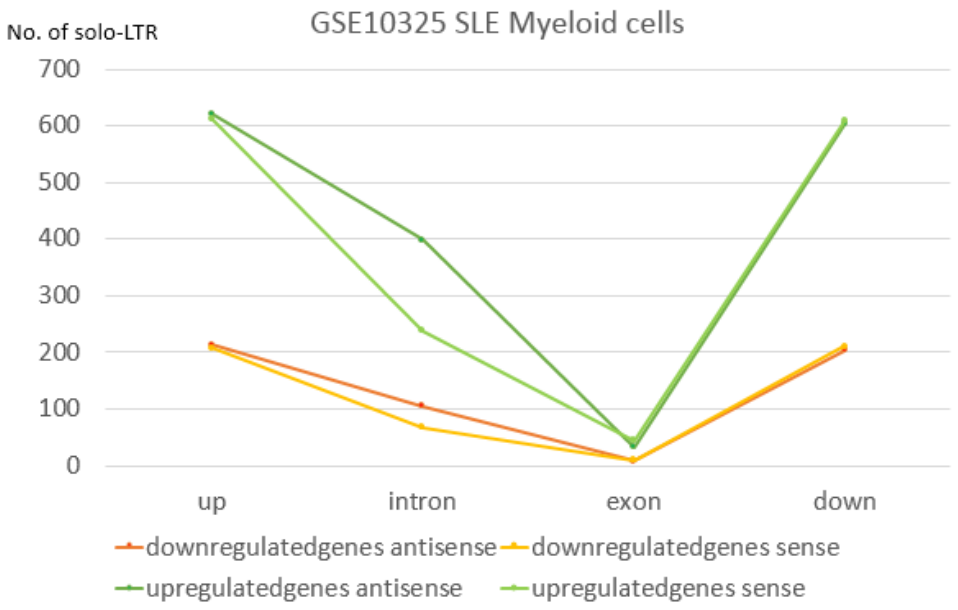
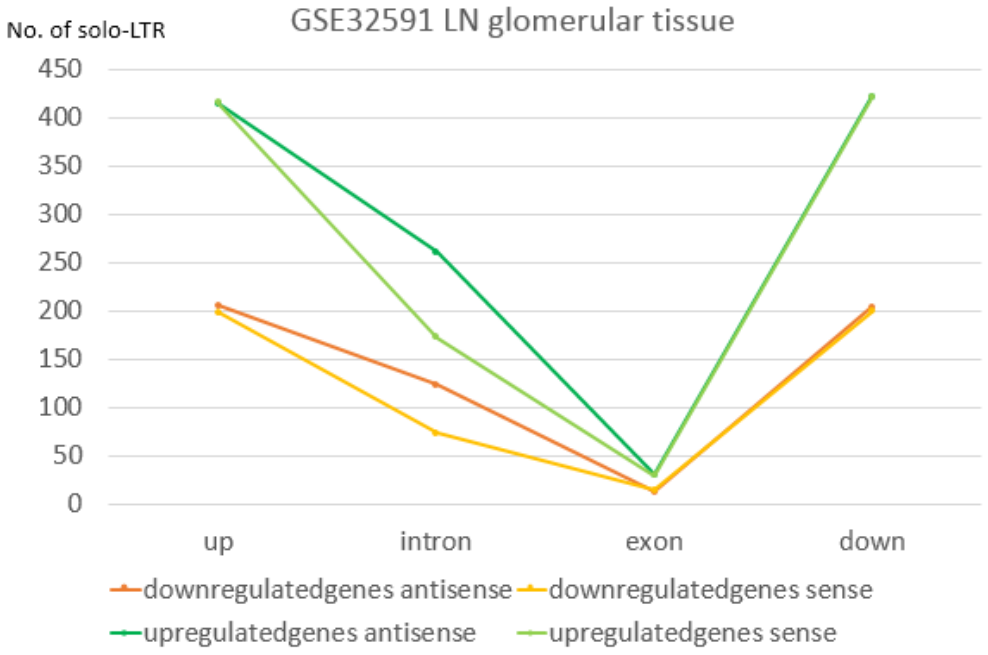
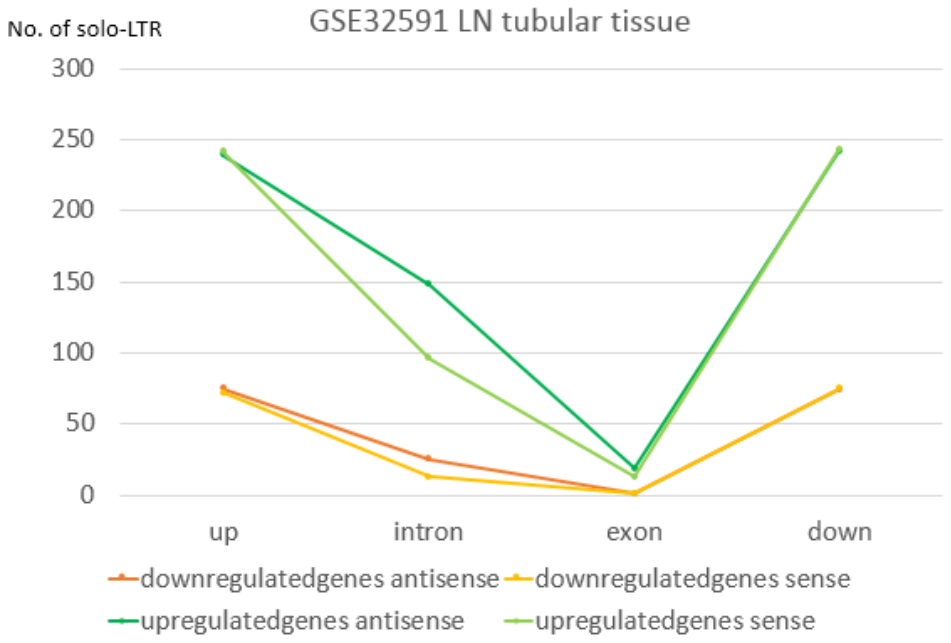
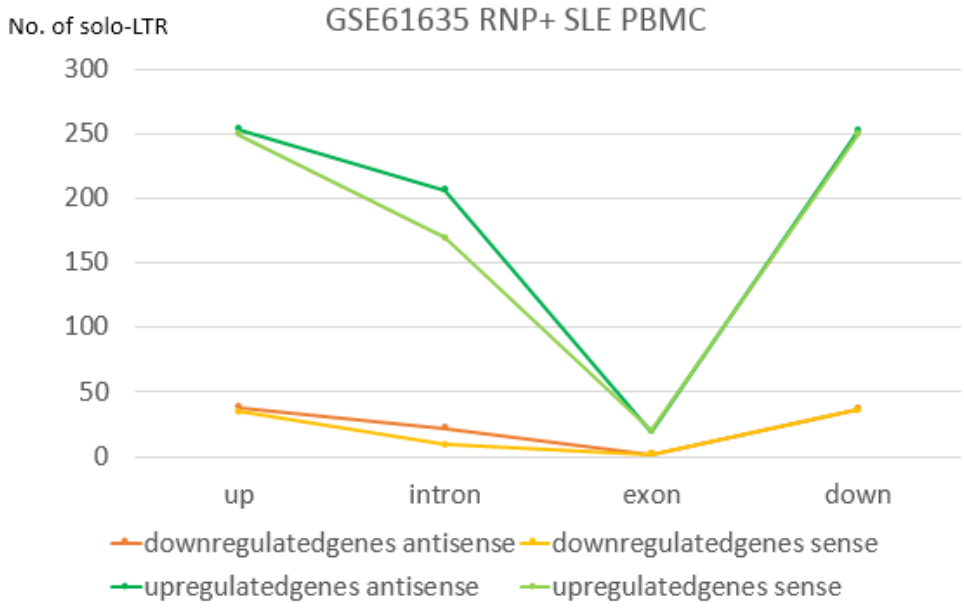


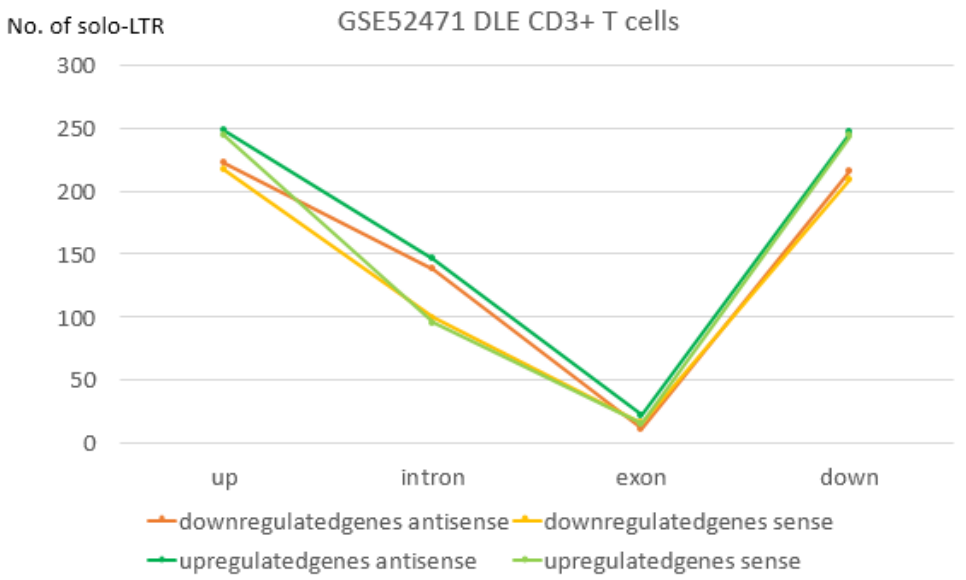
Figure 26 The solo-LTR distributions ratio in different part of genes under up- and down-regulated gene expression conditions.


GSE10325 SLE Myeloid cells
No. of solo-LTR
700
600
500
400
300
200
100
0
up
intron
exon
down
downregulatedgenes antisense
downregulatedgenes sense
upregulatedgenes antisense
upregulatedgenes sense


GSE32591 LN glomerular tissue
No. of solo-LTR
450
400
350
300
250
200
150
100
50
0
up
intron
exon
down
downregulatedgenes antisense
downregulatedgenes sense
upregulatedgenes antisense
upregulatedgenes sense


GSE32591 LN tubular tissue
No. of solo-LTR
300
250
200
150
100
50
0
up
intron
exon
down
downregulatedgenes antisense
downregulatedgenes sense
upregulatedgenes antisense
upregulatedgenes sense


GSE61635 RNP+ SLE PBMC
No. of solo-LTR
300
250
200
150
100
50
0
up
intron
exon
down
downregulatedgenes antisense
downregulatedgenes sense
upregulatedgenes antisense
upregulatedgenes sense



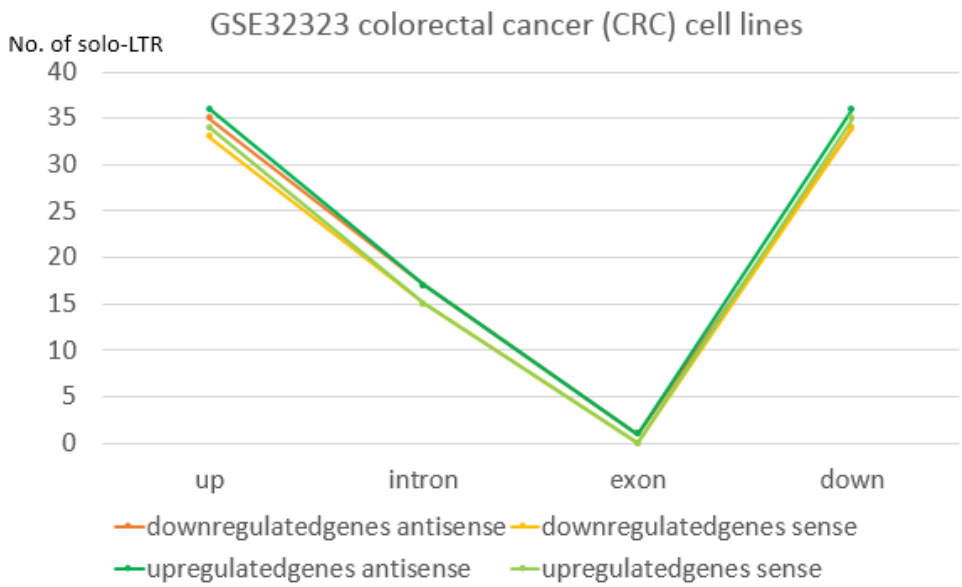


Figure 27 Number of solo-LTR in different part of gene.

Association analysis of solo LTR in cancer and autoimmune disease

HERVs have been reports to be active due to hypomethylation in cancer and autoimmune diseases [18, 155, 156]. Therefore, distinct isoforms or gene silencing due to HERV were reported in various diseases with global hypomethylation events particularly in cancers [157-160]. Alternative transcript of CD5 by HERV-E was also reported in SLE B cells [161]. However, there are still limited comprehensive data and tool available to analyze HERV in relation to other genes. Since we hypothesize that HERV can control neighboring genes by either up-regulation or down-regulation, we develop a model to identify genes that associated with HERV in the genome that were differential expression in various diseases. This information can serve as a screening tool to further study candidate genes that might be under the regulation of HERV. The association analysis results between all HERV neighboring genes in various gene expression conditions are showed in Table 18. Including with the solo-LTR distribution analysis results, there are some LTR patterns that associated with certain differentially expressed genes pattern in specific tissue and disease conditions, the association analysis were performed for all solo-LTR in 56 condition as listed in Table 6 in the method section.

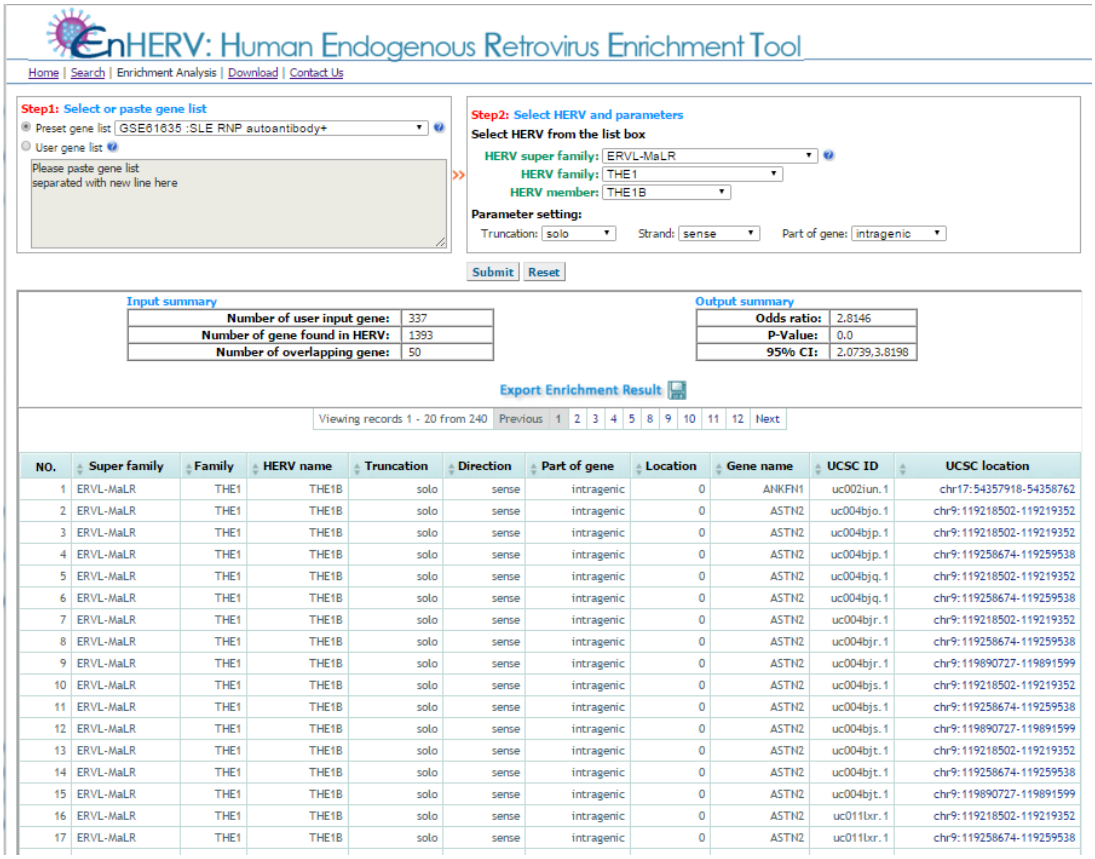
The example of enrichment analysis result of sense intragenic THEB LTR against up-regulated RNP+ SLE gene were illustrated in Figure 28. The result shows the association of 50 up-regulated genes in SLE PBMC RNP+ condition containing sense intragenic THE1B LTR with significant level as P-value 0 and OR ratio 2.83.

We also performed the association analysis between all 4 HERV superfamilies and 25 selected HERVs with the same disease conditions as mention in method section. The list of significantly association between HERV superfamilies and various condition were reported in Table 19 while Table 20 shows the significantly association in individual HERV. The results clearly showed that different type of LTRs were differentially associated with certain gene expression profiles in particular disease conditions.

There is no association between hypomethylation and HERV neighboring genes found in entire HERVs and superfamilies level, with association cut-off $p < 0.001$ and OR >1. The hypomethylation event was represented by DEG in GSE9764 5-azacytidine treated human mesenchymal stem cells. However, we found the association between

down-regulated genes in hypomethylation and some individual HERVs only in ERVL and ERVL-MaLR. These association pattern is correspond with the association pattern of HERVs and gene expression in most of cancer cells. This analysis result also supports the role of HERVs association in diseases as tissues and HERVs type specifics manner according to the hypomethylation event in our analysis was represented by 5-aza treated cancer cells.

.



NO.	Super family	Family	HERV name	Truncation	Direction	Part of gene	Location	Gene name	UCSC ID	UCSC location
1	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ANKFN1	uc002iun.1	chr17:54357918-54358762
2	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc004bjo.1	chr9:119218502-119219352
3	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc004bjp.1	chr9:119218502-119219352
4	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc004bjp.1	chr9:119258674-119259538
5	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc004bjq.1	chr9:119218502-119219352
6	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc004bjq.1	chr9:119258674-119259538
7	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc004bjr.1	chr9:119218502-119219352
8	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc004bjr.1	chr9:119258674-119259538
9	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc004bjr.1	chr9:119890727-119891599
10	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc004bjs.1	chr9:119218502-119219352
11	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc004bjs.1	chr9:119258674-119259538
12	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc004bjs.1	chr9:119890727-119891599
13	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc004bjt.1	chr9:119218502-119219352
14	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc004bjt.1	chr9:119258674-119259538
15	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc004bjt.1	chr9:119890727-119891599
16	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc011lxr.1	chr9:119218502-119219352
17	ERVL-MaLR	THE1	THE1B	solo	sense	intragenic	0	ASTN2	uc011lxr.1	chr9:119258674-119259538

Figure 28 Enrichment analysis result of sense intragenic THEB LTR against up-regulated RNP+ SLE gene.

Table 18 Association analysis results at entire HERV solo-LTR level (Significant data with OD > 1 and P < 0.001, indicated in bold letter)

GEO accession	HERV pattern	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE14905 Psoriasis, skin	All	0.1411	0	0.1426	0
	Sense	0.151	0	0.1547	0
	Anti-sense	0.1529	0	0.1566	0
	Intragenic	**1.2202**	**0.00053**	0.9747	0.667739
	Intergenic	0.1411	0	0.1426	0
GSE52471 psoriasis, skin	All	0.1428	0	0.2326	0
	Sense	0.1518	0	0.2471	0
	Anti-sense	0.1536	0	0.2533	0
	Intragenic	**1.1944**	**0.000797**	1.0886	0.176576
	Intergenic	0.1428	0	0.2326	0
GSE12453 Hodgkin Lymphoma vs centroblasts	All	0.0904	0	0.2135	0
	Sense	0.0993	0	0.2263	0
	Anti-sense	0.101	0	0.2313	0
	Intragenic	**1.4913**	**0**	0.6844	9.90E-06
	Intergenic	0.0904	0	0.2135	0
GSE12453 Hodgkin Lymphoma vs centrocytes	All	0.1052	0	0.1991	0
	Sense	0.1139	0	0.2088	0
	Anti-sense	0.116	0	0.2159	0
	Intragenic	**1.564**	**0**	0.539	0
	Intergenic	0.1052	0	0.1991	0
GSE12453 Hodgkin Lymphoma vs memory B cells	All	0.0902	0	0.1798	0
	Sense	0.0956	0	0.1894	0
	Anti-sense	0.0983	0	0.1959	0
	Intragenic	**1.3944**	**0.000174**	0.6036	0
	Intergenic	0.0902	0	0.1798	0
GSE12453 Hodgkin Lymphoma vs plasma Cells	All	0.0973	0	0.1864	0
	Sense	0.1036	0	0.1944	0
	Anti-sense	0.1058	0	0.2001	0
	Intragenic	**1.2982**	**0.000208**	0.5668	0
	Intergenic	0.0973	0	0.1864	0

GEO accession	HERV pattern	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE12453 Diffuse Large B cells Lymphoma vs centroblasts	All	0.1095	0	0.1528	0
	Sense	0.1189	0	0.1604	0
	Anti-sense	0.1195	0	0.1629	0
	Intragenic	**1.6974**	**0**	0.5042	0
	Intergenic	0.1095	0	0.1528	0
GSE12453 Diffuse Large B cells Lymphoma vs centrocytes	All	0.1038	0	0.1363	0
	Sense	0.1123	0	0.1445	0
	Anti-sense	0.1128	0	0.1452	0
	Intragenic	**1.7304**	**3.00E-08**	0.4389	0
	Intergenic	0.1038	0	0.1363	0
GSE12453 Diffuse Large B cells Lymphoma vs memory B cells	All	0.1061	0	0.1175	0
	Sense	0.1143	0	0.1256	0
	Anti-sense	0.1148	0	0.1284	0
	Intragenic	**1.6155**	**0.00025**	0.5491	0
	Intergenic	0.1061	0	0.1175	0
GSE45829 EBV Infected B cells	All	0.2749	0	0.2021	0
	Sense	0.2873	0	0.2154	0
	Anti-sense	0.2886	0	0.2215	0
	Intragenic	0.7178	0.001509	**1.4435**	**0**
	Intergenic	0.2749	0	0.2021	0
GSE1299 Breast cancer cells	All	0.1202	0	0.1425	0
	Sense	0.1281	0	0.1489	0
	Anti-sense	0.1305	0	0.154	0
	Intragenic	**1.3762**	**0.000808**	1.0505	0.744504
	Intergenic	0.1202	0	0.1425	0
GSE9764 5-azacydine treated human mesenchymal stem cells	All	0.0896	0	0.0912	0
	Sense	0.0973	0	0.0988	0
	Anti-sense	0.0968	0	0.0992	0
	Intragenic	0.8994	0.195535	0.9395	0.504043
	Intergenic	0.0896	0	0.0912	0

Table 19 Association analysis results at HERV superfamily level with OR > 1 and p < 0.001, indicated in bold letter (Only association data were shown in this table)

19A. ERV1/HERVE solo-LTR

GEO accession	HERV pattern	ERV1/HERVE solo-LTR			
		Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE13887 SLE T cells	All	0.2312	0	0.1894	0
	Sense	0.3922	0	0.3648	0
	Anti-sense	0.364	0	0.3746	0
	Intragenic	0.6386	0	**1.3478**	**1.02E-06**
	Intergenic	0.2448	0	0.1985	0
GSE20864 SLE PBMC	All	0.3412	0.32715351	0.3091	0
	Sense	0.3412	0.21366473	0.5722	0.00035791
	Anti-sense	0.7739	0.57836251	0.4959	8.10E-06
	Intragenic	1.1661	1	**1.7429**	**1.57E-05**
	Intergenic	0.3611	0.34199072	0.3026	0
GSE61635 SLE PBMC RNP+	All	0.1805	0	0.116	0
	Sense	0.3277	0	0.2681	0
	Anti-sense	0.3668	0	0.2506	0
	Intragenic	1.2134	0.0368163	**1.3443**	**2.39E-05**
	Intergenic	0.188	0	0.1221	0
GSE13355 Psoriasis, skin	All	0.4136	8.00E-08	0.3885	0
	Sense	0.6945	0.00579991	0.659	0.00011074
	Anti-sense	0.6615	0.00255271	0.6092	6.55E-06
	Intragenic	**1.5684**	**2.61E-05**	0.9349	0.49353108
	Intergenic	0.4289	1.80E-07	0.4122	0
GSE14905 Psoriasis, skin	All	0.2766	0	0.2843	0
	Sense	0.526	0	0.5003	0
	Anti-sense	0.5025	0	0.4823	0
	Intragenic	**1.4588**	**0**	0.8767	0.03968541
	Intergenic	0.2867	0	0.2978	0
GSE32407 Psoriasis, skin	All	inf	1	0.2273	0.04598508
	Sense	inf	1	0.3838	0.10873814
	Anti-sense	inf	1	0.5158	0.40229897
	Intragenic	inf	0.02702273	0.4239	0.36775098
	Intergenic	inf	1	0.2405	0.05249269
GSE52471 psoriasis, skin	All	0.2873	0	0.4111	0
	Sense	0.5098	0	0.649	5.00E-08
	Anti-sense	0.4688	0	0.6327	2.00E-08
	Intragenic	1.2675	2.10E-05	0.9724	0.68830239
	Intergenic	0.3007	0	0.43	0

GSE12453 Hodgkin Lymphoma vs centroblasts	All	0.203	0	0.3687	0
	Sense	0.4032	0	0.6228	1.54E-05
	Anti-sense	0.4098	0	0.5571	1.10E-07
	Intragenic	**1.2306**	**0.00015277**	0.7329	1.75E-03
	Intergenic	0.2094	0	0.3801	0
GSE12453 Hodgkin Lymphoma vs centrocytes	All	0.2293	0	0.3436	0
	Sense	0.4296	0	0.5681	2.00E-08
	Anti-sense	0.4392	0	0.5182	0
	Intragenic	**1.2922**	**3.64E-05**	0.6118	2.20E-07
	Intergenic	0.2354	0	0.3603	0
GSE12453 Hodgkin Lymphoma vs memory B cells	All	0.1837	0	0.3214	0
	Sense	0.363	0	0.5546	0
	Anti-sense	0.3558	0	0.4765	0
	Intragenic	**1.3666**	**0.00073039**	0.6234	0
	Intergenic	0.1913	0	0.3391	0
GSE12453 Diffuse Large B cells Lymphoma vs centroblasts	All	0.2267	0	0.2782	0
	Sense	0.4305	0	0.4966	0
	Anti-sense	0.4199	0	0.4515	0
	Intragenic	**1.5358**	**1.10E-07**	0.5131	0
	Intergenic	0.2345	0	0.2931	0
GSE12453 Diffuse Large B cells Lymphoma vs centrocytes	All	0.2173	0	0.2581	0
	Sense	0.4308	0	0.45	0
	Anti-sense	0.4176	0	0.3921	0
	Intragenic	**1.6832**	**1.30E-07**	0.4605	0
	Intergenic	0.2247	0	0.2728	0
GSE12453 Diffuse Large B cells Lymphoma vs memory B cells	All	0.2172	0	0.2313	0
	Sense	0.4319	2.00E-08	0.4073	0
	Anti-sense	0.4236	2.00E-08	0.3833	0
	Intragenic	**1.6833**	**6.49E-05**	0.5158	0
	Intergenic	0.2252	0	0.2446	0
GSE12453 Diffuse Large B cells Lymphoma vs naive B cells	All	0.157	0	0.2164	0
	Sense	0.3137	0	0.3952	0
	Anti-sense	0.3144	0	0.386	0
	Intragenic	**1.6229**	**3.97E-05**	0.5586	0
	Intergenic	0.1614	0	0.2309	0
GSE45829 EBV Infected B cells	All	0.4396	1.11E-06	0.3909	0
	Sense	0.6833	0.00508797	0.6027	0
	Anti-sense	0.6402	0.00103782	0.6472	0
	Intragenic	0.7393	0.01269972	**1.2568**	**4.60E-07**
	Intergenic	0.4659	6.04E-06	0.408	0

GSE1299 Breast cancer cells	All	0.2377	0	0.258	0
	Sense	0.4624	0	0.4316	5.00E-08
	Anti-sense	0.4276	0	0.4553	1.18E-06
	Intragenic	**1.4685**	**7.62E-05**	1.0116	0.943421
	Intergenic	0.2523	0	0.2735	0
GSE9750 Cervical cancer	All	0.2022	0	0.3485	0
	Sense	0.4755	0	0.5399	0
	Anti-sense	0.4464	0	0.4537	0
	Intragenic	**1.2148**	**1.00E-08**	0.6896	2.00E-08
	Intergenic	0.2177	0	0.3688	0
GSE9764 5-azacydine treated human mesenchymal stem cells	All	0.1738	0	0.1767	0
	Sense	0.3613	0	0.3495	0
	Anti-sense	0.3212	0	0.3402	0
	Intragenic	1.1446	0.12298673	1.0424	0.66259312
	Intergenic	0.1801	0	0.1878	0

19B. ERV3/HERVL solo-LTR

GEO accession	HERV pattern	**ERV3/HERVL solo-LTR**			
		Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE13887 SLE T cells	All	0.205	0	0.1856	0
	Sense	0.3484	0	0.4186	0
	Anti-sense	0.335	0	0.4338	0
	Intragenic	0.7302	1.39E-05	**1.3342**	**2.08E-06**
	Intergenic	0.214	0	0.194	0
GSE10325 SLE myeloid cells	All	0.3421	5.10E-07	0.3645	0
	Sense	0.574	0.001236	0.5904	0.000108
	Anti-sense	0.5918	0.003464	0.6594	0.003091
	Intragenic	1.134	0.404069	**1.4597**	**0.000949**
	Intergenic	0.3615	2.42E-06	0.3681	0
GSE61635 SLE PBMC RNP+	All	0.1782	0	0.1095	0
	Sense	0.3633	0	0.2893	0
	Anti-sense	0.3523	0	0.277	0
	Intragenic	1.2727	0.008475	**1.4665**	**3.00E-08**
	Intergenic	0.187	0	0.1162	0
GSE10500 RA macophage cells	All	0.3632	0	0.3851	0
	Sense	0.512	0	0.7011	3.14E-06
	Anti-sense	0.4702	0	0.7284	5.45E-05
	Intragenic	0.9208	0.260191	**1.2356**	**0.000658**
	Intergenic	0.3689	0	0.4087	0

GSE13355 Psoriasis, skin	All	0.3959	2.00E-08	0.331	0
	Sense	0.7187	0.012891	0.6036	1.33E-06
	Anti-sense	0.7454	0.028713	0.5446	1.00E-08
	Intragenic	**1.4256**	**0.000888**	0.9818	0.856629
	Intergenic	0.4186	7.00E-08	0.3371	0
GSE14905 Psoriasis, skin	All	0.2812	0	0.2631	0
	Sense	0.5802	0	0.4673	0
	Anti-sense	0.576	0	0.4569	0
	Intragenic	**1.4997**	**0**	0.9298	0.253534
	Intergenic	0.2957	0	0.2674	0
GSE52471 psoriasis, skin	All	0.2732	0	0.3646	0
	Sense	0.5416	0	0.5761	0
	Anti-sense	0.4868	0	0.5867	0
	Intragenic	**1.4214**	**0**	1.1084	0.119536
	Intergenic	0.2852	0	0.3722	0
GSE12453 Hodgkin Lymphoma vs centroblasts	All	0.2076	0	0.3605	0
	Sense	0.4503	0	0.5311	0
	Anti-sense	0.4324	0	0.5752	3.60E-07
	Intragenic	**1.4091**	**0**	0.689	0.000148
	Intergenic	0.2179	0	0.3816	0
GSE12453 Hodgkin Lymphoma vs centrocytes	All	0.2275	0	0.3401	0
	Sense	0.4716	0	0.5055	0
	Anti-sense	0.4693	0	0.4981	0
	Intragenic	**1.4627**	**0**	0.5368	0
	Intergenic	0.2381	0	0.3602	0
GSE12453 Hodgkin Lymphoma vs memory B cells	All	0.1796	0	0.3218	0
	Sense	0.3819	0	0.4982	0
	Anti-sense	0.3674	0	0.4655	0
	Intragenic	**1.3955**	**0.00024**	0.5655	0
	Intergenic	0.1865	0	0.3278	0
GSE12453 Diffuse Large B cells Lymphoma vs centroblasts	All	0.232	0	0.2795	0
	Sense	0.498	0	0.4725	0
	Anti-sense	0.4842	0	0.4373	0
	Intragenic	**1.703**	**0**	0.5133	0
	Intergenic	0.2371	0	0.291	0
GSE12453 Diffuse Large B cells Lymphoma vs centrocytes	All	0.2153	0	0.2466	0
	Sense	0.5	0	0.4008	0
	Anti-sense	0.4714	0	0.3718	0
	Intragenic	**1.7546**	**1.00E-08**	0.3843	0
	Intergenic	0.225	0	0.2582	0

GSE12453 Diffuse Large B cells Lymphoma vs memory B cells	All	0.2125	0	0.2294	0
	Sense	0.4276	1.00E-08	0.4088	0
	Anti-sense	0.4428	8.00E-08	0.3643	0
	Intragenic	**1.6786**	**5.86E-05**	0.5071	0
	Intergenic	0.215	0	0.2387	0
GSE12453 Diffuse Large B cells Lymphoma vs naive B cells	All	0.156	0	0.2102	0
	Sense	0.3556	0	0.3842	0
	Anti-sense	0.3423	0	0.3472	0
	Intragenic	**1.7831**	**7.60E-07**	0.5589	0
	Intergenic	0.1574	0	0.2201	0
GSE45829 EBV infected B cells	All	0.4626	7.91E-06	0.3621	0
	Sense	0.6517	0.001178	0.5733	0
	Anti-sense	0.709	0.011829	0.6013	0
	Intragenic	0.798	0.056784	**1.1838**	**0.000195**
	Intergenic	0.4771	1.34E-05	0.3758	0
GSE1299 Breast cancer cells	All	0.2485	0	0.2397	0
	Sense	0.5814	2.38E-06	0.4615	4.90E-07
	Anti-sense	0.5106	0	0.4954	1.03E-05
	Intragenic	**1.7121**	**2.00E-08**	0.942	0.725264
	Intergenic	0.2561	0	0.2471	0
GSE5816 Lung Adenocarcinoma	All	0.1137	0	0.2313	0
	Sense	0.3274	0	0.4716	0
	Anti-sense	0.3159	0	0.4451	0
	Intragenic	1.0615	0.122261	**1.2789**	**0.000367**
	Intergenic	0.125	0	0.2425	0
GSE6919 Metastasis prostate cancer	All	0.4188	0	0.2975	0
	Sense	0.6889	3.79E-05	0.4986	0
	Anti-sense	0.6744	2.23E-05	0.5285	0
	Intragenic	**1.2899**	**0.000589**	0.9843	0.836614
	Intergenic	0.4436	0	0.3089	0
GSE9750 cervical cancer	All	0.1815	0	0.3426	0
	Sense	0.4718	0	0.4749	0
	Anti-sense	0.4788	0	0.4665	0
	Intragenic	**1.2248**	**0**	0.8118	0.001078
	Intergenic	0.1969	0	0.341	0
GSE9764 5-azacydine treated human mesenchymal stem cells	All	0.2158	0	0.1822	0
	Sense	0.417	0	0.3709	0
	Anti-sense	0.4198	0	0.3443	0
	Intragenic	0.9308	0.336312	1.0843	0.38765109
	Intergenic	0.2247	0	0.189	0

19C. ERVL-MaLR

GEO accession	HERV pattern	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE13887 SLE T cells	All	0.1296	0	0.094	0
	Sense	0.1998	0	0.1743	0
	Anti-sense	0.1793	0	0.1561	0
	Intragenic	0.6399	0	**1.2717**	**3.22E-05**
	Intergenic	0.1304	0	0.0949	0
GSE10500 RA macophage cells	All	0.291	0	0.2242	0
	Sense	0.3753	0	0.3682	0
	Anti-sense	0.3535	0	0.3363	0
	Intragenic	0.8763	0.049067	**1.2469**	**0.000185**
	Intergenic	0.2911	0	0.2233	0
GSE14905 Psoriasis, skin	All	0.1578	0	0.1571	0
	Sense	0.2653	0	0.2571	0
	Anti-sense	0.2449	0	0.2217	0
	Intragenic	**1.4491**	**0**	0.9605	0.490139
	Intergenic	0.16	0	0.1594	0
GSE52471 psoriasis, skin	All	0.1586	0	0.2465	0
	Sense	0.2502	0	0.3885	0
	Anti-sense	0.2406	0	0.3389	0
	Intragenic	**1.3569**	**1.00E-08**	1.0493	0.44043
	Intergenic	0.1609	0	0.2496	0
GSE12453 Hodgkin Lymphoma vs centroblasts	All	0.1029	0	0.2362	0
	Sense	0.189	0	0.3533	0
	Anti-sense	0.1668	0	0.3271	0
	Intragenic	**1.5389**	**0**	0.7286	0.000346
	Intergenic	0.1045	0	0.239	0
GSE12453 Hodgkin Lymphoma vs centrocytes	All	0.1182	0	0.2206	0
	Sense	0.2095	0	0.3148	0
	Anti-sense	0.1846	0	0.308	0
	Intragenic	**1.6388**	**0**	0.5651	0
	Intergenic	0.12	0	0.2233	0
GSE12453 Hodgkin Lymphoma vs memory B cells	All	0.1005	0	0.1984	0
	Sense	0.168	0	0.2935	0
	Anti-sense	0.1495	0	0.2667	0
	Intragenic	**1.5216**	**1.65E-06**	0.5852	0
	Intergenic	0.1006	0	0.201	0

GEO accession	HERV pattern	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE12453 Hodgkin Lymphoma vs plasma Cells	All	0.1083	0	0.2045	0
	Sense	0.1856	0	0.2924	0
	Anti-sense	0.1723	0	0.2659	0
	Intragenic	**1.4366**	**1.70E-07**	0.5664	0
	Intergenic	0.1098	0	0.207	0
GSE12453 Diffuse Large B cells Lymphoma vs centroblasts	All	0.121	0	0.1683	0
	Sense	0.2072	0	0.2567	0
	Anti-sense	0.1874	0	0.2301	0
	Intragenic	**1.8389**	**0**	0.517	0
	Intergenic	0.1226	0	0.1705	0
GSE12453 Diffuse Large B cells Lymphoma vs centrocytes	All	0.1152	0	0.1512	0
	Sense	0.199	0	0.2303	0
	Anti-sense	0.1781	0	0.2034	0
	Intragenic	**1.92**	**0**	0.4441	0
	Intergenic	0.1167	0	0.1532	0
GSE12453 Diffuse Large B cells Lymphoma vs memory B cells	All	0.1172	0	0.1305	0
	Sense	0.1908	0	0.2053	0
	Anti-sense	0.1728	0	0.187	0
	Intragenic	**1.7667**	**6.56E-06**	0.5253	0
	Intergenic	0.1186	0	0.1326	0
GSE12453 Diffuse Large B cells Lymphoma vs naive B cells	All	0.0856	0	0.1132	0
	Sense	0.1394	0	0.1954	0
	Anti-sense	0.1314	0	0.1701	0
	Intragenic	**1.5074**	**0.000373**	0.5802	0
	Intergenic	0.0867	0	0.1149	0
GSE45829 EBV infected B cells	All	0.2945	0	0.2214	0
	Sense	0.4038	1.10E-07	0.3274	0
	Anti-sense	0.3851	7.00E-08	0.3153	0
	Intragenic	0.6928	0.000674	**1.3029**	**0**
	Intergenic	0.2979	0	0.2223	0
GSE1299 Breast cancer cells	All	0.1334	0	0.1571	0
	Sense	0.2207	0	0.257	0
	Anti-sense	0.1923	0	0.2411	0
	Intragenic	**1.5123**	**8.75E-06**	1.1619	0.265316
	Intergenic	0.135	0	0.159	0
GSE9750 cervical cancer	All	0.0537	0	0.2427	0
	Sense	0.1685	0	0.3198	0
	Anti-sense	0.1426	0	0.3066	0
	Intragenic	**1.1485**	**1.34E-05**	0.8536	0.006343
	Intergenic	0.0562	0	0.2417	0

GEO accession	HERV pattern	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE9764 5-azacydine treated human mesenchymal stem cells	All	0.0981	0	0.1015	0
	Sense	0.1659	0	0.1705	0
	Anti-sense	0.1499	0	0.1502	0
	Intragenic	1.1619	0.0672580	1.0948	0.3246718
	Intergenic	0.0994	0	0.1028	0

The analysis results in Table 18 showed that there was no association between HERV and gene expression in SLE at entire HERV neighboring genes but we found the association between intragenic ERV1/ERVE, ERV3/ERVL, and ERVL-MaLR superfamilies with the up-regulated gene in SLE T cell and PBMC with RNP+ conditions, while there was no association found in ERV2/ERVK superfamily (Table 19). This observation is interestingly correlated with our previous studies that hypomethylation of HERV-E was detected in SLE CD4+ T cell but not HERV-K [18]. This specific hypomethylation is also associated with up-regulation of HERV-E transcript in CD4+ T cells [172]. Moreover, our result showed a strong association particularly with SLE with RNP+. This is consistent with the fact that there is sequence homology between HRES-1 and the 70-kDa *gag*-related region of the sn-RNP supporting that possible mechanism in etiopathogenesis of SLE by inducing the cross-reaction between the two proteins by autoantibodies.

We found that intragenic HERV in both entire HERV and superfamilies classification level were associated with down-regulated gene in most of cancer conditions. This finding might suggest that not only LINE-1 are associated with cancer genome wide hypomethylation down-regulating genes as previous report [173], but genes containing intragenic HERVs are also highly associate with down-regulated genes under cancer conditions. Another interesting point is the pattern of association. In contrary to the association with SLE, which were with up-regulated genes suggesting a clear different in pathogenesis of how HERV affect gene regulation in these 2 groups of disease. It should be noted that HERV also associated with up-regulted genes in B cells with EBV infection. This observation is interesting due to the fact that EBV has been implicated as a major risk factor for SLE. In addition, there was no striking

association between HERV and immune cells from other immune-mediated diseases that we analyzed e.g., asthma, graves' disease, rheumatoid arthritis (except for macrophage).

As for the targeted tissue in autoimmune diseases, we found that the pattern of association in psoriasis skin tissue was similar to the pattern in cancer suggesting similar mechanism. However, we did not see any association between HERV in kidney tissue from SLE. It seems that the mechanism that HERV might have in SLE is mainly in the immune cells and have some specificity with certain HERV as well.

Table 20 Association analysis results at individual HERV with OR >_1 and p <0.001, indicated in bold letter (Only association data were shown in this table)

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
LTR12					
GSE45829 EBV infected B cells	All	0.588	0.057009	0.9963	1
	Sense	0.5965	0.225575	1.0031	0.950761
	Anti-sense	0.6318	0.278349	0.9708	0.898923
	Intragenic	0.2979	0.270368	**1.0969**	**0.647804**
	Intergenic	0.6483	0.159074	1.0146	0.885635
LTR12C					
GSE6919 Metastasis prostate cancer	All	1.0631	0.528835	1.1989	0.029162
	Sense	1.0277	0.799053	**1.4276**	**0.000687**
	Anti-sense	1.1335	0.286361	1.0914	0.428592
	Intragenic	1.8283	0.004252	1.1755	0.41821
	Intergenic	1.011	0.920935	1.2282	0.015741
LTR7					
GSE10325 SLE myeloid cells	All	0.6176	0.315517	1.9198	0.00177
	Sense	0.7588	0.822676	2.1462	0.00468
	Anti-sense	0.406	0.248602	1.6445	0.102023
	Intragenic	1.6009	0.359273	1.9322	0.160718
	Intergenic	0.4456	0.116453	**2.1057**	**0.000603**
GSE24706 SLE PBMC, ANA	All	0	1	3.2302	0.00132
	Sense	0	1	2.5107	0.05714
	Anti-sense	0	1	3.9256	0.003117
	Intragenic	0	1	6.2826	0.013876
	Intergenic	0	1	**3.5384**	**0.000655**

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE61635 SLE PBMC RNP+	All	0.7021	0.149413	1.0698	0.638086
	Sense	0.641	0.215454	0.6875	0.142685
	Anti-sense	0.6952	0.388232	1.4125	0.080671
	Intragenic	0.8994	1	**3.1509**	**8.52E-05**
	Intergenic	0.6359	0.085513	0.9862	1
GSE6740 HIV infected cd8 non-pregressive	All	1.7767	0.068882	0.8392	0.847136
	Sense	1.1352	0.78288	1.0273	0.797975
	Anti-sense	2.213	0.046591	0.5477	0.591067
	Intragenic	**7.5414**	**0.000223**	2.1599	0.241396
	Intergenic	1.7539	0.089586	0.6033	0.42085
GSE22859 H3K4me2, HeLa	All	0.9465	0.853674	1.1964	0.339477
	Sense	0.622	0.138268	0.9181	0.887625
	Anti-sense	1.2613	0.304893	1.7086	0.026196
	Intragenic	0.9559	1	**3.7344**	**0.000337**
	Intergenic	1.0012	1	1.1612	0.438858
LTR2					
GSE20864 SLE PBMC	All	7.0415	0.055184	0.9924	1
	Sense	6.2423	0.172558	0.4887	0.202822
	Anti-sense	12.7467	0.019356	1.3547	0.253101
	Intragenic	**65.4884**	**0.000855**	1.0141	0.725561
	Intergenic	3.0157	0.31986	0.9298	0.897101
LTR8					
GSE13887 SLE T cells	All	0.7988	0.012832	0.9589	0.619813
	Sense	0.8937	0.353236	0.8926	0.299009
	Anti-sense	0.7485	0.012909	1.0088	0.923736
	Intragenic	0.48	0.011516	**1.8072**	**0.000332**
	Intergenic	0.8243	0.037693	0.8783	0.117034
GSE9750 cervical cancer	All	0.968	0.450923	0.8761	0.096783
	Sense	0.9802	0.74477	0.9529	0.660108
	Anti-sense	1.0036	0.937214	0.8075	0.037212
	Intragenic	**1.413**	**0.000614**	0.5184	0.008094
	Intergenic	0.9461	0.204692	0.9106	0.261139
MER4C					
GSE6919 Metastasis prostate cancer	All	1.3223	0.022372	0.862	0.275799
	Sense	1.2531	0.185942	0.8784	0.551222
	Anti-sense	1.3631	0.054162	0.8562	0.387764
	Intragenic	**2.7771**	**3.35E-05**	0.8118	0.643995
	Intergenic	1.2065	0.161647	0.8859	0.41281

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
MER39					
GSE61635 SLE PBMC RNP+	All	1.0281	0.835094	1.3053	0.007087
	Sense	0.8961	0.640778	1.1125	0.434509
	Anti-sense	1.0667	0.662871	**1.5765**	**0.00011**
	Intragenic	1.5226	0.115824	**3.1201**	**0**
	Intergenic	1.0261	0.82779	1.1429	0.204549
GSE14905 Psoriasis, skin	All	1.1447	0.124687	0.8717	0.157811
	Sense	1.0551	0.627061	0.7533	0.042565
	Anti-sense	1.3202	0.008843	1.0228	0.818633
	Intragenic	**1.9111**	**9.98E-05**	0.7725	0.305462
	Intergenic	1.0669	0.478771	0.9016	0.316747
GSE36474 myeloma, bone marrow	All	0.8048	0.214075	**1.6484**	**0.000398**
	Sense	0.7584	0.294309	**1.8778**	**0.000484**
	Anti-sense	1.0212	0.921962	1.3852	0.069275
	Intragenic	0.8173	0.72804	1.7934	0.044624
	Intergenic	0.8522	0.415841	1.5913	0.001839
GSE6919 Metastasis prostate cancer	All	1.3157	0.009408	1.1032	0.310102
	Sense	1.2719	0.081014	1.2698	0.05571
	Anti-sense	1.3566	0.018969	0.9665	0.848034
	Intragenic	**2.2615**	**3.41E-05**	0.8269	0.567387
	Intergenic	1.2538	0.039063	1.1245	0.242868
GSE9750 cervical cancer	All	**1.2051**	**0.000136**	0.7365	0.002259
	Sense	1.2093	0.003817	0.8379	0.197778
	Anti-sense	**1.2356**	**0.000525**	0.6407	0.000709
	Intragenic	**1.7392**	**6.00E-08**	0.4917	0.007568
	Intergenic	1.1684	0.002579	0.7608	0.008753
GSE13911 Microsatellite i nstable gastric cancer	All	1.0478	0.505376	0.8147	0.06519
	Sense	1.0783	0.422201	0.8525	0.301262
	Anti-sense	1.0596	0.50805	0.8711	0.333987
	Intragenic	**1.6944**	**0.000105**	0.4858	0.020551
	Intergenic	0.9881	0.912398	0.8278	0.107861
MER52A					
GSE61635 SLE PBMC RNP+	All	0.7389	0.118091	1.2736	0.039187
	Sense	0.4491	0.007372	1.206	0.236743
	Anti-sense	1.0608	0.740231	1.281	0.109835
	Intragenic	1.3062	0.435661	**2.5372**	**2.18E-05**
	Intergenic	0.7104	0.085716	1.1894	0.15098

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE22859 H3K4me2, HeLa	All	0.6984	0.033059	1.0409	0.799671
	Sense	0.5753	0.028201	0.9318	0.908788
	Anti-sense	0.7945	0.323749	1.1299	0.568387
	Intragenic	1.1507	0.604114	**2.7278**	**0.000368**
	Intergenic	0.6004	0.005513	0.9402	0.791115
GSE41040 H3K9me3 primary fibroblasts	All	0.7543	0.046102	0.9224	0.456476
	Sense	0.6936	0.063892	0.8217	0.201587
	Anti-sense	0.8768	0.50327	1.0064	0.946699
	Intragenic	0.6403	0.243601	**2.1863**	**2.12E-05**
	Intergenic	0.7931	0.117455	0.8412	0.131389
MER11C					
GSE32591 LN glomolular	All	0.6762	0.559803	**2.0077**	**0.000612**
	Sense	0.8736	1	1.2444	0.470698
	Anti-sense	0.488	0.443655	**2.7174**	**3.45E-05**
	Intragenic	0	0.629955	2.508	0.037807
	Intergenic	0.7458	0.686787	1.9789	0.001418
LTR16A1					
GSE61635 SLE PBMC RNP+	All	0.8284	0.217784	1.0021	0.958538
	Sense	0.6825	0.083285	1.0822	0.53428
	Anti-sense	0.9355	0.788665	1.0411	0.734174
	Intragenic	1.0908	0.757381	**2.5649**	**7.00E-08**
	Intergenic	0.7724	0.114719	0.9256	0.514124
GSE52471 SLE/DLE, skin	All	1.0323	0.728231	0.9859	0.959313
	Sense	1.0599	0.644252	0.9062	0.542037
	Anti-sense	1.1564	0.242536	0.9945	1
	Intragenic	**1.8645**	**0.000722**	1.0411	0.818851
	Intergenic	1.0302	0.755389	0.9608	0.749064
GSE13355 Psoriasis, skin	All	1.1037	0.519014	0.8574	0.291982
	Sense	1.0188	0.914653	1.0449	0.792755
	Anti-sense	1.2974	0.170418	0.6602	0.047095
	Intragenic	**2.6631**	**0.000165**	0.4086	0.037415
	Intergenic	0.8939	0.612707	0.8889	0.448736
GSE14905 Psoriasis, skin	All	1.0078	0.929001	0.8185	0.038477
	Sense	1.031	0.767577	0.9545	0.765204
	Anti-sense	1.0321	0.771093	0.667	0.002678
	Intragenic	**1.8052**	**0.000416**	0.6315	0.068406
	Intergenic	0.9606	0.709379	0.8706	0.172436

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE9750 cervical cancer	All	1.1005	0.051774	0.7483	0.003007
	Sense	1.0934	0.164609	0.7995	0.094565
	Anti-sense	1.1239	0.06815	0.7169	0.011551
	Intragenic	**1.7822**	**1.00E-08**	0.2161	2.51E-06
	Intergenic	1.059	0.262267	0.8429	0.090681
GSE13911 Microsatellite i nstable gastric cancer	All	0.9861	0.862713	0.9541	0.686066
	Sense	0.9998	1	0.9942	1
	Anti-sense	1.1252	0.175183	0.9912	1
	Intragenic	**1.719**	**4.95E-05**	0.9179	0.820658
	Intergenic	0.9374	0.405268	0.9954	1
GSE6740 HIV infect CD, acute	All	1.4411	0.04277	0.9181	0.814226
	Sense	1.1552	0.589299	0.7152	0.434822
	Anti-sense	1.6273	0.032892	1.0604	0.759368
	Intragenic	**3.4224**	**4.87E-05**	1.3696	0.423472
	Intergenic	1.2732	0.204212	0.7924	0.460108
GSE9764 5-aza Human mesenchymal stem cells	All	0.9617	0.799527	0.8722	0.400549
	Sense	0.9938	1	0.7829	0.264599
	Anti-sense	0.984	1	1.0164	0.927034
	Intragenic	**2.2079**	**0.000298**	1.4462	0.156837
	Intergenic	0.8362	0.231131	0.8538	0.341195
GSE41040 H3K9me3 primary fibroblasts	All	0.787	0.033341	1.0495	0.562006
	Sense	0.8681	0.370229	0.9991	1
	Anti-sense	0.6602	0.008097	1.162	0.144064
	Intragenic	0.8124	0.485087	**1.9861**	**6.45E-06**
	Intergenic	0.8063	0.073003	0.9555	0.633868
LTR33					
GSE13887 SLE T cells	All	0.5788	0	0.9476	0.39222
	Sense	0.6473	1.15E-06	0.9648	0.660431
	Anti-sense	0.5932	0	1.0588	0.411931
	Intragenic	0.6178	0.002659	**1.5445**	**7.37E-05**
	Intergenic	0.6	0	0.9444	0.367537
GSE61635 SLE PBMC RNP+	All	0.7531	0.003186	0.9968	1
	Sense	0.7278	0.008059	0.9848	0.899908
	Anti-sense	0.9095	0.419935	1.1664	0.05465
	Intragenic	1.3428	0.076966	**2.2835**	**0**
	Intergenic	0.7696	0.007561	0.928	0.318866
GSE52471 SLE/DLE, skin	All	1.0141	0.84115	0.9353	0.354274
	Sense	0.9959	1	0.9082	0.267219
	Anti-sense	1.0884	0.274659	0.9667	0.718433
	Intragenic	**1.7442**	**1.46E-06**	0.7864	0.127286
	Intergenic	1.0199	0.785755	0.9609	0.600688

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE10500 RA macophage cells	All	0.7193	5.10E-06	1.1711	0.010493
	Sense	0.7496	0.001251	1.0656	0.390045
	Anti-sense	0.7321	0.000331	**1.3025**	**0.000149**
	Intragenic	0.7218	0.043088	**1.676**	**2.74E-06**
	Intergenic	0.7445	6.20E-05	1.1479	0.028514
GSE13355 Psoriasis, skin	All	**1.4185**	**0.000923**	0.84	0.056759
	Sense	1.3337	0.019228	0.8668	0.203045
	Anti-sense	**1.5021**	**0.000604**	0.8245	0.078636
	Intragenic	**2.5649**	**2.00E-08**	0.722	0.12761
	Intergenic	1.3879	0.002006	0.8738	0.149327
GSE14905 Psoriasis, skin	All	1.167	0.009112	0.853	0.010123
	Sense	1.0426	0.565719	0.834	0.016757
	Anti-sense	**1.3348**	**1.40E-05**	0.8452	0.023709
	Intragenic	**2.0279**	**0**	0.6744	0.004891
	Intergenic	1.1387	0.030464	0.8761	0.036377
GSE52471 psoriasis, skin	All	1.063	0.268151	0.9089	0.153149
	Sense	1.0239	0.715307	0.8766	0.110724
	Anti-sense	1.1269	0.060725	0.8774	0.102993
	Intragenic	**1.8729**	**0**	0.6942	0.01516
	Intergenic	1.0579	0.310849	0.9535	0.488124
GSE1299 Breast cancer cells	All	1.1136	0.279978	0.8242	0.190319
	Sense	1.1279	0.289802	0.847	0.364553
	Anti-sense	1.1861	0.121627	0.8984	0.573273
	Intragenic	**1.8112**	**0.000328**	1.091	0.677073
	Intergenic	1.065	0.516625	0.8721	0.362541
GSE3167 Bladder carcinoma Situ	All	1.1645	0.204846	0.7976	0.000431
	Sense	1.0459	0.772263	0.8011	0.005121
	Anti-sense	1.2726	0.077565	0.8142	0.00712
	Intragenic	**2.4267**	**2.34E-06**	0.3601	0
	Intergenic	1.1301	0.325738	0.8492	0.012329
GSE5764 Ductal and lobular breast cancer	All	0.9736	0.696914	0.9407	0.457752
	Sense	1.0124	0.876385	0.922	0.426029
	Anti-sense	1.0293	0.704636	1.0569	0.552916
	Intragenic	**1.6294**	**1.89E-05**	1.105	0.478299
	Intergenic	0.9325	0.306059	0.9508	0.551025
GSE6919 Metastasis prostate cancer	All	1.2119	0.008807	0.8679	0.039818
	Sense	1.1261	0.178631	0.8365	0.035761
	Anti-sense	1.3024	0.001633	0.9338	0.408679
	Intragenic	**1.8298**	**2.30E-06**	0.9538	0.785668
	Intergenic	1.1958	0.016492	0.8923	0.107127

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE9750 cervical cancer	All	1.0075	0.827404	0.6142	0
	Sense	1.0312	0.445202	0.7043	8.17E-06
	Anti-sense	1.0662	0.099951	0.6036	0
	Intragenic	**1.4915**	**0**	0.4816	1.30E-06
	Intergenic	0.9946	0.877992	0.6349	0
GSE13911 Microsatellite instable gastric cancer	All	0.8917	0.015501	0.6966	3.90E-07
	Sense	0.8839	0.031838	0.676	1.06E-05
	Anti-sense	0.9856	0.806834	0.7835	0.003742
	Intragenic	**1.5809**	**6.00E-08**	0.7342	0.046528
	Intergenic	0.8606	0.001926	0.7138	3.48E-06
GSE9764 5-aza Human mesenchymal stem cells	All	0.8854	0.17117	0.9846	0.887743
	Sense	0.9315	0.538691	1.1031	0.366957
	Anti-sense	0.999	1	0.9164	0.4747
	Intragenic	**1.7753**	**8.97E-05**	1.4892	0.017116
	Intergenic	0.8284	0.036824	0.9475	0.598481
GSE22859 H3K4me2, HeLa	All	0.8033	0.010604	0.796	0.020196
	Sense	0.7839	0.023082	0.8533	0.191701
	Anti-sense	0.9147	0.386723	0.9205	0.506452
	Intragenic	**1.6586**	**0.00044**	1.3561	0.08387
	Intergenic	0.7681	0.002543	0.8443	0.091549
GSE41040 H3K9me3 primary fibroblasts	All	0.7016	1.38E-06	0.8248	0.000747
	Sense	0.8371	0.044654	0.8756	0.057115
	Anti-sense	0.6967	4.43E-05	0.9481	0.435065
	Intragenic	0.7605	0.090378	**1.8866**	**0**
	Intergenic	0.7036	2.66E-06	0.7888	5.04E-05
LTR52					
GSE61635 SLE PBMC RNP+	All	0.5389	0.034318	1.2284	0.178167
	Sense	0.4592	0.084173	1.1859	0.406151
	Anti-sense	0.5815	0.180523	1.2688	0.257472
	Intragenic	0.8691	1	**3.442**	**1.12E-05**
	Intergenic	0.5466	0.047286	0.9852	1
GSE13355 Psoriasis, skin	All	1.3771	0.151546	1.4936	0.032182
	Sense	1.4587	0.198488	0.7734	0.547764
	Anti-sense	1.3087	0.378401	**2.1262**	**0.000796**
	Intragenic	2.5009	0.038427	1.3564	0.427121
	Intergenic	0.5466	0.047286	0.9852	1

MLT1D					
GSE4588 SLE CD4 CELLS	All	0.6251	0	0.5544	0
	Sense	0.7361	5.70E-05	0.6886	2.40E-07
	Anti-sense	0.6926	1.28E-06	0.7043	8.50E-07
	Intragenic	**1.5068**	**4.75E-05**	1.0556	0.602307
	Intergenic	0.6351	0	0.5612	0
GSE13887 SLE T cells	All	0.5382	0	0.7972	0.000106
	Sense	0.6462	0	0.929	0.223066
	Anti-sense	0.5835	0	0.9452	0.345224
	Intragenic	0.575	1.50E-06	**1.6388**	**0**
	Intergenic	0.5734	0	0.7905	5.11E-05
GSE61635 SLE PBMC RNP+	All	0.7067	8.00E-05	0.5806	0
	Sense	0.8561	0.097383	0.7262	5.77E-06
	Anti-sense	0.6927	7.33E-05	0.784	0.000437
	Intragenic	1.3504	0.015155	**1.8107**	**0**
	Intergenic	0.7042	6.30E-05	0.582	0
GSE52471 SLE/DLE, skin	All	0.8884	0.067435	0.7975	0.000538
	Sense	0.9355	0.313984	0.8013	0.001203
	Anti-sense	1.0781	0.245613	0.8924	0.091203
	Intragenic	**1.4915**	**7.98E-06**	0.877	0.228007
	Intergenic	0.874	0.035636	0.8238	0.002927
GSE4588 RA B CELLS	All	0.6872	0	0.768	6.47E-05
	Sense	0.733	3.00E-06	0.7948	0.000955
	Anti-sense	0.9039	0.125289	0.8803	0.061172
	Intragenic	**1.4217**	**8.42E-05**	1.1882	0.075823
	Intergenic	0.6946	1.00E-08	0.7958	0.000526
GSE10500 RA macophage cells	All	0.7511	1.39E-05	0.9982	0.976111
	Sense	0.7924	0.0007	0.9981	1
	Anti-sense	0.8806	0.062151	1.0937	0.139932
	Intragenic	1.0295	0.761934	**1.3521**	**0.000426**
	Intergenic	0.7688	5.99E-05	0.9928	0.905399
GSE13355 Psoriasis, skin	All	1.0157	0.917126	0.8286	0.026867
	Sense	1.1632	0.154589	0.7715	0.003844
	Anti-sense	1.1245	0.27178	0.9873	0.897926
	Intragenic	**1.8159**	**1.54E-05**	0.7753	0.079562
	Intergenic	0.9884	0.917816	0.8677	0.091425
GSE14905 Psoriasis, skin	All	0.8796	0.024826	0.6957	0
	Sense	0.9151	0.137179	0.756	3.76E-06
	Anti-sense	0.993	0.930837	0.7801	3.30E-05
	Intragenic	**1.6377**	**0**	0.8593	0.110741
	Intergenic	0.8643	0.011036	0.7123	0

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE52471 psoriasis, skin	All	0.8582	0.003758	0.8484	0.008041
	Sense	0.8908	0.035673	0.8463	0.009609
	Anti-sense	1.034	0.538712	0.9684	0.616162
	Intragenic	**1.5304**	**1.00E-08**	0.7852	0.019755
	Intergenic	0.8386	0.000806	0.8966	0.07792
GSE12453 Diffuse Large B cells Lymphoma vs naive B cells	All	0.6346	7.96E-05	0.5924	0
	Sense	0.6755	0.001533	0.6348	0
	Anti-sense	0.7797	0.041782	0.6773	0
	Intragenic	**1.8252**	**9.09E-05**	0.4857	0
	Intergenic	0.6169	2.62E-05	0.6285	0
GSE3167 Bladder carcinoma Situ	All	0.9647	0.772034	0.6654	0
	Sense	0.9555	0.724828	0.7095	5.00E-08
	Anti-sense	1.1424	0.267964	0.7497	3.08E-06
	Intragenic	**1.802**	**0.000126**	0.4347	0
	Intergenic	0.9603	0.730288	0.6945	0
GSE5764 Ductal and lobular breast cancer	All	0.7293	3.00E-07	0.778	0.000725
	Sense	0.8393	0.006605	0.8748	0.087126
	Anti-sense	0.8942	0.079233	0.8632	0.059101
	Intragenic	**1.4571**	**1.46E-05**	1.3294	0.007122
	Intergenic	0.7482	2.50E-06	0.7991	0.002543
GSE5816 Lung Adenocarcinoma	All	0.65	0	0.8063	0.001141
	Sense	0.7683	0	0.9338	0.329899
	Anti-sense	0.763	0	0.9054	0.142357
	Intragenic	**1.1993**	**0.000715**	**1.6143**	**1.30E-07**
	Intergenic	0.6594	0	0.7738	8.88E-05
GSE6919 Metastasis prostate cancer	All	1.041	0.590095	0.8276	0.00322
	Sense	1.0119	0.884289	0.7713	0.000123
	Anti-sense	1.0835	0.263031	1.0201	0.769649
	Intragenic	**1.4149**	**0.000546**	1.111	0.277505
	Intergenic	1.0738	0.335873	0.8229	0.002279
GSE9750 cervical cancer	All	0.8175	0	0.7269	3.00E-08
	Sense	0.9055	0.002546	0.7338	3.50E-07
	Anti-sense	0.9371	0.0464	0.8468	0.005085
	Intragenic	**1.4869**	**0**	0.7107	0.000465
	Intergenic	0.8256	0	0.7567	1.11E-06
GSE13911 Microsatellite i nstable gastric cancer	All	0.6558	0	0.6507	0
	Sense	0.7866	2.10E-07	0.7039	3.40E-07
	Anti-sense	0.8125	5.59E-06	0.7138	6.90E-07
	Intragenic	**1.4879**	**0**	0.8014	0.041173
	Intergenic	0.6706	0	0.6677	0

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE6740 HIV infected CD4, chronic	All	1.1511	0.467167	0.6236	0.000319
	Sense	1.4127	0.064342	0.5782	0.000153
	Anti-sense	1.2402	0.234641	0.9928	1
	Intragenic	**2.4799**	**6.44E-05**	1.1321	0.542474
	Intergenic	1.0762	0.71817	0.6118	0.000208
GSE6740 HIV infected cd8acute	All	1.0507	0.56373	0.6607	5.54E-06
	Sense	1.0779	0.380205	0.7037	0.000306
	Anti-sense	1.0797	0.361499	0.7812	0.010477
	Intragenic	**1.5386**	**0.000151**	1.0595	0.672164
	Intergenic	1.0372	0.682347	0.6736	1.53E-05
GSE9764 5-aza Human mesenchymal stem cells	All	0.7757	0.001791	0.6591	3.01E-06
	Sense	0.872	0.113916	0.6778	5.33E-05
	Anti-sense	0.9521	0.563501	0.8982	0.255736
	Intragenic	**1.6851**	**2.34E-06**	1.4907	0.001457
	Intergenic	0.7642	0.000954	0.6329	2.90E-07
GSE59695 H3K4me1, HepG2	All	0.6533	0.005627	0.6999	0.002476
	Sense	0.5703	0.000805	0.8338	0.149986
	Anti-sense	0.8639	0.396453	0.8413	0.170756
	Intragenic	1.786	0.004612	**1.7042**	**0.00077**
	Intergenic	0.6019	0.000795	0.7141	0.00396
GSE41040 H3K9me3 primary fibroblasts	All	0.6551	0	0.6739	0
	Sense	0.7023	5.70E-07	0.759	7.00E-07
	Anti-sense	0.7889	0.000582	0.832	0.000771
	Intragenic	1.1246	0.239605	**1.5402**	**1.00E-08**
	Intergenic	0.649	0	0.6671	0
MLT2B3					
GSE61635 SLE PBMC RNP+	All	1.0079	0.945104	1.1858	0.084327
	Sense	0.8136	0.346161	1.0157	0.88592
	Anti-sense	1.2233	0.211426	1.2651	0.066404
	Intragenic	1.3211	0.31275	**2.3812**	**8.61E-06**
	Intergenic	1.0364	0.774395	1.0956	0.383253
GSE9750 cervical cancer	All	1.0237	0.633888	0.6845	0.000172
	Sense	1.0123	0.863515	0.6976	0.010944
	Anti-sense	1.0541	0.416181	0.6377	0.000998
	Intragenic	**1.6021**	**2.23E-05**	0.3503	0.000854
	Intergenic	0.9971	0.979169	0.7374	0.003485
GSE6740 HIV infected cd8 non-pregressive	All	1.1855	0.498438	1.0611	0.796546
	Sense	1.3135	0.355953	1.0569	0.858927
	Anti-sense	1.2833	0.381855	1.0423	0.866794
	Intragenic	**4.579**	**0.000134**	1.1621	0.74662
	Intergenic	0.9888	1	1.0305	0.89313

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
MSTD					
GSE13887 SLE T cells	All	0.6884	3.40E-07	0.9926	0.925094
	Sense	0.7589	0.002699	0.9674	0.698192
	Anti-sense	0.6705	1.14E-05	1.0533	0.477425
	Intragenic	0.6178	0.002659	**1.6122**	**1.07E-05**
	Intergenic	0.7101	5.37E-06	0.9925	0.923225
GSE61635 SLE PBMC RNP+	All	0.7506	0.003933	0.9066	0.184234
	Sense	0.7713	0.040789	0.8679	0.131087
	Anti-sense	0.8152	0.090998	1.0344	0.699864
	Intragenic	1.1635	0.40185	**1.7424**	**5.22E-06**
	Intergenic	0.74	0.003623	0.8486	0.032749
GSE10500 RA macophage cells	All	0.7617	0.000231	1.1879	0.005926
	Sense	0.801	0.018015	1.0401	0.606869
	Anti-sense	0.7268	0.000454	**1.3239**	**0.000113**
	Intragenic	0.7034	0.030536	1.4477	0.001039
	Intergenic	0.7954	0.002592	1.1205	0.076597
GSE13355 Psoriasis, skin	All	1.4036	0.001511	0.9534	0.617397
	Sense	1.3361	0.023196	0.8914	0.339489
	Anti-sense	1.409	0.005287	0.9988	1
	Intragenic	**2.2202**	**4.77E-06**	0.6006	0.019653
	Intergenic	1.2745	0.030159	1.006	0.96284
GSE14905 Psoriasis, skin	All	1.1545	0.017922	0.8014	0.000569
	Sense	1.1843	0.02036	0.8683	0.083918
	Anti-sense	1.1947	0.012591	0.778	0.001441
	Intragenic	**1.82**	**1.00E-08**	0.6051	0.00048
	Intergenic	1.1121	0.08883	0.8369	0.006839
GSE52471 psoriasis, skin	All	1.011	0.842488	0.9673	0.641128
	Sense	0.9816	0.83302	0.9085	0.26618
	Anti-sense	1.0674	0.325912	0.9971	1
	Intragenic	**1.6124**	**1.36E-06**	0.8255	0.189948
	Intergenic	0.9853	0.815895	0.9996	1
GSE12453 Diffuse Large B cells Lymphoma vs naive B cells	All	1.0714	0.578657	0.7483	0
	Sense	1.0247	0.878534	0.7363	1.90E-07
	Anti-sense	1.1116	0.461325	0.7693	2.92E-06
	Intragenic	**1.9287**	**0.000923**	0.531	0
	Intergenic	1.0213	0.849682	0.7882	4.60E-07
GSE3167 Bladder carcinoma Situ	All	1.2159	0.107336	0.9312	0.277678
	Sense	1.0328	0.81826	0.8871	0.154769
	Anti-sense	1.3841	0.017929	0.9379	0.423772
	Intragenic	**1.9427**	**0.000865**	0.7503	0.044021
	Intergenic	1.1669	0.227536	0.9561	0.512704

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE5764 Ductal and lobular breast cancer	All	0.998	1	0.9073	0.245035
	Sense	1.0019	0.96715	0.8644	0.165118
	Anti-sense	1.0936	0.249292	0.9798	0.885887
	Intragenic	**1.8442**	**4.00E-08**	1.0546	0.693529
	Intergenic	0.9582	0.561721	0.9195	0.324232
GSE5816 Lung Adenocarcinoma	All	0.8113	1.50E-07	0.9525	0.523639
	Sense	0.8458	0.000764	0.981	0.860836
	Anti-sense	0.8618	0.001896	0.957	0.641937
	Intragenic	1.2634	0.001311	**1.5385**	**0.000467**
	Intergenic	0.794	2.00E-08	0.93	0.344973
GSE9750 cervical cancer	All	0.9826	0.617894	0.7364	2.39E-06
	Sense	0.9548	0.287256	0.6834	5.79E-06
	Anti-sense	1.0322	0.435496	0.7857	0.002131
	Intragenic	**1.4422**	**1.00E-08**	0.6173	0.000765
	Intergenic	0.9688	0.377944	0.7734	0.000105
GSE13911 Microsatellite i nstable gastric cancer	All	0.7646	6.00E-08	0.8012	0.002147
	Sense	0.8061	0.000545	0.7998	0.015631
	Anti-sense	0.8676	0.016718	0.8259	0.03046
	Intragenic	**1.4707**	**6.15E-06**	0.7163	0.033224
	Intergenic	0.7293	0	0.8311	0.012355
GSE9764 5-aza Human mesenchymal stem cells	All	0.8592	0.095628	0.9723	0.809207
	Sense	0.884	0.3023	1.0076	0.952336
	Anti-sense	1.043	0.675569	0.8994	0.41961
	Intragenic	**1.6429**	**0.00096**	0.895	0.635279
	Intergenic	0.8335	0.053146	1.0271	0.804472
GSE41040 H3K9me3 primary fibroblasts	All	0.7534	0.000167	0.9005	0.071931
	Sense	0.7037	0.000276	0.983	0.832064
	Anti-sense	0.8634	0.104462	0.9138	0.207189
	Intragenic	0.954	0.832456	**1.4674**	**0.000158**
	Intergenic	0.7566	0.000317	0.8628	0.015025
THE1A					
GSE61635 SLE PBMC RNP+	All	0.8296	0.256099	0.9937	1
	Sense	0.7756	0.28022	0.8451	0.331968
	Anti-sense	0.8719	0.579061	1.1826	0.206243
	Intragenic	1.0833	0.76731	**2.4455**	**1.50E-07**
	Intergenic	0.7606	0.11465	0.776	0.045783
GSE14905 Psoriasis, skin	All	1.1072	0.248873	0.718	0.001289
	Sense	0.9544	0.798302	0.6704	0.00673
	Anti-sense	1.248	0.047357	0.7326	0.024616
	Intragenic	**2.3026**	**3.00E-08**	0.5411	0.011661
	Intergenic	0.9692	0.807162	0.7753	0.01795

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE5764 Ductal and lobular breast cancer	All	1.1285	0.211289	0.8542	0.229043
	Sense	0.787	0.127703	0.8341	0.359198
	Anti-sense	**1.4717**	**0.000881**	0.9395	0.754261
	Intragenic	1.6835	0.003402	1.3265	0.209454
	Intergenic	1.0729	0.491749	0.859	0.279318
GSE6919 Metastasis prostate cancer	All	1.1943	0.106487	0.8457	0.132531
	Sense	0.9813	1	0.7567	0.084888
	Anti-sense	1.3313	0.035557	0.9382	0.685639
	Intragenic	**1.9934**	**0.000297**	1.099	0.591126
	Intergenic	1.0536	0.669517	0.8364	0.138118
GSE9750 cervical cancer	All	0.9822	0.756674	0.5833	5.80E-07
	Sense	0.9234	0.283911	0.5154	2.63E-05
	Anti-sense	1.0333	0.614252	0.6208	0.000682
	Intragenic	**1.6977**	**4.00E-08**	0.6457	0.065709
	Intergenic	0.9013	0.06718	0.6039	8.87E-06
GSE13911 Microsatellite i nstable gastric cancer	All	0.9426	0.430778	0.7283	0.006425
	Sense	0.9371	0.552002	0.8322	0.276744
	Anti-sense	0.9947	1	0.6523	0.007709
	Intragenic	**1.6529**	**0.000102**	0.564	0.038409
	Intergenic	0.9026	0.197625	0.7714	0.035169
THE1B					
GSE4588 SLE CD4 CELLS	All	0.8864	0.098368	0.7324	6.05E-06
	Sense	1.0489	0.552182	0.8239	0.012682
	Anti-sense	0.8792	0.111602	0.8218	0.009716
	Intragenic	**1.5224**	**9.42E-05**	1.2344	0.049074
	Intergenic	0.8348	0.014041	0.7347	8.80E-06
GSE13887 SLE T cells	All	0.5689	0	0.9826	0.772239
	Sense	0.6812	3.60E-07	1.0158	0.798441
	Anti-sense	0.5476	0	1.1066	0.103997
	Intragenic	0.5664	5.51E-06	**1.6812**	**0**
	Intergenic	0.5782	0	0.9687	0.600523
GSE61635 SLE PBMC RNP+	All	0.7092	0.00012	0.9957	0.973555
	Sense	0.6879	0.000299	1.1232	0.107578
	Anti-sense	0.8326	0.066064	1.1308	0.085656
	Intragenic	1.3613	0.019896	**2.235**	**0**
	Intergenic	0.6807	2.03E-05	1.0068	0.920416
GSE52471 SLE/DLE, skin	All	0.8819	0.049751	0.8346	0.00631
	Sense	0.9843	0.834188	0.8896	0.114742
	Anti-sense	0.9904	0.918354	0.833	0.011519
	Intragenic	**1.7463**	**0**	0.7492	0.017044
	Intergenic	0.8233	0.002514	0.849	0.014267

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE4588 RA B CELLS	All	0.829	0.003213	0.8418	0.009535
	Sense	0.9691	0.678377	1.0116	0.88492
	Anti-sense	0.9248	0.264299	0.8116	0.004577
	Intragenic	**1.4512**	**8.73E-05**	1.2563	0.02503
	Intergenic	0.8279	0.003072	0.8436	0.011113
GSE10500 RA macophage cells	All	0.5884	0	1.0878	0.156091
	Sense	0.6763	3.50E-07	1.1793	0.009988
	Anti-sense	0.5551	0	1.1303	0.05159
	Intragenic	0.5847	2.59E-05	**1.5447**	**9.90E-07**
	Intergenic	0.5962	0	1.0607	0.327081
GSE13355 Psoriasis, skin	All	1.3535	0.003386	0.782	0.004272
	Sense	**1.4389**	**0.00081**	0.8945	0.247279
	Anti-sense	1.4148	0.001268	0.7822	0.009732
	Intragenic	**2.2856**	**0**	0.6447	0.006335
	Intergenic	1.2748	0.02012	0.8075	0.012794
GSE14905 Psoriasis, skin	All	**1.206**	**0.000986**	0.6869	0
	Sense	1.1984	0.003231	0.7633	4.46E-05
	Anti-sense	**1.3058**	**7.49E-06**	0.6721	0
	Intragenic	**2.1892**	**0**	0.6319	2.52E-05
	Intergenic	1.1696	0.006152	0.6956	0
GSE52471 psoriasis, skin	All	0.9803	0.713543	0.7866	0.000119
	Sense	1.0706	0.235602	0.8317	0.00897
	Anti-sense	1.0884	0.133354	0.8028	0.001424
	Intragenic	**1.9241**	**0**	0.5924	9.94E-06
	Intergenic	0.9367	0.225354	0.8164	0.001298
GSE12453 Diffuse Large B cells Lymphoma vs centroblasts	All	1.0752	0.351061	0.5596	0
	Sense	1.0522	0.548372	0.6775	7.04E-06
	Anti-sense	1.1024	0.239905	0.5525	0
	Intragenic	**1.5772**	**7.33E-05**	0.3596	0
	Intergenic	1.0387	0.639226	0.5919	0
GSE12453 Diffuse Large B cells Lymphoma vs centrocytes	All	1.2187	0.039516	0.481	0
	Sense	1.109	0.315989	0.5889	0
	Anti-sense	1.2256	0.043761	0.4926	0
	Intragenic	**1.9157**	**1.96E-06**	0.2688	0
	Intergenic	1.172	0.10183	0.5093	0
GSE12453 Diffuse Large B cells Lymphoma vs naive B cells	All	0.9114	0.425341	0.512	0
	Sense	1.1229	0.34585	0.5817	0
	Anti-sense	0.9354	0.622633	0.5329	0
	Intragenic	**2.1282**	**1.45E-06**	0.3978	0
	Intergenic	0.8681	0.229236	0.5358	0

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE1299 Breast cancer cells	All	1.1306	0.192859	0.7909	0.078136
	Sense	1.1171	0.281229	0.9179	0.614178
	Anti-sense	1.1971	0.070455	0.7227	0.034153
	Intragenic	**2.0155**	**7.00E-08**	1.1123	0.586428
	Intergenic	1.0766	0.427288	0.7973	0.100706
GSE3167 Bladder carcinoma Situ	All	1.213	0.097089	0.6311	0
	Sense	1.0333	0.800839	0.6507	0
	Anti-sense	1.3422	0.013466	0.6424	0
	Intragenic	**1.9214**	**5.18E-05**	0.3803	0
	Intergenic	1.114	0.35793	0.6692	0
GSE5764 Ductal and lobular breast cancer	All	0.8731	0.028991	0.9993	1
	Sense	1.0394	0.564262	1.0161	0.838279
	Anti-sense	0.9194	0.219259	0.9677	0.718991
	Intragenic	**1.6357**	**6.00E-08**	1.1088	0.387551
	Intergenic	0.8692	0.026028	0.9934	0.940704
GSE5816 Lung Adenocarcinoma	All	0.7209	0	1.0088	0.895908
	Sense	0.8022	6.00E-08	0.9628	0.613521
	Anti-sense	0.7973	1.00E-08	1.1273	0.089805
	Intragenic	**1.2732**	**2.17E-05**	**1.6047**	**1.09E-06**
	Intergenic	0.72	0	1.0051	0.947572
GSE6919 Metastasis prostate cancer	All	0.981	0.804102	0.9236	0.225031
	Sense	1.0377	0.639019	0.8906	0.112987
	Anti-sense	1.1031	0.193314	0.889	0.09797
	Intragenic	**1.7448**	**6.00E-08**	1.0025	0.957604
	Intergenic	0.9591	0.568899	0.924	0.222915
GSE9750 cervical cancer	All	0.9769	0.464907	0.5202	0
	Sense	1.0819	0.023773	0.4963	0
	Anti-sense	1.0272	0.430084	0.5962	0
	Intragenic	**1.5939**	**0**	0.5653	3.00E-07
	Intergenic	0.9686	0.322228	0.5207	0
GSE13911 Microsatellite i nstable gastric cancer	All	0.8318	3.39E-05	0.6263	0
	Sense	0.947	0.273889	0.6742	2.00E-07
	Anti-sense	0.9336	0.160171	0.712	3.61E-06
	Intragenic	**1.6002**	**0**	0.7013	0.003126
	Intergenic	0.8138	4.46E-06	0.6492	0
GSE9764 5-aza Human mesenchymal stem cells	All	0.92	0.311386	0.7843	0.00741
	Sense	1.0463	0.623056	0.8407	0.094268
	Anti-sense	1.0309	0.726499	0.914	0.385911
	Intragenic	**1.7888**	**6.80E-07**	1.1303	0.372789
	Intergenic	0.9002	0.207008	0.8163	0.027953

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE41040 H3K9me3 primary fibroblasts	All	0.7756	0.000161	0.7936	1.49E-05
	Sense	0.7527	0.000225	0.8931	0.058494
	Anti-sense	0.831	0.012401	0.9208	0.162625
	Intragenic	0.8967	0.377848	**1.7067**	**0**
	Intergenic	0.757	4.26E-05	0.7824	5.01E-06
THE1C					
GSE13887 SLE T cells	All	0.7138	5.62E-06	0.9539	0.48393
	Sense	0.8156	0.02737	0.9596	0.638471
	Anti-sense	0.6709	2.04E-05	1.0699	0.375858
	Intragenic	0.8283	0.220196	**1.8179**	**2.00E-08**
	Intergenic	0.7403	8.33E-05	0.9432	0.396038
GSE61635 SLE PBMC RNP+	All	0.8931	0.270087	0.9941	0.970935
	Sense	0.8589	0.238013	0.9637	0.719864
	Anti-sense	0.9714	0.861834	1.2176	0.01942
	Intragenic	1.414	0.040166	**2.3049**	**0**
	Intergenic	0.9005	0.325182	0.9545	0.549978
GSE10500 RA macophage cells	All	0.5225	0	1.0857	0.205221
	Sense	0.6008	3.40E-07	1.1644	0.05009
	Anti-sense	0.5143	0	1.1566	0.054201
	Intragenic	0.6296	0.005235	**1.5498**	**8.90E-05**
	Intergenic	0.5465	0	1.0797	0.243581
GSE14905 Psoriasis, skin	All	1.1008	0.118818	0.7566	2.38E-05
	Sense	1.0642	0.420092	0.7648	0.001282
	Anti-sense	**1.2744**	**0.00077**	0.7678	0.001187
	Intragenic	**1.8345**	**0**	0.6875	0.00715
	Intergenic	1.1113	0.096007	0.7929	0.000608
GSE52471 psoriasis, skin	All	0.9072	0.100374	0.8583	0.02804
	Sense	0.9314	0.338372	0.8036	0.014071
	Anti-sense	1.0021	0.972182	0.9656	0.712807
	Intragenic	**1.5124**	**3.40E-05**	0.8912	0.432224
	Intergenic	0.9212	0.183253	0.8798	0.076923
GSE36474 myeloma, bone marrow	All	0.8367	0.119288	0.9323	0.580165
	Sense	0.9931	1	0.8671	0.340998
	Anti-sense	0.7367	0.032046	1.1257	0.349194
	Intragenic	0.8755	0.603903	**1.9062**	**0.000242**
	Intergenic	0.8939	0.348991	0.9078	0.427197
GSE12453 Diffuse Large B cells Lymphoma vs naive B cells	All	0.9597	0.802104	0.6118	0
	Sense	0.8438	0.316977	0.6804	0
	Anti-sense	1.2416	0.129399	0.5757	0
	Intragenic	**2.1282**	**0.000126**	0.4061	0
	Intergenic	0.9612	0.797261	0.6433	0

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE1299 Breast cancer cells	All	1.3341	0.003529	0.9415	0.719989
	Sense	1.2186	0.101651	1.0444	0.79113
	Anti-sense	**1.4645**	**0.000814**	0.7343	0.117889
	Intragenic	**1.8088**	**0.000333**	0.8585	0.78004
	Intergenic	1.3364	0.003891	0.97	0.88307
GSE3167 Bladder carcinoma Situ	All	1.2005	0.131797	0.7036	2.90E-07
	Sense	1.2644	0.104336	0.6953	3.24E-05
	Anti-sense	1.3544	0.03291	0.7033	3.72E-05
	Intragenic	**2.3527**	**7.05E-06**	0.529	3.13E-05
	Intergenic	1.1771	0.19709	0.7166	2.25E-06
GSE5764 Ductal and lobular breast cancer	All	0.9486	0.457169	0.9208	0.328951
	Sense	0.9217	0.360205	0.9282	0.514981
	Anti-sense	1.1081	0.204751	0.8357	0.09391
	Intragenic	**1.7433**	**7.70E-07**	0.9784	1
	Intergenic	0.9485	0.466589	0.9529	0.587481
GSE5816 Lung Adenocarcinoma	All	0.7592	0	0.8488	0.028206
	Sense	0.749	2.00E-08	0.917	0.353018
	Anti-sense	0.8268	0.000122	0.9363	0.486296
	Intragenic	1.1614	0.042854	**1.5577**	**0.000356**
	Intergenic	0.768	0	0.8331	0.018153
GSE6919 Metastasis prostate cancer	All	1.0196	0.815214	0.8434	0.018723
	Sense	1.0579	0.533574	0.8407	0.057473
	Anti-sense	1.0224	0.813699	0.949	0.580959
	Intragenic	**1.5976**	**0.000368**	1.2152	0.134159
	Intergenic	1.0273	0.749202	0.8475	0.028034
GSE9750 cervical cancer	All	1.037	0.294831	0.6282	0
	Sense	1.0987	0.026879	0.6301	8.00E-08
	Anti-sense	1.0855	0.049428	0.6327	6.00E-08
	Intragenic	**1.7552**	**0**	0.5754	0.000148
	Intergenic	1.0136	0.706794	0.6498	0
GSE13911 Microsatellite i nstable gastric cancer	All	0.8681	0.004575	0.7198	1.11E-05
	Sense	0.8739	0.029523	0.7463	0.001964
	Anti-sense	1.0413	0.481784	0.7218	0.00044
	Intragenic	**1.5584**	**1.30E-07**	0.6445	0.005991
	Intergenic	0.8609	0.003351	0.7413	0.0001
GSE9764 5-aza Human mesenchymal stem cells	All	0.889	0.21246	0.8436	0.095691
	Sense	0.9801	0.912745	0.9069	0.469148
	Anti-sense	1.0115	0.914202	0.9158	0.51401
	Intragenic	**1.84**	**2.88E-05**	1.4103	0.045198
	Intergenic	0.8634	0.1316	0.8433	0.107549

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE22859 H3K4me2, HeLa	All	0.8935	0.213564	0.6821	0.000306
	Sense	1.0026	0.957953	0.688	0.006167
	Anti-sense	0.8999	0.350293	0.8354	0.169402
	Intragenic	**1.8106**	**2.74E-05**	1.1645	0.388781
	Intergenic	0.8492	0.078332	0.6707	0.00026
GSE41040 H3K9me3 primary fibroblasts	All	0.7989	0.003216	0.899	0.076829
	Sense	0.8194	0.035608	0.9318	0.353458
	Anti-sense	0.8573	0.103559	0.9746	0.752134
	Intragenic	0.9922	1	**1.7505**	**1.00E-08**
	Intergenic	0.8176	0.010045	0.8945	0.069511
THE1D					
GSE13887 SLE T cells	All	0.583	0	0.9347	0.276093
	Sense	0.5809	0	0.8889	0.11409
	Anti-sense	0.6586	8.80E-07	1.0391	0.572841
	Intragenic	0.5498	8.19E-05	**1.5511**	**1.60E-05**
	Intergenic	0.6053	0	0.9021	0.103052
GSE61635 SLE PBMC RNP+	All	0.81	0.025469	1.114	0.118893
	Sense	0.7978	0.051332	1.1862	0.032352
	Anti-sense	0.8934	0.338757	1.2303	0.007669
	Intragenic	1.5419	0.004225	**2.6444**	**0**
	Intergenic	0.7901	0.015859	1.0315	0.672537
GSE52471 SLE/DLE, skin	All	0.9923	0.920985	0.8343	0.009805
	Sense	1.0311	0.694598	0.8032	0.009859
	Anti-sense	1.0426	0.589475	0.8784	0.122567
	Intragenic	**1.4789**	**0.000541**	0.7719	0.071427
	Intergenic	0.9682	0.662057	0.8558	0.029557
RA_GSE4588_CD4	All	0.8958	0.112169	0.7496	0.001274
	Sense	0.8828	0.137683	0.8358	0.102875
	Anti-sense	0.987	0.905939	0.8291	0.07996
	Intragenic	**1.4671**	**0.000841**	1.2303	0.166532
	Intergenic	0.8646	0.039244	0.7207	0.000343
GSE10500 RA macophage cells	All	0.6206	0	1.1205	0.063827
	Sense	0.6971	4.10E-05	1.1366	0.072387
	Anti-sense	0.6055	1.00E-08	1.134	0.071976
	Intragenic	0.6704	0.007819	**1.5844**	**1.03E-05**
	Intergenic	0.6252	0	1.096	0.140094

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE13355 Psoriasis, skin	All	1.1409	0.217221	0.8139	0.022559
	Sense	1.028	0.8484	0.7736	0.021854
	Anti-sense	1.2072	0.117889	0.781	0.024386
	Intragenic	**2.1106**	**5.66E-06**	0.4818	0.000603
	Intergenic	1.0172	0.870143	0.8585	0.098956
GSE14905 Psoriasis, skin	All	1.1106	0.075064	0.7381	1.02E-06
	Sense	1.0411	0.57305	0.7541	0.000187
	Anti-sense	1.2348	0.00164	0.7186	1.06E-05
	Intragenic	**1.8547**	**0**	0.5794	4.19E-05
	Intergenic	1.0439	0.469364	0.7787	7.44E-05
GSE52471 psoriasis, skin	All	1.0029	0.956293	0.8398	0.008369
	Sense	1.0473	0.47407	0.8648	0.072796
	Anti-sense	1.0491	0.443449	0.7995	0.004933
	Intragenic	**1.6095**	**2.40E-07**	0.779	0.067383
	Intergenic	0.9454	0.329452	0.8615	0.026285
GSE36474 myeloma, bone marrow	All	0.9122	0.385769	1.2023	0.073871
	Sense	0.7682	0.045338	1.2891	0.033518
	Anti-sense	0.9711	0.858345	1.0805	0.500829
	Intragenic	0.7401	0.207241	**1.9672**	**3.66E-05**
	Intergenic	0.9164	0.436248	1.1338	0.239474
GSE3167 Bladder carcinoma Situ	All	1.1228	0.338877	0.6945	2.00E-08
	Sense	1.0838	0.570142	0.7506	0.000248
	Anti-sense	1.3073	0.043158	0.7162	1.58E-05
	Intragenic	**1.8683**	**0.000906**	0.5194	3.61E-06
	Intergenic	1.0958	0.465873	0.7356	2.53E-06
GSE5764 Ductal and lobular breast cancer	All	0.9507	0.440652	0.8595	0.057903
	Sense	0.9714	0.731341	0.8409	0.073265
	Anti-sense	1.0539	0.476744	0.9805	0.856919
	Intragenic	**1.4458**	**0.000786**	1.4416	0.005243
	Intergenic	0.9249	0.252769	0.8441	0.03706
GSE5816 Lung Adenocarcinoma	All	0.7843	0	0.9617	0.584338
	Sense	0.8276	4.18E-05	0.988	0.903109
	Anti-sense	0.8594	0.000764	0.9685	0.719868
	Intragenic	1.182	0.014221	**1.5457**	**0.00016**
	Intergenic	0.7856	0	0.9608	0.578209
GSE6919 Metastasis prostate cancer	All	1.1532	0.054002	0.8361	0.009217
	Sense	1.2626	0.005564	0.8652	0.081322
	Anti-sense	1.1029	0.243949	0.8047	0.008156
	Intragenic	**1.6234**	**7.80E-05**	0.8956	0.447807
	Intergenic	1.0947	0.228238	0.8728	0.053113

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
GSE9750 cervical cancer	All	0.9621	0.251887	0.6137	0
	Sense	0.9855	0.72273	0.6901	1.55E-06
	Anti-sense	1.0372	0.343088	0.5861	0
	Intragenic	**1.5831**	**0**	0.4738	1.10E-07
	Intergenic	0.9332	0.042805	0.6388	0
GSE13911 Microsatellite i nstable gastric cancer	All	0.8569	0.000984	0.7162	2.00E-06
	Sense	0.9244	0.163047	0.6125	3.00E-08
	Anti-sense	0.9537	0.390362	0.8343	0.028035
	Intragenic	**1.5852**	**0**	0.6487	0.003293
	Intergenic	0.8403	0.000282	0.7398	2.51E-05
GSE41040 H3K9me3 primary fibroblasts	All	0.8383	0.012818	0.8558	0.005803
	Sense	0.8719	0.109665	0.8276	0.00593
	Anti-sense	0.8586	0.06998	0.98	0.772458
	Intergenic	0.9244	0.163047	0.6125	3.00E-08
	Intragenic	0.9368	0.694207	**1.5721**	**1.08E-06**

Intragenic HERVs associated with up-regulated genes in SLE

There are total 852 and 896 over-expressed genes in SLE were associated with intragenic ERV1 and ERV3 containing genes, respectively. While the number of intragenic HERVs neighboring gene associated with over-expressed genes in SLE were listed in Table 21. As mention previously, there was no HERVK association with SLE in our analysis. In summary, our result showed that the association was restricted to certain LTR subtype (Table 20). It is likely that this specificity was mediated through sequence-specific and cell-specific transcription factor [174]. We have reviewed literatures about the factors that regulate methylation of HERV e.g., SETDB1[175], KAP1 [176], TRIM28 [177] and then analyzed gene expression in cells that lack these genes by knockout experiments as showed in Appendix Table B2 and Table B3 for entire HERVs and superfamily level, respectively. However, we did not see significant changes of gene expression that associated with HERV. It is possible that these transcription factors might be important in embryonic stem cells but not in mature somatic cells especially in blood cells that we are interested in SLE.

Table 21 Number of up-regulated genes in SLE associated with intragenic HERVs

Superfamily	Family	HERV name	Number of intragenic HERV genes
ERV1/ERVE	HERVH	LTR7	45
ERV3/ERVL	HERVL33	LTR33	212
	MLT	MLT1D	384
	MST	MSTD	191
ERVL-MaLR		THE1B	354
	THE1	THE1C	228
		THE1D	264

We further performed functional analysis for all list of discovered associated genes. The analysis results were listed in Appendix Table B4. The gene expression level of cell adhesion molecules was reports as a good prognostic of lupus nephritis previously [5]. Many genes are also involved in biological processes and molecular functions that potentially play a role in SLE pathogenesis. Furthermore, our analysis help reveals signaling pathways that were reported to involve in SLE. The abnormal signaling activity of phosphatidylinositol 3-kinase pathway in SLE patients was reported to be a part of the disease factors [178, 179]. Moreover, the over-expression of EGF gene family was also observed. The EGF proteins play a role in EGFR signaling pathway, which is responding to RNP complex. There is a report on the activation of EGFR pathway by LTR polymorphism [180, 181]. Including with a computational protein-protein interaction network analyses of quantitative phosphoproteome data indicated the linked of RNP complex and EGFR/HER2 signaling networks [182]. This information from our analysis suggest the important role of HERV related to RNP formation due to both the cross-reactivity with HERV protein and the aberrant regulation of EGFR/HER2 signaling.

The enrichment analysis results show that there are associations between certain LTR patterns and gene-upregulation mainly in SLE T cells as shown above [183, 184]. It is most likely that these LTRs from HERV were activated by hypomethylation. We confirm this hypothesis by analysis the association of gene expression profile of cells treated with hypomethylating agent, 5 azacythidine. If hypomethylation is the main

cause, we expect to see up-regulation of genes containing LTR. In fact, this is what we found, both in CD4+ T cells, T cells and B cells as showed in Table 22.

Therefore, we hypothesize that aberrant HERV regulation in SLE will lead to upregulation of genes due to function of LTR as promoter, enhancer or alternative splicing. We analyzed the transcriptome sequencing to detect chimeric sequences. There are many reports showed that TE-derived alternative promoters, which generate a chimeric mRNA with an adjacent gene, are arguably one of the more straightforward scenarios to link a TE with a functional product. Particularly when that gene encodes a protein of known function and LTR located immediately at 5' of protein-coding region frequently function as alternative promoters and/or also express noncoding RNAs. [174, 185, 186]. Since, all chimeric patterns were listed, further analysis in structure visualization of candidate chimerics are required to confirm those chimerics.

Table 22 Association analysis between 5 azacythidine treated mesenchymal stem cells and genes in various SLE conditions (significant with OR > 1 and p < 0.001, indicated in bold letter)

Gene expression condition	OR	p-value
GSE32591 LN glomolular	**3.232**	**0.00023694**
GSE32591 LN tubulolar	2.7351	0.04193993
GSE10325 SLE B cells	**4.0474**	**4.82E-05**
GSE10325 SLE myeloid cells	2.1846	0.01156959
GSE10325 SLE T cells	2.8833	0.02195973
GSE13887 SLE T cells	**2.6678**	**0**
GSE20864 SLE PBMC	1.2979	0.4764354
GSE24706 SLE PBMC, ANA	2.0974	0.18050457
GSE27427 SLE neutrophils	1.5061	0.03732643
GSE4588 SLE B cells	**3.3814**	**0**
GSE4588 SLE CD4 CELLS	**4.6134**	**0**
GSE52471 SLE/DLE, skin	**3.3241**	**0**
GSE61635 SLE PBMC RNP+	**3.3246**	**0**
GSE30153 SLE inactive condition, B cells	6.5257	0.00154786

Chimeric identification in RNA-Seq data analysis

There are 355 and 1678 NGS datasets are available in GEO database that sequenced by Illumina® HiSeq2500 and HIseq2000, respectively (MAY 2016). Unfortunately, with the limitation of computational capability unit, we selected 10 SLE and 3 healthy control RNA-Seq samples from published GSE72509 in NCBI GEO database as the source materials for detecting chimeric transcripts in this analysis. We chosed the samples that have a top three lowest HERV expression signal based on calculated RPKM of RNA-Seq data for representing control group. In contrast we selected top 10 SLE samples that have the highest level of HERV expression to represent SLE group. The number of raw read and assembled transcripts using both TopHat and trinity method was shown in Table 23. The assemble result show that the number of transcripts constructed by Trinity, the de novo, assembly are higher that mapping with TopHat method. This might occurred because the complex of the repeat sequence.

Table 23 RNA-seq analysis statistic

Sample	total read	Number of transcripts		No of gene (RefSeq annotation)	
		Tophat	Trinity	Tophat	Trinity
C5	91422694	45289	51219	21137	15190
C7	95262500	45066	50725	21556	15379
C10	82631888	45174	52818	21589	15719
SLE4	96825506	45350	60493	21634	18744
SLE9	108708983	46789	89903	21757	16256
SLE27	98661930	45249	64057	21639	16690
SLE34	97581203	44620	51244	21587	15326
SLE35	96325010	45963	75887	21696	18744
SLE50	94582459	44334	46567	21554	14801
SLE54	96252132	44330	45867	21549	14132
SLE72	92762536	44067	40278	21618	14140
SLE75	98652153	44815	58582	21628	16170
SLE76	91254856	44387	55442	21537	15711

By using both chimeric detection approaches as mentioned in method section, we selected 4 candidate chimeric transcript based on chimeric pattern and gene related to SLE/LN pathogenesis for further validation by Sanger sequencing method. Figure 30-33 showed the chimeric pattern of selected candidate chimeric transcripts.

There were reported previously that IFI44L promoter methylation can be used as a blood biomarker for SLE [170]. Interestingly, we detected the IFI44L chimeric transcript as shown in Figure 29. The chimeric was found as a part of the last exon of the gene which is normally there is no effect to their neighboring gene. Besides IFI44L, we also discovered IFI44-THE1C chimeric transcript from this study as shown in Figure 30. The THE1C LTR elongates the transcript of IFI44 by fusing to the first exon of the gene. There are about 22 times higher expression level of IFI44 in SLE when compare to healthy people found in this study. Second example, the predicted CLEC2D-THE1C chimeric were shown in Figure 31. This chimeric event possible to use the THE1C LTR as the alternative promoter which is located inside the intron region resulting in shorten the transcript by missing the up-stream exon of the LTR location. This made a high possibility that the chimeric sequence of this CLEC2D might lose their function in SLE. Third example shown in Figure 32 is the predicted CLEC4E-MER52C transcript as MER52C LTR is transcribe to be the alternative first exon of CLEC4E gene. Our last observed chimeric TOP3A-LTR5B indicated LTR5B is located in intron region of the gene as shown in Figure 33. This chimeric shows the same alternative pattern as CLEC2D-THE1C chimeric since they might generate a new transcript isoform by using LTR inside the intron region as the alternative promoter that results in shorten transcript by missing the up-stream exons. We also found that there are difference chimeric transcripts in OSCAR-LTR12B between SLE and normal group as showed in Figure 34. There are many LTRs that located in the exon as part of the genes. But this assembly event normally not affect the expression of their neighboring genes. So, we did not investigate this event in detail in this study. Furthermore, we also can identify a full-range HERV from this RNA-Seq data such as LTR2-HERVE-LTR2 that illustrated in Figure 35.

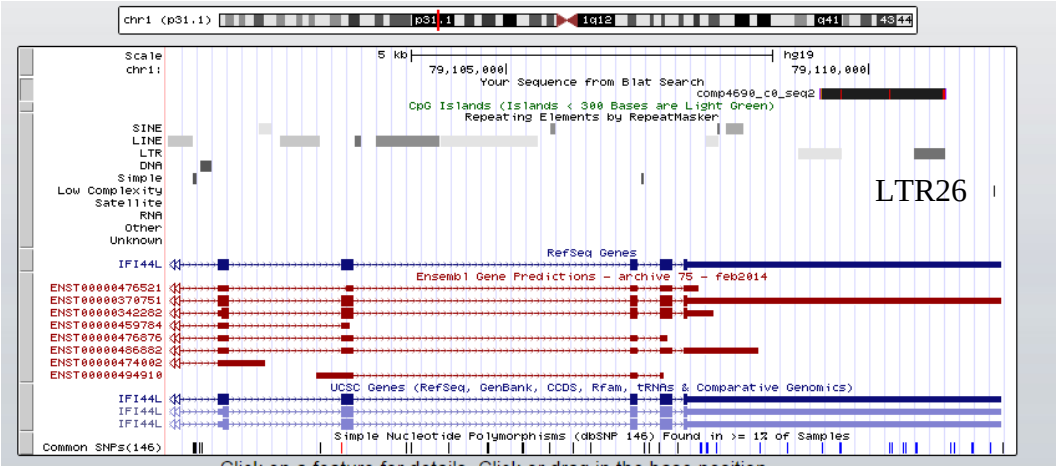


Figure 29 predicted IFI44L-LTR26 chimeric transcript

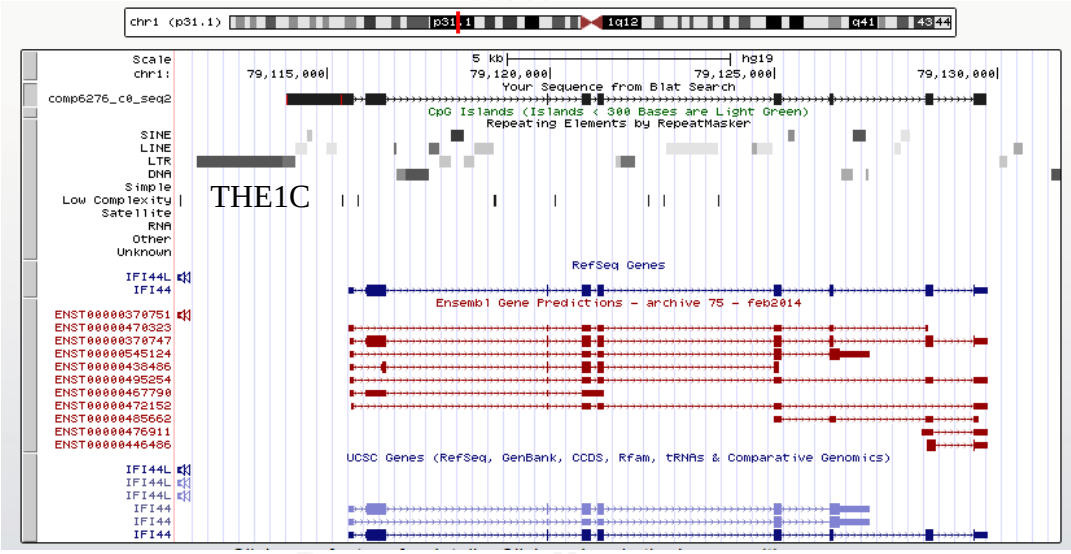


Figure 30 predicted IFI44-THE1C chimeric transcript

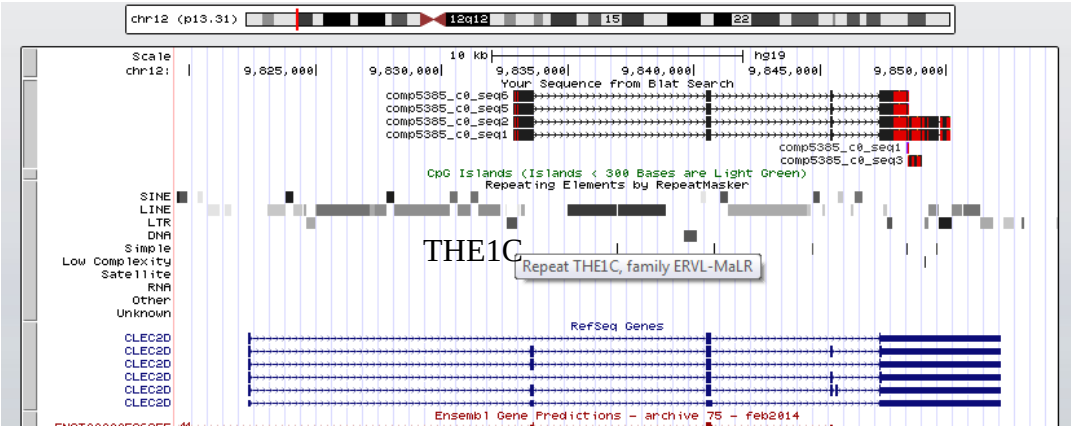


Figure 31 predicted CLEC2D-THE1C chimeric transcript

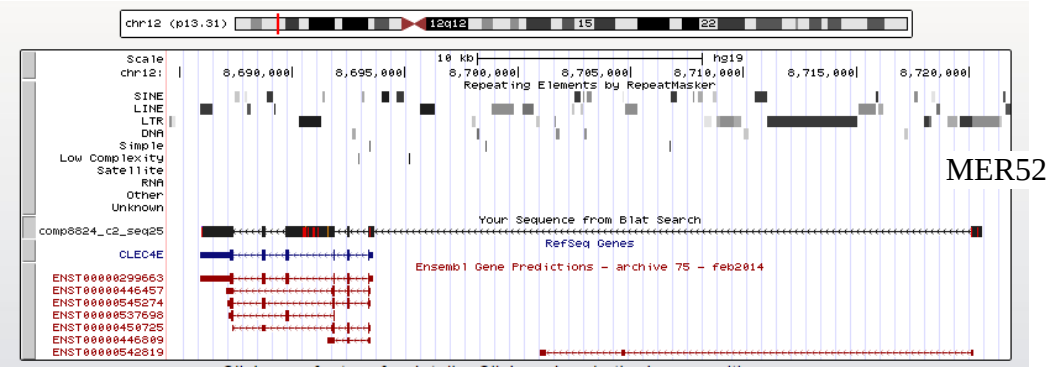


Figure 32 predicted CLEC4E-MER52C chimeric transcript

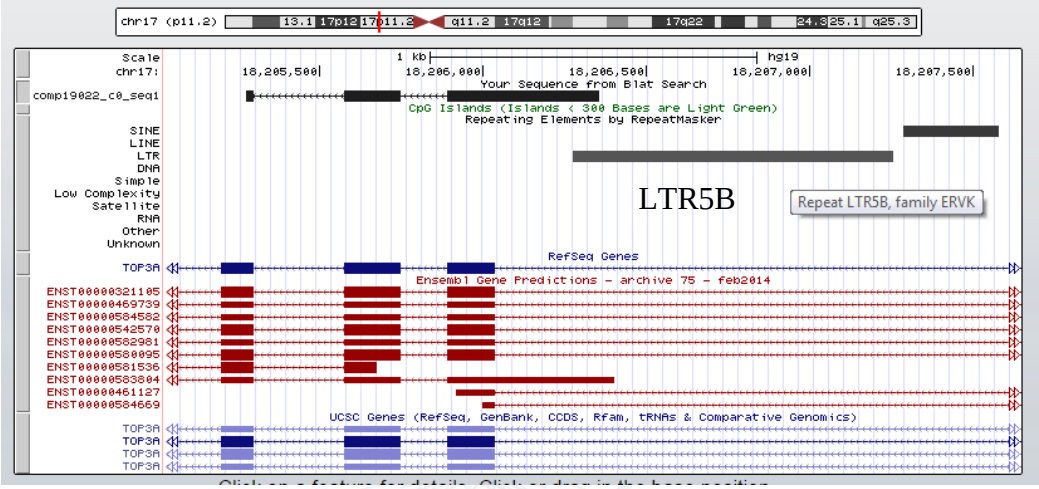


Figure 33 predicted TOP3A-LTR5B chimeric transcript

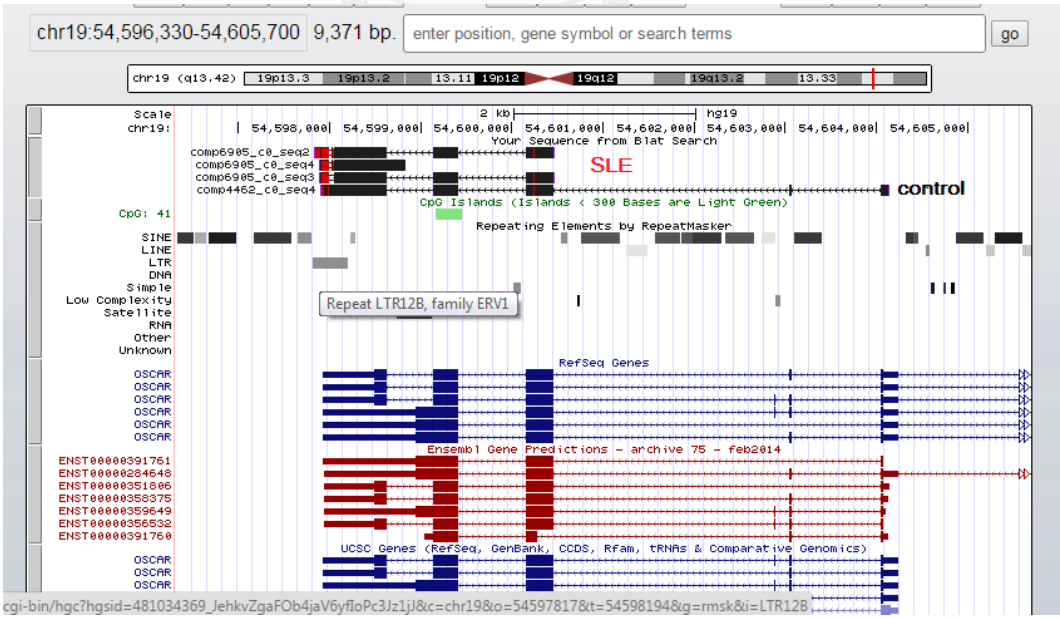


Figure 34 predicted OSCAR-LTR12B chimeric transcript

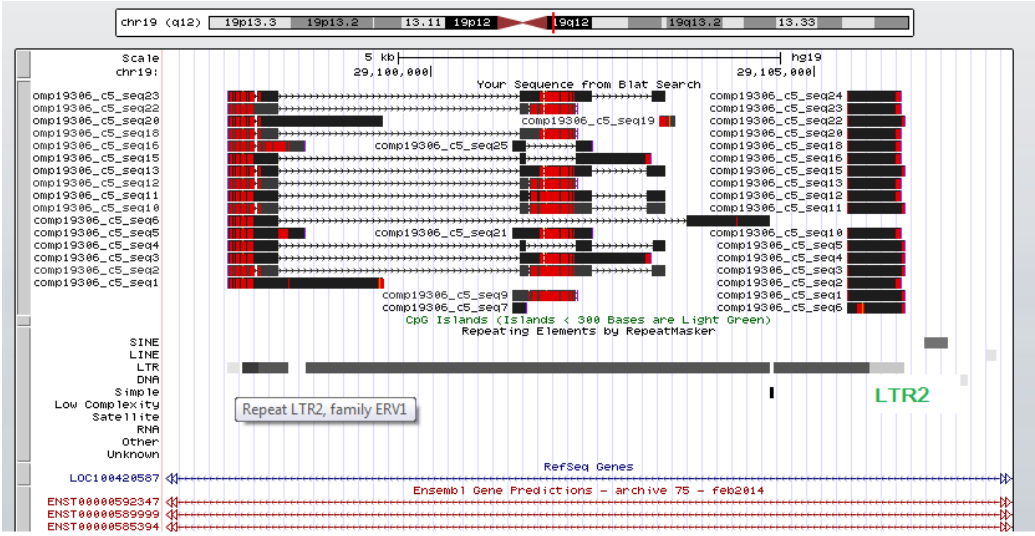


Figure 35 a full-range LTR2-intHERVE-LTR2 structure

For the validation procedure, the forward primers were designed to locate at LTR regions while reverse primers were located at next exon of the chimeric genes. The primers used in this study were listed in Table 24. We have further validated the amplicons that meet the expected size of 3 candidates with Sanger sequencing. Interestingly, the sequencing results show that all LTR forward amplicons were unreadable. It might because THE1C and also other LTR are globally located in the genome so the LTR forward primers might bind to many region subsequently to the mixed amplicons at the LTR regions.

Table 24 Primer list uses for detecting chimeric transcripts

Chimeric transcript	Forward primer	Reverse primer	Expected product size
IFI44-THE1B	CAGCAGGAGAGGGAAATGAG	GTGCAACCATGTCAAACGAG	1470
CLEC2D-THE1C	CCCCAGCCATGTAGAACTGT	CAACCTGAGCAAGATCAGCA	932
CLEC4E-MER52C	TCTCTGCTGAGAGCTGGACA	TGTGCATTGTGTTTCAGATGAT	1380
TOP3A-LTR5B	CTGATCTCTCTTTTCCCCACA	CCAGCACCTCAGGAAAAATC	465

In addition, LOR1a-IRF5 chimeric transcript was reported in Hodgkin lymphoma (HL) by Babaian A *et al.* [186]. They purpose this finding as a key regulator of the aberrant transcriptome characteristic of this disease which IRF5 up-regulation in

HL is driven by LOR1a LTR upstream of IRF5. Therefore we have tested the published primers for detecting this chimeric transcript in the SLE and K562 RNA-Seq data. Unfortunately there is no any chimeric transcript found in our analysis results.

As the limitation of whole genome RNA-Seq analysis in term of the amount of reads in specific region, Stone RC et al. performed RNA-Seq for targeted enrichment for IRF5-SLE risk haplotype analysis [187]. This target enrichment approach will also help in chimeric sequence detection as well. At last, we have to note that the SLE RNA-Seq data that we used in this study is the single end sequencing technique since the aim of the author this RNA-Seq publication was just only to measure the gene expression level. So we suggest that if the pair-end SLE/LN RNA-Seq data are available in the future, this will help to improve the chimeric transcript detection in term of accuracy and the orientation of the assembly as well.

CHAPTER V
CONCLUSIONS

Although renal pathology provides the best diagnostic/prognostic values for lupus nephritis, a non-invasive monitoring remains an unmet need. As urine and serum may be the best sources of repeatable and safely tests. We proposed to identify lists of candidate genes/proteins for diagnostic /prognostic biomarkers using systems biology using two differences studies. The first part we investigated the integration of gene and protein layers of refractory LN. We can identify 5 compacted protein cluster underlying pathways of resistant to treatment such as tight junction, complement and TNF pathways.

Since, many reported support that epigenetics also play an import role in SLE/LN pathogenesis. With the previous publications that shows how TEs can alter the expression of their nearby genes, which resulted from the methylation imbalance. Therefore the second part of this study investigated the association of HERV and gene expression under various disease conditions especially in SLE/LN. By using repeat and human genome information from Repbase and UCSC table browser, respectively, we have developed a HERV database and enrichment tool called EnHERV. EnHERV is available at http://sysbio.chula.ac.th/enherv/. EnHERV provides searching by gene names or HERV characteristics. EnHERV also allows user to do enrichment analysis between user interested genes list and selected specific HERV characteristics. Thousands of enrichment analysis was calculated in this study. The results showed that there seems to have LTR type specific to certain disease conditions. For example, we found that THE1B and THE1D in both orientations were significantly enrichment in over-expression of SLE RNP+ genes. Many THE1B and THE1D intragenic genes tended to involve in molecular functions that played a role in SLE pathogenesis such as cell-cell adhesion, phosphatidylinositol and EGFR pathways. By using EnHERV, it might help us to further understand the pathogenesis of not only SLE in our analysis result but also with other disease that HERVs might involve in their pathogenesis as well.

Based on the hypothesis that HERV used their own regulatory region such as promoter or enhancer in their LTR as alternative regulator of the neighboring genes under imbalance methylation condition. We further investigated the mechanism of those LTR using available SLE RNA-Seq data for detecting chimeric transcript. As a result 4 candidate chimeric transcripts were identified with the computational and RNA-Seq sequences. Further investigate in other RNA-Seq data or construct a precise validation process might help to confirm our hypothesis about these chimeric events in the future.

In conclusion, we report an un-biased discovery of biomarkers of lupus nephritis. These biomarkers may not only be used for diagnostic purpose but may also lead to the new therapeutic targets in the future. However, it is in crucial needs to validate these findings in a well-designed clinical trial before clinical practice guideline implementation.

REFERENCES

1. Avihingsanon, Y. and N. Hirankarn, *Major lupus organ involvement: severe lupus nephritis.* Lupus, 2010. **19**(12): p. 1391-8.
2. Mok, M.Y. and W.L. Li, *Do Asian patients have worse lupus?* Lupus, 2010. **19**(12): p. 1384-90.
3. Jeffries, M.A. and A.H. Sawalha, *Epigenetics in systemic lupus erythematosus: leading the way for specific therapeutic agents.* Int J Clin Rheumtol, 2011. **6**(4): p. 423-439.
4. Wu, H., et al., *The real culprit in systemic lupus erythematosus: abnormal epigenetic regulation.* Int J Mol Sci, 2015. **16**(5): p. 11013-33.
5. Benjachat, T., et al., *Biomarkers for Refractory Lupus Nephritis: A Microarray Study of Kidney Tissue.* Int J Mol Sci, 2015. **16**(6): p. 14276-90.
6. Somparn, P., et al., *Urinary proteomics revealed prostaglandin H(2)D-isomerase, not Zn-alpha2-glycoprotein, as a biomarker for active lupus nephritis.* J Proteomics, 2012. **75**(11): p. 3240-7.
7. Treamtrakanpon, W., et al., *APRIL, a proliferation-inducing ligand, as a potential marker of lupus nephritis.* Arthritis Res Ther, 2012. **14**(6): p. R252.
8. Ou, K., et al., *Novel breast cancer biomarkers identified by integrative proteomic and gene expression mapping.* J Proteome Res, 2008. **7**(4): p. 1518-28.
9. Shibuya, M., et al., *Proteomic and transcriptomic analyses of retinal pigment epithelial cells exposed to REF-1/TFPI-2.* Invest Ophthalmol Vis Sci, 2007. **48**(2): p. 516-21.
10. Goymer, P., *Network biology: why do we need hubs?* Nat Rev Genet, 2008. **9**(9): p. 650.
11. Rodenhiser, D. and M. Mann, *Epigenetics and human disease: translating basic biology into clinical applications.* CMAJ, 2006. **174**(3): p. 341-8.
12. Wu, H., et al., *The key culprit in the pathogenesis of systemic lupus erythematosus: Aberrant DNA methylation.* Autoimmun Rev, 2016.
13. Bannert, N. and R. Kurth, *Retroelements and the human genome: new perspectives on an old relation.* Proc Natl Acad Sci U S A, 2004. **101 Suppl 2**: p. 14572-9.
14. Buzdin, A., *Human-specific endogenous retroviruses.* ScientificWorldJournal, 2007. **7**: p. 1848-68.
15. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nat Rev Genet, 2009. **10**(10): p. 691-703.
16. Okada, M., et al., *Role of DNA methylation in transcription of human endogenous retrovirus in the pathogenesis of systemic lupus erythematosus.* J Rheumatol, 2002. **29**(8): p. 1678-82.
17. Piotrowski, P.C., S. Duriagin, and P.P. Jagodzinski, *Expression of human endogenous retrovirus clone 4-1 may correlate with blood plasma concentration of anti-U1 RNP and anti-Sm nuclear antibodies.* Clin Rheumatol, 2005. **24**(6): p. 620-4.
18. Nakkuntod, J., et al., *DNA methylation of human endogenous retrovirus in systemic lupus erythematosus.* J Hum Genet, 2013. **58**(5): p. 241-9.

19. Lea, J.P., *Lupus nephritis in African Americans.* Am J Med Sci, 2002. **323**(2): p. 85-9.
20. Eguchi, K., *Apoptosis in autoimmune diseases.* Intern Med, 2001. **40**(4): p. 275-84.
21. Pisetsky, D.S. and I.A. Vrabie, *Antibodies to DNA: infection or genetics?* Lupus, 2009. **18**(13): p. 1176-80.
22. van Bavel, C.C., J. van der Vlag, and J.H. Berden, *Glomerular binding of anti-dsDNA autoantibodies: the dispute resolved?* Kidney Int, 2007. **71**(7): p. 600-1.
23. Christopher-Stine, L., et al., *Renal biopsy in lupus patients with low levels of proteinuria.* J Rheumatol, 2007. **34**(2): p. 332-5.
24. Fagerholm, S.C., et al., *The CD11b-integrin (ITGAM) and systemic lupus erythematosus.* Lupus, 2013. **22**(7): p. 657-63.
25. Musone, S.L., et al., *Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus.* Nat Genet, 2008. **40**(9): p. 1062-4.
26. Oishi, T., et al., *A functional SNP in the NKX2.5-binding site of ITPR3 promoter is associated with susceptibility to systemic lupus erythematosus in Japanese population.* J Hum Genet, 2008. **53**(2): p. 151-62.
27. Kamatani, Y., et al., *Identification of a significant association of a single nucleotide polymorphism in TNXB with systemic lupus erythematosus in a Japanese population.* J Hum Genet, 2008. **53**(1): p. 64-73.
28. Peterson, K.S., et al., *Characterization of heterogeneity in the molecular pathogenesis of lupus nephritis from transcriptional profiles of laser-captured glomeruli.* J Clin Invest, 2004. **113**(12): p. 1722-33.
29. Alcorta, D.A., et al., *Leukocyte gene expression signatures in antineutrophil cytoplasmic autoantibody and lupus glomerulonephritis.* Kidney Int, 2007. **72**(7): p. 853-64.
30. Berthier, C.C., et al., *Cross-species transcriptional network analysis defines shared inflammatory responses in murine and human lupus nephritis.* J Immunol, 2012. **189**(2): p. 988-1001.
31. Santucci, L., et al., *Urinary proteome in a snapshot: normal urine and glomerulonephritis.* J Nephrol, 2013. **26**(4): p. 610-6.
32. Vlahou, A., et al., *Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine.* Am J Pathol, 2001. **158**(4): p. 1491-502.
33. Bratt, O., *Hereditary prostate cancer: clinical aspects.* J Urol, 2002. **168**(3): p. 906-13.
34. Liska, V., et al., *Matrix metalloproteinases and their inhibitors in correlation to proliferative and classical tumour markers during surgical therapy of colorectal liver metastases.* Bratisl Lek Listy, 2012. **113**(2): p. 108-13.
35. Wu, T., et al., *Urinary angiostatin--a novel putative marker of renal pathology chronicity in lupus nephritis.* Mol Cell Proteomics, 2013. **12**(5): p. 1170-9.
36. Wu, T. and C. Mohan, *Lupus nephritis - alarmins may sound the alarm?* Arthritis Res Ther, 2012. **14**(6): p. 129.
37. Oates, J.C., et al., *Prediction of urinary protein markers in lupus nephritis.* Kidney Int, 2005. **68**(6): p. 2588-92.

38. Zhang, X., et al., *Biomarkers of lupus nephritis determined by serial urine proteomics.* Kidney Int, 2008. **74**(6): p. 799-807.
39. Bollain, Y.G.J.J., et al., *Increased excretion of urinary podocytes in lupus nephritis.* Indian J Nephrol, 2011. **21**(3): p. 166-71.
40. Avihingsanon, Y., et al., *Decreased renal expression of vascular endothelial growth factor in lupus nephritis is associated with worse prognosis.* Kidney Int, 2009. **75**(12): p. 1340-8.
41. Zhang, S., et al., *Discovering functions and revealing mechanisms at molecular level from biological networks.* Proteomics, 2007. **7**(16): p. 2856-69.
42. Marbach, D., et al., *Wisdom of crowds for robust gene network inference.* Nat Methods, 2012. **9**(8): p. 796-804.
43. McRedmond, J.P., et al., *Integration of proteomics and genomics in platelets: a profile of platelet proteins and platelet-specific genes.* Mol Cell Proteomics, 2004. **3**(2): p. 133-44.
44. Bader, G.D., D. Betel, and C.W. Hogue, *BIND: the Biomolecular Interaction Network Database.* Nucleic Acids Res, 2003. **31**(1): p. 248-50.
45. Chatr-Aryamontri, A., et al., *The BioGRID interaction database: 2015 update.* Nucleic Acids Res, 2015. **43**(Database issue): p. D470-8.
46. Xenarios, I., et al., *DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.* Nucleic Acids Res, 2002. **30**(1): p. 303-5.
47. Gene Ontology, C., *The Gene Ontology project in 2008.* Nucleic Acids Res, 2008. **36**(Database issue): p. D440-4.
48. Chen, J.Y., S. Mamidipalli, and T. Huan, *HAPPI: an online database of comprehensive human annotated and predicted protein interactions.* BMC Genomics, 2009. **10 Suppl 1**: p. S16.
49. Kerrien, S., et al., *The IntAct molecular interaction database in 2012.* Nucleic Acids Res, 2012. **40**(Database issue): p. D841-6.
50. Tarcea, V.G., et al., *Michigan molecular interactions r2: from interacting proteins to pathways.* Nucleic Acids Res, 2009. **37**(Database issue): p. D642-6.
51. Chatr-aryamontri, A., et al., *MINT: the Molecular INTeraction database.* Nucleic Acids Res, 2007. **35**(Database issue): p. D572-4.
52. Pagel, P., et al., *The MIPS mammalian protein-protein interaction database.* Bioinformatics, 2005. **21**(6): p. 832-4.
53. Ficenec, D., et al., *Computational knowledge integration in biopharmaceutical research.* Brief Bioinform, 2003. **4**(3): p. 260-78.
54. Li, J., et al., *The Molecule Pages database.* Nature, 2002. **420**(6916): p. 716-7.
55. Caspi, R., et al., *The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.* Nucleic Acids Res, 2016. **44**(D1): p. D471-80.
56. Kanehisa, M., et al., *KEGG for representation and analysis of molecular networks involving diseases and drugs.* Nucleic Acids Res, 2010. **38**(Database issue): p. D355-60.
57. Ekins, S., et al., *Pathway mapping tools for analysis of high content data.* Methods Mol Biol, 2007. **356**: p. 319-50.

58. Matthews, L., et al., *Reactome knowledgebase of human biological pathways and processes.* Nucleic Acids Res, 2009. **37**(Database issue): p. D619-22.
59. Pico, A.R., et al., *WikiPathways: pathway editing for the people.* PLoS Biol, 2008. **6**(7): p. e184.
60. Le Novere, N., et al., *BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems.* Nucleic Acids Res, 2006. **34**(Database issue): p. D689-91.
61. Ananko, E.A., et al., *GeneNet in 2005.* Nucleic Acids Res, 2005. **33**(Database issue): p. D425-7.
62. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.
63. Letunic, I., et al., *iPath: interactive exploration of biochemical pathways and networks.* Trends Biochem Sci, 2008. **33**(3): p. 101-3.
64. Mi, H., et al., *PANTHER version 10: expanded protein families and functions, and analysis tools.* Nucleic Acids Res, 2016. **44**(D1): p. D336-42.
65. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.
66. Zhao, Y. and Y. Yang, *Frex and FrexH: Indicators of metabolic states in living cells.* Bioeng Bugs, 2012. **3**(3): p. 181-8.
67. Kutmon, M., et al., *WikiPathways: capturing the full diversity of pathway knowledge.* Nucleic Acids Res, 2016. **44**(D1): p. D488-94.
68. Kamburov, A., et al., *ConsensusPathDB: toward a more complete picture of cell biology.* Nucleic Acids Res, 2011. **39**(Database issue): p. D712-7.
69. Griffiths-Jones, S., et al., *miRBase: tools for microRNA genomics.* Nucleic Acids Res, 2008. **36**(Database issue): p. D154-8.
70. Mathelier, A., et al., *JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles.* Nucleic Acids Res, 2014. **42**(Database issue): p. D142-7.
71. Marinescu, V.D., I.S. Kohane, and A. Riva, *MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes.* BMC Bioinformatics, 2005. **6**: p. 79.
72. Wingender, E., *The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation.* Brief Bioinform, 2008. **9**(4): p. 326-32.
73. Hoffmann, R., *Using the iHOP information resource to mine the biomedical literature on genes, proteins, and chemical compounds.* Curr Protoc Bioinformatics, 2007. **Chapter 1**: p. Unit1 16.
74. Hu, Y., et al., *Analysis of genomic and proteomic data using advanced literature mining.* J Proteome Res, 2003. **2**(4): p. 405-12.
75. Pan, Y. and A.H. Sawalha, *Epigenetic regulation and the pathogenesis of systemic lupus erythematosus.* Transl Res, 2009. **153**(1): p. 4-10.
76. Zhao, M., et al., *Epigenetic dynamics in immunity and autoimmunity.* Int J Biochem Cell Biol, 2015. **67**: p. 65-74.
77. Murakami, H., et al., *Detection of inter-spread repeat sequence in genomic DNA sequence.* Genome Inform, 2004. **15**(1): p. 170-9.

78. Hsieh, J. and A. Fire, *Recognition and silencing of repeated DNA.* Annu Rev Genet, 2000. **34**: p. 187-204.
79. Javierre, B.M., et al., *Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus.* Genome Res, 2010. **20**(2): p. 170-9.
80. Wang, G.S., et al., *Ultraviolet B exposure of peripheral blood mononuclear cells of patients with systemic lupus erythematosus inhibits DNA methylation.* Lupus, 2009. **18**(12): p. 1037-44.
81. Richardson, B., et al., *Evidence for impaired T cell DNA methylation in systemic lupus erythematosus and rheumatoid arthritis.* Arthritis Rheum, 1990. **33**(11): p. 1665-73.
82. Luo, Y., et al., *Abnormal DNA methylation in T cells from patients with subacute cutaneous lupus erythematosus.* Br J Dermatol, 2008. **159**(4): p. 827-33.
83. Richardson, B., *Effect of an inhibitor of DNA methylation on T cells. II. 5-Azacytidine induces self-reactivity in antigen-specific T4+ cells.* Hum Immunol, 1986. **17**(4): p. 456-70.
84. Cribbs, A., M. Feldmann, and U. Oppermann, *Towards an understanding of the role of DNA methylation in rheumatoid arthritis: therapeutic and diagnostic implications.* Ther Adv Musculoskelet Dis, 2015. **7**(5): p. 206-19.
85. Lu, Q., et al., *Demethylation of ITGAL (CD11a) regulatory sequences in systemic lupus erythematosus.* Arthritis Rheum, 2002. **46**(5): p. 1282-91.
86. Liang, J., et al., *A correlation study on the effects of DNMT1 on methylation levels in CD4(+) T cells of SLE patients.* Int J Clin Exp Med, 2015. **8**(10): p. 19701-8.
87. Lu, Q., A. Wu, and B.C. Richardson, *Demethylation of the same promoter sequence increases CD70 expression in lupus T cells and T cells treated with lupus-inducing drugs.* J Immunol, 2005. **174**(10): p. 6212-9.
88. Blikstad, V., et al., *Evolution of human endogenous retroviral sequences: a conceptual account.* Cell Mol Life Sci, 2008. **65**(21): p. 3348-65.
89. Gifford, R. and M. Tristem, *The Evolution, Distribution and Diversity of Endogenous Retroviruses.* Virus Genes, 2003. **26**(3): p. 291-315.
90. Griffiths, D., *Endogenous retroviruses in the human genome sequence.* Genome Biology, 2001. **2**(6): p. reviews1017.1 - reviews1017.5.
91. Nelson, P.N., et al., *Dymystified...Human endogenous retroviruses.* Journal of Clinical Pathology, 2003. **56**(1): p. 11-18.
92. Nelson, P.N., et al., *Human endogenous retroviruses: transposable elements with potential?* Clinical & Experimental Immunology, 2004. **138**(1): p. 1-9.
93. Kazazian, H.H., Jr., *Mobile Elements: Drivers of Genome Evolution.* Science, 2004. **303**(5664): p. 1626-1632.
94. Jern, P. and J.M. Coffin, *Effects of Retroviruses on Host Genome Function.* Annual Review of Genetics, 2008. **42**(1): p. 709-732.
95. Donovan, T., *Endogenous Retroviruses and the Human Genome: Implications for Human Disease.* SPCV, 2010. **2**(1): p. 1-33.
96. Blomberg, J., et al., *Classification and nomenclature of endogenous retroviral sequences (ERVs): Problems and recommendations.* Gene, 2009. **448**(2): p. 115-123.

97. Jurka, J., et al., *Repbase Update, a database of eukaryotic repetitive elements.* Cytogenet Genome Res, 2005. **110**(1-4): p. 462-7.
98. Mayer, J. and E. Meese, *Human endogenous retroviruses in the primate lineage and their influence on host genomes.* Cytogenetic and Genome Research, 2005. **110**(1-4): p. 448-456.
99. Smit, A.F.A., *Identification of a new, abundant superfamily of mammalian LTR-transposons.* Nucleic Acids Research, 1993. **21**(8): p. 1863-1872.
100. Oja, M., et al., *Methods for estimating human endogenous retrovirus activities from EST databases.* BMC Bioinformatics, 2007. **8**(Suppl 2): p. S11.
101. Blomberg, J., D. Ushameckis, and P. Jern, *Evolutionary Aspects of Human Endogenous Retroviral Sequences (HERVs) and Disease*. Retroviruses and Primate Genome Evolution, ed. E.D. Sverdlov. 2005: Eurekah.com.
102. Kurth, R. and N. Bannert, *Beneficial and detrimental effects of human endogenous retroviruses.* International Journal of Cancer, 2010. **126**(2): p. 306-314.
103. Buzdin, A., *Human-Specific Endogenous Retroviruses.* TheScientificWorldJOURNAL, 2007. **7**: p. 1848-1868.
104. Long, Q., et al., *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, 1998. **54**(3): p. 542-555.
105. Dunn, C.A., P. Medstrand, and D.L. Mager, *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, 2003. **100**(22): p. 12841-12846.
106. Medstrand, P., J.-R. Landry, and D.L. Mager, *Long Terminal Repeats Are Used as Alternative Promoters for the Endothelin B Receptor and Apolipoprotein C-I Genes in Humans.* Journal of Biological Chemistry, 2001. **276**(3): p. 1896-1903.
107. Dunn, C.A., et al., *Transcription of two human genes from a bidirectional endogenous retrovirus promoter.* Gene, 2006. **366**(2): p. 335-342.
108. Di Cristofano, A., et al., *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, 1995. **23**(15): p. 2823-2830.
109. Feuchter-Murthy, A.E., J.D. Freeman, and D.L. Mager, *Splicing of a human endogenous retrovirus to a novel phospholipase A2 related gene.* Nucleic Acids Research, 1993. **21**(1): p. 135-143.
110. Shirai, T. and S. Hirose, *Molecular pathogenesis of SLE.* Springer Seminars in Immunopathology, 2006. **28**(2): p. 79-82.
111. Ines, C. and F.G. Robert, *Role of Endogenous Retroviruses in Autoimmune Diseases.* Infectious disease clinics of North America, 2006. **20**(4): p. 913-929.
112. Adelman, M.K. and J.J. Marchalonis, *Endogenous Retroviruses in Systemic Lupus Erythematosus: Candidate Lupus Viruses.* Clinical Immunology, 2002. **102**(2): p. 107-116.
113. Sekigawa, I., et al., *Systemic lupus erythematosus and human endogenous retroviruses.* Modern Rheumatology, 2003. **13**(2): p. 107-113.

114. Naito, T., et al., *Immune Abnormalities Induced by Human Endogenous Retroviral Peptides: With Reference to the Pathogenesis of Systemic Lupus Erythematosus.* Journal of Clinical Immunology, 2003. **23**(5): p. 371-376.
115. Chu, J.L., et al., *The defect in Fas mRNA expression in MRL/lpr mice is associated with insertion of the retrotransposon, ETn.* The Journal of Experimental Medicine, 1993. **178**(2): p. 723-730.
116. Balada, E., J. Ordi-Ros, and M. Vilardell-Tarrés, *Molecular mechanisms mediated by human endogenous retroviruses (HERVs) in autoimmunity.* Reviews in Medical Virology, 2009. **19**(5): p. 273-286.
117. Balada, E., M. Vilardell-Tarres, and J. Ordi-Ros, *Implication of Human Endogenous Retroviruses in the Development of Autoimmune Diseases.* International Reviews of Immunology, 2009. **29**(4): p. 351-370.
118. Wicker, T., et al., *A universal classification of eukaryotic transposable elements implemented in Repbase.* Nat Rev Genet, 2008. **9**(5): p. 414-414.
119. Blikstad, V., et al., *Endogenous retroviruses.* Cellular and Molecular Life Sciences, 2008. **65**(21): p. 3348-3365.
120. Bergman, C.M. and H. Quesneville, *Discovering and detecting transposable elements in genome sequences.* Briefings in Bioinformatics, 2007. **8**(6): p. 382-392.
121. Pereira, V., *Automated paleontology of repetitive DNA with REANNOTATE.* BMC Genomics, 2008. **9**(1): p. 614.
122. Rivals, I., et al., *Enrichment or depletion of a GO category within a class of genes: which test?* Bioinformatics, 2007. **23**(4): p. 401-7.
123. McDonald, J.H., *Handbook of Biological Statistics*. 2nd ed. ed. 2009, Maryland: Sparky House Publishing.
124. Ott, R.L. and M. Longnecker, *An Introduction to Statistical Methods and Data Analysis*. 6th ed. ed. 2010: Brooks/Cole Cengage Learning.
125. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-95.
126. Li, R., et al., *SOAP: short oligonucleotide alignment program.* Bioinformatics, 2008. **24**(5): p. 713-4.
127. Li, R., et al., *SOAP2: an improved ultrafast tool for short read alignment.* Bioinformatics, 2009. **25**(15): p. 1966-7.
128. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
129. Grabherr, M.G., et al., *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat Biotechnol, 2011. **29**(7): p. 644-52.
130. Giardine, B., et al., *Galaxy: a platform for interactive large-scale genome analysis.* Genome Res, 2005. **15**(10): p. 1451-5.
131. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Brief Bioinform, 2013. **14**(2): p. 178-92.
132. Rosenbloom, K.R., et al., *ENCODE data in the UCSC Genome Browser: year 5 update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D56-63.
133. Perco, P., et al., *Linking transcriptomic and proteomic data on the level of protein interaction networks.* Electrophoresis, 2010. **31**(11): p. 1780-9.

134. Dahlquist, K.D., et al., *GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways.* Nat Genet, 2002. **31**(1): p. 19-20.
135. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
136. Alfarano, C., et al., *The Biomolecular Interaction Network Database and related tools 2005 update.* Nucleic Acids Res, 2005. **33**(Database issue): p. D418-24.
137. Hermjakob, H., et al., *IntAct: an open source molecular interaction database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D452-5.
138. Gao, J., et al., *Integrating and annotating the interactome using the MiMI plugin for cytoscape.* Bioinformatics, 2009. **25**(1): p. 137-8.
139. Bader, G.D. and C.W. Hogue, *An automated method for finding molecular complexes in large protein interaction networks.* BMC Bioinformatics, 2003. **4**: p. 2.
140. Saito, R., et al., *A travel guide to Cytoscape plugins.* Nat Methods, 2012. **9**(11): p. 1069-76.
141. Thomas, P.D., et al., *PANTHER: a library of protein families and subfamilies indexed by function.* Genome Res, 2003. **13**(9): p. 2129-41.
142. Dennis, G., Jr., et al., *DAVID: Database for Annotation, Visualization, and Integrated Discovery.* Genome Biol, 2003. **4**(5): p. P3.
143. Cline, M.S., et al., *Integration of biological networks and gene expression data using Cytoscape.* Nat Protoc, 2007. **2**(10): p. 2366-82.
144. Bao, W., K.K. Kojima, and O. Kohany, *Repbase Update, a database of repetitive elements in eukaryotic genomes.* Mob DNA, 2015. **6**: p. 11.
145. Smit, A., R. Hubley, and G. Glusman. *RepeatMasker*. 2011, 8 June]; Available from: http://www.repeatmasker.org/.
146. De Santis, M. and C. Selmi, *The therapeutic potential of epigenetics in autoimmune diseases.* Clin Rev Allergy Immunol, 2012. **42**(1): p. 92-101.
147. Ihaka, R. and R. Gentleman, *R: A Language for Data Analysis and Graphics.* Journal of Computational and Graphical Statistics, 1996. **5**(3): p. 299-314.
148. Schumacher, R. and A. Lentz. *Dispelling the Myths*. 2011, 8 June]; Available from: http://dev.mysql.com/tech-resources/articles/dispelling-the-myths.html.
149. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool.* Nucleic Acids Research, 2004. **32**(suppl 1): p. D493-D496.
150. Hsu, F., et al., *The UCSC Known Genes.* Bioinformatics, 2006. **22**(9): p. 1036-1046.
151. Bansal, V., et al., *An MCMC algorithm for haplotype assembly from whole-genome sequence data.* Genome Research, 2008. **18**(8): p. 1336-1346.
152. Aporntewan, C., Mutirangura, A., *Connection up- and down-regulation expression*

analysis of microarrays (CU-DREAM): a physiogenomic discovery tool. Asian Biomedicine, 2011. **5**(2): p. 257-262.
153. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Res, 2002. **30**(1): p. 207-10.

154. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D991-5.
155. Bhardwaj, N. and J.M. Coffin, *Endogenous retroviruses and human cancer: is there anything to the rumors?* Cell Host Microbe, 2014. **15**(3): p. 255-9.
156. Absher, D.M., et al., *Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations.* PLoS Genet, 2013. **9**(8): p. e1003678.
157. Lock, F.E., et al., *Distinct isoform of FABP7 revealed by screening for retroelement-activated genes in diffuse large B-cell lymphoma.* Proc Natl Acad Sci U S A, 2014. **111**(34): p. E3534-43.
158. Sarot, E., et al., *Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the Drosophila melanogaster flamenco gene.* Genetics, 2004. **166**(3): p. 1313-21.
159. Landry, J.R., et al., *The Opitz syndrome gene Mid1 is transcribed from a human endogenous retroviral promoter.* Mol Biol Evol, 2002. **19**(11): p. 1934-42.
160. Feuchter-Murthy, A.E., J.D. Freeman, and D.L. Mager, *Splicing of a human endogenous retrovirus to a novel phospholipase A2 related gene.* Nucleic Acids Res, 1993. **21**(1): p. 135-43.
161. Renaudineau, Y., et al., *Characterization of the human CD5 endogenous retrovirus-E in B lymphocytes.* Genes Immun, 2005. **6**(8): p. 663-71.
162. Hung, T., et al., *The Ro60 autoantigen binds endogenous retroelements and regulates inflammatory gene expression.* Science, 2015. **350**(6259): p. 455-9.
163. Slavoff, S.A., et al., *Peptidomic discovery of short open reading frame-encoded peptides in human cells.* Nat Chem Biol, 2013. **9**(1): p. 59-64.
164. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.
165. Langmead, B., *Aligning short sequencing reads with Bowtie.* Curr Protoc Bioinformatics, 2010. **Chapter 11**: p. Unit 11 7.
166. Rosenbloom, K.R., et al., *The UCSC Genome Browser database: 2015 update.* Nucleic Acids Res, 2015. **43**(Database issue): p. D670-81.
167. Pruitt, K.D., et al., *RefSeq: an update on mammalian reference sequences.* Nucleic Acids Res, 2014. **42**(Database issue): p. D756-63.
168. Yates, A., et al., *Ensembl 2016.* Nucleic Acids Res, 2016. **44**(D1): p. D710-6.
169. Haas, B.J., et al., *De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.* Nat Protoc, 2013. **8**(8): p. 1494-512.
170. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.
171. Untergasser, A., et al., *Primer3--new capabilities and interfaces.* Nucleic Acids Res, 2012. **40**(15): p. e115.
172. Wu, Z., et al., *DNA methylation modulates HERV-E expression in CD4+ T cells from systemic lupus erythematosus patients.* J Dermatol Sci, 2015. **77**(2): p. 110-6.

173. Aporntewan, C., et al., *Hypomethylation of intragenic LINE-1 represses transcription in cancer cells through AGO2.* PLoS One, 2011. **6**(3): p. e17934.
174. Mak, K.S., et al., *Repression of chimeric transcripts emanating from endogenous retrotransposons by a sequence-specific transcription factor.* Genome Biol, 2014. **15**(4): p. R58.
175. Collins, P.L., et al., *The histone methyltransferase SETDB1 represses endogenous and exogenous retroviruses in B lymphocytes.* Proc Natl Acad Sci U S A, 2015. **112**(27): p. 8367-72.
176. Lukic, S., J.C. Nicolas, and A.J. Levine, *The diversity of zinc-finger genes on human chromosome 19 provides an evolutionary mechanism for defense against inherited endogenous retroviruses.* Cell Death Differ, 2014. **21**(3): p. 381-7.
177. Turelli, P., et al., *Interplay of TRIM28 and DNA methylation in controlling human endogenous retroelements.* Genome Res, 2014. **24**(8): p. 1260-70.
178. Besliu, A.N., et al., *PI3K/Akt signaling in peripheral T lymphocytes from systemic lupus erythematosus patients.* Roum Arch Microbiol Immunol, 2009. **68**(2): p. 69-79.
179. Lee, T.P., et al., *Anti-ribosomal phosphoprotein autoantibody triggers interleukin-10 overproduction via phosphatidylinositol 3-kinase-dependent signalling pathways in lipopolysaccharide-activated macrophages.* Immunology, 2009. **127**(1): p. 91-102.
180. Kahyo, T., et al., *Identification and association study with lung cancer for novel insertion polymorphisms of human endogenous retrovirus.* Carcinogenesis, 2013. **34**(11): p. 2531-8.
181. Planque, S., et al., *Autoantibodies to the epidermal growth factor receptor in systemic sclerosis, lupus, and autoimmune mice.* FASEB J, 2003. **17**(2): p. 136-43.
182. Imami, K., et al., *Temporal profiling of lapatinib-suppressed phosphorylation signals in EGFR/HER2 pathways.* Mol Cell Proteomics, 2012. **11**(12): p. 1741-57.
183. Faulkner, G.J., et al., *The regulated retrotransposon transcriptome of mammalian cells.* Nat Genet, 2009. **41**(5): p. 563-71.
184. Reichmann, J., et al., *Microarray analysis of LTR retrotransposon silencing identifies Hdac1 as a regulator of retrotransposon expression in mouse embryonic stem cells.* PLoS Comput Biol, 2012. **8**(4): p. e1002486.
185. Tallack, M.R., et al., *A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells.* Genome Res, 2010. **20**(8): p. 1052-63.
186. Babaian, A., et al., *Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma.* Oncogene, 2015.
187. Stone, R.C., et al., *RNA-Seq for enrichment and analysis of IRF5 transcript expression in SLE.* PLoS One, 2013. **8**(1): p. e54487.

APPENDIX A

Integrative analysis

Table A1. List of genes in MCODE clusters.

MCODE_cluster	gene_symbol	description
C1	REL	v-rel reticuloendotheliosis viral oncogene homolog (avian)
C1	CDC6	cell division cycle 6 homolog (S. cerevisiae)
C1	BRD2	bromodomain containing 2
C1	CKS2	CDC28 protein kinase regulatory subunit 2
C1	CCND1	cyclin D1
C1	SRPR	signal recognition particle receptor ('docking protein')
C1	HIST1H2AC	histone cluster 1, H2ac
C1	CDK2	cyclin-dependent kinase 2
C1	CCNB1	cyclin B1
C1	SMAD2	SMAD family member 2
C1	PPP2CA	protein phosphatase 2 (formerly 2A), catalytic subunit, alpha isoform
C1	GTF2H3	general transcription factor IIH, polypeptide 3, 34kDa
C1	FAS	Fas (TNF receptor superfamily, member 6)
C1	ESR1	estrogen receptor 1
C1	TAF1	TAF1 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 250kDa
C1	TSC2	tuberous sclerosis 2
C1	RB1	retinoblastoma 1 (including osteosarcoma)
C1	PPP2R1B	protein phosphatase 2 (formerly 2A), regulatory subunit A, beta isoform
C1	MDM2	Mdm2, transformed 3T3 cell double minute 2, p53 binding protein (mouse)
C1	JARID1A	jumonji, AT rich interactive domain 1A
C1	HNRNPA2B1	heterogeneous nuclear ribonucleoprotein A2/B1
C1	TMPO	thymopoietin
C1	ALG5	asparagine-linked glycosylation 5 homolog (S. cerevisiae, dolichyl-phosphate beta-glucosyltransferase)
C1	MYBL2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2
C1	JUND	jun D proto-oncogene
C1	JUN	jun oncogene
C1	TUBA3C	tubulin, alpha 3c
C1	TIMM50	translocase of inner mitochondrial membrane 50 homolog (S. cerevisiae)
C1	TERF2	telomeric repeat binding factor 2
C1	SP1	Sp1 transcription factor

MCODE_cluster	gene_symbol	description
C1	SNAPC3	small nuclear RNA activating complex, polypeptide 3, 50kDa
C1	SNAPC1	small nuclear RNA activating complex, polypeptide 1, 43kDa
C1	SHC1	SHC (Src homology 2 domain containing) transforming protein 1
C1	RUVBL2	RuvB-like 2 (E. coli)
C1	PSMB2	proteasome (prosome, macropain) subunit, beta type, 2
C1	PPIA	peptidylprolyl isomerase A (cyclophilin A)
C1	PIK3R1	phosphoinositide-3-kinase, regulatory subunit 1 (alpha)
C1	PCNA	proliferating cell nuclear antigen
C1	NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (p105)
C1	MTHFD1	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1, methenyltetrahydrofolate cyclohydrolase, formyltetrahydrofolate synthetase
C1	MAD2L1	MAD2 mitotic arrest deficient-like 1 (yeast)
C1	HSPA9	heat shock 70kDa protein 9 (mortalin)
C1	HSPA8	heat shock 70kDa protein 8
C1	HSPA1L	heat shock 70kDa protein 1-like
C1	HIF1A	hypoxia-inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)
C1	HDAC1	histone deacetylase 1
C1	GRB2	growth factor receptor-bound protein 2
C1	ERBB3	v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)
C1	EGFR	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)
C1	CTNNB1	catenin (cadherin-associated protein), beta 1, 88kDa
C1	CD82	CD82 molecule
C1	BIRC5	baculoviral IAP repeat-containing 5 (survivin)
C1	BCL3	B-cell CLL/lymphoma 3
C1	ACTG1	actin, gamma 1
C1	NFKBIB	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, beta
C1	E2F3	E2F transcription factor 3
C1	HNF4A	hepatocyte nuclear factor 4, alpha
C1	IKBKG	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma

MCODE_cluster	gene_symbol	description
C1	MYC	v-myc myelocytomatosis viral oncogene homolog (avian)
C1	HIST1H2BJ	histone cluster 1, H2bj
C1	MAP3K14	mitogen-activated protein kinase kinase kinase 14
C1	MSH2	mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)
C1	E2F4	E2F transcription factor 4, p107/p130-binding
C1	E2F1	E2F transcription factor 1
C1	NFKBIA	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha
C1	CDK6	cyclin-dependent kinase 6
C1	CDK4	cyclin-dependent kinase 4
C1	MT1G	metallothionein 1G
C1	NFIL3	nuclear factor, interleukin 3 regulated
C1	CREB3L3	cAMP responsive element binding protein 3-like 3
C1	CREB3L1	cAMP responsive element binding protein 3-like 1
C1	CREB3	cAMP responsive element binding protein 3
C1	RBL1	retinoblastoma-like 1 (p107)
C1	BATF3	basic leucine zipper transcription factor, ATF-like 3
C1	BATF	basic leucine zipper transcription factor, ATF-like
C1	TRAF2	TNF receptor-associated factor 2
C1	RBL2	retinoblastoma-like 2 (p130)
C1	IKBKB	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase beta
C1	DDIT3	DNA-damage-inducible transcript 3
C1	CEBPG	CCAAT/enhancer binding protein (C/EBP), gamma
C1	UXT	ubiquitously-expressed transcript
C1	E2F2	E2F transcription factor 2
C1	TEF	thyrotrophic embryonic factor
C1	HLF	hepatic leukemia factor
C1	DBP	D site of albumin promoter (albumin D-box) binding protein
C1	CEBPE	CCAAT/enhancer binding protein (C/EBP), epsilon
C1	CEBPD	CCAAT/enhancer binding protein (C/EBP), delta
C1	CEBPB	CCAAT/enhancer binding protein (C/EBP), beta

MCODE_cluster	gene_symbol	description
C2	PPT2	palmitoyl-protein thioesterase 2
C2	NEK2	NIMA (never in mitosis gene a)-related kinase 2
C2	IGFBP6	insulin-like growth factor binding protein 6
C2	IGF2	insulin-like growth factor 2 (somatomedin A)
C2	RECQL5	RecQ protein-like 5
C2	UCP2	uncoupling protein 2 (mitochondrial, proton carrier)
C2	BGLAP	bone gamma-carboxyglutamate (gla) protein (osteocalcin)
C2	RANBP9	RAN binding protein 9
C2	HTATIP	HIV-1 Tat interacting protein, 60kDa
C2	ATF1	activating transcription factor 1
C2	POLA2	polymerase (DNA directed), alpha 2 (70kD subunit)
C2	FOXO1	forkhead box O1
C2	NCOA6	nuclear receptor coactivator 6
C2	RIPK2	receptor-interacting serine-threonine kinase 2
C2	PCK1	phosphoenolpyruvate carboxykinase 1 (soluble)
C2	GEMIN4	gem (nuclear organelle) associated protein 4
C2	RFC2	replication factor C (activator 1) 2, 40kDa
C2	MYH10	myosin, heavy chain 10, non-muscle
C2	SERPINA1	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1
C2	TGFBR2	transforming growth factor, beta receptor II (70/80kDa)
C2	POU2F1	POU class 2 homeobox 1
C2	PEX6	peroxisomal biogenesis factor 6
C2	FADD	Fas (TNFRSF6)-associated via death domain
C2	PPP1CC	protein phosphatase 1, catalytic subunit, gamma isoform
C2	MITF	microphthalmia-associated transcription factor
C2	MGST3	microsomal glutathione S-transferase 3
C2	LIG1	ligase I, DNA, ATP-dependent
C2	SLC25A5	solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 5
C2	FASLG	Fas ligand (TNF superfamily, member 6)
C2	BCL2	B-cell CLL/lymphoma 2
C2	TPMT	thiopurine S-methyltransferase
C2	ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)

MCODE_cluster	gene_symbol	description
C2	AURKB	aurora kinase B
C2	TROVE2	TROVE domain family, member 2
C2	CASP3	caspase 3, apoptosis-related cysteine peptidase
C2	PPOX	protoporphyrinogen oxidase
C2	KHDRBS1	KH domain containing, RNA binding, signal transduction associated 1
C2	PLEKHO1	pleckstrin homology domain containing, family O member 1
C2	ZFYVE9	zinc finger, FYVE domain containing 9
C2	GAPDH	glyceraldehyde-3-phosphate dehydrogenase
C2	SKP2	S-phase kinase-associated protein 2 (p45)
C2	PXN	paxillin
C2	ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)
C2	G6PD	glucose-6-phosphate dehydrogenase
C2	NOL3	nucleolar protein 3 (apoptosis repressor with CARD domain)
C2	ATF2	activating transcription factor 2
C2	EP300	E1A binding protein p300
C2	NFKBIE	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, epsilon
C2	TRAF6	TNF receptor-associated factor 6
C2	MAP3K7IP1	mitogen-activated protein kinase kinase kinase 7 interacting protein 1
C2	SMAD4	SMAD family member 4
C2	BRCA1	breast cancer 1, early onset
C2	CDC25A	cell division cycle 25 homolog A (S. pombe)
C2	REL	v-rel reticuloendotheliosis viral oncogene homolog (avian)
C2	CDC6	cell division cycle 6 homolog (S. cerevisiae)
C2	BRD2	bromodomain containing 2
C2	CCND1	cyclin D1
C2	CKS2	CDC28 protein kinase regulatory subunit 2
C2	HIST1H2AC	histone cluster 1, H2ac
C2	CCNB1	cyclin B1
C2	CDK2	cyclin-dependent kinase 2
C2	SRPR	signal recognition particle receptor ('docking protein')
C2	PPP2CA	protein phosphatase 2 (formerly 2A), catalytic subunit, alpha isoform
C2	SMAD2	SMAD family member 2
C2	ESR1	estrogen receptor 1
C2	GTF2H3	general transcription factor IIH, polypeptide 3

MCODE_cluster	gene_symbol	description
C2	FAS	Fas (TNF receptor superfamily, member 6)
C2	TAF1	TAF1 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 250kDa
C2	TSC2	tuberous sclerosis 2
C2	RB1	retinoblastoma 1 (including osteosarcoma)
C2	JARID1A	jumonji, AT rich interactive domain 1A
C2	HNRNPA2B1	heterogeneous nuclear ribonucleoprotein A2/B1
C2	PPP2R1B	protein phosphatase 2 (formerly 2A), regulatory subunit A, beta isoform
C2	MDM2	Mdm2, transformed 3T3 cell double minute 2, p53 binding protein (mouse)
C2	TMPO	thymopoietin
C2	ALG5	asparagine-linked glycosylation 5 homolog (S. cerevisiae, dolichyl-phosphate beta-glucosyltransferase)
C2	JUND	jun D proto-oncogene
C2	MYBL2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2
C2	JUN	jun oncogene
C2	SNAPC1	small nuclear RNA activating complex, polypeptide 1, 43kDa
C2	SHC1	SHC (Src homology 2 domain containing) transforming protein 1
C2	ERBB3	v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)
C2	HSPA1L	heat shock 70kDa protein 1-like
C2	NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (p105)
C2	TIMM50	translocase of inner mitochondrial membrane 50 homolog (S. cerevisiae)
C2	BCL3	B-cell CLL/lymphoma 3
C2	PPIA	peptidylprolyl isomerase A (cyclophilin A)
C2	HSPA8	heat shock 70kDa protein 8
C2	HIF1A	hypoxia-inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)
C2	TERF2	telomeric repeat binding factor 2
C2	PCNA	proliferating cell nuclear antigen
C2	ACTG1	actin, gamma 1
C2	CTNNB1	catenin (cadherin-associated protein), beta 1, 88kDa
C2	SP1	Sp1 transcription factor
C2	SNAPC3	small nuclear RNA activating complex, polypeptide 3, 50kDa

MCODE_cluster	gene_symbol	description
C2	HDAC1	histone deacetylase 1
C2	MTHFD1	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1, methenyltetrahydrofolate cyclohydrolase, formyltetrahydrofolate synthetase
C2	TUBA3C	tubulin, alpha 3c
C2	RUVBL2	RuvB-like 2 (E. coli)
C2	GRB2	growth factor receptor-bound protein 2
C2	PSMB2	proteasome (prosome, macropain) subunit, beta type, 2
C2	EGFR	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)
C2	HSPA9	heat shock 70kDa protein 9 (mortalin)
C2	PIK3R1	phosphoinositide-3-kinase, regulatory subunit 1 (alpha)
C2	BIRC5	baculoviral IAP repeat-containing 5 (survivin)
C2	MAD2L1	MAD2 mitotic arrest deficient-like 1 (yeast)
C2	NFKBIB	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, beta
C2	E2F3	E2F transcription factor 3
C2	CDC2	cell division cycle 2, G1 to S and G2 to M
C2	HNF4A	hepatocyte nuclear factor 4, alpha
C2	IKBKG	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma
C2	MYC	v-myc myelocytomatosis viral oncogene homolog (avian)
C2	HIST1H2BJ	histone cluster 1, H2bj
C2	MAP3K14	mitogen-activated protein kinase kinase kinase 14
C2	MSH2	mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)
C2	E2F4	E2F transcription factor 4, p107/p130-binding
C2	E2F1	E2F transcription factor 1
C2	NFKBIA	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha
C2	CDK4	cyclin-dependent kinase 4
C2	CDK6	cyclin-dependent kinase 6
C2	MT1G	metallothionein 1G
C2	CREB3	cAMP responsive element binding protein 3
C2	CREB3L3	cAMP responsive element binding protein 3-like 3
C2	CREB3L1	cAMP responsive element binding protein 3-like 1

MCODE_cluster	gene_symbol	description
C2	BATF	basic leucine zipper transcription factor, ATF-like
C2	RBL1	retinoblastoma-like 1 (p107)
C2	BATF3	basic leucine zipper transcription factor, ATF-like 3
C2	TRAF2	TNF receptor-associated factor 2
C2	RBL2	retinoblastoma-like 2 (p130)
C2	IKBKB	inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase beta
C2	CEBPG	CCAAT/enhancer binding protein (C/EBP), gamma
C2	DDIT3	DNA-damage-inducible transcript 3
C2	E2F2	E2F transcription factor 2
C2	UXT	ubiquitously-expressed transcript
C2	HLF	hepatic leukemia factor
C2	TEF	thyrotrophic embryonic factor
C2	CEBPE	CCAAT/enhancer binding protein (C/EBP), epsilon
C2	CEBPB	CCAAT/enhancer binding protein (C/EBP), beta
C2	DBP	D site of albumin promoter (albumin D-box) binding protein
C2	CEBPD	CCAAT/enhancer binding protein (C/EBP), delta
C3	MARK4	MAP/microtubule affinity-regulating kinase 4
C3	PRKCI	protein kinase C, iota
C3	PARD6A	par-6 partitioning defective 6 homolog alpha (C. elegans)
C3	PRKCZ	protein kinase C, zeta
C3	PARD6B	par-6 partitioning defective 6 homolog beta (C. elegans)
C3	YWHAH	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, eta polypeptide
C3	PARD3	par-3 partitioning defective 3 homolog (C. elegans)
C4	SMAD4	SMAD family member 4
C4	ATF2	activating transcription factor 2
C4	IGFBP6	insulin-like growth factor binding protein 6
C4	PPP1CC	protein phosphatase 1, catalytic subunit, gamma isoform
C4	PCK1	phosphoenolpyruvate carboxykinase 1 (soluble)
C4	IGF2	insulin-like growth factor 2 (somatomedin A)

MCODE_cluster	gene_symbol	description
C4	UCP2	uncoupling protein 2 (mitochondrial, proton carrier)
C4	G6PD	glucose-6-phosphate dehydrogenase
C4	GAPDH	glyceraldehyde-3-phosphate dehydrogenase
C4	LIG1	ligase I, DNA, ATP-dependent
C4	RIPK2	receptor-interacting serine-threonine kinase 2
C4	BRCA1	breast cancer 1, early onset
C4	HTATIP	HIV-1 Tat interacting protein, 60kDa
C4	EP300	E1A binding protein p300
C5	PLEKHO1	pleckstrin homology domain containing, family O member 1
C5	CASP3	caspase 3, apoptosis-related cysteine peptidase

APPENDIX B

Human Endogenous Retrovirus analysis

Table B1. Full list of HERV superfamilies, families and HERV names

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
	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
1. ERV1 (cont.)	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

Superfamily	Family	Name/group
2. ERVK	HERVK10/ HERVK (HML-2)	LTR5, LTR5_Hs, LTR5A, LTR5B, HERVK
	HERVK14/ HERVK (HML-1)	HERVK14, HERVK14C, LTR14, LTR14A, LTR14B, LTR14C
	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

Superfamily	Family	Name/group
3. ERVL (cont.)	LTR84	LTR84a, LTR84b
	LTR86	LTR86A1, LTR86A2, LTR86B1, LTR86B2, LTR86C
	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

Table B2. Association analysis results of all solo-LTRs in gene knockdown studies.

GEO accession	HERV pattern	Down DEG		Up DEG	
		OR	p-value	OR	p-value
H3K4me1_GSE59695	All	0.0813	0	0.1023	0
	Sense	0.0875	0	0.1103	0
	Anti-sense	0.0879	0	0.1108	0
	Intragenic	0.8569	0.32634303	1.2136	0.10259369
	Intergenic	0.0813	0	0.1023	0
H3K4me2_GSE22859	All	0.0737	0	0.0779	0
	Sense	0.0803	0	0.0836	0
	Anti-sense	0.08	0	0.084	0
	Intragenic	0.9258	0.33007014	0.7913	0.00926265
	Intergenic	0.0737	0	0.0779	0
H3K9_GSE44084	All	0	0	0.0002	0
	Sense	0	0	0.0003	0
	Anti-sense	0	0	0.0003	0
	Intragenic	0	0	0	0
	Intergenic	0	0	0.0002	0
H3K9me3_GSE25282	All	0.1177	0.0327894	0.1679	0.01952111
	Sense	0.1262	0.03714177	0.1802	0.02331161
	Anti-sense	0.1267	0.03741682	0.1809	0.02355726
	Intragenic	0.7097	0.74286987	0.7097	0.48876999
	Intergenic	0.1177	0.0327894	0.1679	0.01952111
H3K9me3_GSE41040	All	0.1169	0	0.0706	0
	Sense	0.1264	0	0.0772	0
	Anti-sense	0.127	0	0.0777	0
	Intragenic	0.8875	0.07448786	0.9484	0.31674741
	Intergenic	0.1169	0	0.0706	0
SETDB1_GSE45175	All	0.1455	0.01405811	0.1177	0
	Sense	0.1561	0.01684388	0.1264	0
	Anti-sense	0.1568	0.01702487	0.1269	0
	Intragenic	1.9527	0.22299657	0.4346	0.00033427
	Intergenic	0.1455	0.01405811	0.1177	0
SETDB1_GSE73231	All	0	0	0	0
	Sense	0	0	0	0
	Anti-sense	0	0	0	0
	Intragenic	0.0027	0	0	0
	Intergenic	0	0	0	0

Table B3. Association analysis results at HERV superfamily level in gene knockdown studies.

ERV1/HERVE solo-LTR

GEO accession	HERV pattern	Down DEG		Up DEG	
		OR	p-value	OR	p-value
H3K4me1_GSE59695	All	0.1602	0	0.1927	0
	Sense	0.3296	0	0.3892	0
	Anti-sense	0.2838	0	0.3523	0
	Intragenic	1.0799	0.6225193	1.2162	0.11243188
	Intergenic	0.1698	0	0.1971	0
H3K4me2_GSE22859	All	0.1559	0	0.1611	0
	Sense	0.3422	0	0.3521	0
	Anti-sense	0.3044	0	0.2928	0
	Intragenic	1.1494	0.09783741	1.0342	0.73252917
	Intergenic	0.1634	0	0.1695	0
H3K9_GSE44084	All	0	0	0.0005	0
	Sense	0	0	0.0014	0
	Anti-sense	0	0	0.0012	0
	Intragenic	0	0	0	0
	Intergenic	0	0	0.0006	0
H3K9me3_GSE25282	All	0.2387	0.10893156	0.3409	0.10377159
	Sense	0.5973	0.62800843	0.8533	0.73862702
	Anti-sense	0.3094	0.10849076	0.7738	0.72506798
	Intragenic	0.6662	0.73308789	1.1662	0.79825837
	Intergenic	0.2526	0.11913333	0.3608	0.11699692
H3K9me3_GSE41040	All	0.2137	0	0.1574	0
	Sense	0.3971	0	0.3373	0
	Anti-sense	0.3737	0	0.3104	0
	Intragenic	0.911	0.20823849	1.1729	0.00489918
	Intergenic	0.2277	0	0.1649	0
SETDB1_GSE45175	All	0.2955	0.07817655	0.2098	1.60E-07
	Sense	0.3752	0.07143696	0.4111	0.00074238
	Anti-sense	0.6706	0.46481864	0.3956	0.00045769
	Intragenic	2.3334	0.10046444	0.4385	0.00511858
	Intergenic	0.3127	0.08857196	0.2221	3.70E-07
SETDB1_GSE73231	All	0.0001	0	0	0
	Sense	0.0005	0	0.0001	0
	Anti-sense	0.0004	0	0.0001	0
	Intragenic	0.0073	0	0	0
	Intergenic	0.0002	0	0	0

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
TRIM28_gse61639	All	inf	1	0.4094	0.37015791
	Sense	1.0241	1	1.0241	1
	Anti-sense	inf	0.60429054	0.3869	0.23918297
	Intragenic	0	0.11106357	3.1105	0.20838854
	Intergenic	inf	1	0.4333	0.38633196

ERV2/HERVK solo-LTR

H3K4me1_GSE59695	All	0.7441	0.06655765	0.75	0.01822931
	Sense	0.8406	0.38116924	0.8236	0.19818313
	Anti-sense	0.6858	0.04193788	0.7341	0.03038839
	Intragenic	0.6627	0.29184958	1.3459	0.1592962
	Intergenic	0.7178	0.04488832	0.7511	0.0203067
H3K4me2_GSE22859	All	0.6458	9.00E-08	0.7226	0.00055385
	Sense	0.7124	0.00048932	0.7414	0.00787723
	Anti-sense	0.7027	0.00016332	0.7845	0.02332106
	Intragenic	1.0725	0.6392624	0.8471	0.47155876
	Intergenic	0.6329	4.00E-08	0.7377	0.00134659
H3K9_GSE44084	All	0	0	0	0
	Sense	0	0	0	0
	Anti-sense	0	0	0	0
	Intragenic	0	0.00011923	0	0.00037728
	Intergenic	0	0	0	0
H3K9me3_GSE25282	All	0.6912	0.74268879	1.7289	0.33941363
	Sense	0.8794	1	1.184	0.78444887
	Anti-sense	0.7691	1	2.1548	0.1120043
	Intragenic	0	1	2.809	0.11373028
	Intergenic	0.757	1	1.5146	0.47094426
H3K9me3_GSE41040	All	0.7614	6.91E-05	0.5894	0
	Sense	0.7869	0.00314985	0.6751	0
	Anti-sense	0.7792	0.00140923	0.6031	0
	Intragenic	0.9593	0.84259056	1.1175	0.26775267
	Intergenic	0.771	0.00016059	0.5716	0
SETDB1_GSE45175	All	1.3829	0.6145428	0.678	0.11622941
	Sense	1.026	1	0.6884	0.24640907
	Anti-sense	1.6157	0.39825251	0.8114	0.4581713
	Intragenic	2.0054	0.28864514	0.1728	0.04376174
	Intergenic	1.1778	0.80090074	0.7427	0.21581801

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
SETDB1_GSE73231	All	0.0021	0	0	0
	Sense	0.0048	0	0	0
	Anti-sense	0	0	0	0
	Intragenic	0.022	0	0	0
	Intergenic	0.0023	0	0	0
TRIM28_gse61639	All	3.4576	0.13870927	1.8438	0.46310899
	Sense	4.1057	0.06620708	4.1057	0.06620708
	Anti-sense	0.4486	0.68244559	0.4486	0.68244559
	Intragenic	0	1	2.3391	0.38244404
	Intergenic	3.7866	0.12278983	2.0192	0.44698496

ERV3/HERVL solo-LTR

H3K4me1_GSE59695	All	0.1524	0	0.1919	0
	Sense	0.3808	2.00E-08	0.4232	0
	Anti-sense	0.342	0	0.3898	0
	Intragenic	1.2886	0.121324	1.3982	0.005604
	Intergenic	0.1612	0	0.1994	0
H3K4me2_GSE22859	All	0.1559	0	0.1558	0
	Sense	0.3566	0	0.3348	0
	Anti-sense	0.3394	0	0.3193	0
	Intragenic	1.1677	0.058753	1.029	0.77176
	Intergenic	0.1656	0	0.1652	0
H3K9_GSE44084	All	0	0	0.0005	0
	Sense	0	0	0.0015	0
	Anti-sense	0	0	0.0014	0
	Intragenic	0	0	0	0
	Intergenic	0	0	0.0006	0
H3K9me3_GSE25282	All	0.2337	0.105291	0.3337	0.09913
	Sense	0.638	0.636504	0.6379	0.507711
	Anti-sense	0.3363	0.128247	0.4371	0.166142
	Intragenic	1.1084	1	0.6332	0.611038
	Intergenic	0.2467	0.114772	0.2465	0.027327
H3K9me3_GSE41040	All	0.2158	0	0.1611	0
	Sense	0.417	0	0.3679	0
	Anti-sense	0.4198	0	0.3352	0
	Intragenic	0.9308	0.336312	1.2266	0.000245
	Intergenic	0.2247	0	0.1703	0

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
SETDB1_GSE45175	All	0.2001	0.015209	0.1923	2.00E-08
	Sense	0.4008	0.087008	0.37	7.86E-05
	Anti-sense	0.2801	0.019479	0.3059	1.76E-06
	Intragenic	2.2178	0.110567	0.4555	0.00565
	Intergenic	0.2113	0.018047	0.2031	5.00E-08
SETDB1_GSE73231	All	0.0001	0	0	0
	Sense	0.0005	0	0.0001	0
	Anti-sense	0.0005	0	0.0001	0
	Intragenic	0.0069	0	0	0
	Intergenic	0.0002	0	0	0
TRIM28_gse61639	All	inf	1	0.4008	0.364187
	Sense	inf	0.604783	1.094	1
	Anti-sense	inf	0.603397	1.0094	1
	Intragenic	0	0.107354	2.9565	0.214325
	Intergenic	inf	1	0.4232	0.379514

ERV3/HERVL-MaLR solo-LTR

H3K4me1_GSE59695	All	0.0897	0	0.1131	0
	Sense	0.1431	0	0.1768	0
	Anti-sense	0.1341	0	0.1718	0
	Intragenic	1.0176	0.93954	1.3002	0.025795
	Intergenic	0.0908	0	0.1145	0
H3K4me2_GSE22859	All	0.0827	0	0.0869	0
	Sense	0.1456	0	0.1399	0
	Anti-sense	0.1276	0	0.131	0
	Intragenic	1.0549	0.505659	0.9249	0.392258
	Intergenic	0.0831	0	0.088	0
H3K9_GSE44084	All	0	0	0.0003	0
	Sense	0	0	0.0005	0
	Anti-sense	0	0	0.0004	0
	Intragenic	0	0	0	0
	Intergenic	0	0	0.0003	0
H3K9me3_GSE25282	All	0.1292	0.038735	0.1845	0.024744
	Sense	0.2198	0.095387	0.2196	0.01913
	Anti-sense	0.1955	0.078759	0.1953	0.0132
	Intragenic	0.631	0.739804	0.4853	0.23468
	Intergenic	0.1307	0.039506	0.1866	0.025446

GEO accession	hervList	Down DEG		Up DEG	
		OR	p-value	OR	p-value
SETDB1_GSE45175	All	0.1599	0.0179	0.1295	0
	Sense	0.272	0.064896	0.193	4.00E-08
	Anti-sense	0.1674	0.008541	0.1714	1.00E-08
	Intragenic	2.7784	0.074748	0.4619	0.001657
	Intergenic	0.1617	0.018418	0.131	0
SETDB1_GSE73231	All	0	0	0	0
	Sense	0.0001	0	0	0
	Anti-sense	0.0001	0	0	0
	Intragenic	0.0039	0	0	0
	Intergenic	0	0	0	0
TRIM28_gse61639	All	inf	1	0.2218	0.224513
	Sense	inf	1	0.3769	0.347344
	Anti-sense	inf	1	0.3354	0.316772
	Intragenic	0	0.02012	1.6832	0.706967
	Intergenic	inf	1	0.2242	0.226683

Table B4. Functional annotation analysis of intragenic HERV associated over-expressed genes in SLE
ERV1/ERVE

Term - Biological Function	PValue	FDR
GO:0007155~cell adhesion	1.79E-05	0.031974
GO:0022610~biological adhesion	1.80E-05	0.032142
GO:0006793~phosphorus metabolic process	2.62E-04	0.465859
GO:0006796~phosphate metabolic process	2.62E-04	0.465859
GO:0048666~neuron development	4.63E-04	0.822743
GO:0030182~neuron differentiation	6.18E-04	1.096611
GO:0030030~cell projection organization	8.91E-04	1.578573
GO:0007409~axonogenesis	0.00102	1.805016
GO:0007167~enzyme linked receptor protein signaling pathway	0.001112	1.965951
GO:0048667~cell morphogenesis involved in neuron differentiation	0.001147	2.028433
GO:0007268~synaptic transmission	0.001204	2.127704
GO:0048812~neuron projection morphogenesis	0.001457	2.569179
GO:0015698~inorganic anion transport	0.00163	2.870146
GO:0007242~intracellular signaling cascade	0.001669	2.938353
GO:0048858~cell projection morphogenesis	0.001707	3.004174
GO:0016337~cell-cell adhesion	0.001809	3.180801
GO:0006468~protein amino acid phosphorylation	0.001993	3.498344
GO:0007214~gamma-aminobutyric acid signaling pathway	0.002155	3.777331
GO:0051056~regulation of small GTPase mediated signal transduction	0.002454	4.29141
Term - Cellular Component		
GO:0044459~plasma membrane part	4.44E-08	6.25E-05
GO:0045202~synapse	1.67E-07	2.34E-04
GO:0044456~synapse part	1.16E-06	0.001629
GO:0030054~cell junction	3.33E-06	0.004693
GO:0005626~insoluble fraction	3.78E-06	0.005322
GO:0005886~plasma membrane	5.92E-06	0.008338
GO:0005624~membrane fraction	5.56E-05	0.07818
GO:0042995~cell projection	2.42E-04	0.340513
GO:0031226~intrinsic to plasma membrane	2.81E-04	0.395478
GO:0005856~cytoskeleton	2.95E-04	0.41425
GO:0000267~cell fraction	3.98E-04	0.559286
GO:0045211~postsynaptic membrane	5.04E-04	0.707481
GO:0005912~adherens junction	8.37E-04	1.17214
GO:0044463~cell projection part	0.00126	1.759215
GO:0005887~integral to plasma membrane	0.001434	1.999142
GO:0043005~neuron projection	0.002449	3.392555
GO:0070161~anchoring junction	0.00261	3.611965
GO:0009986~cell surface	0.003131	4.317599

Term - Biological Function	PValue	FDR
GO:0015629~actin cytoskeleton	0.003381	4.655253
GO:0034707~chloride channel complex	0.00359	4.936643
Term - Molecular Function		
GO:0030695~GTPase regulator activity	3.21E-06	0.004994
GO:0060589~nucleoside-triphosphatase regulator activity	5.56E-06	0.008641
GO:0030554~adenyl nucleotide binding	9.09E-06	0.014128
GO:0001883~purine nucleoside binding	1.11E-05	0.017225
GO:0001882~nucleoside binding	1.54E-05	0.02393
GO:0005524~ATP binding	1.69E-05	0.026263
GO:0032559~adenyl ribonucleotide binding	1.78E-05	0.02767
GO:0017076~purine nucleotide binding	7.04E-05	0.109432
GO:0032553~ribonucleotide binding	1.30E-04	0.202057
GO:0032555~purine ribonucleotide binding	1.30E-04	0.202057
GO:0005083~small GTPase regulator activity	1.87E-04	0.290786
GO:0017124~SH3 domain binding	2.30E-04	0.356573
GO:0043167~ion binding	2.43E-04	0.377774
GO:0019904~protein domain specific binding	4.09E-04	0.633491
GO:0005096~GTPase activator activity	4.97E-04	0.770519
GO:0000166~nucleotide binding	7.17E-04	1.108071
GO:0008047~enzyme activator activity	0.00105	1.620249
GO:0019899~enzyme binding	0.001095	1.68884
GO:0003779~actin binding	0.0014	2.153693
GO:0008237~metallopeptidase activity	0.001614	2.478916
GO:0008092~cytoskeletal protein binding	0.00178	2.73101
GO:0043169~cation binding	0.001864	2.858312
GO:0046872~metal ion binding	0.00235	3.59183
GO:0005230~extracellular ligand-gated ion channel activity	0.002552	3.893704
GO:0005254~chloride channel activity	0.002552	3.893704
GO:0008081~phosphoric diester hydrolase activity	0.002559	3.90452
GO:0005089~Rho guanyl-nucleotide exchange factor activity	0.003133	4.760087
GO:0005085~guanyl-nucleotide exchange factor activity	0.003181	4.831961

ERV3-ERVL

Term - Biological Function		
GO:0022610~biological adhesion	1.46E-06	0.002627
GO:0007155~cell adhesion	1.47E-06	0.002638
GO:0051270~regulation of cell motion	1.50E-05	0.026957
GO:0007214~gamma-aminobutyric acid signaling pathway	3.63E-05	0.065231
GO:0042692~muscle cell differentiation	3.71E-05	0.066667
GO:0040012~regulation of locomotion	3.99E-05	0.07173
GO:0015698~inorganic anion transport	6.31E-05	0.113489

Term - Biological Function	PValue	FDR
GO:0030334~regulation of cell migration	1.30E-04	0.233112
GO:0006897~endocytosis	1.34E-04	0.240103
GO:0010324~membrane invagination	1.34E-04	0.240103
GO:0007167~enzyme linked receptor protein signaling pathway	1.36E-04	0.244618
GO:0016337~cell-cell adhesion	1.63E-04	0.292148
GO:0007229~integrin-mediated signaling pathway	1.86E-04	0.334262
GO:0006821~chloride transport	2.16E-04	0.387625
GO:0048666~neuron development	2.53E-04	0.454706
GO:0030182~neuron differentiation	4.55E-04	0.815529
GO:0051146~striated muscle cell differentiation	4.67E-04	0.836333
GO:0034330~cell junction organization	5.22E-04	0.93459
GO:0007268~synaptic transmission	5.89E-04	1.053337
GO:0007044~cell-substrate junction assembly	7.99E-04	1.427642
GO:0034329~cell junction assembly	9.05E-04	1.614692
GO:0006820~anion transport	9.19E-04	1.639583
GO:0007242~intracellular signaling cascade	0.001008	1.796532
GO:0051056~regulation of small GTPase mediated signal transduction	0.001046	1.864234
GO:0030030~cell projection organization	0.001073	1.911849
GO:0001932~regulation of protein amino acid phosphorylation	0.001244	2.213311
GO:0051240~positive regulation of multicellular organismal process	0.00147	2.610331
GO:0045087~innate immune response	0.001674	2.968469
GO:0007160~cell-matrix adhesion	0.001727	3.060784
GO:0019226~transmission of nerve impulse	0.001827	3.23439
GO:0007169~transmembrane receptor protein tyrosine kinase signaling	0.002305	4.064838
GO:0009187~cyclic nucleotide metabolic process	0.002666	4.686284
Term - Cellular Component		
GO:0044459~plasma membrane part	8.54E-14	1.20E-10
GO:0005886~plasma membrane	5.23E-11	7.37E-08
GO:0045202~synapse	2.06E-10	2.91E-07
GO:0030054~cell junction	8.39E-09	1.18E-05
GO:0044456~synapse part	7.98E-08	1.12E-04
GO:0031226~intrinsic to plasma membrane	2.62E-07	3.69E-04
GO:0005887~integral to plasma membrane	1.16E-06	0.00164
GO:0005626~insoluble fraction	1.70E-06	0.002393
GO:0005912~adherens junction	4.64E-06	0.006539
GO:0042995~cell projection	6.75E-06	0.009522
GO:0005624~membrane fraction	1.35E-05	0.019065
GO:0016323~basolateral plasma membrane	1.39E-05	0.019591
GO:0005856~cytoskeleton	2.33E-05	0.032855
GO:0070161~anchoring junction	2.50E-05	0.035281
GO:0045211~postsynaptic membrane	7.80E-05	0.109871
GO:0030055~cell-substrate junction	8.40E-05	0.118403

Term - Biological Function	PValue	FDR
GO:0031410~cytoplasmic vesicle	1.36E-04	0.192024
GO:0009986~cell surface	1.43E-04	0.20137
GO:0000267~cell fraction	1.61E-04	0.226132
GO:0031982~vesicle	2.14E-04	0.300944
GO:0005925~focal adhesion	3.52E-04	0.494558
GO:0044463~cell projection part	3.68E-04	0.517971
GO:0005924~cell-substrate adherens junction	5.24E-04	0.73656
GO:0019898~extrinsic to membrane	8.23E-04	1.153741
GO:0016023~cytoplasmic membrane-bounded vesicle	0.001068	1.496011
GO:0034707~chloride channel complex	0.001113	1.557589
GO:0015629~actin cytoskeleton	0.001197	1.674262
GO:0031091~platelet alpha granule	0.001271	1.776693
GO:0009898~internal side of plasma membrane	0.001299	1.816339
GO:0014069~postsynaptic density	0.001963	2.733011
GO:0031988~membrane-bounded vesicle	0.001981	2.757046
GO:0043005~neuron projection	0.002085	2.900262
GO:0012505~endomembrane system	0.002572	3.565734
Term - Molecular Function		
GO:0030695~GTPase regulator activity	2.97E-07	4.61E-04
GO:0060589~nucleoside-triphosphatase regulator activity	5.61E-07	8.70E-04
GO:0017124~SH3 domain binding	1.53E-06	0.002371
GO:0003779~actin binding	4.71E-06	0.007302
GO:0008092~cytoskeletal protein binding	1.20E-05	0.018609
GO:0008047~enzyme activator activity	2.13E-05	0.032982
GO:0019904~protein domain specific binding	3.88E-05	0.060177
GO:0005096~GTPase activator activity	6.14E-05	0.095207
GO:0005509~calcium ion binding	6.25E-05	0.096838
GO:0005254~chloride channel activity	6.52E-05	0.10106
GO:0005083~small GTPase regulator activity	7.27E-05	0.112596
GO:0031404~chloride ion binding	1.33E-04	0.206744
GO:0016917~GABA receptor activity	1.46E-04	0.225534
GO:0005253~anion channel activity	1.53E-04	0.236636
GO:0043167~ion binding	1.61E-04	0.25009
GO:0004890~GABA-A receptor activity	2.72E-04	0.420876
GO:0043168~anion binding	8.01E-04	1.234747
GO:0004222~metalloendopeptidase activity	8.38E-04	1.290964
GO:0004725~protein tyrosine phosphatase activity	8.38E-04	1.290964
GO:0008509~anion transmembrane transporter activity	0.001414	2.169419
GO:0005085~guanyl-nucleotide exchange factor activity	0.002035	3.108747
GO:0046872~metal ion binding	0.002544	3.872071
GO:0043169~cation binding	0.002626	3.995685
GO:0008237~metallopeptidase activity	0.002713	4.124443
GO:0001883~purine nucleoside binding	0.00311	4.715255

Term - Biological Function	PValue	FDR
LTR7		
Term - Molecular Function		
GO:0046870~cadmium ion binding	1.77E-06	0.0021640
GO:0005507~copper ion binding	6.99E-04	0.8497616
GO:0031267~small GTPase binding	0.002104	2.5375235
GO:0051020~GTPase binding	0.002614	3.1438198
LTR33		
Term - Biological Function		
GO:0048666~neuron development	4.06E-05	0.06711
GO:0007155~cell adhesion	1.73E-04	0.286508
GO:0030182~neuron differentiation	1.76E-04	0.291357
GO:0022610~biological adhesion	1.77E-04	0.291991
GO:0009187~cyclic nucleotide metabolic process	9.48E-04	1.556741
GO:0009123~nucleoside monophosphate metabolic process	9.98E-04	1.638803
GO:0046058~cAMP metabolic process	0.001749	2.85506
GO:0046928~regulation of neurotransmitter secretion	0.001749	2.85506
GO:0007268~synaptic transmission	0.002512	4.075657
Term - Cellular Component		
GO:0005856~cytoskeleton	3.65E-04	0.468627
GO:0031012~extracellular matrix	3.74E-04	0.480741
GO:0045202~synapse	4.91E-04	0.630079
GO:0044459~plasma membrane part	6.53E-04	0.837969
GO:0005578~proteinaceous extracellular matrix	6.57E-04	0.842559
GO:0030054~cell junction	7.28E-04	0.933463
GO:0016010~dystrophin-associated glycoprotein complex	0.00113	1.444845
GO:0005886~plasma membrane	0.002405	3.052403
GO:0019898~extrinsic to membrane	0.003448	4.349105
GO:0044456~synapse part	0.003648	4.596318
Term - Molecular Function		
GO:0005509~calcium ion binding	1.18E-06	0.001638
GO:0043167~ion binding	6.25E-05	0.086697
GO:0008081~phosphoric diester hydrolase activity	6.33E-05	0.087909
GO:0004115~3',5'-cyclic-AMP phosphodiesterase activity	1.33E-04	0.184525
GO:0043169~cation binding	4.66E-04	0.645097
GO:0046872~metal ion binding	6.23E-04	0.861255
GO:0004114~3',5'-cyclic-nucleotide phosphodiesterase activity	0.00281	3.832194
GO:0008092~cytoskeletal protein binding	0.002847	3.881698
GO:0004112~cyclic-nucleotide phosphodiesterase activity	0.003165	4.307123

Term - Biological Function	PValue	FDR

MLT1D

Term - Biological Function		
GO:0007155~cell adhesion	5.70E-08	9.68E-05
GO:0022610~biological adhesion	5.81E-08	9.87E-05
GO:0019226~transmission of nerve impulse	4.60E-05	0.078081
GO:0007268~synaptic transmission	5.42E-05	0.092046
GO:0009187~cyclic nucleotide metabolic process	1.23E-04	0.208011
GO:0016337~cell-cell adhesion	2.25E-04	0.381773
GO:0050808~synapse organization	2.65E-04	0.448588
GO:0048666~neuron development	2.75E-04	0.465837
GO:0030182~neuron differentiation	3.44E-04	0.582589
GO:0007156~homophilic cell adhesion	4.15E-04	0.702244
GO:0007519~skeletal muscle tissue development	4.32E-04	0.731964
GO:0060538~skeletal muscle organ development	4.32E-04	0.731964
GO:0050806~positive regulation of synaptic transmission	6.44E-04	1.088676
GO:0030030~cell projection organization	7.24E-04	1.222972
GO:0051971~positive regulation of transmission	9.59E-04	1.616694
GO:0031646~positive regulation of neurological system process	0.001224	2.059853
GO:0031644~regulation of neurological system process	0.001383	2.324126
GO:0050804~regulation of synaptic transmission	0.002181	3.642539
GO:0043062~extracellular structure organization	0.002215	3.696958
GO:0009123~nucleoside monophosphate metabolic process	0.002479	4.130301
GO:0007242~intracellular signaling cascade	0.002561	4.263534
Term - Cellular Component		
GO:0045202~synapse	1.92E-07	2.56E-04
GO:0030054~cell junction	2.37E-06	0.003151
GO:0044459~plasma membrane part	4.93E-06	0.00657
GO:0044456~synapse part	6.19E-06	0.008243
GO:0005886~plasma membrane	1.42E-05	0.018907
GO:0005856~cytoskeleton	1.88E-05	0.024995
GO:0019898~extrinsic to membrane	2.41E-05	0.032138
GO:0015629~actin cytoskeleton	7.29E-05	0.096976
GO:0044463~cell projection part	1.65E-04	0.220011
GO:0005912~adherens junction	5.24E-04	0.695764
GO:0031226~intrinsic to plasma membrane	7.44E-04	0.985549
GO:0014069~postsynaptic density	8.12E-04	1.075371
GO:0005913~cell-cell adherens junction	8.50E-04	1.126367
GO:0042995~cell projection	0.001079	1.427751
GO:0005626~insoluble fraction	0.001148	1.517216
GO:0070161~anchoring junction	0.00124	1.638708
GO:0043197~dendritic spine	0.001248	1.64846

Term - Biological Function	PValue	FDR
GO:0005887~integral to plasma membrane	0.001615	2.129118
GO:0030425~dendrite	0.002772	3.628239
GO:0043005~neuron projection	0.003025	3.953829
GO:0042734~presynaptic membrane	0.003259	4.252815
GO:0009986~cell surface	0.003588	4.673085
Term - Molecular Function		
GO:0043167~ion binding	2.11E-05	0.030787
GO:0005509~calcium ion binding	2.18E-05	0.031814
GO:0030695~GTPase regulator activity	2.77E-05	0.040375
GO:0060589~nucleoside-triphosphatase regulator activity	3.91E-05	0.057097
GO:0043169~cation binding	7.78E-05	0.113479
GO:0046872~metal ion binding	8.12E-05	0.118447
GO:0005096~GTPase activator activity	1.00E-04	0.146583
GO:0060090~molecular adaptor activity	6.31E-04	0.916449
GO:0017124~SH3 domain binding	0.001193	1.727193
GO:0004114~3',5'-cyclic-nucleotide phosphodiesterase activity	0.001686	2.433324
GO:0004112~cyclic-nucleotide phosphodiesterase activity	0.001973	2.841981
GO:0003779~actin binding	0.002221	3.193667
GO:0005083~small GTPase regulator activity	0.002867	4.104468

MSTD

Term - Biological Function		
GO:0007155~cell adhesion	5.69E-06	0.009175
GO:0022610~biological adhesion	5.82E-06	0.009379
GO:0016337~cell-cell adhesion	2.04E-05	0.032842
Term - Cellular Component		
GO:0044459~plasma membrane part	6.09E-04	0.771758
Term - Molecular Function		
GO:0008092~cytoskeletal protein binding	0.001085	1.482524
GO:0004114~3',5'-cyclic-nucleotide phosphodiesterase activity	0.002104	2.856019
GO:0004112~cyclic-nucleotide phosphodiesterase activity	0.002372	3.214559
GO:0003779~actin binding	0.002794	3.77533

THE1B

Term - Biological Function		
GO:0007155~cell adhesion	2.37E-07	4.03E-04
GO:0022610~biological adhesion	2.41E-07	4.09E-04
GO:0019226~transmission of nerve impulse	1.63E-04	0.277736
GO:0007268~synaptic transmission	2.33E-04	0.395441
GO:0060538~skeletal muscle organ development	2.77E-04	0.470567
GO:0007519~skeletal muscle tissue development	2.77E-04	0.470567
GO:0048666~neuron development	3.33E-04	0.56555
GO:0030182~neuron differentiation	9.02E-04	1.524226

Term - Biological Function	PValue	FDR
Term - Cellular Component		
GO:0030054~cell junction	6.37E-07	8.57E-04
GO:0044459~plasma membrane part	6.92E-07	9.31E-04
GO:0045202~synapse	8.47E-07	0.001139
GO:0044456~synapse part	2.52E-06	0.003391
GO:0005626~insoluble fraction	3.32E-05	0.044703
GO:0005856~cytoskeleton	2.22E-04	0.298402
GO:0005886~plasma membrane	3.37E-04	0.4525
GO:0005624~membrane fraction	4.33E-04	0.581448
GO:0009986~cell surface	7.12E-04	0.952974
GO:0000267~cell fraction	0.001052	1.405875
GO:0005912~adherens junction	0.001193	1.592242
GO:0009897~external side of plasma membrane	0.002373	3.145102
GO:0070161~anchoring junction	0.002576	3.409182
GO:0015629~actin cytoskeleton	0.003254	4.289384
Term - Molecular Function		
GO:0003779~actin binding	4.04E-04	0.580517
GO:0005509~calcium ion binding	5.65E-04	0.809919
GO:0004115~3',5'-cyclic-AMP phosphodiesterase activity	6.52E-04	0.934412
GO:0008092~cytoskeletal protein binding	6.98E-04	0.999885
GO:0004222~metalloendopeptidase activity	0.00132	1.883978
GO:0004114~3',5'-cyclic-nucleotide phosphodiesterase activity	0.001325	1.890629
GO:0008237~metallopeptidase activity	0.00139	1.98271
GO:0004112~cyclic-nucleotide phosphodiesterase activity	0.001552	2.211509
GO:0019899~enzyme binding	0.002558	3.620145

THE1C

Term - Biological Function		
GO:0007155~cell adhesion	3.02E-06	0.00493
GO:0022610~biological adhesion	3.09E-06	0.005053
GO:0050808~synapse organization	8.02E-04	1.302657
GO:0007185~transmembrane receptor protein tyrosine phosphatase signaling pathway	0.002877	4.599809
Term - Cellular Component		
GO:0005886~plasma membrane	2.61E-04	0.336618
GO:0044459~plasma membrane part	2.66E-04	0.343818
GO:0016010~dystrophin-associated glycoprotein complex	0.001193	1.531411
GO:0031224~intrinsic to membrane	0.00214	2.732228
GO:0030054~cell junction	0.002424	3.090474
GO:0016021~integral to membrane	0.002829	3.597619
Term - Molecular Function		
GO:0005216~ion channel activity	9.81E-04	1.348007

Term - Biological Function	PValue	FDR
GO:0022838~substrate specific channel activity	0.001295	1.77532
GO:0019198~transmembrane receptor protein phosphatase activity	0.001308	1.793287
GO:0030695~GTPase regulator activity	0.001481	2.027636
GO:0015267~channel activity	0.001763	2.409153
GO:0022803~passive transmembrane transporter activity	0.001801	2.460709
GO:0060589~nucleoside-triphosphatase regulator activity	0.001801	2.460709
GO:0005509~calcium ion binding	0.002021	2.757363

THE1D

Term - Biological Function		
GO:0050808~synapse organization	2.39E-04	0.397623
GO:0008104~protein localization	3.92E-04	0.65175
GO:0043112~receptor metabolic process	9.71E-04	1.608948
GO:0043062~extracellular structure organization	0.002499	4.091787
GO:0060538~skeletal muscle organ development	0.002593	4.242038
GO:0007519~skeletal muscle tissue development	0.002593	4.242038
Term - Cellular Component		
GO:0045202~synapse	1.52E-06	0.001972
GO:0044456~synapse part	1.94E-05	0.02517
GO:0044459~plasma membrane part	1.81E-04	0.233886
GO:0030054~cell junction	2.78E-04	0.359693
GO:0005886~plasma membrane	8.86E-04	1.141762
GO:0042995~cell projection	0.001719	2.203694
GO:0043005~neuron projection	0.001898	2.429726
GO:0016010~dystrophin-associated glycoprotein complex	0.001914	2.450228
GO:0009986~cell surface	0.002212	2.827497
GO:0015629~actin cytoskeleton	0.0024	3.06347
GO:0044463~cell projection part	0.002731	3.479362
Term - Molecular Function		
GO:0005509~calcium ion binding	5.91E-05	0.083317
GO:0043167~ion binding	3.25E-04	0.456931
GO:0004114~3',5'-cyclic-nucleotide phosphodiesterase activity	4.39E-04	0.616967
GO:0030695~GTPase regulator activity	4.66E-04	0.65522
GO:0004112~cyclic-nucleotide phosphodiesterase activity	5.16E-04	0.725414
GO:0060589~nucleoside-triphosphatase regulator activity	5.94E-04	0.834066
GO:0043169~cation binding	0.001482	2.069051
GO:0003779~actin binding	0.001493	2.083938
GO:0046872~metal ion binding	0.001764	2.457321
GO:0008081~phosphoric diester hydrolase activity	0.001768	2.462905
GO:0019198~transmembrane receptor protein phosphatase activity	0.002408	3.340632
GO:0017016~Ras GTPase binding	0.002655	3.678445
GO:0005083~small GTPase regulator activity	0.003237	4.466659

VITA

Mr. Pumipat was born on March 4th, 1983 in Sing Buri, Thailand. He graduated with the Bachelor of Science degree in Agricultural Biotechnology with first class honor from Kasetsart University in 2005 and Master of Science degree in Bioinformatics from King Mongkut's University of Technology Thonburi in 2008. He got a Royal Golden Jubilee (RGJ) Ph.D. Scholarship from the Thailand Research Fund (TRF) and participated in Biomedical Science program, Graduate School, Chulalongkorn University for philosophy degree in 2010.