COMPUTER BASED SCREENING OF ENDOGENOUS RETROVIRUS OR RT LIKE SEQUENCES FROM THE MOUSE GENOME PROJECT DATABASE

SARATHA THEVI A/P THRICHELVAM

BIOTECHNOLOGY PROGRAMME
SCHOOL OF SCIENCE AND TECHNOLOGY
UNIVERSITI MALAYSIA SABAH

MAC 2005

COMPUTER BASED SCREENING OF ENDOGENOUS RETROVIRUS OR RT LIKE SEQUENCES FROM MOUSE GENOME PROJECT DATABASE

SARATHA THEVI A/P THRICHELVAM

THIS DISSERTATION IS SUBMITTED
IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE
DEGREE OF BACHELOR OF SCIENCE WITH HONOURS

BIOTECHNOLOGY PROGRAMME
SCHOOL OF SCIENCE AND TECHNOLOGY
UNIVERSITI MALAYSIA SABAH

Mac 2005

DECLARATION

I hereby declare that this dissertation is based on my original work except for certain citations, quotations and summaries, which have been duly acknowledged.

31st MAC 2005

SARATHA THEVI A/P THRICHELVAM

HS2002-3116

AUTHENTICATION

Signature

1. **SUPERVISOR**

   (DR ROZIAH HAJI KAMBOL) _____

2. **EXAMINER 1**

   (DR WONG NYET KUI) _____

3. **EXAMINER 2**

   (DR ZALEHA ABDUL AZIZ) _____

4. **DEAN**

   (ASSOC.PROF DR.AMRAN AHMED) _____

PUMS 99:1

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN STATUS TESIS@

JUDUL: Computer based screening of Endogenous Retrovirus or RT like sequences from mouse genome database

Ijazah: SARJANA MUDA KEPUJIAN

SESI PENGAJIAN: 2002 - 2005

Saya SARATHA THEVI A/P THRICHELVAM

(HURUF BESAR)

mengaku membenarkan tesis (LPS/Sarjana/Doktor Falsafah)* ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:

1. Tesis adalah hakmilik Universiti Malaysia Sabah.
2. Perpustakaan Universiti Malaysia Sabah dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. **Sila tandakan (/)

☐	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
☐	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
☑	TIDAK TERHAD	

Disahkan oleh

(TANDATANGAN PENULIS)

(TANDATANGAN PUSTAKAWAN)

Alamat Tetap: 18 JLN 3, Taman Desa Temiang, 70200 Seremban, N.Sembilan.

DR. RODIAH KAMBOL

Nama Penyelia

Tarikh: 29/03/2005

Tarikh: _____

CATATAN: * Potong yang tidak berkenaan.

** Jika tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh tesis ini perlu dikelaskan sebagai SULIT dan TERHAD.

@ Tesis dimaksudkan sebagai tesis bagi Ijazah Doktor Falsafah dan Sarjana secara penyelidikan, atau disertasi bagi pengajian secara kerja kursus dan penyelidikan, atau Laporan Projek Sarjana Muda (LPSM).

ACKNOWLEDGEMENTS

I would like to take this great opportunity to extend my appreciation and gratitude to those who have contributed in completion of this project. In preparing for this study, I owe my gratitude to many people. First of all, I would like to express my gratitude to my supervisor Dr Roziah Kambol for her guidance throughout this project and also for her helpful and supportive comments on my writing for this project. Without her, it is not possible for me to complete this project successfully. I would like to thank my parents and family who have been a pillar of strength and comfort me throughout the years. I would like to thank my beloved parents Mr and Mrs Thrichelvam. My special thanks to Mr and Mrs Marimuthu for being a constant source of encouragement during my study. Lastly, I would like to express my thanks to my dear friends for their continuous encouragement and support. Thanks for everyone whom I mentioned and to who indirectly involved in this project.

ABSTRACT

In this study the reverse transcriptase gene is studied as it shows the most conserved motif in each domain. Two probes namely *Mouse Mammary Tumor Virus* (MMTV) and *Mus Dunni Endogenous Virus* (MDEV) are used to screen the mouse genome database. The four main steps involved in this study are obtaining sequences which act as probe, screening sequences inside the mouse genome database, sequence alignment and finally constructing a phylogenetic tree. A total of 100 sequences that have all the entire domain of reverse transcriptase gene have been successfully screened from mouse genome database. Only 16 sequences are selected from 100 sequence that have been screened based on several criteria for sequence alignment. This selected sequence is then aligned manually and using computer software Clustal W program. Two phylogenetic trees were constructed to study the connection of sequences from mouse endogenous retrovirus and also to investigate the relationship with 7 established genera of retroviruses with viruses obtained from this study. The major result of this study indicate that the endogenous retroviruses are present more in germ line which has 63 sequences compared to 37 sequences in autosome chromosome in both MMTV and MDEV probe. The data presented through this study suggest that endogenous retroviruses are distributed in 19 autosome and 1 sex mouse chromosome. Furthermore this tree support previous categorization where MMTV belong betaretroviruses genera and MDEV belongs to gammaretroviruses genera. From both phylogenetic trees, the sequences are classified according similarities in the reverse transcriptase gene.

ABSTRAK

Dalam kajian ini, gen transkriptase terbalik dikaji kerana ia banyak mengandungi jujukan motif terpelihara dalam setiap domain. Dua jujukan retrovirus endogenous yang bertindak sebagai probe iaitu *Mouse Mammary Tumor Virus* (MMTV) and *Mus Dunni Endogenous Retrovirus* (MDEV) telah digunakan untuk menyaring genom tikus. Terdapat empat langkah utama dalam proses penyaringan iaitu mendapatkan jujukan retrovirus endogenous tikus yang mempunyai motif terpelihara dari gen transkriptase berbalik yang bertindak sebagai probe, menyaring jujukan – jujukan yang terdapat dalam Genbank tikus, penyusunan jujukan dan akhirnya pembinaan pokok filogenetik. Sebanyak 100 jujukan yang mempunyai kesemua motif terpelihara dalam setiap domain berjaya disaringkan daripada genom tikus. Daripada 100 jujukan yang telah disaringkan hanya 16 jujukan dipilih untuk tujuan penyusunan jujukan berdasarkan beberapa criteria. Jujukan yang dipilih ini disusun secara manual dan menggunakan program komputer, Clustal W. Dua pokok filogenetik dibina untuk menggambarkan perhubungan antara jujukan – jujukan dalam retrovirus endogenous tikus dan juga untuk mengaji perhubungan antara virus yang diperolehi daripada kajian dengan 7 genera retrovirus. Daripada hasil kajian ini didapati endogenous retrovirus wujud lebih banyak dalam kromosom pembiakan iaitu terdapat 63 jujukan berbanding hanya 37 jujukan dalam kromosom autosom dalam kedua-dua probe. Data yang diperolehi melalui kajian ini menunjukkan endogenous retrovirus tersebar dalam 19 kromosom autosom dan 1 kromosom seks tikus. Kajian ini telah membuktikan bahawa MMTV tergolong dalam genera betaretovirus manakala MDEV tergolong dalam genera gammaretrovirus. Daripada kedua–dua pokok filogenetik, didapati jujukan dikelaskan berdasarkan persamaan dalam gen transkriptase berbalik.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Retroelements are classified as a mobile genetic elements which contain a reverse transcriptase gene and exist as DNA or RNA molecules. Elements that are transpose duplicatively via retrotransposition by an RNA intermediate called retroelements, for example the element is transcribed to RNA, then reverse-transcribed to DNA, and integrated elsewhere in the genome. It is known that viruses that encode for reverse transcriptase enzyme are known as retrovirus and the mobile elements that use reverse transcription in their propagation are called retroelements.

Retroelements can be classified into two major groups, viral and non-viral. Retroviruses are the known viral retroelements. Where else the non-viral retroelements can be divided into several groups such as Long Interspersed Elements (LINEs) and Short Interspersed Elements (SINEs). The total mass of SINE and LINE elements probably accounts for less than 15% of the mouse genome (Silver, 1995).

Retroviruses have the ability to transcribe their RNA genome into double stranded DNA intermediates by reverse transcriptase enzyme in order to propagate into the host organism. The enzyme reverse transcriptase (RT) is a virally encoded

protein (Patterson, 1994). Among the viruses, only Retrovirus has the capability to transform cells and also produce virus at the same time.

All retrovirus genome encode three types of protein which are known as structural protein, enzymatic and envelope proteins. Typically, there are 3 major regions in the retrovirus genome that are involved in their replication cycle. It is known as *gag* (group specific antigens), *pol* (polymerase), and *pro* (protease) (Swanstrom and Vogt, 1990). There is also *env* gene that encodes envelope glycoprotein for virus infectivity.

In addition endogenous retroviruses are included in Long Terminal Repeat (LTR) retrotransposons (Wilkinson *et al*, 1994). Retroviruses that have complete *gag*, *pro*, and *pol* genes are able to replicate efficiently and it is called as replication – competent retrovirus and it includes all the exogenous viruses that are transmitted by somatic cells.

In contrast, viruses that lack part of their *gag*, *pol* and *env* genes require another virus to provide them with those genes to complete their replication cycle. Thereby, these groups of viruses are defective in their gene arrangement and represent a group of viruses found in endogenous form where recombination or mutations have disrupted their viral genes. The endogenous form is transmitted through the germ line (sex cells) and is inherited from one generation to another generation as a stable gene.

The *Mus* genus has been divided into four subgenera, of which one is also called *Mus*. This subgenus contains all the house mouse *M.Musculus*. In this study, house mouse namely Mouse Mammary Tumor Virus (*MMTV*) and wild type mouse called Mus Dunni Endogenous Retroviruses (*MDEV*) are studied. The genome size of a mouse is 2600Mb and it has 20 chromosome including chromosome X and chromosome Y and 19 autosomes chromosome. The mouse is an excellent tool for genetics investigations because of its short generation time and small size.

Genomic resources for the mouse genome have increased greatly. Since mouse and humans are mammals, I choose mouse as model organisms because there is a high possibility that endogenous retrovirus in mouse may have similar genomic organizations like endogenous retrovirus in human. The DNA sequences are sufficiently related between mouse and human because most human genes and proteins will function in a mouse and the regulation of gene expression is highly conserved (Hall and Dunlop, 1997).

The endogenous retroviruses are located at different sites in the mouse genome and are vertically transmitted. One example of endogenous retrovirus is called Mouse Mammary Tumor Virus (MMTV). Today there exist model organism databases for the mouse, the Mouse Genome Database to provide comprehensive access to experimental and consensus data about mouse.

The objectives of this study are:-

- To screen the endogenous retrovirus in mouse genome using bioinformatics tools.
- To investigate the distribution of endogenous retrovirus in mouse genome by screening the entire mouse chromosome in mouse genome database.
- To perform sequence alignment from the gathered data.
- To build a phylogenetic tree from the gathered data in order to investigate the relationship of endogenous retrovirus in mouse with other established endogenous retrovirus genera.

CHAPTER 2

LITERATURE REVIEW

2.1 Retroelements

Retroelements are mobile genetic elements containing a reverse transcriptase gene and exist as DNA molecules and it is estimated to represent 10% in the mammalian genome (Taruscia and Mantovani, 1998). Retroelements are widely assumed to have an ancient origin.Through the process of retrotransposition, retroelements transpose duplicatively to an RNA intermediate where the element is transcribed to RNA or reverse-transcribed to DNA and it is integrated elsewhere in the genome. Retroelements also has a gene organization and replication cycle that is similar to retroviruses. Yet it is differ from retroviruses where they do not have an obligatory extra cellular phase and *env* gene in their genome (Campbell, 1990).

Retroelements can be classified into two major groups, viral and non-viral. Retroviruses are the known viral retroelements. This viral retroelements shares large structural homologies with the proviral form of retroviruses, which comprises retrotransposons and endogenous retroviruses. Non-viral retroelements can be divided into several groups. This non-viral retroelements contain three types of elements which are known as discussed in the following page:-

- Long interspersed elements (LINEs)

LINEs are known as autonomous retroelements because it has ability to produce their own reverse transcriptase. The examples are LINE 1 (L1) in human genome. LINEs contain up to 6,000 DNA base pairs long.

- Short interspersed elements (SINEs)

SINEs are known as non-autonomous retroelements which derived from small functional RNAs and usually contain 90 - 300 base pairs. Due to its lack of coding capacity it depends on long interspersed elements (LINEs) for their amplification (McCarthy and McDonald, 2004).

- Long Terminal Repeat (LTR).

The third major class is LTR retroelements which were included the endogenous retroviruses because it encodes their own reverse transcriptase and known as autonomous retroelements.

The Long Terminal Repeat (LTR) can be subclassified further into

I. Non LTR-retrotransposons.

The nonLTR-retrotransposons elements utilize a promoter sequence located within the 5' end of the coding sequence and make a polyadenylated RNA. The example of non LTR- retrotransposons is L1 elements in humans. They are similar to retroviruses in structure, with transcriptional regulatory sequences located in the flanking LTRs, some sort of priming site to allow priming of the reverse transcription that is usually located just downstream of the first LTR, and several open reading frames (ORFs) encoding proteins necessary for retrotransposition. These elements have no protein coding capacity but dependent on the L1 proteins for their retroposition.

II. LTR-retrotransposon,

LTR-retrotransposons are mobile genetic elements that make up a large fraction in the most mammalian genome where it exist about 8% in the human genome and approximately 10% in the mouse genome (Taruscia and Mantovani, 1998). Typically all families of mouse LTR retrotransposons are classified as a member of the *gypsy*-like superfamily of retroviral-like elements and generally the mouse LTR retrotransposons were classified into three distinct classes (class I, Class II and Class III) (Swanstrom and Vogt, 1990). Yet LTR retrotransposons are different from retroviruses because it has envelope genes and genomic components required for making a functional viral capsule.

2.2 Retroviruses

Retroviruses are infectious particles and it is grouped under one taxonomic unit, the family *Retroviridae.* The first retrovirus must have been able to free itself from the confines of the cell nucleus. It is done through association of retrovirus with a small number of proteins that allowed it to coat, and protect itself from the harsh extra cellular environment. Retroviruses are RNA-containing viruses that can convert their RNA genome into DNA molecules through a viral-associated reverse transcriptase which becomes activated upon cell infection.

According to Becker (1979), Dr.Peyton Rous was successfully isolated the first retrovirus from a connective tissue tumor (sarcoma) in a domestic chicken in 1910. While spumaviruses were the first retroviruses isolated from humans.

Traditionally, retroviruses were divided into groups based on their morphology which is listed below:-

- **A -type:** Also known as intracisternal particles. It is found in the genome of rodents as retrovirus like elements. It is a non-enveloped, immature particle which is only seen inside cells and believed to result from endogenous retrovirus-like genetic elements (Kuff and Leuders, 1998). Type A particles were generally 60 to 90nm in diameter with a nucleotide surrounded by a double shell.
- **B-type:** It is enveloped, extracellular particles with a condensed, acentric core and prominent envelope spikes, e.g. Mouse Mammary Tumor Virus (MMTV).
- **C-type:** Same as B-type, but it has a central core and barely visible spikes. It is found in most mammalian and avian retroviruses such as avian Leukaemia Virus (ALV), Human T cell Virus (HTLV), and Human Immunodeficiency Viruses (HIV).
- **D-type:** Usually slightly larger (to 120nm) and spikes less prominent. Mostly found in primates include mason Pfizer Monkey Virus (MPMV). (Roman, 1983).

According to Teich (1972), the family of *Retroviridae* consists of three subfamilies, *Oncovirinae* (found in avian and mammalian species), *Spumavirinae* (are foamy viruses found in mammalian species), and *Lentivirinae* (found in primates, sheep and goats). The term "onkos" (*oncovirinae*) means tumor in greek and it is separated into genera based on its morphologies and features of their maturation process during replication (Holland, 1992). The second subfamily, the "spuma"

REFERENCES

Attwood, T. K., and Smith, D.J.P., 1999. *Cell and molecular biology in action series: Introduction to Bioinformatics*. Pearson Education Limited, England.

Baillie, G. J., and Wilkins, R. J., 2001. Endogenous type D retrovirus in a marsupial, the common brushtail possum (*Trichosurus vulpecula*). *J. Virol.* **75:**2499-2507.

Barner, M. R., and Gary. I. C., 1999. *Bioinformatics for Geneticists*. Glaxosmithklie Pharmaceuticals, United Kingdom.

Baxevanis, A. D. and Ouellette, B. F. f., 2001. Bioinformatics: *A Practical Guide to the Analysis of Genes and Proteins*. In: Brinkman, F. S. L. and Leipe, D. D. *Phylogenetic Analysis*. British Columbia, Canada, 323 – 327.

Becker, Y. 1979. *Replication of Viral and cellular Genomes*. Martinus Nijhoff Publishing.

Benveniste, R.E., 1983. *The contributions of Retroviruses to the study of mammalian evolution*, Laboratory of Viral Carcinogenesis, National Cancer Institute, Maryland, 360-362.

Campbell, A., 1990. *Integration specificity of Retrotransposons and Retroviruses* - Annual Review of Genetics **24**, 491 – 518.

Coffin M.J., Hughes, S.H., and Varmus, H.E., 1997. *Retrotransposons, Endogenous Retroviruses and the Evolution of Retroelement*. Cold Spring Harbour laboratory Press.

Deininger, P. L., and Batzer M. A., 2002. *Mammalian retroelements*. Department of Biological Sciences, Biological Computation and Visualization Center, Louisiana State University, USA.

Foster, H. L., Small, J. D. and Fox, J. G., 1981. The Mouse in Biomedical Research, – *History, Genetics and Wild Mice* **1**, 1 -5.

Hall, J. C. and Dunlop, J. C. 1997. *Incorporating Molecular Genetics Medicin: Advances in Genetics* Department of Biology, Brandeis University, Waltham Massachusetts.

Herniou, E., 1998. *Retroviral Diversity and distribution in vertebrates*: Journal of virology. Department of biology, Imperial College, United Kingdom, **72**, 5955-5966.

Holland, J.J.1992. *Genetic Diversity of RNA Viruses:* Current topics in microbiology and Immunology. Springer – Verlag publisher, USA.

Kuff, E. L and Lueders, K. K., 1988. The intracisternal A-particle gene family: *structure and functional aspects – Adv Cancer* **51**,183–276.

Laughlin, M. A. and Pomerantz, R. J., 1994. *Retroviral Latency*. Academic Press, California.

McCarthy, E. M. and McDonald, J. F., 2004. Long terminal repeat retrotransposons of *Mus musculus*. http:// genomebiology.com/2004/5/3/R14.

Menzies, J. I. and Pennis. E., 1979. Journal of Virolgy: *Handbook of New Guinea Rodent* **6**, 21 – 25.

Ostel, J. M., Wheelan, S. J and Kans, J. A. *The NCBI Data Model*. National Center for Biotechnology Information, Maryland, 19 -32.

Patterson, M., 1997. *Why map the rat?* Trends in Genetics, Vol **13**(1), 152 – 165.

Perron, H., Garson, J.A., Beseme, F., Paranhos-Baccala, G., komurian-Pradel, F., Mallet, F., Tuke, P.W., Voisset, C., Blond, J.L., Lalandes, B., Seigneurind, J. M., Mandrand, B., 1997. *Molecular identification of a novel retrovirus repeatedly isolated from patients with multiple sclerosis*. Proc. Natl. Academic Scient, USA.

Roman, H. L., 1983. *Mapping in mouse*: Annual review of Genetics **17**, 91 – 97.

Schupbach, J. 1989. *Human Retrovirology Facts and Concepts*. Springer : Current topics in Microbiology and Immunology – Verlag publisher, USA.

Silver, L. M., 1995. *Mouse Genetics Concepts and Applications*. Oxford University Press, New York.

Swanstrom, R. and Vost, P.K., 1990. *Retroviruses Strategies of Replication* - Current Topics in Microbiology and Immunology **157**, 619 – 622.

Weitzman, J., 2002. *The mouse genome: An international consortium has generated a high-quality draf sequence of the mouse genome. Nature.420:520-562.*

Wilkinson, D.A, Mager D.L, Leong J.C., 1994. *Endogenous human retroviruses*. In: Levy J, Ed. The Retroviridae, New York: Plenum Press, pp. 465-535.

Wilson, J.F., 2003.*The model organism*: *Diseases by design, the scientist supplement*, Elsevier Science Ltd. 32-336.

Taruscio, D., and Mantovani, A., 1998. Human Endogenous retroviral sequences: *Posssible roles in Reproductive Physiopathology* **59**, 713-724.

Teich, N., 1982. Taxonomy of retroviruses In (eds) Weiss, R., Teich, N., Varmus, H and Coffin, J., Molecular Biology Of Tumor Viruses. 2nd edition. Cold Spring Harbor Lab Press, USA.

Tristem, M., 2000. *Identification and Characterization of Novel Human Endogenous Retrovirus Families by Phylogenetic Screening of the Human Genome Mapping Project Database:* Journal of Virology. **74**: 3715-3730