The Effects of Human Immunodeficiency Virus (HIV) Infection on Host DNA Methylation and Plasma Microbiome

LI, Sai Kam

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy
in
Biochemistry (Medicine)

The Chinese University of Hong Kong
September 2011

Thesis/ Assessment Committee

Professor Mary M.Y. Waye (Chair)

Professor Stephen K.W. Tsui (Thesis Supervisor)

Professor T.F. Chan (Committee Member)

Professor Patrick C.Y. Woo (External Examiner)

Abstract

Abstract of thesis entitled:

The Effects of Human Immunodeficiency Virus (HIV) Infection on Host DNA Methylation and Plasma Microbiome

Submitted by **LI, Sai Kam**

for the degree of **Doctor of Philosophy in Biochemistry (Medicine)**

at The Chinese University of Hong Kong in Sept, 2011.

In English

The global dissemination of human immunodeficiency virus (HIV) has been a continuing health threat to mankind after 30 years since its discovery. At the same time, the next-generation sequencing has transformed today's research. In this study, we have made use of the next-generation high-throughput sequencing technology to study two aspects of HIV-1 induced biological changes. The HIV-associated DNA methylation was elucidated by comparing the DNA methylation pattern of T-lymphocytes between HIV-infected and uninfected identical twins. We found that genes regulating RNA splicing and cell cycle regulation were hypo-methylated whereas the genes regulating nerve functions and signal transduction were hyper-methylated. The molecular study suggested that HIV-1 increases the transcription level of DNA methyltransferase (*DNMT*) *3A* but not the *DNMT1* or *DNMT3B*. The up-regulation of *DNMT3A* was not associated with the methylation of HIV-1 promoter LTR and latency in chronically infected cells with actively replicating viruses. In studying the more advanced HIV disease with AIDS and opportunistic infections, HIV/AIDS patients with very low CD4+ T-cell counts were recruited.

We have profiled the plasma microbiome and virome of these AIDS patients in parallel with healthy adults. HIV/AIDS plasma microbiome was dominated by bacteria from the order Pseudomonadales while healthy control subjects carried few bacterial DNA in the blood. We have found that many of the microbes in HIV/AIDS plasma are similar to some of the microbes found in the human gut. The HIV/AIDS and normal plasma virome share some similarity in the presence of common ubiquitous eukaryotic viruses. The normal virome was mainly composed of viruses from Anelloviridae. The HIV/AIDS viromes was contrasted by the presence of a large proportion of bacteriophages, typical eukaryotic viruses and untypical non-human viruses. In addition, by means of sequencing, we have found several sequences which might belong to novel bacteria or endogenous retroviruses. The results described in this thesis manifest the use of the high-throughput technology in studying cellular genome and microbial metagenomics. The insights gained into the HIV-associated DNA methylation and the specific spectrum of microbes found in these patients may facilitate future research studies in combating HIV/AIDS.

摘要

Abstract in Chinese of thesis entitled:

The Effects of Human Immunodeficiency Virus (HIV) Infection on Host DNA Methylation and Plasma Microbiome

Submitted by **LI, Sai Kam**

for the degree of **Doctor of Philosophy in Biochemistry (Medicine)**

at The Chinese University of Hong Kong in Sept, 2011.

In Chinese

人類免疫缺陷病毒（HIV）自三十年前爲人所發現後，至今仍一直威脅人類健康。與此同時，新一代基因測序方法正逐步轉化今天的科學研究。在這項研究中，我們利用了這種新一代測序方法來研究兩個方面有關 HIV-1 誘導的生物學變化。我們透過比較 HIV 感染和未感染的同卵雙胞胎，找出 T-淋巴細胞內與 HIV 相關的甲基化基因。我們發現在受 HIV 感染情況下，調控 RNA 剪接與細胞週期調控的基因甲基化的程度減少；相對地，負責調節神經功能和信號轉導的基因甲基化程度增加。分子研究表明，HIV-1 提高 DNA 甲基轉移酶 3A 的轉錄水平，但沒有影響轉移酶 1(*DNMT1*)或轉移酶 3B (*DNMT3B*)的轉錄水平。該上調的 *DNMT3A* 轉錄水平，與在包含活躍複製的 HIV -1 的慢性感染細胞中的病毒啓動子 LTR 的甲基化，及病毒的潛伏沒有關聯。爲了解更後期的感染病變，特別是針對艾滋病者的機會性感染，我們對 CD4+淋巴細胞計數很低的艾滋病患者進行了研究。我們分析了這些艾滋病患者和健康成人血漿裏的微生物群系和病毒群系。艾滋病患者血漿裏的微生物群系含大量來自 Pseudomonadale 次的細菌，而健康成人的血則漿裏只有少量的細菌 DNA。我們發現這個血漿微生物群系與已公佈的人類腸道微生物組有不少相似之處。另一方面，我們發現艾滋病者和正常血漿病

毒群系有一些相似之處，包括一些普通的人類病毒。正常的血漿病毒群系主要由 Anelloviridae 病毒組成。艾滋病者的血漿病毒群包含大比例的噬菌體、典型的真核細胞病毒和非人類病毒。此外，通過鑽研測序的細節，我們發現了一些可能屬於新的細菌和內源性逆轉錄病毒 K 的序列。本論文中描述的研究結果，體現了高通量基因測序技術在研究細胞的基因組及微生物宏基因組學的用途。這些與艾滋病毒相關的 DNA 甲基化和微生物群系的研究結果，將有利於未來對於艾滋病毒/艾滋病防治的研究。

Acknowledgement

I am deeply grateful to be given the opportunity to be the student of Professor Tsui, one the few professors I adore so much. He has been my teacher of life, Christian faith and research since 2002 when we met each other in the cell group of the Department of Biochemistry. Without his supervision, friendship, trust, encouragement and help, I will have to spend many more years to finish this PhD. study. I will never forget such a wonderful supervisor in my research life.

A very special recognition needs to be given to another teacher of life--Professor C.K. Lau. By working in the same laboratory day-to-day for two years, Professor Lau has helped me achieve several additional insights into research and life. He is a very patient and compassionate person who cares for every of his team members. His extensive support and encouragement during my research is indispensable. I especially acknowledge his help in the part of RNA splicing gene analysis.

I would like to acknowledge and extend my heartfelt gratitude to one the internal examiners Professor M.Y. Waye who was also my supervisor during the M.Phil study. Her enthusiasm and perseverance has shaped my attitude in research (and in flute music). I would also like to thank Professor T.F. Chan for sharing his experience in Bioinformatics in the part of metagenome analysis. He is always so friendly and cares a lot of my research progress. He is much like an adviser more than an examiner. I would like to warmly acknowledge Professor C.Y. Woo for being my external examiner. I read several of his papers on 16S sequencing but did not imagine such a great microbiologist would one day be my examiner. I am so grateful that as a Biochemistry graduate, I could have the chance to meet the professionals from Microbial genomics and proteomics (Professor Tsui and Professor Lau), Molecular neurology (Professor Waye), Bioinformatics (Professor Chan) and Microbiology (Professor Woo). I would also like to extend my thanks to Professor W.P. Fong (SLS, CUHK) and Professor C.C. Wan (SBS, CUHK) for sharing the cell lines. And thank Dr. M. Peterlin for sharing the plasmid.

Of the many coworkers who have been enormously helpful in my research, I especially thank our post-doctoral fellow Dr. Patrick K.S. Ng for his unconditional support in my experiments. He has spent a lot of time with me in the BSL3 facility working on the 'dangerous' virus culture work. We have been working closely for the laboratory management in the past three years and have together faced lots of trouble. I thank him for every discussion (research or non-research) in which he exudes his intelligence and wisdom. I would also like to thank Dr. Wayne Zhou and Dr. Candy Li for their help in some of the experiments. I would never forget the help from the people in the Hong Kong Bioinformatics Centre. They are Mr. K.K. Leung, Mr. K.T. Kwong, Mr. S.K. Lou, Mr. Y. Yang and Mr. Allen Yue. These projects on HIV will not be possible without the help from our collaborators Dr. J.H. Wang from Institut Pasteur of Shanghai, Dr. Chiyi Zhang from Jiangsu University, Dr. J.F. Wei from First Affiliated Hospital of Nanjing Medical University, Dr. H.X. Guo from Jiangsu Provincial Center for Disease Prevention and Control and Professor S.S. Lee from Department of Microbiology, CUHK. I am grateful to have met all these great virologists who work hard for years to combat the HIV/AIDS disease. And thanks to Professor Dennis Lo and Professor Rossa Chiu (Chemical Pathology, CUHK) for their collaboration.

Most special thanks to my family, Mr. Matthew Ma and his family. Words alone cannot express what I owe them for their encouragement and patient love which enabled me to complete this PhD. study. I thank my mother for her unchanging love, care and support. She did not receive a lot of education in schools and cannot read any English word written here. But I know she understands my passion in research better than anyone else in this world.

I thank God for giving me a loving family, caring teachers, trustworthy collaborators, helpful coworkers and supportive friends. May God make use of our effort to save more people's lives both bodily and spiritually, and to demonstrate His greatness.

Table of Content

Chapter 2 p. 26-81

Analysis of HIV-1 associated DNA cytosine methylation in HIV carriers

Abbreviations

Abbreviations	Full name
a.a.	amino acid(s)
ADC	AIDS dementia complex
AIDS	acquired immunodeficiency syndrome
AKAP17A	A kinase (PRKA) anchor protein 17, SR protein
ARHGEF3	Rho guanine nucleotide exchange factor (GEF) 3
BLAST	Basic Local Alignment Search Tool
BLASTn	nucleotide BLAST
bp	base pair(s)
BSL	biosafety level
BT	Bluetongue
CD	cluster of differentiation
cDNA	complementary DNA
CNS	central nervous system
CpG	cytosine-phosphate-guanosine
CRF	circulating recombinant forms
Ct	threshold cycle
CUHK	The Chinese University of Hong Kong
Da	Dalton
DAMP	damage-associated molecular pattern
DAVID	Database for Annotation, Visualization and Integrated Discovery
DDX1	DEAD (Asp-Glu-Ala-Asp) box polypeptide 1
DMEM	Dulbecco's modified Eagle's medium
DMPK	Dystrophia myotonica-protein kinase
DNA	deoxyribonucleic acid
DNMT	DNA methyltransferase
dNTP	deoxyribonucleoside triphosphate
ELISA	enzyme-linked immunosorbent assay
env	envelope
G3PDH	glyceraldehyde-3-phosphate dehydrogenase
Gemin8	gem (nuclear organelle) associated protein 8
GI	gastrointestinal
GO	gene ontology
GR	glucocorticoid receptor
GRE	glucocorticoid response element
HAART	highly active antiretroviral therapy
HBV	hepatitis B virus
HCV	hepatitis C virus
HD	Huntington's disease
HDF	host dependency factor

HERV	human endogenous retrovirus
HIV	human immunodeficiency virus
HIVE	HIV-1 encephalitis
HIV-PI	HIV-protease inhibitors
HK	Hong Kong
HMGB	high mobility group box
HMP	human Microbiome Project
HPV	human papillomavirus
HRP	horse radish peroxidase
HTT	huntingtin
IP	immunoprecipitation
KAT	lysine acetyltransfease
kb	kilo base pairs
KDAC	lysine deacetylase
KMT	lysine methyltransferase
LTR	long terminal repeat
MAGOHB	Mago-nashi homolog B (Drosophila), Exon junction complex
MBD5	methyl-CpG binding domain protein 5
MeDIP	methylated DNA immunoprecipitation
MHC	major histocompatibility complex
miRNA	micro RNA
MND1	meiotic nuclear divisions 1 homolog (*S. cerevisiae*)
mRNA	messenger RNA
MYCBP2	MYC binding protein 2
MZ	monozygotic
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
NDUFA2	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 2, 8kDa
nef	negative regulatory factor
NIH	National Institute of Health
NNRTI	non-nucleoside reverse-transcriptase (RT) inhibitors
NRTI	nucleoside reverse-transcriptase (RT) inhibitors
ORF	open reading frame
PAMP	pathogen-associated molecular pattern
PBMC	peripheral blood mononuclear cell
PCR	polymerase chain reaction
PMMV	Pepper Mild Mottle Virus
pol	polymerase
rev	regulator of expression of virion protein
RNA	ribonucleic acid
RNP	ribonucleoproteins
RT	reverse transcription
SDS	sodium dodecyl sulphate
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis

SFPQ	splicing factor proline/glutamine-rich binding protein
SIRS	systemic inflammation response syndrome
SKA3	spindle and kinetochore associated complex subunit 3
SLE	systemic lupus erythematosus
snRNA	small nuclear RNA
SNRNP48	small nuclear ribonucleoprotein 48kDa (U11/U12)
SNRPA1	small nuclear ribonucleoprotein polypeptide A'
SPC25	NDC80 kinetochore complex component, homolog (*S. cerevisiae*)
SR	arginine/serine rich
tat	transactivator of transcription
TBP	TATA-binding protein
TF	transcription factor
TIMELESS	timeless homolog (Drosophila)
TSS	transcriptional start site
TTMDV	Torque teno midi virus
TTV	Torque teno virus
TXNL4A	spliceosomal U5 snRNP-specific
UNAIDS	United Nations Program on HIV/AIDS
URF	unique recombinant forms
UsnRNP	uridine-rich small nuclear ribonucleoproteins
UV	ultra violet
vif	virion infectivity factor
vpr	viral protein R
vpu	viral protein U
WHO	World Health Organization
WSSV	white spot syndrome virus

List of Figures

List of Tables

Chapter 1

Introduction

1.1 HIV/AIDS

1.1.1 Global and local statistics of HIV/AIDS pandemic

The global dissemination of human immunodeficiency virus (HIV) has been a threat to mankind for more than 30 years. The HIV-mediated immunodeficiency and neurological diseases were first recognized and described as acquired immunodeficiency syndrome (AIDS) in 1981 by the US Centre for Disease Control and Prevention, when a group of man developed a common feature of T-cell dysfunction. This then led to the discovery of HIV in 1983 (Barre-Sinoussi *et al.*, 1983; Gallo *et al.*, 1984; Montagnier, 1985). Ever since its discovery, the virus has been spreading over the globe and has claimed millions of lives. The joint United Nations Program on HIV/AIDS (UNAIDS/ WHO) estimates that in 2009 there were 33.4 million (31.4 million – 35.3 million) individuals living with HIV/AIDS (UNAIDS/WHO, 2010). About 1.8 million [1.6 million – 2.1 million] HIV/AIDS patients died of the diseases in 2009. The life-threatening HIV/AIDS pandemic continues to expand globally at a rate of 2.6 million [2.3 million – 2.8 million] annually (more than 7000 new cases per day).

Figure 1.1 shows the latest HIV/AIDS pandemic information provided by UNAIDS. Sub-Saharan Africa continues to be the epicenter of the pandemic with 22.5 million [20.9 million -- 24.2 million] HIV carrying adults and children in the

year 2009 and 1.8 million [1.6 million -- 2.0 million] new infections per year. Not surprisingly, Asia is becoming the second epicenter with a total of ~ 4.8 million carriers and ~ 350 thousand annual new infections.

Number of people living with HIV/AIDS in 2009	33.3 (31.4 – 35.3) million
New infection during 2009	2.6 (2.3 – 2.8) million
Total AIDS death in 2009	1.8 (1.6 – 2.1) million

Figure 1.1 Distribution and estimated number of adults and children infected with HIV in different geographical regions of the world in 2009. The estimated total number of people living with HIV/AIDS worldwide in 2009 was about 33.3 million. The number of new infection and number of AIDS death were shown in the lower boxes. The Arabic numbers in the circles indicates the prevalence of HIV with "1" being the most prevalent area.

There are currently an estimated 740,000 people living with HIV in China. During 2009 around 26,000 people died from AIDS. By considering the large population of 1.3 billion in China, the prevalence was regarded as low (0.1% among adults). Yet, the infection rate is alarmingly high in certain provinces such as Yunnan, Guangxi, Sichuan, Xinjiang, Guangdong and Henan (Figure 1.2) (Ministry of Health, 2010). They represent around 70 to 80 percent of the national infection in total. In 2009, China reported that AIDS had become the country's leading cause of death among infectious diseases for the first time ever, surpassing both tuberculosis and rabies (AVERT, 2010).

Figure 1.2 Distribution and estimated number of people infected with HIV in different provinces in China in 2009 (modified from Annual report 2009 from the Ministry of Health, People's Republic of China, Joint United Nations Programme on HIV/AIDS World Health Organization). Highest prevalence of HIV was found in Yunnan and Guangxi followed by Sichuan, Xinjiang, Guangdong and Henan.

In Hong Kong, there were 4,730 HIV carriers and among them 1,171 were AIDS patients according to the statistics in September 2010 (Virtual AIDS Office of Hong Kong, 2010). The details of the statistics are shown in Table 1.1.

Table 1.1 The current HIV/AIDS situation in Hong Kong.

		Cumulative Sept 2010	
Sex		**HIV**	**AIDS**
	Male	3,789	999
	Female	941	172
Ethnicity			
	Chinese	3,143	910
	Non-Chinese	1,587	261
Route of transmission			
	Heterosexual contacts	2,082	714
	Homosexual contacts	1,249	250
	Bisexual contacts	193	47
	Injecting drug use	291	46
	Blood / blood product recipients	79	24
	Perinatal	25	7
	Undetermined	811	83
Total		**4,730**	**1,171**

1.1.2 Genotype classification of HIV

AIDS etiologies can be attributed to two related human lentiviruses, namely HIV type 1 (HIV-1) and HIV type 2 (HIV-2). HIV-1 is distributed worldwide and accounts for the majority of HIV infections. In contrast HIV-2 is confined to Africa, India and Korea and only subtype A and B are disseminated into significant numbers of human population (Grez *et al.*, 1994; Nam *et al.*, 2006; Pfutzner *et al.*, 1992). It is known that sexual and perinatal transmission of HIV-2 is less efficient than HIV-1 (Kanki *et al.*, 1994). HIV-1 can be classified into three distinct types, M (Main) group, O (Outlier) group and N (New, or non-M/non-O) group, which represent separate introductions into human from chimpanzees (Ayouba *et al.*, 2000;

Loussert-Ajaka *et al.*, 1995; Robertson, 1999; Vanden Haesevelde *et al.*, 1994) (Figure 1.3).

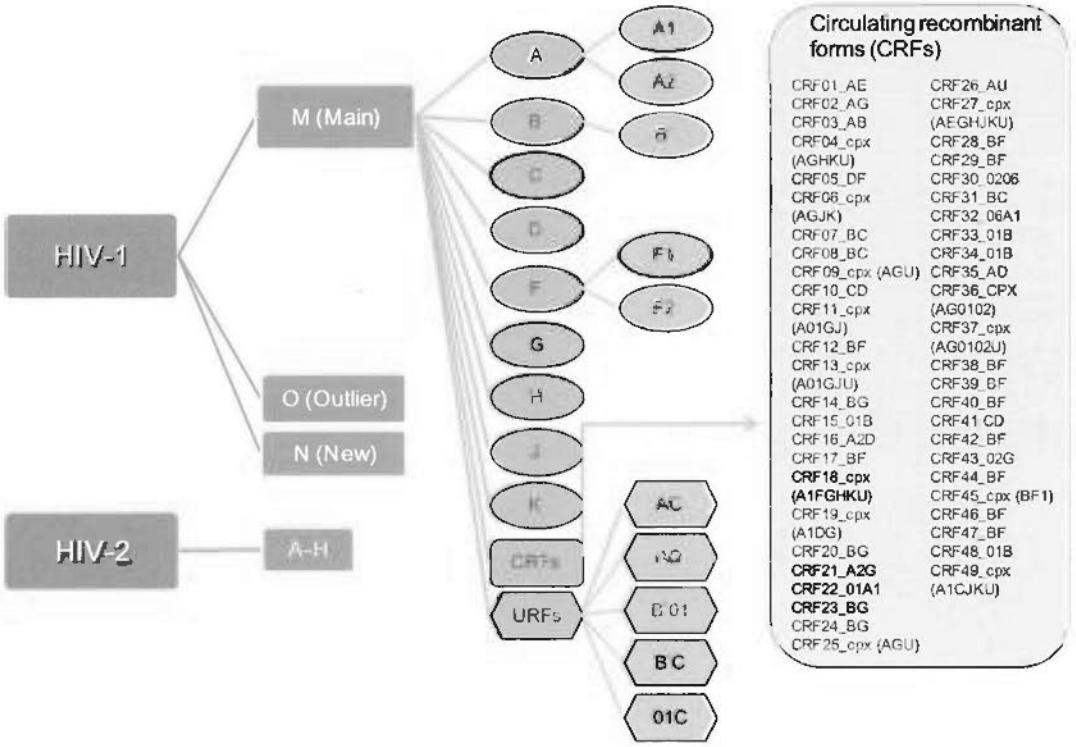


Figure 1.3 Classification of HIV genotypes and their origins (drawn bases on the information updated in Feb 2011). HIVs are classified into types (types 1 and 2), groups (M, O and N for HIV-1; A to H for HIV-2), subtypes (A to K for HIV-1) and subsubtypes (A1 and A2, F1 and F2, B'). HIV-1 recombinants are categorized into circulating recombinant forms (CRFs) and unique recombinant forms (URFs) by their magnitude of dissemination. Up to Feb 2011, 49 CRFs were described. The alphabets show the subtypes constituting the CRFs.

The M (Main) group can be further classified into 9 genetic subtypes, namely A, B, C, D, F, G, H, J and K with about 25-35% sequence variation among subtypes (Robertson, 1999). In addition to the discrete subtypes, the M group also comprises >40 circulating recombinant forms (CRFs) and at least 6 unique recombinant forms (URFs). Circulating recombinant forms (CRFs) represents the HIV-1 strains clustered with genome from different subtypes. The database of CRFs (http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html) is updated frequently from time to time due to the practical reason that new CRFs are generated rapidly. By February 2011, there are in total 49 CRFs of HIV-1 (NCBI, 2011b). Unique recombinant forms (URFs) are diverse forms of HIV-1 inter-subtype recombinants with unique mosaic structures, found only in a single person or in a few individuals (McCutchan *et al.*, 2005). The emergence of CRFs and URFs indicates that multiple infections with different lineages of HIV-1 strains in the same person at the same time are not rare events. The likelihood of generating new recombinant viruses will continue to increase because of the rapid human migration or travelling. Group Outlier (O) comprises a pool of highly divergent, genetically related strains which are limited to people living in central Africa (Loussert-Ajaka *et al.*, 1995). Only a few cases of group N infections are identified in a limited number of patients from Cameroon (Simon *et al.*, 1998).

The classification of these subtypes is based on the complete viral genome composition. In practice, classification of any HIV-1 subtypes that might be new could be done on the database available from the list on National Center for Biotechnology Information (NCBI) Tools for Bioinformatics Research – the Viral Genotyping tool (http://www.ncbi.nlm.nih.gov/projects/genotyping/formpage.cgi)

(NCBI, 2011a). It works by comparing the submitted sequences to the reference HIV-1 genome of different subtypes.

1.1.3 Emergence of unique clusters in China

HIV-1 subtype B is the most common subtype that infects people in China including Hong Kong. However, the real situation can often be complicated in cities like Hong Kong where people from all over the world come and go. HIV was therefore carried from a city to another by the frequent human traffic. This might be the reason that the HIV subtype B found in Hong Kong and nearby regions has been changing rapidly genetically. It was previously found by our group that the prevalence of subtype BF was not uncommon in Macau, a previous colony of Portugal, which might be attributed to be immigration of the Portugueses (Leung *et al.*, 2010). Although the infecting HIV-1 in Hong Kong was broadly described as subtype B, different new clusters of subtype B were continuously identified. In a previous study of the full-length genome sequencing and characterization of three HIV-1 isolates selected from one of the emerging clusters of subtype B, we have identified several subtype-D related mutations in the *vif* gene which could disrupt its interplay with the host anti-viral protein (Leung *et al.*, 2008; Tsui *et al.*, 2010). This further indicates that HIV-1 is not a static molecule that can be easily targeted and eliminated.

1.2 Human immunodeficiency virus (HIV)

1.2.1 Virus structure and genome organization

HIV-1 belongs to the Lentivirus subfamily of the family Retroviridae. Figure 1.4 shows the structure of a HIV-1 virion. The virus particle is a sphere of

approximately 100 nm in diameter (Wang *et al.*, 2000). Each virion is composed of two molecules of single stranded RNA surrounded by the gag gene product: the p17 outer matrix protein, the p24 major capsid protein, which forms the capsid shell; and the p7 nucleoprotein, which binds tightly to the viral RNA. Many other viral proteins are incorporated with the virion including the protease, reverse transcriptase and integrase.

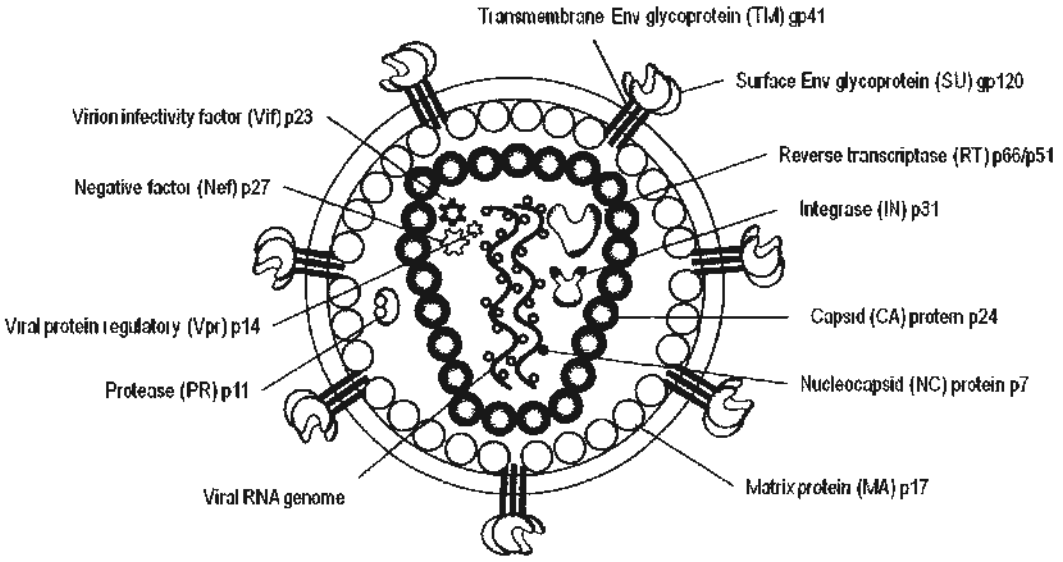


Figure 1.4 Structure of a mature HIV-1 virion depicting the key viral proteins and their arrangement within the virion. The viral envelope is formed from the host cell membrane, into which the HIV-1 envelope proteins gp41 and gp120 penetrate. The matrix between the envelope and the core is formed predominantly from Gag protein p17. The major structural proteins of the core are Gag proteins p24 and p6 (not shown). The core contains the viral RNA genome, closely associated with nucleocapsid protein p7, in addition to RT and IN. Viral accessory proteins Vif, Nef and Vpr are also packaged within the virion and are thought to localize within the core. The viral protease is also incorporated in the mature virion.

The HIV-1 proviral DNA integrated into the host genome is 9.7 kilobases (kb) in length. The basic genome organization of HIV genome was shown in Figure 1.5. As illustrated, the provirus is symmetrically flanked at either end by the viral long terminal repeat (LTR). These LTRs contains transcriptional regulatory sequences, RNA processing signals, packaging sites and the integration sites. The 5' end begins with *gag* gene which encodes for the core and matrix protein; the *pol* gene, which begins in an overlapping frame encoding viral protease, reverse transcriptase and integrase; and the *env* gene, which encodes for the outer and transmembrane envelope protein. In addition to these structure proteins, HIV has six accessory genes: *vif*, *vpr*, *tat*, *rev*, *vpu* and *nef* (Wang *et al* , 2000). The detailed gene products and their major functions are given in Table 1.2.

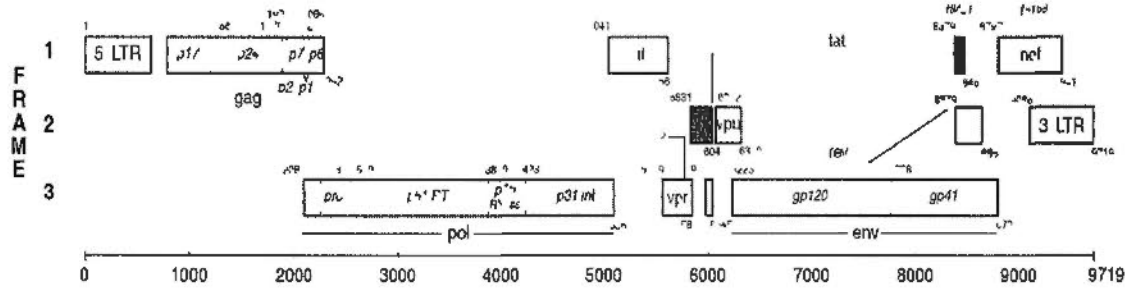


Figure 1.5 Schematic diagram of HIV-1 genome, based on the HXB2 strain (adapted from NCBI GenBank accession NC_001802). The open reading frames are shown as rectangles with the encoded proteins indicated.

Table 1.2 Genes and gene products of HIV-1.

Gene	Proteins	Functions
gag	p24	Core protein
	p17	Matrix protein, interacts with gp41
	p7	Nucleocapsid, binds viral RNA
	p6	Core protein, binds to Vpr
	p2	Cleavage products of Gag precusor protein
	p1	Cleavage products of Gag precusor protein
pol	Protease	Proteolytic cleavage of Gag and Pol
	Reverse transcriptase	Polymerase and RNase H activity (p66 subunit only)
	Integrase	Integration of proviral DNA into host chromosome
env	gp120	Envelope protein for viral entry into host cell
	gp41	Transmembrane protein for cell fusion
vif	Virion infectivity factor	Efficient cell-free transmission
vpr	Viral protein R	Enhances viral replication, cell G2/M phase arrest
tat	Transactivator of transcription	Major viral transactivator
rev	Regulator of expression of virion protein	Exports unspliced and singly spliced viral RNAs
vpu	Viral protein U	Enhances virion release from cells, downregualtes CD4
nef	Negative regulatory factor	Inhibits or enhances viral replication depending on strain and cell type; down-regulates CD4

1.2.2 Virus replication cycle

1.2.2.1 Viral entry and integration

HIV begins its life cycle by binding to the cell surface-associated receptor CD4 and co-receptors using its envelope glycoprotein gp120 (Chan and Kim, 1998) (Figure 1.6). This interaction, which is mediated by the HIV envelope (Env) protein gp120, leads to subsequent interaction of the gp120 V3 loop region with a co-receptor, usually a member of the seven-membrane-spanning chemokine-receptor families (Cocchi *et al.*, 1996; Monini *et al.*, 2004). Two distinct predominant chemokine co-receptors used by different HIV-1 are known to be the α-chemokine receptor CXCR4 (X4 virus) on T-lymphocytes and dendritic cells, and β-chemokine receptor CCR5 (R5 virus) on macrophage (Alkhatib, 2009). After virus adsorption,

the viral and cell membranes fuse together, and the viral 'core' (which includes the diploid viral genome) is released into the cytoplasm. The viral genomic positive RNA is converted into negative DNA by the viral reverse transcriptase (Herschhorn and Hizi, 2010). Human transfer RNAs specific for lysine ($tRNA^{lys}$) bind to the viral promoter LTR as the primers at this step. After the synthesis of the single strand DNA, the human $tRNA^{lys}$ binds to the 3' end LTR and starts synthesizing the second strand DNA. As this process continues, a double-stranded viral DNA is formed. This proviral DNA can then be integrated into the cellular genome particularly at gene-rich regions, mediated by the viral integrase. Apparently, the viral integration is affected by several cellular factors such as the status of the target T-lymphocytes. It is shown that the viral DNA can only integrate into the genome of activated T-cells but not the resting ones (Chan and Kim, 1998; Wang *et al.*, 2000).

1.2.2.2 Viral transcription and package

After genome integration, the HIV-1 provirus may either remain dormant or become transcriptionally active. The transcription of the provirus is mediated by the cellular RNA polymerase II which produces a polycistronic pre-mRNA of 9 kb containing more than 40 different splicing sites. This long pre-mRNA can be the unspliced genomic RNAs (9.0 kb) or be singly spliced to form 4 kb mRNA or multiply spliced to form 2.2 kb mRNAs. The unspliced and singly spliced RNAs accumulate in the nucleus before the viral RNA export channel is formed by the viral protein Rev (Monini *et al.*, 2004). In contrast, the multiply spliced 2.2 kb mRNA can be exported into the cytoplasm by the cellular exportin-1 (CRM-1) (Nakielny and Dreyfuss, 1999; Stade *et al.*, 1997). This small mRNA is further spliced for the

mRNA of the early essential viral proteins tat, rev and nef. As Rev proteins accumulate, nuclear export of the unspliced and singly mRNA is facilitated (Kim *et al.*, 1989; Klotman *et al.*, 1991). These mRNA are then translated to form vif, vpr, vpu, env, gag and gag-pol polyproteins. The new viral particle is formed by the assembly of the structural proteins and the genomic RNA. The newly synthesized viral particles then leave the cells by budding motion. Three typical classes of anti-HIV drugs have been used to disturb this highly-orchestrated viral life cycle. They are nucleoside and non-nucleoside reverse-transcriptase (RT) inhibitors (NRTIs and NNRTIs, respectively), and HIV-protease inhibitors (HIV-PIs) (Monini *et al.*, 2004).

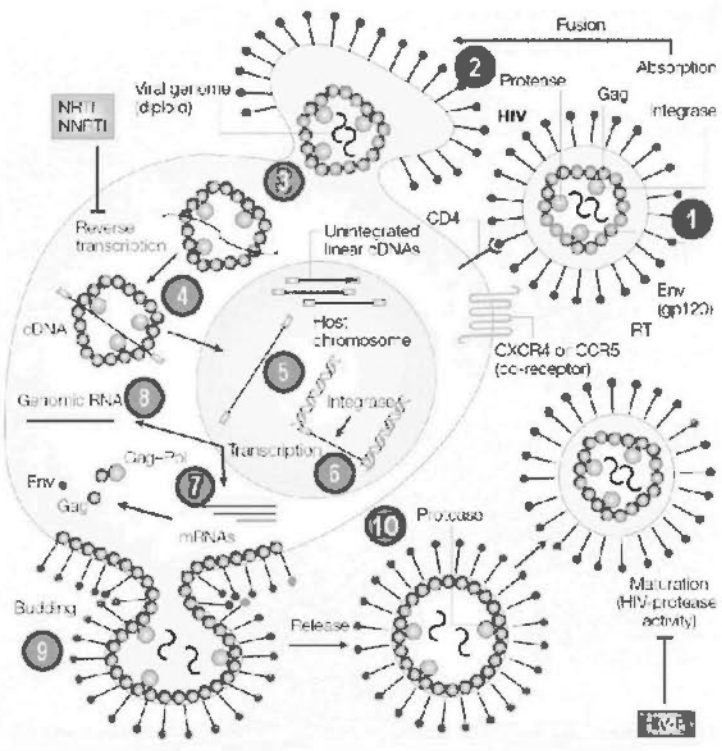


Figure 1.6 The life cycle of HIV-1, modified from (Monini *et al.*, 2004). The life cycle begins with the binding of the virus to CD4 receptor and CXCR4 or CCR5 co-receptor on target cell surface (1), followed by the fusion of the viral envelope with the cell membrane (2) and the release of the viral core into the cytoplasm (3). The viral RNA genome is then reverse transcribed to DNA (4) which is then integrated into the host cell genome to form the provirus (5). Upon the stimulation by the appropriate viral and cellular transcription initiation complex, the viral LTR is transcriptionally active and viral mRNAs are therefore generated (6). These mRNAs are spliced to form the structure protein as well as accessory proteins or remained unspliced as the genomic RNAs (7, 8). New virions are formed by incorporating the mature viral proteins and the newly synthesized viral RNA genome (9). The viruses then leave the cell by budding (10). These crucial steps of the HIV life cycle are repressed by nucleoside and non-nucleoside reverse-transcriptase (RT) inhibitors (NRTIs and NNRTIs, respectively), and with HIV-protease inhibitors (HIV-PIs).

1.3 Host-HIV interactions

1.3.1 HIV and the human methylome

1.3.1.1 Introduction to human methylome

On top of the four genetic letters A, T, G, C, there is an additional genetic letter for making up the genetic code of life. This is 5-methylcytosine. This new genetic family member is estimated to account for 1-6 % of the nucleotides in mammals and plants (Lister and Ecker, 2009; Montero *et al* , 1992). The presence of such 5-methylcytosine is usually referred as cytosine methylation or DNA methylation which impacts an additional layer of inheritable genetic information upon DNA code. Methylated DNA sequence are less expressed by inducing a repressive chromatin structure that is are not recognized by transcription factors (Jaenisch and Bird, 2003). Such silencing effect is usually in proportion to the degree of methylation (Siedlecki and Zielenkiewicz, 2006). In mammals, DNA methylation occurs almost exclusively at the fifth carbon of cytosine residues found within the cytosine-phosphate-guanosine (CpG) dinucleotides. There are 56 million CpGs in most human cells (Rollins *et al* , 2006). The majority (70-80%) of CpG dinucleotides in human are methylated. However CpG dinucleotides are under-represented in the human genome with the exception of CpG islands. About 50% of these CpG islands are located in the promoter regions around transcription start sites of mainly housekeeping genes (Siedlecki and Zielenkiewicz, 2006). Such epigenetic process is important in guarding several biological events in development including embryonic development, X-chromosome inactivation, genomic imprinting, silencing of transposable elements such as endogenous retroviruses or other repetitive sequences, and transcriptional regulation of gene expression (Goll and Bestor, 2005; Henderson and Jacobsen, 2007; Jaenisch and Bird, 2003; Li, 2002;

Lister and Ecker, 2009; Lister *et al.*, 2008; Reik, 2007; Weber *et al.*, 2005). Throughout the past thirty years, intense research findings have widened our understanding on DNA methylation in eukaryotes. The human methylome was finally resolved with the availability of scalable and cost-effective methylation profiling methods. The fantastic human methylome with single bases information was published by the Lister R. *et al.* in year 2009 after sequencing almost the whole genome of embryonic stem cells and fibroblasts (Lister *et al.*, 2009).

1.3.1.2 Implication of altered DNA methylation and human diseases

In addition to its indispensible roles in maintaining normal development, DNA methylation may also serve as the hallmarks of human diseases, by its degree of methylation. As early as in 1983, a decreased total amount of cytosine methylation was reported for many human neoplastic tissues (Feinberg and Vogelstein, 1983). With the advanced techniques in methylation study, it was suggested global hypo-methylation usually happens in repetitive DNA sequences (Cheung *et al.*, 2009). At the same time, aberrant hyper-methylation at gene promoter region was observed in many cancer cells and was thought to contribute to carcinogenesis. Thus, it is now commonly accepted that aberrant DNA hyper-methylation is a hallmark of cancers in which the tumor suppressor genes are silenced (Cheung *et al.*, 2009; Jones and Baylin, 2002; Weber *et al.*, 2005).

1.3.1.3 Molecular basis of DNA methylation

Mammalian CpG methylation is carried out by three active DNA methyltransferases (DNMTs), DNMT1, DNMT3A, and DNMT3B (Henderson and Jacobsen, 2007; Li, 2002). Thus, in studying DNA methylation, studying the

expression level of these enzymes is almost equally important as studying the overall methylated-DNA content. DNMT1 has a higher preference for hemi- methylated DNA as the substrate. It copies the DNA methylation patterns from parental DNA strand into the daughter DNA strand during replication. Therefore, it is usually referred as a maintenance methyltransferase. DNMT3A and DNMT3B are mainly regarded as *de novo* DNA methyltransferases, whose roles are critical in the dynamic DNA methylation process during embryogenesis and pathogenesis. However, it is suggested that DNA methyltransferases do not work solely in maintenance or *de novo* methylation. DNMT1 has been shown to have a high activity and specificity on unmethylated DNA substrate whereas some reported that DNMT3A and DNMT3B played a role in the maintenance of methylation (Hsieh, 2005; Liang *et al.*, 2002; Okano *et al.*, 1998).

1.3.1.4 Alteration of human methylome by eukaryote viruses

An aberrant DNA methylation in cancer cells may be attributed to the effect of cancer-inducing viruses (Ferrari *et al.*, 2009). Several viral oncoproteins were found to interact with cellular DNA methyltransferases (DNMTs) and histone-modifying enzymes, including lysine acetyltransfease (KATs), lysine deacetylases (KDACs) and lysine methyltransferases (KMTs). Table 1.3 summaries the effect of several reported cellular epigenetic modifying viruses. In many tumor viruses, the viral proteins prompt normal cells to replicate when they should remain quiescent and that is actually a hallmark of cancer. The viral proteins can mediate such a non-quiescent status of cell by interacting with the epigenetic modifiers such as DNA methyltransferases. HIV-1 was not regarded as a tumor virus but it was reported that HIV potentiated hyper-methylation in some of the the host genes.

HIV-1 infection of lymphoid cells resulted in increased DNMTs expression and their enzymatic activities (Youngblood and Reich, 2008). This resulted in hyper-methylation and a reduced expression of IFN-γ and p16INK4A (Fang *et al.*, 2001; Mikovits *et al.*, 1998).

Table 1.3 Interaction of viral proteins with cellular epigenetic modifiers, modified from (Adhya and Basu, 2010; Ferrari *et al.*, 2009).

Virus	Viral protein	Epigenetic modifier	Description
Adenovirus	e1a	p300/CBP	KAT: H3 (K14, K18), H4 (K5, K8), H2A (K5), H2B (K12, K15)
		PCAF	KAT: H3 (K9, K14, K18)
		DNMT1	DNA (cytosine-5-)-methyltransferase
		p400	Chromatin remodelling
		TrrAP (GCN5)	KAT: H3 (K9, K14, K18)
Epstein–Barr virus	BZLF1 (also known as Zta or Zebra)	CBP	
	EBNA2	p300/CBP	
	EBNA3	PCAF	KAT: H3 (K14, K18), H4 (K5, K8), H2A (K5), H2B (K12, K15)
	BrLF1	CBP	KAT: H3 (K14, K18), H4 (K5, K8), H2A (K5), H2B (K12, K15)
	LMP1	DNMT1	DNA (cytosine-5-)-methyltransferase
Kaposi's sarcoma associated herpesvirus	virF	p300	KAT: H3 (K14, K18), H4 (K5, K8), H2A (K5), H2B (K12, K15) 30
	K8	CBP	
	LANA	DNMT3a	DNA (cytosine-5-)-methyltransferase
Hepatitis B virus	HBx	DNMT3a	DNA (cytosine-5-)-methyltransferase
	orF50	CPB, HDAC1	KAT, HDAC
Papillomavirus	E2	p300	KAT: H3 (K14, K18), H4 (K5, K8), H2A (K5), H2B (K12, K15)
		CBP	
	E6	p300	
		CBP	
		PCAF	KAT: H3 (K9, K14, K18)
	E7	Mi2	Chromatin remodelling
		p300	KAT: H3 (K14, K18), H4 (K5, K8), H2A (K5), H2B (K12, K15)
		DNMT1	DNA (cytosine-5-)-methyltransferase
Paramecium bursaria Chlorella virus 1	vSET1	Self	KMT: H3 (K27me3)
Simian virus 40	Simian virus 40 large T antigen	p300	KAT: H3 (K14, K18), H4 (K5, K8), H2A (K5), H2B (K12, K15)
		CBP	
Human immunodeficiency virus	Tat	HDAC	Histone deacetylase
Bovine leukemia virus	Tax	HDAC	Histone deacetylase
Herpes simplex virus	Not determined	p300/ CBP	HAT H3
		HDAC	Histone deacetylase
Varicella zoster virus	Not determined	HDAC 1	Histone deacetylase
		HDAC 2	
Dengue virus	Not determined	Histone H3, H4	Histone

1.3.1.5 Current approaches in studying DNA methylation

Several methods have been used to study DNA methylation. In the early years, enzymatic digestion of DNA by methylation sensitive enzymes followed by DNA amplification of the target genes was applied. A better method is called methylation specific polymerase chain reaction (PCR) which allows the study of the methylation status of a specific gene at single base resolution. It is a combination of the chemical conversion of unmethylated cytosines to uracils followed by PCR amplification using primers designed for methylated or unmethylated sequences. Followed by cloning and sequencing of the amplified DNA fragment, the methylation status of specific cytosine of the cloned fragments can be revealed. However, this method is only useful in studying a limited number of target genes because of the tedious experimental procedures. Nowadays several high-throughput methods have been used to study the methylation of genes at a genome-wide level. Widely used approaches include enzyme digestion of methylated samples followed by hybridization onto oligo arrays (Lippman *et al.*, 2005; Martienssen *et al.*, 2005). Alternatively, methylated DNA can be immunoprecipitated using antibodies against 5-methylated cytosines, dubbed MeDIP (Weber *et al.*, 2005). The enriched methylated DNA can be identified by hybridization onto oligo arrays or by the new sequencing technologies (Down *et al.*, 2008). Methods involving the use of oligo arrays have several limitations including low resolution of detection single base methylation status, difficulty in discrimination of similar sequences, inability to determine the sequence context of DNA methylation sites and the array bias towards abundant fragments (Lister and Ecker, 2009). A more advanced but costly method for study the whole genome methylation status at single bases resolution is the whole genome sequencing of the

human methylome. It incorporates the use of bisulfite treatment which translates the epigenetic difference into a genetic one which then allows amplification of DNA for the next-generation DNA sequencing. This is the method used to sequence the whole methylome described in the previous section (Lister *et al.*, 2009).

1.3.2 HIV and human microbiome

1.3.2.1 Introduction to human microbiome

Microbiome is a term used to describe the genetic information of a microbial complex, particularly from an environmental sample that contains different kinds of microbes. It generally refers to the genetic information of bacteria, viruses and fungi. However, it is practically used to describe bacterial population rather than viruses and fungi. The population of viruses is separately described as virome. Human microbiome – genome of the microbiota living on/ in the human body, now interests the mankind than ever in the history of microbiology. By taking into the account of the vast number of the symbiotic microbes which contribute to 100 times more genes, human can be re-described as a 'super-organism'. Up to date, the normal microbiome of several body parts have been described including that for the gut (Qin *et al.*, 2010), distal gut (Gill *et al.*, 2006), and oral cavity (Zaura *et al.*, 2009) and conjunctiva (Dong *et al.*, 2011).

The human microbiota is often taken as commensals, yet they can engage in mutualistic interactions with host directly or indirectly. Nelsons J.H. *et al.* have described the microbiome as an ecosystem in which an equilibrium of various members is required to keep a person healthy (Nelsons, 2010). Apparently, some diseases are associated with the presence of an abnormal proportion of bacteria from

the same taxonomic groups. This is illustrated in Crohn's disease which is associated with a lower than normal proportion of Bacteriodetes in the gastrointestinal tract (Gophna U, 2006), and in active celiac disease which is caused by an higher than normal frequency of Bacteriodetes (Nadal *et al.*, 2007).

1.3.2.2 The Human Microbiome Project

The Human Microbiome Project (HMP) was commenced in 2008 as one of the common fund programs in the National Institute of Health (NIH), USA (http://commonfund.nih.gov/hmp) (Peterson *et al.*, 2009). It aims to dissect the microbiota found in five anatomical sites of the human body, including the oral cavity, skin, nasal cavity, gastrointestinal tract and vagina. The project is not in a scale smaller than the Human Genome Project because it involves sequencing the enormous number of bacterial/ viral genomes in ~300 human subjects. Members of NIH HMP Jumpstart Consortium include the Baylor College of Medicine Human Genome Sequencing Center, the Broad Institute, the J. Craig Venter Institute, and the Genome Center at Washington University. By May 2010, genomes of 117 selected bacteria were available (HMP, 2010). These bacteria are distributed among gastrointestinal tract, oral cavity, urogenital/vaginal tract, skin and respiratory tract. The broad phylogenetic distribution of the sequenced strains was analyzed according to the 16S ribosomal RNA sequences. The analysis was presented in their recently published paper and shown here for a glimpse (Figure 1.7). This includes genome from two kingdom (Bacteria and Archaea), nine phyla, 18 classes, and 24 orders.

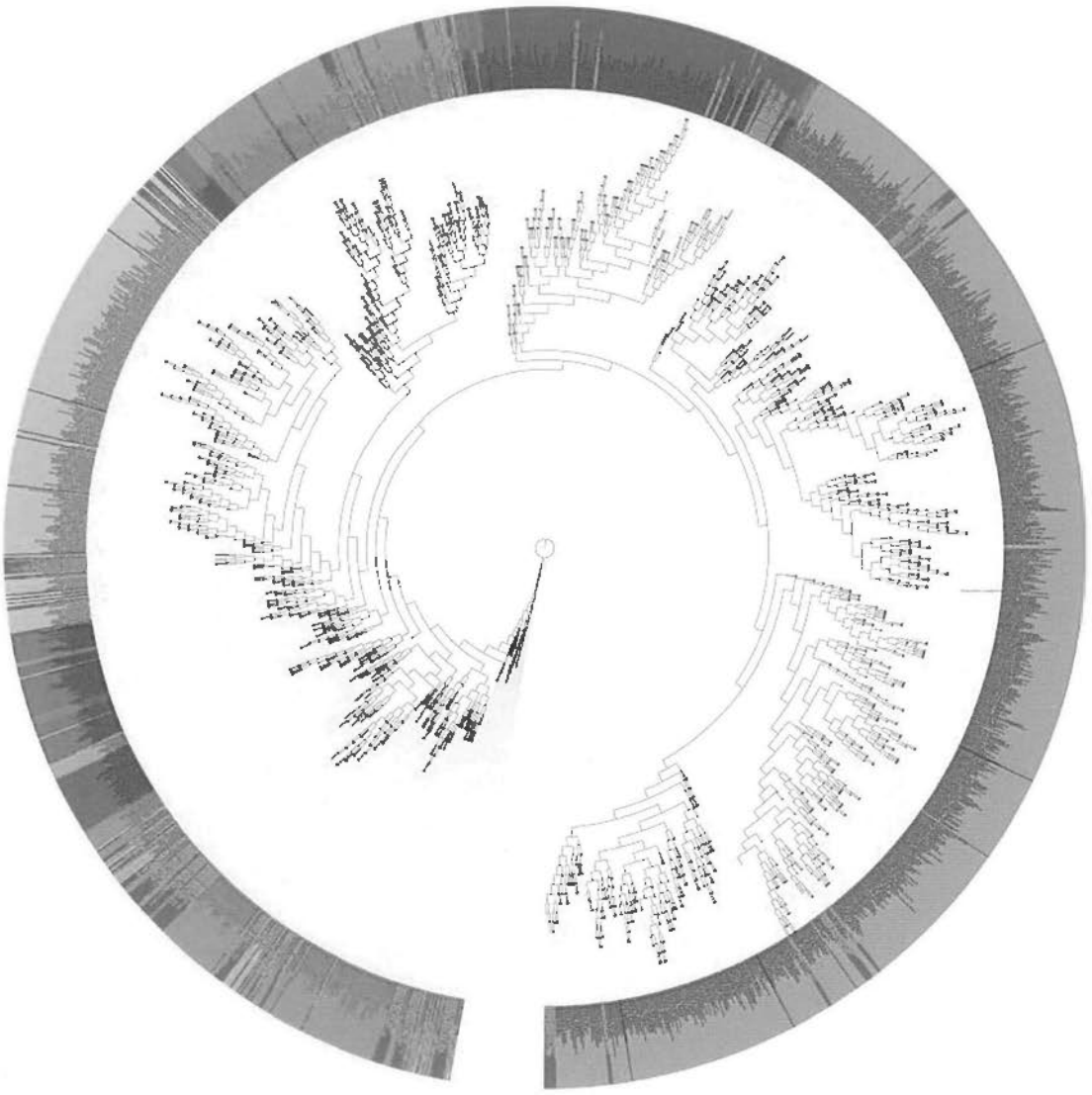


Figure 1.7 Phylogenetic tree of 16S rDNA sequences (adapted from HMP, 2010). The original figure can be enlarged for details and is available at (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2940224/figure/F1/). Blue, organisms sequenced in HMP; yellow, Actinobacteria; dark green, Bacteriodetes; light green, Cyanobacteria; red, Firmicutes; cyan, Fusobacteria; dark red, Planctomycetes; grey, Proteobacteria; magenta, Spirochaetes; light pink, TM7; tan, Tenericutes

1.3.2.3 Human microbiome in HIV/AIDS patients

Depletion of CD4+ T-lymphocytes is a distinctive feature of HIV infection. Development of the acquired immunodeficiency syndrome (AIDS) is defined when the CD4+ cell count of the HIV-carrier drops to <200 cells per microliter (Stebbing *et al.*, 2004). While it is still not known whether HIV infection would alter the population of normal microbiota on the skin, in the gastrointestinal tract or oral cavity of the infected person, it has been known to directly increase the host's susceptibility to opportunistic infections (Kapogiannis *et al.*, 2008; Mootsikapun, 2007). Such infections can be caused by bacteria or viruses which result in bacteremia or viremia (presence of these species in the blood stream). Bacteremia and viremia are the most important causes of morbidity and mortality among HIV carriers. Thus, it is important to identify the pathogens in the blood stream of the patients. However, it usually takes several days to two weeks to culture the blood microbes before appropriate medications can be applied. Moreover, not all bacteria or viruses can be cultured using currently available techniques (Handelsman, 2004).

1.3.2.4 Current approaches in studying microbiota

Nowadays, 16S rDNA sequencing is the most commonly used method in molecular identification of bacteria. It involves automated sequencing of partial or the full-length 16S rDNA gene (Gill *et al.*, 2006; Woo *et al.*, 2003; Zaura *et al.*, 2009). The method is especially useful in identifying bacteria which cannot be identified using traditional methodologies such as phenotype recognition, staining methods and biochemical tests (Petti *et al.*, 2005; Woo *et al.*, 2008). Bacteria which present atypical phenotypes or are rarely encountered can be identified by 16S sequencing unambiguously. The method can be extended to identify the

slow-growing bacteria by skipping the long culture time since a trace amount of starting microbes will be sufficient. Identification of bacteria which requires special culturing equipment is usually not possible in laboratories without such equipment. But by incorporating 16S sequencing in the routine microbiological tests which need not the culture step, these bacteria can be identified (Woo *et al.*, 2008).

16S sequencing is useful in clinical laboratories for practical reasons. Yet, it is limited to bacteria of a pure culture, or to bacteria in a relative larger proportion if not pure. To study a complex sample such an environmental sample, the typical 16S sequencing will be too labour-intensive. Moreover, amplification of 16S rDNA gene may not always be feasible in clinical samples such as plasma (Ferri *et al.*, 2010). High-throughput next-generation sequencing thus has been used for sequencing an enormous number of 16S clones (Gill *et al.*, 2006). More strikingly, the primer-free sequencing that surpasses the PCR step can be applied directly on metagenome, as illustrated by an elegant study on the metagenome found on the windshield splatter (Kosakovsky Pond *et al.*, 2009). In their pioneering study, the scientists has sequenced two environmental samples by the 454 Life Science (Roche) FLX sequencing. The metagenome of the constituting bacteria was then resolved by comparing the sequences against the available bacterial genome database, as a method termed 'GALAXY'. Another profound study is the generation of the gut microbial gene catalogue by use of Illumina GA short read sequencing, published in *Nature* 2010 (Qin *et al.*, 2010).

1.4 Objectives of the research

In this thesis, two projects on HIV were described. In the first project, the

HIV-host relationship at the moderately advanced stage of the disease was studied, as exemplified by the determination of DNA methylation induced by HIV-1. The second project focuses on elucidating the common opportunistic infections in HIV/AIDS patients, which refers to the advanced stage of the disease.

1.4.1 Sequencing of HIV-1 associated cytosine methylation

Knowing how HIV alters the human cytosine methylation pattern will potentially unveil the mechanism of viral induced cellular changes. Given the fact that the normal DNA methylation pattern in man to man can be very different, we have intended to study the HIV-induced methylation by comparing the cytosine methylome between a pair of monozygotic twins with one infected with HIV and the other uninfected. The methylated DNA patterns between the twins were resolved by the next generation high-throughput sequencing. The sequencing results together with the molecular basis of the regulation and consequences of methylation were described in Chapter 2.

1.4.2 Analysis of plasma microbiome in AIDS patients

Knowing the plasma microbiome information in AIDS patients will be beneficial to the identification of the common bacteria and viruses associated with these patients. Together with the clinical symptom description of the patients, the reference pathogen list will help to facilitate a faster screening of infection in these patients thus a prompt administration of appropriate therapeutic agents and a better disease prognosis. Thus, we have performed a metagenome analysis on the nucleic acids isolated from the plasma of HIV/AIDS patients in order to reveal a whole spectrum of all the microbes including the bacteria, fungi and viruses. The

discovered microbes may include cultivatable and non-cultivatable microbes that are now known to or unknown to human. The plasma microbiome and virome in AIDS patients were described in Chapter 3 and Chapter 4 respectively. Apart from establishing a reference data-set for the commonly AIDS-associated microbes, we also ought to identify some novel species of pathogens. The sequence reads with a significant difference between the unknown species and those most closely related ones will be investigated to discover the potential novel species. The findings on bacteria and viruses were also described in Chapter 3 and 4. In order to elucidate the micro-organisms that can survive in the blood of a normal healthy adult, we also set out to compare the plasma microbiome from two separate sets of cohort in Hong Kong. The results were given in Chapter 5.

Chapter 2

Analysis of HIV-1 associated DNA cytosine methylation in HIV carriers

Summary

The knowledge in DNA methylation-mediated gene regulation has brought the scientists a step closer to understanding the virus-host interplay in the context of genome alteration. Many viruses including HIV have been shown to regulate the host genome methylation. We have therefore carried out a large-scale evaluation on HIV-1 associated genome-wide DNA methylation pattern using identical twin samples. The methylation pattern in the HIV-positive subject was compared to the HIV-negative twin sibling. Analysis of the gene list showed that upon HIV infection, the genes regulating RNA splicing and cell cycle regulation were hypo-methylated whereas the genes regulating nerve functions and signal transduction were hyper-methylated. The molecular study suggested that HIV-1 increases the transcription level of DNA methyltransferase (DNMT) 3A but not the DNMT1 or DNMT3B. The up-regulation of DNMT3A was not associated with the methylation of HIV-1 promoter LTR and latency in chronically infected cells with actively replicating viruses.

2.1 Introduction

2.1.1 HIV and DNA methylation

As described in chapter 1, many eukaryotic viruses can interact with cellular factors to alter DNA methylation and therefore to affect host genes transcription. In contrast to the myriad studies on HIV-induced mRNA differential expression, the HIV-induced methylation alteration is rarely described. An up-regulated expression level of DNA methyltransferases (DNMT) 1 was observed in *in vitro* HIV acute infection model (Fang *et al.*, 2001; Youngblood and Reich, 2008). However, a detailed description on the mechanism and consequence is still lacking.

It is known that endogenous retroviruses in the human genome can be silenced epigenetically by the host DNA methyltransferases. These include transcriptional repression of the proviral DNA, modifying the chromatin conformation, and epigenetically methylating the DNA at the enhancer/ modulatory region (Bednarik *et al.*, 1990; Lassen *et al.*, 2004; Tanaka *et al.*, 2003). Such hyper-methylation can also happen to the integrated HIV genome, such that it may explain the virus latency in chronic infections. The HIV long terminal repeats (LTR) was hyper-methylated at the CpG sites in the 5'LTR but not the 3'LTR. Demethylation of CREB/ATF sites in the LTR has also been shown as a crucial step for the reactivation of latent HIV-1 genome *in vivo* (Blazkova *et al.*, 2009; Ishida *et al.*, 2006). Similar methylation suppression was observed in HBV virus. HBV can induce an over-expression of the host DNMTs, which on the other hand, methylates the viral DNA to cause a decreased viral gene expression and replication (Vivekanandan *et al.*, 2010). Until now, there was no published data on the inter-methylation effect between HIV and the infected host.

2.1.2 Monozygotic twins as subjects to study DNA methylation and transcriptome

In this part of the study, we ought to resolve the DNA methylation pattern in HIV infected subjects in order to identify the genes affected by HIV-1 by means of methylation. To study DNA methylation at single bases resolution, the best approach is to perform a bisulfite converted full genome of the same study subjects. However, such artificial genome can be effectively composed of just 3 bases which leads to a high error rate when base-calling is performed based on the bisulfite converted sequence. Therefore, a reference control full genome of the same study subject without bisulfite treatment should be included, adding an extra cost on sequencing (Lister and Ecker, 2009). The second problem is the person-to-person epigenetic variations which make it difficult to associate a methylome with a disease. Identical twins have been commonly used to rule out genetic variation such that, the difference between the identical twins were attributed to the exogenous factors such as viruses. In the context of epigenetic variations, it was shown that monozygotic monochorionic twins have similar methylation content in their genome, in contrast to the non-similar dizygotic twins (Kaminsky *et al.*, 2009).

In view of this, we have recruited a pair of monozygotic (MZ) twins in mainland China, with one infected by HIV-1 and the other not. Due to the limited amount of samples which is insufficient for generating a full bisulfite converted genome and a reference genome, we have performed methylated DNA immunoprecipitation followed by high-throughput sequencing for the study purpose. While a fair gene normalization control is not available for this kind of MeDIP-Seq, an overall hypo-methylation was observed in the HIV infected twin subject. The sequencing

results were validated by examining the differential mRNA expression level of the selected genes including genes for RNA splicing, cell cycle control, nerve functions and cell signaling.

2.2 Materials and Methods

2.2.1 Ethics statement

This study was conducted according to the principles expressed in the Declaration of Helsinki. The study was approved by the Institutional Review Board of Jiangsu University, China and the Institutional Review Board of the Chinese University of Hong Kong. All the studied subjects provided written informed consent for the collection of samples and subsequent analysis.

2.2.2 Detection of HIV infection in monozygotic twin samples and their mother

Twin subjects were recruited from an AIDS village in mainland China. Screening of potential HIV infection was done by detecting the HIV-1/2 proteins using the Colloidal Gold kit according to the manufacturer's instruction. A pair of female twins was found in the screening. The basic information of the twins was not disclosed as a privacy measure on subject protection. A confirmation of HIV-1 infection was determined using the HIV Blot 2.2 Western Blot confirmation assay (Genelabs). The HIV viral load was determined using HIV real-time PCR assay (Shenzhou PG Biotech). The CD4+, CD8+ and CD3+ cell-counts were determined by FACS counting using BD Tritest with BD Trucount tubes (BD Biosciences).

2.2.3 Isolation of genomic DNA from PBMCs

Peripheral blood mononuclear cells (PBMCs) which contain the CD4+

T-lymphocytes were separated from the EDTA-blood collected from the monozygotic twins. Total genomic DNA was extracted from the PBMCs using QIAamp DNA Mini/ blood mini Kit (QIAGEN) according to the manufacturer's instructions. The extracted genomic DNA was dissolved in 300 µl nuclease-free water.

2.2.4 Immunoprecipitation of cytosine methylated DNA (MeDIP) and amplification of DNA

Immunoprecipitation of cytosine methylated DNA from twins genomic DNA was performed using the Methylated-DNA IP Kit (Zymo Research). The genomic DNA was first sonicated to form fragments of 250-500 b.p. in length. A monoclonal antibody against methylated-cytosine was used to capture the methylated DNA fragments. In order to get 10 µg of DNA as required for Illumina Solexa sequencing, the immunoprecipitated methylated-DNA was amplified using random primers and Phi29 enzyme as instructed in GenomiPhi V2 kit (GE Medical Systems) (Reyes *et al.*, 2010). Five microliter immunoprecipitated DNA sample was mixed with 5 µl sample buffer and heat denatured at 95°C for 3 minutes. The samples were then cooled on ice. Nine microliter reaction buffer and 1 µl enzyme mixture were added to the DNA sample and incubated at 30°C for 90 minutes. The reaction was terminated by heating at 65°C for 10 minutes. The amplified samples were analyzed on DNA agarose gel before purification and purified afterward.

2.2.5 Illumina Solexa sequencing of MeDIP samples

The amplified and purified plasma DNA samples were sequenced by paired-end Illumina Solexa short-read sequencing technology with insert size of 250 base pairs

(b.p.) and read length of 75 b.p. The sequence signals were analyzed using an Illumina Genome Analyzer.

2.2.6 Analysis of sequencing results

2.2.6.1 Sequence reads assembly and mapping

The sequence reads were assembled into contigs using Velvet version 1.0 (Zerbino and Birney, 2008). The sequence reads were mapped to the *Homo sapiens* (human) genome Build 37.2 by an in-house algorithm of Hong Kong Bioinformatics Centre of The Chinese University of Hong Kong. The contigs aligned to the 5' regulatory regions of annotated genes with reference to human genomic/ plus transcript (G+T) database were selected. The 5' regulatory region in study was defined as the 1,000 nucleotide upstream of the +1 transcription start site 1,000 nucleotides downstream of the transcription start site of the corresponding genes. The degree of methylation of the gene promoter region was determined by counting the number of mapped reads. Since the total sequence reads from HIV positive sample and HIV negative sample differed quite a lot, statistical analysis for the significance of the methylation was performed and presented as p-value (Appendix Table 1) (Klipper-Aurbach *et al.*, 1995; Storey and Tibshirani, 2003).

2.2.6.2 Gene function annotation and pathways analyses

The gene functions were annotated using The Database for Annotation, Visualization and Integrated Discovery (DAVID) (http://david.abcc.ncifcrf.gov/) as of 11 March 2011 (Huang da *et al.*, 2009). We have adapted the official symbols for the gene names and annotation sources GOERM-BP (biological process), and GOTERM-MF (molecular function) for the pathway analyses.

2.2.7 Cell culture and plasmids

2.2.7.1 Cell culture

H9 T-lymphocyte cell line (ATCC HTB-176) and H9/HIV_{IIIB} T-lymphocyte cell line (ATCC CRL-8543) were cultured in ATCC formulated RPMI medium supplemented with 10% heat inactivated fetal bovine serum, 50 U/ml penicillin and 50 μg/ml streptomycin. The former was cultured in a biosafety level (BSL) 2 laboratory whereas the later was cultured in the BSL3 facilities of The Chinese University of Hong Kong. The HIV_{IIIB} lab-adapted strain is more heterogeneous than just the reference HIV clone HXBc2 and thus has been used to represent a population of viral genotype (Brass *et al*, 2008). The human embryonic kidney cell-line HEK293 is a gift from Dr. W.P. Fong from the School of Life Sciences, The Chinese University of Hong Kong. The HEK293 cells were cultured in DMEM medium (Gibco, Invitrogen) supplemented with 10% heat inactivated fetal bovine serum, 50 U/ml penicillin and 50 μg/ml streptomycin. Transfection of HEK293 was performed on 60 mm culture dish using Lipofectamine LTX (Invitrogen) according to the manufacturer's instructions.

2.2.7.2 Plasmids

Plasmid pcDNA3.1 was modified to form pcDNA3.1-3X-FLAG by the insertion of three repeats of FLAG tag sequence at the *Hin*dIII site. Plasmid pcDNA-Tat construct was obtained from Dr. M. Peterlin (Cujec *et al.*, 1997) through Addgene Incorporation (http://www.addgene.org/pgvec1). The HIV-1 *Tat* gene was subcloned from pcDNA-Tat. HIV-1 *Vpr* and *Nef* genes was amplified from the RNA samples used in a previously described study (Tsui *et al*, 2010). The HIV-1 genes were cloned into pcDNA3.1-3X-FLAG. pGL3-basic luciferase reporter

vector used in the promoter study and the pGEM-1 Easy System used in bisulfite sequencing was obtained from Promega. Primer sequences used in cloning were summarized in Table 2.1.

2.2.8 Extraction of genomic DNA, total RNA and protein from cell lines

The extraction of genomic DNA, total RNA and proteins were done separately for H9 and H9/HIV_{IIIB} in BSL 2 and BSL 3 laboratory respectively. Genomic DNA and total RNA was extracted from H9 and H9/HIV_{IIIB} cell lines using QIAamp DNA Mini/ blood mini Kit (QIAGEN) and RNeasy Mini kit (QIAGEN) respectively according to the manufacturer's instructions. The total protein was extracted by firstly lysing the cells in Radioimmunoprecipitation assay (RIPA) buffer supplemented with protease inhibitors (Roche) for 15 min and then centrifuged at 12,000 x g at 4°C to get the protein lysate in the supernatant.

2.2.9 Determination of relative transcription by real-time PCR

Quantitative real-time PCR with Power SYBR® Green PCR Master Mix (Applied Biosystem) was used to measure the mRNA expression of specific target genes in H9 and H9/HIV_{IIIB} cell line. Real-time PCR was performed in triplicate in a 10 µl reaction on an ABI Fast 7500 real-time PCR machine (Applied Biosystem). The threshold cycle number (Ct) value defines the number of cycles for reaching the threshold was designated to the samples by auto-detection. After normalization with the endogenous housekeeping gene GAPDH, the ΔΔCt value was calculated for the relative expression of the target transcripts. The sequences of primers used for real-time PCR were given in Table 2.2.

2.2.10 Western blot analysis

Whole-cell extracts were resolved by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and then transferred to Immobilon-P membrane (Millipore). The membranes were probed with the indicated antibodies. Monoclonal antibodies from mouse targeting Tat (ab42359), DNMT1 (ab13537), DNMT3A (ab13888) and DNMT3B (ab13604) were obtained from Abcam and used at 1 μg/ml of 5% blocking solution (Bio-Rad) with 0.1% Tween-20 (USB). Mouse antibody conjugated with horse-radish peroxidase (HRP) for β-actin protein input control was obtained from Santa Cruz (sc-47778) and used at 1:20,000 dilution. Mouse anti-FLAG antibody was obtained from Sigma Aldrich (F3165) and used at 1:1,000 dilution. Goat anti-mouse-HRP was obtained from Bio-Rad and used at 1:1,000 dilution.

2.2.11 Promoter studies

2.2.11.1 Analysis of DNMT3A promoter

The reference sequence of DNA methyltransferases (DNMT) 3A was retrieved from NCBI GenBank ID: 1788. The promoter of DNMT3A was amplified from the genomic DNA extracted from H9 or H9/HIV_{IIIB} T-lymphocytes and cloned into pGL3-basic luciferase reporter vector (Promega). The sequence of promoter was determined by conventional Sanger sequencing. The primers used for cloning was summarized in Table 2.1.

2.2.11.2 Determination of methylation status of promoters

The methylation status of the target genes was determined by bisulfite sequencing. The total genomic DNA was treated with bisulfite using EZ DNA

Methylation-Gold™ Kit (Zymo Research) as instructed. The unmethylated cytosines were converted to uracils in bisufite treatment whereas the methylated cytosines were not affected. The target promoter region was amplified using bisulfite-matched primers. The amplified fragment was cloned in pGEM-T easy system (Promega) and sequenced. The sequence of the bisulfite treated DNA was aligned with the reference gene sequence to determine the site of methylated CpG dinucleotide using BioEdit (Hall, 1999). The summary of the bisulfite converted sequence was further aligned and presented in graph using the CpG Viewer (Carr *et al.*, 2007).

Table 2.1 Primers used in cloning.

Gene	Target	Primer name	Direction	Sequence 5' -> 3'
DNMT3A	Promoter	Kpn-D3A-pro-1254F	Forward	TAGGTACCGCAGAGGTATCTCAGTCATCCTTC
		Kpn-D3A-pro-977F	Forward	TA GGTACCCACTTGAGTGAGGGTGACACTGAA
		Kpn-D3A-pro-604F	Forward	TA GGTACCGAATCTTCCAGCTGGAAGAGACCTT
		Kpn-D3A-pro+235F	Reverse	TA GGTACCTACGACTCGAGGGGCTCTGAGT
		Kpn-D3A-pro+554F	Reverse	TA GGTACCGGAAGGAGGAAGGCGGAGAGAA
		BglII-D3A-pro+781-R1B	Reverse	TA AGATCTAGCGGCGCTCATTAC CGTATGG
HIV_{IIIB} LTR	1st round PCR	HIV-LTR-F1	Forward	AACTACACACCAGGGCCRGGGRYY A
		HIV-LTR-R1	Reverse	CGGGCACACACYACTTKA AGCACTC
	2nd round PCR	HIV-LTR-F2	Forward	CACCAGGGCCRGGGRYYAGATAYC
		HIV-LTR-R2	Reverse	GACCCAGTACARGCRAAAAGCRGCT
Tat	Tat	HindIII-tat-F	Forward	GAT ATAAAGCTTGAGCCAGTAGATCCTA
		NotI-tat-R	Reverse	GAT ATAGCGGCCGCCTAAGGGACTGGATCTGTC
Vpr	Vpr	HindIII-vpr-F	Forward	GAT ATAAAGCTTGAACAAGCCCCAGAA
		NotI-vpr-R	Reverse	GAT ATAGCGGCCGCCTAGGATCTACTGGCTCCAT
Nef	Nef	HindIII-nef-F	Forward	GAT ATAAAGCTTGGTGGCAAGTGGTCA
		NotI-nef-R	Reverse	GAT ATAGCGGCCGCTCAGCAGTTCTTGTAGTACTC

Table 2.2 Primers used in real-time polymerase chain reaction.

Gene	Transcript accession	Primer	Sequence (5'->3')	Length	Start	Stop	Tm	GC%
SFPQ	NM_005066	Forward	GCGGGGTCCCTACCACACCT	20	608	627	60	70%
		Reverse	TTGGGACCACCCGGACCTGG	20	716	697	60	70%
AKAP17A	NM_005088.2	Forward	GCACTCCCGCAGCTGAAGCA	20	299	318	60	65%
		Reverse	CACCTCCCCCTCGAAGCGGA	20	445	426	60	70%
SNRNP48	NM_003090	Forward	TCGCCCCGCCCGTCTTCTAA	20	3940	3959	60	65%
		Reverse	AGCTTTGCGCCAGAACCGTT	20	4019	4000	58	55%
MAGOHB	NM_018048.2	Forward	GGGGACACGTTGGCTGCGTT	20	33	52	60	65%
		Reverse	AGCTTTCCGTCCGGCCGAAA	20	178	159	59	60%
TXNL4A	NM_006701.2	Forward	TCGAGACGGTGTACCGCGGG	20	487	506	61	70%
		Reverse	AGGGCGCCTCAGTAGCGGTA	20	577	558	59	65%
Gemin 8	NM_001042480.1	Forward	AGAAGAACGACGGCGGCAGC	20	455	474	60	65%
	(NM_001042479.1)	Reverse	TTCTACCGACCGGCGGGTGT	20	554	535	60	65%
	(NM_017856.2)							
SNRPA1	NM_003090.2	Forward	CAGGCGGCGCAGTACACCAAC	21	108	128	61	67%
		Reverse	TTTATACCCCCGGAGGTCCAGCTC	24	167	144	59	58%
SKA3	NM_001166017	Forward	ATGAGGCCGCCCCTGACCTT	20	331	350	60	65%
	(NM_145061.5)	Reverse	GGAAGGCAGAGCTGGCTGGC	20	390	371	60	70%
TIMELESS	NM_003920.3	Forward	GCAGCCCAGATCCTACAGAG	20	169	188	60	60%
		Reverse	CAGATTGCCAAAACAGAGCA	20	300	281	60	45%
SPC25	NM_020675.3	Forward	GCAAAAAGCAGGAATTGGAA	20	272	291	60	40%
		Reverse	TTCTGCAGCCTTTTCAACCT	20	398	379	60	45%
MND1	NM_032117	Forward	CCGCAAGTTGTGGAAGAAAT	20	445	464	60	45%
		Reverse	TTGGCCCAAGATTTTATTGC	20	542	523	60	45%
DDX1	NM_004939.1	Forward	TCCCGGGAGTTAGCTGAACAAACT	24	901	924	58	50%
		Reverse	GCTGATCCCGGGCTGCAACA	20	1015	996	60	65%
HTT	NM_002111	Forward	TAGCCAAACAGCAGATGCAC	20	5004	5023	60	50%
		Reverse	CCATTGTGTTTGGAGTGACG	20	5141	5122	60	50%
NDUFA2	NM_002488.4	Forward	GTCCGCGGTTGGTCAGACCG	20	134	153	60	70%
	(NM_001185012.1)	Reverse	CTCCTCGACTTGCTGCGGCC	20	230	211	60	70%
MYCBP2	NM_015057.4	Forward	GCGGACTCCCGGGGTCACTA	20	439	458	60	70%
		Reverse	ATCCTCCGGTAGCGGTCGGC	20	506	487	60	70%
ARHGEF3	NM_00112861.5	Forward	GCTGGCTCCCTTGCCTCAGC	20	386	405	60	70%
	(NM_09555.2)	Reverse	TCCAGCAGAGCTTTGGCGGC	20	455	436	60	65%
	(NM_001081562.1)							
DMPK	NM_00108156.2	Forward	GGATTCCGGCCGAGATGGCG	20	279	298	60	70%
	(NM_001081560.1)	Reverse	CACGTAGCCAAGCCGGTGCA	20	358	339	60	65%
	(NM_004409.3)							
	(NM_001081563.1)							

Table 2.2 (continued) Primers used in real-time polymerase chain reaction.

Gene	Transcript accession	Primer	Sequence (5'->3')	Length	Start	Stop	Tm	GC%
MBD5	NM_018328.4	Forward	GATGGAACATGCAAGTGTGG	20	1166	1185	60	50%
		Reverse	CGGTTCTCTGTTTCACAGCA	20	1256	1237	60	50%
IK	NM_006083.3	Forward	TGTGGGAAAGAGCTTGTCGCTGC	23	35	57	60	57%
		Reverse	GGCCAAAGGGTTGGAGAACGGC	22	152	131	60	64%
PTPLA	NM_014241.3	Forward	CAGACATTTGCCTTGCTTGA	20	390	409	60	45%
		Reverse	CTTGGACCCCAGTCACAATC	20	468	449	60	55%
REPIN1	NM_013400.3	Forward	TCCTTGGTCTGTCTCGCTTT	20	2421	2440	58	50%
	(NM_001099695.1)	Reverse	CAGACAAGAGCCTGGACACA	20	2510	2491	61	55%
	(NM_001099696.2)							
IL4V1	NM_000589.2	Forward	TCTGTGCACCGAGTTGACCGT	21	508	528	63	51%
		Reverse	AACTGCCGGAGCACAGTCGC	20	606	587	65	65%
GAPDH	NM_002046.3	Forward	TTGCCATCAATGACCCCTTCA	21	194	214	60	48%
		Reverse	CGCCCCACTTGATTTTGGA	19	367	349	58	53%
DNMT1	NM_00130823.1	Forward	CGACTACATCAAAGGCAGCAACCTG	25	3159	3183	67	52%
	(NM_001329.2)	Reverse	TGGAGTGGACTTGTGGGTGTTCTC	24	3324	3301	67	54%
DNMT3A	NM_175629.1	Forward	CGAGTCCAACCCTGTGATGATTG	23	2657	2679	65	52%
	(NM_153759.2)	Reverse	GCTGGTCTTTGCCCTGCTTTATG	23	2796	2774	65	52%
	(NM_022552.3)							
DNMT3B	NM_006892.3	Forward	TTGGAATAGGGGACCTCGTGTG	22	989	1010	64	55%
	(NM_175848.1)	Reverse	AGAGACCTCGGAGAACTTGCCATC	24	1140	1117	67	54%
	(NM_175849.1)							
	(NM_175850.1)							

Footnote:

Bracket ( ) : transcript variants of the same gene that can be detected using the same set of primers

2.3 Results

2.3.1 Determination of DNMTs expression in HIV-positive cells

HIV has been proposed to induce an aberrant methylation of the host genome. However, there is no direct evidence on this to date. We therefore began the study by determining the expression level of DNA methyltransferases in H9/HIV_{IIIB} T-lymphocytes, which chronically expresses a lab-adapted strain HIV_{IIIB}. The expression of HIV was detected by anti-Tat antibody (Figure 2.1A). Total RNA was extracted from the H9/HIV_{IIIB} T-lymphocyte and its parental H9 cell-lines. When compared to its parental cell-line H9 T-lymphocytes, *DNMT1* showed a significant (p-value: 0.013) but not impactful (1.13, < 2 fold) decrease in expression. *DNMT3A* showed a significantly higher expression (p-value: 0.001) than its parental cell-line (3.47 fold, > 2 fold). *DNMT3B* also showed a slight increase in transcription level which is not significant (1.79, < 2 fold) (Figure 2.1B). The protein expression of DNA methyltransferase 3A was greatly increased in H9/HIV_{IIIB} as revealed on western blot. No increase or decrease in protein expression of *DNMT1* and *DNMT3B* was observed (Figure 2.1C).

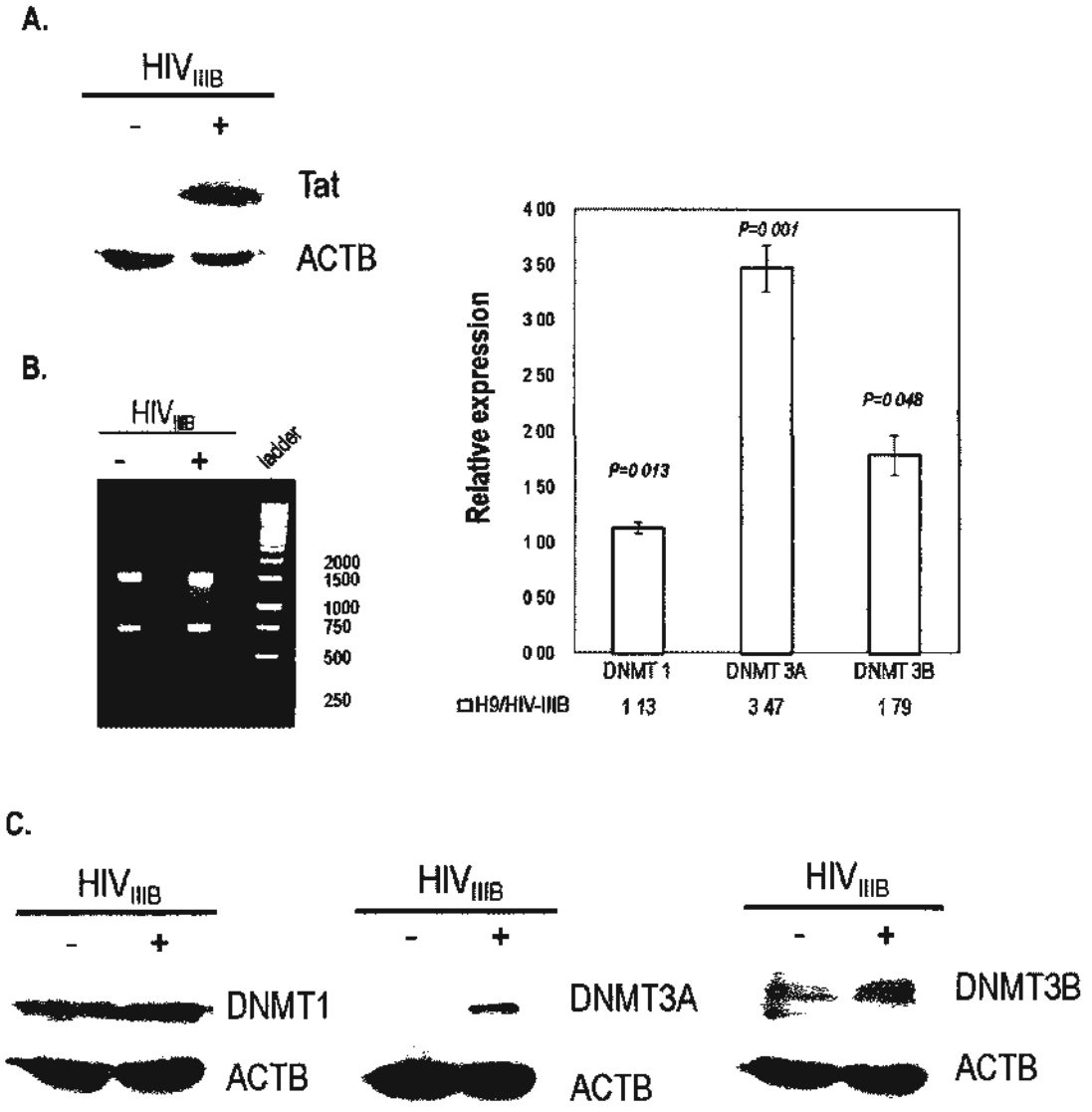


Figure 2.1 Determination of the relative expression of DNA methyltransferases in H9/HIV$_{IIIB}$ compared to that of its parental cell-line H9. **A.** Expression of HIV in H9/HIV$_{IIIB}$ was determined by the detection of Tat protein on western blot. **B.** Total RNA was isolated from H9 T-lymphocyte cell line and H9/HIV$_{IIIB}$ cell line. *DNMT1* mRNA expression of HIV-expressing cell-line was slightly decreased when compared to its parental cell-line (fold change: 1.13; p-value: 0.013). *DNMT3A* mRNA showed a significant increase (fold change: 3.47; p-value: 0.001.) *DNMT3B* showed a slight increase in mRNA expression (fold change: 1.79; p-value: 0.048). **C.** H9 cells infected by HIV$_{IIIB}$ exhibited a higher protein expression level of DNMT3A while no change was observed for DNMT1 and DNMT3B.

2.3.2 Diagnosis of HIV infection

Twins subjects were recruited in an AIDS village in the mainland China. To identify HIV infection, an easy and fast test was performed to detect the HIV-1/2 antibodies (Figure 2.2). The results showed that one of the twins was infected with HIV-1/2 while the other was not. Western blot results confirmed the HIV infection in one of the twins to eliminate false positive results in the screening test (Figure 2.3). The CD4+, CD3+ and CD8+ cell counts were given in Table 2.3. The CD4+ cell count for the twin subject uninfected (T -) was acceptable. The infected subject (T +), however, had a suboptimal count of CD4+ cells of <400 cells/ microliter. The CD4+ cell count suggested the T+ subject was at the moderate advanced stage of HIV disease instead of getting infected recently.

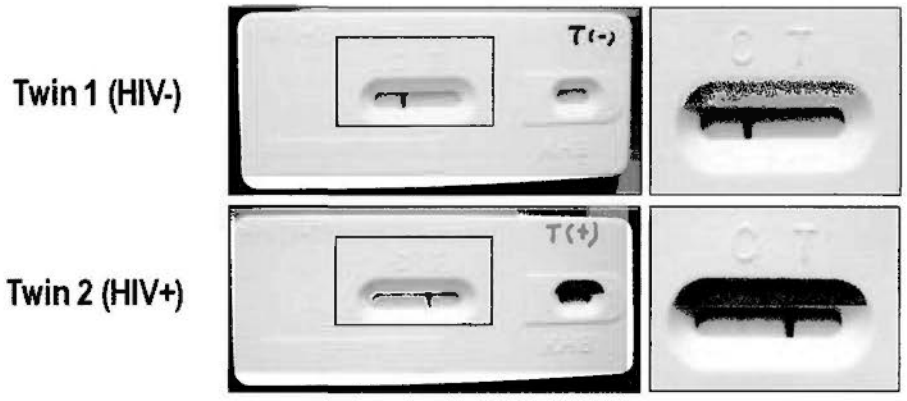


Figure 2.2 Detection of antibodies against HIV-1 and HIV-2. This initial fast test showed that one of the twins was infected with HIV while the other was not.

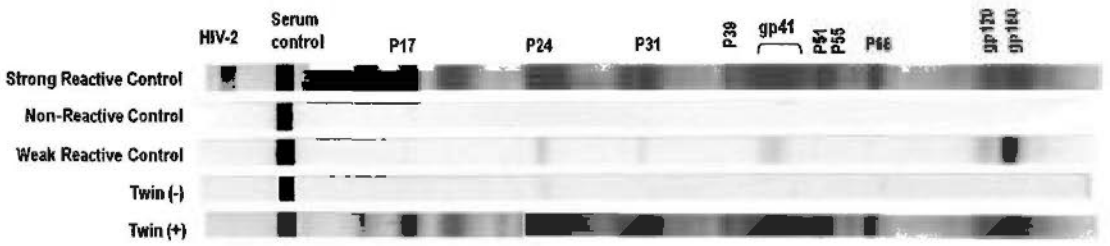


Figure 2.3 Western blot confirmation tests for HIV proteins in the twin samples. This was a confirmation test to avoid false positive results in the simple HIV-1/2 test. The test serum sample was compared to the strong positive control, weakly positive control and negative controls. One of the twins was positive for HIV protein p17, p24, p31, p39 etc, with signal comparable to the strong positive control. The test further showed that the subject was not infected by HIV-2.

Table 2.3 Cell enumeration results of the twin subjects used in this study.

Subject	CD4 /μl	CD8 /μl	CD3 /μl	CD4/CD8	CD4/CD3	CD8/CD3
Twins (HIV-)	665	945	1778	0.7	0.37	0.53
Twins (HIV+)	336	1054	1509	0.32	0.22	0.7
Healthy physical range	706-1125	323-836	1027-2086	1--2	0.30-0.54	0.15-0.34

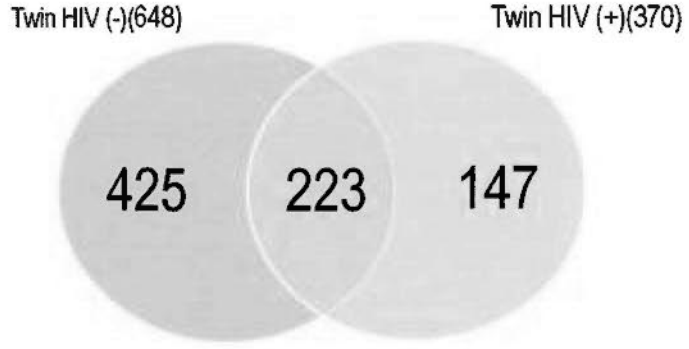
2.3.3 Differential methylation genome associated with HIV

2.3.3.1 Determination of methylated genes

The determination of methylated DNA was done by MeDIP sequencing. Fragments of methylated DNA isolated with an anti-methylated cytosine antibody were evaluated using an Illumina Solexa sequence analyzer. The sequence contigs were then mapped onto the reference human genome Build 37.2. By considering the fact that 70-80% CpG methylation happens in the 5' regulatory/ promoter region of genes, contigs mapped onto the 5' regulatory region of annotated transcripts were extracted for further analysis. The matched genes were screened to remove the genes insignificantly associated with HIV (p-value >0.001), ribosomal proteins,

pseudogenes and genes without known function (e.g. with gene prefix KIAA, LOC) (Appendix Table 1). The resulted gene list was compared and contrasted to generate a specific methylation gene list for HIV negative and another for HIV positive. The overview of methylated genes in the twin subjects was shown in a Venn diagram in Figure 2.4A. A total of 425 genes were methylated in twin negative subject but unmethylated in twin positive subject. Similarly, 147 genes were methylated in twin positive but unmethylated twin negative subject. There were 223 gene methylated in both subjects, irrespective of the degree of methylation. Genes that were methylated in uninfected subject but unmethylated in the infected subject was in Figure 2.4B. These represent the genes that were hypo-methylated upon HIV infection. Genes that were exclusively methylated in infected subject were shown in Figure 2.4C. These represent the genes that were hyper-methylated upon HIV infection.

A.



B.

1	ABCG8	62	CTSLL4	123	GLULP2	184	LILRA3	245	NCRNA00114	306	RHOT1P2	367	TEP1
2	ACAA2	63	CXADRP1	124	GLYATL1P2	185	LIMD1	246	NCRNA00116	307	RHPN2	368	THAP10
3	ACAP1	64	CXorf24	125	GNL1	186	LONP2	247	NCRNA00161	308	RILPL1	369	TIMELESS
4	ACSS2	65	CYB5AP4	126	GP2	187	LPA	248	NCRNA00292	309	RLN1	370	TIMM23
5	ADAM33	66	CYCSP34	127	GPR112	188	LRRN4CL	249	NDUFAF1	310	RN7SL3	371	TJP3
6	ADAMTS13	67	CYP3A5P1	128	GPR18	189	LYPLA1	250	NDUFB3P4	311	RNF222	372	TKTL1
7	ADAMTS6	68	DCAF16	129	GRB7	190	LYPLA2P1	251	NDUFB8P1	312	RNF38	373	TLR12P
8	[redacted]	69	DCTPP1	130	GRTP1	191	MAGOHB	252	NDUFB8P3	313	RNU3P2	374	TM4SF19
9	AFM	70	DDTL	131	GTF2H2	192	MANBAL	253	NDUFS1	314	RNU4-7P	375	TMBIM6
10	[redacted]	71	DDX1	132	GUCA1A	193	MAP1LC3P	254	NDUFV2P1	315	RNU6-2	376	TMEM155
11	AIMP1	72	DDX50P1	133	GZMM	194	MAPK9	255	NEK5	316	RNU6ATAC5P	377	TMEM202
12	AKAP17A	73	DDX58	134	HAPLN4	195	MBD3L5	256	NFXL1	317	RNU7-39P	378	TMEM213
13	ALDOAP2	74	DDX60	135	HAUS1	196	MEMO1P1	257	NPM1P10	318	RNU7-50P	379	TMPRSS3
14	ALG1	75	DECR2	136	HAUS4	197	METT10D	258	NPM1P12	319	RNU7-72P	380	TMSB4XP1
15	ALPK2	76	DEFB118	137	HERC2P5	198	MEX3A	259	NPM1P19	320	RNU7-77P	381	TP53BP2
16	AMAC1L1	77	DERA	138	HKDC1	199	MFAP5	260	NPW	321	RNY5P6	382	TPM5P
17	AMAC1L2	78	DHX57	139	HLA-U	200	MIAT	261	NSL1	322	SALL1P1	383	TPRX1
18	AP1B1P1	79	DMC1	140	HMGB1P1	201	MIR1256	262	NUBPL	323	SCARNA23	384	TPTE
19	AP1M1	80	DNAJA2	141	HMGB1P12	202	MIR1267	263	NUP214	324	SEC14L3	385	TRADD
20	APOL5	81	DNAJB13	142	HMGB1P9	203	[redacted]	264	NUS1P3	325	SEC63	386	TRAF3IP2
21	ARL9	82	DNAJC12	143	HMGB4	204	MIR138-2	265	NUS1P4	326	SERPINF1	387	TRIM16
22	ASCC1	83	DND1	144	HMGN5	205	MIR220A	266	ODF3L2	327	SF3A3P1	388	TRNAE39P
23	ASS1P12	84	DNM1P28	145	HMOX1	206	MIR302F	267	OR1M4P	328	SFPQ	389	TRNAS29P
24	ASS1P3	85	DPP9	146	HNRPA1L3	207	MIR3123	268	OR2M7	329	SFRP2	390	TRNAV33P
25	ATP5A1	86	EDDM3A	147	HPS4	208	MIR3134	269	OR5BJ1P	330	SGK1	391	TRPV4
26	ATP5A1P9	87	EEF1A1P24	148	HRASLS2	209	MIR4284	270	OR7D11P	331	SH3KBP1	392	TSPAN15
27	ATP5J2P5	88	EEF1A1P32	149	HSD17B11	210	[redacted]	271	OR7E100P	332	SIGLEC26P	393	TTC28
28	ATP6V1G1P4	89	EIF3LP1	150	HSP90AB3P	211	MIR4327	272	OR8K5	333	SKA3	394	TTC39C
29	ATP8B1	90	EIF4A1P12	151	HSPA2	212	MIR548D2	273	OSBPL3	334	SLC12A6	395	TXNL4A
30	AVPI1	91	EIF4BP5	152	IDH1	213	MIR548H3	274	OXGR1	335	SLC25A2	396	TXNRD1
31	BOLA3	92	EXD1	153	IFNA16	214	MIR548O	275	PABPC1P8	336	SLC28A1	397	UQCR11
32	C5AR1	93	F13A1	154	IGKV6-21	215	MIR631	276	PABPC3	337	SLC2A11	398	URB1
33	CABYR	94	FABP5P6	155	IGLV1-62	216	MLKL	277	PARP4	338	SLC2A7	399	USP33
34	CACNB4	95	FAM101A	156	IL28RA	217	MND1	278	PATE4	339	SLC7A2	400	[redacted]
35	CAMKK1	96	FAM138D	157	IL3RA	218	MPP7	279	PDZD3	340	SLFN13	401	VDAC1P6
36	CASS4	97	FAM175B	158	IL3RA	219	MPPE1	280	PFAS	341	SNAP23	402	VENTXP7
37	CC2D2B	98	FAM177A1	159	IMMTP1	220	MPTX	281	PHACTR4	342	SNORA2A	403	VHL
38	CCDC11	99	FAM183B	160	IMP4	221	MPZL3	282	PHKA1	343	SNORA69	404	VN1R46P
39	CCDC60	100	[redacted]99YP	161	IMPDH1P9	222	MRP63P9	283	PIGCP1	344	SNORA72	405	VN1R47P
40	CCDC8	101	FAM92A1	162	INSIG2	223	MRPL50P2	284	PLA2G10	345	SNRNP48	406	VN1R7P
41	CCNB1IP1	102	FBP2	163	INSL3	224	MRPL9	285	PLIN2	346	SNRPA1	407	VN1R8P
42	CD163	103	FBXL17	164	ITGAL	225	MRPS33P2	286	POLR2B	347	SPATA24	408	VN1R9P
43	CDC37P2	104	FBXO40	165	ITGB1P1	226	MS4A4E	287	POLR2F	348	SPC25	409	VNN2
44	CDH26	105	FDPSP3	166	ITIH5	227	MSR1	288	POTEM	349	SPINK6	410	VOPP1
45	CECR1	106	FLJ43879	167	IVD	228	MTCO2P1	289	PPM1F	350	SPOCK2	411	WDR73
46	CES3	107	FLJ45256	168	JMJD6	229	MTHFD1L	290	PPP1R13L	351	SPRR2A	412	WDR87
47	CETN4P	108	FLNB	169	KCNN4	230	MTMR3	291	PPP6CP	352	SRCRB4D	413	WFDC8
48	CHRNA6	109	FLOT1	170	KIF2C	231	MTND2P2	292	PRDM6	353	ST13P2	414	XKRX
49	CKS1BP2	110	FMO11P	171	KLHL28	232	MTRF1L	293	PSME2	354	STARD10	415	YDJC
50	CLASP1	111	FOLR2	172	KRT18P48	233	MX1	294	PTCD2	355	SUCLA2	416	[redacted]
51	CLU	112	FRK	173	KRT23	234	MYL6P3	295	PTK6	356	SUCLA2P1	417	YPEL4
52	CLUAP1	113	FTH1P21	174	KRT26	235	MYO5B	296	PTMAP2	357	SUGT1P2	418	YWHAZP4
53	CN5H6.4	114	GAL3ST3	175	KRT80	236	MZT1	297	PTPLA	358	SULT1C3	419	YWHAZP8
54	COASY	115	GAPDHP72	176	KRT8P23	237	NAA20	298	PTPN21	359	SV2A	420	ZNF333
55	COX4I1P2	116	GCNT6	177	KRTAP19-5	238	NAA25	299	PTPN3	360	SYCP3	421	ZNF543
56	CPP	117	GEMIN8	178	KRTAP4-3	239	NAAA	300	PXMP4	361	SYT5	422	ZNF680
57	CRLF2	118	GEMIN8P3	179	LACRT	240	NANOGP8	301	QDPR	362	TAF7L	423	ZNF70
58	CSAD	119	GGNBP1	180	LATS2	241	NCAPH	302	RANP1	363	TBC1D10C	424	ZNF785
59	CSF2RA	120	GIP	181	LCN15	242	NCRNA00032	303	RBM34	364	TCEA1P1	425	ZRSR1
60	CTCFL	121	GJD4	182	LCN1L1	243	NCRNA00092	304	REC8	365	TCTN3		
61	CTSB	122	GLTSCR2	183	LDHBL1	244	NCRNA00111	305	REXO1L3P	366	TDH		

Degree of methylation							
Grey	>10	Pale yellow	>20	Yellow	>30	Orange	>40
Red	>50	Purple	>60	Blue	>70	Brown	>80

C.

1	*ACLY*	31	*CYCSP16*	61	*HTT*	91	*OR2AT1P*	121	*ST13P17*
2	*ACVR2A*	32	*CYP17A1*	62	*IFNNP1*	92	*OR51H1P*	122	*STARD9*
3	*ADAMTS2*	33	*DGKK*	63	*IGHV1OR15-9*	93	*OR5BB1P*	123	*SYNJ2BP*
4	*ANKMY2*	34	*DLX6*	64	*KIF23*	94	*OR7A17*	124	*TFF1*
5	*ANKRA2*	35	*DMPK*	65	*KRT124P*	95	*OR7E101P*	125	*THAP11*
6	*APBB2*	36	*DPM3*	66	*KRT77*	96	*PAICSP7*	126	*TMED2*
7	*APOBEC1*	37	*ECEL1P2*	67	*KRTAP5-8*	97	*PEBP1P1*	127	*TMEM186*
8	*ARHGEF3*	38	*EIF1AX*	68	*KRTAP6-3*	98	*PELP1*	128	*TMEM221*
9	*ATP5EP1*	39	*EMX2*	69	*LAMC3*	99	*PGF*	129	*TSSK1A*
10	*ATP5LP7*	40	*FAM200A*	70	*LANCL1*	100	*PIH1D1*	130	*TTC22*
11	*BET3L*	41	*FDX1L*	71	*LPPR3*	101	*PMS2L3*	131	*TULP2*
12	*BRP44L*	42	*FOXD4L2*	72	*LST-3TM12*	102	*PMS2L4*	132	*UBA52P9*
13	*C1QL1*	43	*FTH1P25*	73	*LYPD3*	103	*POLR1B*	133	*UBR3*
14	*CALM2P3*	44	*FTLP11*	74	*MBD5*	104	*POU3F2*	134	*UPP1*
15	*CBR3*	45	*GABRR3*	75	*MDS2*	105	*PPEF1*	135	*UTS2D*
16	*CCDC102A*	46	*GAPDHP45*	76	*MIR122*	106	*PTCHD3*	136	*VN1R14P*
17	*CCDC148*	47	*GAPDHP54*	77	*MIR1246*	107	*RABL5*	137	*VN1R76P*
18	*CCDC90B*	48	*GCDH*	78	*MIR1294*	108	*RFT1*	138	*VPS11*
19	*CCL22*	49	*GCNT1*	79	*MIR3179-1*	109	*RHBDL2*	139	*XCL2*
20	*CDAN1*	50	*GGT7*	80	*MIR3187*	110	*RNY4P16*	140	*XRCC2*
21	*CDC25B*	51	*GKAP1*	81	*MIR514B*	111	*RTTN*	141	*ZDHHC18*
22	*CDRT4*	52	*GOT2*	82	*MYCBP2*	112	*SEC23IP*	142	*ZG16*
23	*CHIC1*	53	*GPR53P*	83	*MYLK4*	113	*SELPLG*	143	*ZNF415*
24	*CNOT6LP1*	54	*GPR83*	84	*MYLPF*	114	*SETP2*	144	*ZNF480*
25	*COL4A3BP*	55	*GSDMA*	85	*NBPF15*	115	*SF3A1*	145	*ZNF607*
26	*CPLX3*	56	*HCG22*	86	*NDUFA2*	116	*SNRNP27*	146	*ZNF773*
27	*CRYZL1*	57	*HPSE*	87	*NDUFA4P1*	117	*SNX22*	147	*ZSCAN1*
28	*CT45A5*	58	*HR*	88	*NDUFB3P1*	118	*SPANXN1*		
29	*CTAGE5*	59	*HRG*	89	*NPM1P3*	119	*SPECC1L*		
30	*CTPS*	60	*HSPB3*	90	*NUPR1*	120	*SPRY3*		

Degree of methylation	
Grey	>10
Pale yellow	>20
Yellow	>30
Orange	>40
Red	>50
Purple	>60
Blue	>70
Brown	>80

Figure 2.4 The differential methylation gene lists in an HIV negative and an HIV positive subject. **A.** Venn diagram showing the number of methylated genes found in twin subjects. A total of 648 genes were methylated in twin negative subject including 425 genes which were not methylated in the infected subject. 370 genes were methylated in twin positive including 147 genes not found in the uninfected subjects. **B.** Exclusively de-methylated genes upon HIV infection. **C.** Exclusively methylated genes upon HIV infection. The degree of methylation was reflected by the number of sequence reads mapped to the same gene. Genes methylated in both subjects, irrespective of the degree of methylation were not included in B and C.

2.3.3.2 Gene clustering suggested the pathways involved by the methylated genes

The gene functions were annotated using The Database for Annotation, Visualization and Integrated Discovery (DAVID) (http://david.abcc.ncifcrf.gov/) (Huang da *et al.*, 2009). The functional analysis according to the Gene Ontology (GO) using the annotation sources GOTERM-BP (biological process), and GOTERM-MF (Molecular functions) identified functional categories were given in Table 2.4 and Table 2.5. The list was limited with p-value score of functional cluster of p-value <0.1. The HIV-associated de-methylated gene can be categorized into functional groups of mitosis phase, lipid transport, unfolded proteins binding, RNA splicing, etc. The HIV-associated hype-methylated are mainly found in the pathways of cell motility and nerve development.

Table 2.4 Clusters of HIV-associated de-methylated genes (P-value < 0.1).

Category	Count	P-Value	Genes
Mitotic phase	16	1.13E-04	*HAUS4, HAUS1, MND1, LATS2, KIF2C, SPC25, NCAPH, REC8, HSPA2, SYCP3, TIMELESS, NSL1, SKA3, CLASP1, DMC1, TXNL4A*
Lipid transport	10	3.02E-04	*ABCG8, KCNN4, MSR1, LPA, OSBPL3, PLIN2, PLA2G10, CLU, ATP8B1, APOL5*
Unfolded protein binding	7	0.0085	*HSPA2, DNAJC12, HSP90AB3P, DNAJB13, NDUFAF1, SEC63, DNAJA2*
RNA splicing	11	0.0097	*SNRPA1, POLR2F, SNRNP48, JMJD6, SFPQ, DDX1, MAGOHB, GEMIN8, HNRPA1L3, POLR2B, TXNL4A*
Nucleoside binding	37	0.0125	*COASY, GAL3ST3, FRK, ADCY7, ACSS2, POLR2B, PFAS, CAMKK1, MTHFD1L, LATS2, KIF2C, LONP2, HSPA2, IVD, DDX60, PTK6, ATP8B1, MLKL, SUCLA2, SLC28A1, MYO5B, DHX57, SGK1, ALPK2, HKDC1, DDX1, NUBPL, ABCG8, DDX58, HSP90AB3P, SLFN13, MAPK9, TEP1, TXNRD1, ATP5A1, DMC1, NEK5*
Positive regulation of foam cell differentiation	3	0.0152	*MSR1, PLA2G10, MAPK9*
Cell volume homeostasis	3	0.0176	*SLC12A6, KCNN4, TRPV4*
Extracellular region	44	0.0181	*HSD17B11, INSL3, MSR1, ADAMTS13, SPOCK2, CRLF2, F13A1, CLU, TMEM155, SPINK6, LPA, FOLR2, PLIN2, RLN1, LILRA3, HMOX1, GP2, LCN1L1, DEFB118, ITIH5, GNL1, CSF2RA, LCN15, HAPLN4, GIP, EDDM3A, AIMP1, PLA2G10, CECR1, CD163, GZMM, WFDC8, ADAMTS6, AFM, SERPINF1, SFRP2, NPW, LACRT, SRCRB4D, CTSB, IFNA16, MFAP5, APOL5, PATE4*
Src homology-3 domain	9	0.0182	*FRK, CASS4, TP53BP2, PTK6, SH3KBP1, MPP7, CACNB4, TJP3, PPP1R13L*
Scavenger receptor activity	4	0.0333	*MSR1, SRCRB4D, TMPRSS3, CD163*
Protein amino acid dephosphorylation	6	0.0475	*PPM1F, MTMR3, PTPN3, PTPLA, PTPN21, TPTE*
Cellular respiration	5	0.0556	*UQCR11, IDH1, SUCLA2, NDUFAF1, NDUFS1*
Purine nucleotide metabolic process	7	0.0581	*ADCY7, ATP8B1, CECR1, CACNB4, ATP5A1, NDUFS1, PFAS*
Purine nucleotide metabolic process	7	0.0581	*ADCY7, ATP8B1, CECR1, CACNB4, ATP5A1, NDUFS1, PFAS*
Regulation of body fluid levels	6	0.0582	*KCNN4, LPA, ADAMTS13, F13A1, LACRT, HPS4*
ATPase activity, coupled	9	0.0584	*DDX58, LONP2, DDX60, ATP8B1, DDX1, ATP5A1, DMC1, GTF2H2, DHX57*
Carboxylic acid transport	6	0.0671	*PLIN2, FOLR2, SLC7A2, ATP8B1, SLC25A2, CACNB4*
Cell death	17	0.0672	*SGK1, AIMP1, TP53BP2, TMBIM6, CLU, DNAJB13, PPP1R13L, TRADD, PPM1F, GZMM, TRAF3IP2, JMJD6, HMOX1, SH3KBP1, TCTN3, PARP4, NDUFS1*
Transcription initiation from RNA polymerase II promoter	4	0.0794	*POLR2F, POLR2B, GTF2H2, TAF7L*

Table 2.5 Clusters of HIV-associated hyper-methylated genes (P-value< 0.1).

Category	Count	P-Value	Genes
Cell motion	7	0.0505	*CCL22, LYPD3, EMX2, POU3F2, APBB2, SELPLG, MYCBP2*
Telencephalon development	3	0.0516	*HTT, EMX2, POU3F2*
Metal-binding	24	0.0541	*APOBEC1, ANKMY2, LANCL1, HR, DGKK, UBR3, ZDHHC18, ACLY, POLR1B, ZSCAN1, MYCBP2, PMS2L3, DMPK, ACVR2A, CYP17A1, PPEF1, ZNF773, ZNF415, FDX1L, ZNF480, VPS11, ZNF607, THAP11, ADAMTS2*
Zinc-finger	15	0.0892	*ANKMY2, HR, DGKK, UBR3, ZDHHC18, POLR1B, ZSCAN1, PMS2L3, MYCBP2, ZNF773, ZNF415, ZNF480, VPS11, ZNF607, THAP11*

2.3.4 Analysis of genes associated with HIV pathogenesis

2.3.4.1 Expression level of differential methylated genes

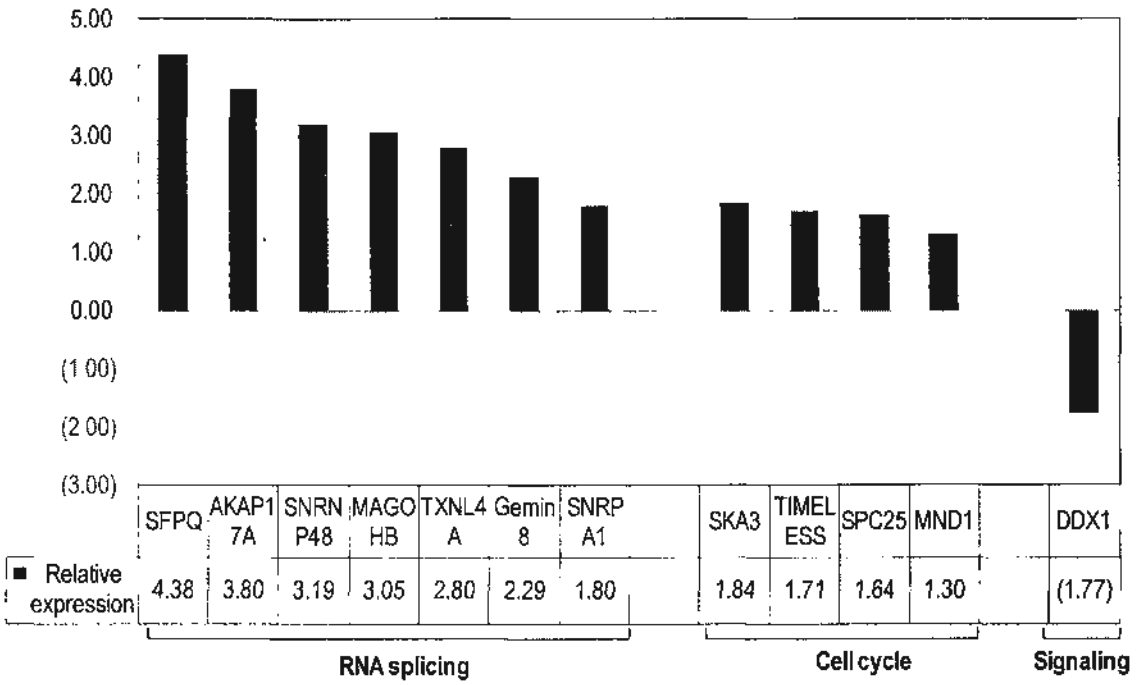
We examined in detail the list of de-methylated and hyper-methylated genes involved in HIV pathogenesis to select genes for validation. These include genes in RNA splicing, cell cycle and nerve development. Figure 2.5 shows the relative expression level of several selected de-methylated genes in H9-HIV_{IIIB} compared to its parental cell line H9. These include RNA splicing genes *SFPQ* (4.38 fold), *AKAP17A* (3.80), *SNRANP48* (3.19), *MAGOHB* (3.05), *TXNL4A* (2.80), *Gemin8* (2.29) and *SNRPA1* (1.80), cell cycle genes *SKA3* (1.84), *TIMELESS* (1.71), *SPC25* (1.64) and *MND1* (1.30) and signaling protein gene *DDX1* (-1.77). However only the genes involved in RNA splicing were up-regulated with a relative expression of more than two folds.

These include genes for splicing factors *SFPQ* and *AKAP17A*. SFPQ is a proline/glutamine-rich binding protein whilst AKAP17A is A-kinase (PRKA) anchor protein for arginine/serine-rich binding. Spliceosome complex members were also up-regulated including *SNRNP48* (small nuclear ribonucleoprotein 48kDa (U11/U12))

in the minor spliceosome complex and *TXNL4A* (2.80) (Spliceosomal U5 snRNP-specific) for the major and minor spliceosome complex. *Gemin8* (gem (nuclear organelle) associated protein) which was a member of SMN complex was also up-regulated for 2.29 folds. Exon-exon junction complex component mago-nashi homolog B (*Drosophila*) *MAGOHB* was up-regulated by 3.05 folds.

Despite the fact that hyper-methylation are usually associated with gene silencing, several selected hyper-methylated genes in this study did not show a down-regulated transcription level in real-time PCR (Figure 2.6). The genes with an increase in their transcriptional level included Huntington's disease-associated gene Huntingtin *HTT* (2.05 fold), cell signaling proteins NADH dehydrogenase (ubiquinone) gene *NDUFA2* (6.06), MYC binding protein gene *MYCBP2* (4.75) and Rho guanine nucleotide exchange factor (GEF) gene *ARHGEF3* (2.19). Two other selected genes *DMPK* (1.99) and *MBD5* (-1.13) did not show significant changes in their transcription level for more than two folds.

A.

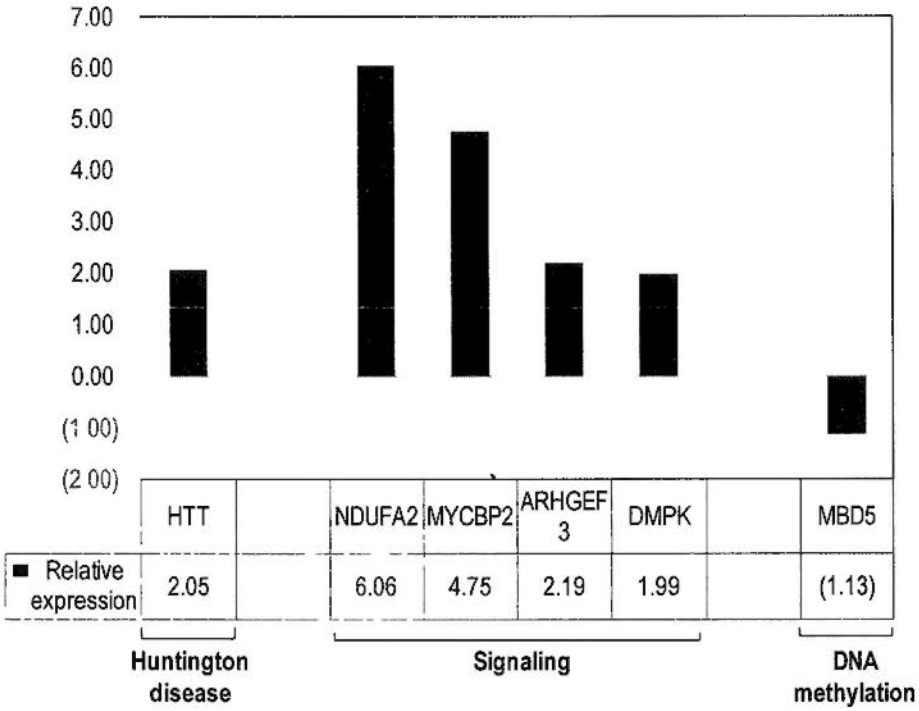


B.

Category	Gene	Relative expression	Description
RNA splicing	*SFPQ*	4.38	Splicing factor proline/glutamine-rich binding protein
	AKAP17A	3.80	A kinase (PRKA) anchor protein 17, SR protein
	SNRNP48	3.19	Small nuclear ribonucleoprotein 48kDa (U11/U12)
	MAGOHB	3.05	Mago-nashi homolog B (Drosophila), Exon junction complex
	TXNL4A	2.80	Spliceosomal U5 snRNP-specific
	Gemin8	2.29	Gem (nuclear organelle) associated protein 8
	SNRPA1	1.80	Small nuclear ribonucleoprotein polypeptide A'
Cell cycle	*SKA3*	1.84	Spindle and kinetochore associated complex subunit 3
	TIMELESS	1.71	Timeless homolog (Drosophila)
	SPC25	1.64	NDC80 kinetochore complex component, homolog *(S. cerevisiae)*
	MND1	1.30	Meiotic nuclear divisions 1 homolog (*S. cerevisiae*)
Signaling	*DDX1*	-1.77	DEAD (Asp-Glu-Ala-Asp) box polypeptide 1

Figure 2.5 Differential expression of the de-methylated genes in HIV_{IIIB}-expressing cell line compared to that of its parental H9 cell line. **A.** The mRNA expression level of genes selected from three categories were determined and normalized with GAPDH mRNA. Genes in RNA splicing and cell cycle were up-regulated while *DDX1* for signal transduction was down-regulated. **B.** Description of the detected genes.

A.



B.

Category	Gene	Relative expression	Description
Huntington's disease	*HTT*	2.05	Huntingtin
Signaling	*NDUFA2*	6.06	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 2, 8kDa
	MYCBP2	4.75	MYC binding protein 2
	ARHGEF3	2.19	Rho guanine nucleotide exchange factor (GEF) 3
	DMPK	1.99	Dystrophia myotonica-protein kinase
Methylation	*MBD5*	-1.13	Methyl-CpG binding domain protein 5

Figure 2.6 Differential expression of the hyper-methylated genes in HIV_{IIIB}-expressing cell line compared to that of its parental H9 cell line. **A.** Huntington disease gene hungtintin *HTT* was up-regulated. Several genes of signal transduction were up-regulated. No significant change in mRNA of *MBD5* was observed. **B.** Description of the detected genes.

2.3.4.2 Methylation may explain the transcriptional level of HIV regulated genes reported in other studies

Although there are several elegant studies on HIV dependency host factors and HIV-induced mRNA expression, there are few studies on HIV induced methylated genes, making it difficult to judge whether our results are representative. Hence, to align our study with other studies, we have scanned through the hypo- and hyper-methylated genes list to select which had been reported by at least one other research group (Brass *et al.*, 2008; de la Fuente *et al.*, 2002; Park *et al.*, 2007; Rotger *et al.*, 2010; Solis *et al.*, 2006; van 't Wout *et al.*, 2003; van 't Wout AB, 2003). We have also included a study on host dependency factors (HDF) for HIV replication for a wider comparison of virus-related genes (Brass *et al.*, 2008). These HDF were essential for the HIV survival such that their silence (by RNA interference) hindered the viral replication. Owing to the limited amount of twin samples, we have adopted a lab-adapted HIV model for the validation of the relative mRNA expression of these selected methylation genes. This is based on the rationale that de-methylation or hypo-methylation usually cause an increase in transcription level whereas hyper-methylation cause genes silencing, though there is no strict relationship in between. The validation results together with the real-time PCR results described in section 2.3.3.1, were compared to the results of other mRNA studies. Table 2.6 and 2.7 showed the genes which were found in other studies and in our methylation study. Some of the genes found in our study have not yet been described in any other previous studies. These genes include de-methylated genes *SFPQ*, *AKAP17A*, *SNRNP48*, *MAGOHB*, *Gemin8*, *SNRPA1*, *SKA3*, *DDX1*, and hyper-methylated genes *HTT*, *MYCBP2*, *ARHGEF3*, *DMPK* and *MBD5*. Differential degree of DNA methylation may be one of the reasons for the

differential expression levels of the genes reported in other studies. As indicated in bold in Table 2.6, *DDX60*, *FBX40*, *MND1*, *PTPLA*, *REPIN1*, etc, were found up-regulated in other studies and found de-methylated in our study. Similarly, some genes were found down-regulated in other studies and hyper-methylated in our study, as shown in Table 2.7.

Table 2.6 Differential expression levels of hypo-methylated genes in this study and other studies.

Host genes in this study	Description	Real time PCR validation in this study		Studies by others	
		Correlation with HIV	Relative expression	Paper	Correlation with HIV
ABCG8	ATP-binding cassette	N.D.		Rogter, 2010	down-regulated(ABCB1)
CCDC11	coiled-coil domain-containing protein	N.D.		Rogter, 2010	down-regulated (CCDC13)
CCDC60	coiled-coil domain-containing protein	N.D.		Rogter, 2010	down-regulated (CCDC13)
CCDC8	coiled-coil domain-containing protein	N.D.		Rogter, 2010	down-regulated (CCDC13)
CCNB1IP1	**cyclin B1 interacting protein 1**	**N.D.**		**Rogter, 2010**	**down-regulated**
DDX60	**DEAD (Asp-Glu-Ala-Asp) box polypeptide 60**	**N.D.**		**Rogter, 2010**	**up-regulated**
DECR2	2,4-dienoyl CoA reductase	N.D.		Van't Wount, 2003	up-regulated (DECR1)
DPP9	dipeptidyl-peptidase	N.D.		Brass, 2008	HDF (DPP4)
FABP5P6	fatty acid binding protein	N.D.		Park, 2007	up-regulated (FABP4)
FBX40	**F-box protein 40**	N.D.		**Rogter, 2010**	**down-regulated**
FDPSP3	farnesyl diphosphate synthase	N.D.		Van't Wount, 2003	up-regulated(FDPS)
FMO11P	flavin containing monooxygenase	N.D.		Park, 2007	up-regulated (FMO3)
HSD17B11	hydroxysteroid (17-beta) dehydrogenase	N.D.		Van't Wount, 2003	up-regulated (HSD17B7)
HSP90AB3P	heat shock protein 90kDa alpha (cytosolic)	N.D.		Brass, 2008	HDF (HSP90AA1)
HSPA2	heat shock 70kDa protein	N.D.		Brass, 2008	HDF (HSPA1A)
INSIG2	insulin induced gene	N.D.		Van't Wount, 2003	up-regulated (INSIG1)
ITGB1P1	integrin beta	N.D.		Park, 2007	up-regulated (ITGB3)
KIF2C	kinesin family member	N.D.		Rogter, 2010	down-regulated (KIF5C)
MAP1LC3P	microtubule-associated protein 1 light chain 3 pseudogene	N.D.		Brass, 2008	HDF

Host genes in this study	Description	Real time PCR validation in this study		Studies by others	
		Correlation with HIV	Relative expression	Paper	Correlation with HIV
MAPK8	mitogen-activated protein kinase	N.D.		Solis, 2006	up-regulated (MAPK1)
MND1	**meiotic nuclear divisions 1 homolog (S. cerevisiae)**	**no change**	**0.93 fold**	**Rogter, 2010**	**up-regulated**
NUP214	nucleoporin	N.D.		Brass, 2008	HDF (NUP62, NUP98)
PABPC3	polyA binding protein	N.D.		Brass, 2008	HDF (PABPC1)
PHKA1	phosphorylase kinase	N.D.		Solis, 2006	up-regulated (PHKA2)
POLR2B	polymerase (RNA) II (DNA directed)	N.D.		Rogter, 2010	down-regulated (POLR2J4)
POLR2F	polymerase (RNA) II (DNA directed)	N.D.		Rogter, 2010	down-regulated (POLR2J4)
PRDM6	PR domain	N.D.		Rogter, 2010	down-regulated (PRDM2)
PSME2	proteasome (prosome, macropain)	N.D.		Brass, 2008	HDF
PTPLA	**protein tyrosine phosphatase-like (proline instead of catalytic arginine), member A**	**up-regulated**	**2.95 fold**	**Solis, 2006**	**up-regulated**
RANP1	RAN, member RAS oncogene family	N.D.		Brass, 2008	HDF (RAN)
REPIN1	**replication initiator 1**	**up-regulated**	**2.79 fold**	**Solis, 2006**	**up-regulated**
SGK1	serum/glucocorticoid regulated kinase	slightly up-regulated	1.84 fold	Rogter, 2010	down-regulated (SGK3)
SPC25	**SPC25, NDC80 kinetochore complex component, homolog (S. cerevisiae)**	**slightly up-regulated**	**1.64fold**	**Rogter, 2010**	**up-regulated**
ST13P2	suppression of tumorigenicity	N.D.		Solis, 2006	up-regulated (ST5)
TAF7L	TAF7 RNA polymerase II	N.D.		De la Fuente, 2002	up-regulated (TAF7)
TIMELESS	**timeless homolog (Drosophila)**	**slightly up-regulated**	**1.71 fold**	**Rogter, 2010**	**up-regulated**

Host genes in this study	Description	Real time PCR validation in this study		Studies by others	
		Correlation with HIV	Relative expression	Paper	Correlation with HIV
TIMM23	translocase of inner mitochondrial membrane	N.D.		Solis, 2006	up-regulated (TIMM44)
TMSB4XP2	thymosin beta	N.D.		Van't Wount, 2003	up-regulated (TMSB10)
TTC28	tetratricopeptide repeat domain	N.D.		Rogter, 2010	down-regulated (TTC8)
TTC39C	tetratricopeptide repeat domain	N.D.		Rogter, 2010	down-regulated (TTC8)
TXNL4A	thioredoxin	up-regulated	2.80 fold	Brass, 2008	HDF (TXN)

Foot note:

Bold: Exactly the same genes were found HYPO-methylated in HIV infected subject as well as in other studies.

Non-bold: Genes in the same family were found in other studies.

HDF: Host dependency factor (Brass *et al.*, 2008)

Table 2.7 Differential transcription levels of hyper-methylated genes in this study and other studies.

Host genes in this study	Description	Validation in this study		Studies by others	
		Correlation with HIV	Relative expression	Paper	Correlation with HIV
APOBEC1	**apolipoprotein B mRNA editing enzyme, catalytic polypeptide 1**			**Park, 2007, Rogter, 2010**	**up-regulated**
CDC25B	cell division cycle 25 homolog B (*S. pombe*)			Rogter, 2010	up-regulated (CDC25A)
DNAJB13	DnaJ (Hsp40) homolog, subfamily B			Solis, 2006	down-regulated (DNAJB2)
GABRR3	gamma-aminobutyric acid (GABA) receptor			De la Fuente, 2002	down-regulated (GABRA5)
IK	**IK cytokine, down-regulator of HLA II**	**no change**	**-1.12**	**Solis, 2006**	**down-regulated**
KIF23	**kinesin family**			**Rogter, 2010**	**up-regulated**
NDUFA2	**NADH dehydrogenase (ubiquinone) 1 alpha subcomplex**	**up-regulated**	**6.06**	**Rogter, 2010**	**up-regulated (NDUFA9)**
POU3F2	DNA octamer-binding proteins			Brass, 2008	HDF (POU2F1)
RABL5	RAS oncogene family			Rogter, 2010, Brass, 2008	up-regulated (RAB8A); HDF (RAB6)
SETP2	SET nuclear oncogene			Van't Wount, 2003	down-regulated (SET)
VPS11	vacuolar protein			Brass, 2008	HDF (VPS53)
XRCC2	X-ray repair complementing defective repair in Chinese hamster cells			De la Fuente, 2002	down-regulated (XRCC9)

Foot note:

Bold: Exactly the same genes were found HYPER-methylated in HIV infected subject as well as in other studies.

Non-bold: Genes in the same family were found in other studies.

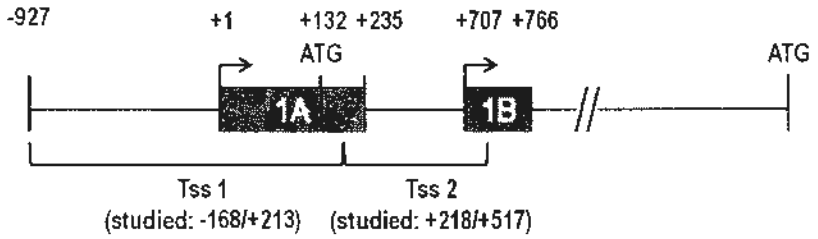
HDF: Host dependency factor (Brass *et al.*, 2008)

2.3.5 Promoter study of DNMT3A suggested for indirect effect of regulation by HIV

2.3.5.1 Methylation status analysis of DNMT3A promoter

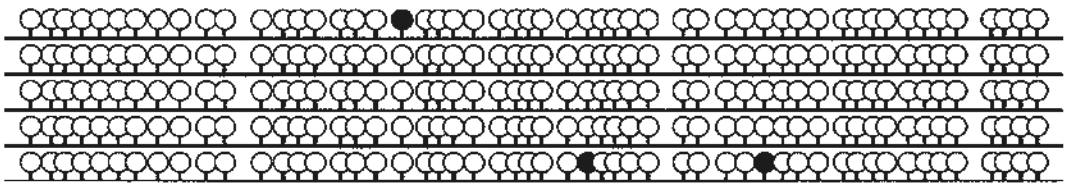
In section 2.3.1, an increase level of *DNMT3A* transcripts and protein was described. We have therefore studied the *DNMT3A* promoter in order to dissect the molecular basis of an up-regulated *DNMT3A* transcription and its relationship with HIV infection. An increase in *DNMT3A* transcription may be attributed to hypo-methylation of its gene promoter region, thus we examined the methylation status of its promoter in H9 cells and H9/HIV_{IIIB}. The amplified region of the bisulfite-treated transcriptional start sites (TSS) was cloned into pGEM-T. It was found that *DNMT3A* has two TSS (Yanagisawa et al., 2002) (Figure 2.7A). One TSS locates in front of exon 1A whereas the other locates at region between 1A and 1B. The methylation status analyzed by bisulfite-sequencing showed that *DNMT3A* TSS2 was unmethylated in both normal H9 T-lymphocyte cell line and H9/HIV_{IIIB} cell line. The sequencing results also showed that the TSS2 sequences in both cell lines are exactly the same as the reference sequence from NCBI (not shown). Figure 2.7B shows the methylation status of each CpG dinucleotide in TSS2, with black representing the methylated CpG whereas white as the unmethylated CpG. Each row represents an individual clone. The TSS1 was also unmethylated in H9 cell-line (Figure 3C). Unfortunately, we failed in determining the methylation of status of *DNMT3A* TSS1 for H9/HIV_{IIIB} after using different pairs of primers and methods. Nevertheless, by deduction as using the results for TSS2, TSS1 in HIV_{IIIB}-expressing cells was thought to be unmethylated.

A.

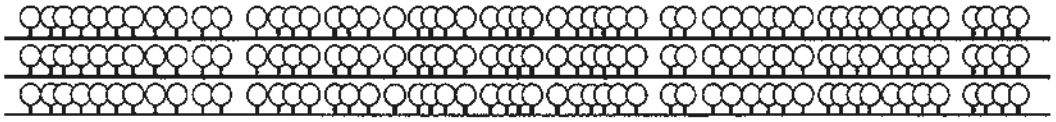


B. TSS2

H9 DNMT3A +218/+517



H9/HIV$_{IIIB}$ DNMT3A +218/+517



C. TSS1

H9 DNMT3A -168/+213

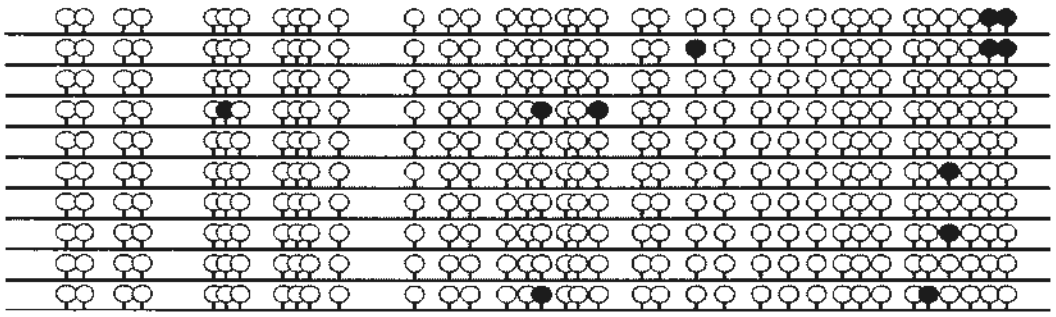
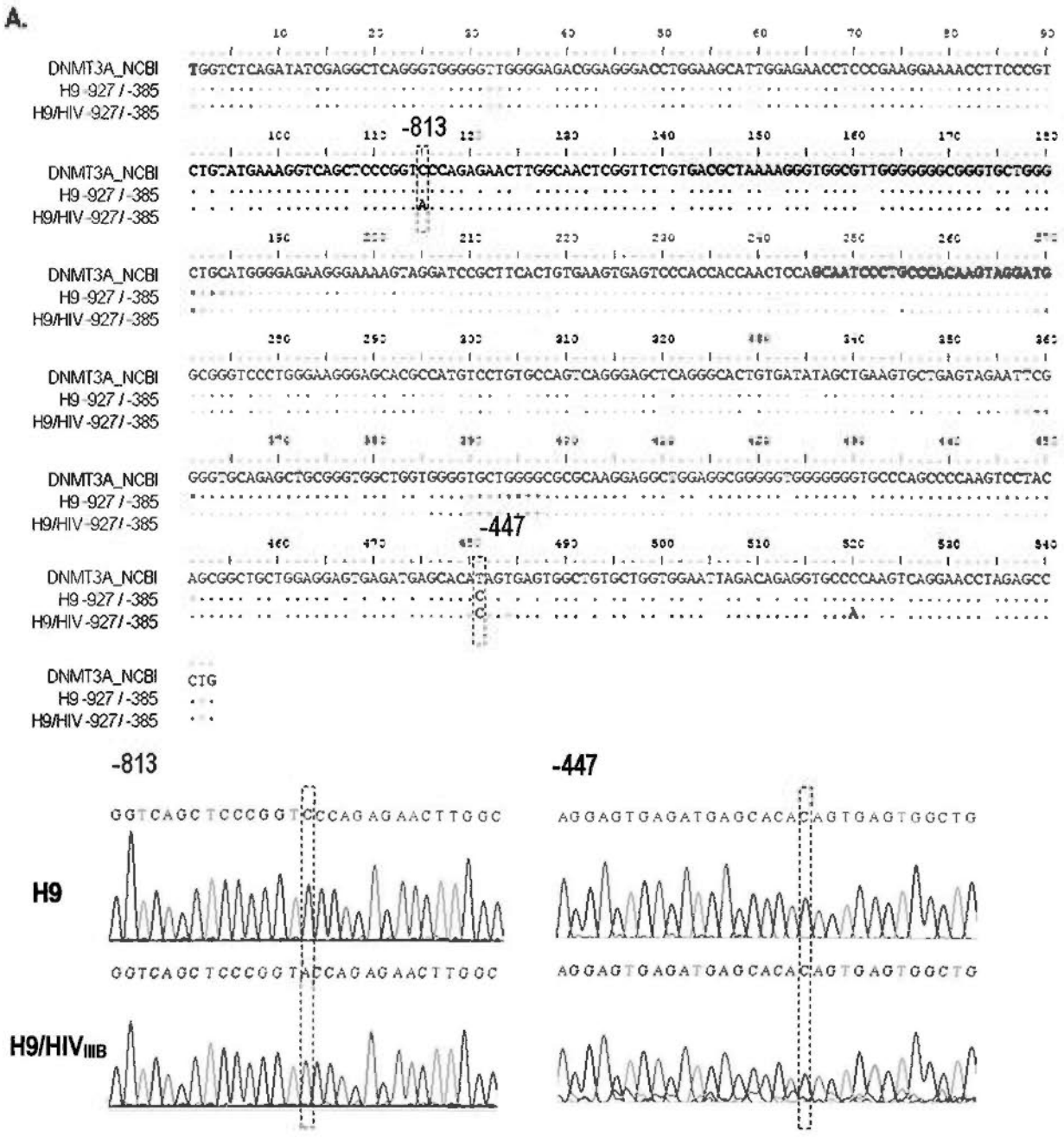


Figure 2.7 Effect of HIV$_{IIIB}$ expression on the methylation status of *DNMT3A* promoter. **A.** The schematic structure of two reported transcription start site (TSS) of DNMT3A gene. **B.** *DNMT3A* promoter is unmethylated in normal H9 parental cell-line at both TSS sites. Black, methylated CpG; White, unmethylated CpG. **C.** Expression of HIV$_{IIIB}$ in H9 does not alter the methylation status of this promoter at TSS2. We unfortunately did not get the methylation status of the TSS1 in this virus-expressing cell-line.

2.3.5.2 Promoter sequence analysis of DNMT3A

In section 2.3.5.1, we failed to amplify *DNMT3A* TSS1 region in HIV_{IIIB}. We therefore suspected the *DNMT3A* promoter sequence might have been mutated in the HIV-expression cell line. The *DNMT3A* promoter sequence in H9 and $H9/HIV_{IIIB}$ was evaluated by molecular cloning. The coordinates of the sequence were annotated with reference to the first nucleotide of transcription which was annotated as +1. As illustrated in Figure 2.8, *DNMT3A* promoter possessed a point mutation C-813A in $H9/HIV_{IIIB}$ cell-line but not in H9 parental cell line. On the other hand, *DNMT3A* promoter was mutated in both cell lines for T-447C. However, no point mutation was identified in *DNMT3A* (-429/+235) which covered the TSS1 sequence (-168/+235) in the methylation analysis, suggesting for alternative reasons of failure in the bisulfite sequencing.


A.
DNMT3A_NCBI
H9-927/-385
H9/HIV-927/-385
-813
-447
-813
-447
H9
H9/HIV$_{IIIB}$

B.

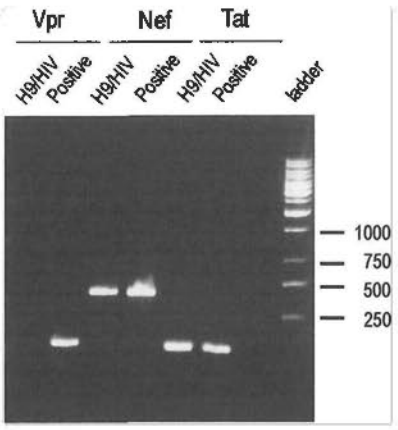
Figure 2.8 Sequence analysis of *DNMT3A* promoter TSS1. *DNMT3A* promoter was cloned into a pGL3-basic vector and sequenced. **A.** Two point mutations C-813A and T-447C were found in the *DNMT3A* -927/-384 regions in H9 and H9/HIV$_{IIIB}$ cells. **B.** The sequence of *DNMT3A* -429/+235 in H9 and H9/HIV$_{IIIB}$ cells. No point mutation was identified including the TSS1 -168/+235 region.

2.3.6 Alteration of DNMT3A expression by HIV proteins

2.3.6.1 Integrity of HIV protein in lab-adapted HIV_{IIIB} strain

It is suggested HIV proteins can transactivate the cellular protein expression. The HIV proteins Vpr, Tat and Nef are suggested to possess such transactivation activity. As described in section 2.3.1, differential expression of DNMT3A was observed in H9 cells expressing HIV_{IIIB}, we therefore ought to resolve the relationship between HIV proteins and up-regulation of this DNA methyltranferase. To begin with, we test the integrity of the HIV proteins in the HIV_{IIIB} strain in the $H9/HIV_{IIIB}$ cell line because it has been suggested that this HIV_{IIIB} did not possess a wild-type Vpr and Nef protein (Brass *et al.*, 2008). Using the primers flanking the full-length *Vpr*, *Tat* and *Nef* which also served as positive controls, we failed to get a PCR product for *Vpr* (Figure 2.9). This suggested that the *Vpr* gene in HIV_{IIIB} was truncated. The successfully amplified *Tat* and *Nef* genes were sequenced for any mutations or frame-shift. *Nef* in HIV_{IIIB} was found to have several deletion and non-synonymous mutation points but did not lead to truncation of the translated Nef protein.

A



B

```
                    10        20        30        40        50        60
                    ....|....|....|....|....|....|....|....|....|....|....|....|
Nef_ref             MGGKWSKRSRDAMDAIRERMRRAEPRVEPAAAGVGAVSRDLEQRGAITSSN--HGNADCV
Nef HIV-IIIB        LGGKWSKSSVVGWPAVRERMRRAEP----AADGVGAASRDLEKHGAITSSNTAANNAACA

                    70        80        90        100       110       120
                    ....|....|....|....|....|....|....|....|....|....|....|....|
Nef_ref             WVESQEEEEEVGFPVRPQVPLRPMTYKEAFDLSHFLKEKGGLDGLVYSQKRKDILDLWVY
Nef HIV-IIIB        WLEAQEEEE-VGFPVTPQVPLRPMTYKAAVDLSHFLKEKGGLEGLIHSQRRQDILDLWIY

                    130       140       150       160       170       180
                    ....|....|....|....|....|....|....|....|....|....|....|....|
Nef_ref             HTQGFFPDWQCYTPGPGVRYPLTFGWCFKLVPLEPEEVEKANEGENNCLLHPISQHGIED
Nef HIV-IIIB        HTQGYFPDWQNYTPGPGIRYPLTFGWCYKLVPVEQEKVEEANEGENTRLLHPVSLHGMDD

                    190       200       210
                    ....|....|....|....|....|....|.
Nef_ref             PEREVLVWKFDSHLALRHMARELHPEYYKNC
Nef HIV-IIIB        PEREVLEWRFDSRLAFHHMARELHPEYYKNC

                    10        20        30        40        50        60
                    ....|....|....|....|....|....|....|....|....|....|....|....|
Tat_ref             MEPVDPRLEPWKHPGSQPKTACTRCYCKKCCLHCQVCFMQKGLGISYGRKKRRQRRRAPP
Tat HIV-IIIB        LEPVDPNLEPWKHPGSQPRTACNNCYCKKCCFHCYACFTRKGLGISYGRKKRRQRRRAPQ

                    70        80        90        100
                    ....|....|....|....|....|....|....|....|..
Tat_ref             ESQNHQVSLSKQPTSQVRGDPTGPEESKKKVETKTETDPVP*
Tat HIV-IIIB        DSQTHQVSLSKQPASQSRGDPTGPTESKKKVER--ETDPVP*
```

Figure 2.9 Analysis of *Vpr*, *Tat* and *Nef* in H9/HIV$_{IIIB}$. **A.** The integrity of *Vpr*, *Tat* and *Nef* genes was determined by PCR using primers flanking for the full-length of the gene. No full length Vpr product was obtained from HIV$_{IIIB}$. **B.** The sequences of *Tat* and *Nef* in HIV$_{IIIB}$ were compared to the reference sequence. No truncation on Tat or Nef protein was found.

2.3.6.2 DNMT3A was not up-regulated by expressing Tat, Vpr or Nef alone

To elucidate the effect of HIV infection on host gene methylation, we have thus studied the effect of over-expressing HIV proteins in T-lymphocyte on *DNMT3A* expression. Instead of using the genes cloned from HIV_{IIIB}, we have cloned the more representing wild-type Nef and Tat genes into a modified pcDNA3.1(+) expression system which expresses the a 3X-FLAG tag at the N-terminal of viral proteins (Figure 2.10A). HIV wild-type Vpr was included for comparison. The recombinant constructs were transfected into HEK293 cells. Figure 2.10B showed the positive detection of the 3X FLAG tagged Vpr, Nef and Tat proteins at 48 hours post-transfection. β-actin (ACTB) was detected for the loading input. Figure 2.10B shows the DNMT3A protein level detected by western blot upon expression of different HIV proteins. Cells transfected with the empty vector was used as a control for the endogenous level of *DNMT3A* in HEK293 cells. Expression of wild-type Vpr proteins apparently increased the DNMT3A expression at protein level but not at transcriptional level (Figure 2.10B & C). Nef and Tat did not show an obvious effect on the expression of DNMT3A.

We further detected the transcription of *SFPQ* in HEK293 cells expressing the HIV protein (Figure 2.11). *SFPQ* showed the highest differential expression level upon HIV expression, as described in section 2.3.4, which meant it may react more sensitively to viral protein expression. However, no significant differential expression of *SFPQ* was found upon expression of different viral proteins.

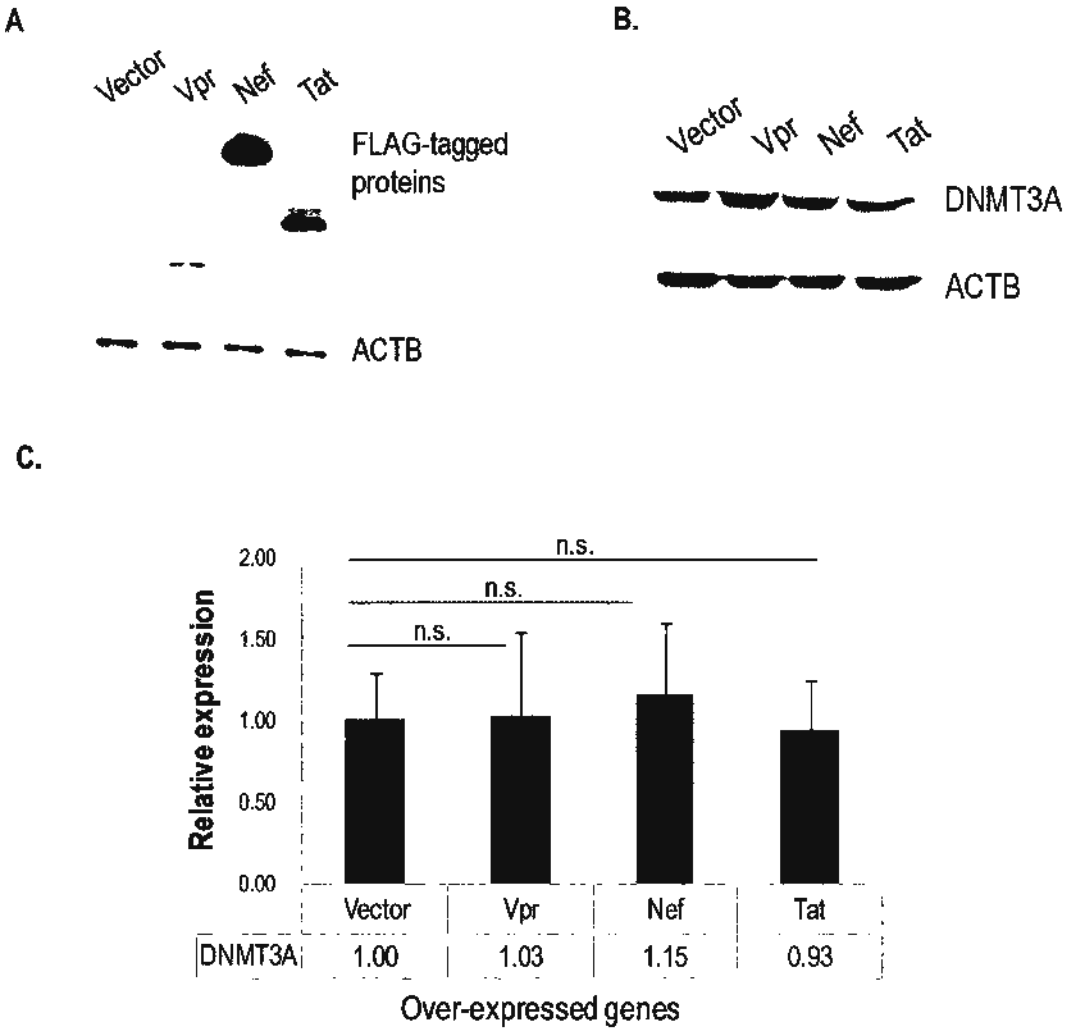


Figure 2.10 Expression of DNMT3A upon expression of HIV proteins. **A.** Western blot analysis was used to detect the over-expression of Vpr, Nef and Tat in HEK293 cells in parallel with the vector only as a control. **B.** Analysis of DNMT3A protein expression in HEK293 over-expressing different HIV proteins. **C.** Relative expression level of *DNMT3A* transcripts in the same cells over-expressing HIV proteins, in comparison to the empty vector control.

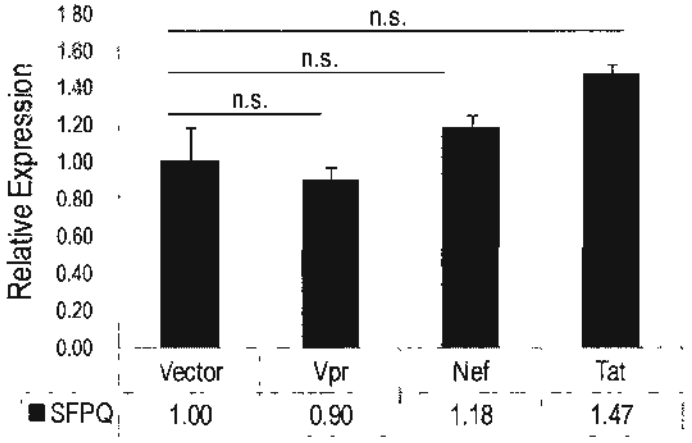


Figure 2.11 Expression of *SFPQ* upon expression of HIV proteins. The transcription level of *SFPQ* was determined in real-time PCR. No significant correlation was found between single HIV protein expression and the mRNA level of *SFPQ*.

2.3.7 Analysis of methylation of HIV promoter

DNA methytransferases have been found significantly up-regulated by hepatitis B virus (HBV), leading to a hyper-methylated HBV genome. The HBV, on the other hand, causes hyper-methylation of the host genome (Vivekanandan *et al.*, 2010). In other words, viruses may directly methylate viral genome by inducing the host DNA methylation mediators. We there determined the methylation of the HIV genome in H9/HIV_{IIIB} cell line which has an up-regulated DNMT3A expression level (as shown in section 2.3.1). The unknown promoter sequence of HIV_{IIIB} was determined by using degenerated PCR. The reference sequences for HIV were retrieved from the HIV reference genome set on NCBI. The LTR promoter sequences of different viral genotyped were aligned. Degenerated primers were designed based on the alignment and were used to amplify the LTR from HIV_{IIIB}. The amplified LTR was cloned into pGEM-T vector for sequencing. The LTR sequence of HIV_{IIIB} was shown in Figure 2.12A. The CpG dinucleotides including those lie on some of the transcription factor-binding region, were marked in grey.

Bisulfite sequencing of the LTR showed that the LTR in HIV_{IIIB} was not methylated (Figure 2.11B).

A.

CACCAGGGCCGGGGGTCAGATACCCACTGACCTTTGGATGGTGCTACAAGCTA
GTACCAGTTGAGCCAGAGAAGTTAGAAGAAGCCAATAAAGGAGAGAACACCAG
CTTGTTACACCCTGTGAGCCTGCATGGGATGGATGACCCGGAGAGAGAAGTGT
TAGAGTGGAGGTTTGACAGCCGCCTAGCATTTCATCACATGGCCCGAGAGCTG
CATCCGGAGTACTTCAAGAACTGCTGATATCGAGCTTGCTACAAGGGACTTTCC
GCTGGGGACTTTCCAGGGAGGCGTGGCCTGGGCGGGACTGGGGAGTGGCGA
GCCCTCAGATCCTGCATATAAGCAGC

NF-κB
Sp1

B.

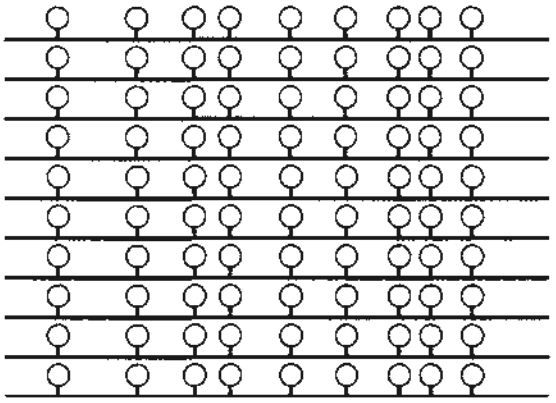


Figure 2.12 Effect of an up-regulated *DNMT3A* expression on methylation status of HIV promoter LTR. **A.** Sequence of HIV_{IIIB} was determined using degenerative PCR. It contains two NF-κB binding sites and two SP1 sites on which laid the CpG dinucleotides. **B.** Methylation status of HIV LTR was determined by bisulfite sequencing. A total of ten clones were analyzed. White represents the unmethylated CpG. Black represents the methylated CpG. All the CpGs of HIV_{IIIB} LTR were unmethylated in H9/HIV_{IIIB} with an up-regulated DNMT3A.

2.4 Discussion

The involvement of DNA methyltransferases in the etiology of viral infection and cancer underscores the importance of DNA methylation in maintaining genome stability. The genome-wide DNA methylation study was greatly facilitated by the use of deep sequencing or gene tiling arrays. Surprisingly, there are limited researches on DNA methylation induced by HIV in contrast to its weighty role in scientific/ medical research. We have also linked our study with many other studies on HIV induced mRNA expressions for which DNA methylation may explain differential mRNA expressions. In the following parts of the discussion, the implication of HIV induced methylation and cellular activities, particularly on RNA splicing and neurology will be described.

2.4.1 Molecular basis of DNA methylation

2.4.1.1 Epigenetic modification of host genome by HIV

Hereby we showed the evidence of a HIV-induced aberrant up-regulation of the *de novo* DNA methyltransferase 3A at both transcriptional and translational level. Many of the differentially methylated genes had been reported in other reports as differentially expressed genes or HIV-dependent host factors. By taking into account the transcriptional level of the differentially methylated genes, we suggest that DNA methylation may explain the HIV-induced differential mRNA expression in other studies. Despite the fact that DNA methylation is associated with gene silencing, we have found that many hyper-methylated genes in this study were up-regulated transcriptionally. The discrepancy may be attributed to the cell types or samples used in various studies for addressing different biological phenomenon (Brass *et al.*, 2008; Park *et al.*, 2007; Rotger *et al.*, 2010; Solis *et al.*, 2006; van 't

Wout AB, 2003). Actually it had been shown by another research group that the proportion of up-regulated genes in a hyper-methylated gene list was more or less the same to the proportion of down-regulated genes; and same case for hypo-methylated genes (Cheung *et al.*, 2010). Paradoxically, the *in vivo* results showed that the overall DNA methylation seems to be higher in HIV negative control when compared to the HIV positive twin subject. Such result implies that HIV may have induced a global hypo-methylation of the chromatin indirectly through DNMT3A. Alternatively, it may be due to the lack of a normalization control in analyzing the sequencing results.

2.4.1.2 Molecular basis of an augmented DNMT3A

The augmented level of DNMT3A mRNA level may be attributed to an increased promoter activity by DNA demethylation or mutation. When analyzed with bisulfite sequencing, no methylation on DNMT3A was found in HIV-infected and HIV-uninfected cells. Promoter sequence analysis reveals no major mutations. The up-regulation of DNMT3A was also observed at protein level which may be due to an increased mRNA stability or a decreased level of protein degradation which awaits further studies. It has been shown that the expression level of DNMTs including DNMT3A can be affected by age and gender which is controlled in this study by using twin samples and cell lines of the same origin (Xiao *et al.*, 2008). This shows that the altered DNMT3A expression level is due to the exogenous HIV particles. Therefore, we over-expressed three viral proteins in HEK293 cells and studied the expression level of DNMT3A. These include the viral proteins that may possess transactivation activity, namely Tat, Vpr and Nef. Tat can initiate viral transcription by recruiting cellular transcriptional elongation complex, p-TEFb which

contains cdk9, cyclin T, 7SK small nuclear RNA and the hexamethylene bisacetamide-induced protein (HEXIM1) (Contreras *et al.*, 2007; Muniz *et al.*, 2010; Romano *et al.*, 1999; Sobhian *et al.*, 2010; Zhu *et al.*, 1997). Thus it was suggested that Tat can also transactivate cellular genes. The HIV-1 Vpr protein can interfere several cellular processes such as inducing cell cycle arrest at G2 phase via the inhibitory phosphorylation of Cdc3 (He *et al.*, 1995; Re *et al.*, 1995). Vpr has been shown to account for the transcription of viral promoter to some extent, by the recruitment of cellular p300/CBP or Sp1 cellular transcriptional activators (Kino *et al.*, 2002; Philippon *et al.*, 1999; Wang *et al.*, 1995; Zhu *et al.*, 2001). By interacting with glucocorticoid receptor (GR), Vpr can induce the glucocorticoid response element (GRE) to transactivate the LTR promoter (Schafer *et al.*, 2006). Nef is well-known for its functions in down-modulation of cellular major histocompatibility complex class I (MHC-1) and therefore helps the virus to evade the immune system during active viral replication (Schwartz *et al.*, 1996). Although the interaction between Nef and cellular proteins has not been linked to a defined effect on cellular gene transcription, it is likely that Nef may regulate cellular factors through several kinases or by cooperative effect with Vpr (Arora *et al.*, 2000; Lahti *et al.*, 2003; Renkema *et al.*, 1999; Ye *et al.*, 2004). We have over-expressed Tat, Vpr and Nef separately in HEK-293 cells. However, no significant alteration of DNMT3A expression was observed upon expression of these proteins suggesting that HIV_{IIIB} up-regulates DNMT3A in a complex instead of by a single viral protein.

2.4.1.3 Epigenetic control of HIV genome by host proteins

Integrated HIV genome in host chromosome is initially transcribed when induced by cellular transcription factors. The transcription process is not efficient

but it generates a trace amount of viral protein Tat which is enough to take over the transcriptional transactivation for the viral genome (Lever and Jeang, 2011; Monini *et al.*, 2004; Wang *et al.*, 2000). Hence, it has been suggested that Tat is the subject of epigenetic modification which in turn represses the viral transcription. DNA methyltransferases 3A (DNMT3A) has been shown to interact with protein methyltransferase SETD1 to repress reporter gene expression (Li *et al.*, 2006). It was then speculated such interaction resulted in the methylation of and repression of HIV Tat protein by recruiting transcriptional silencing factors and hence decreasing viral transcription (Van Duyne *et al.*, 2008). An alternative way to decrease viral transcription was by the methylation of viral promoter (Bednarik *et al.*, 1990; Blazkova *et al.*, 2009). As shown in this study, there are several CpG dinucleotides on the NK-κB binding sites and the SP-1 binding sites on the HIV LTR. DNA methylation of the HIV 5' long terminal repeat (LTR) can result in the formation of heterochromatin that hinders the binding of these transcription factors and hence blocks the transcription. The virus therefore enters the latent period and evades the host immune responses and chemophylaxis. Such methylation repression can be overcome by the addition of methylation inhibitor 5-aza-C or by consistently applying trans-activator tat without significant demethylation of the HIV LTR (Bednarik *et al.*, 1990; Blazkova *et al.*, 2009). This implies that HIV-infected patients with aberrant DNA methylation must not be treated with DNA methyltransferases inhibitors. Such inhibitors can rescue the latent viral genome and results in a drastic increase in viral copy number. The potential reactivation of virus by DNMT inhibitors also raise the concern of using these inhibitors to treat patients with virus-induced cancer, such as lymphomas (Epstein-Barr virus, human polyoma virus) (Shah and Young, 2009; Takacs *et al.*, 2010), hepatocellular

carcinoma (hepatitis B virus) (Brechot *et al.*, 2010), cervical cancer (human papillomavirus) (Stanley, 2010; Szalmas and Konya, 2009), Kaposi's Sarcoma (Human Herpes Virus 8) (Pantry and Medveczky, 2009; Toth *et al.*, 2010), etc. We found an up-regulation of DNMT3A but the HIV_{IIIB} genome in the same cell line was not methylated at the viral promoter LTR. This could reflect the capability of actively replicating HIV at its lytic phase to resist the effect of DNA methyltransferase. This may also be the reason for the under-presented methylated host genes in HIV-infected subject in spite of an increased DNA methyltransferase 3A.

2.4.2 HIV and host RNA splicing machinery

2.4.2.1 HIV and human spliceosome

The host splicing machinery is essential for HIV-1 replication in infected cells. Alternative splicing of the HIV-1 genomic RNA generates more than 40 different mRNAs required for viral replication. The spliced mRNAs are then transported from the nucleus to the cytoplasm for the synthesis of viral proteins. In the early phase of HIV-1 gene expression, *tat*, *rev* and *nef* are transcribed which are completely spliced and exported out of the nucleus for translation by cellular exportin-1 (CRM-1) (Pollard and Malim, 1998; Urcuqui-Inchima *et al.*, 2011). The biological importance of Tat and Nef protein for HIV life cycle has been described in section 2.4.1. Rev protein is an important viral protein in the early phase as it forms the nuclear export channel for the singly spliced and unspliced viral mRNA which are otherwise unrecognized by the cellular export channel and retained in the cytoplasm (Pollard and Malim, 1998). This then allows the export of transcripts for other viral proteins including envelope protein, polymerase, integrase and proteases,

etc. The unspliced viral genomic RNAs are also exported through the Rev channel (Purcell and Martin, 1993; Tazi *et al.*, 2010).

A direct relationship between HIV and RNA spliceosome was shown by identifying *Gemin8* as a de-methylated gene in HIV infected subject. A significant up-regulated level of *Gemin8* was revealed by real-time PCR which implies that HIV highjacked the host RNA splicing pathway. Gemin8 is a component of the SMN complex. SMN protein is the product of the spinal muscular astrophy disease gene called 'survial of motor neurons' (*SMN*) (Lefebvre *et al.*, 1995). SMN protein together with Gemin 2-8 and Unr-interacting proteins (UNRIP) form a large macromolecule complex called SMN complex that localizes in the cytoplasm, nucleoplasms and in nuclear Gems (Carissimi *et al.*, 2006; Gubitz *et al.*, 2004). The primary function of SMN complex in the cell is the biogenesis of uridine-rich small nuclear ribonucleoproteins (UsnRNPs), which are assembled in the cytoplasm and accumulates in gems in the nucleus (Cauchi, 2010; Fischer *et al.*, 1997; Kolb *et al.*, 2007). The SMN associates with Sm proteins and interacts directly with specific domains of these small nuclear RNA (snRNA). Such interaction results in the formation of Sm cores on snRNAS of the splicesosome. The spliceosome complex can be divided into major spliceosome containing snRNP U1, U2, U4 and U5 and the minor spliceosome containing U11 and U12, U4atac/U6atac and U5 snRNP (Cauchi, 2010; Patel and Steitz, 2003).

The cellular SMN complex has been utilized by viral small RNAs encoded by the *lymphotropic herpesvirus saimiri* (Golembe *et al.*, 2005). Making use of the host splicing machinery is especially important to RNA virus like HIV because of the

lack of viral self-splicing mechanism. It is known that the SMN interacting protein Gemin2 binds to HIV integrase at its C-terminal to facilitate viral cDNA synthesis (Hamamoto *et al.*, 2006). It has also been shown as a host dependency factor and is indispensable for the early events post-infection (Brass *et al.*, 2008; Hamamoto *et al.*, 2006; Warren, 2009). In this study, we have further showed the indirect evidence that HIV may use SMN complex by up-regulating *Gemin8*. An increased expression level of Gemin8 may potentiate the activity of the SMN complex and lead to an increase in snRNP biogenesis and mRNA splicing (Feng *et al.*, 2005). In consistence with this speculation, we found that the small nuclear ribonucleoprotein (snRNP) U5 (*TXNLA*) and U11/U12 (SNRNP48) were also up-regulated transcriptionally upon viral infection. *TXNLA* and *SNRNP48* are found de-methylated in this study which explains its increased level of transcription. We speculate that when the splicing units are taken up by the virus for viral mRNA splicing, the host can increase the expression of splicing factors accordingly to replenish the splicing complexes for normal cellular activities.

2.4.2.2 HIV and human splicing factors

Splicing enhances translation via the deposition of the exon-exon junction complex and other multifunctional splicing factors, including arginine/serine rich (SR) proteins. Exon junction complex is implicated in HIV infection by the discovery of an up-regulated *MAGOHB* expression in this study. *MAGOHB* is the second human homolog for the *Drosophila mago nashi* gene (Lau *et al.*, 2003; Newmark and Boswell, 1994; Zhao *et al.*, 1998). SR proteins share a common arginine/serine domain. They play a role in early recognition of splice sites, recruitment of basic splicing factors to the pre-mRNA, and formation of bridging

contacts with other arginine/serine domain-containing splicing factors (Philipps *et al.*, 2003). Viruses, without their own splicing system, profoundly alter the cellular SR proteins and hence control the RNA splicing machinery. HIV-1 in macrophage has been found to alter the expression level of a SR protein SC35 (Maldarelli *et al.*, 1998; Ropers *et al.*, 2004). By doing this, the virus optimizes the cellular environment for the generation of different viral proteins required at different stages of the viral life cycle (Dowling *et al.*, 2008; Ropers *et al.*, 2004). In this study, a host serine/ arginine-rich (SR) protein encoding genes *AKAP17A* was found de-methylated in HIV-positive twin subject and up-regulated in *in vitro* HIV_{IIIB}-expressing cells. A-Kinase (PRKA) anchor protein 17A (*AKAP17A*) by its name, targets to protein kinase A (Jarnaess *et al.*, 2009). It co-localizes with its target as well as with SC35 SR protein. It is speculated that the virus induces a higher expression of cellular AKAP17A protein in order to facilitate the expression of viral proteins in its actively replicating stage. Alternatively, the up-regulation may be the strategy used by the cells to compensate for the SR proteins taken over by the viruses. Another mRNA splicing factor proline/glutamine-rich protein gene *SFPQ* was found to be up-regulated. It was recently shown to bind to the dyslexia susceptibility genes *DYX1C1* in a complex with transcription factor (TFII-I) and poly (ADP-ribose) polymerase 1 (PARP) (Tapia-Paez *et al.*, 2008). SFPQ (PSF) is a DNA/RNA binding protein which facilitates the DNA pairing to RNA, single-stranded DNA or double-stranded DNA to facilitate renaturation of complementary single-stranded DNA molecules (Akhmedov and Lopez, 2000). Although SFPQ does not possess a typical RS domain, it shares some function of SR protein by binding to U1-70K and SR protein (Shav-Tal *et al.*, 2001).

2.4.3 Linking epigenetics with alternative splicing

In human, about 90% genes undergo alternative splicing during mRNA maturation. Histone methylation has been shown as an emerging regulation mechanism for mRNA splicing (Allo *et al.*, 2009). Histone hyper-methylated chromatins recruit different chromatin-splicing adaptor factors, thus result in different splicing patterns (Luco *et al.*, 2010). Although there is no direct evidence by now, there is an indication that DNA methylation may directly or indirectly affect splice site choice via histone modifications (Luco *et al.*, 2011). Such regulation is likely mediated by methyl-CpG binding proteins (MDB) as mutation in methyled-CpG binding proteins 2 have been linked to abnormal splicing pattern (Okitsu and Hsieh, 2007; Young *et al.*, 2005). Indeed the complex regulation circuit mediated by different chromatin modifications may explain the expression of different spliced forms of mRNA in a tissue-specific manner. This avoids including an additional set of splicing factors in different cell types (Hanamura *et al.*, 1998). Hence, it is possible that HIV also indirectly alter the cellular mRNA splicing pattern through epigenetic modulations to produce proteins that favour viral survival. A more comprehensive reading on the interplay of epigenetics and mRNA splicing should refer to Luco's review entitled 'Epigenetics in alternative pre-mRNA splicing' (Luco *et al.*, 2011).

2.4.4 HIV and neurology

2.4.4.1 HIV-1 infection of the central nervous system

One of the distinctive features of AIDS is the development of neurological problems in infected subjects. Central nervous system (CNS) HIV-1 infection begins during primary infection with the detection of viral particles in the

cerebrospinal fluid. The infection can continue as a chronic infection throughout the course of disease if untreated (Ritola *et al.*, 2004; Spudich *et al.*, 2005). The CNS infection can progress to more 'invasive' HIV-1 encephalitis (HIVE) resulting in cognition, motor and behavioral dysfunction which is described as AIDS dementia complex (ADC) (Navia *et al.*, 1986; Price, 1994). It is believed that HIV in the infected lymphocytes and monocytes pass through the blood-brain barrier to infect the CNS. There are two theories as to the means by which HIV damage the blood brain barriers. One is the direct damage of the barrier by infecting the endothelial cells. The other suggests that the infected lymphocytes and monocytes traffic across the barriers as part of their immune surveillance functions (Brew, 2001; Valcour *et al.*, 2011). CNS HIV-1 infection accounts for the majority of neurological disease in HIV/AIDS and contributed to a large of proportion of mortality. The early CNS infection can be reversed by applying the highly active antiretroviral therapy (HAART) whilst later HIV-1 infections sound 'untreatable'. In contrast, fewer cases are attributed to the HIV/AIDS associated brain infection by *Cryptococcal neoformans* (Cryptococcal meningitis) and *Toxoplasma gondii* (Toxoplasmosis) (Brew, 2001; Price and Spudich, 2008).

2.4.4.2 HIV induced methylation of neurology-related genes

In this study, many of the hyper-methylated genes were clustered in the neuron development pathway including the Huntington's disease (HD) gene Huntingtin (*HTT*) (Morell, 1993). Huntington's disease is characterized by a loss of striatal neurons in the brain of the patients. The disease gene *HTT* contains a fairly broad range of the number of trinucleotide CAG repeats in normal people while repeat numbers exceeding 40 is pathogenic (Jayadev and Garden, 2009). Until now, there

is a dearth of study on HIV-1 and commonly known neuron degenerative diseases such as Huntington's disease and Parkinson's disease. There are few cases of Huntington's disease in HIV/AIDS patients reported so far (Sadek *et al.*, 2004; Sevigny *et al.*, 2005). These common CNS diseases complicated by HIV-1 may not present the 'classical' clinical symptoms, making it challenging to differentiate the diseases. In addition, *POU3F2* which is a transcription factor contributing to a Huntington-like disease was found hyper-methylated (Costa Mdo *et al.*, 2006). EMX2, which is important in the brain development, was also found hyper-methylated (Suda *et al.*, 2010). HTT and EMX2 are normally expressed in both brain and lymphocytes while POU3F2 is only expressed in the brain. Although our results suggested that HIV may potentiate up-regulation of HTT and methylation of POU3F2 and EMX2 in brain cells, no conclusion could be drawn before further studies using brain cells as a model.

2.4.5 HIV and cell transduction

Several genes found in this study were related to signal transduction including the Rho-signaling pathway (Rho guanine nucleotide exchange factor GEF 3, *ARHGEF3*), electron transport chain across mitochondrial membrane (NADH dehydrogenase 1 alpha subcomplex, *NDUFA2*), and synaptogenesis (MYC binding protein 2, *MYCBP2*). Interestingly, ARHGEF protein was found to be involved in the control of synapse development (Tolias et al., 2011) whereas MYCBP2 protein is related to several neural degenerative disease via regulating synapsis (Le Guyader *et al.*, 2005). This implies a potential connection between the involvement of signaling molecules and neurological dysfunction in HIV-1 infection.

2.4.6 HIV and non-coding RNAs

In referring to Figure 2.4 and 2.5, we noticed that there are a number of non-coding RNAs (ncRNAs) with prefix of NCRNA- and microRNAs with prefix of MIR-. There are no defined functions of these ncRNAs thus they were not included in the validation. Nevertheless, these could be the targets for future studies. Regulation of viral infection by ncRNA is an emerging field of study for HIV in recent years, when scientists started to know more about the ncRNAs. For example, the role of 7SK small RNAs was implicated in Tat-induced viral transcription (Muniz *et al.*, 2010). Another elegant study showed miR-382 was involved in HIV latency (Huang *et al.*, 2007). Nevertheless, a larger-scale study on cellular miRNA in HIV infection is still lacking. Our results indirectly show that HIV could induce changes in miRNAs by means of DNA methylation. In fact, the epigenetic silencing of tumor suppressor miRNAs and ncRNAs by CpG hyper-methylation has been reported, which may serve as a hall mark of cancer, in addition to the DNA methylation of tumor suppressor proteins (Lujambio *et al.*, 2008; Lujambio *et al.*, 2010).

2.4.7 Potential pitfalls in this study

2.4.7.1 Application of high-throughput sequencing in studying methylated genes

In studying differential mRNA expression induced by exogenous factors, housekeeping genes such as β-actin, GAPDH, TBP, 18S, etc were often used as the endogenous controls to normalize the degree of mRNA expression among different samples. Alternatively, the precise calculation of relative expression could be achieved by 'calibrating' against a standard curve for specific genes. But in studying methylation, there is no such a well-accepted endogenous gene which could

fairly serve as a normalization control. Practically, equal amount of exogenous methylated/ unmethylated DNA such as methylated pUC19 DNA was added to the samples before immunoprecipitating the methylated-DNA. But in the process of high-throughput sequencing, the exogenous pUC19 may not work fairly as a housekeeping control. This may be due to the bias of MeDIP towards genes with CpG rich and longer (relative abundant) promoters. In the whole genome amplification prior to sequencing and emulsifying PCR during sequencing, the bias can be 'amplified' as the DNA increases exponentially.

2.4.7.2 Determination of the degree of methylation

Since there is no reference standard of DNA methylation in this study, the study of differential methylated DNA is limited to the mutually exclusively methylated genes in diseased and non-diseased subjects. Despite the availability of human methylome (Lister *et al.*, 2009), the individual-to-individual difference in methylation make it difficult to compare the degree of methylation in two groups of study subjects. Using identical twin samples in this study has kept off the majority of inter-personal genetic variations. Yet, the problem has been complicated by the lack of a normalization control as described in 2.4.7.1 which restricts our study to the exclusively methylated or unmethylated genes.

2.4.7.3 Mapping limitation

In this study, we defined the promoter region as -1000/+1000 of the transcription start site with reference to the human transcript (G+T) databases in NCBI. However, the strategy may not work well for small genes such that the -1000/+1000 region of these genes may not necessary be their regulatory sequence.

While validation experiments are always needed in showing the methylation status of these genes, the mapping algorithm is also subject to adjustment. Separate algorithms should be applied according to the predicted length of gene promoters. For instance, including of an extra mapping strategies for genes of small RNAs (miRNA, short non-coding RNAs) would increase the preciseness of the analysis.

2.5 Conclusion

It is obvious that viruses can induce DNA methylation of cellular genes. By means of studying DNA methylation genes in identical twins as well as using the cell line models, we discovered that HIV-1 can induce cellular factors related to RNA splicing, neurology and cellular signaling. The identification of non-coding RNAs and microRNAs hint at a role of that, in addition to modulating cellular protein, HIV-1 can induce methylation of cellular small RNAs in infection. The study should also be extended to more twin subjects, though it is practically challenging to find another pair of twin subjects. Extensive studies are needed to define the interaction between HIV-1 and these cellular genes and non-coding RNAs, which have been emerging has a new field of genetic studies. Such findings might hold the key to the breakthrough in understanding virus-host interplay after 30 years since the discovery of the virus.

Chapter 3

Metagenomic comparison of the plasma microbiome of HIV/AIDS patients and healthy adults

Summary

This chapter describes the results of a metagenomic study on plasma microbiome in HIV/AIDS patients compared to normal adults. We have profiled the plasma bacterial DNA of 10 HIV/AIDS patients in parallel with 10 healthy adults using the objective high-throughput Illumina Solexa sequencing technology. HIV/AIDS plasma microbiome was dominated by bacteria from the order Pseudomonadales (72.31%). The validation experiments suggested that, for the first time, the association of several bacteria such as Moraxella osloensis and Psychrobacters with HIV infection. We have also detected a few microbial genetic elements in plasma of healthy people suggesting for the presence of bacterial genome in apparently healthy people. Besides, by comparing with the published gut microbiome, we have found that many of the microbes in HIV/AIDS plasma are similar to some of the microbes found in the human gut. In addition, by means of sequencing, we have found several long sequences which might belong to novel bacteria. The insights gained into the specific spectrum of microbes found in these patients may facilitate the identification of potential infections in HIV/AIDS patients and the use of appropriate prophylaxis to improve the disease prognosis.

3.1 Introduction

Human plasma in immuno-competent individuals is assumed to be sterile and therefore no microbes except viruses should be present. This might not be the case in immuno-compromised hosts, as exemplified by the development of opportunistic infection in human immunodeficiency virus (HIV) infected patients, which signifies the progression to acquired immunodeficiency syndrome (AIDS) (Mootsikapun, 2007). As mentioned in Chapter 1, one of the distinctive features of HIV infection is the depletion of CD4+ T-lymphocytes (Rudnicka and Schwartz, 2009). Development of AIDS is defined when the CD4+ cell count falls below 200 cells per microliter. As a result of the CD4+ lymphocytopenia, HIV/AIDS patients without antiretroviral therapy are highly susceptible to infections by different microbes in the natural environment (Mootsikapun, 2007). Microbial infection is in fact a hallmark of AIDS, which causes significant morbidity and mortality (Antiretroviral Therapy Cohort Collaboration, 2010; Palella *et al.*, 2006). The severe complications of such infections can be preceded by asymptomatic bacteremia or viremia, the characterization of which would be useful for supporting clinical diagnosis and for the introduction of prophylaxis. However, the identification of microbes after the onset of clinical complications takes time, and may not be always possible using currently available techniques (Handelsman, 2004). In this part of the study, we set out to conduct a metagenomic study to profile microbial genetic materials in the bloodstream of treatment naïve HIV/AIDS patients and healthy adults, the results of which may help to improve the disease management.

3.2 Materials and Methods

3.2.1 Ethics Statement

This study was conducted according to the principles expressed in the Declaration of Helsinki. The study was approved by the Institutional Review Board of Jiangsu University, China and the Institutional Review Board of the Chinese University of Hong Kong. All the studied subjects provided written informed consent for the collection of samples and subsequent analysis.

3.2.2 Collection of plasma from HIV/AIDS patients and control groups

Ten treatment naïve HIV patients with a CD4+ cell count of 4-125 cells per microliter (μl) were recruited in Jiangsu Province, China (Table 3.1). These patients were not under antiretroviral therapy by the time they were recruited. Medical follow-up was given afterwards. They were free of any symptoms and were not under treatment of opportunistic infections when their blood samples were taken. An independent control group of 10 healthy adults were recruited from Hong Kong. Another sero-negative control group of 10 healthy adults were later recruited from Jiangsu Province as the environment-corresponding control group. Blood samples were collected from the study subjects by applying standard aseptic techniques and put into tubes containing EDTA.

Table 3.1 Information of adult HIV/AIDS patients recruited in this study.

Code	Gender	Age	CD4 counts/μl	CD8 counts/μl	CD4/CD8 ratio	viral load (copies/ml)	Antiviral therapy	Transmission route
HIV/AIDS patients from Jiangsu								
L-001	M	28	4	697	0.006	786000	no	heterosexual
L-002	F	33	6	361	0.017	426000	no	IDU
L-003	F	45	11	522	0.021	346000	no	blood donor
L-004	F	40	20	329	0.061	101000	no	heterosexual
L-005	F	36	35	263	0.133	303000	no	heterosexual
L-006	M	38	48	667	0.072	165000	no	blood transfusion
L-007	F	24	107	1855	0.058	110000	no	heterosexual
L-008	M	45	124	1271	0.098	410000	no	heterosexual
L-009	M	38	124	1671	0.074	195000	no	MSM
L-010	F	35	125	543	0.230	634000	no	heterosexual

Footnote:

Gender: M-male, F-female;

Transmission route: IDU-infection drug use, MSM-men who have sex with men.

3.2.3 Extraction of plasma DNA

DNA was extracted from 200 μl plasma using QIAamp DNA blood mini kit (QIAGEN). The extracted plasma DNA was dissolved in 200 μl H_2O which has been treated with DNase before use. The extraction process was performed in a hood pre-illuminated with ultra-violet (U.V.) light for 30 minutes. Designated pipettes, filter tips and other necessary equipment were used in this project to avoid contamination of microbial DNA from the environment. Unless otherwise specified, similar precautionary measures were adopted in the other parts of the study, including polymerase chain reactions.

3.2.4 Molecular determination of bacteremia

As a pilot study to validate the hypothesis, we have first determined the presence of bacteria DNA in the plasma of the HIV/AIDS patient samples using 16S ribosomal RNA (rRNA) gene sequencing. The 16S rRNA gene was amplified using

universal 16S rRNA gene primers (de Madaria *et al.*, 2005; Dethlefsen *et al.*, 2008). The sequence of the forward and reverse primers were 5'AGA GTT TGA TCC TGG CTC AG 3' and 5'ACG GCT ACC TTG TTA CGA CTT 3' respectively. Polymerase chain reaction (PCR) was performed in a 30 µl mixture with 200 nM of each of the primers and 10 ng plasma DNA. Negative controls were prepared by replacing the template with DNase-treated H_2O. The amplified and purified 16S rRNA gene from two patients was cloned into pGEM-T easy vector (Promega) according to the manufacturer's instructions. The positive pGEM-T-16S clones were sequenced for the identity of the bacterial 16S rRNA genes.

3.2.5 Amplification of plasma DNA and Illumina Solexa sequencing

The yield of plasma DNA extracted from 200 µl of plasma from HIV/AIDS patients was only about 10 ng/µl, the samples were therefore amplified in order to have 10 µg DNA for Illumina Solexa sequencing. To minimize individual variations of opportunistic infections by heterogeneous sources of pathogens, 5 nanogram (ng) of total plasma DNA was taken from each of the HIV/AIDS patients and mixed together. Similarly, 5 ng plasma DNA was taken from each of the ten healthy individuals and mixed together. The pooled plasma DNA was amplified by multiple displacement amplification method using GenomiPhi V2 kit (GE Medical Systems). Five microlitre pooled plasma DNA sample was mixed with 5 µl sample buffer and heat denatured at 95°C for 3 minutes. The samples were then cooled on ice. Nine microlitre reaction buffer and 1 µl enzyme mixture were added to the DNA sample and incubated at 30°C for 90 minutes. The reaction was terminated by heating at 65°C for 10 minutes. The amplified samples were analyzed on DNA agarose gel before purification and purified afterward. This method has been used

for amplification of a tiny amount of DNA for large scale sequencing (Reyes *et al.*, 2010).

3.2.6 Illumina Solexa sequencing analysis

The amplified and purified plasma DNA samples were sequenced by pair-ended Solexa short-read sequencing technology with insert sizes of 250 base pairs (b.p.). The sequence signals were analyzed using Illumina Genome Analyzer.

3.2.7 Contigs assembly and BLAST analysis

The detailed work flow of the experiments was described in section 3.3. Basically, all sequenced reads were assembled using Velvet (Zerbino and Birney, 2008) with a hash length of K21. The sequence contigs assembled using Velvet with K21were filtered to remove those less than 100 nucleotides. Gene identity was determined by Basic Local Alignment Search Tool (BLAST) nucleotide BLAST (BLASTn) and BLASTx searches against the microbial genomes available on NCBI dataset (http://www.ncbi.nlm.nih.gov/) in Jan 2010 with an e-value of $<1x10^{-10}$ and a positive hit length of >50% as cut off. Taxonomic analysis of the sequenced microbial genome was made based on the NCBI data sets.

3.2.8 Amplification of bacterial genes as validations

The amplification of bacterial genes as part of the validation experiments was done by touch-down PCR (Korbie and Mattick, 2008). Amplification of *Moraxella osloensis* preprotein translocase was performed by MOslo-F 5'AGT TGA GTA ACG CAG CCT CAA '3 and MOslo-R 5'GCC CAA AGA AAA GTG GAA AAC 3'. Amplification of a hypothetical protein gene of *Ralstonia pickettii* was done by RP-F

5' CTG GGG TCG ATG ACG GTA 3' and RP-R 5' ATC TCT GCT TCG TTA GTG GC 3'. Amplifications of *Psychrobacter sp.* ribosomal protein S19 and *Acinetobacter baumannii* transposase subunit were done using Psychr-F 5' AAT TTC ATG CCT CGT TCA TTG 3', Psychr-R 5' GAC CAA CCA TTT GCT CGT TTA 3', ABaum-F 5'TCC TCA GTT TAA TGC CAA TGC 3' and ABaum-R 5' CCA AAA CCA ATT AAA CGC TGA 3' respectively. The PCR was performed in a 15 μl mixture with 200 nM of each of the primers, 1X *Taq* polymerase buffer, 1X additives (0.5 M betaine, 1.3 mM DTT, 1.3% DMSO, 11 mg/ml BSA), 1.5 mM Mg_2Cl, 200 μM dNTP mixture, 2 units of Taq polymerase and 10 ng plasma DNA. DNase-treated H_2O was used in preparing the reagents for PCR use. The reaction was performed by denaturing at 95°C for 5 minutes, followed by 20 cycles of 94°C for 10 seconds, 62°C for 30 seconds (decreasing 0.5°C per cycle) and 72°C for 30 seconds, then 30 cycles of 94°C for 10 seconds, 50°C for 30 seconds and 72°C for 30 seconds, with a final extension at 72°C for 10 minutes and a hold at 4°C before the sample was removed from the machine. Negative controls were prepared by replacing the template with DNase-treated H_2O. Positive controls were prepared by adding 10 ng DNA materials of specified bacteria in the PCR.

3.3 Results

3.3.1 Extraction of nucleic acids from AIDS patients and control groups

Plasma DNA and RNA was extracted from 10 AIDS patients and healthy adults. It was not surprising that the concentration of plasma nucleic acids was very low. The amount and quality of extracted DNAs are shown in Table 3.2.

Table 3.2 Amount and quality of extracted plasma nucleic acids.

Disease status	Code (Reference)	Gender	Age	Plasma DNA		
				ng/μl	A260/280	A260/230
HIV/AIDS	L-001	M	28	2.80	2.96	0.15
HIV/AIDS	L-002	F	33	4.40	1.19	0.20
HIV/AIDS	L-003	F	45	36.60	1.46	0.80
HIV/AIDS	L-004	F	40	3.30	1.86	0.16
HIV/AIDS	L-005	F	36	4.60	2.08	0.19
HIV/AIDS	L-006	M	38	4.10	1.97	0.18
HIV/AIDS	L-007	F	24	3.60	2.38	0.17
HIV/AIDS	L-008	M	45	3.10	2.73	0.18
HIV/AIDS	L-009	M	38	4.60	1.81	0.17
HIV/AIDS	L-010	F	35	3.30	2.31	0.15
Healthy	M-001	M	19	1.20	1.00	0.17
Healthy	M-002	M	21	1.40	0.73	0.18
Healthy	M-003	F	26	1.50	1.16	0.22
Healthy	M-004	M	23	1.50	1.15	0.24
Healthy	M-005	F	27	1.40	0.78	0.19
Healthy	M-006	M	2	2.10	1.13	0.18
Healthy	M-007	M	22	0.90	1.02	0.12
Healthy	M-008	F	25	1.30	1.00	0.16
Healthy	M-009	F	22	2.20	1.25	0.14
Healthy	M-010	M	31	1.60	1.03	0.20

3.3.2 Determination of multiple bacterial infections in AIDS patients

The bacteremia status of the patients was determined with 16S rRNA gene amplification. In the clonal selection, five clones of pGEM-TA-16S were isolated for patient #3 while 13 clones were isolated for patient #6. Conventional sequencing of the 1.5 k.b. long insert of 16S rRNA gene confirmed that both patients carried genetic materials of multiple bacteria which have never been reported as AIDS-associated (Table 3.3). After eliminating the clones with same sequences, bacteria including *Methylobacterium radiotolerans*, *Bradyrhizobium japonicum* and *Rickettsia sp.* were found in patient #3 and *Renibacterium salmoninarum*, *Phyllobacterium myrsinacearum*, *Burkholderia cenocepacia*, *Pseudomonas/ Pelomonas saccharphila*, *Leifsonia sp.*, *Methylobacterium radiotolerans* and *Afipia*

genosp. 1 were found in patient #6.

3.3.3 Metagenomic sequencing of AIDS plasma microbiome

Figure 3.1 shows the workflow of the experiments for analyzing the AIDS plasma microbiome. The plasma DNA extracted from AIDS patients was amplified and sequenced using Illumina Solexa sequencing. The sequence reads were assembled into contigs and aligned against the reference nucleotide collection (nt/ nr). The contigs without matches in nucleotide BLAST (BLASTn) was aligned with the protein collection using BLASTx. The contigs with matches were further analyzed for the identity and relative abundance. A total of 658,535,877 b.p. were generated for AIDS plasma microbiome in the Solexa sequencing including 1.43 % of human sequences. After filtering out the contigs smaller than 100 b.p., about 16,308 contigs were left. These contigs had a length of 100 b.p. to 9.5 k.b., and over 90% were shorter than 1 k.b. (Figure 3.2).

Table 3.3 Multiple bacterial infections in two HIV/AIDS patients.

Source	Bacteria no. (reference)	Identification by 16S rRNA gene sequencing	Nucleotide blast		Diease		Taxonomy	
			Positive length	Identity %	Human	Non-human	Phylum	Family
L003	1	*Methylobacterium radiotolerans*	1168	98	Non-pathogenic, lives inside the human mouth		Alphaproteobacteria	Methylobacteriaceae
	2	*Bradyrhizobium japonium*	2012	97		Symbiotic plant bacteria	Alphaproteobacteria	Bradyrhizobiaceae
	3	*Rickettsia sp.*	1182	98	Rocky Mountain spotted fever, Rickettsial pox, Boutonneuse fever, Siberian tick typhus, Oriental spotted fever		Alphaproteobacteria	Rickettsiaceae
L006	1	*Renibacterium salmoninarum*	1216	93		Salmon fish: kidney disease	Actinobacteria	Micrococcaceae
	2	*Phyllobacterium myrsinacearum*	1182	97		Plant bacteria	Alphaproteobacteria	Phyllobacteriaceae
	3	*Burkholderia cenocepacia*	1216	96	Opportunistic infection in patients with cystic fibrosis and chronic granulomatous disease		Betaproteobacteria	Burkholderiaceae
	4	*Pseudomonas/ Pelomonas saccharphila*	1208	96	Isolated from haemodialysis water		Betaproteobacteria	Comamonadaceae
	5	*Leifsonia sp.*	1200	97		Plant: ratoon stunting disease	Actinobacteria	Microbacteriaceae
	6	*Methylobacterium radiotolerans*	1024	96	Non-pathogenic, lives inside the human mouth		Alphaproteobacteria	Methylobacteriaceae
	7	*Afipia genosp. 1*	1199	96		(Afipia felis is the bacterium causing cat scratch disease)	Alphaproteobacteria	Bradyrhizobiaceae

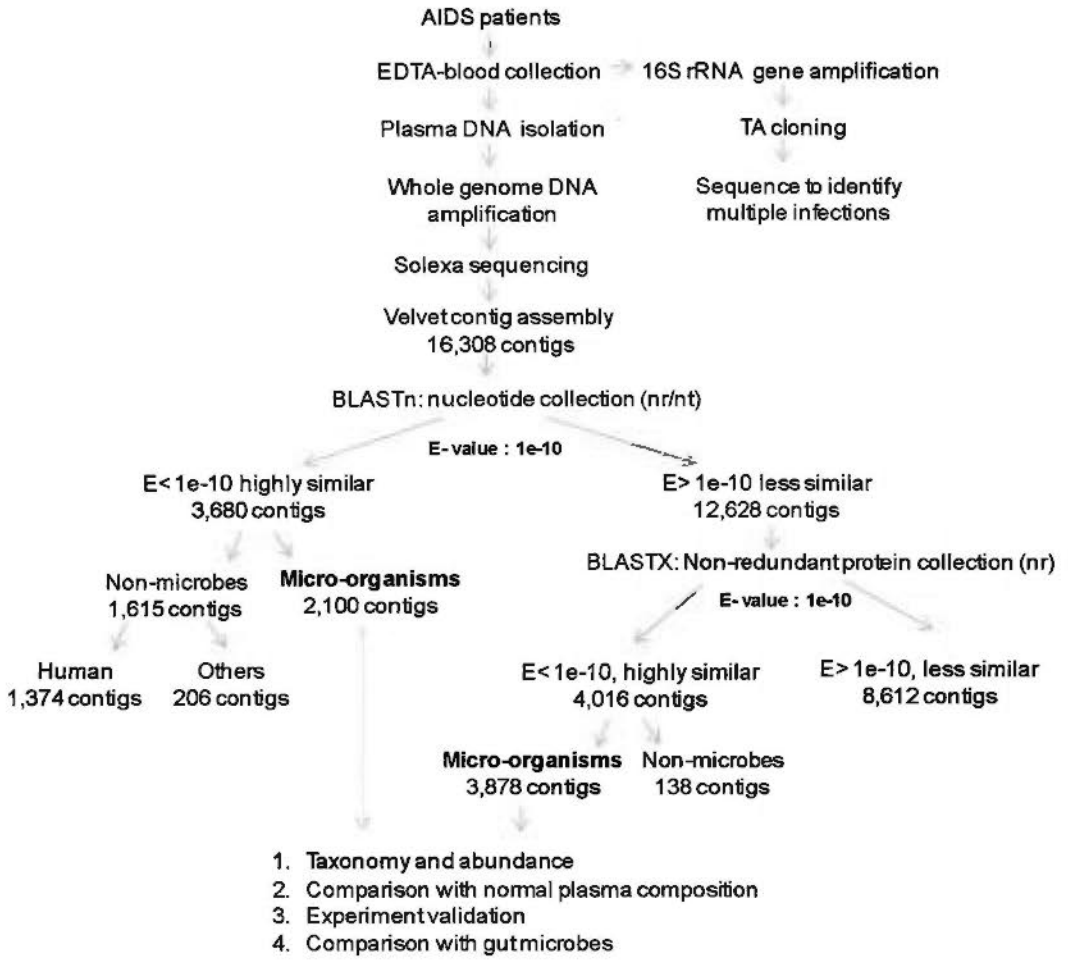


Figure 3.1 Work-flow of this study. The plasma DNA extracted from AIDS patients was amplified and sequenced using Illumina Solexa sequencing. The sequence reads were assembled in contigs and aligned against the reference nucleotide collection (nt/-nr). The contigs without matches in BLASTn was aligned with the protein collection using BLASTx. The contigs with matches were further analyzed for the identity and relative abundance.

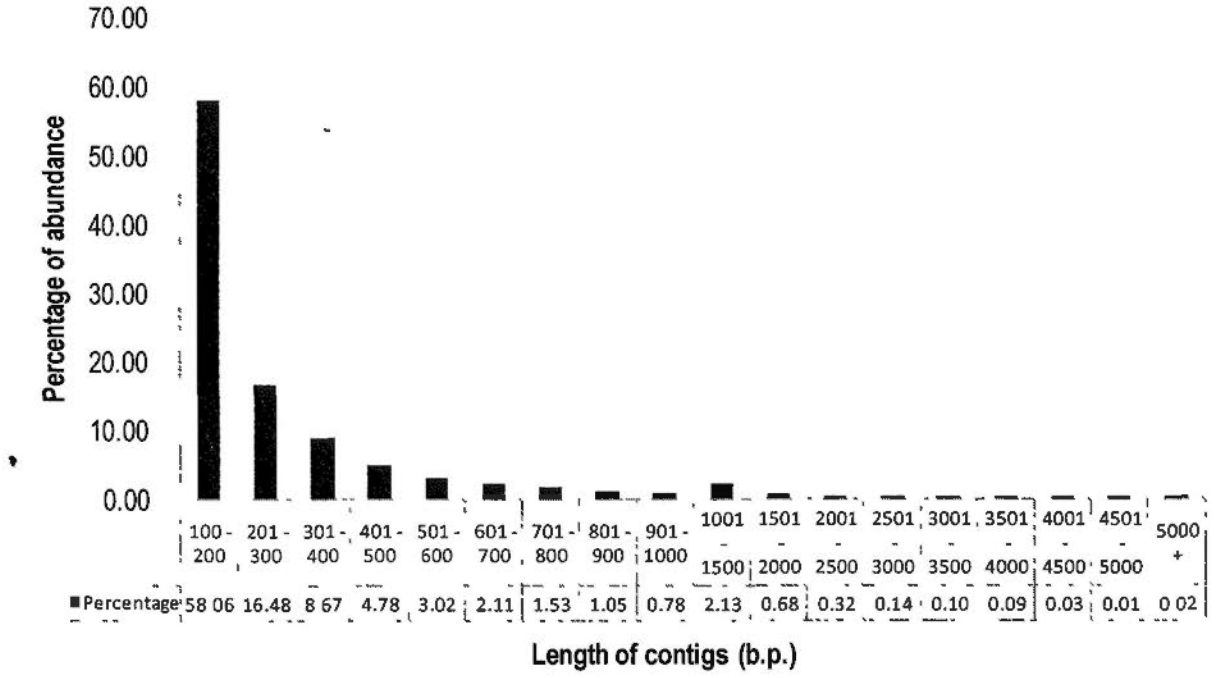


Figure 3.2 Length distributions of assembled contigs. Most of the contigs had a length of 100-200 nucleotides.

Contigs with special features are listed in Table 3.4. Some long contigs (>500 b.p.) were matched with known bacteria at a high nucleotide similarity. These matched long contigs were then successfully amplified by experiments described in later sections. On the other hand, some very long contigs had only <10% of their sequence aligned with any known bacteria. Many contigs, despite their short length, had a noticeably high number of constituting reads (coverage). Of the 16,308 assembled contigs, 3,680 (22.6%) had matches from the NCBI nucleotide collection (nr/nt) using an e-value of $<1x10^{-10}$ as cut off, while 4,016 (24.6% of 16,308 contigs) of the remaining 12,628 contigs were identified in BLASTx search against a non-redundant reference database using the same cut off e-value.

Table 3.4 Selected contigs with special features.

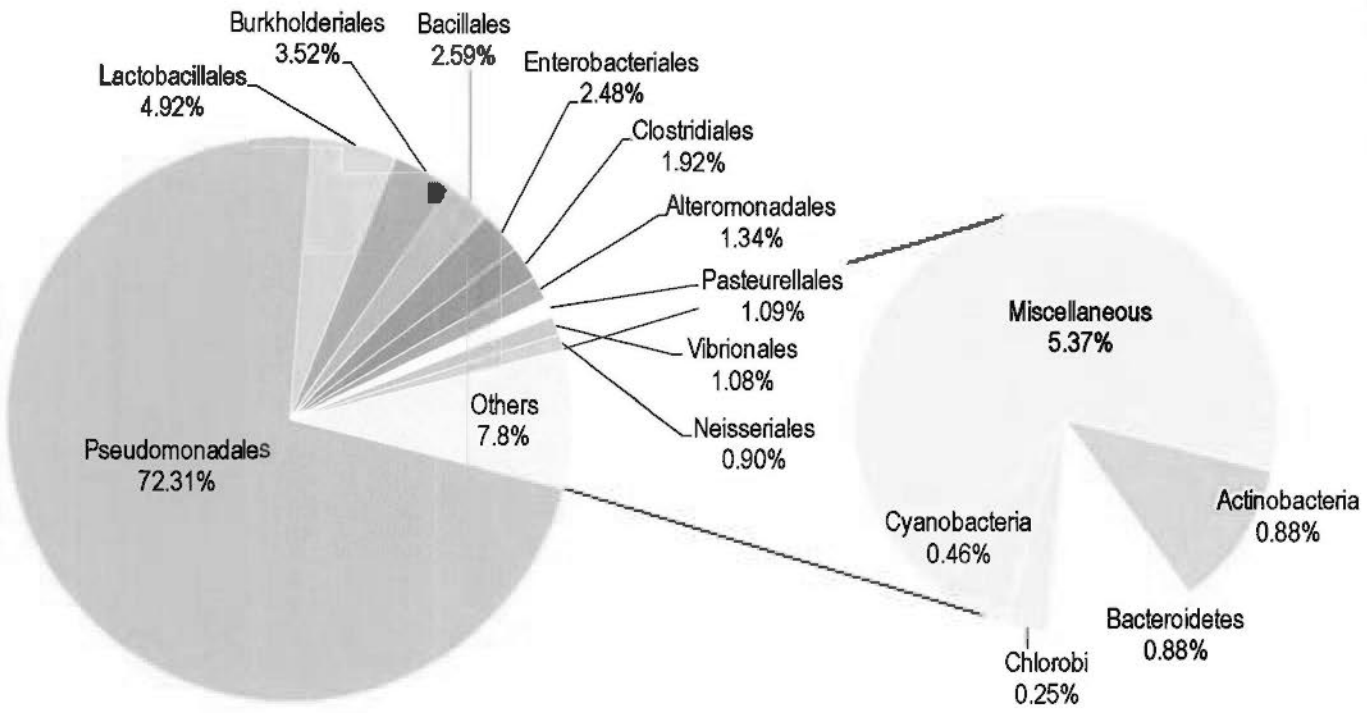
Contigs (Reference)	Length (b.p.)	Coverage (average no. of constituting reads)	Micro-organisms	Longest positive hit length (b.p.)	Hit identity %
Long contigs with high identity matches					
1	4,562	12.80x	*Ralstonia pickettii*	4,362	100.0
2	3,795	11.90x	*Alteromona macleodii*	3,215	99.8
3	2,473	5.77x	*Propionibacterium acnes*	2,269	99.8
4	1,763	11.00x	*Moraxella osloensis*	958	97.2
5	1,688	9.17x	*Listeria innocua*	1,461	98.3
6	1,668	5.21x	*Bordetella bronchiceptica*	1,280	99.9
7	1,384	4.77x	*Psychrobacter PRwf-1*	1,176	99.3
8	667	39.70x	*Acinetobacter baumannii*	466	99.8
Long contigs with short-length matches					
9	9,490	10.65x	*Lactobacillus salivarius*	386	79.9
10	5,530	11.73x	*Synechocystis sp.*	79	81.4
11	4,942	8.57x	*Bacillus licheniformis*	326	79.5
12	4,479	8.11x	*Lactococcus lactis*	141	80.6
13	3,824	12.58x	*Listeria innocua*	57	86.4
14	3,057	8.40x	*Streptococcus sp.*	113	81.9
Short contigs with high coverage					
15	378	228.00x	*Acinetobacter baumannii*	144	84.7
16	333	205.00x	*Lactobacillus lactis*	116	99.1
17	325	214.00x	*Escherichia coli*	60	85.7
18	323	289.00x	*Psychrobacter PRwf-1*	103	100.0
19	304	401.00x	*Acinectobacter sp.*	86	86.0
20	118	108.00x	*Yersinia pseudotubertulosis*	91	85.8

As illustrated in Figure 3.1, the taxonomy ranks of the sequenced and annotated genomes were analyzed. The human and non-microbes contigs were removed, leaving 2,100 contigs and 3,878 contigs from BLASTn and BLASTx research respectively. Taken together, a total of 5,978 contigs of 100 b.p. to 9.5 k.b. long were originated from micro-organisms. The relative abundance of microbial genomes was calculated by taking into account the matched length of the contigs and number of reads constituting the region of matched contigs (coverage). The relative abundance of different bacteria was denoted by the relative percentage of weight (number of bases) per order, class, phylum and super-kingdom (Table 3.5). The total of 2,100 contigs of microbes revealed in BLASTn and 3,878 contigs in BLASTx, equivalent to 2,379 and 4,277 hits in BLAST research were analyzed. The analyzed contigs were equivalent to a total of 48,811,322 b.p. which contributes to 7.41 % of the total sequenced nucleotides. The microbiome and virome of AIDS plasma were analysed. The virome was separately described in Chapter 4.

While environmental bacteria such as *Bacillus bataviensis* (Brodie *et al.*, 2007) were rarely found, the HIV associated microbiome was dominated by bacteria from the orders Pseudomonadales (35,297,430 b.p., 72.31%), Lactobacillus (2,400,333 b.p., 4.92%), Burkholderiales (1,717,225 b.p., 3.52%), Bacillales (1,264,882 b.p., 2.59%) and Enterobacteriales (1,212,526 b.p., 2.48%) (Figure 3.3).

Table 3.5 Summary of taxonomy information for the microbes (excluding endogenous viruses) found in the HIV/AIDS patients group.

Taxonomy				BLASTN (nt)		BLASTX (nr)		TOTAL			
Super-kingdom	Phylum	Class	Order	Frequency of hit	Relative abundance (b.p.)	Frequency of hit	Relative abundance (b.p.)	Frequency of hit	Relative abundance (b.p.)	Percentage of TOTAL per phylum	Percentage of TOTAL per order
Archaea				1	694	12	112,507	13	113,201	0	0
Bacteria				**2,337**	**12,360,376**	**4,144**	**36,337,745**	**6,481**	**48,698,121**		
	Acidobacteria			0	0	1	3,130	1	3,130	0	0
	Actinobacteria			58	92,302	46	335,708	104	428,010	0	0
	Aquificae			0	0	3	38,521	3	38,521	0	0
	Bacteroidetes			10	5,585	63	425,807	73	431,392	0	0
	Chlorobi			2	39,390	12	83,928	14	123,318	0	0
	Chloroflexi			0	0	5	59,730	5	59,730	0	0
	Cyanobacteria			2	4,128	21	219,061	23	223,189	0	0
	Deinococcus-Thermus			0	0	2	52,553	2	52,553	0	0
	Elusimicrobia			0	0	1	18,656	1	18,656	0	0
	Fibrobacteres			1	2,538	3	28,057	4	30,595	0	0
	Firmicutes			248	760,846	721	3,976,556	969	4,737,403	0	
		Bacilli		**242**	**752,840**	**535**	**2,912,375**	**777**	**3,665,215**		
			Bacillales	78	162,295	197	1,102,587	275	1,264,882		0
			Lactobacillales	164	590,544	338	1,809,788	502	2,400,333		0
		Clostridia		**6**	**8,007**	**173**	**1,005,367**	**179**	**1,013,373**		
			Clostridiales	NA		163	938,950	163	938,950		0
			Halanaerobiales	NA		1	2,087	1	2,087		0
			Natranaerobiales	NA		1	865	1	865		0
			Thermoanaerobacterales	NA		8	63,465	8	63,465		0
			undef	6	8,007	0	0	6	8,007		0
		Erysipelotrichi		**0**	**0**	**13**	**58,815**	**13**	**58,815**		0
	Fusobacteria			1	797	11	42,400	12	43,197	0	0
	Gemmatimonadetes			0	0	2	13,244	2	13,244	0	0
	Lentisphaerae			0	0	2	9,079	2	9,079	0	0
	Nitrospirae			0	0	2	34,998	2	34,998	0	0
	Planctomycetes			0	0	2	18,832	2	18,832	0	0
	Proteobacteria			1,997	11,427,014	3,222	30,798,152	5,219	42,225,166	1	
		Alphaproteobacteria		**6**	**20,136**	**68**	**471,203**	**74**	**491,338**		0
		Betaproteobacteria		**153**	**1,144,756**	**123**	**1,190,667**	**276**	**2,335,423**		
			Burkholderiales	130	936,154	76	781,071	206	1,717,225		0
			Gallionellales	0	0	1	7,352	1	7,352		0
			Hydrogenophilales	0	0	1	2,681	1	2,681		0
			Methylophilales	6	29,604	4	31,551	10	61,155		0
			Neisseriales	13	148,770	32	292,303	45	441,072		0
			Nitrosomonadales	1	8,765	4	36,441	5	45,206		0
			Rhodocyclales	2	14,855	5	39,269	7	54,124		0
			undef	1	6,608	0	0	1	6,608		0
		Deltaproteobacteria		**0**	**0**	**28**	**221,190**	**28**	**221,190**		0
		Epsilonproteobacteria		**2**	**2,700**	**18**	**173,318**	**20**	**176,018**		0
		Gammaproteobacteria		**1,838**	**10,259,422**	**2,982**	**28,724,362**	**4,818**	**38,983,784**		
			Aeromonadales	1	2,632	5	15,251	6	17,883		0
			Alteromonadales	51	257,330	42	397,484	93	654,814		0
			Cardiobacteriales	3	11,510	3	39,847	6	51,357		0
			Chromatiales	1	7,853	15	149,714	16	157,567		0
			Enterobacteriales	99	276,489	99	936,038	198	1,212,526		0
			Legionellales	0	0	7	80,734	7	80,734		0
			Oceanospirillales	9	34,537	14	146,122	23	180,659		0
			Pasteurellales	42	156,040	39	375,589	81	531,628		0
			Pseudomonadales	1,609	9,400,391	2,689	25,897,039	4,298	35,297,430		1
			Thiotrichales	1	3,276	4	61,087	5	64,363		0
			undef	3	13,070	14	113,731	17	126,801		0
			Vibrionales	14	90,105	41	436,008	55	526,114		0
			Xanthomonadales	3	6,190	10	75,718	13	81,908		0
		undef		**0**	**0**	**3**	**17,413**	**3**	**17,413**		0
	Spirochaetes			0	0	5	22,211	5	22,211	0	0
	Synergistetes			0	0	3	19,381	3	19,381	0	0
	Tenericutes			5	9,285	4	53,130	9	62,415	0	0
	Thermotogae			0	0	2	3,558	2	3,558	0	0
	Verrucomicrobia			0	0	7	53,541	7	53,541	0	0
	undef			13	18,490	4	27,513	17	46,003	0	0
		Grand total		**2,379**	**12,361,070**	**4,277**	**36,450,252**	**6,656**	**48,811,322**	**1**	1



Order (Phylum)	Nucleotides	Percentage
Pseudomonadales	35,297,430	72.31%
:tobacillales	2,400,333	4.92%
kholderiales	1,717,225	3.52%
3acillales	1,264,882	2.59%
robacteriales	1,212,526	2.48%
lostridiales	938,950	1.92%
'omonadales	654,814	1.34%
steurellales	531,628	1.09%
'ibrionales	526,114	1.08%
aisseriales	441,072	0.90%
tinobacteria	428,010	0.88%
cteroidetes	431,392	0.88%
Chlorobi	123,318	0.25%
anobacteria	223,189	0.46%
scellaneous	2,620,439	5.37%
Total	48,811,322	100.00%

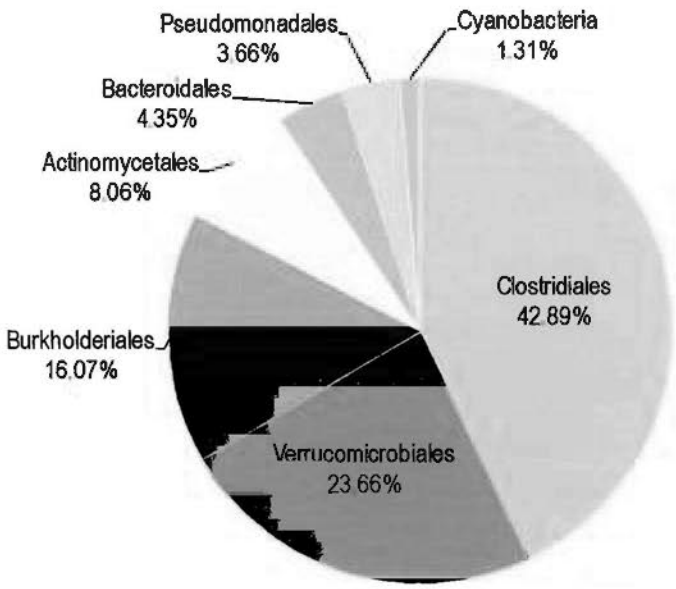
Figure 3.3 The relative abundance of microbial genomes in HIV/AIDS patients group. The relative abundance of microbes in the AIDS patients. The left panel shows the most abundant order of bacteria, dominated by Pseudomonadales (72.31%), Lactobacillales (4.92%) and Burkholderiales (3.52%), etc. The right panel showed the less abundant micro-organisms per phylum.

3.3.4 Analysis of microbial materials in healthy adults

For the control group of healthy adults, a total 410,000,000 b.p. were generated. Contigs assembly was performed to this data set in the same way of analyzing the HIV/AIDS plasma microbiome. A total of 337,857 contigs longer than 100 b.p. were analyzed for their taxonomy and abundance. This includes 140,653 contigs of human sequences which makes up to 141,246,585 b.p. (34.45% of sequenced microbiome). After excluding the contigs for viruses and human, a total of 25 hits contigs of microbes were generated in BLASTn (nt) and BLASTx (nr) (Table 3.6). These microbial contigs made up to 34,408 b.p. which was 0.0084% of the sequenced nucleotides. All the microbial materials in normal plasma come from bacteria. Unlike the AIDS plasma microbiome, the most dominant order of microbes in normal plasma was Clostridiales (14,758 b.p., 42.89%), followed by Verrucomicrobiales (8,141 b.p., 23.66%) and Burkholderiales (5,528, 16.07%) (Figure 3.4). Intriguingly, there was only one contig for Clostridiales in the normal microbiome but it contributed to more than 14 k.b. in weight.

Table 3.6 Summary of taxonomy information for bacterial materials found in the healthy adults.

Taxonomy				BLASTN (nt)		BLASTX (nr)		TOTAL		Percentage of TOTAL per phylum	Percentage of TOTAL per order
Super-kingdom	Phylm	Class	Order	Frequency of hit	Relative abundance (b.p.)	Frequency of hit	Relative abundance (b.p.)	Frequency of hit	Relative abundance (b.p.)		
Bacteria											
	Actinobacteria									8.06%	
		Actinobacteria (class)	Actinomycetales	0	0	1	2,773	1	2,773		**8.06%**
	Bacteroidetes									4.35%	
		Bacteroidia	Bacteroidales	1	951	1	546	2	1,497		**4.35%**
	Cyanobacteria									1.31%	
		undef	Oscillatoriales	1	450	0	0	1	450		**1.31%**
	Firmicutes									42.89%	
		Clostridia	Clostridiales	0	0	1	14,758	1	14,758		**42.89%**
	Proteobacteria									19.73%	
		Betaproteobacteria	Burkholderiales	15	5,062	1	466	16	5,528		**16.07%**
		Gammaproteobacteria	Pseudomonadales	2	817	1	444	3	1,261		**3.66%**
	Verrucomicrobia									23.66%	
		Verrucomicrobiae	Verrucomicrobiales	0	0	1	8,141	1	8,141		**23.66%**
			Grand total	**19**	**7,280**	**6**	**27,128**	**25**	**34,408**	100.00%	**100.00%**

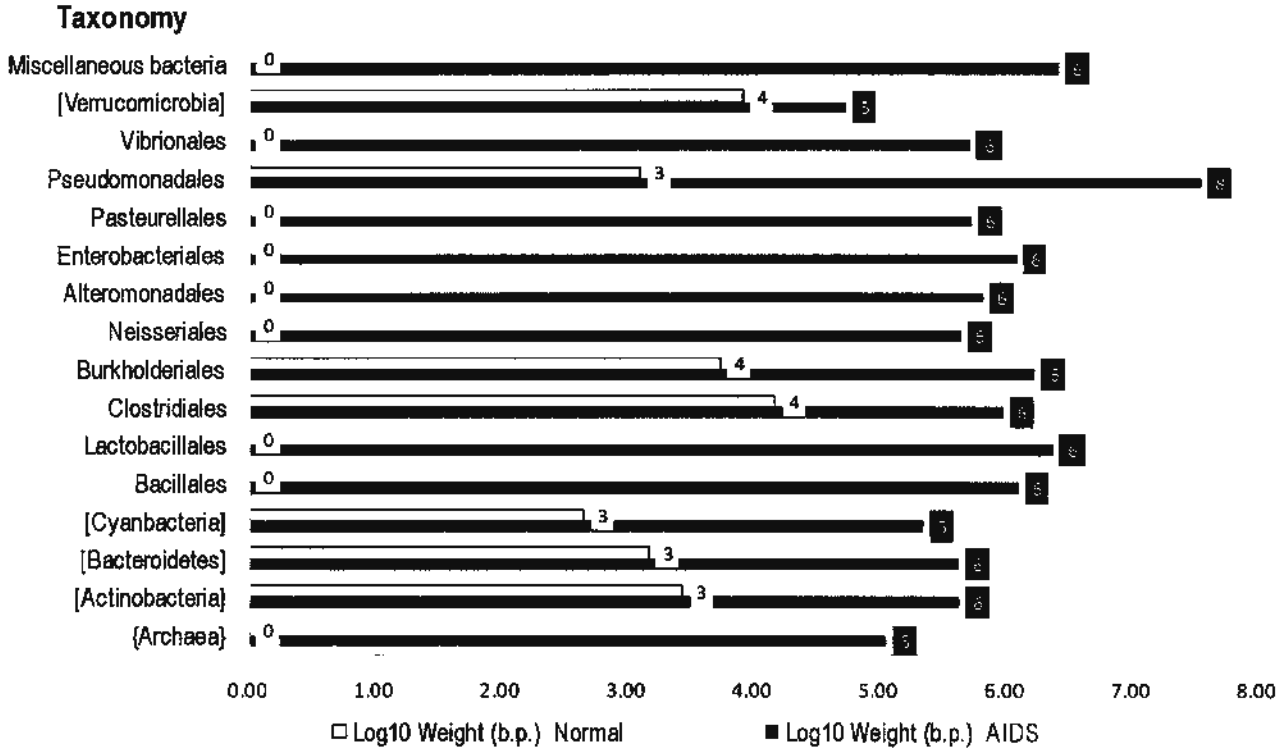


Order	Nucleotides	Percentage per order
Clostridiales	14,758	42.89%
Verrucomicrobiales	8,141	23.66%
Burkholderiales	5,528	16.07%
Actinomycetales	2,773	8.06%
Bacteroidales	1,497	4.35%
Pseudomonadales	1,261	3.66%
Cyanobacteria	450	1.31%
Total	**34,408**	**100.00%**

Figure 3.4 The relative abundance of microbial genomes in healthy adults. The relative abundance of microbes in normal control group was denoted in order and was dominated by Clostridiales (42.89%), followed by Verrucomicrobiales (23.66%) and Burkholderiales (16.07%).

3.3.5 Comparison of microbial materials in HIV/AIDS patients and normal adults

For comparison, HIV/AIDS plasma microbiome and normal plasma microbiome were presented as relative weight in $\log_{10}$ scale against the taxonomy of microbes in the finest rank, given that the preciseness of classification was maintained (Figure 3.5). All the genetic materials of bacteria found in normal healthy adults were also found in HIV/AIDS patients. These include Verrucomicrobia, Pseudomonadales, Burkholderiales, Clostridiales, Cyanobacteria, Bacteroidetes and Actinobacteria. By comparison, HIV/AIDS carry some specific bacteria from Vibrionales, Pasteurellales, Enterobacteriales, Alteromonadales, Neisseriales, Lactobacillales and Bacillales. They also carry some Archaea materials and quite a number of miscellaneous bacteria of some may be unknown to human.

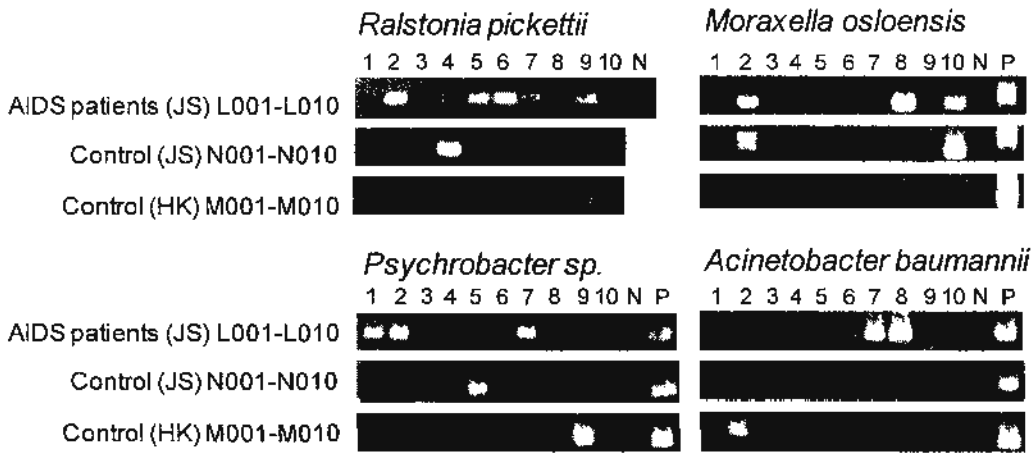


Taxonomy	Weight (b.p.)		Log10 Weight (b.p.)	
	AIDS	Normal	AIDS	Normal
{Archaea}	113,201	0	5.05	0.00
[Actinobacteria]	428,010	2,773	5.63	3.44
[Bacteroidetes]	431,392	1,497	5.63	3.18
[Cyanbacteria]	223,189	450	5.35	2.65
Bacillales	1,264,882	0	6.10	0.00
Lactobacillales	2,400,333	0	6.38	0.00
Clostridiales	938,950	14,758	5.97	4.17
Burkholderiales	1,717,225	5,528	6.23	3.74
Neisseriales	441,072	0	5.64	0.00
Alteromonadales	654,814	0	5.82	0.00
Enterobacteriales	1,212,526	0	6.08	0.00
Pasteurellales	531,628	0	5.73	0.00
Pseudomonadales	35,297,430	1,261	7.55	3.10
Vibrionales	526,114	0	5.72	0.00
[Verrucomicrobia]	53,541	8,141	4.73	3.91
Miscellaneous bacteria	2,577,015	0	6.41	0.00
Grand total	**48,811,322**	**34,408**		

Figure 3.5 Relative abundance of microbes in the HIV/AIDS patients compared to that in healthy aduts. The relative abundance of different groups of microbes was presented in Log_{10} scale in respective to the finest rank of taxonomy, basically in order rank. The broader level of taxonomy includes phylum, indicated in parenthesis ( ) or class indicated in square bracket [ ] or super-kingdom in curly brace { }.

3.3.6 Amplification of bacterial gene materials in HIV/AIDS patients and healthy adults

Given the fact that bacteria are abundant in HIV/AIDS patients, we then examined their presence in individual samples by conventional PCR. The selected targets include several long contigs having a high nucleotide similarity to existing sequences (Table 3.3) such as contigs matched with *Ralstonia pickettii* in the order Burkholderiales. Besides *Ralstonia pickettii* which were consistently detected by PCR (Figure 3.6), *Moraxella osloensis* and *Psychrobacter sp.* from the family Moraxellaceae of Pseudomonadales were also detected when using primers which were both contig sequence specific and bacterial genome specific. In line with the current understanding of opportunistic pneumonia pathogens in HIV/AIDS patients (Manfredi *et al.*, 2001), *Acinetobacter baumannii* was successfully amplified from our patients. Unexpectedly, several of these bacteria genes were amplified from the healthy individuals we used in the Illumina Solexa sequencing (the HK group). We therefore tried to detect the same groups of bacteria in the environment-corresponding healthy group by recruiting another groups of sero-negative healthy people from Jiangsu. As shown in Figure 3.6, bacterial genes were detected in some of the healthy people from Jiangsu. Nevertheless, the relative detection rate of bacterial gene fragments in sero-negative groups from Jiangsu and healthy Hong Kong adults was much lower than that in the HIV/AIDS patients.



Footnote: JS- Jiangsu, HK- Hong Kong, N- negative control, P- positive control

Figure 3.6 Amplification of bacterial genes from HIV/AIDS patients. The presence of *Ralstonia pickettii, Moraxella osloensis, Psychrobacter sp. and Acinetobacter baumannii* in AIDS patients and control individuals was validated. The identities of the products were sequenced for validation.

3.3.7 Identification of gut microbes in AIDS plasma microbiome

We noticed that Lactobacillales and Enterobacteriales which were typically found in the gastrointestinal tract accounted for a significant proportion of the HIV/AIDS plasma microbiome (4.92% and 2.48%). These were not found in the control group of healthy adults. We further investigated the identity of these two groups of microbes by comparing them to the recently published gut microbiome (Brodie *et al.*, 2007). Of the 273 contigs aligned with gut microbiome, *Lactobacillus sp.* (94/273), *Enterococcus sp.* (48/273) and *Clostridium sp.* (29/273) were the most commonly found microbes (Figure 3.7). The detailed identity of these bacteria was given in Appendix 2.

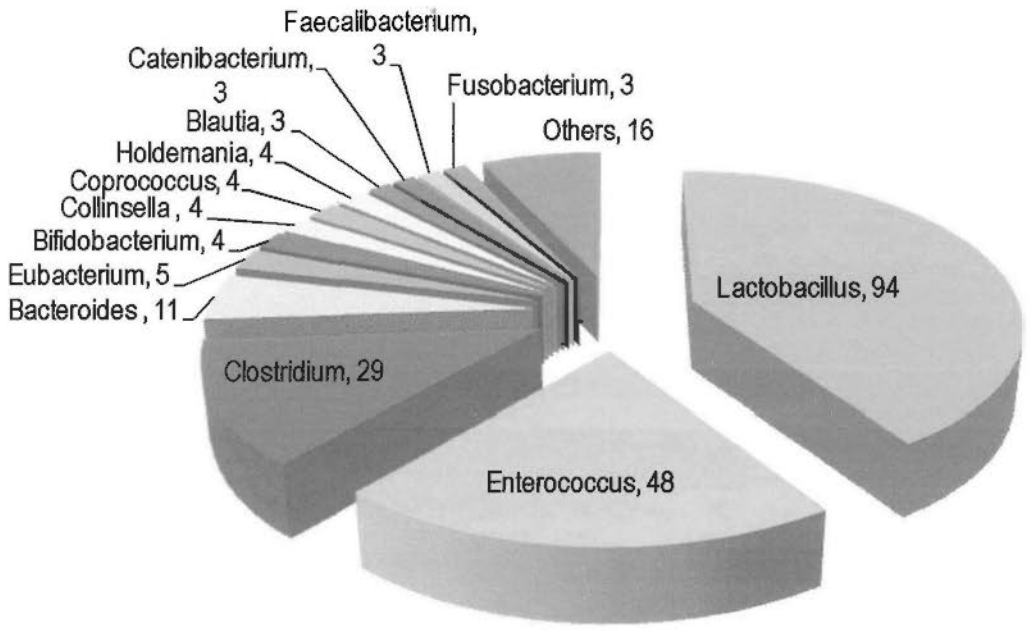


Figure 3.7 Abundance of gut microbes in HIV/AIDS patient plasma. HIV/AIDS plasma microbiome was compared to the gut microbiome catalogue launched by MetHIT Consortium (Qin *et al.*, 2010) and shown by the number of contigs. Out of the 273 contigs aligned with the gut microbiome, *Lactobacillus sp.*, *Enterococcus sp.* and *Closteridum sp.* were most abundant.

3.3.8 Application of high-throughput sequencing in identifying potential novel gene bacterial genes

As described in section 3.3.1, there were more several long contigs in HIV/AIDS plasma microbiome that were longer than 3 k.b. but only less than 400 b.p. on them were matched to existing nucleotide databases when achieved on Jan 2010. We have then analyzed their gene structure in order to determine the novelty of these contigs. For illustration, the longest contig in the dataset was analyzed using MetaGene Annotator (Noguchi *et al.*, 2008). The results showed that there were ten predicted open reading frames (ORF) on the contig (Figure 3.8). The nucleotide sequences of the individual ORF were aligned with the reference sequences in BLASTx search with percentage identity of 41-72%, according to the database in Jan

2010. All the ten aligned targets belong to the phylum Firmicutes in the order Lactobacillales. ORF A, C-E, H-J were aligned with the order Lactobacillales whereas B and F aligned to the order Bacillus and G to the order Clostridiales.

We continued to analyze these long contigs until in Nov 2010 we determined their identity. With the submission of the draft genome for many bacteria by the Human Microbiome Project (HMP) (as described in Chapter 1), we found that the longest contig belongs to *Aerococcus viridans* in the phylum Firmicute. The *A. viridians ATCC 11563* strain was collected from the urogenital tract of an infected individual. The genome (ADNT00000000.1) contains 90% of a core set of bacterial genes with >30% identity and >30% length (Appendix 3). We identified 24 genes of *A. viridans* in 13 contigs found in the HIV/AIDS plasma microbiome. Figure 3.9 shows the location of these genes along the genome, as indicated by numeric. The location of the contigs was presented as alphabets. The details of the mapped genes were given in Table 3.7. The selected 13 contigs in AIDS plasma microbiome contributed to 58,343 b.p. unique sequence in total, equivalent to 2.91% of the whole genome of *A. viridans*.

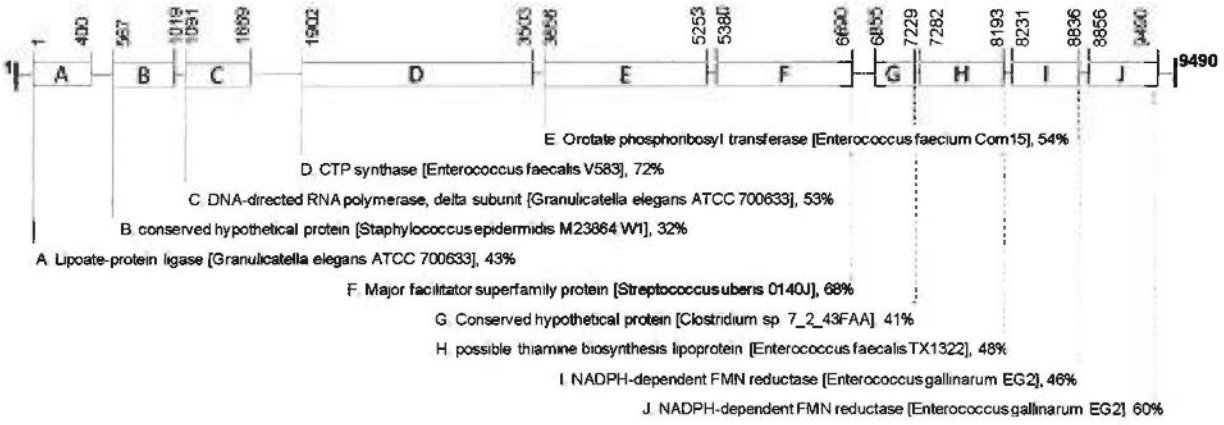


Figure 3.8 Gene structure prediction of the longest contig in HIV/AIDS plasma microbiome by MetaGeneAnnotator. This long contig contains ten predicted open reading frames, annotated as A-J. For each predicted gene, the protein matches with the highest alignment score was described with the identity in percentage.

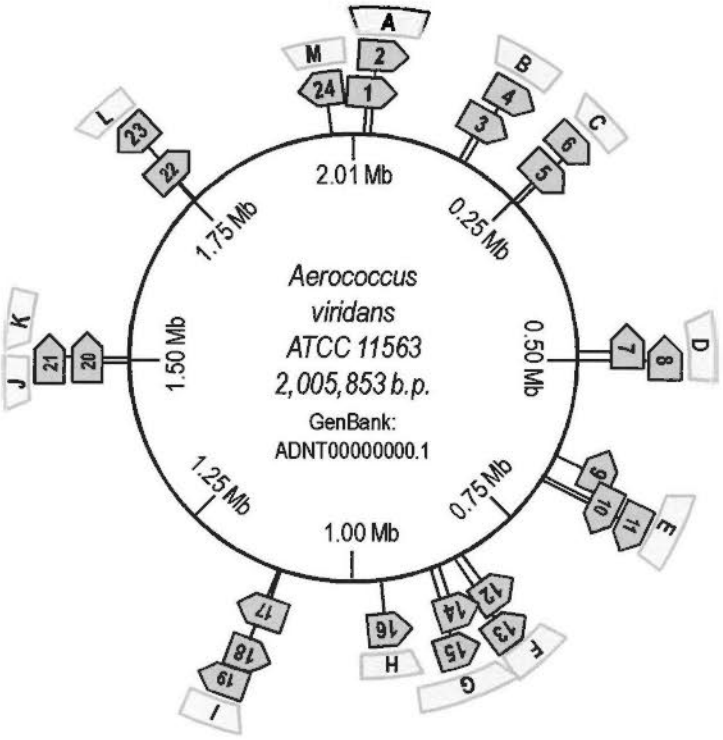


Figure 3.9 Genetic map of *Aerococcus viridans ATCC11563* and the mapped contigs. The localization of the contigs found in AIDS plasma microbiome was shown in grey boxes. The blue arrows indicate the *A. viridans* genes aligned with the contigs with the orientation of transcription.

Table 3.7 Contigs in HIV/AIDS plasma microbiome representing the genes of *A.*

Contig	Length	Coverage	Gene	Accession number	Protein	Product name	Locus	Length	Start	End	Strand
A	4479 b p	8 11x	1	ZP_06806958 1	EFG50643 1	alpha-amylase	amy	631 a a	23,121	24,716	+
			2	ZP_06806960 1	EFG50645 1	DNA repair protein RadA	radA	460 a a	25,406	26,788	+
B	5461 b p	6 15x	3	ZP_06807080 1	EFG50516 1	multidrug ABC superfamily ATP binding cassette transporter, ABC protein	mdlB	603 a a	153,842	155,653	+
			4	ZP_06807081 1	EFG50517 1	multidrug ABC superfamily ATP binding cassette transporter, ABC protein	mdlA	589 a a	155,640	157,409	+
C	3747 b p	9 99x	5	ZP_06807169 1	EFG50421 1	A/G-specific adenine glycosylase	mutY	412 a a	251,167	252,405	+
			6	ZP_06807171 1	EFG50423 1	serine-rich adhesin for platelets	-	420 a a	253,316	254,578	+
D	3824 b p	12 58x	7	ZP_06807404 1	EFG50194 1	TRAP tripartite ATP-independent transporter, binding protein	dctP	400 a a	497,656	498,858	-
			8	ZP_06807406 1	EFG50196 1	alpha-acetolactate decarboxylase	budA	234 a a	499,716	500,420	-
E	5530 b p	11 7x	9	ZP_06807548 1	EFG50009 1	glutathione S-transferase domain protein	-	318 a a	651,243	652,199	-
			10	ZP_06807549 1	EFG50010 1	tetrahydrofolate synthase	folC	425 a a	652,662	653,939	+
			11	ZP_06807551 1	EFG50012 1	dihydropteroate synthase	folP	279 a a	654,573	655,412	+
F	3018 b p	13 37x	12	ZP_06807673 1	EFG49932 1	phosphomethylpyrimidine kinase	thiD	274 a a	783,096	783,920	-
			13	ZP_06807674 1	EFG49933 1	thiamine-phosphate diphosphorylase	thiE	219 a a	783,921	784,580	-
G	9490 b p	10 65x	14	ZP_06807742 1	EFG49882 1	major facilitator family protein	-	450 a a	837,911	839,263	-
			15	ZP_06807745 1	EFG49885 1	CTP synthase	pyrG	533 a a	841,104	842,705	-
H	4057 b p	8 34x	16	ZP_06807854 1	EFG49755 1	glycerol-3-phosphate dehydrogenase	-	452 a a	950,778	952,136	-
I	4942 b p	8 57x	17	ZP_06807984 1	EFG49582 1	transcription termination/antitermination factor NusG	nusG	184 a a	1,100,184	1,100,738	+
			18	ZP_06807985 1	EFG49583 1	phosphate transport system regulatory protein PhoU	phoU	240 a a	1,101,092	1,101 814	-
			19	ZP_06807987 1	EFG49585 1	50S ribosomal protein L1	rplA	229 a a	1,102,562	1,103,251	+
J	3057 b p	8 40x	20	ZP_06808420 1	EFG49120 1	PTS family porter	treP	499 a a	1,528,861	1,530,360	+
K	3536 b p	12 97x	21	ZP_06808421 1	EFG49121 1	alpha,alpha-phosphotrehalase	treC	570 a a	1,530,399	1,532,111	+
L	3470 b p	7 18x	22	ZP_06808608 1	EFG48972 1	PTS family beta-glucosides porter, IIABC component	scrA	665 a a	1,747,902	1,749,899	+
			23	ZP_06808610 1	EFG48974 1	ribosome-associated GTPase EngA	engA	436 a a	1,750,585	1,751,895	-
M	3732 b p	13 30x	24	ZP_06808841 1	EFG48744 1	possible IS1272 transposase	-	339 a a	1,987,622	1988641	-

3.4 Discussion

3.4.1 Determination of HIV/AIDS plasma microbiome by next-generation sequencing

While highly active antiretroviral therapy (HAART) has improved the clinical outcome of HIV patients, access and adherence to treatment remain problems in many places, particularly developing and transitional countries. In 2008, only about 42% of the 10 million HIV/AIDS patients who required immediate life-saving treatment were receiving it (AVERT, 2010). Against the background of declining immunity, treatment naïve HIV/AIDS patients are vulnerable to microbial infections. It is possible that microbes circulating in their blood stream during asymptomatic period could be a source of subsequent diseases. Compounded by suboptimal access to HAART, this predisposition to opportunistic infections could be a cause for concern.

In studying the potential bacteremia status of our subjects, asymptomatic infections in these treatment naïve patients was revealed by the amplification of microbial 16S rRNA gene from plasma DNA. The bacteria identified by 16S rRNA gene sequencing were mostly non-pathogenic and rarely found in healthy adults. Therefore, we suggest that these 'non-pathogenic' bacteria can be a threat to immuno-compromised patients and their pathogenicity should be redefined in the context of HIV co-infection. It has been suggested that 16S rRNA gene sequencing might not be a sensitive method to detect plasma bacterial DNA (Ferri *et al.*, 2010). Therefore, we had used the objective high-throughput Illumina Solexa sequencing technology in profiling the plasma microbiome. The advantages and limitations of next-generation sequencing in metagenome study will be further discusses in Chapter

4. The reliability of the contig assembly was validated by the BLAST results and the PCR results, but still we were surprised to find many of the assembled contigs did not resemble any known genes in microbes. This may imply that HIV/AIDS plasma microbiome contains bacteria which were not included in currently available genome databases or even unknown to humans.

Members of the order Pseudomonadales such as *Psychrobacters*, *Moraxella osloensis* and *Acinetobacter baumannii* were found by sequencing and were later amplified in the validation PCR experiments. Little is known of the pathogenicity of human *Psychrobacters*, though they have been isolated from patients with meningitis (Lloyd-Puryear *et al.*, 1991). *Moraxella osloensis* is an opportunistic human pathogen found to cause several human diseases, including endocarditis and osteomyelitis (Maayan *et al.*, 2004; Tan and Grewal, 2001). Identification of these two bacteria may explain some of the HIV associated bacterial infections which were not characterized in clinical microbiology laboratories. Supported by previous reports on other pneumonia pathogens in the same order (Apisarnthanarak and Mundy, 2005; Huang *et al.*, 2006; Manfredi *et al.*, 2001; Sadikot *et al.*, 2005), we suggest that the substantial proportion of Pseudomonadales in HIV/AIDS plasma microbiome could explain the life-threatening respiratory infections in immuno-compromised patients after the exclusion of well-known pathogens such as *M. tuberculosis* (Comas and Gagneux, 2009). Besides Pseudomonadales, *Ralstonia picketii* was found in more than half of the studied HIV/AIDS patients. Not being associated with any distinctive diseases, *Ralstonia pickettii* is occasionally isolated from a variety of clinical specimens, such as the blood and sputa of patients with cystic fibrosis (Coenye *et al.*, 2002; Pellegrino *et al.*, 2008). The results described

here is the first study to associate this bacterium with HIV infection.

3.4.2 Translocation of gut microbes into systemic circulation

We have also compared the HIV/AIDS plasma microbiome with the reported human gut microbiome (Qin *et al.*, 2010). Despite the fact that many bacteria in the gut are also present in the environment, we suspect that the gut may be a major source of these microbial materials. As we know, HIV preferentially infects CD4+ T-lymphocytes, and the depletion of these CD4+ cells in the circulating system reflects the status of disease development. In fact, the gut mucosal layer is more abundant in activated CD4+ cells than in the circulating system, thus the majority of the depleted CD4+ T-cells in HIV infection reside in the gut (Brenchley *et al.*, 2004; Veazey *et al.*, 1998). These gut CD4+ T-cells express high level of chemokine coreceptor CCR5, one of the major receptors for the entry of HIV (Johnson, 2008). We therefore speculate that when the activated CD4+ cells in the gut mucosal layer are largely depleted in HIV infection, the microbial elements can escape from immune surveillance and enter the blood circulation. It has been recently suggested that the microbial elements from the damaged gut are the drivers for a progressive CD4+ T-cell depletions (Lederman, 2010). On the other hand, as suggested by these authors, the translocation of bacterial DNA could be attributed to the depletion of CD4+ T-cells in the mucosal layer of gastrointestinal tract in chronic HIV infection (Brenchley *et al.*, 2006; Douek, 2007; Jiang *et al.*, 2009; Nelson *et al.*, 2010). Here we provide the evidence that genetic materials of gut microbes are present in the patient plasma. A recent study has reported the microbiome of normal body sites including the gastrointestinal tract, oral cavity, urogenital/ vaginal

tract, skin, and respiratory tract (Nelson *et al.*, 2010). The results show the preferential distribution of microbes according to the growing habitats of the body. Hence, we suggest that the detection of the 'normal' gut microbes, whose pathogenicity is restricted by the nature of their populating tissue, may reflect the breakdown of immunological surveillance and the wandering of the implicated bacteria into more nutrient and oxygen-rich blood niches (Douek, 2007). Indeed it is believed that translocation of gut microbes into the systemic circulation is the driving force for most of the severe sepsis (with unknown causes) in patients with critical illness (Alverdy and Chang, 2008). Nevertheless, these microbes are unlikely to cause mortality unless being accumulated to an excessive population that they start to contend the immunity.

3.4.3 Detection of a considerably large amount of *A. viridans* in HIV/AIDS patients

Based on the BLAST alignment results, the HIV/AIDS plasma microbiome is constituted by unclassified micro-organisms, represented by the large number of contigs that were not matched to existing databases. Notably, we found *A. viridans*, which was unidentified before Nov 2010. Although we were not aiming to determine the complete genome of any microbes, we found around 3% of *A. viridans* whole genome, and possibly in contigs in this study fill some of the gap in the draft genome of this bacterium. *A. viridans* is usually associated with urogenital tract infection (pediatrics) (Leite *et al.*, 2011), endocarditis (Popescu *et al.*, 2005) and septic arthritis (Taylor and Trueblood, 1985). It has also been associated with granulocytopenia (abnormally low concentration of granulocytes in the blood) (Uh *et al.*, 2002) and one case of HIV patients (Razeq *et al.*, 1999). Here we provide the

evidence for the presence in HIV/AIDS patients a considerable large amount of *A. viridans*, which subsequently made up the longest contigs in the HIV/AIDS plasma microbiome.

3.4.4 Detection of microbe materials in apparent healthy adults

Surprisingly, we have also successfully sequenced the genetic materials of bacteria from the healthy adults whose plasma was assumed to be free of microbes (except viruses). This may be due to the presence of circulating bacterial gene fragments which await removal by the immune system. Alternatively, this may imply that previously unattended normal microbes are able to circulate in apparently healthy people yet somehow evade immune elimination (HMP, 2010; Uzonna *et al.*, 2001). Our speculation is further supported by the detection of non-disease-causing isolates such as *Psychrobacter sp.* in human blood by PCR (HMP, 2010).

This normal plasma microbiome is distinctly different from the HIV/AIDS plasma microbiome in the composition and abundance of individual species. The difference in these two sets of microbiome will be important in understanding the risk of getting opportunistic infections faced by the immuno-comprised people when compared to immuno-competent adults. The presence of bacterial materials in the systemic circulation of healthy adults was further studied and described in Chapter 5 in this thesis.

3.4.5 Future work on the unmatched contigs in HIV/AIDS microbiome

The future work in this part of the project shall mainly focus on the analysis of unmatched contigs, particularly on the identification of potential novel bacteria. A

conventional method to identify novel bacteria is called 16S rRNA gene sequencing. In 1987, Carl Woese has successfully showed that 16S rRNA gene was conserved among bacteria of the same genera and same species (Woese, 1987). Since this great discovery, 16S rRNA gene has been used as an evolutionary indicator and for taxonomic classification / reclassification of bacteria. Moreover, 16S rRNA gene sequencing has become a useful tool to identify bacteria which are otherwise unidentified by traditional methods e.g. bacteria that are unculturable or grow very slowly such as *Mycobacteria* (Handelsman, 2004; Heller *et al.*, 1996; Woo *et al.*, 2000). This technique also enables a precise determination of bacteria which usually display ambiguous phenotypes and are usually misidentified (Woo *et al.*, 2008). The common procedure of defining a novel bacterium largely relies on the comparison of the novel 16S rRNA gene to the closely related groups of bacteria, which are usually defined by similar bacteria phenotypes, associated diseases, host, reservoirs and 16S rRNA gene conservation(Woo *et al.*, 2008). By characterizing the phenotypic features of the bacterium using typical microbiological methods, the novel bacterium could be defined down to the genera level. Genomic comparison could be done by restriction digestion of the genomic DNA followed by gel electrophoresis. Nowadays, genomic comparison is usually done by comparing 16S rRNA gene sequence. This process is further facilitated by the availability of a lot of 16S rRNA gene sequences of a wide variety of bacteria on Ribosomal Database Project (RDP) (Cole *et al.*, 2009). Bacteria which showed very similar but different from the existing bacteria will be defined as novel and name according to the defined genera. For example, using phenotype characterization and 16S rRNA gene sequencing, *Streptococcus sinensis sp. nov.*, *Laribacter hongkongensis gen. nov., sp. nov.*, *Nitratireductor lucknowense sp. nov.* were identified and named (Manickam *et*

al., 2011; Woo *et al.*, 2002; Yuen *et al.*, 2001).

In comparison to 16S rRNA gene sequencing, next-generation sequencing is primer-independent, thus allows the detection of genome sequences other than the rRNA gene (de Madaria *et al.*, 2005; Woo *et al.*, 2003). Nevertheless, the database of bacterial complete genomes covers much fewer bacteria when compared to the vast 16S rRNA gene sequences on RDP. Thus, usually there are a lot of metagenome sequences found unmatched in the nucleotide BLAST or protein BLAST if the analysis was not done by 16S rRNA sequencing, such as that in our project. Further studies on these unmatched contigs in HIV/AIDS microbiome shall begin with the alignment to the 16S rRNA database. In fact, we had performed such an alignment but did not found any matches. Automated gene structure study could be performed to predict the coding region on these contigs as illustrated in the identification of the longest contig as a member in the phylum Firmicute in our project. Nevertheless, unlike typical isolation of bacteria from a diseased patient with distinct symptoms, metagenomic analysis involves the study of a complex of bacteria which thus is much more complicated in terms of data analysis. Thus, this project shall be extended to a new study on HIV/AIDS associated novel bacteria. The long unmatched contigs found in HIV/AIDS microbiome could be used as a reference sequence. Fresh bacteria complex samples should be isolated from the HIV/AIDS patients and grown on blood agar plate until they are separated into pure culture of individual bacterium. Molecular detection of the reference sequences will allow the association of a particular unmatched contig to a pure bacterial culture. Then the bacteria could be identified using typical phenotype characterization and 16S rRNA analysis as described above.

3.5 Conclusion

Our results build a knowledge base on common pathogens that might be encountered by immuno-compromised individuals when compared to the healthy adults. By taking reference from the microbiome reported in this study, clinicians will be better equipped when assessing HIV patients to provide tailored treatment or prophylaxis. For instance, Pseudomonadales may serve as the first target of consideration for antibiotics selection for patients with respiratory infection but free from *M. tuberculosis*. By focusing on immuno-compromised hosts, the results of our study fill a gap in the understanding of the dynamics of human microbiome, and demonstrate that blood is the traffic hub and potential medium for normal flora. Furthermore, by studying the long unmatched contigs, we may be able to identify genomes of new microbes and start to know the vast world of unknown microbes that are affecting the HIV/AIDS patients.

Chapter 4

Metagenomic comparison of the plasma virome of HIV/AIDS patients and healthy adults

Summary

Apart from the plasma microbiome described in Chapter 3, we have also elucidated the plasma virome in HIV/AIDS patients in parallel with healthy adults. The HIV//AIDS and normal plasma virome share some similarity in the presence of common ubiquitous eukaryotic viruses. The normal virome was mainly composed of viruses from Anelloviridae. The HIV/AIDS viromes was contrasted by the presence of a large proportion of bacteriophages, typical eukaryotic viruses and untypical non-human viruses. Analysis of the endogenous retrovirus genome K revealed the presence of potential novel retroviral coding genes.

4.1 Introduction

The HIV/AIDS plasma microbiome was described and discussed in Chapter 3. This chapter depicts another dominating population of micro-organisms living in human -- the viruses. The genetic sequences of such a large viral community including pathogenic viruses, endogenous viruses and bacteriophages are collectively termed as virome (Haynes M. & Rhower, 2010). Elucidation of the HIV/AIDS plasma virome against the normal virome is important because of the newly emerging paradigm that an illness may be defined by a disruption of the normal 'healthy' virome.

The classical, golden method of viral detection is by culturing the virus in susceptible cells and detecting the virus signatures by different means (Leland and Ginocchio, 2007). Molecular detection of the viral DNA or RNA by polymerase chain reaction (PCR) has been widely used in virology laboratories in diagnosing viral infection (Elnifro *et al.*, 2000). This method, though is much more costly, speeds up the diagnosis and allows detection of uncultivable viruses and. The third method used in the recognition of viral antigens is by enzyme-linked immunosorbent assay (ELISA), western blot or immunofluorescence. All these methods are useful in routine testing in the virology laboratories. But in studying a metagenome of a complex viral sample, these methods will not give a comprehensive picture. A fantastic thought on higher throughput detection is to detect the virus by means of microarray (Chou *et al.*, 2006). But that will involve bioinformatics detection of specific epitopes of the large diversity of viral antigens and is limited to known viruses and the designated genotypes. Unlike bacteria which can be studied using universal primers for the 16S rRNA gene that present in every bacterial genome, viruses do not carry any conserved genes for a 'one-for-all' detection. Taking into the consideration of the great variety of genotypes of viruses and rapid emergence of new genotypes, microarray may not be practically useful at an early stage of investigation. These may account for the reason that a reference human virome is not available until year 2010 when the study of viral metagenomic was greatly facilitated by the advances in next-generation pyrosequencing. These technologies include long read 454 sequencing methods especially useful for full genome study (Roche/ 454 Life Sciences) and short-read sequencing for genome identification purpose (Illumina Solexa, SOLiD) (Schuster, 2008).

4.2 Materials and Methods

4.2.1 Patients information and ethics Statement

The HIV/AIDS patients and healthy adults from Hong Kong were described in Chapter 3. The ethics statement should also refer to section 3.2.1.

4.2.2 Extraction of plasma DNA/ RNA and reverse transcription

Extraction of plasma viral DNA and RNA was done using QIAamp MinElute Virus Spin kit (QIAGEN). Same precautionary measures were adopted as that described in section 3.2.3. The extracted viral RNA was reverse transcribed into complementary DNA by Quantiscript Reverse Transcriptase and random primers as instructed by the manufacturer (QIAGEN).

4.2.3 Amplification of plasma cDNA and sequencing

Same as that described in Chapter 3, the extracted viral DNA and viral cDNA was far from enough for the next-generation sequencing. Therefore, these samples were amplified using GenomiPhi V2 kit (GE Medical Systems). The amplified sample was prepared by pooling 5 ng of plasma cDNA/RNA from each of the ten subjects in the patient group or the control group. The amplification was performed as previously described. Solexa sequencing was performed for the plasma cDNA. The sequence reads were assembled into contigs and aligned with the existing databases using BLASTn and BLASTx. The viral DNA data described in this chapter came from the same set of the plasma DNA sequenced for the microbiome. Chapter 3 only described the portion of bacterial DNA and the viral DNA will be depicted in this chapter, together with the viral RNA/cDNA data-set. All the data analysis was performed in the same way as that used in analyzing the microbiome.

4.2.4 Open reading frame finder

Open reading frames of unknown sequences were predicted using the ORF Finder provided by NCBI at http://www.ncbi.nlm.nih.gov/projects/gorf/.

4.2.5 Phylogenetic tree analysis

Phylogenetic analysis was performed on BioEdit version 7.0.9.0 (Hall, 1999). The input sequences were first aligned with each other. The phylogenetic analysis was evaluated using DNADIST Neighbor Phylogenetic tree version 3.5C with default settings.

4.3 Results

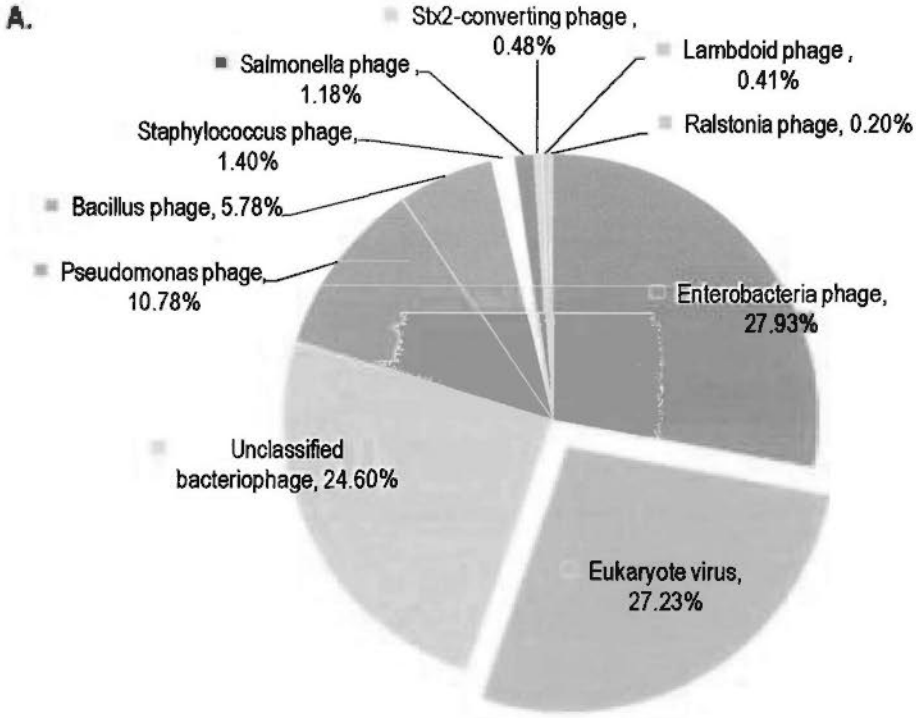
4.3.1 Analysis of HIV/AIDS plasma virome

To determine the genetic materials in a viral community, the genetic materials of DNA viruses and RNA viruses were extracted. Apart from the plasma DNA used for the study described in Chapter 3, we have also isolated the plasma RNA samples from the HIV/AIDS patients. The plasma RNA samples isolated was reverse transcribed into complementary DNA and then sequenced using Illumina Solexa as described in Chapter 2 and 3. A total of 859,104,378 b.p. were generated for HIV/AIDS plasma RNA/cDNA in the Solexa sequencing including 18.70 % for human sequences. Contig assembly and taxonomy analysis was performed with the same method used for analyzing plasma DNA (Chapter 3). The BLASTn and BLASTx results for the viral DNA together with the viral RNA were combined in calculation for the total virome.

Figure 4.1 shows the plasma virome for HIV/AIDS patients. The HIV

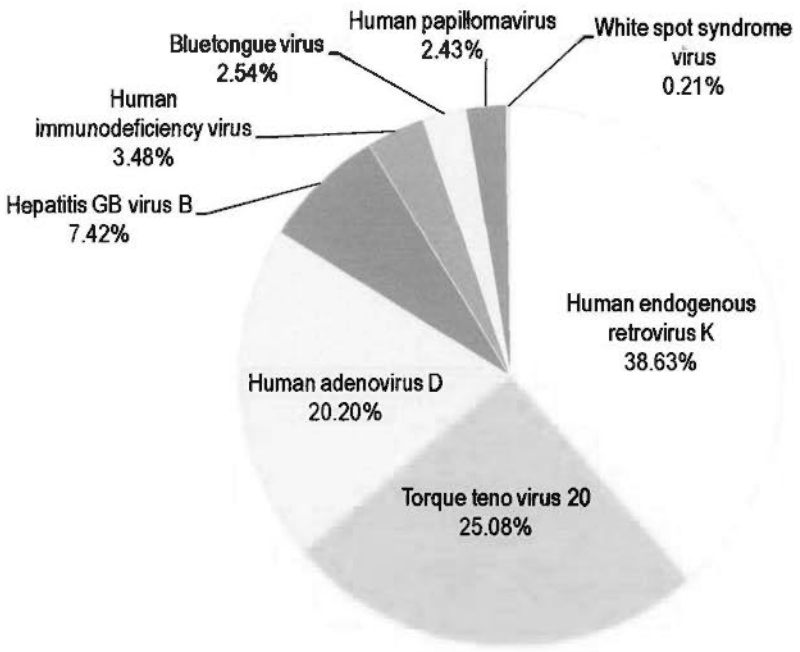
infection was indicated by the detection of HIV sequences which contributes to 0.95% of the total viral population by base pairs. The virome was dominated by bacterial phages which contributed to 72.77% of total plasma viral groups. Eukaryotes viruses contributed to 27.23% of the virome. The most abundant phage group was Enterobacteria phage (27.93%) followed by unclassified bacteriophage (24.60%) and Pseudomonas phage (10.78%) (Figure 4.1A). The most abundant eukaryote virus groups was endogenous retrovirus K (10.52%) followed by Torque teno virus (6.83%). Human adenovirus contributed to considerably large portion of the HIV/AIDS virome (6.60%). The sequencing results also showed that one or more patients in this group were co-infected by hepatitis GB virus B (2.02% of total virome) or human papillomavirus (0.66%). Surprisingly, several non-human viruses were found in this group of immuno-compromised patients. They include the bluetongue virus (0.69%) which usually infects ruminants and the white spot syndrome virus that infects shrimps (0.06%) (Figure 4.1B) (Escobedo-Bonilla *et al.*, 2008; Zhang *et al.*, 2010).

A.



Virus/phage	Weight (b.p.)	% virome
Enterobacteria phage	90,040	27.93%
Eukaryote virus	87,793	27.23%
Unclassified bacteriophage	79,318	24.60%
Pseudomonas phage	34,743	10.78%
Bacillus phage	18,633	5.78%
Staphylococcus phage	4,518	1.40%
Salmonella phage	3,804	1.18%
Stx2-converting phage	1,555	0.48%
Lambdoid phage	1,325	0.41%
Ralstonia phage	650	0.20%
Grand total	**322,380**	**100.00%**

B.



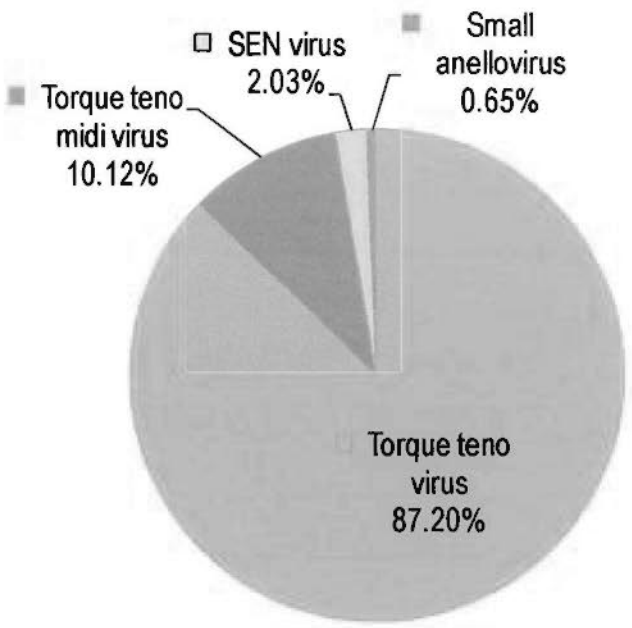
Eukaryote virus	Weight (b.p.)	% virome	% eukaryotes virus
Human endogenous retrovirus K	33,915	10.52%	38.63%
Torque teno virus	22,019	6.83%	25.08%
Human adenovirus	17,737	5.50%	20.20%
Hepatitis GB virus B	6,512	2.02%	7.42%
Human immunodeficiency virus 1	3,058	0.95%	3.48%
Bluetongue virus	2,234	0.69%	2.54%
Human papillomavirus	2,133	0.66%	2.43%
White spot syndrome virus 1	185	0.06%	0.21%
Grand total	**87,793**	**27.23%**	**100.00%**

Figure 4.1 The relative abundance of viral species in HIV/AIDS plasma virome. **A.** The major population of the HIV/AIDS plasma virome was bacterial phages (72.77%), dominated by Enterobacteria phage (27.93%) and unclassified bacteriophage (24.60%). Eukaryotes viruses made up 27.23% of the virome. **B.** Human endogenous retrovirus K contributed to the largest portion of eukaryotic virus genome (38.63%) [10.52% of virome], followed by Torque teno virus (25.08%) [6.83% of virome] and adenovirus D (20.20%) [5.5% of virome].

4.3.2 Comparison between normal plasma virome and HIV/AIDS virome

In contrast, the normal plasma virome did not contain any bacterial phages. It consisted of a few species of viruses. The dominant group of virus was Torque teno virus (TTV) (87.20%) followed by Torque teno midi virus (TTMDV) (10.12%). SEN virus and small anellovirus constituted to 2.03% and 0.65% of the normal virome respectively. No human endogenous retrovirus K was sequenced. No hepatitis viruses or pathogenic viruses are found in this group of healthy people.

In comparison to the HIV/AIDS plasma virome, healthy adults carried more TTV by weight and proportion. TTMDV, SEN virus and small anellovirus were found in the normal virome only. On the other hand, HIV/AIDS virome contained some unique viruses such as human endogenous retrovirus K, human adenovirus, human papillomavirus and hepatitis GB virus B in addition to HIV and a large amount of phages. Non-human bluetongue virus and white spot syndrome virus were also found in the HIV/AIDS virome but not in the normal one.



Virus/ phage	Weight (b.p.)	%
Torque teno virus	439,394	87.20%
Torque teno midi virus	51,003	10.12%
SEN virus	10,224	2.03%
Small anellovirus	3,270	0.65%
Phage	0	0.00%
Total	**503,892**	**100.00%**

Figure 4.2 Plasma virome of healthy adults. Unlike the HIV/AIDS plasma virome, the normal virome did not contain bacterial phages. Torque teno virus (87.2%) and Torque teno midi virus constituted the largest viral communities in this virome. SEN virus (2%) and small anellovirus (0.65%) were also found in the plasma virome.

4.3.3 Identification of Chinese specific human endogenous retrovirus genome

In studying the HIV/AIDS plasma virome, we noticed that many of the contigs representing the human endogenous retrovirus (HERV) K actually had less than 60% similarity to the current HERV genome on NCBI nt/nr protein reference database. As shown in Table 4.1, the nucleotide BLAST results of all the five contigs of HERV showed a nearly 100% identity to human chromosome. However, when the contigs was BLAST against the reference nucleotide translated database using BLASTx, only human endogenous retrovirus K were found. This might represent the HERV integrated in the chromosome of person selected for building the reference human genome. Open reading frame prediction was predicted for these sequences. However, unlike the BLASTx results, no coding frame was found for contig 4 which might be due to the lack of general start and stop codons in these viral sequences (Figure 4.4).

We have compared the phylogenetic distance of these contigs with the known HERV genomes using Cluster W program using Neighbour-joining (Hall, 1999). By considering the length and accuracy of phylogenetic analysis, only contig 2 >NODE_491839 (the longest one) was analyzed. It was found to be polymerase protein of HERV by BLASTx. Figure 4.4 shows the phylogenic tree plotted for this longest contig against the polymerase gene of matched HERV-K genome. The results suggested that this contig was distinctively different from all the existing HERV-K genomes.

Table 4.1 Details of BLAST results for all the contigs representing the human endogenous retrovirus K in HIV/AIDS plasma virome.

Contig	Reference no.	Length	Coverage	Nucleotide BLAST				BLASTX			
				Matched target	Identity	E-value	Target details	Matched target	Identity	E-value	Target details
1	NODE_432919	392	2.90x	gi|21728162|dbj|AP 005209.3|	99.74	0	Homo sapiens genomic DNA, chromosome 18 clone:RP11-138E9, complete sequence	sp/P63135.1/POK 12_HUMAN	56.67	4.00E-16	Human endogenous retrovirus K 1q22 provirus ancestral pol protein
2	NODE_491839	843	4.03x	gi|23396223|gb|AC 130456.2|	99.27	0	Homo sapiens chromosome 16 clone CTA-363E6, complete sequence	sp/P63136.1/POK 17_HUMAN	65.83	4.00E-71	Human endogenous retrovirus K 11q22.1 provirus pol protein
3	NODE_253046	261	7.57x	gi|23396223|gb|AC 130456.2|	99.20	5.00E-121	Homo sapiens chromosome 16 clone CTA-363E6, complete sequence	sp/Q9WJR5.2/PO K_HUMAN	66.67	6.00E-24	Human endogenous retrovirus K 19q12 provirus ancestral pol protein
4	NODE_501411	212	4.65x	gi|23396223|gb|AC 130456.2|	100.00	9.00E-99	Homo sapiens chromosome 16 clone CTA-363E6, complete sequenc	gb/AAD51797.1/A F164614_1	66.67	1.00E-15	Human endogenous retrovirus K 6 Gal-Pro-Pol protein
5	NODE_478229	149	4.03x	gi|29498303|emb|A L391099.12|	100.00	2.00E-62	Human DNA sequence from clone RP11-178A10 on chromosome 10 Contains a novel gene (FLJ34785 FLJ33470)	gb/ABD29044.1	70.45	2.00E-12	Human endogenous retrovirus K envelope glycoprotein

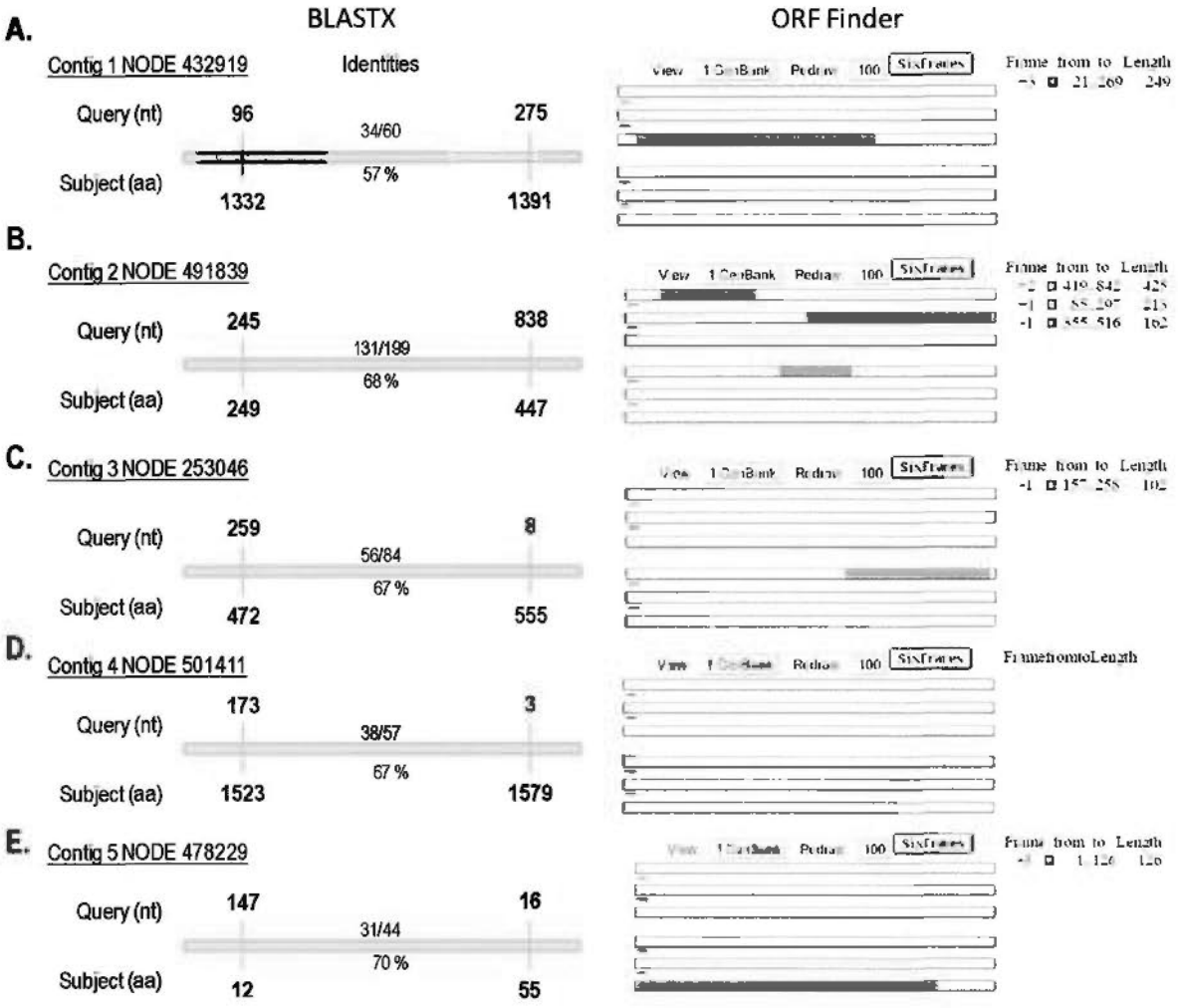


Figure 4.3 Schematic plots for the BLASTx and open reading frame prediction results found for HERV-K. A-E represents contigs 1-5 respectively. Except for contig 4, open reading frames were predicted for the contigs found for the HERV-K. The details of the BLASTx results should refer to Table 4.1.

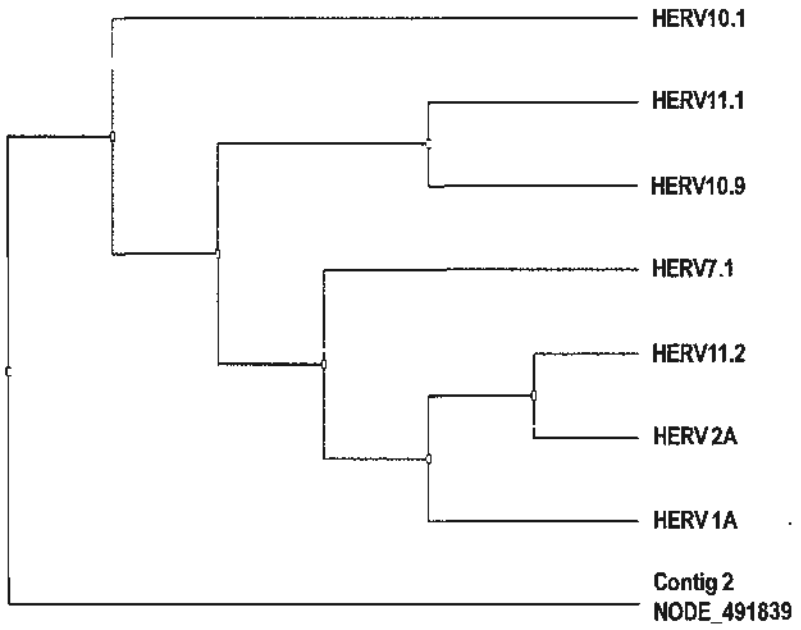


Figure 4.4 Phylogenic relationship of HERV contig 2 (NODE_491839) and its related HERV polymerase reference sequences. The results showed that this contig was distinctly far away from the matched HERV polymerases.

4.4 Discussion

4.4.1 Complex viral community in HIV/AIDS patients and healthy adults

4.4.1.1 Comparison between HIV/AIDS virome and normal virome

The plasma virome in HIV/AIDS patients was depicted and compared to the one in healthy adults. Strikingly, the HIV/AIDS plasma virome contained a large portion of bacteriophage by constituting nucleotides (72.77%). This phage community shared more sequencing space in the short-read sequencing thus subsequently relatively fewer eukaryotes viruses were sequenced. Therefore, the total weight contribution of eukaryotic viruses (87,793 b.p.) in HIV/AIDS virome was even less than the total weight of eukaryotes viruses in healthy adults (506,892 b.p.). This uneven distribution sequencing space may explain the artifacts that 'more' eukaryotes viruses were present in normal healthy people.

The presence of bacteriophage in HIV/AIDS virome may in other way reflect the presence of bacteria in patients' plasma. These bacteria can possibly be living bacteria that are circulating with the systemic blood flow that allows the propagation and host such a large number of bacteriophages. Nevertheless, it is also possible that the phages enter the blood stream by passing through the digestive system as that described for gut microbes in Chapter 3. This can be exemplified by plant virus Pepper Mild Mottle Virus (PMMV) which was proven to had been taken in food and enter the circulation via the gastrointestinal system (Zhang *et al.*, 2006). On the contrary, no bacteriophage was found in the normal virome, suggesting that the microbial elements found in the healthy adults (Chapter 3) may be the remains of the bacteria causing non-symptomatic, transient bacteremia in these healthy study subjects. Phages made up the largest community living at different sites of our human body. It is estimated that the total number of phages on human body is about 10^{15} phages, compared to the estimated number of 10^{14} microbial cells and 10^{13} human cells (Angly *et al.*, 2009; Haynes M. & Rhower, 2010). It is also estimated that a typical healthy adult carries 1,500 genotypes of viruses. Haynes and Rhower suggest there were a total of about 10^{8} virus particles, as estimated by the number of virus-like particles (VLPs) formed. This directly shows that blood can be a favourable niche for virus growth.

4.4.1.2 Potential reactivation of human endogenous retrovirus K genome by HIV

Taken 13 years time of the collective effort from scientists over the globe, the completely reference human genome was available in early 21st century (Lander *et al.*, 2001). The completed human genome consists of 3 billion base pairs.

However, it is estimated that only 3-8 % the human genome is functional (Siepel *et al.*, 2005). These functional regions constitute to 20,000-25,000 genes, encoding for recognizable human proteins. With the sequence for individual human genomes, approximately 3 to 4 million variations were found with respect to the reference genome (Frazer *et al.*, 2009). It was estimated that approximately 45% of our genome is retro-transposons, DNA transposons and viral sequences. It is found that human endogenous retroviruses comprise about 8% of the human genome with 98,000 elements and fragments (Belshaw *et al.*, 2004; Garrison *et al.*, 2007). These viruses are thought to have infected human and integrated in the human genome many millions of years ago. The most actively studied HERV is HERV-K which makes up less than 1% of HERV elements in the human genome. This HERV-K might have been active in the past few hundred thousand years as some individuals carry more copies of this viral genome than other people. However, as human genome replicated and pass from one generation to another, the viral genome was inactivated by deletion, insertion or mutation and is unlikely active at present (Belshaw *et al.*, 2005).

However, with the succeed in reconstituting the HERV-K, it was implicated that the 'dormant' retroviral gene fragment may still possess the ability to encode proteins such as the retrovirus polymerase and envelope protein (Dewannieux *et al.*, 2006; Lee and Bieniasz, 2007). These expressed retrovirus fragment may be associated with some current human diseases. Taken as an example, a group of most studied HERVs is the MS-associated retrovirus W. It was so termed because of its association with an autoimmune disease called Multiple Sclerosis (Mameli *et al.*, 2007). In our project, we have found five contigs which might be the coding genes

of HERV. These HERV genes were not defined in the current human genome but may be the novel sequences of ancestral HERV.

More strikingly, it has been shown that the HERV-K fragment in the human genome can be 'reactivated' by other exogenous retrovirus such as the human immunodeficiency virus (HIV). Being a retrovirus, HIV can be reverse-transcribed by its own reverse transcriptase to form the viral DNA and then get integrated into the human genome. The neighbour 'dormant' HERV genome (if any) can therefore be transcribed together with HIV integrates when the HIV promoter is activated by appropriate transcription factors (Garrison *et al.*, 2007). In addition, the HERV transcript can be exported from the nucleus to the cytoplasm by using the HIV mRNA export machinery Rev (Yang *et al.*, 1999). However, this discovery is not necessarily bad to humankind. It was shown that the co-expressed HERV protein can be used as a tool to recognize the cells infected by HIV. Moreover the cytotoxic immune response induced by the HERV has a certain degree of effective killing of HIV/HERV positive cells. These render the HERV as a surrogate immuno-therapeutic target against HIV (Garrison *et al.*, 2007). Without exogenous activator factors, healthy adults did not express the HERV in their genome, which explains the absence of HERV genome in the normal virome.

4.4.1.3 HIV/AIDS plasma virome

Apart from the HIV genome, common eukaryotic adenoviruses and human papillomavirus were also found in HIV/AIDS plasma virome. Adenoviruses are a group of non-enveloped icosahedral double-strand DNA viruses. Infection by adenovirus is usually asymptomatic in healthy people. In contrast, it can cause

severe disease in immuno-deficient people or in children. Some of the subtypes were proposed to associate with severe pneumonia (subtype 7 and subtype 14) (Halstead *et al*, 2010; Tang *et al*, 2010). Human papillomavirus is a non-enveloped, DNA virus commonly found the plasma of infected female (Sanclemente and Gill, 2002; Stanley, 2010). It was a major causative agent for cervical cancer. Both adenoviruses and human paillomavirus can be found in normal people. However, in the control group used in this study, seemingly none of these ten subjects got infected by either viruses. Hepatitis GB virus B constitutes to 7.42% of eukaryote viruses found in HIV/AIDS patients (equivalent to 2.02% of HIV/AIDS plasma virome). GB virus B belongs to group of GB viruses suspected to cause chronic hepatitis and jaundice (Martin *et al.*, 2003; Simon and Malim, 1996). Persistent infection of GB virus B does not seem to cause a serious problem in these subjects. Actually the virus was found in a small portion of healthy adults. The hepatitis GB virus subtype C, or called hepatitis G virus nowadays, drawn much more attention than the GB virus B because it causes hepatitis C like symptoms. They are more closely related to hepatitis C virus which is a RNA virus. The GB virus C was proposed to inhibit the replication of HIV in a person with co-infection, and thus delay the development of AIDS (Tillmann and Manns, 2001). This interaction between the two viruses does not seem immediately apparent. Instead, it may be mediated by cellular factors. Yet, we did not find any GB virus C in the HIV/AIDS subjects in this study.

We have also found in the HIV/AIDS virome the presence of two non-human viruses. The white spot syndrome virus (WSSV) was originally thought as ubiquitous in marine organisms until it caused a massive death of shrimps in Asia

(Sanchez-Paz, 2010). The pathogenesis of WSSV-induced disease in marine organisms in not well understood, despite the publication of the viral membrane protein complex recently (Chang *et al.*, 2010). Bluetongue (BT) virus is a virus that infects ruminants (Maclachlan *et al.*, 2009). BT infection was frequently described in cattle and less but still often reported in sheep, horses, goats and even zoo carnivores. It causes a characteristic vascular injury, leading to excessive haemorrhage and ulceration in gastrointestinal tract, pulmonary and skeletal muscles (Maclachlan and Guthrie, 2010). Interestingly, both WSSV and BT can be transmitted by arthropods (shrimps and insects) and can be found in food. The roles of these viruses in HIV/AIDS patients shall be further studied.

4.4.1.4 Implication of the presence of Torque teno virus in normal and HIV/AIDS subjects

There are a few groups of low pathogenicity found in the virome of HIV/AIDS and healthy adults. Torque teno virus (TTV) made up a large part of viromes in both HIV/AIDS patients group and healthy adults. It is a small, non-enveloped DNA anellovirus with a circular single-stranded circular DNA genome. TTV was originally thought to cause hepatitis (Nishizawa *et al.*, 1997) and acute respiratory infection (Maggi *et al.*, 2003), but now it seems that it is non-pathogenic in infected healthy people. It can be found in 50-90% of healthy adults and apparently more in Asian populations such as Japanese and Russians accompanied by higher viral load (Shibayama *et al.*, 2001; Vasilyev *et al.*, 2009). This virus can be transmitted by transfusion, sexual, mother-to-child, faecal-oral or even through contaminated drinking water (Gerner *et al.*, 2000; Griffin *et al.*, 2008; Krekulova *et al.*, 2001; Nishizawa *et al.*, 1997). It can be easily detected in the blood and faecal sample of

the infected subjects who did not show any symptomatic presentation. Therefore, the higher prevalence of TTV in certain geographic areas may be attributed to the environmental factors and living styles in that particular area.

The only concrete association of TTV and diseases can be found in immuno-compromised or immuno-deficient people. HIV/AIDS patients with a lower CD4+ cell counts was associated with a higher TTV load compared to patients with higher CD4+ cell count and healthy people, and that contributed to a higher mortality rate in the patient population (Shibayama *et al.*, 2001). The role of TTV in autoimmune disorders was also implicated in patients with systemic lupus erythematosus (SLE) (Gergely *et al.*, 2005) and bullous pemphigoid (BP) (Blazsek *et al.*, 2008). In this chapter, we have described the amount of nucleotides of viruses in healthy people and HIV/AIDS but did not aim to determine the rate of infection of TTV infection in these two groups of subjects. Unfortunately, we could neither compare the viral load in these subjects due to the fact that the sequencing preference analyzed the dominant bacteriophages and other viral groups.

4.4.1.5 Healthy adults are living with some small viruses

Noticeably, we found three other groups of viruses in healthy adults including Torque teno midi virus (TTMDV), SEN virus and small anellovirus. In fact, these three viruses, together with the TTV, constitute the whole virome in healthy adults. Besides, all of them belong to the Family Anelloviridae. TTMDV is a small virus that is distantly similar to TTV, which was discovered recently (Ninomiya *et al.*, 2007). It was thought to be the intermediate between TTV and another similar virus called Torque teno mini virus (TTMV). TTMV were found to affect human

population in a way similar to that of TTV (Thom and Petrik, 2007). Thus, it is speculated that TTMDV also worked like TTV and TTMV.

SEN virus was first reported in 1999 in lay press for a post-transfusion hepatitis case and in 2000 in scientific journal for its genome information (Akiba *et al.*, 2005). It was firstly suspected to cause non-A-E acute hepatitis and might be associated with hepatocellular carcinoma (Pirovano *et al.*, 2002b). It was suggested to be a co-infection agent in HIV/AIDS patients who acquire HIV through intravenous drug injection. Nevertheless, it was later found that a certain percentage of healthy adults carry this virus in their blood but do not cause any diseases, which is in consistence with our findings. Like TTV and TTMDV, SEN virus was suggested to be a ubiquitous normal blood virus which can be transmitted through blood or mother-to-child (Pirovano *et al.*, 2002a). The prevalence of SEN virus is higher in Asia and Europe with about 10-20 % healthy people infected when compared to <2% in America (Kao *et al.*, 2002; Shibata *et al.*, 2001; Umemura *et al.*, 2003; Umemura *et al.*, 2001).

Interestingly (and might not coincidentally), small anellovirus are also similar to TTV viruses and had been associated with hepatitis C (Andreoli *et al.*, 2006). Since its discovery in 2006, small anellovirus was found in different groups of peoples (Chung *et al.*, 2007). Yet, by now there is no much information of the pathogenicity of this virus. Detecting it in the normal virome as described here suggests that it may be a virus that is carried by the general population.

As mentioned previously, TTV, TTMDV, SEN viruses and small anellovirus all

belong to Anelloviridae and share a great similarity in functions. Thus, there might be a possibility that one virus was mistaken as another, which cannot be proved technically. Nevertheless, we could still conclude that normal virome is mainly consists of anelloviruses.

4.4.2 Use of next-generation sequencing in metagenomics

4.4.2.1 Limitations

As described in Chapter 3 and this chapter, next-generation sequencing technology was used in determining the microbial and viral metagenomes. This kind of high-throughput analysis method has been emerging as a new tool for studying the old diseases in a new perspective. But still there are quite a number of limitations in using this method. First of all, there are always potential contaminated sequences in the data set. Although human factors can be stringently controlled by getting rid of the sources, there can always be some exogenous sequences samples coming from the air, tubes, reagents and water. And unfortunately, we cannot 'kill' DNA. Despite having destroyed the DNA into fragments by U.V. light or DNase, there can still be contamination at the step of sample transfer and sequencing. In addition, if the fragmented DNA were sequenced, they would likely be assembled into contigs and included in the metagenome. Desperately, there is no standard of adequate precaution. Thus, one of the biggest problems faced by sequence analysts is 'how to discriminate the contaminated sequences from the real sequences'. This is technically very difficult and is further complicated by the mixing of real sequences which have no similarity to any exist micro-organisms. In this manner, metagenomics are always presented in relative proportion of different species and only the dominating population is

analyzed critically. In this project, we only analyzed the part of data on microbes that are already known to mankind.

Another limitation of this technology is its 'low' sensitivity towards tiny amount of DNA. The samples put for sequencing must be in a (unreasonably) large amount, usually 10 μg for DNA and 5 μg for RNA. Therefore, the samples extracted from the clinical samples such as blood and tissues are usually subjected to amplification. The amplification bias can then artificially increase the difference in the amount of various species, and favour the longer and relative abundant ones. Apart from this, the commonly use amplification enzyme *Phi29* (as that used in this project) preferentially amplifies double stranded circular genome such as viral genomes and is less efficient in amplifying non-circular or single stranded cDNA (RNA). Nevertheless, this short-coming can be overcome by replacing the *Phi29* method with traditional cDNA/DNA library construction, although the later method is less convenient to use.

4.4.2.2 Advantages

Even though there are some limitations, next-generation sequencing is still the best method in studying complex genetic samples. By considering its high-throughput capability, this technology is regarded as a cost-effective method. The traditional high-throughput method involves the use of hybridization chip. It favours the detection of sequence that can hybridize onto the chip under a fixed hybridization conditions. However, it is practically impossible to optimize the hybridization on the same chip for all the target sequences. The most outstanding advantage of sequencing is that it saves the time for tedious cloning. It is not easy

to clone the sequences which are long, repetitive or with a higher GC content. But the sequencing is basically unbiased such that it detects whatever has been included in the sequencing reaction.

4.4.2.3 Potential pitfalls in this project

Although we have used the most innovative method to study microbiome and virome, there are still some pitfalls in this project. Firstly, as described in section 4.3.1, the sequencing depth was not sufficient to sequence all the species in a one to one basis. Instead, the longer and more abundant species are favoured. As illustrated by the abundance of TTV, it is difficult to judge whether there was really more TTV in healthy people or that was just an artifact due to the uneven distribution of the sequencing space. In addition to that, the experiment design shall be further benefited by spiking the samples with a known concentration of standard sequences to check the efficiency of nucleic acid extraction in preparing the two sets of samples. Seemingly, the sequence resolution is in proportion to the sequencing depth. Choosing sequencing service of larger depth may solve the problem, if the cost is not a matter of concern. Another problem in this project is the presence of a large amount of sequences that do not resemble any known sequences with any significance. Obviously bioinformatics analysis method must be available to analyze these sequences. Above all, this study shall be benefited by sequencing the plasma microbiome or virome in individuals instead of pooling the samples. It shall also be benefited by including people whose healthy status shall represent the general population. For example, to study the plasma microbiome/ virome in Chinese, a considerable portion of the HIV-uninfected but HBV-infected subjects shall be included because it is known that about >8% of the population are HBV carriers

(Chan and Jia, 2011).

4.4.3 Future work on the unmatched contigs in HIV/AIDS virome

Future work on this part of the project shall mainly focus on the study of the unmatched contigs and the identification of potential novel viruses. A novel virus is defined as a virus which was not found in any organisms before. Nowadays, the most innovative way to identify a potential virus is by a virus chip in combination of whole viral genome sequencing (Chen *et al.*, 2011; Wang *et al.*, 2002). Taken as an example, Chen et al. has made use of the viral chip in combination with whole viral genome sequence to identify a novel adenovirus (Chen *et al.*, 2011). The genome of suspected novel virus was hybridized against the conserved viral sequences of different family of viruses. By the result of the viral chip, the identity of the potential novel virus can be predicted down to a few families for example Hepadnaviridae, Adenoviridae and Retroviridae. The suspected virus was then sequenced. Instead of sequencing a particular gene as that applied to a bacterial genome, the whole viral genome was sequenced due to the fact that viral genomes are thousand times smaller than bacterial genomes. The complete genome sequence was then compared to the known members in the families of viruses. This would allow us to further narrow down the viral target to a particular family such as Adenoviridae. A phylogenetic tree was then constructed to compare the sequenced viruses with the existing genotypes of adenoviruses. By studying the phylogenetic conservation, the suspected novel virus is grouped into a particular phylogenetic blade or defined as a new blade which indicates its novelty. In this project, by means of next generation sequencing we have saved the extra time for viral sequencing (Gaynor *et al.*, 2007). However, unlike typical complete genome

sequencing of a pure virus, complete genome information is usually not available in the dataset of metagenome. Thus, to identify potential novel viruses in HIV/AIDS patients, plasma samples have to be recollected and cultured for full genome sequencing.

This project can also be continued by studying the non-human virus bluetongue (BT) virus and white spot syndrome virus (WSSV). These viruses can be studied in the same way as described in the previous section. Alternatively, the viral genes can be amplified from HIV/AIDS patients by means of overlapping PCR. The gene structure can be compared to other existing BT virus and WSSV for the potential recombination of other viruses with these two viruses. In fact, it is even more worthwhile to study the potential recombination of RNA viruses. It is reported that RNA viruses possess a high degree of interspecies jumping due to the infidelity of their RNA polymerases and coexpression of pre-integrated retroviruses (Geuking *et al.*, 2009; Woo *et al.*, 2009). Taken as an example, RNA virus coronaviruses were originally found on bats and birds. But by means of the continued replication, the viral genome was gradually modified which allows it to infect other organisms (Woo *et al.*, 2009).

4.5 Conclusion

With the increasing enormous amount of data on human microbiome and virome, we are convinced that we human are not alone. There are unaccountable tiny organisms living with us day to day. This study on plasma microbiome and virome as described in Chapter 3 and 4, shall help us to better understand the large community living in our blood, though their mode of symbiosis await further studies.

In addition, the metagenomic information found for immuno-deficient people can indirectly reveal the microbes that are normally eliminated by the power of our immunity. In regards to the virome, the data described here may open up further studies on plasma bacteriophages and novel human endogenous retrovirus.

Chapter 5

Study of normal plasma microbiome in two separate cohorts

Summary

In view of the belief that there are no bacteria in the systemic circulation of healthy normal adults, we then elucidated the microbiome for a separate cohort of healthy people. The data suggested that both cohorts contained bacterial elements in their plasma samples. The bacterial and viral species found in the two cohorts were quite different. The difference may be attributed to variables in the two independent studies, including environmental and human factors. While presence or absence of bacterial elements in blood was not justified, we discuss the potential roles and mode of bacterial elements in the circulation. We have also discovered that genome of Toxoplasma gondii was somewhat similar to the human genome, rendering it impossible to differentiate the T. gondii in the sequencing data-sets.

5.1 Introduction

Human plasma is assumed to be sterile. Although it had once been proposed that normal people may live with some bacteria (Domingue and Schlegel, 1977), the general belief is that only viruses can present in our blood. Bacteria are likely to cause microbial pathogen-associated molecular patterns (PAMPs) in which the innate immunocytes are activated (Bianchi, 2007).

In this project, we originally set out to study the microbiome and virome in immuno-deficient HIV/AIDS patients. We incorporated the healthy adults as a control in the study and we expected to get nothing meaningful in sequencing the normal plasma. However, even with very stringent experiment procedures and data analysis (setting BLAST e-value gate and positive hit length gate), we surprisingly found some bacterial sequences in the healthy adults who were definitely not suffering from any immunity problems, and not even from common viral infection by hepatitis viruses or human papillomavirus, as revealed in Chapter 4. The 'normal' plasma microbiome may simply be the contaminants, which sounds impossible to get rid of or to prove their presence. But what if these really present in our blood stream? To elucidate whether the microbiome comes from the environment which should be quiet specific to each laboratory, we have elucidated the normal microbiome from another cohort of healthy adults recruited by another research group. Instead of pooling the plasma DNA of this group of healthy adults, the DNA was sequenced separately for each individual.

5.2 Materials and Methods

5.2.1 Study subjects

A separate cohort of 17 subjects was recruited by Dr. Dennis Y.M. Lo and Dr. Rossa W.K. Chiu of the Department of Chemical Pathology in The Chinese University of Hong Kong (CUHK). The cohort was recruited as control for their profound study in fetal RNAs in maternal circulation.

5.2.2 Sequencing plasma DNA

The plasma DNA was extracted from the control groups described in section 5.2.1. Instead of pooling the samples, these extracted DNA was then sequenced one by one using short-read sequencing. The DNA extraction, sample preparation and sequencing preparation were all done by Dr. Lo and Dr. Chiu in a laboratory in the CUHK affiliated hospital – The Prince of Wales Hospital.

5.2.3 Data analysis

The 17 sets of sequence data were analyzed using the same pipeline as that described in Chapter 3 with an e-value gate of $1x10^{-10}$ and positive hit length gate to filter the sequences that were matched for >50% length of the input sequences. This was done by the same bioinformatics team in The Hong Kong Bioinformatics Center of the Chinese University of Hong Kong.

5.2.4 Retrieval of *Toxoplasma gondii* reference protein sequences

BLASTP (search protein databases using a protein query) was performed using default setting against the non-redundant protein database (nr) to evaluate the similarity between *T. gondii* with other organisms. . The reference protein sequence as obtained from Genome database from NCBI from *T. gondii ME49* whole genome shotgun sequence (RefSeq: NZ_ABPA00000000).

5.3 Results

5.3.1 Comparison of normal microbiome in two separate cohorts

The microbial sequences presence in the normal individuals in two separate cohorts were compared and summarized in Table 5.1. The data-set 1 was the same set of data described in Chapter 3 and 4 for the normal microbiome and virome. It was generated by sequencing the pooled plasma nucleic acids for the ten studied normal subjects. The data set 2 belonged to the separate cohort as described in section 5.2. Individual plasma samples were sequenced and analyzed separately.

Comparison between the two cohorts showed that the only co-existing microbial group was uncultured bacteria. Set 1 contained several unique bacteria such as Actinomycetales, Bacteroidales, Oscillatoriales, Cyanobacteria, Clostridiales, Burkholderiales, Pseudomonadales and Verrucomicrobiales. On the other hand, unique bacteria found in set 2 included Rhodobacteriles, Sphingomonadales, Acholeplasmatales. Intriguingly, several individuals in set 2 carry a group of protists which belong to the phylum Apicomplexa which was not found in set 1.

The viruses found in the two cohorts were quite different. In our data-set 1, only viruses in the family Anelloviridae were found while no human endogenous virus were found. In contrast, set 2 did not contain any anelloviruses but quite a number of endogenous viral sequences. Some of the subjects in set 2 were infected by hepatitis B virus (HBV) or hepatitis C virus (HCV).

Table 5.1 The micro-organisms found in normal healthy adults in two separate cohorts.

Taxonomy				Datasets		Individuals in set2 cohort														
Super-kingdom	Phyfm	Class	Order	Set 1 Normal microiome (Chapter 3)	Set_2 Cohort 2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Bacteria				Y																		
	Actinobacteria	Actinobacteria (class)	Actinomycetales	Y																		
	Bacteroidetes	Bacteroidia	Bacteroidales	Y																		
	Chlamydiae	undef	Oscillatoriales	Y																		
	Cyanobacteria	undef	undef	Y																		
	Firmicutes	Clostridia	Clostridiales	Y																		
	Proteobacteria	Alphaproteobacteria	Rhodobacterales		Y																Y	
			Sphingomonadales		Y																Y	
		Betaproteobacteria	Burkholderiales	Y																		
		Gammaproteobacteria	Pseudomonadales	Y																		
	Tenericutes	Mollicutes	Acholeplasmatales		Y	Y	Y	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Y																		
	Uncultured bacteria			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Eukaryota																						
	Apicomplexa	Aconoidasida	Haemosporida		Y												Y	Y	Y	Y		Y
Virus	Retro-transcribing viruses	Hepadnaviridae	Hepatitis B virus		Y	Y																
		Retroviridae	Huamn endogeneous virus		Y	Y	Y	Y	Y			Y		Y	Y	Y	Y	Y	Y	Y		Y
	ssRNA viruses	ssRNA positive-strand viruses	Hepatitis C virus		Y														Y			
	Anelloviridae	unclassified Anelloviridae	Torque Teno virus	Y																		
			Torque Teno midi virus	Y																		
			SEN virus	Y																		
			Small anellovirus	Y																		

Footnote: Y- presence of a particular group of microbes in the individual.

5.3.2 Comparison of *Toxoplasma gondii* with human genome

We noticed that there was quite a number of contigs mapped to *Toxoplasma gondii* for both groups of cohorts (Table 5.2). *T. gondii* belongs to the order Eucoccoidiorida in the Phylum Apicocomplexa. This group of organisms was finally excluded in Table 5.1 because of its high similarity to human genome, which made it impossible to judge whether they are real microbial sequences. As illustrated in Figure 5.2, *T. gondii* proteins dual specificity phosphatase, catalytic domain containing protein (XP_002364154.1) and U2 small nuclear ribonucleoprotein (RNP) auxiliary factor U2AF (XP_002364161.1) were aligned with human proteins with a high similarity with significance, and were not filtered by exerting the e-value ($1x10^{-10}$) and positive hit length gate (50%). Therefore, all the sequences mapped to *T. gondii* were excluded in the analysis for all the microbiome described in this project, including that in Chapter 3, 4 and 5.

Table 5.2 Potential existence of *T. gondii* in healthy individuals.

Taxonomy					Datasets		Individuals in set 2 cohort														
Super-kingdom	Phylm	Class	Order	Species	Set 1 Normal microlome (Chapter 3)	Set_2 Cohort 2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Eukaryota	Apicomplexa	Coccidia	Eucoccoidiorida	*T. gondii*	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y

Footnote: Y- presence of a particular group of microbes in the individual.

A.

BLASTP

Input: U2 small nuclear ribonucleoprotein auxiliary factor U2AF [Toxoplasma gondii ME49] XP_002364161.1

Output:
GENE ID: 11338 U2AF2 | U2 small nuclear RNA auxiliary factor 2 [Homo sapiens]
(Over 10 PubMed links)

```
Score = 89.0 bits (219),  Expect = 3e-15, Method: Compositional matrix adjust.
Identities = 53/156 (34%), Positives = 80/156 (51%), Gaps = 6/156 (4%)

Query  43   KSEKKKQYR--FDSPPRTGETPEEQQKRLLQQHMGAAPTVAGTTLLPTVAPSAAFSGPAS  100
            + EKKK+ R  +D PP   E     Q + +Q    AA  +  T LLPT+ P       P
Sbjct  81   RHEKKKKVRKYWDVPPPGFEHITPMQYKAMQ----AAGQIPATALLPTMTPDGLAVTPTP  136

Query  101  SLALDAAATKAARELYVGNLPPSLEVPQLMEFLNAAMAAVGGALLPGPPAVKAWRSTDGH  160
               + +  T+ AR LYVGN+P  +    +M+F NA M   G    PG P +    + D +
Sbjct  137  VPVVGSQMTRQARRLYVGNIPFGITEEAMMDFFNAQMRLGGLTQAPGNPVLAVQINQDKN  196

Query  161  YAFVEFRTMEEASNGMQLNGLNCMGFNLRIGRPKTY  196
            +AF+EFR+++E +  M  +G+   G +L+I RP  Y
Sbjct  197  FAFLEFRSVDETTQAMAFDGIIFQGQSLKIRRPHDY  232
```

B.

BLASTP

Input:
DNA polymerase, putative [Toxoplasma gondii ME49] XP_002364159.1

Output:
>gb|AAF63383.1|AF245438_1 Gene info linked to AAF63383.1 DNA polymerase iota [Homo sapiens] Length=715
GENE ID: 11201 POLI | polymerase (DNA directed) iota [Homo sapiens]

```
Score = 93.2 bits (230),  Expect = 3e-16, Method: Compositional matrix adjust.
Identities = 44/91 (48%), Positives = 64/91 (70%), Gaps = 2/91 (2%)

Query  203  SRVIIHVDMDCFYAQVEELLDPSIVGKPVAVRQKQLIVTSNLVARQKPWQVRKGIYMPEA  262
            SRVI+HVD+DCFYAQVE + +P -  KP+ V+QK L+VT N  AR+     V+K + + +A
Sbjct  27   SRVIVHVDLDCFYAQVEMISNPELKDKPLGVQQKYLVVTCNYEARK--LGVKKLMNVRDA  84

Query  263  LRRCPSLVVKNGEDLDKYRRVSDQILHALQE  293
              +CP LV+ NGEDL +YR +S ++   L+E
Sbjct  85   KEKCPQLVLVNGEDLTRYREMSYKVTELLEE  115


 Score = 61.6 bits (148),  Expect = 1e-06, Method: Compositional matrix adjust.
 Identities = 31/93 (33%), Positives = 54/93 (58%), Gaps = 1/93 (1%)

Query  616  LAIATHLASAIRRYIGTHMGLTSTAGISTNKSLAKLAGAFNKPSRQTLLLPQHRRGFLAP  675
            L + + +A+ +R  +   +GLT  AG+++NK LAKL     KP++QT+LLP+  +  +
Sbjct  173  LLVGSQIAAEMREAMYNQLGLTGCAGVASNKLLAKLVSGVFKPNQQTVLLPESCQHLIHS  232

Query  676  LA-LQKIPGVGSALVRLLSRAGLRTCADLLAVS  707
            L  +++IPG+G    + L   G+ +  DL   S
Sbjct  233  LNHIKEIPGIGYKTAKCLEALGINSVRDLQTFS  265
```

Figure 5.1 Illustration of protein similarity between *T. gondii* and *H. sapiens*. BLASTP was performed for two selected proteins from *T. gondii*. The BLASTP results showed that *T. gondii* was somewhat similar to human genome, and were non-differentiable in our analysis.

5.4 Discussion

5.4.1 Presence of bacteria in healthy adults in two separate cohorts

Bacteria were found in both cohorts of healthy adults. By their unique distribution in the two separate cohorts, it seemed that they were environment/ laboratory specific contaminant rather than living microbes. Alternatively, it may due the variables in human factors or techniques in the two independent studies. Nonetheless, there was no direct evidence on either the presence or absence of bacteria in normal people. Neither was the fair comparison between cohorts justified, as illustrated by the presence of human endogenous retroviral genome in many different subjects in set 2 but none from set 1. Similarly, a lot of *Torque teno viruses* were found in set 1 while none in set 2. We then propose that the study should be expanded to larger scale and to separate laboratories to prove or disprove the presence of bacteria in systemic circulation of healthy adults.

5.4.2. Can bacteria or bacterial elements live with us?

5.4.2.1 Translocation of bacterial genome into systemic circulation

While it sounds preposterous to claim that bacteria can live in our blood, we discuss whether or not the bacteria materials can be found in the circulation. Presence of bacteria in the blood (bacteremia) often leads to sepsis which can, if untreated, results in several systemic inflammation and eventually systemic organ failure and death (Bone, 1992; Fine *et al.*, 1959). The systemic inflammation response syndrome (SIRS) is often more detrimental that the destruction caused by the infection. One of the common causes of SIRS is the bacterial translocation from the gut to the circulation. Mild transient bacteremia, however, is typically

undetectable by microbiological methods due to the strong suppression by the immune system. In this project, bacterial genetic elements were detected by means of sequencing which, unlike traditional culture methods, tells nothing about the viability of these microbes. As discussed in Chapter 3, these may be genetic materials of bacteria that await removal by the immune system. Presence of genetic materials in our circulation is a well-accepted fact in science and medicine. The circulating nucleic acid can be the human DNA (Stroun *et al.*, 1987), viral DNA (Leland and Ginocchio, 2007) or fetal DNA in pregnant women (Chiu *et al.*, 2009) which do not provoke immune attack. This can be further explicated by the use DNA immunization (Koup *et al.*, 2010; Montgomery and Prather, 2006). DNA sequence of target antigens cannot be injected directly into the subject. Instead, the antigen sequence was cloned into a vector incorporated with a viral transcription and translation sequences, due to the fact that immune system mainly recognizes proteins. Therefore it may be possible that the bacteria release their non-immune reactive genetic materials into the circulation.

The most possible way to acquire bacteria materials is through the gut. Our gastrointestinal (GI) tract is the most complex part of our immune system as the GI tract needs to differentiate between food proteins, drugs and oral DNA vaccines from the infectious agents (Mowat, 2003). We speculated that the GI tract may allow the translocation of bacterial DNA into the circulation as they seemingly do not induce any immune attack.

5.4.2.2 Integration of bacterial genome

Viremia causes fewer problems when compared to bacteremia. Viruses can live in our circulation by hiding in our genome of blood cells, as in the case for HIV. Even though all the circulating viral particles, presenting surface proteins can be effectively eliminated in the immune response, the integrated viruses can still survive with our cells (Rathinam and Fitzgerald, 2011). (This also raises the concern to kill the virus-bearing cells rather than just killing viruses themselves.) Is it possible for bacteria, by any means, to integrate their genome into human genome? Bacterial gene transposition is one of the shrew strategies adopted by bacteria to exchange the sequence between genomes and plasmids, in response to the changes in physiology of host cells (Nagy and Chandler, 2004). Intriguingly, the majority of human genome is composed of the repetitive sequences (Belshaw *et al.*, 2004; Lander *et al.*, 2001). If bacterial genetics elements can 'survive' in the circulation, then it might be possible that the bacterial jumping elements are recognized by cellular factors and be transposed into the host genomes.

5.4.3 The mitochondrion theory

Published in year 2010 in *Nature*, the mitochondria were proven to play a leading role in the prolonged innate immune response and inflammation (Zhang *et al.*, 2010). In cellular injury without infection such as trauma and schematic-reperfusion, mitochondria are released from the damaged cells. They are mistaken as bacteria by the immune system which can cause severe immune attack much like sepsis in pathogen-associated molecular patterns (PAMPs). This process is called damage-associated molecular patterns (DAMPs) and it involves the mis-recognition

of human proteins as foreign elements, and can lead to severe organ failure. Examples of DAMPs are high mobility group box 1 (HMGB1) which is a nuclear protein present in most human cells, and S100 which a group of calcium binding proteins (Bianchi, 2007). Mitochondria are thought to have evolved from an alpha-proteobacteria as they share the similarity in chromosome sequence and having a circular genome (Henze and Martin, 2003; Manfredi and Rovere-Querini, 2010). In the HIV/AIDS microbiome, the most abundant bacterial groups was found to be Pseudomonadales which is actually a group of alpha-proteobacteria, rendering it possible to mis-detect the mitochondrial DNA as bacteria. Though the event seems unlikely in healthy people without any injury, it was much likely in HIV/AIDS when the T-lymphocytes were markedly damaged. Nevertheless, in our analysis, we have already carefully removed the mitochondria DNA which were not included in the reported microbiome.

5.4.4 Ambiguous identification of *T. gondii*

As illustrated in the results, *T. gondii* proteins are similar to some of the human proteins. Hence, the genome sequences of *T. gondii* were not analyzed in this study. This finding can be exploited in further experiments for a better comparison of the two genomes. *T. gondii* is a protozoan that can infect all kinds of warm-blooded eukaryotes (Khan *et al.*, 2006). Infection in pregnant women can cause congenital birth defect of the baby, and in immuno-compromised people can cause encephalitis. *T. gondii* causes severe cerebral toxoplamosis in HIV/AIDS (Antinori *et al.*, 2004). Molecular diagnosis has been used as a non-invasive method to detect the plasma *T. gondii* in patients with suspected infection (Colombo *et al.*, 2005; Mesquita *et al.*,

2010). The accuracy of the molecular diagnosis has been improved by sequencing different strains of *T. gondii* (Gajria *et al.*, 2008). Our data further showed that *T. gondii* genome must be carefully analyzed for the design of molecular detection probes to avoid false positive results.

5.5 Conclusion

Through studying two separate cohorts of healthy adults, we proposed that it was very likely that plasma samples could be contaminated by exogenous bacterial sequences. Nonetheless, we speculate it is also possible that bacteria genetic elements are persistent in our circulation. On one hand, we are to eliminate the contaminations; on the other hand, we are not to ignore the presence of endogenous bacterial genes. Collectively work from the professionals in bioinformatics, microbiology, biochemistry and molecular medicine is definitely needed in solving this intricate biological puzzle.

Thesis Conclusion

In this decade, there are a lot of innovative techniques for studying genomics. The next-generation sequencing has transformed today's research. It gives new dimensions to investigate the old, unsolved topics in biology. We have made use of the next-generation short-read sequencing in studying two aspects of HIV-1 induced cellular changes. These include elucidating the effect of HIV infection host DNA methylation and plasma microbiome and virome composition.

By studying a pair of identical twins with one infected by HIV and the other uninfected, we found that genes regulating RNA splicing and cell cycle regulation were hypo-methylated whereas the genes regulating nerve functions and signal transduction were hyper-methylated. We have profiled the plasma microbiome and virome of ten HIV/AIDS patients in parallel with healthy adults. HIV/AIDS plasma microbiome was dominated by bacteria from the order Pseudomonadales while healthy control subjects carried few bacterial DNA in the blood. We have found that many of the microbes in HIV/AIDS plasma are similar to some of the microbes found in the human gut. The HIV/AIDS and normal plasma virome share some similarity in the presence of common ubiquitous eukaryotic viruses. The normal virome was mainly composed of viruses from Anelloviridae. The HIV/AIDS viromes was contrasted by the presence of a large proportion of bacteriophages, typical eukaryotic viruses and untypical non-human viruses. In addition, by means of sequencing, we have found several sequences which might belong to novel bacteria or endogenous retroviruses. We have also studied the microbial elements

in two independent cohorts of uninfected subjects. The results showed that normal people may carry bacteria genetic elements in the blood circulation and these elements are unlikely to be alive.

The results described here in this thesis manifest the use of such high-throughput technology in studying cellular genome and microbial metagenomics. Nevertheless, we emphasize that only a few elements in the methylome and plasma microbial metagenomics have been studied. But even these have revealed the noteworthy complexity of the interplay between HIV-1, human and a long-overlooked party— the micro-organisms.

Appendix

Appendix 1 The complete gene list of DNA methylation in HIV negative and HIV positive twins. The number of reads mapped to the gene was given (sorted by gene symbols).

Gene	HIV negative	HIV positive	log2 (Fold change)	log2 (Fold change) normalized	z-score	p-value
ABCA9	11	14	-0.35	-3.22	-6.89	5.73E-12
ABCG8	56	0	6.81	3.94	5.50	3.71E-08
ACAA2	50	0	6.64	3.77	5.13	2.93E-07
ACAP1	76	0	7.25	4.38	6.61	3.97E-11
ACLY	0	5	-3.32	-6.19	-4.73	2.25E-06
ACSS2	84	0	7.39	4.52	7.00	2.64E-12
ACTG1P1	13	16	-0.30	-3.17	-7.31	2.63E-13
ACVR2A	0	7	-3.81	-6.68	-5.55	2.88E-08
ADAM33	31	0	5.95	3.08	3.72	0.0002
ADAMTS13	44	0	6.46	3.59	4.72	2.31E-06
ADAMTS2	0	6	-3.58	-6.45	-5.16	2.46E-07
ADAMTS6	28	0	5.81	2.94	3.45	0.0006
ADCY7	68	0	7.09	4.22	6.19	6.07E-10
ADH4	26	11	1.24	-1.63	-4.07	4.73E-05
ADRA1A	2	9	-2.17	-5.04	-6.31	2.86E-10
AFM	34	0	6.09	3.22	3.97	7.18E-05
AHNAK	63	0	6.98	4.11	5.91	3.36E-09
AIMP1	30	0	5.91	3.04	3.63	0.0003
AKAP17A	28	0	5.81	2.94	3.45	0.0006
AKR1CL1	9	24	-1.42	-4.28	-9.98	1.87E-23
ALDOAP2	42	0	6.39	3.52	4.58	4.60E-06
ALG1	61	0	6.93	4.06	5.80	6.67E-09
ALMS1P	174	2	6.44	3.57	9.38	6.67E-21
ALPK2	40	0	6.32	3.45	4.44	9.16E-06
AMAC1L1	185	0	8.53	5.66	10.67	1.42E-26
AMAC1L2	286	0	9.16	6.29	13.22	6.97E-40
ANKMY2	0	5	-3.32	-6.19	-4.73	2.25E-06
ANKRA2	0	5	-3.32	-6.19	-4.73	2.25E-06
AP1B1P1	53	0	6.73	3.86	5.32	1.04E-07
AP1M1	27	0	5.75	2.89	3.36	0.0008
APBB2	0	9	-4.17	-7.04	-6.23	4.70E-10
APOBEC1	0	7	-3.81	-6.68	-5.55	2.88E-08
APOL5	78	0	7.29	4.42	6.71	2.01E-11
APOL6	20	21	-0.07	-2.94	-8.09	6.15E-16
ARHGEF3	0	28	-5.81	-8.68	-10.20	2.04E-24
ARL9	48	0	6.58	3.72	5.00	5.83E-07
ASCC1	42	0	6.39	3.52	4.58	4.60E-06
ASH1L	2	13	-2.70	-5.57	-7.64	2.15E-14
ASS1P12	69	0	7.11	4.24	6.24	4.31E-10
ASS1P3	36	0	6.17	3.30	4.13	3.62E-05
ATF4P1	5	46	-3.20	-6.07	-14.37	8.55E-47
ATP5A1	33	0	6.04	3.18	3.89	0.0001
ATP5A1P9	41	0	6.36	3.49	4.51	6.49E-06
ATP5EP1	0	11	-4.46	-7.33	-6.82	9.18E-12
ATP5J2P5	34	0	6.09	3.22	3.97	7.18E-05

ATP5JP1	51	1	5.67	2.80	4.55	5.33E-06
ATP5LP7	0	13	-4.70	-7.57	-7.34	2.08E-13
ATP6V1G1P4	30	0	5.91	3.04	3.63	0.0003
ATP8B1	27	0	5.75	2.89	3.36	0.0008
AVPI1	36	0	6.17	3.30	4.13	3.62E-05
BAAT	36	18	1.00	-1.87	-5.73	9.99E-09
BCAT1	36	46	-0.35	-3.22	-12.49	8.00E-36
BCYRN1	1	4	-2.00	-4.87	-4.18	2.87E-05
BET3L	0	39	-6.29	-9.15	-11.69	1.46E-31
BOLA3	150	0	8.23	5.36	9.58	9.43E-22
BRP44L	0	7	-3.81	-6.68	-5.55	2.88E-08
C1QL1	0	13	-4.70	-7.57	-7.34	2.08E-13
C1QTNF6	3	13	-2.12	-4.98	-7.57	3.77E-14
C5AR1	234	0	8.87	6.00	11.99	3.91E-33
CABYR	60	0	6.91	4.04	5.74	9.40E-09
CACNB4	82	0	7.36	4.49	6.90	5.18E-12
CALM2P3	0	6	-3.58	-6.45	-5.16	2.46E-07
CAMKK1	47	0	6.55	3.69	4.93	8.23E-07
CASS4	27	0	5.75	2.89	3.36	0.0008
CBR3	0	7	-3.81	-6.68	-5.55	2.88E-08
CC2D2B	80	0	7.32	4.45	6.80	1.02E-11
CCDC102A	0	8	-4.00	-6.87	-5.90	3.59E-09
CCDC11	58	0	6.86	3.99	5.62	1.87E-08
CCDC148	0	10	-4.32	-7.19	-6.53	6.43E-11
CCDC60	30	0	5.91	3.04	3.63	0.0003
CCDC8	40	0	6.32	3.45	4.44	9.16E-06
CCDC90B	0	5	-3.32	-6.19	-4.73	2.25E-06
CCL22	0	8	-4.00	-6.87	-5.90	3.59E-09
CCNB1IP1	303	0	9.24	6.37	13.59	4.78E-42
CD163	35	0	6.13	3.26	4.05	5.10E-05
CD200R1L	109	2	5.77	2.90	6.77	1.30E-11
CD24	24	22	0.13	-2.74	-7.99	1.33E-15
CD24P4	24	22	0.13	-2.74	-7.99	1.33E-15
CD69	9	17	-0.92	-3.79	-8.09	5.86E-16
CDAN1	0	6	-3.58	-6.45	-5.16	2.46E-07
CDC25B	0	7	-3.81	-6.68	-5.55	2.88E-08
CDC37P2	48	0	6.58	3.72	5.00	5.83E-07
CDH26	76	0	7.25	4.38	6.61	3.97E-11
CDR2L	53	2	4.73	1.86	3.62	0.0003
CDRT4	0	9	-4.17	-7.04	-6.23	4.70E-10
CECR1	54	0	6.75	3.89	5.38	7.39E-08
CEP76	5	18	-1.85	-4.72	-8.83	1.07E-18
CES3	41	0	6.36	3.49	4.51	6.49E-06
CETN4P	34	0	6.09	3.22	3.97	7.18E-05
CHIC1	0	10	-4.32	-7.19	-6.53	6.43E-11
CHRFAM7AP2	4	9	-1.17	-4.04	-6.01	1.84E-09
CHRNA6	146	0	8.19	5.32	9.45	3.41E-21
CIAPIN1	5	14	-1.49	-4.35	-7.65	1.96E-14
CICP2	4	5	-0.32	-3.19	-4.10	4.13E-05
CKS1BP2	271	0	9.08	6.21	12.88	5.88E-38
CLASP1	28	0	5.81	2.94	3.45	0.0006
CLK4	3	4	-0.42	-3.28	-3.71	0.0002
CLU	100	0	7.64	4.77	7.71	1.22E-14
CLUAP1	36	0	6.17	3.30	4.13	3.62E-05
CMTM1	4	7	-0.81	-3.68	-5.14	2.77E-07

CN5H6.4	41	0	6.36	3.49	4.51	6.49E-06
CNOT6LP1	0	5	-3.32	-6.19	-4.73	2.25E-06
COASY	43	0	6.43	3.56	4.65	3.26E-06
COL4A3BP	0	9	-4.17	-7.04	-6.23	4.70E-10
COLEC12	73	3	4.60	1.74	4.06	4.98E-05
COMMD6	1	5	-2.32	-5.19	-4.72	2.40E-06
COPZ1	65	1	6.02	3.15	5.44	5.34E-08
COQ6	1	6	-2.58	-5.45	-5.19	2.14E-07
COX4I1P2	29	0	5.86	2.99	3.54	0.0004
CPLX3	0	6	-3.58	-6.45	-5.16	2.46E-07
CPP	45	0	6.49	3.62	4.79	1.64E-06
CRKL	56	2	4.81	1.94	3.83	0.0001
CRLF2	156	0	8.29	5.42	9.78	1.38E-22
CRYZL1	0	8	-4.00	-6.87	-5.90	3.59E-09
CSAD	79	0	7.30	4.43	6.75	1.43E-11
CSF2RA	78	0	7.29	4.42	6.71	2.01E-11
CST7	38	28	0.44	-2.43	-8.43	3.48E-17
CT45A5	0	29	-5.86	-8.73	-10.35	4.32E-25
CTAGE5	0	5	-3.32	-6.19	-4.73	2.25E-06
CTCFL	35	0	6.13	3.26	4.05	5.10E-05
CTHRC1	3	5	-0.74	-3.61	-4.31	1.62E-05
CTPS	0	7	-3.81	-6.68	-5.55	2.88E-08
CTSB	57	0	6.83	3.96	5.56	2.63E-08
CTSLL4	37	0	6.21	3.34	4.21	2.57E-05
CXADRP1	35	0	6.13	3.26	4.05	5.10E-05
CXorf24	65	0	7.02	4.15	6.02	1.69E-09
CXorf66	10	23	-1.20	-4.07	-9.63	5.81E-22
CYB5AP4	245	0	8.94	6.07	12.26	1.41E-34
CYCSP16	0	17	-5.09	-7.96	-8.25	1.53E-16
CYCSP34	39	0	6.29	3.42	4.36	1.29E-05
CYCSP43	97	40	1.28	-1.59	-7.63	2.28E-14
CYP11B2	12	12	0.00	-2.87	-6.04	1.55E-09
CYP17A1	0	9	-4.17	-7.04	-6.23	4.70E-10
CYP3A5P1	29	0	5.86	2.99	3.54	0.0004
DCAF16	59	0	6.88	4.01	5.68	1.33E-08
DCP1B	19	24	-0.34	-3.21	-9.00	2.19E-19
DCTPP1	51	0	6.67	3.80	5.19	2.08E-07
DDTL	59	0	6.88	4.01	5.68	1.33E-08
DDX1	39	0	6.29	3.42	4.36	1.29E-05
DDX50P1	48	0	6.58	3.72	5.00	5.83E-07
DDX58	34	0	6.09	3.22	3.97	7.18E-05
DDX60	82	0	7.36	4.49	6.90	5.18E-12
DDX6P2	3	5	-0.74	-3.61	-4.31	1.62E-05
DECR2	29	0	5.86	2.99	3.54	0.0004
DEFB118	58	0	6.86	3.99	5.62	1.87E-08
DERA	121	0	7.92	5.05	8.56	1.16E-17
DGKK	0	17	-5.09	-7.96	-8.25	1.53E-16
DHX57	81	0	7.34	4.47	6.85	7.27E-12
DIRAS2	1	3	-1.58	-4.45	-3.56	0.0004
DLX6	0	79	-7.30	-10.17	-15.50	3.73E-54
DMC1	88	0	7.46	4.59	7.18	6.83E-13
DMPK	0	80	-7.32	-10.19	-15.57	1.13E-54
DNAH5	17	22	-0.37	-3.24	-8.66	4.67E-18
DNAJA2	57	0	6.83	3.96	5.56	2.63E-08
DNAJB13	153	0	8.26	5.39	9.68	3.60E-22

DNAJC12	38	0	6.25	3 38	4.29	1.82E-05
DND1	36	0	6.17	3.30	4.13	3.62E-05
DNM1P28	64	0	7.00	4.13	5.97	2.38E-09
DPM3	0	23	-5.52	-8.39	-9.39	5.99E-21
DPP9	41	0	6.36	3.49	4.51	6.49E-06
ECEL1P2	0	13	-4.70	-7.57	-7.34	2.08E-13
EDDM3A	36	0	6.17	3.30	4.13	3.62E-05
EEF1A1P13	4	9	-1.17	-4.04	-6.01	1.84E-09
EEF1A1P24	80	0	7.32	4.45	6.80	1.02E-11
EEF1A1P32	40	0	6.32	3.45	4.44	9.16E-06
EIF1AX	0	5	-3.32	-6.19	-4.73	2.25E-06
EIF3LP1	81	0	7.34	4.47	6.85	7.27E-12
EIF4A1P12	34	0	6.09	3.22	3.97	7.18E-05
EIF4BP5	42	0	6.39	3.52	4.58	4.60E-06
EMX2	0	8	-4.00	-6.87	-5.90	3.59E-09
ENPP6	41	1	5.36	2.49	3.83	0.0001
EPB42	23	53	-1.20	-4.07	-14.63	1.88E-48
EXD1	42	0	6.39	3.52	4.58	4.60E-06
F13A1	73	0	7.19	4.32	6.45	1.10E-10
FABP5P6	32	0	6.00	3.13	3.80	0.0001
FAM101A	43	0	6.43	3.56	4.65	3.26E-06
FAM106A	13	7	0.89	-1.98	-3.71	0.0002
FAM138D	37	0	6.21	3.34	4.21	2.57E-05
FAM175B	28	0	5.81	2.94	3.45	0.0006
FAM177A1	59	0	6.88	4.01	5.68	1.33E-08
FAM182A	11	51	-2.21	-5.08	-15.03	4.69E-51
FAM183B	48	0	6.58	3.72	5.00	5.83E-07
FAM197Y1	3	45	-3.91	-6.78	-14.03	9.64E-45
FAM199YP	61	0	6.93	4.06	5.80	6.67E-09
FAM200A	0	7	-3.81	-6.68	-5.55	2.88E-08
FAM32B	52	30	0.79	-2.08	-7.93	2.23E-15
FAM75A1	19	22	-0.21	-3.08	-8.46	2.57E-17
FAM78A	3	5	-0.74	-3.61	-4.31	1.62E-05
FAM81B	85	22	1.95	-0.92	-3.67	0.0002
FAM8A2P	2	4	-1.00	-3.87	-3.95	7.70E-05
FAM92A1	37	0	6.21	3.34	4.21	2.57E-05
FAT1P1	41	121	-1.56	-4.43	-22.60	4.15E-113
FAUP2	1	10	-3.32	-6.19	-6.69	2.25E-11
FBP2	29	0	5.86	2.99	3.54	0.0004
FBXL17	40	0	6.32	3.45	4.44	9.16E-06
FBXO40	65	0	7.02	4.15	6.02	1.69E-09
FDPSP3	70	0	7.13	4.26	6.30	3.07E-10
FDX1L	0	44	-6.46	-9.33	-12.28	1.23E-34
FLJ43879	97	0	7.60	4.73	7.59	3.32E-14
FLJ45256	42	0	6.39	3.52	4.58	4.60E-06
FLNB	63	0	6.98	4.11	5.91	3.36E-09
FLOT1	30	0	5.91	3.04	3.63	0.0003
FMO11P	30	0	5.91	3.04	3.63	0.0003
FOLR2	39	0	6.29	3.42	4.36	1.29E-05
FOXD4L2	0	7	-3.81	-6.68	-5.55	2.88E-08
FRAT1	107	6	4.16	1.29	3.94	8.24E-05
FRK	27	0	5.75	2.89	3.36	0.0008
FTH1P21	43	0	6.43	3.56	4.65	3.26E-06
FTH1P25	0	5	-3.32	-6.19	-4.73	2.25E-06
FTLP11	0	12	-4.58	-7.45	-7.09	1.36E-12

FTLP2	39	19	1.04	-1.83	-5.81	6.34E-09
GABRR3	0	13	-4.70	-7.57	-7.34	2.08E-13
GAL3ST3	65	0	7.02	4.15	6.02	1.69E-09
GALC	24	12	1.00	-1.87	-4.68	2.88E-06
GAPDHP45	0	13	-4.70	-7.57	-7.34	2.08E-13
GAPDHP54	0	13	-4.70	-7.57	-7.34	2.08E-13
GAPDHP69	197	57	1.79	-1.08	-6.76	1.38E-11
GAPDHP72	27	0	5.75	2.89	3.36	0.0008
GCDH	0	5	-3.32	-6.19	-4.73	2.25E-06
GCNT1	0	7	-3.81	-6.68	-5.55	2.88E-08
GCNT1P2	2	4	-1.00	-3.87	-3.95	7.70E-05
GCNT6	34	0	6.09	3.22	3.97	7.18E-05
GEMIN8	54	0	6.75	3.89	5.38	7.39E-08
GEMIN8P3	32	0	6.00	3.13	3.80	0.0001
GGNBP1	57	0	6.83	3.96	5.56	2.63E-08
GGT7	0	8	-4.00	-6.87	-5.90	3.59E-09
GIP	32	0	6.00	3.13	3.80	0.0001
GJD4	70	0	7.13	4.26	6.30	3.07E-10
GKAP1	0	6	-3.58	-6.45	-5.16	2.46E-07
GLTSCR2	48	0	6.58	3.72	5.00	5.83E-07
GLUD1P5	37	1	5.21	2.34	3.51	0.0005
GLULP2	32	0	6.00	3.13	3.80	0.0001
GLYATL1P2	74	0	7.21	4.34	6.50	7.84E-11
GNL1	36	0	6.17	3.30	4.13	3.62E-05
GOLGA6L13P	7	8	-0.19	-3.06	-5.09	3.60E-07
GOLGA6L16P	39	1	5.29	2.42	3.67	0.0002
GOT2	0	9	-4.17	-7.04	-6.23	4.70E-10
GP2	123	0	7.94	5.07	8.63	6.03E-18
GPR112	31	0	5.95	3.08	3.72	0.0002
GPR18	136	0	8.09	5.22	9.10	8.67E-20
GPR53P	0	16	-5.00	-7.87	-8.04	8.93E-16
GPR83	0	11	-4.46	-7.33	-6.82	9.18E-12
GRB7	70	0	7.13	4.26	6.30	3.07E-10
GRTP1	33	0	6.04	3.18	3.89	0.0001
GSDMA	0	9	-4.17	-7.04	-6.23	4.70E-10
GTF2H2	43	0	6.43	3.56	4.65	3.26E-06
GTF2H2C	1	6	-2.58	-5.45	-5.19	2.14E-07
GTF2IRD2P1	65	21	1.63	-1.24	-4.58	4.71E-06
GUCA1A	143	0	8.16	5.29	9.35	8.99E-21
GYG2P1	1	9	-3.17	-6.04	-6.36	2.08E-10
GYPE	34	39	-0.20	-3.07	-11.25	2.37E-29
GZMM	36	0	6.17	3.30	4.13	3.62E-05
H2BFS	52	29	0.84	-2.03	-7.68	1.65E-14
HAPLN4	72	0	7.17	4.30	6.40	1.55E-10
HAUS1	35	0	6.13	3.26	4.05	5.10E-05
HAUS4	44	0	6.46	3.59	4.72	2.31E-06
HAVCR2	18	128	-2.83	-5.70	-24.00	2.93E-127
HCG18	29	129	-2.15	-5.02	-23.87	5.65E-126
HCG22	0	27	-5.75	-8.62	-10.04	9.79E-24
HDHD1P1	60	28	1.10	-1.77	-6.89	5.75E-12
HERC2P4	135	103	0.39	-2.48	-16.37	3.24E-60
HERC2P5	49	0	6.61	3.75	5.06	4.13E-07
HK2P1	86	2	5.43	2.56	5.63	1.84E-08
HKDC1	31	0	5.95	3.08	3.72	0.0002
HLA-U	289	0	9.17	6.31	13.28	2.88E-40

HMGB1P1	43	0	6.43	3.56	4.65	3.26E-06
HMGB1P12	64	0	7.00	4.13	5.97	2.38E-09
HMGB1P9	29	0	5.86	2.99	3.54	0.0004
HMGB4	51	0	6.67	3.80	5.19	2.08E-07
HMGN5	83	0	7.38	4.51	6.95	3.70E-12
HMOX1	38	0	6.25	3.38	4.29	1.82E-05
HNRPA1L3	29	0	5.86	2.99	3.54	0.0004
HPS4	41	0	6.36	3.49	4.51	6.49E-06
HPSE	0	5	-3.32	-6.19	-4.73	2.25E-06
HR	0	5	-3.32	-6.19	-4.73	2.25E-06
HRASLS2	155	0	8.28	5.41	9.75	1.90E-22
HRG	0	8	-4.00	-6.87	-5.90	3.59E-09
HS3ST6	1	5	-2.32	-5.19	-4.72	2.40E-06
HSD17B11	49	0	6.61	3.75	5.06	4.13E-07
HSD17B3	5	21	-2.07	-4.94	-9.61	7.44E-22
HSP90AB2P	13	22	-0.76	-3.63	-9.07	1.23E-19
HSP90AB3P	39	0	6.29	3.42	4.36	1.29E-05
HSPA2	32	0	6.00	3.13	3.80	0.0001
HSPB3	0	14	-4.81	-7.68	-7.59	3.29E-14
HSPD1P17	22	49	-1.16	-4.02	-14.01	1.29E-44
HTT	0	81	-7.34	-10.21	-15.65	3.44E-55
IDH1	58	0	6.86	3.99	5.62	1.87E-08
IDH2	2	6	-1.58	-4.45	-5.04	4.73E-07
IDS	42	1	5.39	2.52	3.90	9.52E-05
IFNA16	41	0	6.36	3.49	4.51	6.49E-06
IFNNP1	0	6	-3.58	-6.45	-5.16	2.46E-07
IGFL4	37	100	-1.43	-4.30	-20.40	1.64E-92
IGHV1OR15-9	0	6	-3.58	-6.45	-5.16	2.46E-07
IGHV3-42	7	6	0.22	-2.65	-4.09	4.24E-05
IGKV6-21	84	0	7.39	4.52	7.00	2.64E-12
IGLV1-62	52	0	6.70	3.83	5.26	1.47E-07
IGLV4-60	77	30	1.36	-1.51	-6.36	2.02E-10
IL28RA	35	0	6.13	3.26	4.05	5.10E-05
IL3RA	79	0	7.30	4.43	6.75	1.43E-11
IL3RA	40	0	6.32	3.45	4.44	9.16E-06
IL9R	2	11	-2.46	-5.33	-7.01	2.35E-12
IMMTP1	64	0	7.00	4.13	5.97	2.38E-09
IMP4	36	0	6.17	3.30	4.13	3.62E-05
IMPDH1P9	55	0	6.78	3.91	5.44	5.24E-08
INSIG2	43	0	6.43	3.56	4.65	3.26E-06
INSL3	27	0	5.75	2.89	3.36	0.0008
ITGAL	68	0	7.09	4.22	6.19	6.07E-10
ITGB1P1	82	0	7.36	4.49	6.90	5.18E-12
ITIH5	29	0	5.86	2.99	3.54	0.0004
ITSN2	3	8	-1.42	-4.28	-5.76	8.34E-09
IVD	32	0	6.00	3.13	3.80	0.0001
JMJD6	82	0	7.36	4.49	6.90	5.18E-12
KCNN4	65	0	7.02	4.15	6.02	1.69E-09
KDSR	12	7	0.78	-2.09	-3.85	0.0001
KIF23	0	13	-4.70	-7.57	-7.34	2.08E-13
KIF2C	32	0	6.00	3.13	3.80	0.0001
KLHL28	35	0	6.13	3.26	4.05	5.10E-05
KRT124P	0	8	-4.00	-6.87	-5.90	3.59E-09
KRT18P19	33	28	0.24	-2.63	-8.82	1.16E-18
KRT18P35	43	34	0.34	-2.53	-9.51	1.92E-21

KRT18P48	27	0	5.75	2.89	3.36	0.0008
KRT23	70	0	7.13	4.26	6.30	3.07E-10
KRT26	58	0	6.86	3.99	5.62	1.87E-08
KRT77	0	8	-4.00	-6.87	-5.90	3.59E-09
KRT80	27	0	5.75	2.89	3.36	0.0008
KRT8P23	87	0	7.44	4.57	7.14	9.57E-13
KRTAP10-8	3	77	-4.68	-7.55	-17.89	1.49E-71
KRTAP19-5	146	0	8.19	5.32	9.45	3.41E-21
KRTAP2-2	11	36	-1.71	-4.58	-12.41	2.20E-35
KRTAP4-3	35	0	6.13	3.26	4.05	5.10E-05
KRTAP5-8	0	8	-4.00	-6.87	-5.90	3.59E-09
KRTAP6-3	0	12	-4.58	-7.45	-7.09	1.36E-12
LACRT	78	0	7.29	4.42	6.71	2.01E-11
LAMC3	0	7	-3.81	-6.68	-5.55	2.88E-08
LANCL1	0	11	-4.46	-7.33	-6.82	9.18E-12
LATS2	27	0	5.75	2.89	3.36	0.0008
LCN15	56	0	6.81	3.94	5.50	3.71E-08
LCN1L1	28	0	5.81	2.94	3.45	0.0006
LDHBL1	54	0	6.75	3.89	5.38	7.39E-08
LETM1	2	4	-1.00	-3.87	-3.95	7.70E-05
LETM1P3	102	6	4.09	1.22	3.68	0.0002
LETMD1	6	24	-2.00	-4.87	-10.25	1.20E-24
LILRA3	69	0	7.11	4.24	6.24	4.31E-10
LIMD1	29	0	5.86	2.99	3.54	0.0004
LIPA	6	5	0.26	-2.61	-3.71	0.0002
LONP2	43	0	6.43	3.56	4.65	3.26E-06
LPA	28	0	5.81	2.94	3.45	0.0006
LPPR3	0	11	-4.46	-7.33	-6.82	9.18E-12
LRRN4CL	44	0	6.46	3.59	4.72	2.31E-06
LST-3TM12	0	11	-4.46	-7.33	-6.82	9.18E-12
LYPD3	0	9	-4.17	-7.04	-6.23	4.70E-10
LYPD4	1	7	-2.81	-5.68	-5.61	2.02E-08
LYPLA1	28	0	5.81	2.94	3.45	0.0006
LYPLA2P1	35	0	6.13	3.26	4.05	5.10E-05
MAGOHB	43	0	6.43	3.56	4.65	3.26E-06
MANBAL	139	0	8.12	5.25	9.21	3.28E-20
MAP1LC3P	29	0	5.86	2.99	3.54	0.0004
MAPK9	50	0	6.64	3.77	5.13	2.93E-07
MBD3L5	43	0	6.43	3.56	4.65	3.26E-06
MBD4	1	22	-4.46	-7.33	-9.64	5.24E-22
MBD5	0	80	-7.32	-10.19	-15.57	1.13E-54
MCART4P	4	6	-0.58	-3.45	-4.65	3.40E-06
MCM9	14	8	0.81	-2.06	-4.08	4.59E-05
MDS2	0	17	-5.09	-7.96	-8.25	1.53E-16
MEMO1P1	31	0	5.95	3.08	3.72	0.0002
METT10D	38	0	6.25	3.38	4.29	1.82E-05
MEX3A	34	0	6.09	3.22	3.97	7.18E-05
MFAP5	30	0	5.91	3.04	3.63	0.0003
MIAT	36	0	6.17	3.30	4.13	3.62E-05
MID1IP1	1	6	-2.58	-5.45	-5.19	2.14E-07
MIR122	0	43	-6.43	-9.30	-12.16	4.99E-34
MIR1246	0	13	-4.70	-7.57	-7.34	2.08E-13
MIR1256	27	0	5.75	2.89	3.36	0.0008
MIR1267	38	0	6.25	3.38	4.29	1.82E-05
MIR1277	1	12	-3.58	-6.45	-7.30	2.91E-13

MIR1294	0	13	-4.70	-7.57	-7.34	2.08E-13
MIR1305	64	0	7.00	4.13	5.97	2.38E-09
MIR1308	30	13	1.21	-1.66	-4.49	7.12E-06
MIR138-2	29	0	5.86	2.99	3.54	0.0004
MIR220A	87	0	7.44	4.57	7.14	9.57E-13
MIR302F	169	0	8.40	5.53	10.19	2.20E-24
MIR3123	108	0	7.75	4.89	8.05	8.50E-16
MIR3134	57	0	6.83	3.96	5.56	2.63E-08
MIR3172	20	16	0.32	-2.55	-6.55	5.89E-11
MIR3179-1	0	6	-3.58	-6.45	-5.16	2.46E-07
MIR3187	0	11	-4.46	-7.33	-6.82	9.18E-12
MIR4284	30	0	5.91	3.04	3.63	0.0003
MIR4293	66	0	7.04	4.18	6.08	1.20E-09
MIR4306	45	17	1.40	-1.46	-4.68	2.85E-06
MIR4327	28	0	5.81	2.94	3.45	0.0006
MIR4330	73	1	6.19	3.32	5.90	3.70E-09
MIR500B	1	7	-2.81	-5.68	-5.61	2.02E-08
MIR501	1	7	-2.81	-5.68	-5.61	2.02E-08
MIR514B	0	6	-3.58	-6.45	-5.16	2.46E-07
MIR548D2	286	0	9.16	6.29	13.22	6.97E-40
MIR548H3	50	0	6.64	3.77	5.13	2.93E-07
MIR548O	46	0	6.52	3.65	4.86	1.16E-06
MIR552	6	6	0.00	-2.87	-4.27	1.95E-05
MIR628	8	7	0.19	-2.68	-4.45	8.64E-06
MIR631	36	0	6.17	3.30	4.13	3.62E-05
MIR659	135	3	5.49	2.62	7.15	8.75E-13
MIR941-1	12	9	0.42	-2.45	-4.81	1.53E-06
MLKL	33	0	6.04	3.18	3.89	0.0001
MND1	29	0	5.86	2.99	3.54	0.0004
MORF4	76	2	5.25	2.38	5.07	3.88E-07
MPP7	28	0	5.81	2.94	3.45	0.0006
MPPE1	82	0	7.36	4.49	6.90	5.18E-12
MPTX	62	0	6.95	4.08	5.86	4.73E-09
MPZL3	40	0	6.32	3.45	4.44	9.16E-06
MRO	55	2	4.78	1.91	3.76	0.0002
MRP63P9	131	0	8.03	5.16	8.93	4.41E-19
MRPL50P2	35	0	6.13	3.26	4.05	5.10E-05
MRPL9	42	0	6.39	3.52	4.58	4.60E-06
MRPS16P	9	7	0.36	-2.51	-4.29	1.78E-05
MRPS33P2	141	0	8.14	5.27	9.28	1.72E-20
MS4A4E	224	0	8.81	5.94	11.74	8.22E-32
MSR1	54	0	6.75	3.89	5.38	7.39E-08
MTCO2P1	51	0	6.67	3.80	5.19	2.08E-07
MTHFD1L	64	0	7.00	4.13	5.97	2.38E-09
MTMR3	38	0	6.25	3.38	4.29	1.82E-05
MTND2P2	52	0	6.70	3.83	5.26	1.47E-07
MTRF1L	68	0	7.09	4.22	6.19	6.07E-10
MTRNR2L9	2	4	-1.00	-3.87	-3.95	7.70E-05
MX1	132	0	8.04	5.18	8.96	3.18E-19
MX2	1	44	-5.46	-8.33	-13.03	7.94E-39
MYCBP2	0	42	-6.39	-9.26	-12.05	2.04E-33
MYL6P3	89	0	7.48	4.61	7.23	4.88E-13
MYLK4	0	13	-4.70	-7.57	-7.34	2.08E-13
MYLPF	0	5	-3.32	-6.19	-4.73	2.25E-06
MYO5B	40	0	6.32	3.45	4.44	9.16E-06

MZT1	79	0	7.30	4.43	6.75	1.43E-11
N6AMT1	24	35	-0.54	-3.41	-11.17	5.89E-29
NAA20	34	0	6.09	3.22	3.97	7.18E-05
NAA25	36	0	6.17	3.30	4.13	3.62E-05
NAAA	78	0	7.29	4.42	6.71	2.01E-11
NANOGP1	12	6	1.00	-1.87	-3.31	0.0009
NANOGP8	98	0	7.61	4.75	7.63	2.38E-14
NBPF15	0	7	-3.81	-6.68	-5.55	2.88E-08
NCAPH	36	0	6.17	3.30	4.13	3.62E-05
NCOA1	37	21	0.82	-2.05	-6.58	4.58E-11
NCOR1	110	2	5.78	2.91	6.81	9.46E-12
NCRNA00032	81	0	7.34	4.47	6.85	7.27E-12
NCRNA00092	45	0	6.49	3.62	4.79	1.64E-06
NCRNA00111	31	0	5.95	3.08	3.72	0.0002
NCRNA00114	43	0	6.43	3.56	4.65	3.26E-06
NCRNA00116	61	0	6.93	4.06	5.80	6.67E-09
NCRNA00161	58	0	6.86	3.99	5.62	1.87E-08
NCRNA00230B	1	6	-2.58	-5.45	-5.19	2.14E-07
NCRNA00292	56	0	6.81	3.94	5.50	3.71E-08
NDUFA2	0	43	-6.43	-9.30	-12.16	4.99E-34
NDUFA4P1	0	9	-4.17	-7.04	-6.23	4.70E-10
NDUFAF1	36	0	6.17	3.30	4.13	3.62E-05
NDUFB3P1	0	21	-5.39	-8.26	-9.04	1.63E-19
NDUFB3P4	35	0	6.13	3.26	4.05	5.10E-05
NDUFB8P1	31	0	5.95	3.08	3.72	0.0002
NDUFB8P3	108	0	7.75	4.89	8.05	8.50E-16
NDUFS1	28	0	5.81	2.94	3.45	0.0006
NDUFV2P1	93	0	7.54	4.67	7.41	1.27E-13
NEBL	3	11	-1.87	-4.74	-6.91	4.92E-12
NEK5	303	0	9.24	6.37	13.59	4.78E-42
NFXL1	34	0	6.09	3.22	3.97	7.18E-05
NHP2L1	24	9	1.42	-1.45	-3.39	0.0007
NIPA1	1	3	-1.58	-4.45	-3.56	0.0004
NOL9	1	4	-2.00	-4.87	-4.18	2.87E-05
NPM1P10	50	0	6.64	3.77	5.13	2.93E-07
NPM1P12	66	0	7.04	4.18	6.08	1.20E-09
NPM1P19	45	0	6.49	3.62	4.79	1.64E-06
NPM1P2	2	9	-2.17	-5.04	-6.31	2.86E-10
NPM1P3	0	16	-5.00	-7.87	-8.04	8.93E-16
NPW	29	0	5.86	2.99	3.54	0.0004
NSL1	132	0	8.04	5.18	8.96	3.18E-19
NUBPL	37	0	6.21	3.34	4.21	2.57E-05
NUP214	36	0	6.17	3.30	4.13	3.62E-05
NUPR1	0	7	-3.81	-6.68	-5.55	2.88E-08
NUS1P3	35	0	6.13	3.26	4.05	5.10E-05
NUS1P4	57	0	6.83	3.96	5.56	2.63E-08
OCIAD2	20	15	0.42	-2.45	-6.21	5.42E-10
ODF3L2	50	0	6.64	3.77	5.13	2.93E-07
OFD1P12Y	94	0	7.55	4.69	7.45	9.08E-14
OR10V7P	2	11	-2.46	-5.33	-7.01	2.35E-12
OR1M4P	33	0	6.04	3.18	3.89	0.0001
OR2AT1P	0	12	-4.58	-7.45	-7.09	1.36E-12
OR2M1P	160	2	6.32	3.45	8.87	7.11E-19
OR2M7	36	0	6.17	3.30	4.13	3.62E-05
OR4A14P	16	22	-0.46	-3.33	-8.76	1.94E-18

OR4A3P	3	32	-3.42	-6.28	-11.95	6.34E-33
OR51H1P	0	95	-7.57	-10.44	-16.65	2.97E-62
OR5BB1P	0	8	-4.00	-6.87	-5.90	3.59E-09
OR5BJ1P	72	0	7.17	4.30	6.40	1.55E-10
OR7A17	0	5	-3.32	-6.19	-4.73	2.25E-06
OR7A18P	3	16	-2.42	-5.28	-8.45	2.89E-17
OR7A8P	1	25	-4.64	-7.51	-10.21	1.84E-24
OR7D11P	88	0	7.46	4.59	7.18	6.83E-13
OR7E100P	28	0	5.81	2.94	3.45	0.0006
OR7E101P	0	22	-5.46	-8.33	-9.22	3.10E-20
OR7G15P	11	7	0.65	-2.22	-3.99	6.59E-05
OR8K5	28	0	5.81	2.94	3.45	0.0006
OSBPL3	79	0	7.30	4.43	6.75	1.43E-11
OXGR1	28	0	5.81	2.94	3.45	0.0006
PABPC1P8	28	0	5.81	2.94	3.45	0.0006
PABPC3	128	0	8.00	5.13	8.82	1.17E-18
PAICSP7	0	17	-5.09	-7.96	-8.25	1.53E-16
PAPSS1	58	53	0.13	-2.74	-12.40	2.73E-35
PARK2	162	5	5.02	2.15	6.97	3.27E-12
PARP4	115	0	7.85	4.98	8.33	8.38E-17
PATE4	30	0	5.91	3.04	3.63	0.0003
PDE6A	2	5	-1.32	-4.19	-4.53	5.95E-06
PDZD3	131	0	8.03	5.16	8.93	4.41E-19
PEBP1P1	0	11	-4.46	-7.33	-6.82	9.18E-12
PELP1	0	12	-4.58	-7.45	-7.09	1.36E-12
PFAS	40	0	6.32	3.45	4.44	9.16E-06
PGF	0	8	-4.00	-6.87	-5.90	3.59E-09
PGM5P2	55	2	4.78	1.91	3.76	0.0002
PHACTR4	32	0	6.00	3.13	3.80	0.0001
PHKA1	47	0	6.55	3.69	4.93	8.23E-07
PIGCP1	122	0	7.93	5.06	8.59	8.37E-18
PIH1D1	0	5	-3.32	-6.19	-4.73	2.25E-06
PKIG	1	5	-2.32	-5.19	-4.72	2.40E-06
PLA2G10	38	0	6.25	3.38	4.29	1.82E-05
PLIN2	61	0	6.93	4.06	5.80	6.67E-09
PMS2L3	0	13	-4.70	-7.57	-7.34	2.08E-13
PMS2L4	0	29	-5.86	-8.73	-10.35	4.32E-25
PNPLA4P1	59	1	5.88	3.01	5.07	3.89E-07
POLR1B	0	13	-4.70	-7.57	-7.34	2.08E-13
POLR2B	50	0	6.64	3.77	5.13	2.93E-07
POLR2F	51	0	6.67	3.80	5.19	2.08E-07
POTEM	27	0	5.75	2.89	3.36	0.0008
POU3F2	0	82	-7.36	-10.23	-15.72	1.05E-55
POU5F1P3	6	5	0.26	-2.61	-3.71	0.0002
PPEF1	0	9	-4.17	-7.04	-6.23	4.70E-10
PPIAP15	56	42	0.42	-2.45	-10.39	2.83E-25
PPM1F	33	0	6.04	3.18	3.89	0.0001
PPP1R12BP2	49	121	-1.30	-4.17	-22.26	8.79E-110
PPP1R13L	32	0	6.00	3.13	3.80	0.0001
PPP2R2A	1	3	-1.58	-4.45	-3.56	0.0004
PPP2R3C	14	23	-0.72	-3.59	-9.23	2.74E-20
PPP6CP	36	0	6.17	3.30	4.13	3.62E-05
PRDM14	13	19	-0.55	-3.42	-8.23	1.87E-16
PRDM6	35	0	6.13	3.26	4.05	5.10E-05
PSIP1P1	22	101	-2.20	-5.07	-21.15	2.94E-99

PSME2	42	0	6.39	3.52	4.58	4.60E-06
PTCD2	54	0	6.75	3.89	5.38	7.39E-08
PTCHD3	0	6	-3.58	-6.45	-5.16	2.46E-07
PTK6	30	0	5.91	3.04	3.63	0.0003
PTMAP2	113	0	7.82	4.95	8.25	1.62E-16
PTPLA	54	0	6.75	3.89	5.38	7.39E-08
PTPN21	70	0	7.13	4.26	6.30	3.07E-10
PTPN3	164	0	8.36	5.49	10.03	1.07E-23
PXMP4	277	0	9.11	6.24	13.02	9.93E-39
QDPR	77	0	7.27	4.40	6.66	2.83E-11
RAB3D	17	17	0.00	-2.87	-7.19	6.56E-13
RABL5	0	9	-4.17	-7.04	-6.23	4.70E-10
RAC1P2	21	87	-2.05	-4.92	-19.55	4.30E-85
RAET1F	10	23	-1.20	-4.07	-9.63	5.81E-22
RANP1	33	0	6.04	3.18	3.89	0.0001
RARG	40	1	5.32	2.45	3.75	0.0002
RBM22	2	7	-1.81	-4.68	-5.50	3.88E-08
RBM34	35	0	6.13	3.26	4.05	5.10E-05
REC8	28	0	5.81	2.94	3.45	0.0006
RESP18	1	5	-2.32	-5.19	-4.72	2.40E-06
REXO1L3P	88	0	7.46	4.59	7.18	6.83E-13
RFT1	0	5	-3.32	-6.19	-4.73	2.25E-06
RHAG	10	6	0.74	-2.13	-3.61	0.0003
RHBDL2	0	9	-4.17	-7.04	-6.23	4.70E-10
RHOT1P2	45	0	6.49	3.62	4.79	1.64E-06
RHOXF1	4	5	-0.32	-3.19	-4.10	4.13E-05
RHPN2	42	0	6.39	3.52	4.58	4.60E-06
RILPL1	37	0	6.21	3.34	4.21	2.57E-05
RLN1	51	0	6.67	3.80	5.19	2.08E-07
RN7SL3	35	0	6.13	3.26	4.05	5.10E-05
RNF222	50	0	6.64	3.77	5.13	2.93E-07
RNF38	35	0	6.13	3.26	4.05	5.10E-05
RNU3P2	276	0	9.11	6.24	12.99	1.34E-38
RNU4-7P	135	0	8.08	5.21	9.07	1.20E-19
RNU6-2	35	0	6.13	3.26	4.05	5.10E-05
RNU6ATAC3P	101	4	4.66	1.79	4.87	1.10E-06
RNU6ATAC5P	74	0	7.21	4.34	6.50	7.84E-11
RNU7-33P	6	6	0.00	-2.87	-4.27	1.95E-05
RNU7-35P	11	7	0.65	-2.22	-3.99	6.59E-05
RNU7-39P	40	0	6.32	3.45	4.44	9.16E-06
RNU7-50P	37	0	6.21	3.34	4.21	2.57E-05
RNU7-64P	38	1	5.25	2.38	3.59	0.0003
RNU7-66P	5	11	-1.14	-4.01	-6.63	3.36E-11
RNU7-72P	63	0	6.98	4.11	5.91	3.36E-09
RNU7-75P	57	1	5.83	2.96	4.95	7.51E-07
RNU7-77P	52	0	6.70	3.83	5.26	1.47E-07
RNU7-82P	4	9	-1.17	-4.04	-6.01	1.84E-09
RNY4P16	0	8	-4.00	-6.87	-5.90	3.59E-09
RNY5P6	66	0	7.04	4.18	6.08	1.20E-09
RTTN	0	7	-3.81	-6.68	-5.55	2.88E-08
SAA2	35	23	0.61	-2.26	-7.33	2.36E-13
SALL1P1	55	0	6.78	3.91	5.44	5.24E-08
SAR1B	1	3	-1.58	-4.45	-3.56	0.0004
SCARNA23	43	0	6.43	3.56	4.65	3.26E-06
SDCCAG3L	12	20	-0.74	-3.61	-8.62	6.43E-18

SEC14L3	31	0	5.95	3.08	3.72	0.0002
SEC23IP	0	14	-4.81	-7.68	-7.59	3.29E-14
SEC63	33	0	6.04	3.18	3.89	0.0001
SELENBP1	13	19	-0.55	-3.42	-8.23	1.87E-16
SELPLG	0	9	-4.17	-7.04	-6.23	4.70E-10
SERBP1P2	105	39	1.43	-1.44	-7.00	2.49E-12
SERPIND1	36	21	0.78	-2.09	-6.66	2.65E-11
SERPINF1	43	0	6.43	3.56	4.65	3.26E-06
SETP2	0	16	-5.00	-7.87	-8.04	8.93E-16
SF3A1	0	5	-3.32	-6.19	-4.73	2.25E-06
SF3A3P1	69	0	7.11	4.24	6.24	4.31E-10
SFPQ	70	0	7.13	4.26	6.30	3.07E-10
SFPQP1	4	11	-1.46	-4.33	-6.77	1.26E-11
SFRP2	29	0	5.86	2.99	3.54	0.0004
SGK1	38	0	6.25	3.38	4.29	1.82E-05
SH3KBP1	32	0	6.00	3.13	3.80	0.0001
SHCBP1	5	43	-3.10	-5.97	-13.90	6.45E-44
SHFM1P1	11	29	-1.40	-4.27	-10.96	5.98E-28
SIGLEC26P	146	0	8.19	5.32	9.45	3.41E-21
SIGLEC8	17	25	-0.56	-3.43	-9.45	3.35E-21
SIRPB1	8	25	-1.64	-4.51	-10.31	6.19E-25
SKA2L	203	8	4.67	1.80	6.93	4.24E-12
SKA3	70	0	7.13	4.26	6.30	3.07E-10
SLC12A6	64	0	7.00	4.13	5.97	2.38E-09
SLC14A2	51	61	-0.26	-3.13	-14.20	9.63E-46
SLC25A2	41	0	6.36	3.49	4.51	6.49E-06
SLC26A1	51	2	4.67	1.80	3.48	0.0005
SLC28A1	217	0	8.76	5.89	11.55	7.00E-31
SLC2A11	192	0	8.58	5.72	10.87	1.59E-27
SLC2A7	80	0	7.32	4.45	6.80	1.02E-11
SLC33A1	2	6	-1.58	-4.45	-5.04	4.73E-07
SLC4A9	6	9	-0.58	-3.45	-5.69	1.28E-08
SLC6A16	27	104	-1.95	-4.81	-21.30	1.09E-100
SLC7A2	29	0	5.86	2.99	3.54	0.0004
SLCO1B1	3	11	-1.87	-4.74	-6.91	4.92E-12
SLFN13	27	0	5.75	2.89	3.36	0.0008
SNAP23	50	0	6.64	3.77	5.13	2.93E-07
SNORA12	53	1	5.73	2.86	4.69	2.78E-06
SNORA2A	32	0	6.00	3.13	3.80	0.0001
SNORA69	42	0	6.39	3.52	4.58	4.60E-06
SNORA72	33	0	6.04	3.18	3.89	0.0001
SNORA74B	11	44	-2.00	-4.87	-13.88	8.62E-44
SNORD116-27	41	14	1.55	-1.32	-3.92	8.73E-05
SNORD82	6	8	-0.42	-3.28	-5.25	1.50E-07
SNRNP27	0	11	-4.46	-7.33	-6.82	9.18E-12
SNRNP48	31	0	5.95	3.08	3.72	0.0002
SNRPA1	66	0	7.04	4.18	6.08	1.20E-09
SNRPBP1	11	36	-1.71	-4.58	-12.41	2.20E-35
SNX22	0	8	-4.00	-6.87	-5.90	3.59E-09
SPAG1	2	13	-2.70	-5.57	-7.64	2.15E-14
SPANXN1	0	6	-3.58	-6.45	-5.16	2.46E-07
SPATA24	143	0	8.16	5.29	9.35	8.99E-21
SPC25	33	0	6.04	3.18	3.89	0.0001
SPECC1L	0	7	-3.81	-6.68	-5.55	2.88E-08
SPINK6	78	0	7.29	4.42	6.71	2.01E-11

SPOCK2	42	0	6.39	3.52	4.58	4.60E-06
SPRR2A	40	0	6.32	3.45	4.44	9.16E-06
SPRY3	0	11	-4.46	-7.33	-6.82	9.18E-12
SRCRB4D	36	0	6.17	3.30	4.13	3.62E-05
ST13P17	0	5	-3.32	-6.19	-4.73	2.25E-06
ST13P2	59	0	6.88	4.01	5.68	1.33E-08
ST13P5	11	9	0.29	-2.58	-4.94	7.64E-07
STARD10	89	0	7.48	4.61	7.23	4.88E-13
STARD9	0	9	-4.17	-7.04	-6.23	4.70E-10
STGC3	42	1	5.39	2.52	3.90	9.52E-05
STMN3	6	10	-0.74	-3.61	-6.10	1.07E-09
SUCLA2	49	0	6.61	3.75	5.06	4.13E-07
SUCLA2P1	46	0	6.52	3.65	4.86	1.16E-06
SUGT1P2	36	0	6.17	3.30	4.13	3.62E-05
SULT1C3	54	0	6.75	3.89	5.38	7.39E-08
SV2A	36	0	6.17	3.30	4.13	3.62E-05
SYCP3	139	0	8.12	5.25	9.21	3.28E-20
SYNJ2BP	0	6	-3.58	-6.45	-5.16	2.46E-07
SYT5	58	0	6.86	3.99	5.62	1.87E-08
TAF7L	114	0	7.83	4.96	8.29	1.17E-16
TBC1D10C	35	0	6.13	3.26	4.05	5.10E-05
TBC1D9	2	4	-1.00	-3.87	-3.95	7.70E-05
TCEA1P1	42	0	6.39	3.52	4.58	4.60E-06
TCEB1P22	12	63	-2.39	-5.26	-16.77	4.38E-63
TCTN3	116	0	7.86	4.99	8.36	6.03E-17
TDH	28	0	5.81	2.94	3.45	0.0006
TEC	25	11	1.18	-1.68	-4.17	3.06E-05
TEP1	59	0	6.88	4.01	5.68	1.33E-08
TERF2IPP1	6	11	-0.87	-3.74	-6.48	8.97E-11
TEX14	49	1	5.61	2.75	4.41	1.02E-05
TFF1	0	5	-3.32	-6.19	-4.73	2.25E-06
THAP10	31	0	5.95	3.08	3.72	0.0002
THAP11	0	82	-7.36	-10.23	-15.72	1.05E-55
THSD1	2	9	-2.17	-5.04	-6.31	2.86E-10
THUMPD1P1	68	3	4.50	1.63	3.75	0.0002
TIMELESS	29	0	5.86	2.99	3.54	0.0004
TIMM23	64	0	7.00	4.13	5.97	2.38E-09
TJP3	32	0	6.00	3.13	3.80	0.0001
TKTL1	36	0	6.17	3.30	4.13	3.62E-05
TLR12P	36	0	6.17	3.30	4.13	3.62E-05
TM2D2	35	17	1.04	-1.83	-5.48	4.15E-08
TM4SF19	65	0	7.02	4.15	6.02	1.69E-09
TMBIM6	58	0	6.86	3.99	5.62	1.87E-08
TMED2	0	6	-3.58	-6.45	-5.16	2.46E-07
TMEM155	63	0	6.98	4.11	5.91	3.36E-09
TMEM179B	3	13	-2.12	-4.98	-7.57	3.77E-14
TMEM186	0	5	-3.32	-6.19	-4.73	2.25E-06
TMEM202	230	0	8.85	5.98	11.89	1.32E-32
TMEM213	139	0	8.12	5.25	9.21	3.28E-20
TMEM221	0	5	-3.32	-6.19	-4.73	2.25E-06
TMEM60	3	7	-1.22	-4.09	-5.32	1.03E-07
TMPRSS2	3	11	-1.87	-4.74	-6.91	4.92E-12
TMPRSS3	58	0	6.86	3.99	5.62	1.87E-08
TMSB4XP2	61	0	6.93	4.06	5.80	6.67E-09
TMSL7	13	9	0.53	-2.34	-4.67	2.96E-06

TMTC4	5	13	-1.38	-4.25	-7.33	2.33E-13
TNS4	3	43	-3.84	-6.71	-13.74	5.75E-43
TP53BP2	31	0	5.95	3.08	3.72	0.0002
TPM5P	55	0	6.78	3.91	5.44	5.24E-08
TPRX1	67	0	7.07	4.20	6.13	8.54E-10
TPTE	29	0	5.86	2.99	3.54	0.0004
TPTE2P1	1	13	-3.70	-6.57	-7.58	3.48E-14
TRADD	43	0	6.43	3.56	4.65	3.26E-06
TRAF3IP2	53	0	6.73	3.86	5.32	1.04E-07
TRAV19	63	22	1.52	-1.35	-5.01	5.41E-07
TRAV8-1	15	16	-0.09	-2.96	-7.08	1.39E-12
TRBV7-1	25	13	0.94	-1.93	-4.97	6.74E-07
TRIM16	37	0	6.21	3.34	4.21	2.57E-05
TRIM34	1	6	-2.58	-5.45	-5.19	2.14E-07
TRIM4	1	4	-2.00	-4.87	-4.18	2.87E-05
TRNAA44P	1	4	-2.00	-4.87	-4.18	2.87E-05
TRNAE39P	92	0	7.52	4.65	7.36	1.78E-13
TRNAS29P	92	0	7.52	4.65	7.36	1.78E-13
TRNASUP6P	137	39	1.81	-1.06	-5.49	3.99E-08
TRNAV33P	45	0	6.49	3.62	4.79	1.64E-06
TRPV4	151	0	8.24	5.37	9.62	6.84E-22
TSG1	22	15	0.55	-2.32	-6.00	1.97E-09
TSPAN15	28	0	5.81	2.94	3.45	0.0006
TSPYL1	39	1	5.29	2.42	3.67	0.0002
TSSK1A	0	5	-3.32	-6.19	-4.73	2.25E-06
TTC22	0	8	-4.00	-6.87	-5.90	3.59E-09
TTC28	32	0	6.00	3.13	3.80	0.0001
TTC39C	55	0	6.78	3.91	5.44	5.24E-08
TUBBP1	98	4	4.61	1.75	4.72	2.37E-06
TULP2	0	7	-3.81	-6.68	-5.55	2.88E-08
TXNL4A	93	0	7.54	4.67	7.41	1.27E-13
TXNRD1	33	0	6.04	3.18	3.89	0.0001
UBA52P9	0	8	-4.00	-6.87	-5.90	3.59E-09
UBR3	0	44	-6.46	-9.33	-12.28	1.23E-34
UPP1	0	5	-3.32	-6.19	-4.73	2.25E-06
UQCR11	29	0	5.86	2.99	3.54	0.0004
UQCRHP1	65	3	4.44	1.57	3.56	0.0004
URB1	144	0	8.17	5.30	9.38	6.51E-21
USP12PY	1	3	-1.58	-4.45	-3.56	0.0004
USP33	27	0	5.75	2.89	3.36	0.0008
USP45	65	0	7.02	4.15	6.02	1.69E-09
USP9YP7	9	35	-1.96	-4.83	-12.36	4.27E-35
UTS2D	0	6	-3.58	-6.45	-5.16	2.46E-07
VDAC1P6	27	0	5.75	2.89	3.36	0.0008
VENTXP7	52	0	6.70	3.83	5.26	1.47E-07
VHL	28	0	5.81	2.94	3.45	0.0006
VN1R107P	10	9	0.15	-2.72	-5.09	3.67E-07
VN1R14P	0	6	-3.58	-6.45	-5.16	2.46E-07
VN1R45P	2	4	-1.00	-3.87	-3.95	7.70E-05
VN1R46P	44	0	6.46	3.59	4.72	2.31E-06
VN1R47P	64	0	7.00	4.13	5.97	2.38E-09
VN1R68P	115	51	1.17	-1.70	-9.02	1.83E-19
VN1R76P	0	9	-4.17	-7.04	-6.23	4.70E-10
VN1R7P	41	0	6.36	3.49	4.51	6.49E-06
VN1R8P	51	0	6.67	3.80	5.19	2.08E-07

VN1R9P	53	0	6.73	3.86	5.32	1.04E-07
VNN2	43	0	6.43	3.56	4.65	3.26E-06
VOPP1	33	0	6.04	3.18	3.89	0.0001
VPS11	0	7	-3.81	-6.68	-5.55	2.88E-08
WDR13	3	21	-2.81	-5.68	-9.72	2.55E-22
WDR73	59	0	6.88	4.01	5.68	1.33E-08
WDR87	36	0	6.17	3.30	4.13	3.62E-05
WFDC5	1	43	-5.43	-8.30	-12.91	4.13E-38
WFDC8	50	0	6.64	3.77	5.13	2.93E-07
XCL2	0	9	-4.17	-7.04	-6.23	4.70E-10
XKRX	44	0	6.46	3.59	4.72	2.31E-06
XRCC2	0	6	-3.58	-6.45	-5.16	2.46E-07
YDJC	40	0	6.32	3.45	4.44	9.16E-06
YKT6	69	0	7.11	4.24	6.24	4.31E-10
YPEL4	36	0	6.17	3.30	4.13	3.62E-05
YWHAZP4	54	0	6.75	3.89	5.38	7.39E-08
YWHAZP6	1	43	-5.43	-8.30	-12.91	4.13E-38
YWHAZP8	59	0	6.88	4.01	5.68	1.33E-08
ZDHHC18	0	9	-4.17	-7.04	-6.23	4.70E-10
ZG16	0	5	-3.32	-6.19	-4.73	2.25E-06
ZNF333	110	0	7.78	4.91	8.13	4.38E-16
ZNF37BP	15	57	-1.93	-4.80	-15.76	6.14E-56
ZNF410	2	9	-2.17	-5.04	-6.31	2.86E-10
ZNF415	0	6	-3.58	-6.45	-5.16	2.46E-07
ZNF445	2	11	-2.46	-5.33	-7.01	2.35E-12
ZNF480	0	13	-4.70	-7.57	-7.34	2.08E-13
ZNF543	33	0	6.04	3.18	3.89	0.0001
ZNF607	0	11	-4.46	-7.33	-6.82	9.18E-12
ZNF630	2	9	-2.17	-5.04	-6.31	2.86E-10
ZNF680	95	0	7.57	4.70	7.50	6.50E-14
ZNF70	40	0	6.32	3.45	4.44	9.16E-06
ZNF75BP	21	40	-0.93	-3.80	-12.43	1.85E-35
ZNF773	0	16	-5.00	-7.87	-8.04	8.93E-16
ZNF785	40	0	6.32	3.45	4.44	9.16E-06
ZRSR1	108	0	7.75	4.89	8.05	8.50E-16
ZSCAN1	0	7	-3.81	-6.68	-5.55	2.88E-08
ZSCAN4	28	0	5.81	2.94	3.45	0.0006

Appendix 2 Complete list gut microbes found in AIDS patients (arranged in alphabetical order).

Genus	Species	Frequency per genus	Frequency per species
Akkermansia		1	
	Akkermansia muciniphila ATCC BAA-835		1
Anaerococcus		2	
	Anaerococcus hydrogenalis DSM 7454		2
Anaerofustis		2	
	Anaerofustis stercorihominis DSM 17244		2
Bacteroides		11	
	Bacteroides caccae ATCC 43185		1
	Bacteroides capillosus ATCC 29799		2
	Bacteroides cellulosilyticus DSM 14838		1
	Bacteroides coprophilus DSM 18228		1
	Bacteroides fragilis 3_1_12		1
	Bacteroides intestinalis DSM 17393		1
	Bacteroides plebeius DSM 17135		2
	Bacteroides sp. 4_3_47FAA		1
	Bacteroides sp. 9_1_42FAA		1
Bifidobacterium		4	
	Bifidobacterium adolescentis L2-32		2
	Bifidobacterium animalis subsp. lactis AD011		1
	Bifidobacterium gallicum DSM 20093		1
Blautia		3	
	Blautia hansenii DSM 20583		1
	Blautia hydrogenotrophica DSM 10507		2
Butyrivibrio		2	
	Butyrivibrio crossotus DSM 2876		2
Catenibacterium		3	
	Catenibacterium mitsuokai DSM 15897		3
Citrobacter		1	
	Citrobacter sp. 30_2		1
Clostridiales		2	
	Clostridiales bacterium 1_7_47FAA		2
Clostridium		29	
	Clostridium bartlettii DSM 16795		4
	Clostridium hathewayi DSM 13479		3
	Clostridium methylpentosum DSM 5476		3
	Clostridium nexile DSM 1787		4
	Clostridium asparagiforme DSM 15981		1
	Clostridium bolteae ATCC BAA-613		2
	Clostridium hiranonis DSM 13275		2
	Clostridium hylemonae DSM 15053		2
	Clostridium ramosum DSM 1402		1
	Clostridium scindens ATCC 35704		1
	Clostridium sp. L2-50		2
	Clostridium sp. M62/1		1
	Clostridium sp. SS2/1		2

	Clostridium sporogenes ATCC 15579		1
Collinsella		4	
	Collinsella aerofaciens ATCC 25986		1
	Collinsella intestinalis DSM 13280		1
	Collinsella stercoris DSM 13279		2
Coprococcus		4	
	Coprococcus comes ATCC 27758		2
	Coprococcus eutactus ATCC 27759		2
Desulfovibrio		1	
	Desulfovibrio piger ATCC 29098		1
Dorea		1	
	Dorea longicatena DSM 13814		1
Enterobacter		1	
	Enterobacter cancerogenus ATCC 35316		1
Enterococcus		48	
	Enterococcus casseliflavus EC10		10
	Enterococcus casseliflavus EC20		21
	Enterococcus faecalis D6		5
	Enterococcus faecalis TX1322		2
	Enterococcus faecium Com15		5
	Enterococcus faecium TX1330		4
	Enterococcus faecalis TX0104		1
Escherichia		1	
	Escherichia coli ED1a		1
Eubacterium		5	
	Eubacterium dolichum DSM 3991		2
	Eubacterium biforme DSM 3989		1
	Eubacterium hallii DSM 3353		1
	Eubacterium ventriosum ATCC 27560		1
Faecalibacterium		3	
	Faecalibacterium prausnitzii A2-165		3
Fusobacterium		3	
	Fusobacterium mortiferum ATCC 9817		2
	Fusobacterium sp. 7_1		1
Helicobacter		1	
	Helicobacter bilis ATCC 43879		1
Holdemania		4	
	Holdemania filiformis DSM 12042		4
Klebsiella		1	
	Klebsiella pneumoniae 342		1
Lactobacillus		94	
	Lactobacillus acidophilus NCFM		2
	Lactobacillus antri DSM 16041		2
	Lactobacillus brevis ATCC 367		4
	Lactobacillus buchneri ATCC 11577		2
	Lactobacillus casei BL23		3
	Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365		2
	Lactobacillus fermentum ATCC 14931		3
	Lactobacillus gasseri ATCC 33323		2
	Lactobacillus helveticus DPC 4571		3
	Lactobacillus helveticus DSM 20075		4
	Lactobacillus hilgardii ATCC 8290		3
	Lactobacillus johnsonii NCC 533		3

	Lactobacillus paracasei subsp. paracasei ATCC 25302		2
	Lactobacillus plantarum subsp. plantarum ATCC 14917		6
	Lactobacillus rhamnosus LMS2-1		8
	Lactobacillus ruminis ATCC 25644		2
	Lactobacillus sakei subsp. sakei 23K		6
	Lactobacillus salivarius ATCC 11741		3
	Lactobacillus salivarius UCC118		13
	Lactobacillus ultunensis DSM 16047		3
	Lactobacillus brevis subsp. gravesensis ATCC 27305		9
	Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842		1
	Lactobacillus fermentum IFO 3956		1
	Lactobacillus paracasei subsp. paracasei 8700:2		1
	Lactobacillus plantarum WCFS1		1
	Lactobacillus reuteri CF48-3A		1
	Lactobacillus reuteri JCM 1112		1
	Lactobacillus reuteri MM2-3		1
	Lactobacillus reuteri MM4-1A		1
	Lactobacillus reuteri SD2112		1
Leuconostoc		1	
	Leuconostoc mesenteroides subsp. cremoris ATCC 19254		1
Listeria		6	
	Listeria grayi DSM 20601		6
Oxalobacter		1	
	Oxalobacter formigenes HOxBLS		1
Parabacteroides		1	
	Parabacteroides johnsonii DSM 18315		1
Parvimonas		2	
	Parvimonas micra ATCC 33270		2
Prevotella		2	
	Prevotella copri DSM 18205		2
Proteus		2	
	Proteus penneri ATCC 35198		2
Providencia		5	
	Providencia stuartii ATCC 25827		2
	Providencia alcalifaciens DSM 30120		1
	Providencia rettgeri DSM 1131		1
	Providencia rustigianii DSM 4541		1
Roseburia		4	
	Roseburia ignavus ATCC 29149		1
	Roseburia intestinalis L1-82		1
	Roseburia inulinivorans DSM 16841		2
Ruminococcus		8	
	Ruminococcus lactaris ATCC 29176		1
	Ruminococcus obeum ATCC 29174		2
	Ruminococcus sp. 5_1_39BFAA		2
	Ruminococcus torques ATCC 27756		3
Streptococcus		6	
	Streptococcus thermophilus LMD-9		4

	Streptococcus infantarius subsp. infantarius ATCC BAA-102		2
Subdoligranulum		1	
	Subdoligranulum variabile DSM 15176		1
Weissella		3	
	Weissella paramesenteroides ATCC 33313		3
	Total	**273**	**273**

Appendix 3 Genome information of *Aerococcus viridans* ATCC 11563 (adapted from NCBI Genome database on 14 May 2011.)

A. Taxonomy

Super-kingdom: Bacteria
Phylum: Firmicutes
Class: Bacilli
Order: Lactobacillales
Family: Aerococcaceae
Genus: *Aerococcus*
Species: *Aerococcus viridans*
Strain: *Aerococcus viridans ATCC 11563*

B. Genome information

Genome Information	Features	Review Information
RefSeq: NZ_ADNT00000000	Genes: 1978	Publications: None
GenBank:ADNT00000000	Protein coding: 1929	Refseq Status: WGS
Length: 2,005,853 nt	Structural RNAs: 49	Seq.Status: Draft
GC Content: 39%	Pseudo genes: None	Sequencing center: Baylor College of Medicine
% Coding: 85%	Others: 16	Completed: 2010/04/29
Topology: other	Contigs: 150	
Molecule: DNA		

References

1. Adhya D, Basu A (2010). Epigenetic modulation of host: new insights into immune evasion by viruses. J Biosci 35: 647-63.

2. Akhmedov AT, Lopez BS (2000). Human 100-kDa homologous DNA-pairing protein is the splicing factor PSF and promotes DNA strand invasion. Nucleic Acids Res 28: 3022-30.

3. Akiba J, Umemura T, Alter HJ, Kojiro M, Tabor E (2005). SEN virus: epidemiology and characteristics of a transfusion-transmitted virus. Transfusion 45: 1084-8.

4. Alkhatib G (2009). The biology of CCR5 and CXCR4. Curr Opin HIV AIDS 4: 96-103.

5. Allo M, Buggiano V, Fededa JP, Petrillo E, Schor I, de la Mata M et al (2009). Control of alternative splicing through siRNA-mediated transcriptional gene silencing. Nat Struct Mol Biol 16: 717-24.

6. Alverdy JC, Chang EB (2008). The re-emerging role of the intestinal microflora in critical illness and inflammation: why the gut hypothesis of sepsis syndrome will not go away. J Leukoc Biol 83: 461-6.

7. Andreoli E, Maggi F, Pistello M, Meschi S, Vatteroni M, Nelli LC et al (2006). Small Anellovirus in hepatitis C patients and healthy controls. Emerg Infect Dis 12: 1175-6.

8. Angly FE, Willner D, Prieto-Davo A, Edwards RA, Schmieder R, Vega-Thurber R et al (2009). The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comput Biol 5: e1000593.

9. Antinori A, Larussa D, Cingolani A, Lorenzini P, Bossolasco S, Finazzi MG et al (2004). Prevalence, associated factors, and prognostic determinants of AIDS-related toxoplasmic encephalitis in the era of advanced highly active antiretroviral therapy. Clin Infect Dis 39: 1681-91.

10. Antiretroviral Therapy Cohort Collaboration A (2010). Causes of death in HIV-1-infected patients treated with antiretroviral therapy, 1996-2006: collaborative analysis of 13 HIV cohort studies. Clin Infect Dis 50: 1387-96.

11. Apisarnthanarak A, Mundy LM (2005). Etiology of community-acquired pneumonia. Clin Chest Med 26: 47-55.

12. Arora VK, Molina RP, Foster JL, Blakemore JL, Chernoff J, Fredericksen BL et al (2000). Lentivirus Nef specifically activates Pak2. J Virol 74: 11081-7.

13. AVERT. (2010). Vol. 2011.

14. AVERT (2010). World estimates of the HIV & AIDS epidemics at the end of 2008.

15. Ayouba A, Souquieres S, Njinku B, Martin PM, Muller-Trutwin MC, Roques P et al (2000). HIV-1 group N among HIV-1-seropositive individuals in Cameroon. AIDS 14: 2623-5.

16. Barre-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J et al (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220: 868-71.

17. Bednarik DP, Cook JA, Pitha PM (1990). Inactivation of the HIV LTR by DNA CpG methylation: evidence for a role in latency. EMBO J 9: 1157-64.

18. Bell JT, Spector TD (2011). A twin approach to unraveling epigenetics. Trends Genet 27: 116-25.

19. Belshaw R, Dawson AL, Woolven-Allen J, Redding J, Burt A, Tristem M (2005). Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. J Virol 79: 12507-14.

20. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A et al (2004). Long-term reinfection of the human genome by endogenous retroviruses. Proc Natl Acad Sci U S A 101: 4894-9.

21. Bianchi ME (2007). DAMPs, PAMPs and alarmins: all we need to know about danger. J Leukoc Biol 81: 1-5.

22. Blazkova J, Trejbalova K, Gondois-Rey F, Halfon P, Philibert P, Guiguen A et al (2009). CpG methylation controls reactivation of HIV from latency. PLoS Pathog 5: e1000554.

23. Blazsek A, Sillo P, Ishii N, Gergely P, Jr., Poor G, Preisz K et al (2008). Searching for foreign antigens as possible triggering factors of autoimmunity: Torque Teno virus DNA prevalence is elevated in sera of patients with bullous pemphigoid. Exp Dermatol 17: 446-54.

24. Bone RC (1992). Toward an epidemiology and natural history of SIRS (systemic inflammatory response syndrome). JAMA 268: 3452-5.

25. Brass AL, Dykxhoorn DM, Benita Y, Yan N, Engelman A, Xavier RJ et al (2008). Identification of host proteins required for HIV infection through a functional genomic screen. Science 319: 921-6.

26. Brechot C, Kremsdorf D, Soussan P, Pineau P, Dejean A, Paterlini-Brechot P et al (2010). Hepatitis B virus (HBV)-related hepatocellular carcinoma (HCC): molecular mechanisms and novel paradigms. Pathol Biol (Paris) 58: 278-87.

27. Brenchley JM, Price DA, Douek DC (2006). HIV disease: fallout from a mucosal catastrophe? Nat Immunol 7: 235-9.

28. Brenchley JM, Schacker TW, Ruff LE, Price DA, Taylor JH, Beilman GJ et al (2004). CD4+ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. J Exp Med 200: 749-59.

29. Brew BJ (2001). AIDS Dementia Complex. HIV Neurology. Oxford University Press: New York.

30. Brodie EL, DeSantis TZ, Parker JP, Zubietta IX, Piceno YM, Andersen GL (2007). Urban aerosols harbor diverse and dynamic bacterial populations. Proc Natl Acad Sci U S A 104: 299-304.

31. Carissimi C, Saieva L, Baccon J, Chiarella P, Maiolica A, Sawyer A et al (2006). Gemin8 is a novel component of the survival motor neuron complex and functions in small nuclear ribonucleoprotein assembly. J Biol Chem 281: 8126-34.

32. Carr IM, Valleley EM, Cordery SF, Markham AF, Bonthron DT (2007). Sequence analysis and editing for bisulphite genomic sequencing projects. Nucleic Acids Res 35: e79.

33. Cauchi RJ (2010). SMN and Gemins: 'we are family' ... or are we?: insights into the partnership between Gemins and the spinal muscular atrophy disease protein SMN. Bioessays 32: 1077-89.

34. Chan DC, Kim PS (1998). HIV entry and its inhibition. Cell 93: 681-4.

35. Chan HL, Jia J (2011). Chronic hepatitis B in Asia-new insights from the past decade. J Gastroenterol Hepatol 26 Suppl 1: 131-7.

36. Chang YS, Liu WJ, Lee CC, Chou TL, Lee YT, Wu TS et al (2010). A 3D model of the membrane protein complex formed by the white spot syndrome virus structural proteins. PLoS One 5: e10718.

37. Chan EC, Yagi S, Kelly KR, Mendoza SP, Maninger N, Rosenthal A et al (2011). Cross-species transmission of a novel adenovirus associated with a fulminant pneumonia outbreak in a new world monkey colony. PLoS Pathog 7: e1002155.

38. Cheung HH, Lee TL, Davis AJ, Taft DH, Rennert OM, Chan WY (2010). Genome-wide DNA methylation profiling reveals novel epigenetically regulated genes and non-coding RNAs in human testicular cancer. Br J Cancer 102: 419-27.

39. Cheung HH, Lee TL, Rennert OM, Chan WY (2009). DNA methylation of cancer genome. Birth Defects Res C Embryo Today 87: 335-50.

40. Chiu RW, Cantor CR, Lo YM (2009). Non-invasive prenatal diagnosis by single molecule counting technologies. Trends Genet 25: 324-31.

41. Chou CC, Lee TT, Chen CH, Hsiao HY, Lin YL, Ho MS et al (2006). Design of microarray probes for virus identification and detection of emerging viruses at the genus level. BMC Bioinformatics 7: 232.

42. Chung JY, Han TH, Koo JW, Kim SW, Seo JK, Hwang ES (2007). Small anellovirus infections in Korean children. Emerg Infect Dis 13: 791-3.

43. Cocchi F, DeVico AL, Garzino-Demo A, Cara A, Gallo RC, Lusso P (1996). The V3 domain of the HIV-1 gp120 envelope glycoprotein is critical for chemokine-mediated blockade of infection. Nat Med 2: 1244-7.

44. Coenye T, Vandamme P, LiPuma JJ (2002). Infection by Ralstonia species in cystic fibrosis patients: identification of R. pickettii and R. mannitolilytica by polymerase chain reaction. Emerg Infect Dis 8: 692-6.

45. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ et al (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37: D141-5.

46. Colombo FA, Vidal JE, Penalva de Oliveira AC, Hernandez AV, Bonasser-Filho F, Nogueira RS et al (2005). Diagnosis of cerebral toxoplasmosis in AIDS patients in Brazil: importance of molecular and immunological methods using peripheral blood samples. J Clin Microbiol 43: 5044-7.

47. Comas I, Gagneux S (2009). The past and future of tuberculosis research. PLoS Pathog 5: e1000600.

48. Contreras X, Barboric M, Lenasi T, Peterlin BM (2007). HMBA releases P-TEFb from HEXIM1 and 7SK snRNA via PI3K/Akt and activates HIV transcription. PLoS Pathog 3: 1459-69.

49. Control CfD (1981).

50. Costa Mdo C, Teixeira-Castro A, Constante M, Magalhaes M, Magalhaes P, Cerqueira J et al (2006). Exclusion of mutations in the PRNP, JPH3, TBP, ATN1, CREBBP, POU3F2 and FTL genes as a cause of disease in Portuguese patients with a Huntington-like phenotype. J Hum Genet 51: 645-51.

51. Cujec TP, Okamoto H, Fujinaga K, Meyer J, Chamberlin H, Morgan DO et al (1997). The HIV transactivator TAT binds to the CDK-activating kinase and

activates the phosphorylation of the carboxy-terminal domain of RNA polymerase II. Genes Dev 11: 2645-57.

52. de la Fuente C, Santiago F, Deng L, Eadie C, Zilberman I, Kehn K et al (2002). Gene expression profile of HIV-1 Tat expressing cells: a close interplay between proliferative and differentiation signals. BMC Biochem 3: 14.

53. de Madaria E, Martinez J, Lozano B, Sempere L, Benlloch S, Such J et al (2005). Detection and identification of bacterial DNA in serum from patients with acute pancreatitis. Gut 54: 1293-7.

54. Dethlefsen L, Huse S, Sogin ML, Relman DA (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biol 6: e280.

55. Dewannieux M, Harper F, Richaud A, Letzelter C, Ribet D, Pierron G et al (2006). Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Res 16: 1548-56.

56. Domingue GJ, Schlegel JU (1977). Novel bacterial structures in human blood: cultural isolation. Infect Immun 15: 621-7.

57. Dong Q, Brulc JM, Iovieno A, Bates B, Garoutte A, Miller D et al (2011). Diversity of bacteria at healthy human conjunctiva. Invest Ophthalmol Vis Sci.

58. Douek D (2007). HIV disease progression: immune activation, microbes, and a leaky gut. Top HIV Med 15: 114-7.

59. Dowling D, Nasr-Esfahani S, Tan CH, O'Brien K, Howard JL, Jans DA et al (2008). HIV-1 infection induces changes in expression of cellular splicing factors that regulate alternative viral splicing and virus production in macrophages. Retrovirology 5: 18.

60. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E et al (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol 26: 779-85.

61. Elnifro EM, Ashshi AM, Cooper RJ, Klapper PE (2000). Multiplex PCR: optimization and application in diagnostic virology. Clin Microbiol Rev 13: 559-70.

62. Escobedo-Bonilla CM, Alday-Sanz V, Wille M, Sorgeloos P, Pensaert MB, Nauwynck HJ (2008). A review on the morphology, molecular characterization, morphogenesis and pathogenesis of white spot syndrome virus. J Fish Dis 31: 1-18.

63. Fang JY, Mikovits JA, Bagni R, Petrow-Sadowski CL, Ruscetti FW (2001). Infection of lymphoid cells by integration-defective human immunodeficiency virus type 1 increases de novo methylation. J Virol 75: 9753-61.

64. Feinberg AP, Vogelstein B (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature 301: 89-92.

65. Feng W, Gubitz AK, Wan L, Battle DJ, Dostie J, Golembe TJ et al (2005). Gemins modulate the expression and activity of the SMN complex. Hum Mol Genet 14: 1605-11.

66. Ferrari R, Berk AJ, Kurdistani SK (2009). Viral manipulation of the host epigenome for oncogenic transformation. Nat Rev Genet 10: 290-4.

67. Ferri E, Novati S, Casiraghi M, Sambri V, Genco F, Gulminetti R et al (2010). Plasma levels of bacterial DNA in HIV infection: the limits of quantitative polymerase chain reaction. J Infect Dis 202: 176-7; author reply 178.

68. Fine J, Frank ED, Ravin HA, Rutenberg SH, Schweinburg FB (1959). The bacterial factor in traumatic shock. N Engl J Med 260: 214-20.

69. Fischer U, Liu Q, Dreyfuss G (1997). The SMN-SIP1 complex has an essential role in spliceosomal snRNP biogenesis. Cell 90: 1023-9.

70. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009). Human genetic variation and its contribution to complex traits. Nat Rev Genet 10: 241-51.

71. Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X et al (2008). ToxoDB: an integrated Toxoplasma gondii database resource. Nucleic Acids Res 36: D553-6.

72. Gajria B, Bahl A, Brestelli J, Dommer J, Fischer S, Gao X et al (2008). ToxoDB: an integrated Toxoplasma gondii database resource. Nucleic Acids Res 36: D553-6.

73. Gallo RC, Salahuddin SZ, Popovic M, Shearer GM, Kaplan M, Haynes BF et al (1984). Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. Science 224: 500-3.

74. Garrison KE, Jones RB, Meiklejohn DA, Anwar N, Ndhlovu LC, Chapman JM et al (2007). T cell responses to human endogenous retroviruses in HIV-1 infection. PLoS Pathog 3: e165.

75. Gaynor AM, Nissen MD, Whiley DM, Mackay IM, Lambert SB, Wu G et al (2007). Identification of a novel polyomavirus from patients with acute respiratory tract infections. PLoS Pathog 3: e64.

76. Gergely P, Jr., Pullmann R, Stancato C, Otvos L, Jr., Koncz A, Blazsek A et al (2005). Increased prevalence of transfusion-transmitted virus and cross-reactivity with immunodominant epitopes of the HRES-1/p28 endogenous retroviral autoantigen in patients with systemic lupus erythematosus. Clin Immunol 116: 124-34.

77. Gerner P, Oettinger R, Gerner W, Falbrede J, Wirth S (2000). Mother-to-infant transmission of TT virus: prevalence, extent and mechanism of vertical transmission. Pediatr Infect Dis J 19: 1074-7.

78. Geuking MB, Weber J, Dewannieux M, Gorelik E, Heidmann T, Hengartner H et al (2009). Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. Science 323: 393-6.

79. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS et al (2006). Metagenomic analysis of the human distal gut microbiome. Science 312: 1355-9.

80. Golembe TJ, Yong J, Battle DJ, Feng W, Wan L, Dreyfuss G (2005). Lymphotropic Herpesvirus saimiri uses the SMN complex to assemble Sm cores on its small RNAs. Mol Cell Biol 25: 602-11.

81. Goll MG, Bestor TH (2005). Eukaryotic cytosine methyltransferases. Annu Rev Biochem 74: 481-514.

82. Gophna U SK, Gophna S, Doolittle WF, Veldhuyzen van Zanten SJ (2006). Differences between tissue-associated intestinal microfloras of patients with Crohn's disease and ulcerative colitis. J Clin Microbiol (11):. Epub 2006 Sep 20.: 4136-41.

83. Grez M, Dietrich U, Balfe P, von Briesen H, Maniar JK, Mahambre G et al (1994). Genetic analysis of human immunodeficiency virus type 1 and 2 (HIV-1 and HIV-2) mixed infections in India reveals a recent spread of HIV-1 and HIV-2 from a single ancestor for each of these viruses. J Virol 68: 2161-8.

84. Griffin JS, Plummer JD, Long SC (2008). Torque teno virus: an improved indicator for viral pathogens in drinking waters. Virol J 5: 112.

85. Gubitz AK, Feng W, Dreyfuss G (2004). The SMN complex. Exp Cell Res 296: 51-6.

86. Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41: 95-98.

87. Halstead DC, Gray GC, Meyer KS, Stanciu SR, Gorospe WC (2010). Recombinant Adenovirus (AdV) Type 3 and Type 14 Isolated From a Fatal Case of Pneumonia. Rev Med Microbiol 21: 28-30.

88. Hamamoto S, Nishitsuji H, Amagasa T, Kannagi M, Masuda T (2006). Identification of a novel human immunodeficiency virus type 1 integrase interactor, Gemin2, that facilitates efficient viral cDNA synthesis in vivo. J Virol 80: 5670-7.

89. Hanamura A, Caceres JF, Mayeda A, Franza BR, Jr., Krainer AR (1998). Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors. RNA 4: 430-44.

90. Handelsman J (2004). Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68: 669-85.

91. Haynes MR, F. (ed.) (2010) The Human Virome.

92. He J, Choe S, Walker R, Di Marzio P, Morgan DO, Landau NR (1995). Human immunodeficiency virus type 1 viral protein R (Vpr) arrests cells in the G2 phase of the cell cycle by inhibiting p34cdc2 activity. J Virol 69: 6705-11.

93. Heller R, Jaulhac B, Charles P, De Briel D, Vincent V, Bohner C et al (1996). Identification of Mycobacterium shimoidei in a tuberculosis-like cavity by 16S ribosomal DNA direct sequencing. Eur J Clin Microbiol Infect Dis 15: 172-5.

94. Henderson IR, Jacobsen SE (2007). Epigenetic inheritance in plants. Nature 447: 418-24.

95. Henze K, Martin W (2003). Evolutionary biology: essence of mitochondria. Nature 426: 127-8.

96. Herschhorn A, Hizi A (2010). Retroviral reverse transcriptases. Cell Mol Life Sci 67: 2717-47.

97. HMP HMPDACC (2010). Reference genomes of the Human Microbiome Project. .

98. Hsieh CL (2005). The de novo methylation activity of Dnmt3a is distinctly different than that of Dnmt1. BMC Biochem 6: 6.

99. Huang da W, Sherman BT, Lempicki RA (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.

100. Huang HH, Zhang YY, Xiu QY, Zhou X, Huang SG, Lu Q et al (2006). Community-acquired pneumonia in Shanghai, China: microbial etiology and implications for empirical therapy in a prospective study of 389 patients. Eur J Clin Microbiol Infect Dis 25: 369-74.

101. Huang J, Wang F, Argyris E, Chen K, Liang Z, Tian H et al (2007). Cellular microRNAs contribute to HIV-1 latency in resting primary CD4+ T lymphocytes. Nat Med 13: 1241-7.

102.Ishida T, Hamano A, Koiwa T, Watanabe T (2006). 5' long terminal repeat (LTR)-selective methylation of latently infected HIV-1 provirus that is demethylated by reactivation signals. Retrovirology 3: 69.

103.Jaenisch R, Bird A (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 33 Suppl: 245-54.

104.Jarnaess E, Stokka AJ, Kvissel AK, Skalhegg BS, Torgersen KM, Scott JD et al (2009). Splicing factor arginine/serine-rich 17A (SFRS17A) is an A-kinase anchoring protein that targets protein kinase A to splicing factor compartments. J Biol Chem 284: 35154-64.

105.Jayadev S, Garden GA (2009). Host and viral factors influencing the pathogenesis of HIV-associated neurocognitive disorders. J Neuroimmune Pharmacol 4: 175-89.

106.Jiang W, Lederman MM, Hunt P, Sieg SF, Haley K, Rodriguez B et al (2009). Plasma levels of bacterial DNA correlate with immune activation and the magnitude of immune restoration in persons with antiretroviral-treated HIV infection. J Infect Dis 199: 1177-85.

107.Johnson RP (2008). How HIV guts the immune system. N Engl J Med 358: 2287-9.

108.Jones PA, Baylin SB (2002). The fundamental role of epigenetic events in cancer. Nat Rev Genet 3: 415-28.

109.Kaminsky ZA, Tang T, Wang SC, Ptak C, Oh GH, Wong AH et al (2009). DNA methylation profiles in monozygotic and dizygotic twins. Nat Genet 41: 240-5.

110.Kanki PJ, Travers KU, S MB, Hsieh CC, Marlink RG, Gueye NA et al (1994). Slower heterosexual spread of HIV-2 than HIV-1. Lancet 343: 943-6.

111.Kao JH, Chen W, Chen PJ, Lai MY, Chen DS (2002). Prevalence and implication of a newly identified infectious agent (SEN virus) in Taiwan. J Infect Dis 185: 389-92.

112.Kapogiannis BG, Soe MM, Nesheim SR, Sullivan KM, Abrams E, Farley J et al (2008). Trends in bacteremia in the pre- and post-highly active antiretroviral therapy era among HIV-infected children in the US Perinatal AIDS Collaborative Transmission Study (1986-2004). Pediatrics 121: e1229-39.

113.Khan A, Bohme U, Kelly KA, Adlem E, Brooks K, Simmonds M et al (2006). Common inheritance of chromosome Ia associated with clonal expansion of Toxoplasma gondii. Genome Res 16: 1119-25.

114.Kim S, Ikeuchi K, Byrn R, Groopman J, Baltimore D (1989). Lack of a negative influence on viral growth by the nef gene of human immunodeficiency virus type 1. Proc Natl Acad Sci U S A 86: 9544-8.

115.Kino T, Gragerov A, Slobodskaya O, Tsopanomichalou M, Chrousos GP, Pavlakis GN (2002). Human immunodeficiency virus type 1 (HIV-1) accessory protein Vpr induces transcription of the HIV-1 and glucocorticoid-responsive promoters by binding directly to p300/CBP coactivators. J Virol 76: 9724-34.

116.Klipper-Aurbach Y, Wasserman M, Braunspiegel-Weintrob N, Borstein D, Peleg S, Assa S et al (1995). Mathematical formulae for the prediction of the residual beta cell function during the first two years of disease in children and adolescents with insulin-dependent diabetes mellitus. Med Hypotheses 45: 486-90.

117.Klotman ME, Kim S, Buchbinder A, DeRossi A, Baltimore D, Wong-Staal F (1991). Kinetics of expression of multiply spliced RNA in early human immunodeficiency virus type 1 infection of lymphocytes and monocytes. Proc Natl Acad Sci U S A 88: 5011-5.

118.Kolb SJ, Battle DJ, Dreyfuss G (2007). Molecular functions of the SMN complex. J Child Neurol 22: 990-4.

119.Korbie DJ, Mattick JS (2008). Touchdown PCR for increased specificity and sensitivity in PCR amplification. Nat Protoc 3: 1452-6.

120.Kosakovsky Pond S, Wadhawan S, Chiaromonte F, Ananda G, Chung WY, Taylor J et al (2009). Windshield splatter analysis with the Galaxy metagenomic pipeline. Genome Res 19: 2144-53.

121.Koup RA, Roederer M, Lamoreaux L, Fischer J, Novik L, Nason MC et al (2010). Priming immunization with DNA augments immunogenicity of recombinant adenoviral vectors for both HIV-1 specific antibody and T-cell responses. PLoS One 5: e9015.

122.Krekulova L, Rehak V, Killoran P, Madrigal N, Riley LW (2001). Genotypic distribution of TT virus (TTV) in a Czech population: evidence for sexual transmission of the virus. J Clin Virol 23: 31-41.

123.Lahti AL, Manninen A, Saksela K (2003). Regulation of T cell activation by HIV-1 accessory proteins: Vpr acts via distinct mechanisms to cooperate with Nef in NFAT-directed gene expression and to promote transactivation by CREB. Virology 310: 190-6.

124.Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J et al (2001). Initial sequencing and analysis of the human genome. Nature 409: 860-921.

125.Lassen K, Han Y, Zhou Y, Siliciano J, Siliciano RF (2004). The multifactorial nature of HIV-1 latency. Trends Mol Med 10: 525-31.

126.Lau CK, Diem MD, Dreyfuss G, Van Duyne GD (2003). Structure of the Y14-Magoh core of the exon junction complex. Curr Biol 13: 933-41.

127.Lau KA, Wang B, Saksena NK (2007). Emerging trends of HIV epidemiology in Asia. AIDS Rev 9: 218-29.

128.Le Guyader S, Maier J, Jesuthasan S (2005). Esrom, an ortholog of PAM (protein associated with c-myc), regulates pteridine synthesis in the zebrafish. Dev Biol 277: 378-86.

129.Lederman M (2010). Why suppression of HIV replication does not always make everything better. . Journal of the International AIDS Society 13 (Suppl 4):K1. .

130.Lee YN, Bieniasz PD (2007). Reconstitution of an infectious human endogenous retrovirus. PLoS Pathog 3: e10.

131.Lefebvre S, Burglen L, Reboullet S, Clermont O, Burlet P, Viollet L et al (1995). Identification and characterization of a spinal muscular atrophy-determining gene. Cell 80: 155-65.

132.Leite A, Vinhas-Da-Silva A, Felicio L, Vilarinho AC, Ferreira G (2011). Aerococcus viridans urinary tract infection in a pediatric patient with secondary pseudohypoaldosteronism. Rev Argent Microbiol 42: 269-70.

133.Leland DS, Ginocchio CC (2007). Role of cell culture for virus detection in the age of technology. Clin Microbiol Rev 20: 49-78.

134.Leung RK, Fong FN, Au TC, Lau IF, Chan PK, Zhang C et al (2010). An unusual cluster of HIV-1 B/F recombinants in an Asian population. Int J Infect Dis 14 Suppl 3: e294-8.

135.Leung TW, Mak D, Wong KH, Wang Y, Song YH, Tsang DN et al (2008). Molecular epidemiology demonstrated three emerging clusters of human immunodeficiency virus type 1 subtype B infection in Hong Kong. AIDS Res Hum Retroviruses 24: 903-10.

136.Li E (2002). Chromatin modification and epigenetic reprogramming in mammalian development. Nat Rev Genet 3: 662-73.

137.Li H, Rauch T, Chen ZX, Szabo PE, Riggs AD, Pfeifer GP (2006). The histone methyltransferase SETDB1 and the DNA methyltransferase DNMT3A interact directly and localize to promoters silenced in cancer cells. J Biol Chem 281: 19489-500.

138.Liang G, Chan MF, Tomigahara Y, Tsai YC, Gonzales FA, Li E et al (2002). Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. Mol Cell Biol 22: 480-91.

139.Lippman Z, Gendrel AV, Colot V, Martienssen R (2005). Profiling DNA methylation patterns using genomic tiling microarrays. Nat Methods 2: 219-24.

140.Lister R, Ecker JR (2009). Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res 19: 959-66.

141.Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH et al (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133: 523-36.

142.Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J et al (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462: 315-22.

143.Lloyd-Puryear M, Wallace D, Baldwin T, Hollis DG (1991). Meningitis caused by Psychrobacter immobilis in an infant. J Clin Microbiol 29: 2041-2.

144.Loussert-Ajaka I, Chaix ML, Korber B, Letourneur F, Gomas E, Allen E et al (1995). Variability of human immunodeficiency virus type 1 group O strains isolated from Cameroonian patients living in France. J Virol 69: 5640-9.

145.Luco RF, Allo M, Schor IE, Kornblihtt AR, Misteli T (2011). Epigenetics in alternative pre-mRNA splicing. Cell 144: 16-26.

146.Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T (2010). Regulation of alternative splicing by histone modifications. Science 327: 996-1000.

147.Lujambio A, Calin GA, Villanueva A, Ropero S, Sanchez-Cespedes M, Blanco D et al (2008). A microRNA DNA methylation signature for human cancer metastasis. Proc Natl Acad Sci U S A 105: 13556-61.

148.Lujambio A, Portela A, Liz J, Melo SA, Rossi S, Spizzo R et al (2010). CpG island hypermethylation-associated silencing of non-coding RNAs transcribed from ultraconserved regions in human cancer. Oncogene 29: 6390-401.

149.Maayan H, Cohen-Poradosu R, Halperin E, Rudensky B, Schlesinger Y, Yinnon AM et al (2004). Infective endocarditis due to Moraxella lacunata: report of 4 patients and review of published cases of Moraxella endocarditis. Scand J Infect Dis 36: 878-81.

150.Maclachlan NJ, Drew CP, Darpel KE, Worwa G (2009). The pathology and pathogenesis of bluetongue. J Comp Pathol 141: 1-16.

151.Maggi F, Pifferi M, Fornai C, Andreoli E, Tempestini E, Vatteroni M et al (2003). TT virus in the nasal secretions of children with acute respiratory diseases: relations to viremia and disease severity. J Virol 77: 2418-25.

152.Maldarelli F, Xiang C, Chamoun G, Zeichner SL (1998). The expression of the essential nuclear splicing factor SC35 is altered by human immunodeficiency virus infection. Virus Res 53: 39-51.

153.Mameli G, Astone V, Arru G, Marconi S, Lovato L, Serra C et al (2007). Brains and peripheral blood mononuclear cells of multiple sclerosis (MS) patients hyperexpress MS-associated retrovirus/HERV-W endogenous retrovirus, but not Human herpesvirus 6. J Gen Virol 88: 264-74.

154.Manfredi AA, Rovere-Querini P (2010). The mitochondrion--a Trojan horse that kicks off inflammation? N Engl J Med 362: 2132-4.

155.Manfredi R, Nanetti A, Valentini R, Chiodo F (2001). Acinetobacter infections in patients with human immunodeficiency virus infection: microbiological and clinical epidemiology. Chemotherapy 47: 19-28.

156.Manickam N, Pareek S, Kaur I, Singh NK, Mayilraj S (2011). Nitratireductor lucknowense sp. nov., a novel bacterium isolated from a pesticide contaminated soil. Antonie Van Leeuwenhoek.

157.Marchand V, Santerre M, Aigueperse C, Fouillen L, Saliou JM, Van Dorsselaer A et al (2010). Identification of protein partners of the human immunodeficiency virus 1 tat/rev exon 3 leads to the discovery of a new HIV-1 splicing regulator, protein hnRNP K. RNA Biol 8.

158.Martienssen RA, Doerge RW, Colot V (2005). Epigenomic mapping in Arabidopsis using tiling microarrays. Chromosome Res 13: 299-308.

159.Martin A, Bodola F, Sangar DV, Goettge K, Popov V, Rijnbrand R et al (2003). Chronic hepatitis associated with GB virus B persistence in a tamarin after intrahepatic inoculation of synthetic viral RNA. Proc Natl Acad Sci U S A 100: 9962-7.

160.Masucci MG, Contreras-Salazar B, Ragnar E, Falk K, Minarovits J, Ernberg I et al (1989). 5-Azacytidine up regulates the expression of Epstein-Barr virus nuclear antigen 2 (EBNA-2) through EBNA-6 and latent membrane protein in the Burkitt's lymphoma line rael. J Virol 63: 3135-41.

161.Matthew Haynes FR (2010). The Human Virome. In: Nelson KE (ed). Metagenomics of the Human Body.

162.McCutchan FE, Hoelscher M, Tovanabutra S, Piyasirisilp S, Sanders-Buell E, Ramos G et al (2005). In-depth analysis of a heterosexually acquired human immunodeficiency virus type 1 superinfection: evolution, temporal fluctuation, and intercompartment dynamics from the seronegative window period through 30 months postinfection. J Virol 79: 11693-704.

163.Mesquita RT, Vidal JE, Pereira-Chioccola VL (2010). Molecular diagnosis of cerebral toxoplasmosis: comparing markers that determine Toxoplasma gondii by PCR in peripheral blood from HIV-infected patients. Braz J Infect Dis 14: 346-50.

164.Mikovits JA, Young HA, Vertino P, Issa JP, Pitha PM, Turcoski-Corrales S et al (1998). Infection with human immunodeficiency virus type 1 upregulates DNA methyltransferase, resulting in de novo methylation of the gamma interferon (IFN-gamma) promoter and subsequent downregulation of IFN-gamma production. Mol Cell Biol 18: 5166-77.

165.Ministry of Health P, Joint United Nations Programme on HIV/AIDS, World Health Organization, Beijing, China. (2010). Beijing, China. .

166.Monini P, Sgadari C, Toschi E, Barillari G, Ensoli B (2004). Antitumour effects of antiretroviral therapy. Nat Rev Cancer 4: 861-75.

167.Montagnier L (1985). Lymphadenopathy-associated virus: from molecular biology to pathogenicity. Ann Intern Med 103: 689-93.

168.Montero LM, Filipski J, Gil P, Capel J, Martinez-Zapater JM, Salinas J (1992). The distribution of 5-methylcytosine in the nuclear genome of plants. Nucleic Acids Res 20: 3207-10.

169.Montgomery DL, Prather KJ (2006). Design of plasmid DNA constructs for vaccines. Methods Mol Med 127: 11-22.

170.Mootsikapun P (2007). Bacteremia in adult patients with acquired immunodeficiency syndrome in the northeast of Thailand. Int J Infect Dis 11: 226-31.

171.Morell V (1993). Huntington's gene finally found. Science 260: 28-30.

172.Mowat AM (2003). Anatomical basis of tolerance and immunity to intestinal antigens. Nat Rev Immunol 3: 331-41.

173.Muniz L, Egloff S, Ughy B, Jady BE, Kiss T (2010). Controlling cellular P-TEFb activity by the HIV-1 transcriptional transactivator Tat. PLoS Pathog 6: e1001152.

174.Nadal I, Donat E, Ribes-Koninckx C, Calabuig M, Sanz Y (2007). Imbalance in the composition of the duodenal microbiota of children with coeliac disease. J Med Microbiol 56: 1669-74.

175.Nagy Z, Chandler M (2004). Regulation of transposition in bacteria. Res Microbiol 155: 387-98.

176.Nakielny S, Dreyfuss G (1999). Transport of proteins and RNAs in and out of the nucleus. Cell 99: 677-90.

177.Nam JG, Kim GJ, Baek JY, Suh SD, Kee MK, Lee JS et al (2006). Molecular investigation of human immunodeficiency virus type 2 subtype a cases in South Korea. J Clin Microbiol 44: 1543-6.

178.Navia BA, Cho ES, Petito CK, Price RW (1986). The AIDS dementia complex: II. Neuropathology. Ann Neurol 19: 525-35.

179.NCBI. (2011).

180.NCBI. (2011).

181.Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR et al (2010). A catalog of reference genomes from the human microbiome. Science 328: 994-9.

182.Nelsons KE (ed.) (2010) Megagenomics of the Human Boday. Springer Science and Business Media, 1-14pp.

183.Newmark PA, Boswell RE (1994). The mago nashi locus encodes an essential product required for germ plasm assembly in Drosophila. Development 120: 1303-13.

184.Ninomiya M, Nishizawa T, Takahashi M, Lorenzo FR, Shimosegawa T, Okamoto H (2007). Identification and genomic characterization of a novel human torque teno virus of 3.2 kb. J Gen Virol 88: 1939-44.

185.Nishizawa T, Okamoto H, Konishi K, Yoshizawa H, Miyakawa Y, Mayumi M (1997). A novel DNA virus (TTV) associated with elevated transaminase levels in posttransfusion hepatitis of unknown etiology. Biochem Biophys Res Commun 241: 92-7.

186.Noguchi H, Taniguchi T, Itoh T (2008). MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res 15: 387-96.

187.Okano M, Xie S, Li E (1998). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. Nat Genet 19: 219-20.

188.Okitsu CY, Hsieh CL (2007). DNA methylation dictates histone H3K4 methylation. Mol Cell Biol 27: 2746-57.

189.Palella FJ, Jr., Baker RK, Moorman AC, Chmiel JS, Wood KC, Brooks JT et al (2006). Mortality in the highly active antiretroviral therapy era: changing causes of death and disease in the HIV outpatient study. J Acquir Immune Defic Syndr 43: 27-34.

190.Pantry SN, Medveczky PG (2009). Epigenetic regulation of Kaposi's sarcoma-associated herpesvirus replication. Semin Cancer Biol 19: 153-7.

191.Park SE, Lee MJ, Yang MH, Ahn KY, Jang SI, Suh YJ et al (2007). Expression profiles and pathway analysis in HEK 293 T cells overexpressing HIV-1 Tat and nucleocapsid using cDNA microarray. J Microbiol Biotechnol 17: 154-61.

192.Patel AA, Steitz JA (2003). Splicing double: insights from the second spliceosome. Nat Rev Mol Cell Biol 4: 960-70.

193.Pellegrino FL, Schirmer M, Velasco E, de Faria LM, Santos KR, Moreira BM (2008). Ralstonia pickettii bloodstream infections at a Brazilian cancer institution. Curr Microbiol 56: 219-23.

194.Pellizzoni L, Yong J, Dreyfuss G (2002). Essential role for the SMN complex in the specificity of snRNP assembly. Science 298: 1775-9.

195.Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA et al (2009). The NIH Human Microbiome Project. Genome Res 19: 2317-23.

196.Petti CA, Polage CR, Schreckenberger P (2005). The role of 16S rRNA gene sequencing in identification of microorganisms misidentified by conventional methods. J Clin Microbiol 43: 6123-5.

197.Pfutzner A, Dietrich U, von Eichel U, von Briesen H, Brede HD, Maniar JK et al (1992). HIV-1 and HIV-2 infections in a high-risk population in Bombay, India: evidence for the spread of HIV-2 and presence of a divergent HIV-1 subtype. J Acquir Immune Defic Syndr 5: 972-7.

198.Philippon V, Matsuda Z, Essex M (1999). Transactivation is a conserved function among primate lentivirus Vpr proteins but is not shared by Vpx. J Hum Virol 2: 167-74.

199.Philipps D, Celotto AM, Wang QQ, Tarng RS, Graveley BR (2003). Arginine/serine repeats are sufficient to constitute a splicing activation domain. Nucleic Acids Res 31: 6502-8.

200.Pirovano S, Bellinzoni M, Ballerini C, Cariani E, Duse M, Albertini A et al (2002). Transmission of SEN virus from mothers to their babies. J Med Virol 66: 421-7.

201.Pirovano S, Bellinzoni M, Matteelli A, Ballerini C, Albertini A, Imberti L (2002). High prevalence of a variant of SENV in intravenous drug user HIV-infected patients. J Med Virol 68: 18-23.

202.Pollard VW, Malim MH (1998). The HIV-1 Rev protein. Annu Rev Microbiol 52: 491-532.

203.Popescu GA, Benea E, Mitache E, Piper C, Horstkotte D (2005). An unusual bacterium, Aerococcus viridans, and four cases of infective endocarditis. J Heart Valve Dis 14: 317-9.

204.Price RW (1994). Understanding the AIDS dementia complex (ADC). The challenge of HIV and its effects on the central nervous system. Res Publ Assoc Res Nerv Ment Dis 72: 1-45.

205.Price RW, Spudich S (2008). Antiretroviral therapy and central nervous system HIV type 1 infection. J Infect Dis 197 Suppl 3: S294-306.

206.Purcell DF, Martin MA (1993). Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity. J Virol 67: 6365-78.

207.Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C et al (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59-65.

208.Rathinam VA, Fitzgerald KA (2011). Innate immune sensing of DNA viruses. Virology 411: 153-62.

209.Razeq JH, Thomas GM, Alexander D (1999). The first reported case of Aerococcus bacteremia in a patient with HIV infection. Emerg Infect Dis 5: 838-9.

210.Re F, Braaten D, Franke EK, Luban J (1995). Human immunodeficiency virus type 1 Vpr arrests the cell cycle in G2 by inhibiting the activation of p34cdc2-cyclin B. J Virol 69: 6859-64.

211.Reik W (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. Nature 447: 425-32.

212.Renkema GH, Manninen A, Mann DA, Harris M, Saksela K (1999). Identification of the Nef-associated kinase as p21-activated kinase 2. Curr Biol 9: 1407-10.

213.Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F et al (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466: 334-8.

214.Ritola K, Pilcher CD, Fiscus SA, Hoffman NG, Nelson JA, Kitrinos KM et al (2004). Multiple V1/V2 env variants are frequently present during primary infection with human immunodeficiency virus type 1. J Virol 78: 11208-18.

215.Robertson DL, Anderson, J.P., Bradac, J.A., Carr, J.K., Koley, B., funkhouser, R.K., Gao, F., Hahn, B.H., Kalish, M.L., Kuiken, C., Learn, G.H., Leitner, T., et al. (1999). Human retroviruses and AIDS. Los Alamos National laboratory: NM.

216.Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK et al (2000). HIV-1 nomenclature proposal. Science 288: 55-6.

217.Rollins RA, Haghighi F, Edwards JR, Das R, Zhang MQ, Ju J et al (2006). Large-scale structure of genomic methylation patterns. Genome Res 16: 157-63.

218.Romano G, Kasten M, De Falco G, Micheli P, Khalili K, Giordano A (1999). Regulatory functions of Cdk9 and of cyclin T1 in HIV tat transactivation pathway gene expression. J Cell Biochem 75: 357-68.

219.Ropers D, Ayadi L, Gattoni R, Jacquenet S, Damier L, Branlant C et al (2004). Differential effects of the SR proteins 9G8, SC35, ASF/SF2, and SRp40 on the utilization of the A1 to A5 splicing sites of HIV-1 RNA. J Biol Chem 279: 29963-73.

220.Rotger M, Dang KK, Fellay J, Heinzen EL, Feng S, Descombes P et al (2010). Genome-wide mRNA expression correlates of viral control in CD4+ T-cells from HIV-1-infected individuals. PLoS Pathog 6: e1000781.

221.Rudnicka D, Schwartz O (2009). Intrusive HIV-1-infected cells. Nat Immunol 10: 933-4.

222.Sadek JR, Johnson SA, White DA, Salmon DP, Taylor KI, Delapena JH et al (2004). Retrograde amnesia in dementia: comparison of HIV-associated dementia, Alzheimer's disease, and Huntington's disease. Neuropsychology 18: 692-9.

223.Sadikot RT, Blackwell TS, Christman JW, Prince AS (2005). Pathogen-host interactions in Pseudomonas aeruginosa pneumonia. Am J Respir Crit Care Med 171: 1209-23.

224.Sanchez-Paz A (2010). White spot syndrome virus: an overview on an emergent concern. Vet Res 41: 43.

225.Sanclemente G, Gill DK (2002). Human papillomavirus molecular biology and pathogenesis. J Eur Acad Dermatol Venereol 16: 231-40.

226.Schafer EA, Venkatachari NJ, Ayyavoo V (2006). Antiviral effects of mifepristone on human immunodeficiency virus type-1 (HIV-1): targeting Vpr and its cellular partner, the glucocorticoid receptor (GR). Antiviral Res 72: 224-32.

227.Schuster SC (2008). Next-generation sequencing transforms today's biology. Nat Methods 5: 16-8.

228.Schwartz O, Marechal V, Le Gall S, Lemonnier F, Heard JM (1996). Endocytosis of major histocompatibility complex class I molecules is induced by the HIV-1 Nef protein. Nat Med 2: 338-42.

229.Sevigny JJ, Chin SS, Milewski Y, Albers MW, Gordon ML, Marder K (2005). HIV encephalitis simulating Huntington's disease. Mov Disord 20: 610-3.

230.Shah KM, Young LS (2009). Epstein-Barr virus and carcinogenesis: beyond Burkitt's lymphoma. Clin Microbiol Infect 15: 982-8.

231.Shav-Tal Y, Cohen M, Lapter S, Dye B, Patton JG, Vandekerckhove J et al (2001). Nuclear relocalization of the pre-mRNA splicing factor PSF during apoptosis involves hyperphosphorylation, masking of antigenic epitopes, and changes in protein interactions. Mol Biol Cell 12: 2328-40.

232.Shibata M, Wang RY, Yoshiba M, Shih JW, Alter HJ, Mitamura K (2001). The presence of a newly identified infectious agent (SEN virus) in patients with liver diseases and in blood donors in Japan. J Infect Dis 184: 400-4.

233.Shibayama T, Masuda G, Ajisawa A, Takahashi M, Nishizawa T, Tsuda F et al (2001). Inverse relationship between the titre of TT virus DNA and the CD4 cell count in patients infected with HIV. AIDS 15: 563-70.

234.Siedlecki P, Zielenkiewicz P (2006). Mammalian DNA methyltransferases. Acta Biochim Pol 53: 245-56.

235.Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K et al (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-50.

236.Simon F, Mauclere P, Roques P, Loussert-Ajaka I, Muller-Trutwin MC, Saragosti S et al (1998). Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nat Med 4: 1032-7.

237.Simon JH, Malim MH (1996). The human immunodeficiency virus type 1 Vif protein modulates the postpenetration stability of viral nucleoprotein complexes. J Virol 70: 5297-305.

238.Sobhian B, Laguette N, Yatim A, Nakamura M, Levy Y, Kiernan R et al (2010). HIV-1 Tat assembles a multifunctional transcription elongation complex and stably associates with the 7SK snRNP. Mol Cell 38: 439-51.

239.Solis M, Wilkinson P, Romieu R, Hernandez E, Wainberg MA, Hiscott J (2006). Gene expression profiling of the host response to HIV-1 B, C, or A/E infection in monocyte-derived dendritic cells. Virology 352: 86-99.

240.Spudich SS, Nilsson AC, Lollo ND, Liegler TJ, Petropoulos CJ, Deeks SG et al (2005). Cerebrospinal fluid HIV infection and pleocytosis: relation to systemic infection and antiretroviral treatment. BMC Infect Dis 5: 98.

241. Stade K, Ford CS, Guthrie C, Weis K (1997). Exportin 1 (Crm1p) is an essential nuclear export factor. Cell 90: 1041-50.

242. Stanley M (2010). Pathology and epidemiology of HPV infection in females. Gynecol Oncol 117: S5-10.

243. Stebbing J, Gazzard B, Douek DC (2004). Where does HIV live? N Engl J Med 350: 1872-80.

244. Storey JD, Tibshirani R (2003). Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440-5.

245. Stroun M, Anker P, Lyautey J, Lederrey C, Maurice PA (1987). Isolation and characterization of DNA from the plasma of cancer patients. Eur J Cancer Clin Oncol 23: 707-12.

246. Suda Y, Kokura K, Kimura J, Kajikawa E, Inoue F, Aizawa S (2010). The same enhancer regulates the earliest Emx2 expression in caudal forebrain primordium, subsequent expression in dorsal telencephalon and later expression in the cortical ventricular zone. Development 137: 2939-49.

247. Szalmas A, Konya J (2009). Epigenetic alterations in cervical carcinogenesis. Semin Cancer Biol 19: 144-52.

248. Takacs M, Banati F, Koroknai A, Segesdi J, Salamon D, Wolf H et al (2010). Epigenetic regulation of latent Epstein-Barr virus promoters. Biochim Biophys Acta 1799: 228-35.

249. Tan L, Grewal PS (2001). Pathogenicity of Moraxella osloensis, a bacterium associated with the nematode Phasmarhabditis hermaphrodita, to the slug Deroceras reticulatum. Appl Environ Microbiol 67: 5010-6.

250. Tanaka J, Ishida T, Choi BI, Yasuda J, Watanabe T, Iwakura Y (2003). Latent HIV-1 reactivation in transgenic mice requires cell cycle -dependent demethylation of CREB/ATF sites in the LTR. AIDS 17: 167-75.

251. Tang L, Wang L, Tan X, Xu W (2010). Adenovirus serotype 7 associated with a severe lower respiratory tract disease outbreak in infants in Shaanxi Province, China. Virol J 8: 23.

252. Tapia-Paez I, Tammimies K, Massinen S, Roy AL, Kere J (2008). The complex of TFII-I, PARP1, and SFPQ proteins regulates the DYX1C1 gene implicated in neuronal migration and dyslexia. FASEB J 22: 3001-9.

253. Taylor PW, Trueblood MC (1985). Septic arthritis due to Aerococcus viridans. J Rheumatol 12: 1004-5.

254.Tazi J, Bakkour N, Marchand V, Ayadi L, Aboufirassi A, Branlant C (2010). Alternative splicing: regulation of HIV-1 multiplication as a target for therapeutic action. FEBS J 277: 867-76.

255.Thom K, Petrik J (2007). Progression towards AIDS leads to increased Torque teno virus and Torque teno minivirus titers in tissues of HIV infected individuals. J Med Virol 79: 1-7.

256.Tillmann HL, Manns MP (2001). GB virus-C infection in patients infected with the human immunodeficiency virus. Antiviral Res 52: 83-90.

257.Tolias KF, Duman JG, Um K (2011). Control of synapse development and plasticity by Rho GTPase regulatory proteins. Prog Neurobiol.

258.Toth Z, Maglinte DT, Lee SH, Lee HR, Wong LY, Brulois KF et al (2010). Epigenetic analysis of KSHV latent and lytic genomes. PLoS Pathog 6: e1001013.

259.Tsui SK, Fong NY, Li SK, Leung KK, Chan DP, Chan PK et al (2010). Full genome analysis of an emerging cluster of human immunodeficiency virus type 1 subtype B infection in Hong Kong. AIDS Res Hum Retroviruses 26: 117-22.

260.Uh Y, Son JS, Jang IH, Yoon KJ, Hong SK (2002). Penicillin-resistant Aerococcus viridans bacteremia associated with granulocytopenia. J Korean Med Sci 17: 113-5.

261.Umemura T, Tanaka E, Ostapowicz G, Brown KE, Heringlake S, Tassopoulos NC et al (2003). Investigation of SEN virus infection in patients with cryptogenic acute liver failure, hepatitis-associated aplastic anemia, or acute and chronic non-A-E hepatitis. J Infect Dis 188: 1545-52.

262.Umemura T, Yeo AE, Sottini A, Moratto D, Tanaka Y, Wang RY et al (2001). SEN virus infection and its relationship to transfusion-associated hepatitis. Hepatology 33: 1303-11.

263.UNAIDS/WHO. (2010).

264.Urcuqui-Inchima S, Patino C, Zapata X, Garcia MP, Arteaga J, Chamot C et al (2011). Production of HIV particles is regulated by altering sub-cellular localization and dynamics of Rev induced by double-strand RNA binding protein. PLoS One 6: e16686.

265.Uzonna JE, Wei G, Yurkowski D, Bretscher P (2001). Immune elimination of Leishmania major in mice: implications for immune memory, vaccination, and reactivation disease. J Immunol 167: 6967-74.

266.Valcour V, Sithinamsuwan P, Letendre S, Ances B (2011). Pathogenesis of HIV in the central nervous system. Curr HIV/AIDS Rep 8: 54-61.

267. Van Duyne R, Easley R, Wu W, Berro R, Pedati C, Klase Z et al (2008). Lysine methylation of HIV-1 Tat regulates transcriptional activity of the viral LTR. Retrovirology 5: 40.

268. van 't Wout AB LG, Mikheeva SA, O'Keeffe GC, Katze MG, Bumgarner RE et al (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4(+)-T-cell lines. . J Virol 77: 1392-402.

269. Vanden Haesevelde M, Decourt JL, De Leys RJ, Vanderborght B, van der Groen G, van Heuverswijn H et al (1994). Genomic cloning and complete sequence analysis of a highly divergent African human immunodeficiency virus isolate. J Virol 68: 1586-96.

270. Vasilyev EV, Trofimov DY, Tonevitsky AG, Ilinsky VV, Korostin DO, Rebrikov DV (2009). Torque Teno Virus (TTV) distribution in healthy Russian population. Virol J 6: 134.

271. Veazey RS, DeMaria M, Chalifoux LV, Shvetz DE, Pauley DR, Knight HL et al (1998). Gastrointestinal tract as a major site of CD4+ T cell depletion and viral replication in SIV infection. Science 280: 427-31.

272. Virtual AIDS Office of Hong Kong DoH, Hong Kong. (2010).

273. Vivekanandan P, Daniel HD, Kannangai R, Martinez-Murillo F, Torbenson M (2010). Hepatitis B virus replication induces methylation of both host and viral DNA. J Virol 84: 4321-9.

274. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D et al (2002). Microarray-based detection and genotyping of viral pathogens. Proc Natl Acad Sci U S A 99: 15687-92.

275. Wang L, Mukherjee S, Jia F, Narayan O, Zhao LJ (1995). Interaction of virion protein Vpr of human immunodeficiency virus type 1 with cellular transcription factor Sp1 and trans-activation of viral long terminal repeat. J Biol Chem 270: 25564-9.

276. Wang WK, Chen MY, Chuang CY, Jeang KT, Huang LM (2000). Molecular biology of human immunodeficiency virus type 1. J Microbiol Immunol Infect 33: 131-40.

277. Warren K, Warrilow, D., Meredith, L., Harrich, D. (2009). Reverse Transcriptase and Cellular Factors: Regulators of HIV-1 Reverse Transcription. Viruses 1: 873-894.

278. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL et al (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet 37: 853-62.

279. Woese CR (1987). Bacterial evolution. Microbiol Rev 51: 221-71.

280. Woo PC, Lau SK, Huang Y, Yuen KY (2009). Coronavirus diversity, phylogeny and interspecies jumping. Exp Biol Med (Maywood) 234: 1117-27.

281. Woo PC, Lau SK, Teng JL, Tse H, Yuen KY (2008). Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. Clin Microbiol Infect 14: 908-34.

282. Woo PC, Ng KH, Lau SK, Yip KT, Fung AM, Leung KW et al (2003). Usefulness of the MicroSeq 500 16S ribosomal DNA-based bacterial identification system for identification of clinically significant bacterial isolates with ambiguous biochemical profiles. J Clin Microbiol 41: 1996-2001.

283. Woo PC, Tam DM, Leung KW, Lau SK, Teng JL, Wong MK et al (2002). Streptococcus sinensis sp. nov., a novel species isolated from a patient with infective endocarditis. J Clin Microbiol 40: 805-10.

284. Woo PC, Tsoi HW, Leung KW, Lum PN, Leung AS, Ma CH et al (2000). Identification of Mycobacterium neoaurum isolated from a neutropenic patient with catheter-related bacteremia by 16S rRNA sequencing. J Clin Microbiol 38: 3515-7.

285. Xiao Y, Word B, Starlard-Davenport A, Haefele A, Lyn-Cook BD, Hammons G (2008). Age and gender affect DNMT3a and DNMT3b expression in human liver. Cell Biol Toxicol 24: 265-72.

286. Yanagisawa Y, Ito E, Yuasa Y, Maruyama K (2002). The human DNA methyltransferases DNMT3A and DNMT3B have two types of promoters with different CpG contents. Biochim Biophys Acta 1577: 457-65.

287. Yang J, Bogerd HP, Peng S, Wiegand H, Truant R, Cullen BR (1999). An ancient family of human endogenous retroviruses encodes a functional homolog of the HIV-1 Rev protein. Proc Natl Acad Sci U S A 96: 13404-8.

288. Ye H, Choi HJ, Poe J, Smithgall TE (2004). Oligomerization is required for HIV-1 Nef-induced activation of the Src family protein-tyrosine kinase, Hck. Biochemistry 43: 15775-84.

289. Yong J, Golembe TJ, Battle DJ, Pellizzoni L, Dreyfuss G (2004). snRNAs contain specific SMN-binding domains that are essential for snRNP assembly. Mol Cell Biol 24: 2747-56.

290. Yong J, Pellizzoni L, Dreyfuss G (2002). Sequence-specific interaction of U1 snRNA with the SMN complex. EMBO J 21: 1188-96.

291.Youngblood B, Reich NO (2008). The early expressed HIV-1 genes regulate DNMT1 expression. Epigenetics 3: 149-56.

292.Yuen KY, Woo PC, Teng JL, Leung KW, Wong MK, Lau SK (2001). Laribacter hongkongensis gen. nov., sp. nov., a novel gram-negative bacterium isolated from a cirrhotic patient with bacteremia and empyema. J Clin Microbiol 39: 4227-32.

293.Zaura E, Keijser BJ, Huse SM, Crielaard W (2009). Defining the healthy "core microbiome" of oral microbial communities. BMC Microbiol 9: 259.

294.Zerbino DR, Birney E (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821-9.

295.Zhang Q, Raoof M, Chen Y, Sumi Y, Sursal T, Junger W et al (2010). Circulating mitochondrial DAMPs cause inflammatory responses to injury. Nature 464: 104-7.

296.Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SW et al (2006). RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol 4: e3.

297.Zhang X, Boyce M, Bhattacharya B, Schein S, Roy P, Zhou ZH (2010). Bluetongue virus coat protein VP2 contains sialic acid-binding domains, and VP5 resembles enveloped virus fusion proteins. Proc Natl Acad Sci U S A 107: 6292-7.

298.Zhao XF, Colaizzo-Anas T, Nowak NJ, Shows TB, Elliott RW, Aplan PD (1998). The mammalian homologue of mago nashi encodes a serum-inducible protein. Genomics 47: 319-22.

299.Zhu Y, Gelbard HA, Roshal M, Pursell S, Jamieson BD, Planelles V (2001). Comparison of cell cycle arrest, transactivation, and apoptosis induced by the simian immunodeficiency virus SIVagm and human immunodeficiency virus type 1 vpr genes. J Virol 75: 3791-801.

300.Zhu Y, Pe'ery T, Peng J, Ramanathan Y, Marshall N, Marshall T et al (1997). Transcription elongation factor P-TEFb is required for HIV-1 tat transactivation in vitro. Genes Dev 11: 2622-32.