COMPUTER-BASED CHARACTERIZATION OF PRIMATE ENDOGENOUS RETROVIRUS IN HUMAN GENOME PROJECT

CHOW JIA LONG

THIS DISSERTATION IS TO FULFIL PART OF THE REQUIREMENT FOR A BACHELOR OF SCIENCE DEGREE WITH HONOURS

PROGRAM BIOTEKNOLOGI
SEKOLAH SAINS DAN TEKNOLOGI
UNIVERSITI MALAYSIA SABAH
2008

PUMS99:1

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN STATUS TESIS@

JUDUL: Computer-based characterization of primate endogenous retrovirus in human genome project

IJAZAH: Sarjana muda Bioteknologi

SAYA CHOW JIA LONG (HURUF BESAR) SESI PENGAJIAN: 2005/2008

mengaku membenarkan tesis (LPSM/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 institutsi 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	

(TANDATANGAN PENULIS)

Alamat Tetap: 34, tingkat Ipoh 3, Ipoh Garden, 31400 Ipoh, Perak.

Tarikh: 15-5-08

Disahkan Oleh

NURULAIN BINTI ISMAIL
LIBRARIAN
UNIVERSITI MALAYSIA SABAH

(TANDATANGAN PUSTAKAWAN)

Dr. Roziah bte Kamhol
Nama Penyelia

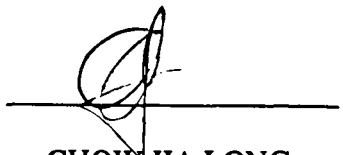
Tarikh: 13-5-08

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 disertai bagi pengajian secara kerja kursus dan Laporan Projek Sarjana Muda (LPSM).

DECLARATION

I hereby declare that the works in this dissertation are from my own except for the quotation and summaries which have been duly acknowledged.

11th May 2008



CHOW JIA LONG

HS2005-2359

APPROVAL OF EXAMINERS

Signature

1. **SUPERVISOR**

   (DR. ROZIAH HJ. KAMBOL)

   _____

2. **EXAMINER 1**

   (DR. WONG NYET KUI)

   _____

3. **DEAN**

   (ASSOC. PROF. DR. SHARIFF A. KADIR S. OMANG)

   _____

ACKNOWLEDGEMENT

First of all, I would like to take this chance to express my gratitude towards the lecturers of the Biotechnology program for their kind help and advice. Special thanks go out to Dr. Roziah Bte. Haji Kambol, my supervisor, for her constant guidance and expertise in helping me complete my dissertation. Dr. Roziah's words of encouragement helped to motivate me to work harder and to perform at my full potential. I also sincerely apologize to Dr. Roziah for causing any unwanted trouble, and ask for forgiveness for any wrongdoings that I may have committed. I also express my gratitude to Dr. Wong Nyet Kui, my examiner for giving me sound advice and opinions regarding my dissertation.

I would also like to take this chance to thank the staff at School of Science and Technology, Universiti Malaysia Sabah, for allowing me to use the computer lab to complete my dissertation, and also for allowing me access to the school's library to search for related dissertations for my reference.

Last but not least, I would like to thank all my coursemates, friends and family members who gave me support and gave their precious time to help me complete my research; I would like to assure them that their sacrifices will not be in vain. Thank you.

ABSTRAK

Melalui experiment ini, jujukan penuh primate endogenous retrovirus telah dijumpai dari Pangkalan Data Projek Genom Manusia. Nombor akses jujukan tersebut ialah gb|AC025420.26|, dan telah dipancing keluar oleh prob virus monyet Mason-Pfizer. Bit skor untuk gb|AC025420.26| ialah 211 dan Nilai E ialah $1e^{53}$. Endogenous retrovirus tersebut terdiri daripada LTR-gag-pol-env-LTR, dengan jumlah panjang 9289 pasangan bes. Panjang LTR ialah 801 bes, manakala panjang jujukan bergabung gag-pol-env ialah 7687 bes. Major homologous region di dalam bahagian gag mempunyai motif terpelihara $\varphi XQX_2EXF\varphi XRLX_3\varphi$, di mana φ ialah jujukan hidrofobik. Kotak Cysteine- Histidine yang juga dalam bahagian gag mempunyai motif terpelihara $CX_2CX_4HX_4C$, dan X ialah mana-mana asid amino. Beberapa motif terpelihara terdapat di bahagian pol, seperti motif transkripsi berbalik dengan jujukan W/CNTP, QDLR, KDAFF, LPQG, dan YXDD. Bahagian env mempunyai banyak jujukan asid amino dengan motif NXS/T, yang merupakan isyarat untuk glycosylation. Daripada pangkalan data Pfam, bahagian gag mempunyai gen yang kod untuk protein gag p10, gag p24, dan jejari zink. Skor bit protein tersebut ialah 68.5, 339.9,dan 34.1. E value protein tersebut ialah 9.3e-20, , 4.4e-99, dan 5e-07. Bahagian pol mempunyai gen yang kod untuk retroviral aspartyl protease, protein G patch, RnaseH, pengikat zink integrase domain, domain utama integrase, domain pengikat DNA integrase, dan transcriptase berbalik. Skor bit protein tersebut ialah 24.6, 72.9, 52.5, 61.8, 153.4, 65.6, and

160.7.. E value mereka ialah 2.6e-06, 2.5e-20, 2.2e-15, 2.4e-15, 6e-43, 1.6e-16, and 3.8e-45. Di bahagian env, terdapat gen untuk protein gp36 dengan skor bit 62.5 dan E value 1.4e-15. Bahagian LTR yang mengapit jujukan gag-pol-env mempunyai kesamaan sebanyak 99% di antara mereka. Semua keputusan telah disahkan dengan menggunakan beberapa program bioinfomatik, seperti BLAST, ORF Finder, BCM Search Launcher, Pfam, dan BLAST2seq.

ABSTRACT

Through this research, the full length of the primate endogenous retrovirus had been found from the Human Genome Project Database. The accession number for this sequence is gb|AC025420.26|, and is found using a Mason-Pfizer monkey Virus probe. The bit score of gb|AC025420.26| is 211 and the E-value is $1e^{53}$. The full length of the endogenous retrovirus comprises of LTR-*gag-pol-env*-LTR, with a total length of 9289 bp. The length of the LTR is 801 bp, while the combined sequence of *gag-pol-env* is 7687 bp long. The major homologous region in the *gag* region contains the conserved motif of $\varphi XQX_2EXF\varphi XRLX_3\varphi$, where where φ stands for hydrophobic sequences. The Cysteine- Histidine box conserved motif is also found in the *gag* region, with the sequence of $CX_2CX_4HX_4C$, where X could be any amino acid. Several conserved motifs had been found in the pol region, such as the reverse transcriptase conserved motif of W/CNTP, QDLR, KDAFF, LPQG, and YXDD. The env region contained a high number of amino acid sequences with the motif NXS/T, which is the signal for glycosylation. From the Pfam database, the *gag* region contains genes that code for gag p10 protein, gag p24 protein, and zinc knuckle. Their respective bit scores are 68.5, 339.9, and 34.1. Their respective E values are 9.3e-20, 4.4e-99, and 5e-07. *pol* region contains genes coding for retroviral aspartyl protease, G patch protein, RnaseH, integrase zinc binding domain, integrase core domain, integrase DNA binding domain, and reverse transcriptase.

Their respective bit scores are 24.6, 72.9, 52.5, 61.8, 153.4, 65.6, and 160.7. Their respective E values are 2.6e-06, 2.5e-20, 2.2e-15, 2.4e-15, 6e-43, 1.6e-16, and 3.8e-45. From Pfam results, there are genes that code for gp36 protein in the *env* region with bit score of 62.5 and E value of 1.4e-15. The flanking LTR region has a high similarity of 99% between them. All the results in this research had been confirmed using a variety of bioinformatics tools such as BLAST, ORF Finder, BCM Search Launcher, Pfam, and BLAST2seq.

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The Human Genome Project is a massive government subsidized-effort to map out the entire sequence of the human genome, which comprises of approximately 3.2 billion base pairs. This global effort was started in 1990 with a projected 15 year time scale and a $3 billion budget for completing the project, and the whole sequence of the human genome has since been mapped.

In the human genome, there is around 1% which comprises of the human endogenous retrovirus, which defective due to numerous deletions and mutations in the transcription process. This project deals with primate endogenous retroviruses found in the human genome. The primate endogenous retrovirus is most likely transmitted around 30 million years ago from prehistoric primates to humans when they have not diverged from the rhesus monkey linage.

Based on phylogenetic evidence, the analysis of the human genome has identified at least 30 HERV families. Identification of the endogenous retrovirus can be done by searching the particular genome for the *pol*, *gag* and *env* gene region of the retrovirus. Inside these genes are conserved motifs which are used to identify the location of the virus genome. One conserved motif is found in the *pol* gene, which codes for the enzyme reverse transcriptase. The motif is found in reverse transcriptase's gene, with the amino acid sequence CNTP, QDLR, KDAFF, FDF, LPQD, and YVDD.

Using this information, we can use bioinformatics to find out the similar sequence in the human genome in order to locate the endogenous retrovirus. There are several programs that we can use, such as the BLAST search, which searches for homologous sequences in an online database, the ORF finder, which gives the probable open reading frame, and the Pfam program, which is used to determine the protein coded by the amino acids. After using these programs for analysis of the human genome, we would expect to characterize the primate endogenous retrovirus, knowing the full sequence, the location in the chromosomes, and the protein it codes for.

The 2 main objectives of this project are to:

1) Identify the sequence of primate endogenous retrovirus in the human genome using BLAST search program

2) To obtain a full length sequence primate endogenous retrovirus in human genome

CHAPTER 2

LITERATURE REVIEW

2.1 Retrovirus

2.1.1 Retrovirus Classification

Retroviruses belong to the virus family of Retroviridae, which consists of RNA-containing viruses. These enveloped, single stranded RNA viruses use reverse transcriptase to complete their replication cycle, by reverse transcription of their viral genome into the host's DNA (Levy et al., 1994). The name retrovirus came for the Latin word for 'backward', *retro*, due to the fact that the viruses in this family contain the reverse transcriptase enzyme. One characteristic that sets it apart from other RNA viruses is that their RNA genome, once transcribed to DNA, can be incorporated into the host's cell genome (Levy et al., 1994).

Previously, three subfamilies were identified based on their morphological, biological and molecular features, but recent research had prompted the International Committee for Taxonomy of Viruses (ICTV) to subgroup the retroviruses into seven genera. Below is a table of the seven genera, and an example of the retrovirus in each genera.

Table 2.1 The seven genera of Retroviridae and examples of their species

Genus	Species	Host
alpharetrovirus	Avian leucosis virus	Chickens
betaretrovirus	Mouse mammary tumor virus	Mouse
gammaretrovirus	Murine leukemia virus, feline leukemia virus	Vertebrates, Cats
deltaretrovirus	Bovine leukemia virus, Human T-lymphotropic virus	Cattle, Humans
Epsilonretrovirus	Walleye dermal sarcoma virus	Fishes
Lentivirus	Human immunodeficiency virus 1, Simian immunodeficiency virus	Humans, Primates
spumavirus	Chimpanzee foamy virus	Chimpanzees

(source: http://en.wikipedia.org/wiki/Retroviridae)

Alpharetroviruses

This genera of virus is usually found in avian species, and consists of three different groups according to their characteristics (Van Regenmortel et al., 2000). The first group the Avian Leukosis-Sarcoma virus (ALSV), and is usually found in chickens. The second group is the Rous Sarcoma virus, and the third group consists of virus which is oncogenic and non-active. The virion consists of an envelope, a nucleocapsid, and a nucleoid. The virions measured 80 to 100nm in diameter, and the surface projections consists of small, densely dispersed spikes which covers the surface evenly. Of the numerous species in this genus, only one species has been successfully recovered in preparations (http://www.ncbi.nlm.nih.gov/ICTVdb/ICTVdB/).

Betaretrovirus

This genus of virus is mostly found in mice and primates, for example the mouse mammary tumor virus and Mason Pfizer monkey virus (http://www.wrongdiagnosis.com/medical/mason_pfizer_monkey_virus.htm). The monkey virus is a species of betaretrovirus which is found in the mammary carcinoma in rhesus monkeys, and several serologically distinct strains exist. It seems to have evolved from a recombination between a murine B oncovirus and a primate C oncovirus which is related to the baboon endogenous retrovirus. The Mason Pfizer monkey virus induces simian AIDS, and is one of the probes used in this project. The mouse mammary tumor virus (MMTV) is a milk transmitted retrovirus, and causes the majority of mammary

tumors in mice (http://en.wikipedia.org/wiki/Mouse_mammary_tumour_virus). Endogenous MMTV is expressed in several organs in addition to the lactating mammary gland in the mouse, and due to its slow process, it is known as a slow-transforming virus (http://www.steadyhealth.com/encyclopedia/MMTV).

Gammaretrovirus

The gammaretrovirus is usually found in mammals, including humans, but can also be found in avian and reptilian hosts. Some examples are the murine leukemia virus, feline leukemia virus, feline sarcoma virus, and avian reticuloendotheliosis virus. Some are endogenous retroviruses, and are thus present in the DNA of its hosts. The virus is split into two groups, one is a transforming virus, which is oncogenic, will cause diseases spontaneously or in a matter of days. The other group of virus does not undergo the replication process, is not oncogenic, and does not cause diseases during host infection (Kozak and Ruscetti, 1992).

Deltaretrovirus

This genus consists of exogenous horizontally transmitted viruses found in mammals (humans and primates), and is associated with B or T cells leukemia/lymphoma with a very long latency (http://www.wrongdiagnosis.com/medical/deltaretrovirus.htm). Some virus species found in it are bovine leukemia virus, human T-lymphotropic retrovirus (I, II, III, IV), and the simian T cell lymphotropic virus. The human and simian

T-lymphotropic viruses are used as probes in this project. The human T-lymphotropic virus (HTLV) is a single-stranded RNA retrovirus which causes T-cell leukemia and T-cell lymphoma in adults. There are four types of HTLV, and HTLV III and IV were discovered in 2006 in rural Cameroon, and were apparently transmitted from monkeys to hunters of the monkeys through their bites and scratches (Mahieux R, Gessain A 2005). HTLV III is similar to Simian T-lymphotropic virus III, which is also being used as a probe, but HTLV IV does not bears much similarity with any known virus, and not much is known on how much further transmission had happened among humans, or whether the virus can actually cause diseases.

Epsilonretrovirus

The viruses in this genus are viruses that have complex genome structure, due to the fact that there are additional genes in its viral genome and new tRNA primers. For example, Walleye dermal sarcoma virus (WDSV) has three extra genes: *orf-a*, *orf-b*, and *orf-c*. These genes are an addition to the existing viral genes, *gag*, *pol* and *env* genes.

Spumavirus

The spumavirus, from the Greek word *spuma*, meaning to foam, got its name from the vacuolar and soapsuds-like features found in cells infected with this genus of virus (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=rv.section.7999). Even though spumavirus have been associated with cytopathic effects in cell culture, this genus is not

linked to any diseases (Levy et.al., 1994). The spumavirus are exogenous and has prominent spikes on its surface. The virion contains significant amounts of double stranded full-length DNA, which is quite unusual in these viruses. Examples of this virus are the chimpanzee foamy virus, simian foamy virus and the human foamy virus. The spreading of this virus in the host is widespread throughout the body, for example simian spumavirus has been found in various tissue cultures of the brain, umbilical cord, kidneys and lungs (Hooks and Gibbs, 1975).

Lentivirus

The lentivirus, which got its name from the Latin word *lenti*, meaning slow, is a genus of slow viruses which has a long incubation period. It can deliver an impressive amount of genetic information into the DNA of the host cell, which makes it an excellent gene delivery vector. The HIV virus is an example of lentivirus. The lentivirus can be divided into five groups, one for each type of host it infects (primates, sheep and goats, horses, cats, and cattle). On the next page is a table containing the type of hosts and an example of the lentivirus.

REFERENCES

Altschul, Stephen F., Warren G., Webb M., Eugene W. M., and David J. L. 1990. *Basic Local Alignment Search Tool.* J. Mol. Biol. 215:403-10.

Aravind L., Koonin E.V. 1999. *G-Patch: A New Conserved Domain In Eukaryotic RNA-Processing Proteins and Type D Retroviral Polyproteins*, Trends Biochem Sci, 24: 342-344

Cai M., Zheng R., Caffrey M., Craigie R., Clore G.M., Gronenborn A.M. 1997. *Solution Structure Of The N-Terminal Zinc Binding Domain Of HIV-1 Integrase*, Nat Struct Biol, 4: 567-577

Cress A., Triezenberg S.J. 1991. *Nucleotide and deduced amino acid sequences of the gene encoding virion protein 16 of herpes simplex virus type 2*, Gene, 103: 235-238

Fossen T., Wray V., Bruns K., Rachmat J., Henklein P., Tessmer U., Maczurek A., Klinger P., Schubert U. 2005. *Solution Structure Of The Human Immunodeficiency Virus Type 1 p6 Protein*, J Biol Chem, 280: 42515-42527

Hazura,H.J. 2006. *Computer-Based Characterization of ERV-9 Human Endogenous Retrovirus from Human Genome Project.* Science and Technology Faculty, Universiti Malaysia Sabah.

Hooks, J.J., Gibbs, C. J. Jr. 1975.*The foamy viruses.Microbiol. Mol. Biol. Rev.* 39: 169-185

Kambol, R. 2003. *Distribution And Evolution Of Endogenous Retrovirus Within Amphibian And Piscine Hosts*, Unpublished PhD, Thesis, Department Of Biological Science, Imperial College, London, United Kingdom

Kozak, C. A., Ruscetti, S. 1992. *The Retroviridae,* ed. Levy, J. A, Plenum, New York, Vol. 1, pp. 405-481.

Levy J.A., Conrat F.H., Owens R.A. 1994. *Virology 3rd edition.* Prentice Hall, New Jersey

Mahieux R, Gessain A. 2005. *New human retroviruses: HTLV-3 and HTLV-4.* Med Trop (Mars) 65 (6): 525-8. PMID 16555510.

Patarca, R., Haseltine, W. A. 1983. *A Major Retroviral Core Protein Related To EPA and TIMP*, Nature. Vol. 318.390.

Pevsner, J. 2003. *Bioinformatics and Functional Genomics.* Wiley –Liss, United States of America.

Swanstrom, R., Wills, J. W., Eds. J. M. Coffin, S. H. Huges., Varmus, H. E. 1997. *Synthesis, Assembly and Processing of Viral Proteins In Retrovirus*. Cold Spring Harbour Laboratory Press, New York.

Van Regenmortel, M.H.V., Fauquet, C.M., Bishop, D.H.L., Carsten, E.B., Estes, M.K., Lemon, S.M., S.M., Maniloff, J., Mayo, M.A., McGeoch, D.J., Pringle, C. R., Wickner, R. B. 2000. *Virus Taxonomy: The Classification And Nomenclature of Viruses*. The Seventh Report of the International Committee on Taxanomy of Viruses. Academic Press, San Diego.

Webster, T. A., Patarca, R., Lathrop, R., Smith, T.F. 1989. *Potential Structure Motifs For Reverse Transcriptase.* Mol. Biol. Evol.

Wehrel K., Chesebro B. 1997. p24 antigen capture assay for quantification of human immunodeficiency virus using readily available inexpensive reagents, Methods, 12:288-293.

Wills, J.W., Craven, R.C. 1991. *Form, Function, And Uses of Retroviral Gag Protein.* Aids.

van der Kuyl A.C., Mang R., Dekker J.T., Goudsmit J. 1997. *Complete nucleotide sequence of simian endogenous type D retrovirus with intact genome organization:*

evidence for ancestry to simian retrovirus and baboon endogenous virus, J. Virol, 71 (5) : 3666-76