creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

**저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.**

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

보건학 석사 학위논문

# 바이오인포매틱스 기법을 활용한 렌티바이러스 유전자 서열 분석

# GENETIC SEQUENCE ANALYSIS OF LENTIVIRUS USING BIOINFORMATICS METHODS

**2021년 8월**

서울대학교 보건대학원

보건학과 생명정보학 전공

민 향 란

# 바이오인포매틱스 기법을 활용한 렌티바이러스 유전자 서열 분석

지도교수 손 현 석

이 논문을 보건학 석사학위논문으로 제출함

**2021년 5월**

서울대학교 보건대학원

보건학과 생명정보학 전공

민 향 란

민향란의 석사학위논문을 인준함

**2021년 6월**

위 원 장 조 성 일

부 위 원 장 정 효 지

위 원 손 현 석

# 국 문 초 록

## 바이오인포매틱스 기법을 활용한 렌티바이러스 유전자 서열 분석

서울대학교 보건대학원 보건학과
생명정보학 전공

**민 향 란**

Human immunodeficiency virus 1 (HIV1)은 렌티바이러스 속의 대표 바이러스로 전 세계적인 후천성면역결핍증(acquired immune deficiency syndrome, AIDS) 발생을 초래하여 지속적인 공중보건학적 문제가 되고 있으며, 최초 감염 사례 보고 이후 40년 가까이 지난 현재까지 백신 개발 및 완치는 어려운 상황이다. 바이러스 코돈 사용에 대한 이해는 바이러스의 진화 방향 예측이나 코돈 편집의 참고 자료로써 숙주의 항바이러스 기전 및 바이러스 생활사 중 미지의 영역을 밝히는데 중요한 정보가 될 수 있다. Simian immunodeficiency virus (SIV)를 포함한 코돈 사용 비교 연구는 현재 미비한 실정으로, 본 연구에서는 HIV1, HIV2 및 SIV에서 공통적/특이적으로 발현되는 진화적·유전적 특성을 탐색하는 것을 주 목적으로 연구를 설계하였다. 분석에 앞서 JAVA 프로그래밍 언어 기반의 코돈 분석 프로그램 CUIs를 자체 개발하였다. 사용성과 안정성을 고려한 GUI 기반 스탠드얼론 시스템으로서 RSCU, ENC, ICDI, CAI, similarity D(A, B) 인덱스와 함께 PR2 분석을 위한 아미노산별 AT와 GC 편향 값의 산출을 지원하도록 하고 tAI, RCA, GRAVY, AROMO 등의 인덱스도 포함시켰다. 영장류 렌티바이러스 코돈 분석은 우선 영장류 렌티바이러스의 gag, pol 및 env 유전자별로 계통수를

구축하여 근연관계를 확인하고 SIV 숙주 분류체계를 고려하여 코돈 사용 비교 분석의 그룹핑을 진행하였다. 코돈 사용 분석은 세 유전자의 핵산 조성, 전반적 코돈 사용 편향, 진화 압력을 형성하는 돌연변이/선택 압력의 상대적 크기, 아미노산별 사용 편향, 인간의 코돈 사용에 대한 적응 정도에 대하여 비교 분석하였다. 결과적으로 HIV는 인간의 코돈 사용과의 유사성은 낮지 않지만 번역의 효율을 높이기 위하여 인간의 코돈 사용과 비슷한 쪽으로 진화하고 있지 않는 것이 확인되었다. 또한 전반적으로 낮은 코돈 사용 편향이지만 HIV1이 HIV2에 비해 높은 코돈 사용 편향을 보이고 있고 코돈 사용에서도 더 높은 선택 압력을 견뎌내고 있는 것을 확인할 수가 있었다. 이는 HIV1이 더 높은 생존력 및 발현도를 가짐으로서 큰 유행을 초래할 수 있는 근거가 될 수 있다. HIV의 기원 바이러스 SIVcpz_gor와 SIVsm의 gag 유전자에서 인간의 코돈 사용에 대한 유의한 적응성이 관찰되었는데 이는 코돈 사용 측면에서도 HIV의 기원을 뒷받침하는 증거가 될 수 있다. 또한 HIV1과 HIV2의 병원성 차이에 작용하는 잠재적 기작과의 연관성이 유추되는 HIV1의 TGT와 HIV2의 CTA와 GTC 등 특징적인 코돈 사용을 제시할 수 있었다. 영장류 렌티바이러스 전체에서 Pro, Thr, Ala, Arg, Gly 아미노산의 코돈 사용이 극도로 보존적인 것이 확인되었는데, 이는 바이러스 생존 및 생활사 진행에 중요한 영향을 미칠 가능성이 클 것으로 생각된다. 이상 렌티바이러스에 대한 연구결과는 코돈 사용 편집의 참고 자료로 활용되어 HIV의 새로운 치료 타겟 발견에 기여할 수 있으며, 개발된 프로그램 또한 바이러스 코돈 연구에 유용하게 사용될 수 있을 것으로 기대된다.

……

주요어 : 바이오인포매틱스, HIV1, 렌티바이러스, 코돈 사용 패턴

학 번 : 2019-28056

*I dedicate this paper to my professor & family.*

# 목 차

국문 초록 ····· i
목 차 ·····iv
표 목 차 ····· vi
그림 목차 ·····vii

**제 1 장. 서 론**

1.1 연구 배경 ····· 1
1.1.1 렌티바이러스의 분류 및 특징····· 1
1.1.2 HIV와 SIV의 유전적·역학적 특성 및 연구현황····· 2
1.1.3 HIV와 SIV에 대한 생명정보학적 분석 현황····· 5
1.1.4 코돈 사용 인덱스 개발 현황 ·····7
1.1.5 코돈 사용 인덱스 산출 프로그램 현황·····9
1.2 연구의 필요성 ····· 11
1.3 연구의 목적 ····· 13

**제 2 장. 연구 방법**

2.1 코돈 사용 인덱스 산출 프로그램 개발····· 14
2.2 렌티바이러스 데이터 수집 및 가공 ····· 15
2.3 계통발생 분석 및 정확도 평가 ····· 17
2.4 렌티바이러스 코돈 사용 패턴 분석 ····· 18
2.4.1 조성 분석 ····· 18
2.4.2 패리티 규칙 2 분석 ····· 18

2.4.3 중립 진화 분석 ····· 19
2.4.4 RSCU 분석 ····· 20
2.4.5 ENC, ICDI 분석 ····· 21
2.4.6 CAI, similarity D(A, B) 분석 ····· 23

**제 3 장. 연구 결과**

3.1 코돈 사용 인덱스 산출 프로그램 ····· 31
3.2 계통 발생 분석 결과 ····· 34
3.3 렌티바이러스의 코돈 사용 분석 결과 ····· 36
3.3.1 조성 분석 결과 ····· 36
3.3.2 패리티 규칙 2 분석 결과 ····· 37
3.3.3 중립 진화 분석 결과 ····· 38
3.3.4 RSCU 분석 결과 ····· 40
3.3.5 ENC, ICDI 분석 결과 ····· 42
3.3.6 CAI, similarity D(A,B) 분석 결과 ····· 44

**제 4 장. 고 찰**

····· 70

**제 5 장. 결론**

····· 74

참고문헌 ····· 77
Abstract ····· 84

# 표 목 차

**Table 2.1** Program composition and calculation formula of each index ··· 27
**Table 2.2** SIV host information ····· 28
**Table 2.3** The number of collected data for codon usage indices analysis 30
**Table 3.1a** RSCU value of lentiviral gag genes ····· 62
**Table 3.1b** RSCU value of lentiviral pol genes ····· 64
**Table 3.1c** RSCU value of lentiviral env genes ····· 66
**Table 3.2** ICDI value of lentiviral gag, pol and env genes····· 68
**Table 3.3** Correlation test result of ENC, ICDI, CAI and D(A, B) ····· 68
**Table 3.4** CAI and 5% critical eCAI value of lentiviral genes ····· 69
**Table 3.5** Similarity D(A, B) value of lentiviral gag, pol and env genes · 69

# 그 림 목 차

**Figure 2.1** Data flow diagram for CUIs program ····· 26
**Figure 3.1** Sequences check in CUIs program ····· 47
**Figure 3.2** Example of saving .xls or .txt file in CUIs program ····· 48
**Figure 3.3a** Phylogenetic tree constructed with nucleotide … ····· 49
**Figure 3.3b** Phylogenetic tree constructed with nucleotide … ····· 50
**Figure 3.3c** Phylogenetic tree constructed with nucleotide … ····· 51
**Figure 3.4a** Compositional features of lentiviral gag genes ····· 52
**Figure 3.4b** Compositional features of lentiviral pol genes ····· 53
**Figure 3.4c** Compositional features of lentiviral env genes ····· 54
**Figure 3.5a** Average PR2-bias plot of 11 lentiviral groups in gag genes ·· 55
**Figure 3.5b** Average PR2-bias plot of 11 lentiviral groups in pol genes ·· 56
**Figure 3.5c** Average PR2-bias plot of 11 lentiviral groups in pol genes ·· 57
**Figure 3.6** Neutrality plot for lentiviral gag, pol and env genes ····· 58
**Figure 3.7** RSCU plot of lentiviral gag, pol and env genes ····· 59
**Figure 3.8** ENC box plot of lentiviral gag, pol and env genes ····· 60
**Figure 3.9** ENC-GC3s plot of lentiviral gag, pol and env genes ····· 61

# 제 1 장. 서 론

## 1.1 연구 배경

### 1.1.1 렌티바이러스의 분류 및 특징

Human immunodeficiency virus 1 (HIV1) 이 대표 바이러스인 렌티바이러스속은 레트로바이러스과 오소레트로바이러스아과에 속하여 있고 현재 총 10개 종이 포함되어있다(ICTV, 2020). 이중 인간을 감염하여 후천성면역결핍증(acquired immune deficiency syndrome, AIDS)을 일으키는 HIV1, HIV2와 비인간 영장류를 감염하는 simian immunodeficiency virus (SIV)가 영장류 렌티바이러스 그룹을 구성하고, 소면역결핍바이러스(bovine immunodeficiency virus, BIV)와 젬브라나병 바이러스(jembrana disease virus, JDV)가 소 그룹, 고양이 면역 결핍 바이러스(feline immunodeficiency virus, FIV)와 푸마 렌티바이러스(puma lentivirus, PLV)가 고양이 그룹, 염소 관절염 뇌염 바이러스(caprine arthritis encephalitis virus, CAEV)와 비스나매디 바이러스(visna/maedi virus)가 염소 또는 양 그룹을 구성하고, 말 전염성 빈혈 바이러스(equine infectious anemia virus, EIAV)가 말 그룹으로 구성되어있다. 렌티바이러스속은 모두 gag, pol, env 유전자를 공통적으로 갖고 있고 종에 따라 보조단백질의 수량 및 종류가 다양하다. Gag 유전자는 바이러스의 기질단백질(matrix, MA), 캡시드(capsid, CA), 뉴클레오캡시드(nucleocapsid, NC) 를 암호화하고 env 유전자는 외피단백질인 surface protein (SU) 와 transmembrane protein (TM), pol 유전자는 역전사 효소, 인테그레이즈, 프로테아제 등

바이러스 전반적인 생활사에 필요한 세가지 효소를 암호화한다. 세 유전자의 변이 정도가 다양한데 env 에 비하여 gag 와 pol가 더 보존적이다(Foley, 2000). 렌티바이러스의 gag, pol, env 유전자에 대하여 계통수를 구축한 결과 영장류 렌티바이러스 그룹이 별개의 분기군을 형성하고 있는 것이 확인되었다(Sharp et al., 1999). 영장류 렌티바이러스와 기타 렌티바이러스의 가장 큰 차이점은 영장류 렌티바이러스 그룹은 pol 유전자 내에 dUTPase를 암호화하는 구역이 없다는 것이다(Foley, 2000). dUTPase는 dUTP를 분해함으로써 세포내 dUTP/dTTP 비율을 낮추어 DNA복제에서 dUTP가 잘못 사용되는 것을 최소화하여 DNA 합성 및 복제의 정확성을 보장하는 역할을 한다(Hizi & Herzig, 2015). dUTPase를 사용하는 대신, 영장류 렌티바이러스에서는 바이러스 복제 능력을 보장하기 위해 다른 방식을 사용한다. 실제 HIV에서 역전사시 유라실화(uracilization) 정도가 큰 것이 확인되었지만 이런 유라실화가 반대로 바이러스의 자가통합(autointegration)을 막고 숙주 세포 염색체에 바이러스 DNA의 삽입 및 바이러스 복제를 촉진하는 역할을 한다는 것이 확인되었다(Yan et al., 2011). HIV를 포함한 영장류 렌티바이러스의 이러한 특이성은 숙주(인간)에 대한 감염성을 높이는 중요한 특성으로 작용할 수 있다.

### 1.1.2 HIV와 SIV의 유전적·역학적 특성 및 연구현황

HIV는 CD4+ 림프구를 표적세포로 8~10년의 잠복기를 거쳐 인간의 면역체계를 심각하게 무너뜨려 최종 기회감염으로 사망에 이르게한다. CD4+ T림프구수가 1ml당 200개 이하로 감소된 상태를 후천성면역결핍증(AIDS)으로 정의하는데 이 단계로 들어서면 건강한 사람에게는 문제가 되지 않는 각종 세균, 진균 등에 의해

생명의 위협을 받는다. AIDS 환자가 나타난 1981년부터 2019년까지 전 세계적으로 총 7천5백7십만명이 HIV에 감염되었고 AIDS 관련 사망자수는 총 3천 2백 7십만명에 달하였다(UNAIDS, 2020). 강력항레트로바이러스 치료법(highly active antiretroviral therapy, HAART)의 도입으로 2004년부터 연사망자수는 감소의 추세를 보이고는 있지만 2019년 한 해 동안 아직도 69만명이 AIDS 관련 질병에 의해 사망하였다(UNAIDS, 2020).

HIV는 유전적 차이가 큰 HIV1 과 HIV2로 분류되는데 동일한 바이러스 전파방식, 유사한 병리성 과정을 갖고 있지만 HIV2는 HIV1에 비해 낮은 병원성 및 전염력을 나타내고 잠복기가 길며 AIDS로의 발전 가능성이 낮고 혈중 바이러스 역가가 낮은 특징을 보이고 있다(Vidyavijayan et al., 2017). HIV1의 전 세계적 확산·유행과 달리 HIV2는 주로 서 아프리카 지역에 집중되어 있지만 현재 앙골라, 모잠비크, 포르투갈, 프랑스, 미국, 인도 등 여러 나라에서 HIV2 감염이 보고되고 있다(Visseaux et al., 2016). HIV1과 HIV2의 염기서열에 대한 분석 결과 두 서열은 최대 55%의 상동성만 가지고 있고(Motomura et al., 2008) 유전체상으로도 다른 보조단백질을 코딩한다. HIV1은 gag, pol, env 세 구조단백질 이외에 tat, rev, vif, nef, vpr, vpu 보조 단백질을 암호화하지만 HIV2는 vpu 대신 vpx 유전자를 갖고 있다. HIV1과 HIV2에 교차면역이 존재하지 않아 HIV1에 감염 후 다시 HIV2에 감염될수 있으나 HIV1 감염 전 이미 HIV2에 감염된 상태라면 보다 높은 CD4+ T림프구 수와 AIDS로 발전하는 잠복기가 길어진다(Esbjörnsson et al., 2012). 유전적 거리 분석 및 계통수 분석에 의하면 HIV1은 M (main), O (outlier), N (not M-not O), P 그룹으로 나눌수 있고 M그룹이 글로벌 유행그룹으로 HIV1의 95%를 차지하고 있다. M그룹은 다시 10개(A-D, F-H, J-L)의 서로 같은 유전적 거리를 갖는 아형(subtype)과 서로 다른 아형 간의

재조합으로 형성한 circulating recombination forms (CRFs)로 나뉜다. M 그룹의 아형간 차이는 env 유전자에서 최대 35%까지 나타나며 아형 내에서도 최대 20%정도의 차이를 보이고 있다(Gaschen et al., 2002). HIV2는 8개(A-G, U)의 lineage를 갖고 있으며 HIV2에서의 lineage는 유전적 거리나 기원의 관점에서 HIV1의 그룹과 같은 분류 기준이다(Robertson et al., 2000). HIV의 표적세포는 CD4+ 림프구로 CD4분자가 HIV의 외피단백질 SU에 대한 수용체로 작용하고 세포 진입에 있어 알파 케모카인(chemokine) 수용체인 CRCX4 또는 베타 케모카인 CCR5가 보조 수용체로 작용하고 있다. CCR5수용체를 이용하는 바이러스를 R5 친화성 바이러스라고 하고 감염 초기에 대식세포 또는 단핵구 세포를 숙주세포로 삼고 CRCX4 수용체를 이용하는 바이러스를 X4 친화성 바이러스라고 한다. X4 친화성 바이러스는 주로 감염 말기에 T 세포를 타겟으로 하여 감염 말기 T세포의 급격한 감소의 원인이 된다(Moore et al., 2004).

HIV와 달리 SIV는 현재 40여종의 비인간 영장류에서 확인되었고 sooty mangabey와 같은 자연숙주에서는 HIV 감염과 비슷한 높은 혈중 바이러스 역가를 보이고는 있지만 질병을 일으키지 않고(Rey-Cuillé et al., 1998) 종간이동으로 감염된 침팬지, 아시안 마카크(macaca mulatta)에서는 원숭이 면역결핍질병을 일으키는 것이 확인되었다(Apetrei et al., 2005). 원숭이 면역결핍질병을 일으키지 않는 비병원성 SIV 종의 원숭이 감염은 HIV의 인간 감염에 비해 숙주(원숭이)에서 약한 면역반응을 일으키는 것이 확인되었으며, 이는 CXCR4 보조 수용체에 대한 친화성이 없고 CCR5 보조 수용체만 이용할수 있는 것이 일부 원인으로 작용한 것으로 보인다(Holmes, 2001). 이처럼 HIV1과 HIV2는 유전적·역학적 측면에서 다양한 차이를 나타내며 SIV 또한 숙주의 종에 따라 다른 감염양상을 보이고 있다.

### 1.1.3 HIV와 SIV에 대한 생명정보학적 분석 현황

영장류 렌티바이러스의 gag, pol, env 유전자에 대한 이전의 계통분류학적 연구결과를 살펴보면 숙주에 따라 즉 인간과 비인간 영장류 두 개의 큰 분기군이 형성된 것이 아니라 HIV1, SIVcpz, SIVgor가 한 분기군을 형성하고 HIV2와 SIVsm이 다른 분기군을 형성하는 것을 확인할 수 있다. 또한 HIV2가 집중되어 있는 서 아프리카 지역이 sooty mangabey의 서식지이고 서 아프리카의 농촌 지역에서 sooty mangabey를 애완용으로 키우고 있는 것이 확인되어 HIV2는 SIVsm이 종간장벽을 넘어 인간으로 전파되었을 확률이 매우 높은 것으로 유추되었다(Sharp et al., 1999). 또한 계통수상 HIV2, SIVsm 분기군에서 HIV2의 lineage들이 가까이 클러스터링(clustering) 되어있는 것이 아니라 HIV2와 SIVsm이 섞여있는 양상으로부터 HIV2는 lineage별로 sooty mangabey에서부터 최소 8번의 종간이동 사건으로 인간에게 전파된 것이 확인되었다. 역학 및 계통수 결과를 기반으로 HIV1의 M과 N그룹은 SIVcpz로부터 최소 두 번의 종간이동 사건으로 인간에게 전파된 것이 확인되었고(Gao et al., 1999) O와 P 그룹은 SIVgor로부터 전파된 것이 확인되었다(Sharp & Hahn, 2010; Hemelaar, 2012). SIV의 자연숙주는 구세계원숭이로 알려져 있는데 침팬지가 보유하고 있는 SIVcpz는 SIVrcm과 SIVgsn이 재조합된 것으로 확인되었고(Sharp et al., 2005) SIVgor 또한 침팬지로부터 종간이동으로 전파해온 것으로 확인되었다(Takehisa et al., 2009). 처음에는 침팬지에서 질병을 일으키지 않는 것으로 여겨졌으나, 2009년 Keele등의 연구에서 SIVcpz도 HIV1과 마찬가지로 진행성 CD4+ T세포 손실과 침팬지의 조기 사망을 초래하고 건강과 번식에 부정적인 영향을 미친다는 것이 확인되었다. 이처럼 구세계 원숭이로부터 종간이동으로 SIV를

획득한 침팬지와 아시안 마카크에서는 인간과 비슷한 후천성면역결핍증이 확인되고 있다(Neel et al., 2010). 따라서 현재 마카크는 HIV연구에서의 동물 모델로 많이 사용되고 있다(Gardner & Luciw, 2008).

HIV의 유전체 뉴클레오티드 조성을 확인한 결과 높은 GC편향을 보이는 인간의 유전체 조성과 달리 HIV의 유전체 뉴클레오티드 구성은 오랜 기간 동안 극도로 보존적이었고 높은 아데닌 조성(최대 40%)과 낮은 시토신 조성을 유지해 왔다(van der Kuyl & Berkhout, 2012). 특히 렌티바이러스의 gag 유전자보다 pol 유전자에서 더 높은 A편향이 관찰되고 있고 다른 보조 단백질 코딩 유전자와 겹치는 구역에서는 상대적으로 낮은 A편향이 관찰되고 있다(Van & Berkhout, 1995). 이런 아데닌이 풍부한 유전체 뉴클레오티드의 구성은 아데닌이 풍부한 코돈에 대한 선택성을 높이고 특히 변화성이 강한 코돈 세 번째 자리를 주로 아데닌으로 치환하는 경향을 가진다. 실제로 HIV1에 대한 연구에서 모든 유전자에서 A또는 T로 끝나는 코돈을 50%이상 사용하고 gag, pol, env, vif, tat 유전자에서는 89%이상으로 극도의 AT편향을 보이고 있다(Pandit & Sinha, 2011). 이런 높은 아데닌 조성은 주로 역전사 반응시 dNTP 풀(pool)의 불균형에 의한 HIV1 양성 가닥 유전자의 G→A 고빈도 돌연변이 원인으로 제안되고 있다(Vartanian et al., 2002). 인간의 숙주제한인자인 apolipoprotein B mRNA editing enzyme 3 (APOBEC3) 패밀리의 APOBEC3G 또한 G→A 고빈도 돌연변이에 기여하는 것으로 알려졌으나, 이는 dNTP 풀의 불균형에 의한 작용보다는 약한 것으로 확인되었다(Yu et al., 2004). 따라서 HIV의 vif 유전자가 APOBEC3G에 결합하여 이를 분해하더라도 HIV유전체는 아직도 높은 A편향 패턴을 보인다(Sheehy et al., 2002; Desimmie et al., 2014).

여러 생물에서 확인되었듯이 synonymous 코돈 즉 동일 아미노산을 암호화하는 서로 다른 코돈은 랜덤으로 사용되는 것이 아니라 코돈 사용 편향(codon usage bias)을 가지고 있다. 이런 코돈 사용 편향은 주로 돌연변이 압력, 번역 압력, 자연 선택 압력, 약물 압력 등으로 형성된 것이다(Jenkins & Holmes, 2003). 코돈 사용 및 뉴클레오티드 구성 편향의 정도와 원인에 대한 이해는 바이러스 진화 연구, 특히 바이러스와 숙주 세포 또는 면역 반응 사이의 상호작용을 이해하는데 필수적이다(Shackelton et al., 2006). 바이러스는 숙주의 '번역 공장'을 이용하여 번역하므로 바이러스의 코돈 사용은 감염 세포의 tRNA구성의 압력도 받는다. 하지만 HIV1에서 A 말단 코돈을 해독하는 여러 종류의 tRNA가 비리온내에 패키징된 것이 확인됨으로써 HIV1은 숙주의 tRNA pool를 조절해가고 있는 것이 발견되었고(Van et al., 2011) HIV1의 이런 전형적인 A편향의 생물학적 기능은 아직 밝혀지지 않았다. 또한 HIV1에 대하여 추출서열의 지리적 분포에 상관없이 9개 유전자의 시간의 흐름에 따른(연도별) 코돈 사용변화를 보기 위한 principal component analysis (PCA) 기반 군집 분석에서 gag, pol, env 유전자의 코돈 사용은 다른 유전자에 비하여 시간적 차이가 적은것으로 관찰되었다(Pandit & Sinha, 2011).

### 1.1.4 코돈 사용 인덱스 개발 현황

동의 코돈 분석에서 기본적으로 사용되고 있는 인덱스로는 코돈별 사용 편향을 나타내주는 상대 동의 코돈 사용(relative synonymous codon usage, RSCU) (Sharp & Li, 1986), 유전자의 전반적인 코돈 사용 편향을 평가하는 유효 코돈수(effective number of codons, ENC) (Wright, 1990), 바이러스와 레퍼런스 셋 즉 호스트의 코돈 사용의 유사성을 정량화는 코돈 적응 지수(codon adaptation index, CAI)

(Sharp & Li, 1987)가 있다. 이외 frequency of optimal codons (Fop)나 codon bias index (CBI)와 같은 인덱스들은 계산시 'optimal' 코돈의 정보를 필요로 하는데 이에 대한 정보가 부족하여 사용이 점점 줄어들고 있다. 대신 RSCU값을 이용한 intrinsic codon deviation index (ICDI) 가 개발되었고 이는 CBI와 ENC값과 높은 상관관계를 보인다(Freire-Picos, 1994). 코돈 사용 패턴 형성에 있어 번역 관련 요소의 영향을 돌연변이 편향과 구분하기 위하여 몇몇 새로운 인덱스와 기존 CAI의 사용을 개선하는 방법이 개발되었다. 우선 2004년 Reis 등에 의하여 바이러스가 숙주 세포의 tRNA pool에 얼마나 적응하고 있는지 즉 번역 효율을 직접적으로 평가할 수 있는 tRNA adaptation index (tAI)가 개발되었다. 다음 2008년 Puigbò 등에 의해 쿼리 서열의 GC 및 아미노산의 조성에 의하여 500개의 랜덤 서열을 생성하고 CAI의 기대값(eCAI)을 산출하여 계산된 CAI 절대값과 통계적으로 유의한 차이가 있는지를 비교하는 방식으로 바이러스가 받고 있는 번역 압력을 평가하였다. 또한 2010년에 개발된 상대 코돈 적응 지수(relative codon adaptation, RCA)는 레퍼런스 셋에서의 특정 코돈의 관찰값과 동의 코돈의 최대빈도값을 직접 나누는 CAI의 단위체 계산과 달리 레퍼런스 셋의 뉴클레오티드 구성을 고려하여 관찰된 특정 코돈의 빈도값과 코돈 세자리 각각에서 특정 뉴클레오티드가 관찰될 확률을 곱하여 얻은 기대 빈도값과의 편차를 비교하는 방식으로 유전자 발현 정도를 보다 정확하게 예측하였다(Fox & Erill, 2010). 2013년 개발된 similarity D(A, B) 지표는 바이러스와 숙주의 59개 RSCU값을 59차원에서의 두 공간 백터로 변환하여 형성된 각의 코사인 값으로 바이러스와 숙주의 코돈 사용 유사성을 측정하였다(Zhou et al., 2013). 바이러스가 받고 있는 진화 압력 및 진화 속도를 확인할수 있는 지표로는 단백질 코딩 서열로 부터 특정 시간 단위 내의 비 동의

사이트에서의 비 동의 치환율과 동의 사이트에서의 동의 치환율의 비율인 dN/dS (Li, 1993)가 있고 자연 선택 압력을 부분적으로 설명할 수 있는 단백질의 소수성(hydrophobicity)을 나타내는 지표인 GRAVY (Kyte & Doolittle, 1982) 와 방향족 아미노산(Phe, Tyr, Trp)의 상대 빈도를 대표하는 AROMO (Lobry & Gautier, 1994)가 있다.

### 1.1.5 코돈 사용 인덱스 산출 프로그램 현황

현재 코돈 분석에서 많이 사용되고 있는 소프트웨어 및 웹사이트로 CodonW (Peden, 2000), CAIcal (http://genomes.urv.es/CAIcal/)가 있다. CodonW는 1999년에 개발되었으며 codon adaptation index (CAI), frequency of optimal codons (Fop), codon bias index (CBI), effective number of codons (ENC), general average hydrophobicity (GRAVY), aromaticity (AROMO)등 인덱스와 기본적인 뉴클레오티드 조성을 제공하고 있지만 nucleotide composition 중 C1와 T1의 값이 바뀌어 제공되고 있는 것이 확인되었다. CodonW는 계층적 메뉴 기반 프로그램으로 사용자 친화성이 부족하여 편리성과 신속성이 다소 떨어진다는 단점이 있다. 또한 CodonW에 포함되어 있는 산출 인덱스 중 숙주에 대한 바이러스의 적응 정도를 예측하는 지표인 CAI의 경우 reference set로 대장균(escherichia coli), 고초균(bacillus subtilis), 출아형 효모(saccharomyces cerevisiae) 세 가지 만을 제공하고 있어 실제 연구에서의 활용가치가 많이 떨어지고 있다. 이점을 보완한 웹 기반 어플리케이션 CAIcal이 2008년에 개발되었는데 해당 프로그램에서는 reference set의 입력을 Codon Usage Database (CUD) (Nakamura et al., 2000)에서 제공하고 있는 코돈 사용 테이블의 포맷을 지원하여 직접 copy and paste 할수 있게 만들어졌다. CUD 에서는 35,799가지 생물의 코돈 사용 테이블을 보유하고 있으므로 CAI의 활용력을 높였다. CAIcal에서는 nucleotide composition, RSCU, ENC,

RCDI, eCAI도 함께 제공하고 있지만 인덱스 종류가 부족하여 실제 연구에서 자주 사용하는 인덱스들을 커버하지 못하고 있다. 또한 웹 서버 기반 어플리케이션으로 접속이 안되는 상황이 가끔 발생되고 있고 입력 서열 데이터가 클 경우 산출이 되지 않는 문제점이 있다. 이밖에 다른 프로그램들로는 EMBOSS, INCA, DnaSP 및 $stAI_{calc}$ 등이 있는데, EMBOSS (https://www.bioinformatics.nl/emboss-explorer/)는 CAI, ENC의 산출, 뉴클레오티드 서열로부터 코돈 사용 테이블의 생성, 코돈 사용 테이블 리포맷 기능을 제공하고 있다. INCA (http://www.bioinfo-hr.org/software-tools/#INCA)(Supek & Vlahoviček, 2004) 프로그램은 대화식 그래픽 화면으로 ENC, CAI, GC3%, GRAVY 등의 지표 산출과 동시에 산출된 인덱스를 기반으로 산점도를 생성하는 기능을 제공하고 있으나, 데이터 항목별로 산출된 값을 프로그램 내에서 확인할수 없고 정확한 매뉴얼을 제공하고 있지 않아 사용이 어렵다. DnaSP(Rozas et al., 2017)도 RSCU, ENC, CBI, Scaled Chi Square와 GC함량 산출 기능을 제공하고 있으나 기본 FASTA 파일(.fasta)이 아닌 서열 정렬 파일(.fas)만을 입력 파일 형식으로 지원하고 있어 코돈 인덱스 산출 기능의 활용성이 떨어진다. $stAI_{calc}$(http://tau-tai.azurewebsites.net/)는 각종 생물의 wobble 가중치를 산출함으로써 tAI인덱스 값의 정확도를 높이기 위해 개발되었다(Sabi et al., 2017).

## 1.2 연구의 필요성

2019년 기준 HIV1의 연감염자수는 170만명으로 AIDS환자가 처음 나타난 1981년부터 38년이 지난 시점에도 아직 많은 감염자를 초래하여 공중보건학적으로 큰 문제가 되고 있다(UNAIDS, 2020). HAART, 소위 칵테일 요법의 출현으로 HIV1 감염자의 혈중 바이러스 부하를 낮은 수치로 억제함으로써 HIV1 감염자의 기대수명을 대폭 높히고는 있지만 수명이 긴 기억 T 림프구와 같은 바이러스 저장소의 존재로 인해 HIV1을 완전히 치유하는 것은 아직 불가능하다. 또한 HIV1 감염을 예방하는 유효 백신 개발에도 큰 어려움을 겪고 있는데 그 주요원인은 HIV1 바이러스주의 다양성, 빈번한 재조합 및 역전사 반응의 높은 돌연변이율로 모든 바이러스 준종을 중화할수 있는 광범위중화항체 유도가 어렵다는 점이다. 또한 SU 단백질의 구조적차단(conformational masking)으로 항체의 접근이 어려울뿐 아니라 SU단백질의 낮은 밀도로 항체의 이가결합이 어려운것도 해결해야할 문제점이다. 현재 67%의 HIV 감염자가 항레트로바이러스 치료를 받고 있지만 항레트로바이러스 약물 치료의 높은 비용은 HIV감염자가 가장 많이 집중되어 있는 아프리카 지역에서의 치료 불평등성을 초래하고 약물의 부작용 및 짧은 지속시간으로 인한 항레트로바이러스 치료의 의존성도 문제가 되고 있다. 따라서 현재 HIV바이러스 관련 주요 연구는 항레트로바이러스 약물 복용 의존성 문제 해결을 위한 장기지속형(long-acting) 약물의 개발(Link et al., 2020), 치유를 위한 프로바이러스(provirus) 제거(Mancuso et al., 2020), elite controller (longterm nonprogressors, LTNPs)의 광범위 중화 항체(broadly neutralizing antibodies, bNAbs) 생성 기작 연구(Jiang et al., 2020) 또는 조혈모세포에 대한 탈체 유전자 편집 등에 집중되어 있다. 영장류

렌티바이러스 즉 HIV1, HIV2, SIV는 밀접한 관련이 있는 세 종의 바이러스로 현재 HIV연구에서도 HIV와 SIV를 결합한 SHIV를 이용하여 마카크(rhesus macaque)에서 실험을 많이 진행하고 있다. 하지만 렌티바이러스의 심한 A편향 뉴클레오티드 조성으로 인해 HIV1에 대한 코돈 사용 분석과 HIV1과 HIV2에 대한 코돈 사용 비교분석은 활발히 이루어지지 않았으며, 특히 HIV와 진화적으로 긴밀한 관계가 있는 SIV에 대한 코돈 사용 분석은 매우 미비하여, 현재 40여종의 SIV를 모두 커버한 코돈 사용 분석이나 병원성 SIV 및 비병원성 SIV에 대한 코돈 사용 비교 분석도 없는 실정이다. Vidyavijayan 등이 수행한 HIV1과 HIV2에 대한 코돈 사용 비교 분석은 ENC 인덱스 및 ENC-GC3s plot을 중심으로 설명이 이루어졌는데(Vidyavijayan et al., 2017) 인간 코돈 사용에 대한 적응성이나 바이러스 코돈 사용 패턴 형성에서 숙주 코돈 사용의 영향을 정량화하는 등 코돈 사용 형성의 여러 방면을 고려하지 않고있어 이 부분에 대한 연구의 필요성이 보이고 있다. 또한 코돈 사용 분석에서 현재 사용되고 있는 도구들의 사용성, 인덱스 커버, 대용량 데이터 계산에 대한 지원성 및 서버 기반 웹 어플리케이션의 불안정성을 고려하여 새로운 코돈 분석 도구의 개발이 필요한 실정이다.

## 1.3 연구의 목적

본 연구의 목적은 첫째, 현재 바이러스 코돈 연구 진행 시 직면하고 있는 문제점을 해결하고 코돈 연구의 편리성을 높이는 새로운 GUI 기반 코돈 분석 프로그램을 개발하는 것이다. 구체적으로 대용량 데이터에 대한 CAI 인덱스 산출을 가능하게 하고, 또한 parity rule 2 (PR2) 분석을 아미노산별로 진행할 수 있게 아미노산별 AT3와 GC3 편향을 제시하며, 데이터 클리닝 도구로도 활용할 수 있게 문제 서열을 배제하여 서열 데이터를 반환하고, RSCU, ENC, ICDI, CAI, similarity D(A, B), tAI, RCA, GRAVY, AROMO 등 인덱스의 계산을 통합적으로 지원하는 프로그램을 개발하고자 한다.

둘째, 개발된 프로그램 기반으로 40여 종의 SIV를 포함한 영장류 렌티바이러스에 대하여 gag, pol, env 유전자의 전반적인 코돈 사용 분석을 진행하고 서로 다른 질병 양상을 보이고 있는 HIV1, HIV2 및 두 바이러스의 기원인 SIVcpz_gor와 SIVsm 그리고 SIV의 자연 숙주인 구세계원숭이와 비인간 영장류에서도 AIDS 상관 질병을 일으키는 SIVcpz_gor와 SIVmac의 코돈 사용을 비교하여 분석하고자 한다. 코돈 분석을 통하여 렌티바이러스의 진화 방향을 확인하고 HIV1, HIV2 및 SIV에서 공통적/특이적으로 발현되는 진화적·유전적 특성을 탐색하고자 한다. 궁극적으로 코돈 연구 결과 기반으로 HIV1과 HIV2의 병원성 차이를 초래하는 가능한 원인을 파악하여 HIV의 새로운 치료 타겟 발견에 기여하고자 한다. 또한 렌티바이러스에서 공통적으로 보존적인 코돈 사용을 확인하여 다양한 미생물에 대한 면역화나 특정 질병 상태를 치료하는 면역요법에 쓰이는 HIV 백신 벡터의 코돈 재코딩 참고 자료를 제공하여 벡터의 안정성 및 활용성을 높이고자 한다.

# 제 2 장. 연구 방법

## 2.1 코돈 사용 인덱스 산출 프로그램 개발

본 연구를 통해 개발한 프로그램은 기본적인 뉴클레오티드 조성 이외에도 중립 진화 분석시 필요한 GC1s%, GC2s%, GC3s%과 PR2 분석시 필요한 4코돈 아미노산 전체의 A3/(A3+T3), G3/(G3+C3)값과 8개의 아미노산 각각의 A3/(A3+T3), G3/(G3+C3)의 산출을 제공하도록 설계하였다. 코돈 인덱스 지표는 크게 네가지 부분으로 코돈별 사용 편향 지표인 RSCU, 유전자의 전반적인 코돈 사용 편향 지표인 ENC와 ICDI, 레퍼런스 셋의 코돈 사용 테이블이나 tRNA 유전자 복제수가 필요로 하는 적응도 지표인 CAI, RCA, similarity D(A, B), tAI 및 선택압력을 부분적으로 평가할수 있는 GRAVY, AROMO를 포함하였다. 각 인덱스의 산출방법 및 출처는 Table2.1에 정리되어 있다.

개발된 codon usage indices (CUIs) 프로그램은 graphical user interface (GUI) 기반 스탠드 얼론 프로그램으로 JAVA swing으로 작성되었다. Figure2.1에 제시한 바와 같이 입력 파일은 FASTA 형식의 뉴클레오티드 서열 데이터로 받아 파일 선택 즉시 서열 체크를 진행하도록 하였다. 종결 코돈이 정확하지 않은 서열(TAG, TGA, TAA가 아닌 서열), coding sequence (CDS)내에 종결 코돈이 존재하는 서열, ATCG 외에 unknown base 혹은 degenerate base가 섞인 서열, CDS 길이가 3의 배수가 아닌 서열을 제외시키고 제외된 서열 헤드라인을 메시지 창에 제시하도록 하였다. 클리닝된 서열 데이터는 인덱스 계산후 저장 버튼 클릭시 자동으로 FASTA 파일로

저장되어 CUIs는 데이터 클리닝 도구 로도 활용 가능하도록 설계하였다. 인덱스 CAI 계산 알고리즘은 6개의 동의 코돈을 갖고 있는 Arg, Leu, Ser 아미노산의 경우 두 세트의 서로 다른 tRNA에 의하여 번역되므로 코돈의 앞 두자리에 의하여 두개의 코돈 패밀리로 분리하여 계산하였다(Xia, 2007). CAI, RCA, similarity D(A, B) 계산에 필요한 레퍼런스 셋의 코돈 사용 테이블은 CUD(Nakamura et al., 2000)에서 35,799가지 생물의 코돈 사용 테이블 및 HIVE-CUTs 데이터베이스(Athey et al., 2017)로부터 52가지 인간 조직 코돈 사용 테이블을 FTP로 받아 프로그램에 내재하였다.

프로그램으로부터 산출된 결과값은 기존 CodonW, CAIcal 프로그램 결과값과 비교하고 인덱스 산출 지원 프로그램이 없는 경우 인덱스 개발 논문에서 제시된 결과값과 일일이 체크하여 정확도 검증을 진행하였다. HIV1의 env 유전자의 서열수가 70,658개로 가장 많은데 이 179MB 크기의 FASTA 파일을 사용하여 CAIcal, CodonW와 개발된 CUIs를 테스트한 결과 CAIcal는 용량제한으로 FASTA 파일 입력 자체가 불가하였다. CodonW의 경우 용량제한의 문제는 없었으나 인덱스의 종류 및 레퍼런스 셋의 한계가 분명하여, CUIs가 이러한 기존 프로그램의 문제점을 해결할 수 있음을 확인하였다.

## 2.2 렌티바이러스 데이터 수집 및 가공

Gag, pol, env 유전자는 백신 개발 및 치료에 중요한 역할을 갖고 있고 기존 HIV1 연구에서 시간에 따른 코돈 사용 차이가 적은 것으로 확인되어 본 연구에서는 gag, pol, env를 선택하여 분석을 진행하였다. 서열 데이터는 NCBI Virus로부터 HIV1, HIV2, SIV 세 바이러스의 gag, pol, env open reading frame (ORF) 별로 수집하였다.

NCBI Virus에서 protein 옵션을 사용하여 서열 데이터를 다운로드하였으며 이때 데이터가 GenBank 파일의 CDS product 명칭으로 분류되어 각 유전자별 입력란 프롬프트의 가능한 명칭들을 모두 선택하도록 하였다. 예를 들어 HIV1의 gag 폴리단백질의 경우 'gag protein', 'gag polyprotein', 'Gag' 세가지 모두를 선택하였다. Pol의 경우 GenBank 파일 대부분이 직접 pol 폴리단백질로 제시되지만 일부 데이터는 gag-pol 폴리단백질로 제시되어 있다. Gag-pol 폴리단백질은 -1리보솜 틀 이동(ribosomal frameshifting) 발생 위치가 join(280..1563,1563..4562)의 형식으로 제시되어 있으므로 JAVA 언어로 해당위치를 찾아 자르는 방식으로 파싱(parsing) 작업을 수행하여 pol 서열을 획득하였다. SIV 데이터의 경우 호스트 정보가 종으로 되어있는데 속(genus)별 또는 아과(subfamily)별 차이를 보기 위하여 Table2.2에 명시된 종속관계로 FASTA 파일의 헤드라인을 Accession과 호스트 속 또는 아과 정보의 조합으로 파싱 작업을 수행하였다. 이중 호스트 정보가 simiiformes로 되어있는 서열은 종 정보가 없으므로 데이터 셋에서 제외하였고(gag: 343, pol: 186, env: 1950) 호스트가 Homo sapiens인 서열도 제외하였다(pol: 3). NCBI Virus의 nucleotide completeness 옵션이 부분 서열과 전체 서열을 정확하게 구분하고 있지않는 것이 확인되어 모든 데이터를 수집한 후 바이러스별 RefSeq를 참고한 유전자별 서열 길이 최소치를 설정하고 JAVA를 이용하여 기준 미부합 서열을 배제하였다. Pol 서열의 경우 시작 코돈이 ATG가 아니므로 gag, env 서열만 시작 코돈 체크를 진행하였다. 다음 서열마다 길이가 3의 배수인지, 종결 코돈이 정확한지, ATCG외 다른 문자나 부호가 섞여 있는지는 개발된 CUIs 프로그램을 이용하여 클리닝 작업을 수행하였다. 최종 인덱스 계산에 쓰인 데이터 수는 Table2.3에 명시되어 있다.

## 2.3 계통발생 분석 및 정확도 평가

HIV1, HIV2, SIV의 진화적 근연관계를 확인하기 위하여 HIV1, HIV2는 랜덤으로 5개 서열을 추출하고 SIV는 호스트의 종별로 2개 서열을 추출하여 gag, pol, env 유전자 별로 계통수를 작성하였다. SIV 호스트의 분류체계 및 호스트 종에 따른 SIV의 약자는 Table2.2에 제시하였다. SIV의 경우 종별로 하나의 서열만 제공되는 경우도 있어, 최종 계통수 구축에는 gag, pol, env 유전자 별로 각각 60, 59, 53개의 서열이 사용되었다.

계통수를 구축하기 위한 사전 단계로 ClustalW (Thompson et al., 1994)를 이용하여 다중 서열 정렬을 진행하였다. 다음으로 계통수 구축을 위해 MEGA-X (Kumar et al., 2018)를 사용하였으며 치환 모델은 transversion에 비하여 transition이 더 쉽게 발생하는 것을 고려한 Kimura-2 parameter (Kimura, 1980)를 선택하고 서열의 모든 site의 진화 확률을 고려한 확률 기반 maximum likelihood (ML) 방법을 사용하였다(Kishino & Hasegawa, 1989). 부트스트랩은 1000으로 설정하여 계통수의 신뢰도를 측정하였다. 계통수 구축시 발생 가능한 바이어스를 최소화하기 위하여 gag, pol, env 유전자 별로 위의 과정을 각각 5번씩 수행하여 일치도가 큰 대표 트리를 선택하여 제시하였다. 영장류 렌티바이러스의 근연관계를 확인하고 SIV의 경우 계통수 결과를 기반으로 호스트의 속, 족, 아과를 고려하여 아래 코돈 분석을 위한 SIV의 그룹핑 작업을 진행하였다. 또한 코돈 사용 패턴 비교 분석의 결과 해석은 렌티바이러스의 근연관계를 고려하여 진행하였다.

## 2.4 렌티바이러스 코돈 사용 패턴 분석

### 2.4.1 조성 분석

개발된 CUIs 프로그램을 통해 HIV1, HIV2, SIV의 gag, pol, env에 대하여 각 유전자의 전반적인 A, T, C, G 함량과 코돈 세 번째 위치의 A3%, T3%, C3%, G3% 그리고 GC%, AT%, GC3%, AT3% 값을 산출하고 MS Excel 2016로 이를 시각화하여 직관적으로 조성 분석을 진행하였다.

### 2.4.2 패리티 규칙 2 분석

Parity rule 2 (PR2)는 1995년 Sueoka에 의하여 최초로 제안되었는데 돌연변이 압력과 선택 압력이 DNA의 두 가닥에 똑같이 작용한다면, 한 가닥내 뉴클레오티드의 상대 빈도도 A=T, G=C의 법칙을 따른다는 것이다(Sueoka, 1995). 하지만 4 코돈 아미노산의 코돈 세 번째 뉴클레오티드 조성을 관찰하면 보통 parity 2 법칙에 위반되며 동의 코돈 편향이 존재하는데 이는 주로 선택 편향에 의하여 형성된 것이다. 그 이유는 4코돈 아미노산만을 고려함으로써 랜덤 돌연변이의 영향을 제외하기 때문이다. 따라서 PR2 polt은 오롯이 방향성 돌연변이 압력(AT↔GC) 및 선택 압력의 영향만 나타내준다. 또한 방향성 돌연변이가 종별 또는 유전체 내의 GC함량 차이를 조성하는 주요 원인으로 작용하고 있는 것으로 알려져 있는데 인간 유전자를 GC함량으로 그룹화 하여 PR2 plot을 나타낸 결과 각 그룹에서 비슷한 PR2 바이어스 패턴이 확인되었다(Sueoka, 1999). 이는 4코돈 아미노산의 PR2 편향 형성의 주된 요소가 선택 압력임을 보여주는 근거가 될 수 있다. 또한 PR2 편향은 아미노산 특이적인데 방향성 돌연변이에 의한 코돈 사용 편향 형성이 아미노산별로 이루어지기는 어렵지만 tRNA pool에 의한

선택은 아미노산별로 작용할 수 있다는 점도 PR2 바이어스는 주로 선택압력에 의하여 형성되었다는 것을 설명해줄 수 있다. 이상의 이론 지식을 바탕으로 gag, pol, env 유전자에 대하여 4개의 동의 코돈을 갖고 있는 아미노산(Val, Pro, Thr, Ala, Gly) 및 6개의 동의 코돈을 갖고 있는 아미노산의 4 코돈 패밀리(Leu의 CTT, CTC, CTA, CTG, Serine의 TCT, TCC, TCA, TCG 과 Arg의 CGT, CGC, CGA, CGG)별로 각각 PR2 plot을 생성하였다. PR2 plot의 X축은 G3/(G3+C3), Y축은 A3/(A3+T3)로 GC편향과 AT편향을 나타내고 HIV1, HIV2 및 SIV의 그룹별 G3/(G3+C3), A3/(A3+T3) 평균값의 위치를 제시하고 돌연변이 압력과 선택 압력이 오프셋된 (0.5, 0.5) 위치로부터의 상대 위치로 바이러스들의 PR2 편향의 방향과 크기를 확인하였다. PR2 그림에 필요한 아미노산 별 AT 편향 및 GC 편향 값은 개발된 CUIs 프로그램을 이용하여 계산하였다.

### 2.4.3 중립 진화 분석

중립 진화 분석은 코돈 세 자리의 GC함량(GC12s와 GC3s)을 비교함으로써 돌연변이 압력 및 선택 압력을 정량화하는 방법이다. 돌연변이는 보통 GC함량이 높거나 낮은 방향으로 진행되는 '방향성'을 띠고 있는데, 이런 방향성 돌연변이 압력은 기능이 중요한 부분보다 코돈 세 번째 자리와 같은 중성 부분에 더 많이 작용한다(Sueoka., 1988). 6코돈 아미노산 이외 코돈 첫 번째와 두 번째 자리의 돌연변이는 비 동의 돌연변이로 단백질의 1차 구조를 변화시켜 단백질의 기능까지 영향을 미칠 수 있으므로 흔히 발생하지 않고 선택 압력을 크게 받고 있을 시 같은 바이러스의 서열별로 GC12s 함량은 큰 차이가 없을 것이다. 따라서 GC3s의 변화에 따른 GC12s의 변화 정도 즉 회귀직선의 기울기는 방향성 돌연변이 압력의 상대적 크기를 나타내주고 있다. 여기서 GC12s%의

s는 동의 코돈을 의미하여 하나의 코돈으로만 암호화되는 Met과 Try을 제외한 코돈 첫 번째 위치와 두 번째 위치의 GC함량의 평균이고 GC3s% 는 코돈 세 번째 위치의 동의코돈의 GC함량이다. 우선 HIV1, HIV2, SIV의 모든 서열 데이터를 gag, pol, env 유전자별로 GC12s를 종좌표로, GC3s의 값을 횡좌표로 산점도를 그린다. 다음으로 바이러스별로 점들이 상관관계가 있는지를 확인하고 상관관계가 있다면 회귀 직선을 그려 유전자의 코돈 사용 패턴 형성의 주요 압력을 확인한다. 회귀 직선의 기울기가 1일 때 즉 모든 점들이 대각선에 분포되어 있다면 다윈 선택 압력은 없거나 약하고 오롯이 돌연변이 압력에 의하여 코돈 사용 편향이 형성되었다는 의미이다. 반대로 회귀 직선의 기울기가 0, 즉 모든 점들이 횡좌표와 평행한 선에 위치되어 있다면 모든 유전자 서열의 코돈 첫 번째, 두 번째 자리의 GC 함량에 차이가 없는 것으로, 진화는 오롯이 선택 압력에 의해 추진된 것으로 간주된다. 분석에 사용된 모든 유전자 서열의 GC12s 및 GC3s 값은 본 연구에서 개발된 CUIs 프로그램을 통해 산출되었다.

### 2.4.4 RSCU 분석

상대 동의 코돈 사용(relative synonymous codon usage, RSCU) 인덱스는 특정 코돈의 관찰된 빈도를 해당 코돈의 기대빈도 즉 해당 아미노산을 코딩하는 동의 코돈의 평균값으로 나눈 것으로 정의된다(Sharp & Li, 1986). RSCU 값은 서열마다 세 개의 종결코돈 TAG, TGA, TAA와 하나의 코돈만 갖고 있는 아미노산 Met와 Trp를 제외한 59개 코돈별로 제시되고, 이는 유전자의 서열 길이 및 아미노산 조성의 영향을 받지 않는다. 구체적인 산출 공식은 다음과 같다.

$$RSCU_{ij} = \frac{x_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}}$$

여기서 xij는 i 번째 아미노산중 j 번째 코돈의 발생 빈도이고 ni는 i 번째 아미노산이 갖고 있는 동의 코돈 수 이다(2, 3, 4 또는 6). RSCU값이 1이면 해당 코돈이 랜덤으로 사용된 것으로 간주되고 1.6보다 클 경우 과잉 표현, 0.6보다 작을 경우 과소 표현된 것으로 간주된다(Wong et al., 2010). CUIs 프로그램으로 RSCU 값을 산출하고 HIV1, HIV2, SIV 그룹 별로 59개 코돈의 평균값을 각각 구한다. 다음 XLSTAT을 이용하여 gag, pol, env 유전자별로 산점도를 생성하여 비교분석을 진행하였다.

### 2.4.5 ENC, ICDI 분석

**ENC**

유효 코돈 수(effective number of codons, ENC)는 전체 코돈 사용 편향을 나타내는 지표로 코돈 빈도표를 통해 계산할 수 있으며 유전자 길이와 아미노산 구성의 영향이 최소화된 값이다. 산출 공식은 다음과 같다.

$$F = \frac{n\sum_{i=1}^{k} p_i^2 - 1}{n-1} \quad n>1 \qquad p_i = \frac{n_i}{n} \qquad (1)$$

$$ENC = 2 + \frac{9}{\overline{F_2}} + \frac{1}{\overline{F_3}} + \frac{5}{\overline{F_4}} + \frac{3}{\overline{F_6}} \qquad (2)$$

위의 공식에서 ni는 관찰된 코돈의 빈도이고 n은 해당 아미노산을 코딩하는 총 코돈 수 이다. 공식 (2)에서 Fi는 각 i코돈 패밀리 F값의 평균이고 ENC 값은 다섯 그룹의 기여도를 더하여 산출한다. ENC 값의 범위는 20에서 61이고 20의 경우 아미노산별 하나의 코돈만 사용된 것으로 극도의 코돈 사용 편향을 의미하고 61의 경우 모든 동의 코돈이 똑같은 빈도로 사용된 것을 의미한다(Wright, 1990). 보통 ENC값이 35 미만의 경우 상당한 코돈 사용 편향이 있는 것으로 간주되고 50 이상은 코돈이 무작위로 사용된 것으로 간주된다(Jiang et al., 2008). 유전자 서열의 GC3s%의 편향은 ENC값에 영향을 주기에 ENC값의 해석을 돕기 위하여 보통 ENC-GC3s plot이 활용된다. 그래프의 X축은 GC3s%로 Y축은 ENC로 서열들의 코돈 사용 편향을 시각화하여 보여주고, 선택 압력이 없다는 가정 하에 GC3s% 값에 따른 ENC의 기대 값 곡선을 제시하여 바이러스별 유전자들이 받고있는 돌연변이 압력을 확인할 수 있다. 만약 실제 유전자의 ENC 값이 기대 곡선 근처에 위치하여 있다면 해당 유전자가 돌연변이 압력만을 받고 있는 것으로 간주되고, 반대로 기대 곡선과 거리가 멀수록 돌연변이 압력이 아닌 자연 선택이나 기타 압력에 의하여 코돈 사용 패턴이 형성된 것으로 생각된다. ENC의 기대 값 산출 공식은 아래와 같으며 s는 GC3s%를 의미한다.

$$ENC_e = 2 + s + \frac{29}{(s^2 + (1-s)^2)}$$

**ICDI**

고유 코돈 편차 지수(intrinsic codon deviation index, ICDI)는 optimal 코돈이 알려지지 않은 종에서 유전자의 전반적인 코돈 편향을 추정하는데 유용한 척도로 CBI와 높은 상관 관계를

보인다(Freire-Picos, 1994). 산출 공식은 다음과 같다.

$$S_k = \sum (n_i - 1)^2 / k(k-1)$$

$$ICDI = (\sum S_2 + S_3 + \sum S_4 + \sum S_6)/18$$

ni는 i 번째 코돈의 RSCU 값이고 k는 해당 아미노산의 축퇴 정도이다(2,3,4 또는6). ICDI값의 범위는 0에서 1 이고 1은 고도로 편향된 코돈 사용을 의미하고 일반적으로 ICDI 값이 0.3 미만인 유전자는 낮은 코돈 사용 편향, 0.5 이상은 높은 코돈 사용 편향을 가지는 것으로 간주된다(Freire-Picos, 1994). ICDI 지표를 통해 바이러스별 코돈 사용 편향을 평가하고 ENC 분석 결과와 같은 패턴을 보이는지를 확인할 수 있다.

### 2.4.6 CAI, similarity D(A, B) 분석

**CAI**

코돈 적응 지수(codon adaptation index, CAI)는 바이러스의 코돈 사용이 숙주의 코돈 사용과 얼마나 비슷한가를 정량화하는 지표로 특정 숙주에서 바이러스의 발현 수준을 예측할 수 있다(Sharp & Li, 1987). 계산 방식은 다음과 같다.

$$CAI = (\prod_{k=1}^{L} w_k)^{\frac{1}{L}} = \exp \frac{1}{L} \sum_{k=1}^{L} \ln w_k \qquad w_{ij} = \frac{x_{ij}}{x_{i\max}}$$

Xij 는 레퍼런스 셋의 i 아미노산의 j 번째 코돈의 빈도수를 의미하고 Ximax 는 레퍼런스 셋의 i 아미노산의 동의 코돈 중 가장 풍부한 코돈의 빈도수를 의미한다. CAI값은 바이러스 유전자 서열에

따라 레퍼런스 셋으로부터 얻은 코돈의 Wij값을 이용하여 기하학적 평균을 구한 값이다. CAI의 범위는 0부터 1까지이고 값이 클수록 숙주에서의 발현 정도가 클 것으로 기대된다. 표준 유전자 코드를 따를 경우 6개 동의 코돈을 갖는 아미노산 Arg, Leu, Ser은 각각 두 개의 코돈 패밀리에 의하여 암호화되므로 Wij는 두 코돈 패밀리별로 따로 계산한다(Xia, 2007). SIV의 40여종 숙주에 대한 코돈 사용 테이블 정보가 부족하여 HIV1, HIV2, SIV에 대하여 모두 Homo sapiens를 레퍼런스로 하였다. CAI값에서 쿼리 서열의 GC 및 아미노산 조성의 영향을 제외하기 위하여 HIV1, HIV2, SIV그룹별로 Markov 방법으로 500개의 랜덤서열을 생성하고 Kolmogorov-Smirnov 테스트로 생성된 서열의 eCAI 값이 정규분포를 따르는지 확인하고 유의수준 0.05에서 단측 검정의 임계값을 산출하였다. eCAI값의 평균 및 임계값 산출은 CAIcal 웹 서버를 이용하였으며, 다음으로 실제 그룹별 CAI 값이 eCAI값보다 큰지 즉 실제로 번역압력을 받고 있는지 평가 후, Wilcoxon & Mann-Whitney 테스트를 통해 그룹별 eCAI 평균값으로 센터링한 값을 비교하여 통계적으로 차이가 있는지를 확인하였다(Puigbò et al., 2008). CAIcal 웹 서버를 이용하여 eCAI값 계산시 쿼리 서열이 너무 클 경우 계산이 불가하므로 HIV1의 세 유전자(gag, pol, env)와 SIVmacaca의 env 유전자의 경우 랜덤으로 500개 서열을 추출하여 쿼리서열로 입력하였다.

**Similarity D(A, B)**

2013년 Zhou 등에 의하여 제기된 유사성 지수 D(A, B)는 Met와 Trp를 제외한 바이러스와 숙주의 59가지 코돈 사용을 59 차원에서의 공간 벡터로 간주하여 형성된 각의 코사인 값으로 바이러스와 숙주 코돈 사용의 유사성을 정량화하였다.

$$R(A,B)=\frac{\sum_{i=1}^{59} a_i \times b_i}{\sqrt{\sum_{i=1}^{59} a_i^2 \times \sum_{i=1}^{59} b_i^2}} \qquad D(A,B)=\frac{1-R(A,B)}{2}$$

위의 산출 공식에서 ai는 관심있는 바이러스 ORF의 59개 코돈에 대한 RSCU값이고 bi는 숙주에서의 동일한 코돈의 RSCU값이다. R(A, B)값은 두 공간 벡터의 코사인 값이고 D(A, B)는 이 값을 0에서 0.5 범위로 고정한다. RSCU값(≥0)에 따라 R(A, B)와 D(A, B)의 값은 각각 0~1과 0~0.5의 범위를 가진다. 해당 수식으로부터 바이러스와 숙주의 코돈 사용이 유사할수록 R(A, B)는 커지고 반대로 D(A, B)의 값은 작아진다. 현재 similarity D(A, B)를 활용하는 많은 연구들에서 D(A, B)값의 해석에 문제가 있음이 확인되었으며, 2020년 Silverj & Rota-Stabelli의 연구에서도 이 부분의 문제점을 지적한 바 있다. 또한 D(A, B)값의 해석을 숙주의 코돈 사용이 바이러스의 코돈 사용 형성에 미치는 영향으로 해석하고 있는데 수식으로부터 코돈 사용 유사성을 평가하는 지표로 CAI와 같이 숙주의 코돈 사용에 대한 적응정도 평가 지표로 사용하는 것이 더 적절하다. Similarity D(A, B) 인덱스는 해석 면에서만 주의하면 바이러스와 숙주의 코돈 사용 유사성을 정량화하는 좋은 지표로 활용될수 있을 것이다. 본 연구에서 세 종의 바이러스 모두 인간을 숙주로 D(A, B) 지표를 계산하고 인간의 59개 코돈의 RSCU값은 CUD에서 제공된 코돈 사용 테이블로부터 산출된 것이다. CUIs 프로그램에 해당 알고리즘이 내재되어 Homo sapiens를 검색하고 선택하면 바로 D(A, B)값을 얻을 수 있다. 다음 D(A, B)값과 CAI값에 상관관계가 있는지 확인하고 HIV1과 HIV2그룹의 D(A, B)값을 비교 분석하였으며, HIV1과 HIV2그룹 각각의 기원 바이러스인 SIVcpz_gor그룹 및 SIVsm그룹과 비교분석을 진행하였다.

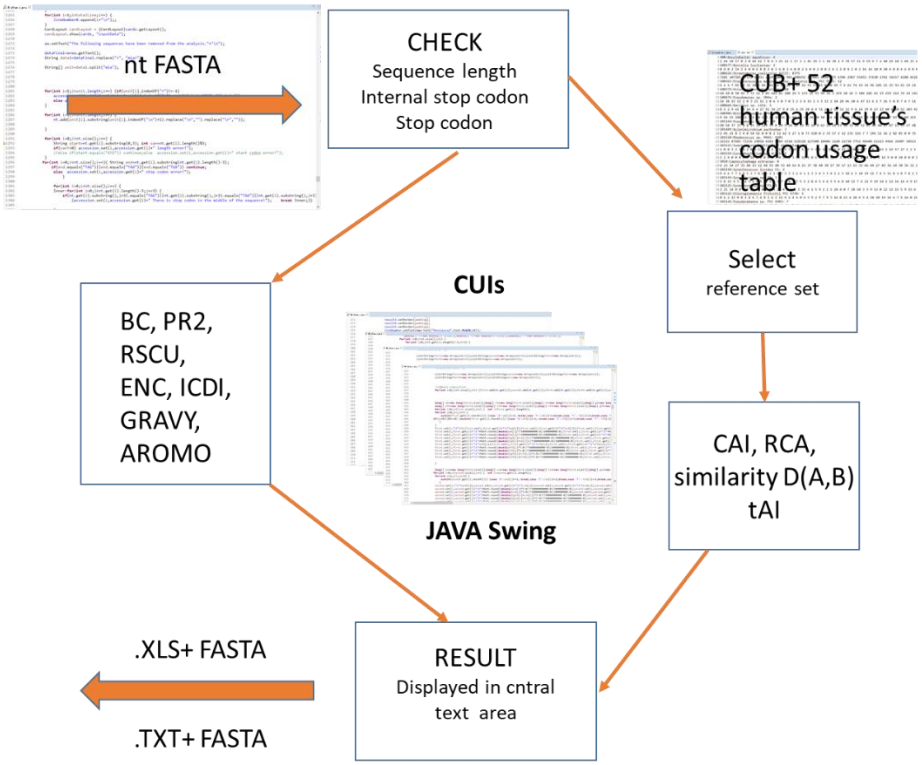


**Figure 2.1 Data flow diagram for CUIs program.** The CUIs is a graphical user program written by JAVA swing. Workflow is divided into two parts, one requires a reference set and the other can calculate the result simply by input sequences.

**Table 2.1 Program composition and calculation formula of each index**

| Categories | Indices | Component or formulae | Ref |
|---|---|---|---|
| Base composition | / | length, A, C, T, G, A%, C%, T%, G%, GC%, GA%, GT%, AT%, AC%, CT% (also respectively for the 1st, 2nd and 3rd codon position) GC1s%, GC2s%, GC3s% (‘s’ means synonymous sites namely do not consider Met and Trp) | |
| PR2 | / | A3/(A3+T3), G3/(G3+C3) for all four-codon amino acid in total and for specific four-codon amino acid | |
| Codon usage bias of specific codon | RSCU | $RSCU_{ij} = \frac{x_{ij}}{\frac{i}{n_i}\sum_{j=1}^{n_i} x_{ij}}$ | (Sharp & Li, 1986) |
| Overall codon usage bias of specific gene | ENC | $\mathrm{F} = \frac{n\sum_{i=1}^{k} {p_i}^2 - 1}{n-1}$ n>1 $p_i = \frac{n_i}{n}$ $\mathrm{ENC} = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}$ | (Wright, 1990) |
| | ICDI | $S_k = \sum (n_i - 1)^2 / k(k-1)$ $\mathrm{ICDI} = (\sum S_2 + S_3 + \sum S_4 + \sum S_6)/18$ | (Freire-Picos, 1994) |
| Codon adaptation to codon usage of host or tRNA pool of host cell | CAI | $w_{ij} = \frac{x_{ij}}{x_{imax}}$ $\mathrm{CAI} = (\prod_{k=1}^{L} w_k)^{\frac{1}{L}} = exp\frac{1}{L}\sum_{k=1}^{L} \ln w_k$ | (Sharp & Li, 1987) |
| | tAI | $W_i = \sum_{j=1}^{n_i}(1 - s_{ij}) \cdot tGCN_{ij}$ $\mathrm{tAI} = (\prod_{i=1}^{L} W_i)^{1/L}$ | (Puigbò et al., 2008) |
| | RCA | $RCA_{xyz} = \frac{X(x,y,z)}{X_1(x)X_2(y)X_3(z)}$ $\mathrm{RCA} = (\prod_{i=1}^{L} RCA_{xyz}(i))^{1/L}$ | (Fox & Erill, 2010) |
| | similarity D (A, B) | $\mathrm{R(A,B)} = \frac{\sum_{i=1}^{59} a_i \times b_i}{\sqrt{\sum_{i=1}^{59} a_i^2 \times \sum_{i=1}^{59} b_i^2}}$ $\mathrm{D(A,B)} = \frac{1 - R(A,B)}{2}$ | (Zhou et al., 2013) |
| Hydrophobicity and aromaticity of protein | GRAVY | $\mathrm{GRAVY} = (\sum_{i=1}^{L} H_i)/L$ | (Kyte & Doolittle, 1982). |
| | AROMO | $\mathrm{AROMO} = \frac{f_{Phe} + f_{Tyr} + f_{Trp}}{L}$ | (Lobry & Gautier, 1994) |

**Table 2.2 SIV host information**

| Subfamily | Tribe | Genus | Species | Common name | Abbreviation |
|---|---|---|---|---|---|
| Cercopithecinae | Cercopithecini | Allochrocebus | lhoesti | L'Hoest monkey | SIVlhoest*^° |
| | | | solatus | sun-tailed monkey | SIVsun*^° |
| | | Chlorocebus | aethiops | grivet | SIVagm-gr*^° |
| | | | cynosuros | malbrouck | SIVagm-ver*^° |
| | | | sabaeus | green monkey | SIVagm-sab*^° |
| | | | tantalus | tantalus monkey | SIVagm-tan* |
| | | Cercopithecus | albogularis | sykesmonkey | SIVsyk*^° |
| | | | ascanius | red-tailed monkey | SIVasc*^ |
| | | | ascanius_whitesidei | yellow-nosed red-tailed monkey | SIVasw* |
| | | | cephus | mustached guenon | SIVmus*^° |
| | | | mitis | blue monkey | SIVblu* |
| | | | mona | mona monkey | SIVmon*^° |
| | | | neglectus | de Brazza's monkey | SIVdeb*^° |
| | | | nictitans | greater spot-nosed monkey | SIVgsn*^° |
| | | Miopithecus | ogouensis | Gabon talapoin | SIVtal*^° |
| | Papioini | Cercocebus | agilis | agile mangabey | SIVagi*^ |
| | | | atys | sooty mangabey | SIVsm*^° |
| | | | torquatus | red-capped mangabey | SIVrcm*^° |
| | | Macaca | arctoides | stump-tailed macaque | SIVstm*^ |
| | | | fascicularis | crab-eating macaque | SIVcem*^° |
| | | | mulatta | rhesus macaque | SIVmac*^° |
| | | | nemestrina | pig-tailed macaque | SIVmne*^° |

**Table 2.2 (continued) SIV host information**

| Subfamily | Tribe | Genus | Species | Common name | Abbreviation |
|---|---|---|---|---|---|
| Cercopithecinae | Papioini | Mandrillus | sphinx | mandrill | SIVmnd-1, SIVmnd-2*^° |
| | | | leucophaeus | drill | SIVdrl*^° |
| Colobinae | | Colobus | guereza | mantled guereza | SIVcol*^° |
| | | Piliocolobus | badius_badius | western red colobus | SIVwrc*^° |
| | | | badius_temminckii | Temminck's red colobus | SIVtrc*^ |
| | | Procolobus | verus | olive colobus | SIVolc*^ |
| Homininae | Gorillini | Gorilla | gorilla gorilla | Western lowland gorilla | SIVgor*^° |
| | Hominini | Pan | troglodytes_schweinfurthii | Eastern chimpanzee | SIVcpz-eas*^° |
| | | | troglodytes_troglodytes | Central African chimpanzee | SIVcpz-cen*^° |
| | | | troglodytes_verus | West African chimpanzee | SIVcpz-wes^° |

Species are shown by the Latin name.

* means gag gene contains that species; ^means pol gene contains that species; °means env gene contains that species

**Table 2.3 The number of collected data for codon usage indices analysis**

| Virus | Polyprotein | The number of data |
|---|---|---|
| HIV1 | gag | 22,699 |
| | pol | 11,021 |
| | env | 70,658 |
| HIV2 | gag | 76 |
| | pol | 45 |
| | env | 170 |
| SIV | gag | 750 |
| | pol | 609 |
| | env | 15,416 |

HIV-1: Human immunodeficiency virus 1; HIV-2: Human immunodeficiency virus 2;
SIV: Simian immunodeficiency virus;

# 제 3 장. 연구 결과

## 3.1 코돈 사용 인덱스 산출 프로그램

본 연구를 통해 개발한 코돈 인덱스 산출 프로그램 CUIs는 기존 코돈 사용 인덱스 계산 프로그램의 불편한 사용성, 인덱스 종류의 한계점 및 웹 서버 기반 어플리케이션의 불안정성 문제를 해결한 GUI 기반 스탠드얼론 프로그램으로, Codon Usage Indices의 약자를 따서 명명하였다. CUIs 프로그램은 RSCU, CAI, RCA, similarity D(A, B), ENC, ICDI, GRAVY, AROMO 인덱스 산출과 함께 PR2 분석을 아미노산 별로 진행가능할 수 있도록 아미노산별 AT편향값 (A3/(A3+T3))과 GC편향값(G3/(G3+C3))을 제시하였다. PR2 분석의 경우 4코돈 아미노산 전체에 대하여 유전자의 전반적인 PR2편향을 확인한 연구들이 있다(Nasrullah et al., 2015). 그러나 PR2 편향은 아미노산 특이적이므로 유전자에 대하여 전반적인 PR2편향만을 보게되면 상반된 결과값이 중화되어 편향의 정도가 작게 측정되거나 또는 방향성까지 문제가 생길수 있다. 따라서 CUIs에서는 아미노산별로 분석을 진행할 수 있도록 PR2 plot의 X축, Y축 값인 G3/(G3+C3), A3/(A3+T3)를 4코돈 전체와 아미노산별로 제시하였다. 그리고 뉴클레오티드 조성의 경우 기존 프로그램에서는 GC3s 만을 계산해주는 경우가 많은데 CUIs에서는 중립 진화 분석시 필요한 GC12s 값과 GC1s, GC2s 값도 함께 제시해주고 있다. 프로그램 메인 화면은 border 레이아웃을 사용하여 상단은 열기, 저장, 인덱스 계산 등 기능 버튼으로 구성되어 있고 중간은 큰 텍스트 구역으로 불러온 데이터와 산출된 결과값 전시의 역할을

하며 하단의 작은 텍스트 구역은 메시지 창 역할로 배제 서열 열거, 경고 메시지, 저장 경로 등을 제시한다. 다음 서열데이터 중 문제 서열이 있을 경우 기존의 CodonW에서는 경고 메시지 창은 띄워주지만 문제 서열의 인덱스도 같이 산출해주고 있고 CAIcal에서는 문제 서열 중 마지막 서열의 헤드 라인만 제시해주고 모든 서열에 대한 인덱스 계산은 실행되지 않고 있다. 실제 활용에서 코딩 작업을 자유자제로 수행하지 못하는 사용자일 경우 방대한 서열데이터중 경고 메시지 창의 문제서열을 일일이 찾아서 FASTA파일로부터 제거하는 것은 쉬운일이 아니므로 본 연구에서 개발된 프로그램은 이런 작업을 대신해주고 있다. Figure3.1와 같이 데이터 서열체크를 진행하여 문제 서열을 인덱스 계산에서 제외시키고 제외된 서열 헤드라인을 메시지 창에 제시하여 확인할 수 있도록 하였다. 그리고 save 버튼 클릭시 문제서열이 제거된 FASTA 파일이 원하는 경로에 자동으로 저장되므로 클리닝 도구로도 유용하게 쓰일 수 있다. 또한 현재 CAI 인덱스의 산출은 거의 모두 CAIcal 웹 서버로 이루어지고 있는데 그 이유는 레퍼런스 셋의 입력을 CUD에서 제공하고 있는 코돈 사용 테이블의 형식으로 받을 수 있게 함으로써 CAI인덱스의 활용성을 많이 높였기 때문이다. 하지만 해당 서버 기반 프로그램에서는 웹 페이지 상의 텍스트박스로 입력서열을 받고 있으며 데이터가 클 경우 입력자체가 실행되지 않는 문제점이 있어 현재 대용량 데이터의 CAI 인덱스 계산이 불가능한 상황이다. 또한 서버 오류로 인해 종종 접속이 불가하여 이용할 수 없는 경우가 생기고 있다. 이런 문제점을 해결하기 위하여 개발된 CUIs 프로그램은 Codon Usage Database에서 제공하는 35,799가지 생물의 코돈 사용 테이블을 프로그램 자체에 내재하였고 바이러스의 조직 친화성을 확인할 수 있게 52가지 인간 조직(tissue)의 코돈 사용 테이블 정보도

프로그램에 내재하여 CAI 인덱스의 활용성을 높였다. CAI | RCA | D(A, B) | tAI 버튼 클릭시 레퍼런스 셋 검색 대화창이 뜨고 생물 종의 라틴 이름을 검색하거나 인간 조직 이름을 검색 하여 원하는 레퍼런스 셋을 더블 클릭하면 인덱스 값이 바로 산출되도록 프로그래밍하였다. 또한 스탠드 얼론 프로그램으로서 서버환경의 안정성에 상관없이 항시 사용할 수 있고 179MB 크기의 대용량 서열데이터 처리도 수행할 수 있다. ICDI, similarity D(A, B), RCA 인덱스의 경우 현재 계산을 제공하고 있는 프로그램이 없는 상황이고 코돈 연구에 많이 사용되고 있는 인덱스이므로 CUIs 프로그램에 포함시켰다. 원하는 인덱스 계산 완료시 결과 값은 탭으로 분리된 형식으로 텍스트 구역에 제시되어 직접 복사하여 쓸 수 있도록 하였으며, 또한 Figure3.2에서 제시한바와 같이 save 버튼 클릭시 원하는 경로에 폴더가 생성되고 그 안에는 엑셀 또는 텍스트 형식의 결과 파일과 인덱스 계산에 쓰인 클리닝된 서열데이터의 FASTA파일이 저장된다. 이중 엑셀 파일은 base composition, PR2, RSCU, CAI_RCA_D_tAI, ENC_ICDI_GRAVY_ AROMO 다섯개의 시트별로 인덱스 결과값이 저장되어 다음 단계 분석의 편리성을 높혔다. 프로그램 상단 오른쪽 끝에 위치한 Help 버튼을 클릭하면 프로그램 사용에 대한 소개와 제공되는 인덱스에 관한 간단한 원리, 용도, 산출공식 및 출처가 제시되어 있다.

## 3.2 계통 발생 분석 결과

Gag, pol, env 유전자별 계통수 구축 결과(Figure3.3) HIV1과 HIV2 모두 각각의 기원으로 밝혀진 SIV 종들과 함께 클러스터링되는 양상을 확인할 수 있었다. 세가지 유전자 모두 HIV1는 SIVcpz 및 SIVgor와, HIV2는 SIVsm 및 SIVmacaca와 같은 clade를 형성하였다. SIVcpz는 SIVgsn(3')과 SIVrcm(5')가 재조합된 것으로 알려져 있는데(Sharp et al., 2005) 본 연구에서 구축한 계통수상에서도 env 유전자 트리에서 SIVgsn과 가장 가까이 위치하고 pol 유전자 트리에서 SIVrcm과 가까이 위치하는 것을 관찰할 수 있었다. 또한 세가지 유전자에서 숙주가 allochrocebus 속에 속해있는 SIVlhoest와 SIVsun이 SIVcolobus와 가까운 근연관계를 갖고 있는 것이 확인되었다. 세가지 유전자의 계통수를 수평으로 관찰할 때 gag와 pol 유전자 보다 env 유전자의 변이정도가 더 높은 것을 확인할 수 있었다.

계통분류학적으로 숙주 원숭이 종들이 밀접하게 관련된 대부분의 경우, 즉 같은 속(genus)에 해당되면 감염된 SIV도 가까운 근연관계를 보이고 있다. 따라서 아래 코돈 분석에서 SIV에 대한 그룹핑은 같은 속을 기본으로 하고 족(tribe)이나 아과(subfamily) 정보도 고려하여 가까운 근연관계를 확인 후 하나의 그룹으로 묶는 것을 원칙으로 하였다. 여기서 SIVmnd1와 SIVsm는 계통수 토폴러지(topology)상의 특이성을 고려하여 그룹핑을 수행하였다. SIVmnd1은 env 유전자에서만 SIVmnd2와 SIVdrl로 구성된 SIVmandrillus과 가깝게 위치하고 gag, pol 에서는 SIVmandrillus보다 SIVlhoest와 서열상 더 가깝다. 이는 기존 연구에서 제시한 내용과 일치하는 결과로 구축된 계통수의 정확성을 확인할 수가 있었다(Peeters & Courgnaud, 2002). 또한 SIVsm은 숙주가 같은

cercocebus속인 SIVagi와 SIVrcm보다 숙주가 macaca속인 SIVmacaca와 더 가깝다. 이 또한 SIVmac가 SIVsm으로부터 종간이동으로 전파된 기존 연구결과와 일치하였다(Apetrei et al., 2005). SIVmnd1은 complete 서열정보로 M27470 하나만 공개 및 제공되고 있고 SIVmandrillus와 gag, pol 서열상 차이가 많이 나므로 SIVmandrillus 그룹에서 제외하였다. SIVsm은 sooty mangabey에서 원숭이 면역 결핍증을 일으키지 않는 것이 명확하게 제시되어 있어 cercocebus 속으로부터 분리하여 따로 보고 나머지 SIVagi와 SIVrcm를 SIVcerco 그룹으로 분석하였다. 세 유전자에서 colobinae 아과의 아프리카 군에 속하여 있는 colobus속, piliocolobus속, procolobus속이 계통수상 같은 clade를 형성하고 있어 아래 코돈 분석 시 SIVcolobus로 그룹핑하였다. 또한 세 가지 유전자에서 SIVcpz와 SIVgor도 같은 homininae과로 HIV1과 clade를 형성하므로 SIVcpz_gor그룹으로 그룹핑하였다. Cercopithecus속과 miopithecus속은 같은 cercopithecini족에 속해 있고 계통수 상으로도 가깝게 위치하므로 SIVpithecus 그룹으로 그룹핑하였다. 마지막으로 chlorocebus속은 SIVagm그룹으로, allochrocebus속은 SIVallo그룹으로 하여 후속 분석을 진행하였다.

## 3.3 렌티바이러스의 코돈 사용 분석 결과

### 3.3.1 조성 분석 결과

세 유전자의 전반적인 뉴클레이티드 조성을 확인한 결과 gag, pol, env 유전자 모두 아데닌(A)이 가장 풍부하고 pol, env 유전자는 사이토신(C) 함량이 가장 적다. gag유전자의 SIVallo, SIVcerco, SIVcolobus 그룹은 사이토신(C) 함량이 가장 적지만 이 세 그룹을 제외한 8개 그룹에서는 티민(T) 함량이 가장 낮다. 세 유전자의 코돈 세 번째 위치의 뉴클레오티드 조성을 확인한 결과 모두 A3 함량이 가장 높지만 pol 유전자는 평균이 45.03%이고 gag는 38.76%, env 유전자는 33.9%로 차이가 많이 나고 있다. 이중 gag는 HIV1의 A3 함량이 가장 높고 pol, env유전자는 모두 SIVcpz_gor 그룹의 A3 함량이 가장 높다. 세 유전자 모두 A3이 T3보다 많이 사용되고 env 유전자의 HIV2그룹을 제외한 모든 그룹에서 G3이 C3 보다 많이 사용되고 있다. Figure3.4의 세 번째 그림에서 유전자별 AT, GC편향과 AT3, GC3편향을 확인할 수 있는데 세 유전자의 그룹 모두 AT가 풍부하고 전반적으로 AT편향이 심하면 AT3 편향은 더 심한 패턴을 보이고 있지만 env 유전자의 HIV2과 SIVpithecus그룹은 상반된 패턴을 보이고 있다. 따라서 이 두 그룹에서의 전반적인 AT편향은 주로 코돈 첫 번째와 두 번째 자리의 AT조성에 의해 형성된 것으로 보인다. Pol과 env 유전자의 AT3 편향의 상위 4위를 확인해 보면 모두 HIV1, SIVcerco, SIVcpz_gor, SIVcolobus를 포함한다. Gag 유전자의 경우 SIVallo가 AT3 편향이 가장 크고 다음으로 HIV1, SIVcerco, SIVcpz_gor가 크게 나타난다. 반면, AT3 편향이 가장 작은 그룹은 pol과 env 유전자 모두 HIV2와 SIVpithecus로 확인된다. Gag 유전자는 SIVmandrillus와 SIVpithecus 그룹의 AT3 편향이 가장 작고 다음으로 HIV2가 작게 나타난다. 전반적으로 HIV1와 SIVcpz_gor는

세 유전자에서 모두 HIV2에 비해 높은 AT3 편향을 관찰할 수 있다.

### 3.3.2 패리티 규칙 2 분석 결과

우선 Arg 아미노산의 경우 세 유전자의 11개 그룹에서 AGA와 AGG 코돈 RSCU값의 평균이 각각 3.46와 1.86로 과표현되어 나머지 4개 코돈의 빈도가 상당히 낮아 서열별 Arg의 AT 편향 값이 0 또는 1로 대표성을 띠지 않으므로, 서열 평균 AT편향과 GC편향 값을 이용한 Arg의 PR2 plot의 결과는 신뢰할 수 없어 결과 해석에서 제외한다. 나머지 4코돈 아미노산의 PR2 분석 결과를 보면(Figure 3.5) Leu 아미노산 이외의 6개 아미노산은 세 유전자 모두 대체적으로 유사한 편향 패턴을 보이고 있다. Val과 Gly의 경우 예외는 있지만 대체적으로 코돈 세번째 자리가 A 또는 G인 코돈을 사용하고 Pro, Thr, Ala, Ser은 A 또는 C를 사용한다. Pro, Thr, Ala, Ser 에서 G보다 C로 끝나는 코돈을 선호하는 것은 CpG 디뉴클리오티드의 사용을 최소화하기 위한 것으로 보이는데 이는 숙주의 방어작용을 회피하여 바이러스의 복제기능을 유지하기 위한 것이다. 실제로 HIV1에서 숙주세포의 zinc finger antiviral protein (ZAP)이 CpG 디뉴클리오티드가 높은 구역에 결합하여 바이러스의 복제를 억제하는 것이 확인되었다(Ficarelli et al., 2020). 유전자별로 보면, gag 유전자의 경우 Thr 아미노산에서 HIV1그룹은 ACG코돈의 사용을 거의 완전히 피하고 Gly 아미노산은 SIVsm그룹만 특이적으로 코돈 세 번째 자리가 G보다 C인 코돈을 편향하고 Ser 아미노산에서 SIVpithecus와 SIVmacaca는 다른 그룹과 다르게 C보다 G인 코돈을 편향한다. Pol 유전자의 Val 아미노산은 9개 그룹의 G3/(G3+C3) 값이 0.52에서 0.62 범위에 위치하는 반면 HIV2와 SIVmandrillus그룹이 각각 0.36과 0.79로 이 범위를 크게 벗어나고 있고 HIV2는 다른 그룹과 다르게 Val 아미노산에서 G 보다 C로

끝나는 GTC코돈을 편향한다. Env 유전자의 Ala 아미노산에서 SIVsm과 SIVmacaca그룹은 G3/(G3+C3) 값이 0.6이상으로 나머지 9개 그룹과 달리 C보다 G코돈을 많이 사용하고 있다. 이 두 그룹에서 특이적으로 CpG 디뉴클레오티드의 사용을 피하지 않는 것이 확인되었고 계통수 상으로 같은 분기군을 형성하고 있는 HIV2와 SIVcerco 그룹도 코돈 세 번째 자리가 C 또는 G인 코돈을 비슷하게 사용하고 있다. Leu 아미노산은 gag, pol 유전자에서는 GC편향이 각각 0.38에서0.8, 0.37에서 0.76 범위내에 다양하게 위치하여 있지만 env 유전자에서는 특이하게 A3/(A3+C3)과 G3/(G3+C3) 값이 (0.5, 0.5)에서 크게 벗어나지 않는 양상을 보이고 있다.

### 3.3.3 중립 진화 분석 결과

Figure 3.6에서 GC3s 함량과 GC12s함량의 관계로 바이러스들이 받고 있는 선택압력과 돌연변이 압력의 상대적 크기를 정량화하여 나타내고 있다. 이론적으로 회귀직선 기울기의 절대값은 0 부터 1의 범위내에 있고 0에 가까울수록 선택압력을 많이 받는다고 간주된다. 이것은 GC3s의 함량 변화에 따라 GC12s의 변화도 똑같이 크면 선택 압력을 적게 받고 있는 것을 의미하고 반면 GC3s의 큰 변화에 상관없이 GC12s 값이 작은 범위내에 국한되어 있다면 그 의미는 해당 바이러스가 높은 선택압력을 받고 있다는 의미이다. 우선 세 유전자별로 바이러스 그룹마다 상관분석을 진행하였다. Gag 유전자의 경우, 유의수준 0.05에서 HIV1 (r=-0.053, p=0), HIV2 (r=0.37, p=0.001), SIVcerco (r=-0.9746, p=0.0048), SIVmacaca (r=-0.2895, p=0), SIVmandrillus (r=-0.9302, p=0.0008), SIVpithecus (r=0.5435, p=0.0363)가 유의한 상관관계를 보였다. Pol 유전자에서는 HIV1 (r=0.2014, p=0), HIV2 (r=0.4803, p=0.0008), SIVagm (r=0.3819, p=0.015), SIVcolobus (r=0.7664, p=0.0445), SIVcpz_gor (r=-0.6292, p=0), SIVpithecus (r=0.589,

p=0.034), env 유전자에서는 HIV1 (r=0.0653, p=0), SIVagm (r=0.2926, p=0.0038), SIVmacaca (r=-0.6714, p=0), SIVpithecus (r=0.6066, p=0.0076), SIVsm (r=-0.6667, p=0)가 유의한 상관관계를 보였다. 세 유전자 모두 대체적으로 X축과 평행한 분포를 보이고 있는데 이는 영장류 렌티바이러스의 코돈 사용 형성에서 선택압력이 주된 역할을 하고 있다는 의미이다. Gag 유전자에서 HIV2를 제외한 그룹은 모두 기울기가 마이너스 값으로 음의 상관관계를 보이고 있다. 유전자 별로 바이러스 그룹에 대한 회귀직선은 Figure3.6에 제시되어 있다. 회귀직선의 기울기는 바이러스가 받고 있는 방향성 돌연변이 압력의 상대적 크기를 나타내고 있는데 gag 유전자의 경우 HIV1, HIV2, SIVcerco, SIVcolobus, SIVmacaca, SIVmandrillus, SIVpithecus 그룹에서 코돈 사용 패턴 형성에서 방향성 돌연변이 압력의 상대 기여도는 각각 2.55%, 17.25%, 41.07%, 4.68%, 33.28%, 13.94% 이다. 즉 gag 유전자에서 HIV1그룹이 받고 있는 선택압력이 97.45%로 가장 크고 다음으로 SIVmacaca 그룹이 95.32%, SIVpithecus 그룹이 86.06%, HIV2 그룹이 82.75%, SIVmandrillus 그룹이 66.72%, SIVcerco 그룹이 58.93% 이다. Pol 유전자에서는 SIVcpz_gor 그룹을 제외한 HIV1, HIV2, SIVagm, SIVcolobus, SIVpithecus 그룹의 기울기는 모두 0보다 큰 값으로, 양의 상관관계를 보이고 있다. Pol 유전자에서 HIV1, HIV2, SIVagm, SIVcolobus, SIVcpz_gor, SIVpithecus 그룹이 받고 있는 방향성 돌연변이 압력은 각각 7.26%, 41.36%, 0.01%, 17.51%, 19.38%, 7.83% 이다. 선택압력을 가장 크게 받고 있는 그룹은 SIVagm으로 99.99%, 다음으로 HIV1, SIVpithecus, SIVcolobus, SIVcpz_gor 그룹이 각각 92.74%, 92.17%, 82.49%, 80.62%의 선택압력을 받고 있으며, HIV2그룹이 58.64%로 선택압력을 가장 적게 받고 있다. Env 유전자에서 HIV1, SIVagm, SIVpithecus그룹은 GC3s, GC12s 값이 양의 상관관계를 보이고 SIVmacaca와 SIVsm 그룹은 음의 상관관계를

보이고 있다. Env 유전자에서 HIV1, SIVagm, SIVmacaca, SIVpithecus, SIVsm 그룹이 받고 있는 방향성 돌연변이 압력의 크기는 각각 3.06%, 9.18%, 15.24%, 7.97%, 40.67% 이고 선택압력을 가장 많이 받고 있는 그룹은 HIV1그룹으로 96.94%, 다음으로 SIVpithecus 그룹이 92.03%, SIVagm 그룹이 90.82%, SIVmacaca 그룹이 84.76%이며, SIVsm 그룹이 59.33%로 선택 압력을 가장 적게 받고 있다.

### 3.3.4 RSCU 분석 결과

RSCU는 유전자내의 특정 코돈에 대한 편향을 확인할 수 있는 지표로 통상적으로 값이 1.6보다 크면 해당 코돈이 과표현된 것으로 인지한다. 바이러스 그룹별 59개 코돈에 대한 모든 서열의 RSCU평균 값을 gag, pol, env 유전자 별로 Table3.1에 제시하였다. Gag 유전자에서 그룹 평균값이 1.6 이상인 코돈은 TTA, GTA, TCA, CCA, ACA, GCA, AGA, AGG, GGA 로 평균 RSCU 값은 각각 1.63, 1.69, 1.67, 2.07, 2.21, 2.07, 3.42, 1.91, 1.87 이다. Pol 유전자에서 과표현된 코돈은 TTA, GTA, TCA, CCA, ACA, GCA, AGA, AGG, GGA 이외 AGT가 추가되었고 평균 RSCU값은 각각 2.12, 2.11, 1.68, 2.11, 2.20, 2.41, 3.81, 1.76, 2.19, 1.67 이다. Env 유전자에서는 CCA, ACA, GCA, AGA, AGG, GGA 등의 코돈이 과표현되고 있으며 RSCU값은 각각 1.99, 1.93, 1.82, 3.16, 1.90, 1.77 이다. 세 유전자의 모든 코돈중 동일한 코돈내에 과표현된(RSCU>1.6) 바이러스 그룹과 저표현된(RSCU<0.6) 바이러스 그룹이 동시에 존재한 코돈은 오직 gag 유전자내의 CTG(Leu) 하나이고 HIV1그룹에서의 RSCU값은 0.51, SIVmacaca그룹에서는 1.68 이다. Gag 유전자에서 가장 과표현된 AGA 코돈 이외에 AGC코돈의 변이 범위가 가장 큰데 HIV1과 HIV2는 평균이 2.2로 가장 많이 과표현되고 SIVmacaca와 SIVsm은 평균값이 1.7로 그 뒤를 잇고 있으며, SIVallo 그룹과 SIVpithecus는 평균이 0.7로 AGC코돈을 적게

사용하고 있다. Figure3.7를 확인해보면 gag 유전자의TTA, ATA, GTA 코돈의 경우 HIV1그룹이 다른 그룹과 비해 RSCU값이 현저히 높다. 2코돈 아미노산 중 Tyr, Cys의 경우 HIV2 그룹의 동의코돈에 대한 평균 사용과 달리 HIV1그룹은 TAT, TGT 코돈의 사용에 편향되어 있다. Pol 유전자중 Leu 아미노산의 코돈 사용 측면에서는 HIV2 그룹이 가장 눈에 띄게 다르게 나타난다. HIV2를 제외한 10개 그룹에서 모두 TTA 코돈을 RSCU값 평균 2.18로 가장 많이 사용하고 있는데 HIV2에서는 TTA의 RSCU값이 1.52로 다른 그룹에 비해 상대적으로 낮고 CTA코돈은 기타 그룹이 평균 1.27에 집중적으로 위치하여 있는 반면 HIV2는 RSCU값이 2.27로 TTA코돈 보다 CTA 코돈을 훨씬 많이 사용하고 있다. Ser 아미노산의 코돈사용에서 TCA 또는 AGT코돈이 많이 사용되고 있는데 HIV2, SIVallo, SIVmacaca, SIVsm 그룹에서는 TCA코돈을 더 선호하고 나머지 그룹에서는 AGT코돈을 더 많이 사용하는 차이를 관찰할 수 있다. 다음 pol 유전자의 Cys 아미노산에서 대부분 큰 동의 코돈 사용 편향이 관찰되지 않고 있는데 HIV1그룹과 SIVcpz_gor 그룹에서 TGT는 과표현, TGC 코돈은 저표현된 패턴을 관찰할 수 있었다. Pol 유전자에서 Cys 를 제외한 2코돈 아미노산에서 모두 A또는 T로 끝나는 코돈을 선호하는 패턴이 관찰되었다. Env 유전자에서 Ser 유전자의 코돈사용은 바이러스 그룹별로 큰 차이를 보이고 있다. SIVmacaca (TCT를 가장 많이 사용)를 제외한 모든 그룹이 TCA또는 AGT를 가장 많이 사용하는데, HIV1, SIVagm, SIVallo, SIVcolobus그룹은 AGT를 선호한다. Env 유전자의 Cys 아미노산에서는 HIV2를 제외한 모든 그룹이 TGT코돈을 선호하고 HIV2그룹만 TGC코돈을 선호한다. Ile아미노산의 경우 pol, env 유전자에서 모두 ATA 코돈의 사용에 편향되어 있고, gag 유전자는 SIVcerco, SIVmacaca, SIVmandrillus, SIVsm 그룹에서만 ATT를

선호하고 나머지 7그룹 모두 ATA 코돈을 선호한다. Thr 아미노산의 경우 세 유전자 모두 ACA코돈과 ACG코돈이 각각 과표현/저표현되는 패턴을 보이고 있다. CCA, ACA, GCA, AGA, AGG, GGA 코돈이 세 유전자에서 모두 과표현되어 해당 아미노산 Pro, Thr, Ala, Arg, Gly도 세 유전자가 비슷한 패턴을 보이고 있다.

### 3.3.5 ENC, ICDI 분석 결과

Figure 3.8에서 gag, pol, env 유전자의 바이러스별 ENC값을 제시하였다. 세 유전자별 평균은 각각 47.80, 43.47, 49.18이고 범위는 44.27-52.52, 40.62-46.62, 46.53-53.42로 pol 유전자의 ENC 값이 가장 낮고 바이러스 그룹별 차이도 가장 적다. ENC값은 낮을수록 높은 코돈 사용 편향을 의미하므로 세 유전자중 pol 유전자의 코돈사용 편향이 상대적으로 가장 크고 env 유전자의 코돈 사용편향이 가장 낮다. Gag 유전자에서 SIVallo, SIVcolobus, HIV1그룹의 ENC값은 각각 44.27, 44.74, 44.75로 하위 3위를 차지하고 SIVmandrillus, SIVpithecus, SIVagm그룹은 52.52, 50.92, 50.63로 상위 3위를 차지한다. Pol 유전자의 하위 세 그룹은 SIVcolobus, HIV1, SIVcpz_gor로 평균은 각각 40.62, 41.53, 41.80 이고 상위 세 그룹은 SIVpithecus, SIVagm, SIVmandrillus로 평균은 각각 46.62, 45.98, 44.58 이다. Env 유전자중 HIV2, SIVagm, SIVmacaca, SIVpithecus, SIVsm 그룹의 ENC 값이 모두 50이상으로 코돈 사용이 무작위로 사용된 것으로 간주되어 낮은 코돈 사용 편향을 보이는 반면, HIV1, SIVcpz_gor, SIVcolobus그룹은 상대적으로 높은 코돈 사용 편향을 보이고 있다. 세 유전자에서 공통적으로 높은 코돈 편향을 보이고 있는 바이러스 그룹은 HIV1과 SIVclobus 그룹이고 낮은 코돈 편향을 보이고 있는 그룹은 SIVpithecus이다. HIV2와 SIVsm 그룹은 env 유전자에서만 특이적으로 낮은 코돈 사용 편향을 보이고 있다. ENC-GC3s plot은 GC함량

차이로 인한 ENC 값의 차이를 제거하기 위해 ENC 기대 곡선과의 수직 거리로 바이러스 그룹들이 받고 있는 선택 압력의 크기를 보여주고 있다. Figure3.9를 보면 gag유전자에서 SIVagm, SIVcpz_gor, SIVmandrillus 그룹이 GC3s% 변이가 작고 ENC값 변이가 큰 수직 형태의 분포를 보이고 있는데 이것은 세 그룹내에서 ENC 값이 작을수록 큰 선택 압력을 받는다는 의미이다. 또한 gag 유전자의 SIVpithecus 그룹 중 서열의 분포가 다양한데 이중 SIVasc가 SIVdeb 보다 큰 선택압력을 받고 있다. Pol 유전자의 경우 바이러스 그룹별로 받고 있는 선택압력의 정도가 가장 비슷한데 SIVagm은 전반적인 패턴과 달리 선택압력을 상대적으로 적게 받고 있는 것을 확인할 수 있고 또한 SIVagm그룹내에서 ENC값이 작을수록 더 큰 선택압력을 받는다. Pol 유전자에서 HIV1 그룹에 비해 HIV2 그룹이 조금 큰 선택 압력을 받고 있고 SIVallo 그룹에서 SIVsun이 SIVlhoest 보다 높은 선택압력을 받는다. Env 유전자의 ENC-GC3s plot을 보면 HIV1그룹이 두 그룹으로 갈려져 있는데 ENC 기대곡선에 위치하여 있는 그룹을 확인해본 결과 복제(clone) 서열로 높은 ENC값을 갖고 있다. env 유전자에서 SIVsm, SIVcolobus그룹이 GC3s% 변이는 적고 ENC값 차이는 큰 수직적 양상을 보이는데, 이는 두 그룹내에서 ENC 값이 작을수록 높은 선택압력을 받고 있다는 의미이다. SIVcolobus 데이터를 확인해본 결과 SIVwrc 가 SIVcol 보다 높은 선택압력을 받고 있는 것이 확인되었고 SIVallo 그룹에서는 SIVlhoest 가 SIVsun 보다 높은 선택 압력을 받고 있는 것을 관찰할 수 있었다. Env 유전자에서 SIVmacaca 그룹은 횡좌표와 평행한 GC3s% 변이가 크고 ENC 값의 변이가 작은 양상을 보이고 있으므로 이 그룹에서는 GC3s%값이 클수록 높은 선택 압력을 받는다. Env 유전자의 SIVpithecus그룹 내 서열들은 다양하게 위치하여 있는데 SIVtal가 다른 종에 비해 선택 압력을 많이 받고

있는 것이 확인되었다. Table3.2에 나타낸 ICDI값에서도 세 유전자의 모든 그룹의 값이 0.3보다 작으므로 전반적으로 낮은 코돈 사용 편향을 보인다. 세 유전자에서 공통으로 코돈 사용 편향이 상대적으로 높은 그룹은 HIV1과 SIVcolobus이고 코돈 사용 편향이 낮은 그룹은 ENC 결과와 달리 모두 HIV2 그룹과 SIVpithecus 그룹을 포함한다. ENC 및 ICDI 인덱스 분석 결과로부터 HIV2에 비해 HIV1의 코돈사용 편향이 높다는 결과를 얻을 수 있었으며 이는 기존 연구와 일치한다(Vidyavijayan et al., 2017). 유전자 별로 ENC, ICDI, CAI, similarity D(A, B) 의 Pearson 상관분석 결과 ENC와 ICDI는 음의 상관관계를 보이고, CAI과 similarity D(A, B) 또한 음의 상관관계를 보이고 있다(Table3.3). 또한 ICDI와 CAI값이 음의 상관관계를 보이고 ENC 또한 similarity D(A, B)와 음의 상관관계를 보이므로 유전자의 전반적인 코돈 사용 편향이 클수록 Homo sapiens의 코돈 사용과 낮은 유사성을 보인다.

### 3.3.6 CAI, similarity D(A,B) 분석 결과

**CAI**

CAI 값은 바이러스가 숙주의 코돈 사용에 대한 적응 정도를 평가하는 지표로 숙주 세포에서의 발현수준을 예측하는데 이용되고 있다. Table3.4를 보면 gag 유전자에서의 CAI 값은 그룹별 평균이 0.715에서 0.746 범위에 있고 SIVmacaca 그룹이 0.746±0.005로 레퍼런스 셋 Homo sapiens에 대한 가장 높은 적응도를 보이고 있다. Pol 유전자에서의 CAI값은 SIVmandrillus 그룹이 0.708±0.004로 가장 높고 HIV1 그룹이 0.691±0.004로 가장 낮다. env 유전자에서는 HIV2 그룹의 CAI 값이 0.730±0.005로 가장 높고 SIVcpz_gor 그룹이 0.700±0.006로 가장 낮다. 세 유전자중 pol 유전자에서의 그룹별 차이가 가장 적은 것으로 관찰되었고 gag 유전자에서의 그룹별

차이가 가장 높다. CAI의 절대값은 서열의 뉴클레오티드 및 아미노산 조성의 영향을 받으므로 실제로 번역 압력을 받고 있는지를 확인하기 위해 Puigbò 등의 방법으로 유의수준 0.05에서의 eCAI값을 산출하였다. 세 유전자의 그룹별 뉴클레오티드 및 아미노산 조성을 반영하여 500개 서열을 생성하였고 Kolmogorov-Smirnov 테스트로 정규성 검정을 실행한 결과 모든 그룹으로부터 생성된 500개 서열이 모두 정규분포를 따르고 있었다(<critical value 0.061). 다음 그룹별 실제 CAI값과 5% 결정수준에서의 eCAI값을 비교한 결과 pol, env 유전자에서 11개 그룹 모두 eCAI값의 평균과 유의한 차이가 없는 것이 확인되었다. 즉 pol, env 유전자에서의 높은 CAI값은 실제 번역 압력에 인하여 형성된 것이 아니라 GC함량 및 아미노산의 조성에 의하여 형성된 것일 수 있다는 의미이다. Gag 유전자에서 SIVallo, SIVcolobus, SIVcpz_gor, SIVmacaca, SIVsm그룹이 Homo sapiens의 코돈 사용에 적응된 것이 확인되었고 나머지 그룹은 모두 산출된 eCAI값의 평균과 유의한 차이를 관찰할 수 없었다. 즉 HIV1과 HIV2는 모든 유전자에서 Homo sapiens와의 코돈 사용 유사성은 높았지만 번역압력을 크게 받고 있지 않는 것이 확인되었다. Gag 유전자의 SIVallo, SIVcolobus, SIVcpz_gor, SIVmacaca, SIVsm 그룹마다 eCAI의 평균으로 센터링한 값을 두 그룹씩 Mann-Whitney 테스트를 한 결과 유의수준 0.05에서 SIVmacaca 그룹이 기타 4개 그룹보다 Homo sapiens에 대한 적응도가 유의하게 높았고 SIVsm 그룹이 SIVcpz_gor 그룹보다 Homo sapiens에 대한 적응도가 높았다.

**Similarity D(A,B)**

Similarity D(A, B)는 바이러스와 숙주의 59개 코돈의 RSCU값을 이용하여 코돈 사용 유사성을 정량화하고 있다. HIV1과 HIV2 각각

Homo sapiens와의 코돈 사용 유사성을 확인해본 결과 gag 유전자에서 HIV1의 similarity D(A, B)값은 0.117 ±0.006이고 HIV2는 0.086±0.008이다(Table3.5). HIV1, HIV2의 origin 인 SIVcpz_gor그룹과 SIVsm 그룹의 similarity D(A, B)값은 SIVcpz_gor와 SIVsm그룹에 대해 각각 0.095±0.008, 0.087±0.007이다. Mann-Whitney 테스트 결과 유의수준 0.05에서 HIV1값이 HIV2보다 유의하게 높았고 HIV1값 또한 이 바이러스의 기원인 SIVcpz_gor그룹보다 유의하게 높았다. 하지만 HIV2와 SIVsm그룹의 similarity D(A, B)값은 유의한 차이가 없는 것으로 확인되었다. Pol 유전자에서 Homo sapiens에 대한 HIV1, HIV2, SIVcpz_gor, SIVsm의 similarity D(A, B)값은 각각 0.132±0.005, 0.121±0.006, 0.132±0.006, 0.119±0.008 이다. Mann-Whitney 테스트 결과 유의수준 0.05에서 gag 유전자와 동일한 결과가 관찰되었다. Env 유전자에서 Homo sapiens에 대한 HIV1, HIV2, SIVcpz_gor, SIVsm의 similarity D(A, B)값은 각각 0.096±0.005, 0.063±0.005, 0.099±0.004, 0.077±0.005 이다. Mann-Whitney 테스트를 해본 결과 유의수준 0.05에서 similarity D(A, B)값은 HIV1그룹이 HIV2보다 유의하게 높았지만 HIV1그룹이 SIVcpz_gor그룹보다 유의하게 낮았고 HIV2그룹 또한 SIVsm그룹보다 유의하게 낮았다. 즉 렌티바이러스는 전반적으로 Homo sapiens의 코돈 사용과 높은 유사성을 보이고 env 유전자에서 가장 높은 유사성을 보이며 pol 유전자에서 상대적으로 낮은 유사성을 보인다. 또한 세 유전자에서 모두 HIV1에 비해 HIV2의 코돈 사용이 인간의 코돈 사용과 더 유사하다. Gag 유전자에서 HIV1은 기원바이러스인 SIVcpz_gor에 비하여 인간의 코돈 사용에 대한 유사성이 상대적으로 낮지만 env 유전자에서는 HIV1과 HIV2 모두 기원바이러스인 SIVcpz_gor와 SIVsm 보다 인간에 더 유사한 코돈 사용 패턴을 보이고 있다.

**Figure 3.1 Sequences check in CUIs program.** The error sequence's headline will be displayed in the bottom text area and will not be included in the indices calculation.

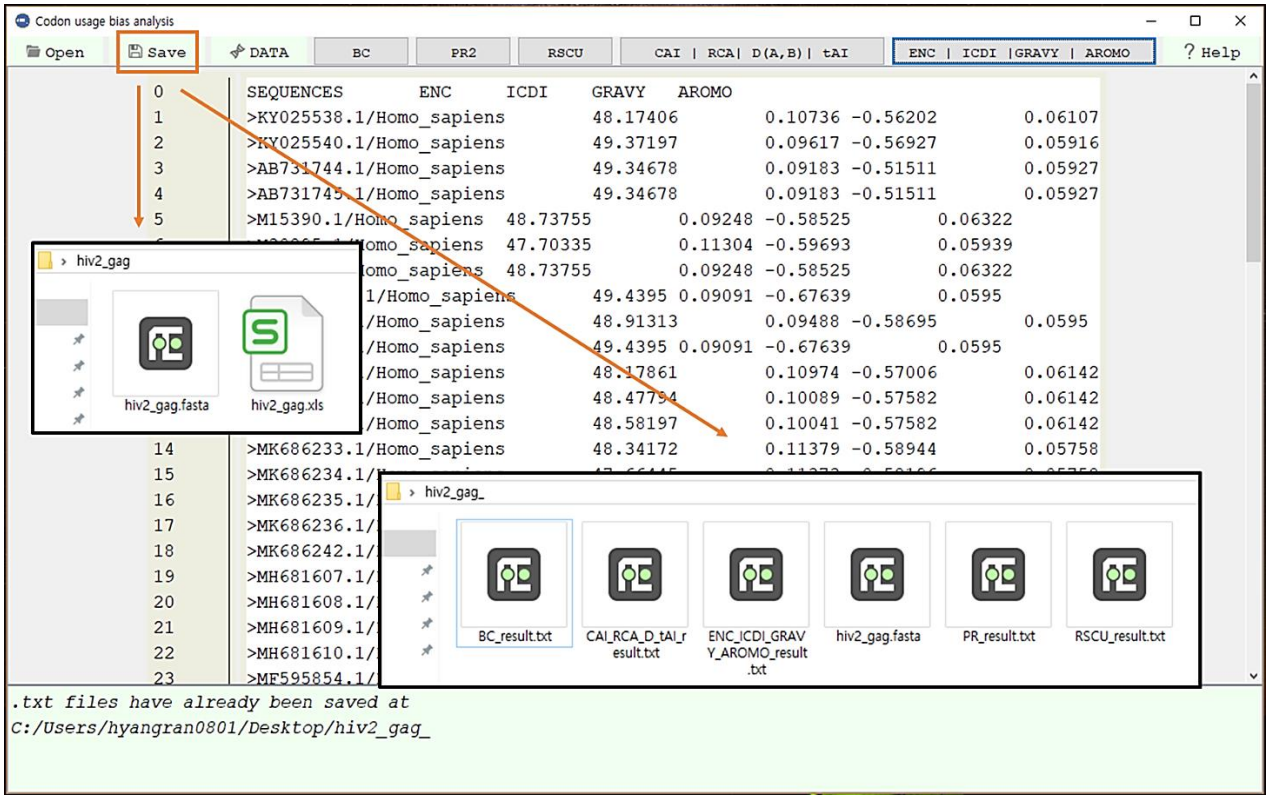


**Figure 3.2 Example of saving .xls or .txt file in CUIs program.** If you choose to output .xls file, a folder containing one xls file and one FASTA sequence file will be saved in the selected path or if you choose .txt file, there will be five txt files and one fasta sequence file appear in the slected path.

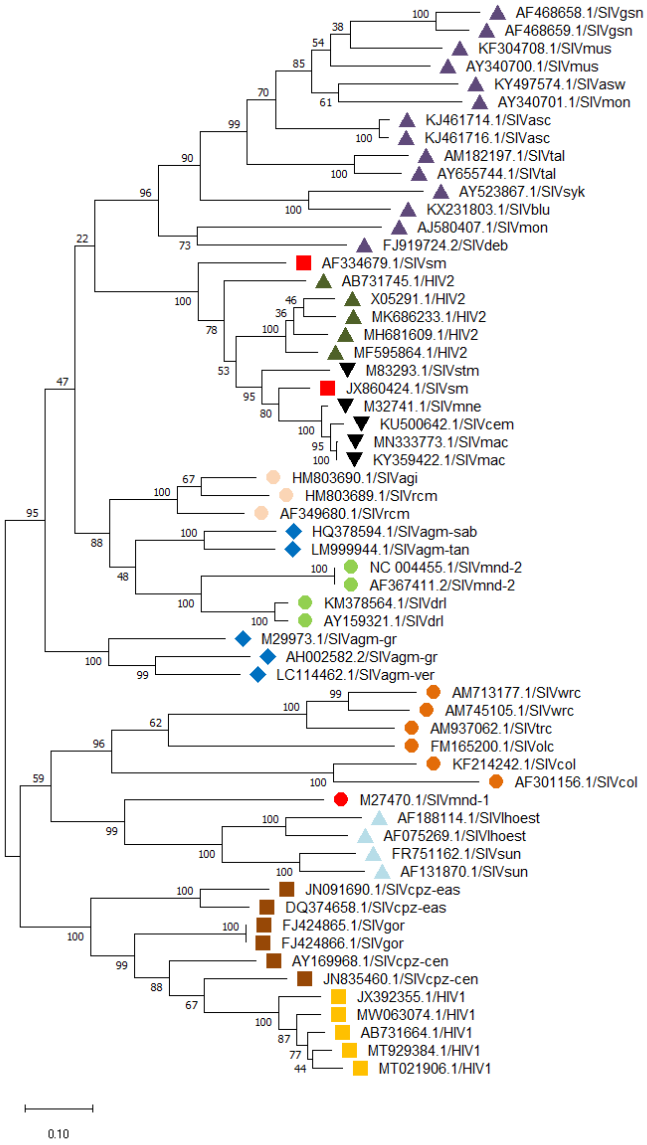


**Figure 3.3a Phylogenetic tree constructed with nucleotide sequences of lentiviral gag genes.** [■]HIV1; [▲]HIV2; [▲]SIVpithecus; [■]SIVsm; [▼]SIVmacaca; [●]SIVcerco; [◆]SIVagm; [●]SIVmandrillus; [●]SIVcolobus; [▲]SIVallo; [■]SIVcpz_gor;

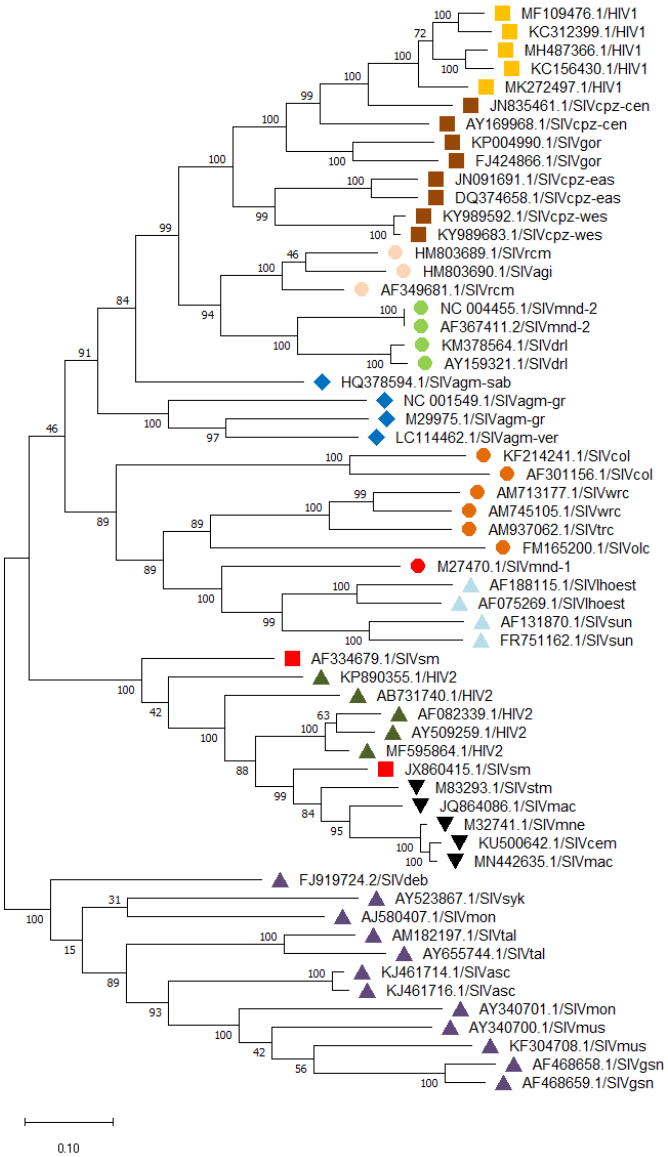


**Figure 3.3b Phylogenetic tree constructed with nucleotide sequences of lentiviral pol genes.** [■]HIV1; [▲]HIV2; [▲]SIVpithecus; [■]SIVsm; [▼]SIVmacaca; [●]SIVcerco; [◆]SIVagm; [●]SIVmandrillus; [●]SIVcolobus; [▲]SIVallo; [■]SIVcpz_gor;

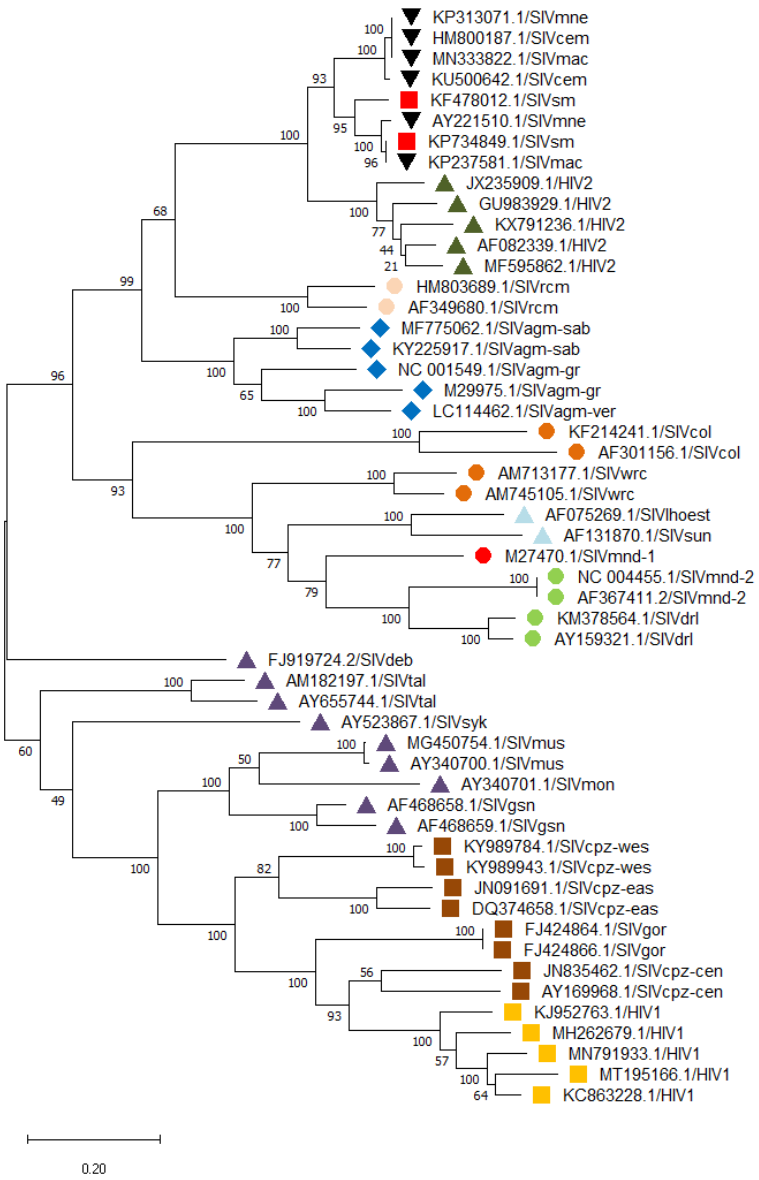


**Figure 3.3c Phylogenetic tree constructed with nucleotide sequences of lentiviral env genes.** [■]HIV1; [▲]HIV2; [▲]SIVpithecus; [■]SIVsm; [▼]SIVmacaca; [●]SIVcerco; [◆]SIVagm; [●]SIVmandrillus; [●]SIVcolobus; [▲]SIVallo; [■]SIVcpz_gor;

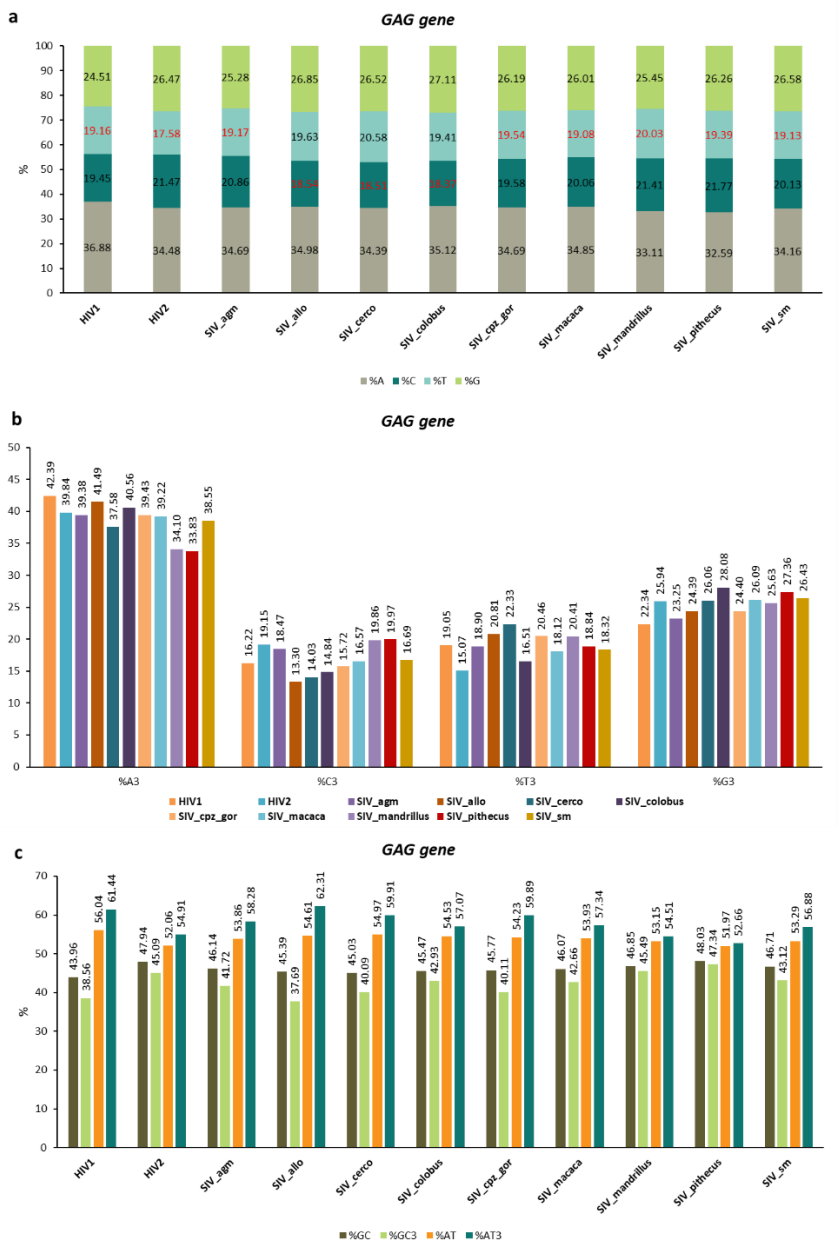


**Figure 3.4a Compositional features of lentiviral gag gene.** a: nucleotide distribution of A, T, C and G in the lentiviral gag genes. b: distribution of A, T, C and G only for the third codon position. c: total GC, AT content and GC, AT content at the third codon positon.

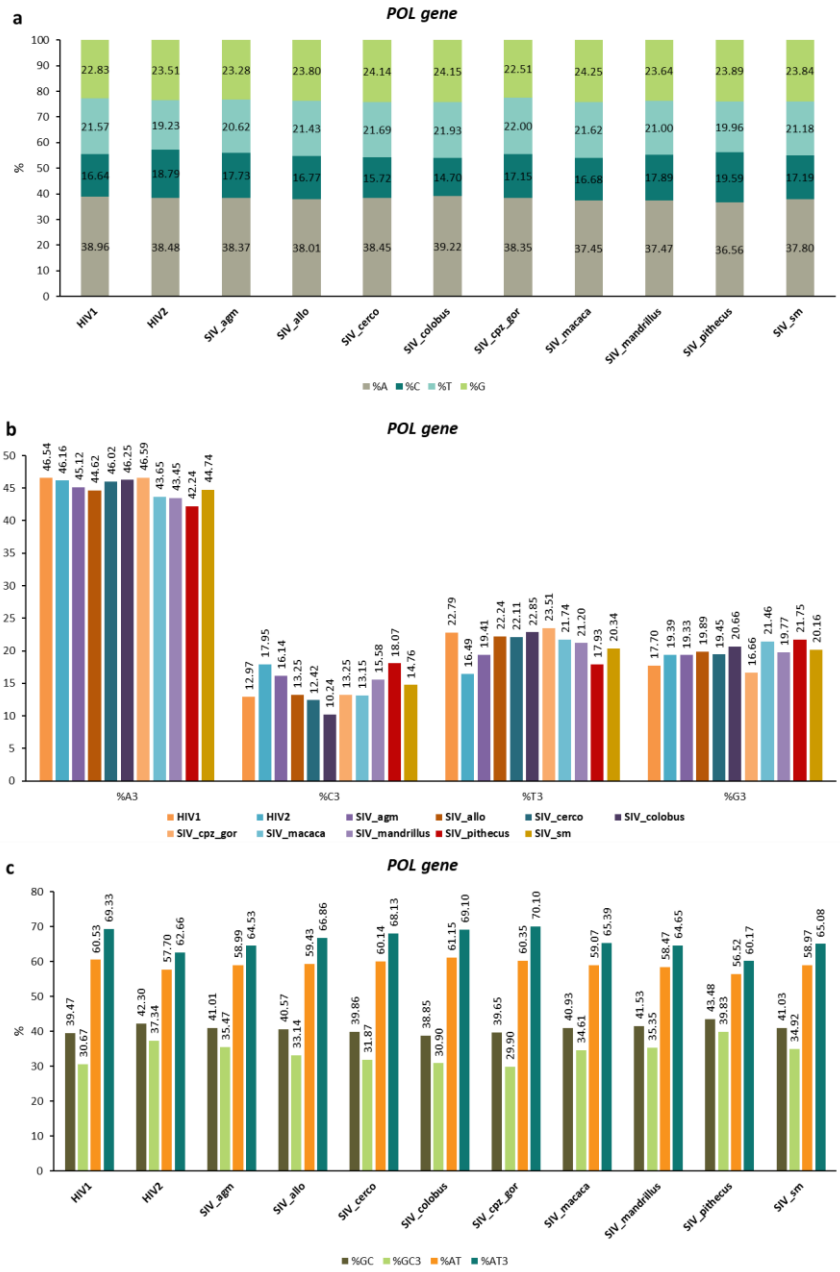


**Figure 3.4b Compositional features of lentiviral pol gene.** a: nucleotide distribution of A, T, C and G in the lentiviral gag genes. b: distribution of A, T, C and G only for the third codon position. c: total GC, AT content and GC, AT content at the third codon positon.

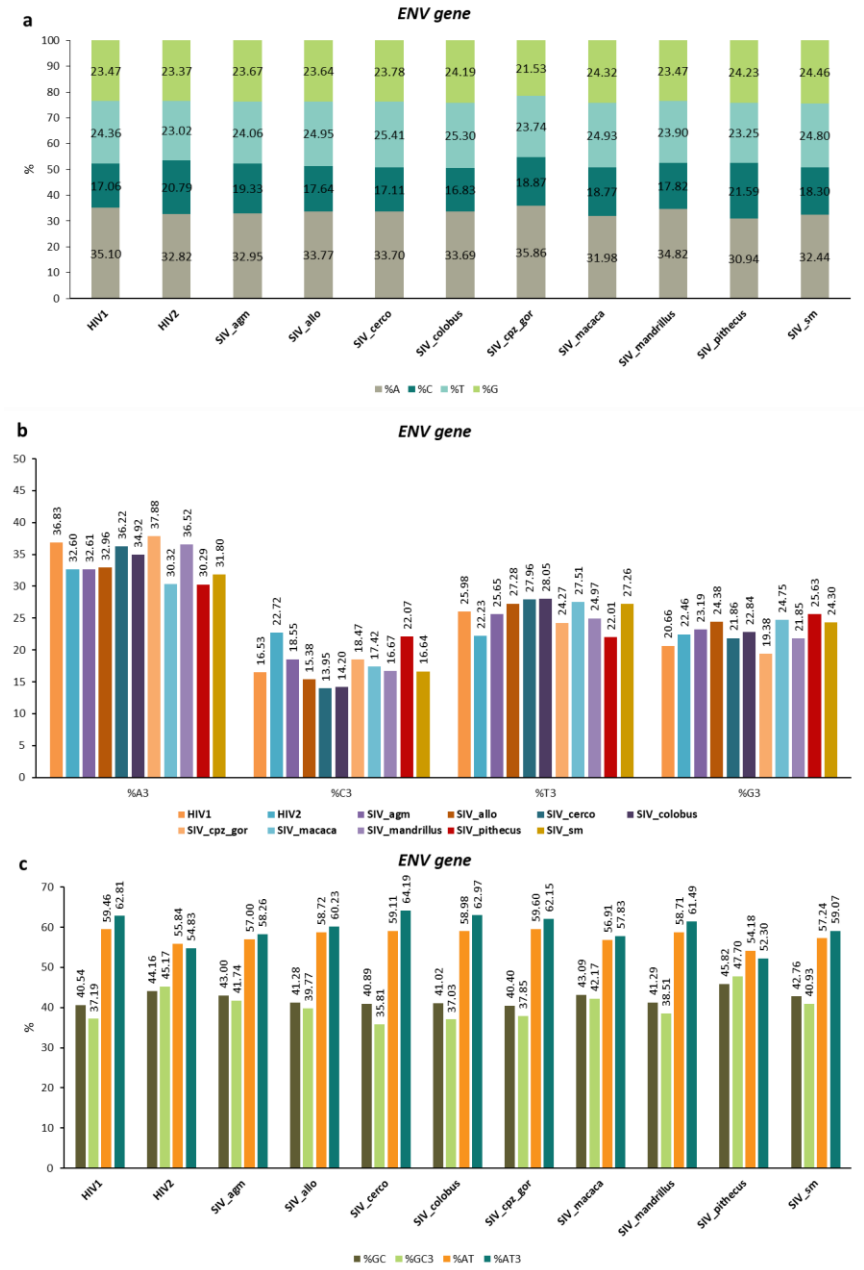


**Figure 3.4c Compositional features of lentiviral env gene.** a: nucleotide distribution of A, T, C and G in the lentiviral gag genes. b: distribution of A, T, C and G only for the third codon position. c: total GC, AT content and GC, AT content at the third codon positon.

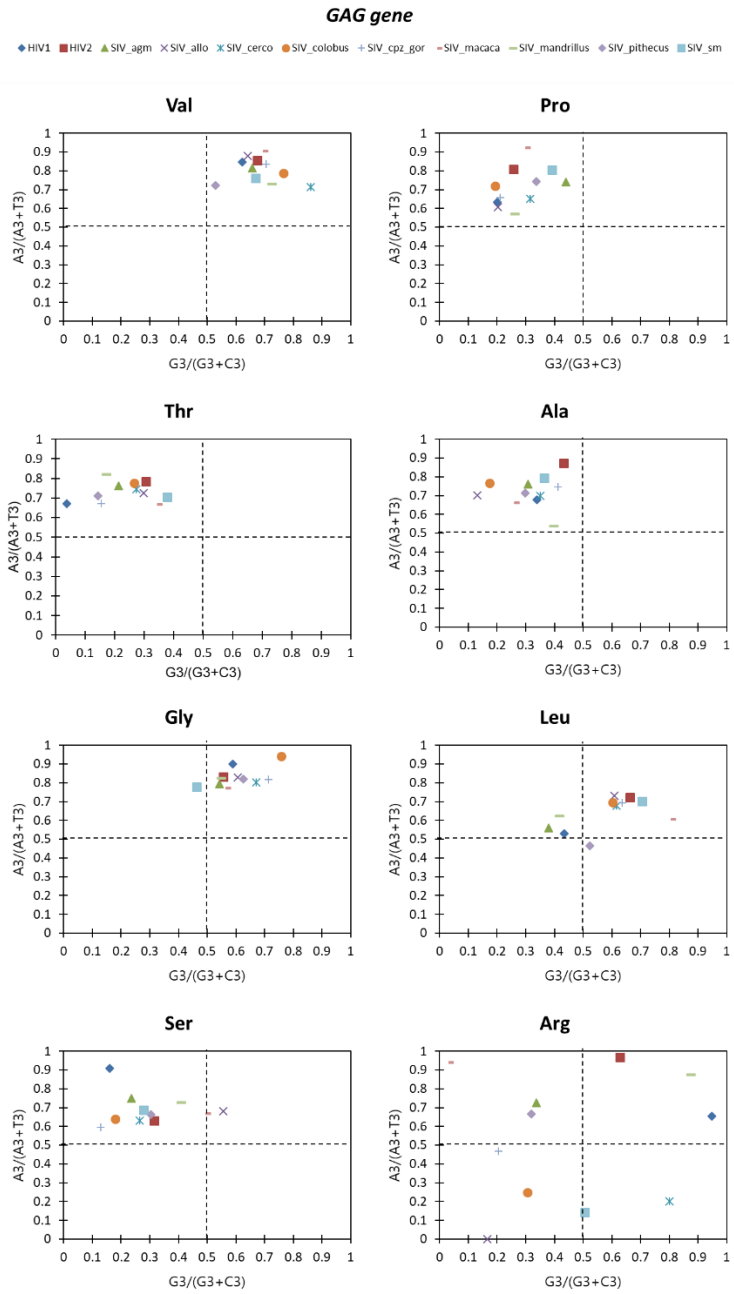


**Figure 3.5a Average PR2-bias plot of 11 lentiviral groups in gag genes.** ATbias(A3/(A3+T3)) and GCbias(G3/(G3+C3)) were calculated by each four-codon amino acids seperately.

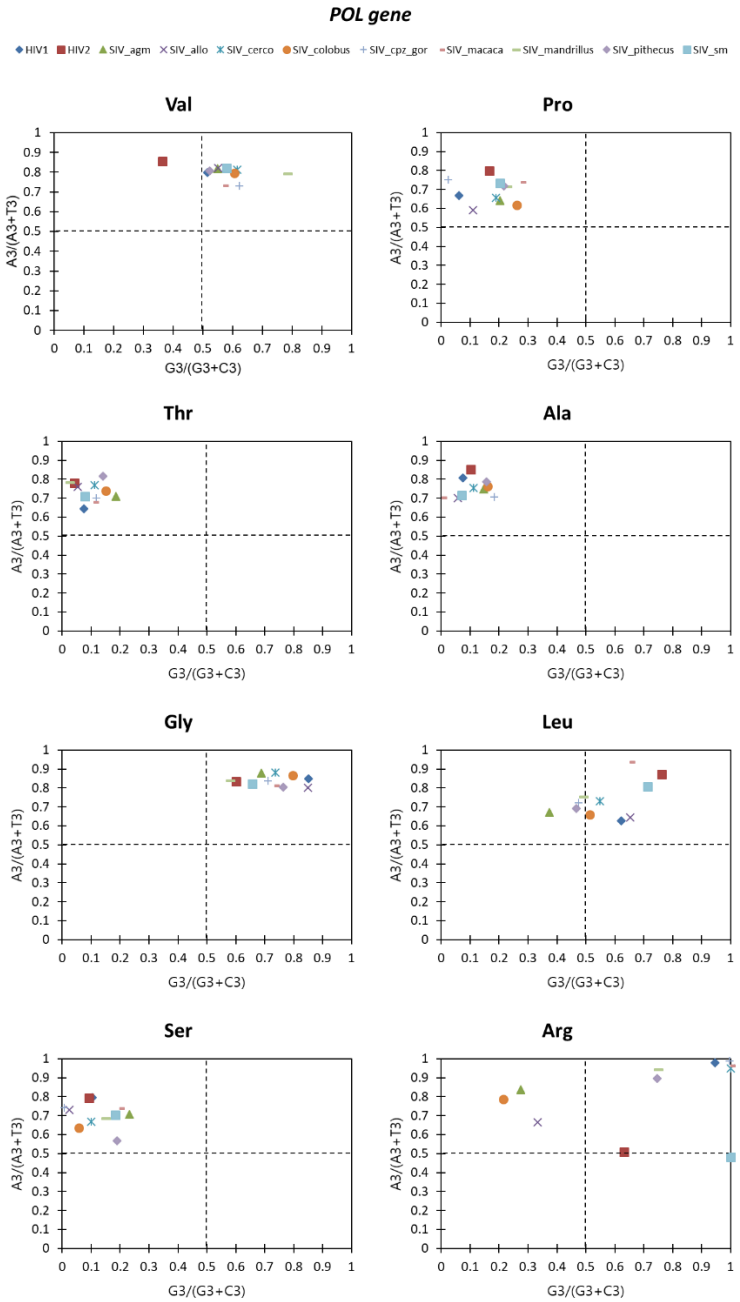


**Figure 3.5b Average PR2-bias plot of 11 lentiviral groups in pol genes.** ATbias(A3/(A3+T3)) and GCbias(G3/(G3+C3)) were calculated by each four-codon amino acids seperately.

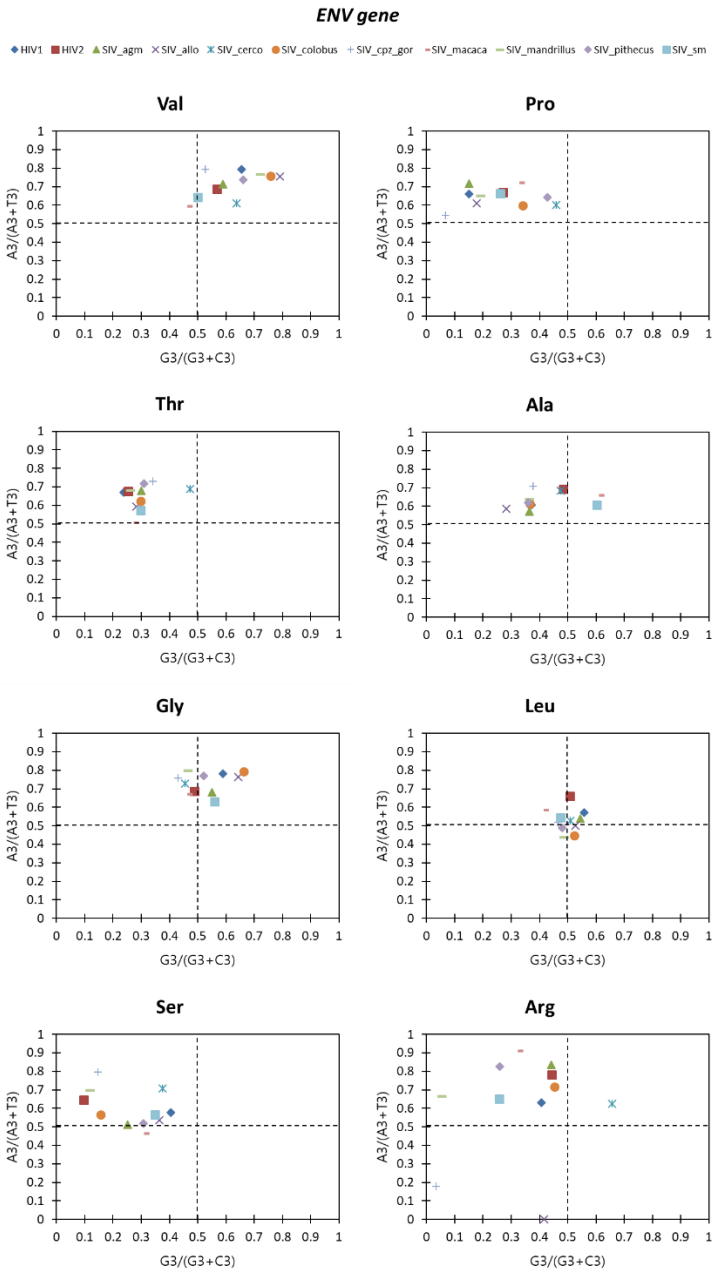


**Figure 3.5c Average PR2-bias plot of 11 lentiviral groups in env genes.** ATbias(A3/(A3+T3)) and GCbias(G3/(G3+C3)) were calculated by each four-codon amino acids seperately.

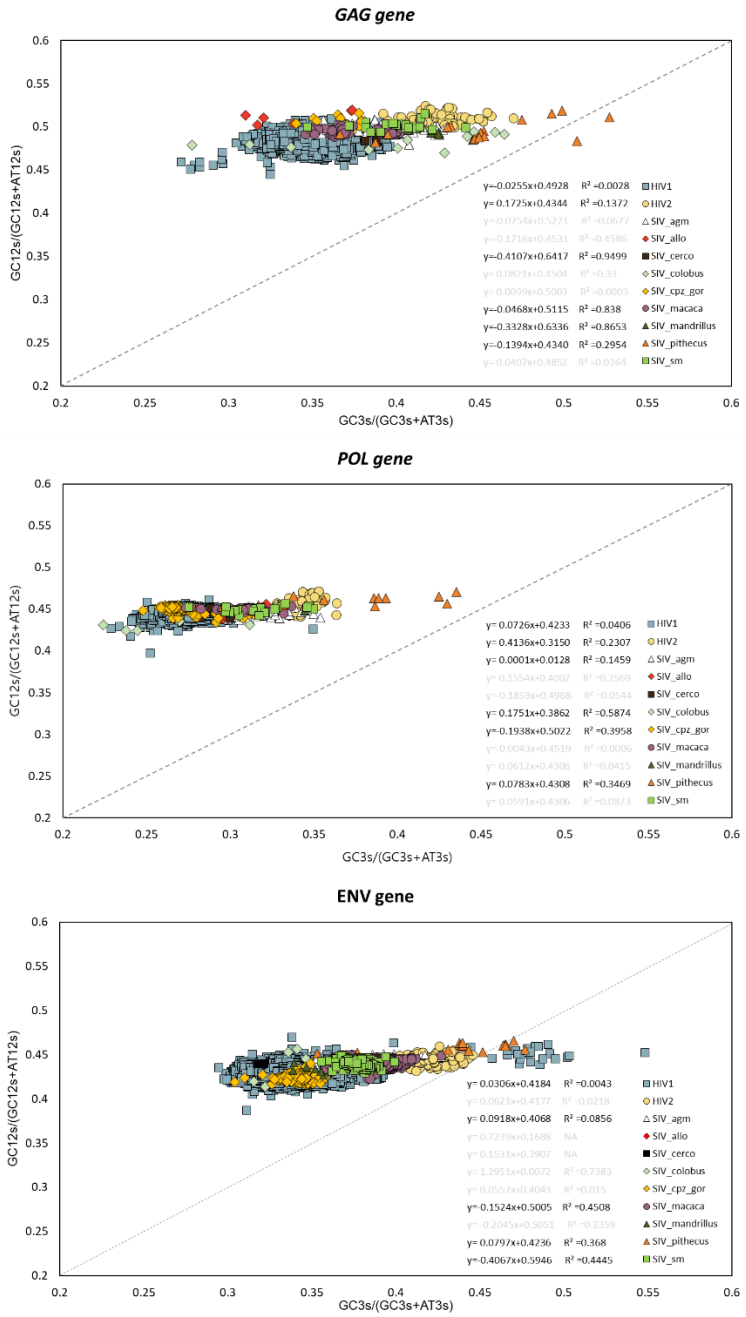


**Figure 3.6 Neutrality plot for lentiviral gag, pol and env genes.** Each icon in the plot represents one sequence. The linear regression of GC12s against GC3s is presented by each lentiviral group.

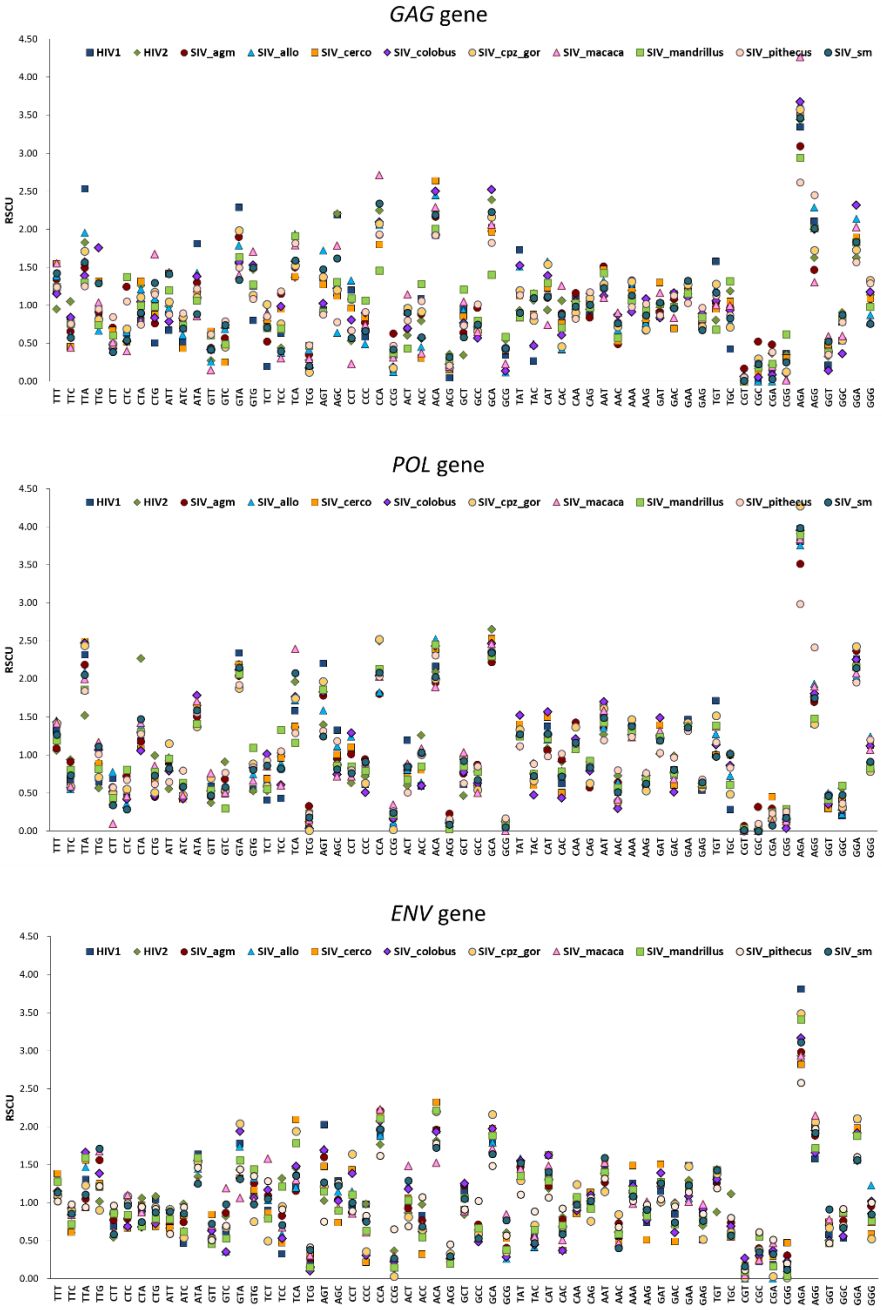


**Figure 3.7 RSCU plot of lentiviral gag, pol and env genes.** It can be found that the degree of variation among the 11 groups in the pol gene is the smallest.

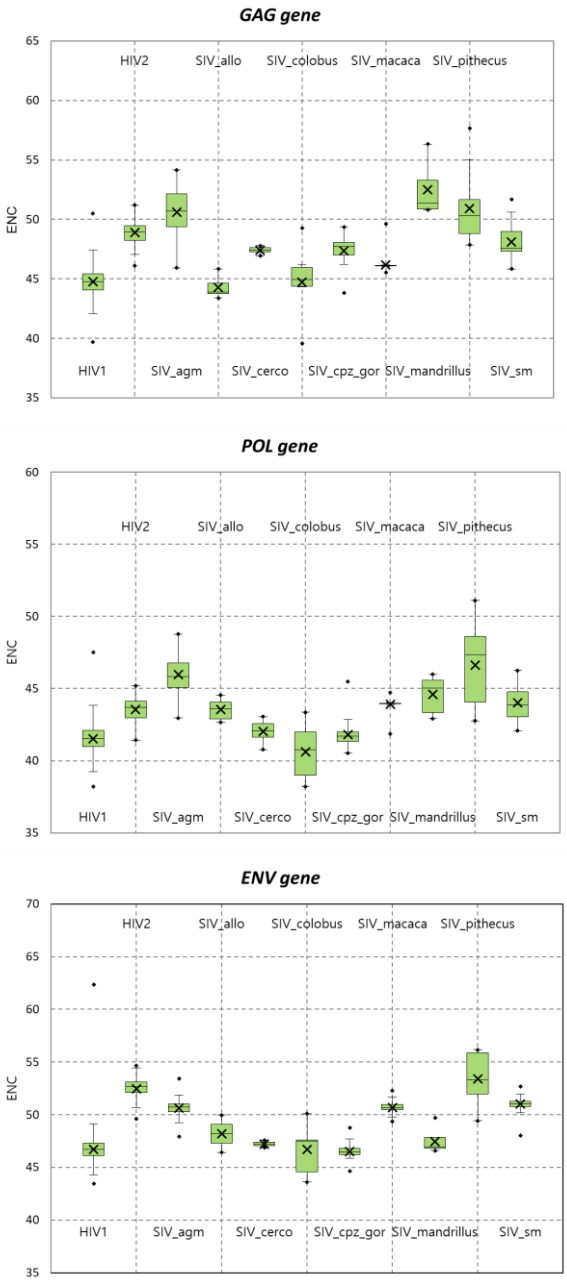


**Figure 3.8 ENC box plot of lentiviral gag, pol and env genes.** the cross represents mean ENC value and dots represent the minimum and maximum ENC value.

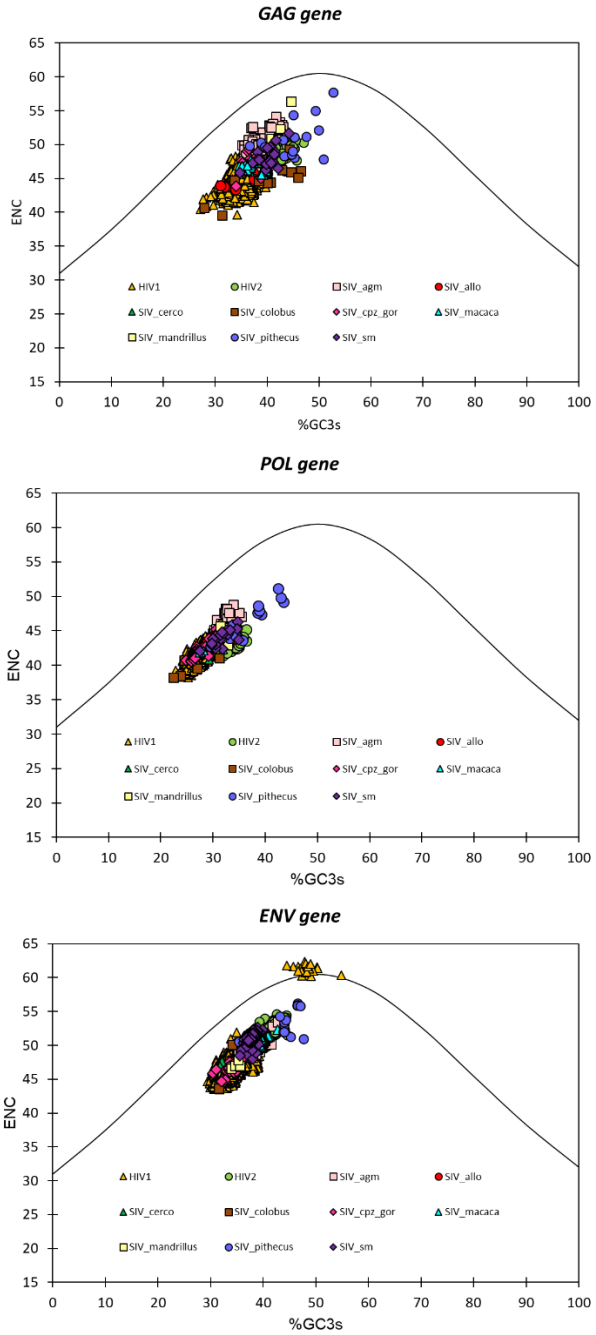


**Figure 3.9 ENC-GC3s plot of lentiviral gag, pol and env genes.** The smooth line is the expected value of ENC calculated based on the GC3s content.

## Table 3.1a RSCU value of lentiviral gag genes

| | | GAG gene | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AA** | **CODON** | **HIV 1** | **HIV 2** | **SIV agm** | **SIV alo** | **SIV cer** | **SIV col** | **SIV cpz** | **SIV mac** | **SIV mnd** | **SIV pit** | **SIV sm** |
| **Phe** | TTT | 1.183 | 0.951 | 1.342 | 1.385 | 1.542 | 1.156 | 1.248 | 1.555 | 1.243 | 1.240 | 1.426 |
| | TTC | 0.817 | 1.049 | 0.658 | 0.615 | 0.458 | 0.844 | 0.752 | 0.445 | 0.758 | 0.760 | 0.574 |
| **Leu** | TTA | **2.534** | **1.827** | 1.493 | **1.958** | 1.291 | 1.390 | **1.716** | 1.578 | 1.305 | 1.253 | 1.568 |
| | TTG | 0.856 | 0.833 | 0.909 | 0.665 | 1.320 | **1.761** | 0.813 | 1.039 | 0.741 | 0.953 | 1.293 |
| | CTT | 0.672 | 0.484 | 0.706 | 0.417 | 0.674 | 0.434 | 0.494 | 0.516 | 0.602 | 0.848 | 0.389 |
| | CTC | 0.631 | 0.533 | 1.243 | 0.657 | 0.516 | 0.526 | 0.692 | 0.399 | 1.373 | 1.052 | 0.545 |
| | CTA | 0.799 | 1.264 | 0.889 | 1.212 | 1.320 | 1.051 | 1.111 | 0.792 | 0.998 | 0.747 | 0.905 |
| | CTG | 0.508 | 1.059 | 0.760 | 1.092 | 0.879 | 0.839 | 1.175 | **1.675** | 0.982 | 1.148 | 1.300 |
| **Ile** | ATT | 0.673 | 0.990 | 0.882 | 0.960 | 1.417 | 0.789 | 1.049 | 1.435 | 1.199 | 0.882 | 1.416 |
| | ATC | 0.514 | 0.832 | 0.817 | 0.610 | 0.435 | 0.832 | 0.849 | 0.706 | 0.739 | 0.898 | 0.698 |
| | ATA | **1.813** | 1.178 | 1.301 | 1.430 | 1.148 | 1.379 | 1.103 | 0.859 | 1.062 | 1.221 | 0.886 |
| **Val** | GTT | 0.420 | 0.271 | 0.440 | 0.258 | 0.659 | 0.421 | 0.422 | 0.149 | 0.602 | 0.623 | 0.418 |
| | GTC | 0.492 | 0.707 | 0.567 | 0.690 | 0.253 | 0.476 | 0.455 | 0.746 | 0.490 | 0.793 | 0.745 |
| | GTA | **2.286** | 1.550 | **1.898** | **1.789** | 1.596 | 1.571 | **1.987** | 1.396 | **1.635** | 1.496 | 1.337 |
| | GTG | 0.802 | 1.471 | 1.094 | 1.263 | 1.491 | 1.532 | 1.136 | **1.710** | 1.273 | 1.088 | 1.500 |
| **Ser** | TCT | 0.196 | 0.794 | 0.524 | 0.906 | 0.807 | 1.002 | 1.013 | 0.888 | 0.706 | 0.857 | 0.711 |
| | TCC | 0.634 | 0.443 | 1.153 | 0.377 | 0.952 | 0.989 | 0.760 | 0.305 | 0.693 | 1.186 | 0.399 |
| | TCA | **1.905** | 1.370 | 1.499 | **1.937** | 1.378 | 1.583 | 1.525 | **1.791** | **1.911** | **1.818** | 1.592 |
| | TCG | 0.138 | 0.278 | 0.350 | 0.417 | 0.466 | 0.175 | 0.123 | 0.300 | 0.472 | 0.479 | 0.204 |
| | AGT | 0.938 | 0.911 | 1.340 | **1.723** | 1.272 | 1.024 | 1.378 | 0.927 | 0.910 | 0.880 | 1.474 |
| | AGC | **2.190** | **2.205** | 1.134 | 0.641 | 1.126 | 1.227 | 1.202 | **1.788** | 1.308 | 0.781 | **1.621** |
| **Pro** | CCT | 1.200 | 0.533 | 0.670 | 1.329 | 0.964 | 0.812 | 1.082 | 0.230 | 1.095 | 0.672 | 0.569 |
| | CCC | 0.587 | 0.901 | 0.774 | 0.491 | 0.846 | 0.906 | 0.663 | 0.742 | 1.067 | 0.917 | 0.667 |
| | CCA | **2.058** | **2.251** | **1.926** | **2.057** | **1.797** | **2.095** | **2.076** | **2.712** | 1.458 | **1.938** | **2.341** |
| | CCG | 0.155 | 0.314 | 0.629 | 0.123 | 0.393 | 0.187 | 0.179 | 0.317 | 0.379 | 0.472 | 0.423 |
| **Thr** | ACT | 0.949 | 0.607 | 0.696 | 0.939 | 0.909 | 0.701 | 0.964 | 1.144 | 0.434 | 0.805 | 0.906 |
| | ACC | 1.082 | 0.797 | 0.884 | 0.454 | 0.306 | 0.572 | 0.920 | 0.370 | 1.282 | 1.059 | 0.580 |
| | ACA | **1.923** | **2.236** | **2.168** | **2.447** | **2.635** | **2.499** | **1.933** | **2.293** | **2.010** | **1.925** | **2.192** |
| | ACG | 0.046 | 0.360 | 0.251 | 0.160 | 0.150 | 0.228 | 0.183 | 0.193 | 0.275 | 0.211 | 0.323 |
| **Ala** | GCT | 0.956 | 0.347 | 0.638 | 0.933 | 0.857 | 0.768 | 0.741 | 1.050 | 1.210 | 0.740 | 0.584 |
| | GCC | 0.679 | 0.720 | 0.965 | 0.729 | 0.769 | 0.573 | 0.659 | 0.653 | 0.798 | 1.016 | 0.751 |
| | GCA | **2.019** | **2.387** | **1.976** | **2.219** | **1.967** | **2.523** | **2.161** | **2.064** | 1.404 | **1.825** | **2.228** |
| | GCG | 0.346 | 0.546 | 0.420 | 0.119 | 0.407 | 0.136 | 0.439 | 0.232 | 0.587 | 0.419 | 0.437 |
| **Tyr** | TAT | **1.729** | 0.935 | 1.198 | 1.504 | 1.124 | 1.528 | 1.197 | 0.843 | 0.846 | 1.136 | 0.904 |
| | TAC | 0.271 | 1.065 | 0.802 | 0.496 | 0.876 | 0.472 | 0.803 | 1.157 | 1.154 | 0.864 | 1.096 |
| **His** | CAT | 1.228 | 0.941 | 1.103 | 1.579 | 1.236 | 1.392 | 1.543 | 0.743 | 1.299 | 1.136 | 1.117 |
| | CAC | 0.772 | 1.059 | 0.897 | 0.421 | 0.764 | 0.608 | 0.457 | 1.257 | 0.701 | 0.864 | 0.883 |
| **Gln** | CAA | 1.091 | 0.845 | 1.162 | 0.941 | 0.932 | 1.051 | 0.902 | 0.990 | 1.029 | 0.825 | 0.989 |
| | CAG | 0.909 | 1.155 | 0.838 | 1.059 | 1.068 | 0.949 | 1.098 | 1.010 | 0.971 | 1.175 | 1.011 |
| **Asn** | AAT | 1.244 | 1.275 | 1.509 | 1.321 | 1.470 | 1.126 | 1.282 | 1.097 | 1.426 | 1.222 | 1.233 |
| | AAC | 0.756 | 0.725 | 0.491 | 0.679 | 0.530 | 0.874 | 0.719 | 0.903 | 0.575 | 0.778 | 0.767 |
| **Lys** | AAA | 1.281 | 1.202 | 1.329 | 1.231 | 1.181 | 0.916 | 1.316 | 1.083 | 1.062 | 0.979 | 1.131 |
| | AAG | 0.719 | 0.798 | 0.671 | 0.769 | 0.819 | 1.084 | 0.684 | 0.917 | 0.938 | 1.021 | 0.869 |

**Table 3.1a (continued) RSCU value of lentiviral gag genes**

| | | GAG gene | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AA** | **CODON** | **HIV 1** | **HIV 2** | **SIV agm** | **SIV alo** | **SIV cer** | **SIV col** | **SIV cpz** | **SIV mac** | **SIV mnd** | **SIV pit** | **SIV sm** |
| **Asp** | GAT | 0.850 | 0.930 | 0.909 | 1.019 | 1.303 | 0.837 | 1.027 | 1.165 | 1.020 | 0.869 | 1.036 |
| | GAC | 1.150 | 1.070 | 1.091 | 0.981 | 0.697 | 1.163 | 0.973 | 0.835 | 0.980 | 1.131 | 0.964 |
| **Glu** | GAA | 1.264 | 1.148 | 1.208 | 1.271 | 1.179 | 1.108 | 1.277 | 1.175 | 1.157 | 1.031 | 1.323 |
| | GAG | 0.736 | 0.852 | 0.792 | 0.730 | 0.821 | 0.892 | 0.723 | 0.825 | 0.843 | 0.969 | 0.677 |
| **Cys** | TGT | 1.576 | 0.809 | 1.035 | 1.220 | 0.949 | 1.061 | 1.283 | 1.003 | 0.683 | 1.131 | 1.168 |
| | TGC | 0.424 | 1.191 | 0.965 | 0.780 | 1.051 | 0.940 | 0.717 | 0.997 | 1.318 | 0.869 | 0.832 |
| **Arg** | CGT | 0.009 | 0.005 | 0.170 | 0.033 | 0.041 | 0.063 | 0.055 | 0.000 | 0.000 | 0.054 | 0.009 |
| | CGC | 0.013 | 0.202 | 0.522 | 0.000 | 0.090 | 0.056 | 0.300 | 0.213 | 0.191 | 0.238 | 0.229 |
| | CGA | 0.162 | 0.370 | 0.485 | 0.000 | 0.044 | 0.091 | 0.216 | 0.202 | 0.234 | 0.385 | 0.030 |
| | CGG | 0.363 | 0.352 | 0.265 | 0.030 | 0.305 | 0.107 | 0.129 | 0.014 | 0.621 | 0.255 | 0.255 |
| | AGA | **3.346** | **3.443** | **3.092** | **3.648** | **3.511** | **3.678** | **3.576** | **4.266** | **2.942** | **2.617** | **3.466** |
| | AGG | **2.107** | **1.628** | 1.467 | **2.288** | **2.008** | **2.004** | **1.723** | 1.305 | **2.012** | **2.450** | **2.012** |
| **Gly** | GGT | 0.211 | 0.328 | 0.450 | 0.436 | 0.483 | 0.141 | 0.394 | 0.592 | 0.380 | 0.354 | 0.526 |
| | GGC | 0.782 | 0.908 | 0.832 | 0.558 | 0.535 | 0.366 | 0.545 | 0.597 | 0.809 | 0.781 | 0.874 |
| | GGA | **1.880** | **1.627** | **1.724** | **2.136** | **1.894** | **2.316** | **1.730** | **2.030** | **1.828** | 1.571 | **1.838** |
| | GGG | 1.127 | 1.137 | 0.994 | 0.871 | 1.087 | 1.177 | 1.331 | 0.780 | 0.983 | 1.294 | 0.761 |

RSCU value greater than 1.6 namely over-represented codons are marked as boldface.

## Table 3.1b RSCU value of lentiviral pol genes

| | | POL gene | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AA** | **CODON** | **HIV 1** | **HIV 2** | **SIV agm** | **SIV alo** | **SIV cer** | **SIV col** | **SIV cpz** | **SIV mac** | **SIV mnd** | **SIV pit** | **SIV sm** |
| **Phe** | TTT | 1.326 | 1.060 | 1.088 | 1.453 | 1.405 | 1.431 | 1.417 | 1.415 | 1.191 | 1.268 | 1.265 |
| | TTC | 0.674 | 0.940 | 0.912 | 0.547 | 0.595 | 0.569 | 0.583 | 0.585 | 0.810 | 0.732 | 0.735 |
| **Leu** | TTA | **2.319** | 1.519 | **2.187** | **2.090** | **2.492** | **2.473** | **2.438** | **2.000** | **1.861** | **1.842** | **2.058** |
| | TTG | 0.644 | 0.567 | 0.887 | 0.716 | 0.888 | 1.068 | 0.705 | 1.166 | 0.818 | 1.013 | 1.113 |
| | CTT | 0.693 | 0.343 | 0.576 | 0.777 | 0.349 | 0.519 | 0.499 | 0.097 | 0.435 | 0.575 | 0.340 |
| | CTC | 0.436 | 0.308 | 0.724 | 0.353 | 0.539 | 0.418 | 0.552 | 0.458 | 0.807 | 0.675 | 0.287 |
| | CTA | 1.182 | **2.271** | 1.178 | 1.386 | 1.082 | 1.057 | 1.299 | 1.424 | 1.300 | 1.277 | 1.471 |
| | CTG | 0.725 | 0.993 | 0.448 | 0.679 | 0.649 | 0.464 | 0.507 | 0.855 | 0.780 | 0.617 | 0.731 |
| **Ile** | ATT | 0.909 | 0.556 | 0.921 | 0.972 | 0.937 | 0.792 | 1.151 | 0.824 | 0.951 | 0.649 | 0.825 |
| | ATC | 0.463 | 0.796 | 0.575 | 0.458 | 0.415 | 0.424 | 0.476 | 0.470 | 0.638 | 0.797 | 0.585 |
| | ATA | **1.628** | **1.648** | 1.504 | 1.570 | **1.648** | **1.784** | 1.373 | **1.706** | 1.411 | 1.555 | 1.590 |
| **Val** | GTT | 0.590 | 0.369 | 0.453 | 0.476 | 0.490 | 0.532 | 0.699 | 0.760 | 0.537 | 0.468 | 0.467 |
| | GTC | 0.518 | 0.910 | 0.681 | 0.599 | 0.489 | 0.515 | 0.536 | 0.500 | 0.297 | 0.768 | 0.577 |
| | GTA | **2.340** | **2.192** | **2.043** | **2.180** | **2.186** | **2.116** | **1.878** | **2.076** | **2.069** | **1.922** | **2.151** |
| | GTG | 0.552 | 0.528 | 0.823 | 0.746 | 0.835 | 0.837 | 0.888 | 0.665 | 1.096 | 0.842 | 0.805 |
| **Ser** | TCT | 0.406 | 0.527 | 0.594 | 0.666 | 0.689 | 1.014 | 0.593 | 0.850 | 0.549 | 0.927 | 0.861 |
| | TCC | 0.427 | 1.029 | 0.982 | 0.889 | 0.957 | 0.605 | 0.816 | 0.601 | 1.332 | 1.050 | 0.823 |
| | TCA | 1.580 | **1.967** | 1.363 | **1.720** | 1.379 | **1.770** | **1.743** | **2.396** | 1.163 | 1.291 | **2.076** |
| | TCG | 0.057 | 0.125 | 0.329 | 0.023 | 0.117 | 0.039 | 0.007 | 0.148 | 0.228 | 0.229 | 0.183 |
| | AGT | **2.205** | 1.401 | **1.782** | 1.584 | **1.854** | **1.827** | **1.967** | 1.288 | **1.864** | 1.318 | 1.246 |
| | AGC | 1.325 | 0.952 | 0.950 | 1.118 | 1.005 | 0.745 | 0.874 | 0.716 | 0.864 | 1.185 | 0.811 |
| **Pro** | CCT | 1.064 | 0.630 | 1.009 | 1.241 | 1.107 | 1.290 | 0.830 | 0.717 | 0.849 | 0.800 | 0.763 |
| | CCC | 0.748 | 0.721 | 0.944 | 0.824 | 0.617 | 0.505 | 0.629 | 0.904 | 0.793 | 0.902 | 0.914 |
| | CCA | **2.137** | **2.499** | **1.804** | **1.823** | **2.107** | **2.043** | **2.522** | **2.033** | **2.134** | **2.032** | **2.083** |
| | CCG | 0.050 | 0.151 | 0.244 | 0.112 | 0.170 | 0.161 | 0.019 | 0.346 | 0.223 | 0.266 | 0.241 |
| **Thr** | ACT | 1.193 | 0.597 | 0.800 | 0.804 | 0.710 | 0.863 | 0.854 | 0.886 | 0.673 | 0.506 | 0.846 |
| | ACC | 0.591 | 1.258 | 1.017 | 0.635 | 0.805 | 0.597 | 1.008 | 1.085 | 0.850 | 1.025 | 1.036 |
| | ACA | **2.165** | **2.089** | **1.953** | **2.522** | **2.391** | **2.431** | **2.001** | **1.893** | **2.452** | **2.309** | **2.030** |
| | ACG | 0.051 | 0.055 | 0.230 | 0.038 | 0.094 | 0.110 | 0.137 | 0.135 | 0.026 | 0.160 | 0.088 |
| **Ala** | GCT | 0.608 | 0.467 | 0.756 | 0.986 | 0.820 | 0.774 | 0.974 | 1.037 | 0.853 | 0.627 | 0.925 |
| | GCC | 0.795 | 0.784 | 0.870 | 0.643 | 0.575 | 0.632 | 0.542 | 0.500 | 0.777 | 0.850 | 0.676 |
| | GCA | **2.530** | **2.655** | **2.220** | **2.326** | **2.525** | **2.467** | **2.355** | **2.461** | **2.303** | **2.354** | **2.347** |
| | GCG | 0.068 | 0.095 | 0.154 | 0.046 | 0.079 | 0.127 | 0.128 | 0.002 | 0.068 | 0.169 | 0.052 |
| **Tyr** | TAT | 1.378 | 1.107 | 1.248 | 1.385 | 1.402 | 1.525 | 1.336 | 1.264 | 1.247 | 1.117 | 1.274 |
| | TAC | 0.622 | 0.893 | 0.752 | 0.615 | 0.598 | 0.475 | 0.664 | 0.736 | 0.753 | 0.883 | 0.726 |
| **His** | CAT | 1.376 | 1.086 | 1.060 | 1.204 | 1.496 | 1.566 | 1.271 | 1.268 | 1.217 | 0.983 | 1.280 |
| | CAC | 0.624 | 0.914 | 0.940 | 0.796 | 0.504 | 0.434 | 0.729 | 0.732 | 0.783 | 1.017 | 0.720 |
| **Gln** | CAA | 1.178 | 1.091 | 1.428 | 1.116 | 1.106 | 1.211 | 1.368 | 1.084 | 1.079 | 1.148 | 1.165 |
| | CAG | 0.822 | 0.909 | 0.572 | 0.885 | 0.894 | 0.789 | 0.632 | 0.916 | 0.922 | 0.852 | 0.835 |
| **Asn** | AAT | 1.488 | 1.273 | 1.360 | 1.362 | **1.630** | **1.702** | **1.609** | 1.583 | 1.431 | 1.196 | 1.487 |
| | AAC | 0.512 | 0.727 | 0.640 | 0.638 | 0.370 | 0.298 | 0.391 | 0.417 | 0.569 | 0.804 | 0.513 |
| **Lys** | AAA | 1.447 | 1.320 | 1.395 | 1.283 | 1.366 | 1.269 | 1.468 | 1.232 | 1.314 | 1.237 | 1.377 |
| | AAG | 0.553 | 0.680 | 0.605 | 0.717 | 0.634 | 0.731 | 0.532 | 0.768 | 0.686 | 0.763 | 0.623 |

**Table 3.1b (continued) RSCU value of lentiviral pol genes**

| | | POL gene | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AA** | **CODON** | **HIV 1** | **HIV 2** | **SIV agm** | **SIV alo** | **SIV cer** | **SIV col** | **SIV cpz** | **SIV mac** | **SIV mnd** | **SIV pit** | **SIV sm** |
| **Asp** | GAT | 1.195 | 1.013 | 1.251 | 1.267 | 1.394 | 1.488 | 1.225 | 1.317 | 1.242 | 1.031 | 1.190 |
| | GAC | 0.805 | 0.987 | 0.749 | 0.733 | 0.606 | 0.512 | 0.775 | 0.683 | 0.758 | 0.969 | 0.810 |
| **Glu** | GAA | 1.462 | 1.374 | 1.398 | 1.360 | 1.408 | 1.422 | 1.425 | 1.386 | 1.377 | 1.320 | 1.359 |
| | GAG | 0.538 | 0.626 | 0.602 | 0.640 | 0.592 | 0.578 | 0.575 | 0.614 | 0.624 | 0.680 | 0.641 |
| **Cys** | TGT | **1.714** | 0.978 | 1.003 | 1.275 | 0.998 | 1.136 | 1.516 | 1.373 | 1.393 | 1.167 | 0.982 |
| | TGC | 0.286 | 1.022 | 0.998 | 0.725 | 1.002 | 0.864 | 0.484 | 0.627 | 0.607 | 0.833 | 1.018 |
| **Arg** | CGT | 0.005 | 0.037 | 0.065 | 0.000 | 0.029 | 0.017 | 0.004 | 0.001 | 0.028 | 0.010 | 0.013 |
| | CGC | 0.015 | 0.032 | 0.315 | 0.000 | 0.000 | 0.019 | 0.001 | 0.000 | 0.055 | 0.096 | 0.000 |
| | CGA | 0.228 | 0.135 | 0.292 | 0.216 | 0.452 | 0.169 | 0.157 | 0.124 | 0.249 | 0.234 | 0.077 |
| | CGG | 0.174 | 0.116 | 0.121 | 0.087 | 0.139 | 0.038 | 0.156 | 0.129 | 0.293 | 0.258 | 0.177 |
| | AGA | **3.811** | **3.914** | **3.511** | **3.759** | **3.972** | **3.952** | **4.276** | **3.844** | **3.895** | **2.985** | **3.985** |
| | AGG | **1.767** | **1.767** | **1.696** | **1.939** | 1.408 | **1.806** | 1.406 | **1.903** | 1.480 | **2.417** | **1.748** |
| **Gly** | GGT | 0.406 | 0.435 | 0.326 | 0.501 | 0.294 | 0.350 | 0.476 | 0.474 | 0.418 | 0.475 | 0.464 |
| | GGC | 0.199 | 0.563 | 0.404 | 0.222 | 0.403 | 0.279 | 0.315 | 0.384 | 0.594 | 0.365 | 0.475 |
| | GGA | **2.266** | **2.189** | **2.369** | **2.032** | **2.166** | **2.254** | **2.426** | **2.074** | **2.161** | **1.958** | **2.146** |
| | GGG | 1.130 | 0.813 | 0.902 | 1.245 | 1.137 | 1.117 | 0.784 | 1.069 | 0.827 | 1.202 | 0.915 |

RSCU value greater than 1.6 namely over-represented codons are marked as boldface.

## Table 3.1c RSCU value of lentiviral env genes

| | | ENV gene | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AA** | **CODON** | **HIV 1** | **HIV 2** | **SIV agm** | **SIV alo** | **SIV cer** | **SIV col** | **SIV cpz** | **SIV mac** | **SIV mnd** | **SIV pit** | **SIV sm** |
| **Phe** | TTT | 1.117 | 1.071 | 1.077 | 1.345 | 1.386 | 1.154 | 1.100 | 1.158 | 1.279 | 1.020 | 1.140 |
| | TTC | 0.883 | 0.929 | 0.923 | 0.655 | 0.614 | 0.846 | 0.900 | 0.842 | 0.721 | 0.980 | 0.860 |
| **Leu** | TTA | 1.304 | 1.236 | 1.052 | 1.469 | 1.561 | **1.663** | 1.231 | 0.943 | 1.591 | 0.941 | 1.109 |
| | TTG | 1.209 | 1.018 | 1.561 | 1.235 | 1.249 | 1.387 | 0.905 | **1.674** | 1.223 | 1.219 | **1.716** |
| | CTT | 0.690 | 0.545 | 0.775 | 0.887 | 0.859 | 0.856 | 0.881 | 0.619 | 0.859 | 0.967 | 0.589 |
| | CTC | 0.835 | 1.055 | 0.793 | 0.712 | 0.671 | 0.689 | 1.086 | 1.103 | 0.842 | 1.001 | 0.963 |
| | CTA | 0.915 | 1.061 | 0.892 | 0.899 | 0.968 | 0.680 | 0.951 | 0.875 | 0.693 | 0.950 | 0.747 |
| | CTG | 1.048 | 1.085 | 0.927 | 0.798 | 0.693 | 0.725 | 0.945 | 0.786 | 0.793 | 0.922 | 0.876 |
| **Ile** | ATT | 0.893 | 0.668 | 0.716 | 0.923 | 0.669 | 0.902 | 0.913 | 0.849 | 0.786 | 0.591 | 0.886 |
| | ATC | 0.469 | 0.985 | 0.746 | 0.635 | 0.833 | 0.540 | 0.543 | 0.871 | 0.619 | 0.941 | 0.860 |
| | ATA | **1.638** | 1.347 | 1.538 | 1.443 | 1.499 | 1.557 | 1.543 | 1.280 | 1.595 | 1.467 | 1.255 |
| **Val** | GTT | 0.464 | 0.654 | 0.547 | 0.572 | 0.848 | 0.634 | 0.529 | 0.706 | 0.463 | 0.509 | 0.724 |
| | GTC | 0.600 | 0.822 | 0.871 | 0.359 | 0.666 | 0.354 | 0.676 | 1.195 | 0.534 | 0.699 | 0.981 |
| | GTA | **1.779** | 1.433 | 1.325 | **1.740** | 1.321 | **1.944** | **2.042** | 1.065 | 1.561 | 1.440 | 1.312 |
| | GTG | 1.157 | 1.090 | 1.258 | 1.331 | 1.166 | 1.068 | 0.753 | 1.034 | 1.443 | 1.352 | 0.983 |
| **Ser** | TCT | 0.899 | 0.801 | 1.098 | 1.035 | 0.871 | 1.169 | 0.497 | 1.580 | 0.792 | 1.286 | 1.053 |
| | TCC | 0.332 | 1.320 | 0.830 | 0.587 | 0.475 | 0.530 | 0.951 | 0.676 | 1.216 | 0.915 | 0.711 |
| | TCA | 1.227 | 1.457 | 1.158 | 1.201 | **2.095** | 1.482 | **1.944** | 1.302 | **1.785** | 1.363 | 1.357 |
| | TCG | 0.224 | 0.146 | 0.277 | 0.326 | 0.344 | 0.103 | 0.155 | 0.313 | 0.157 | 0.415 | 0.381 |
| | AGT | **2.030** | 1.030 | 1.599 | **1.710** | 1.476 | **1.690** | 1.215 | 1.195 | 1.150 | 0.756 | 1.270 |
| | AGC | 1.287 | 1.246 | 1.038 | 1.143 | 0.741 | 1.025 | 1.237 | 0.935 | 0.901 | 1.265 | 1.228 |
| **Pro** | CCT | 0.972 | 0.881 | 0.875 | 1.150 | 1.437 | 1.388 | **1.640** | 0.858 | 1.109 | 0.890 | 1.003 |
| | CCC | 0.977 | 0.984 | 0.772 | 0.800 | 0.217 | 0.311 | 0.361 | 0.609 | 0.634 | 0.834 | 0.757 |
| | CCA | **1.874** | **1.768** | **2.213** | **1.888** | **2.177** | **2.075** | **1.966** | **2.233** | **2.109** | **1.619** | **1.969** |
| | CCG | 0.177 | 0.367 | 0.140 | 0.163 | 0.170 | 0.227 | 0.033 | 0.301 | 0.148 | 0.657 | 0.271 |
| **Thr** | ACT | 0.954 | 0.867 | 0.932 | 1.229 | 1.059 | 1.180 | 0.813 | 1.488 | 1.032 | 0.691 | 1.290 |
| | ACC | 0.834 | 0.981 | 0.769 | 0.699 | 0.324 | 0.620 | 0.651 | 0.713 | 0.551 | 1.074 | 0.688 |
| | ACA | **1.954** | **1.816** | **1.962** | **1.798** | **2.324** | **1.935** | **2.203** | 1.526 | **2.216** | **1.779** | **1.725** |
| | ACG | 0.258 | 0.336 | 0.337 | 0.275 | 0.294 | 0.264 | 0.334 | 0.273 | 0.202 | 0.455 | 0.297 |
| **Ala** | GCT | 1.159 | 0.847 | 1.229 | 1.268 | 0.914 | 1.252 | 0.891 | 0.890 | 1.066 | 0.918 | 1.055 |
| | GCC | 0.656 | 0.644 | 0.714 | 0.676 | 0.571 | 0.487 | 0.588 | 0.544 | 0.663 | 1.026 | 0.530 |
| | GCA | **1.797** | **1.903** | **1.648** | **1.796** | **1.981** | **1.973** | **2.165** | **1.715** | **1.887** | 1.486 | **1.642** |
| | GCG | 0.389 | 0.606 | 0.409 | 0.262 | 0.535 | 0.288 | 0.356 | 0.851 | 0.384 | 0.571 | 0.774 |
| **Tyr** | TAT | 1.417 | 1.304 | 1.474 | 1.584 | 1.303 | 1.552 | 1.291 | 1.446 | 1.355 | 1.111 | 1.526 |
| | TAC | 0.583 | 0.696 | 0.527 | 0.417 | 0.697 | 0.448 | 0.709 | 0.554 | 0.646 | 0.889 | 0.474 |
| **His** | CAT | 1.350 | 1.190 | 1.223 | 1.385 | **1.628** | **1.626** | 1.312 | 1.487 | 1.295 | 1.073 | 1.411 |
| | CAC | 0.650 | 0.810 | 0.777 | 0.616 | 0.373 | 0.374 | 0.688 | 0.513 | 0.705 | 0.927 | 0.589 |
| **Gln** | CAA | 1.031 | 0.978 | 1.074 | 1.052 | 0.859 | 0.898 | 1.242 | 1.060 | 1.075 | 0.937 | 0.974 |
| | CAG | 0.969 | 1.022 | 0.926 | 0.949 | 1.141 | 1.102 | 0.758 | 0.940 | 0.925 | 1.063 | 1.026 |
| **Asn** | AAT | 1.426 | 1.169 | 1.269 | 1.421 | 1.528 | 1.387 | 1.150 | 1.529 | 1.406 | 1.319 | 1.593 |
| | AAC | 0.574 | 0.831 | 0.731 | 0.580 | 0.473 | 0.613 | 0.850 | 0.471 | 0.595 | 0.681 | 0.407 |
| **Lys** | AAA | 1.256 | 1.165 | 1.134 | 1.110 | 1.490 | 1.222 | 1.081 | 0.986 | 1.172 | 1.025 | 1.087 |
| | AAG | 0.744 | 0.835 | 0.866 | 0.890 | 0.511 | 0.778 | 0.919 | 1.014 | 0.828 | 0.975 | 0.913 |

**Table 3.1c (continued) RSCU value of lentiviral env genes**

| | | ENV gene | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AA** | **CODON** | **HIV 1** | **HIV 2** | **SIV agm** | **SIV alo** | **SIV cer** | **SIV col** | **SIV cpz** | **SIV mac** | **SIV mnd** | **SIV pit** | **SIV sm** |
| **Asp** | GAT | 1.151 | 0.999 | 1.264 | 1.268 | 1.507 | 1.392 | 1.018 | 1.261 | 1.279 | 1.048 | 1.257 |
| | GAC | 0.849 | 1.001 | 0.736 | 0.732 | 0.494 | 0.608 | 0.982 | 0.739 | 0.722 | 0.952 | 0.743 |
| **Glu** | GAA | 1.488 | 1.296 | 1.230 | 1.062 | 1.160 | 1.087 | 1.479 | 1.017 | 1.229 | 1.144 | 1.234 |
| | GAG | 0.512 | 0.704 | 0.770 | 0.939 | 0.840 | 0.913 | 0.521 | 0.983 | 0.772 | 0.856 | 0.766 |
| **Cys** | TGT | 1.397 | 0.881 | 1.260 | 1.313 | 1.340 | 1.309 | 1.390 | 1.443 | 1.431 | 1.194 | 1.430 |
| | TGC | 0.603 | 1.119 | 0.740 | 0.688 | 0.661 | 0.691 | 0.610 | 0.557 | 0.569 | 0.806 | 0.570 |
| **Arg** | CGT | 0.054 | 0.108 | 0.063 | 0.000 | 0.174 | 0.270 | 0.014 | 0.054 | 0.099 | 0.114 | 0.164 |
| | CGC | 0.226 | 0.382 | 0.402 | 0.264 | 0.236 | 0.307 | 0.380 | 0.245 | 0.555 | 0.614 | 0.355 |
| | CGA | 0.184 | 0.327 | 0.356 | 0.000 | 0.297 | 0.371 | 0.036 | 0.473 | 0.169 | 0.515 | 0.325 |
| | CGG | 0.145 | 0.300 | 0.310 | 0.201 | 0.476 | 0.230 | 0.015 | 0.146 | 0.047 | 0.204 | 0.124 |
| | AGA | **3.810** | **2.930** | **2.984** | **3.488** | **2.823** | **3.165** | **3.491** | **2.935** | **3.411** | **2.578** | **3.112** |
| | AGG | 1.580 | **1.953** | **1.883** | **2.049** | **1.994** | **1.657** | **2.065** | **2.147** | **1.721** | **1.976** | **1.921** |
| **Gly** | GGT | 0.582 | 0.705 | 0.723 | 0.507 | 0.748 | 0.503 | 0.670 | 0.782 | 0.473 | 0.470 | 0.914 |
| | GGC | 0.538 | 0.887 | 0.761 | 0.695 | 0.681 | 0.558 | 0.696 | 0.865 | 0.880 | 0.923 | 0.673 |
| | GGA | **2.097** | 1.552 | 1.559 | 1.568 | **1.983** | **1.921** | **2.107** | **1.603** | **1.882** | **1.605** | 1.563 |
| | GGG | 0.783 | 0.856 | 0.957 | 1.229 | 0.588 | 1.018 | 0.528 | 0.750 | 0.766 | 1.001 | 0.850 |

RSCU greater than 1.6 namely over-represented codons are marked as boldface.

**Table 3.2 ICDI value of lentiviral gag, pol and env genes**

| | gag | | pol | | env | |
|---|---|---|---|---|---|---|
| Group | N | ICDI | N | ICDI | N | ICDI |
| HIV1 | 22699 | 0.169±0.022 | 11021 | 0.199±0.019 | 70658 | 0.130±0.015 |
| HIV2 | 76 | 0.097±0.013 | 45 | 0.134±0.013 | 170 | 0.065±0.012 |
| SIV_agm | 43 | 0.113±0.023 | 40 | 0.130±0.016 | 96 | 0.088±0.012 |
| SIV_allo | 6 | 0.164±0.012 | 6 | 0.153±0.016 | 2 | 0.114±0.034 |
| SIV_cerco | 5 | 0.132±0.017 | 5 | 0.187±0.017 | 2 | 0.153±0.012 |
| SIV_colobus | 12 | 0.161±0.053 | 7 | 0.223±0.051 | 5 | 0.148±0.018 |
| SIVcpz_gor | 16 | 0.135±0.019 | 172 | 0.189±0.015 | 192 | 0.117±0.007 |
| SIV_macaca | 624 | 0.117±0.002 | 341 | 0.149±0.005 | 14611 | 0.101±0.008 |
| SIV_mandrillus | 8 | 0.081±0.013 | 6 | 0.140±0.016 | 8 | 0.119±0.010 |
| SIV_pithecus | 15 | 0.091±0.018 | 13 | 0.116±0.034 | 18 | 0.061±0.016 |
| SIV_sm | 21 | 0.114±0.013 | 19 | 0.145±0.017 | 482 | 0.105±0.011 |

ICDI value is represented as mean ± standard deviation format.

**Table 3.3 Correlation test result of ENC, ICDI, CAI and D(A, B)**

| GAG gene | | | | |
|---|---|---|---|---|
| Pearson's correlation (N=23525, α=0.05, two-tailed) | ENC | ICDI | CAI | D(A, B) |
| ENC | 1 | -0.783* | 0.129* | -0.794* |
| ICDI | | 1 | -0.289* | 0.593* |
| CAI | | | 1 | -0.509* |
| D(A, B) | | | | 1 |
| **POL gene** | | | | |
| Pearson's correlation (N=11675, α=0.05, two-tailed) | ENC | ICDI | CAI | D(A, B) |
| ENC | 1 | -0.896* | 0.453* | -0.893* |
| ICDI | | 1 | -0.453* | 0.700* |
| CAI | | | 1 | -0.581* |
| D(A, B) | | | | 1 |
| **ENV gene** | | | | |
| Pearson's correlation (N=86244, α=0.05, two-tailed) | ENC | ICDI | CAI | D(A, B) |
| ENC | 1 | -0.826* | 0.594* | -0.928* |
| ICDI | | 1 | -0.559* | 0.762* |
| CAI | | | 1 | -0.714* |
| D(A, B) | | | | 1 |

*means correlation is significant at 0.05 level (two taild).

**Table 3.4 CAI and 5% critical eCAI value of lentiviral genes**

| | gag | | pol | | env | |
|---|---|---|---|---|---|---|
| Group | CAI | eCAI | CAI | eCAI | CAI | eCAI |
| HIV1 | 0.722±0.006 | 0.733 | 0.691±0.004 | 0.708 | 0.711±0.007 | 0.716 |
| HIV2 | 0.735±0.009 | 0.738 | 0.694±0.005 | 0.719 | 0.730±0.005 | 0.739 |
| SIV_agm | 0.715±0.009 | 0.734 | 0.695±0.007 | 0.717 | 0.728±0.009 | 0.733 |
| SIV_allo | **0.727±0.013** | 0.720 | 0.695±0.009 | 0.707 | 0.726±0.006 | 0.734 |
| SIV_cerco | 0.725±0.008 | 0.732 | 0.701±0.007 | 0.713 | 0.716±0.004 | 0.717 |
| SIV_colobus | **0.735±0.016** | 0.731 | 0.698±0.013 | 0.713 | 0.719±0.004 | 0.723 |
| SIVcpz_gor | **0.726±0.009** | 0.724 | 0.691±0.005 | 0.699 | 0.700±0.006 | 0.725 |
| SIV_macaca | **0.746±0.005** | 0.730 | 0.703±0.002 | 0.717 | 0.724±0.006 | 0.734 |
| SIV_mandrillus | 0.729±0.008 | 0.738 | 0.708±0.004 | 0.717 | 0.730±0.009 | 0.734 |
| SIV_pithecus | 0.737±0.017 | 0.747 | 0.703±0.01 | 0.720 | 0.720±0.009 | 0.738 |
| SIV_sm | **0.738±0.008** | 0.737 | 0.702±0.005 | 0.717 | 0.724±0.011 | 0.730 |

CAI value is represented as mean ± standard deviation format and reference set is Homo sapiens. eCAI is represented as 5% critical value of eCAI. Bold type indicates that the CAI value is significantly greater than the mean value of eCAI.

**Table 3.5 Similarity D(A, B) value of lentiviral gag, pol and env genes**

| | gag | | pol | | env | |
|---|---|---|---|---|---|---|
| Group | N | D(A,B) | N | D(A,B) | N | D(A,B) |
| HIV1 | 22699 | 0.117±0.006 | 11021 | 0.132±0.005 | 70658 | 0.096±0.005 |
| HIV2 | 76 | 0.086±0.008 | 45 | 0.121±0.006 | 170 | 0.063±0.005 |
| SIV_agm | 43 | 0.085±0.010 | 40 | 0.110±0.009 | 96 | 0.073±0.006 |
| SIV_allo | 6 | 0.116±0.016 | 6 | 0.120±0.005 | 2 | 0.087±0.011 |
| SIV_cerco | 5 | 0.096±0.004 | 5 | 0.129±0.005 | 2 | 0.099±0.006 |
| SIV_colobus | 12 | 0.108±0.018 | 7 | 0.139±0.016 | 5 | 0.101±0.012 |
| SIVcpz_gor | 16 | 0.095±0.008 | 172 | 0.132±0.006 | 192 | 0.099±0.004 |
| SIV_macaca | 624 | 0.092±0.002 | 341 | 0.120±0.002 | 14611 | 0.078±0.003 |
| SIV_mandrillus | 8 | 0.068±0.007 | 6 | 0.109±0.006 | 8 | 0.089±0.005 |
| SIV_pithecus | 15 | 0.074±0.019 | 13 | 0.100±0.020 | 18 | 0.058±0.011 |
| SIV_sm | 21 | 0.087±0.007 | 19 | 0.119±0.008 | 482 | 0.077±0.005 |

Similarity D(A, B) value is represented as mean ± standard deviation format and reference set is Homo sapiens.

# 제 4 장. 고 찰

코돈 중첩성(degeneracy)에 의해 Met, Trp를 제외한 18개 아미노산은 두 개 이상의 코돈으로 코딩된다. 같은 아미노산을 암호화하는 코돈을 동의 코돈이라고 하며 이런 동의 코돈은 랜덤으로 사용되는 것이 아니라 거의 모든 생물종에서 서로 다른 정도의 코돈 사용 편향이 확인되고 있다. 다른 동의 코돈의 사용은 번역에 직접적으로 영향을 주어 번역의 개시(Kudla et al., 2009; Goodman et al., 2013), 효율 및 정확성(Drummond & Wilke, 2008)에 영향을 줄 뿐만 아니라 RNA의 구조와 접힘(Shabalina et al., 2006)에도 영향을 미치므로 유전자의 기능까지 관여하고 있다. 실제로 여러 연구에서 HIV1 gag유전자에 대하여 코돈 사용 최적화를 함으로써 해당 바이러스의 발현 정도가 향상된 것이 관찰되었고(Deml et al., 2001; Gao et al., 2003; Smith et al., 2004) 이는 번역 효율의 향상 보다 mRNA의 안정성과 핵탈출의 향상이 주요 원인으로 작용된 것이 확인되었다(Kofman et al., 2003). 또한 HIV1의 env 유전자에 대한 연구에서 gp41유전자 구역의 Arg 아미노산을 암호화하는 동의코돈의 치환(AGG→CGU)을 통해 바이러스의 복제 기능 및 env 유전자의 발현이 중지되는 것을 확인할 수 있었는데, 이는 intronic splicing silencer (ISS) RNA의 이차구조의 파괴로 인한 것이었다(Jordan-Paiz et al., 2020). 동의코돈에 대한 사용 패턴 분석은 바이러스 유전체의 진화 방향을 예측 가능하게 하고 바이러스의 코돈 사용에 대한 충분한 이해는 유전체 또는 해당 유전자에 대한 코돈 편집의 참고자료로 활용되어 유전자의 기능 파악, 선천적 면역 반응의 새로운 항 바이러스 기전 및 바이러스 생활사 중 아직까지 규명되지 않은 미지의 영역들을 밝히는데 기여할 수 있다.

본 연구에서 진행된 영장류 렌티바이러스의 코돈 사용 분석에서는 기존 HIV에 대한 연구와 같이(Van Hemert & Berkhout, 1995; Pandit & Sinha, 2011) 세 구조 유전자에서 극도로 심한 A 편향을 확인할 수가 있었다. Pol, env 유전자에서 C 뉴클레오티드를 최소화하는 유전체 조성과는 달리 gag 유전자에서는 T, C 뉴클레오티드의 사용을 최소화하고 있고 SIVallo, SIVcerco, SIVcolobus를 제외한 모든 영장류 렌티바이러스 그룹에서 T 뉴클레오티드 사용이 가장 적은 것으로 확인되었다. 또한 env 유전자에서 HIV2와 SIVpithecus 그룹만이 전반적인 AT 편향에 비해 AT3 편향이 낮은 것으로 확인되어 이 두 그룹에서의 AT편향은 주로 코돈 첫 번째와 두 번째 자리의 영향을 받아 형성된 것으로 생각된다. HIV1에 대한 기존 연구에서 Arg 아미노산의 코돈 사용 패턴은 HIV1 유전체의 극심한 A 편향과 CpG 디뉴클레오티드 사용을 피하는 공통적인 경향을 가장 대표적으로 보여주는 아미노산으로서, AGA 코돈을 가장 많이 사용하고 모든 CGN이 저표현된 것을 확인할 수가 있었다(Van Hemert & Berkhout, 1995). 본 연구에서 Arg 아미노산의 이러한 동의코돈 사용 패턴은 모든 영장류 렌티바이러스 그룹내에서 동일하여 극도로 보존적인 것으로 확인되었다. 또한 기존 연구에서 HIV1의 2코돈 아미노산 Cys의 사용이 인간의 두 가지 코돈에 대한 평균 사용과 달리 TGT코돈에 극히 편향된 것이 확인되었다(Van Hemert & Berkhout, 1995). HIV2와 기타 SIV에 대하여 RSCU 분석을 진행한 결과 gag 유전자에서는 HIV1, pol 유전자에서는 HIV1과 SIVcpz_gor가 이러한 코돈 사용 편향을 갖고 있고 기타 영장류 렌티바이러스는 gag, pol 유전자에서 인간에 가까운 TGT, TGC를 평균적으로 사용하는 패턴이 발견되었다. PR2 분석에서 gag, pol 유전자의 모든 4코돈 아미노산에서 A>T 편향이 관찰되고 있고 Val, Gly에서는 G>C 편향을 보이지만 Pro, Thr, Ala

아미노산에서는 C>G 편향을 보이고 있다. 이는 Pro, Thr, Ala 아미노산에서의 CpG 디뉴클리오티드 사용을 피하기 위한 것으로 생각된다. 또한 pol 유전자의 Val 아미노산에서 HIV2 그룹은 유일하게 GTG보다 GTC를 더 많이 사용하고 있다. Env 유전자의 Ala 아미노산의 코돈 사용에서 SIVsm, SIVmacaca 그룹은 gag와 pol 유전자에서의 심한 GCC 코돈 사용과 달리 GCC 보다 GCG를 많이 사용하고 HIV2와 SIVcerco 그룹에서는 두 코돈의 사용이 비슷한 것이 확인되었다. Env 유전자에 대한 계통수 분석에서도 이 4그룹(SIVsm, SIVmacaca, HIV2, SIVcerco)이 가장 가까운 근연관계를 보이고 있고 Ala의 코돈 사용에서도 특징적인 연관성을 나타내고 있다. 따라서 향후 이상 코돈에 대한 사용 편집으로 생물학적으로 중요한 기능을 하고 있는지를 확인할 수 있다. 또한 ENC-GC3s plot과 중립진화 분석으로부터 모든 영장류 렌티바이러스가 유전자의 GC조성에 의한 돌연변이 보다는 선택압력을 훨씬 많이 받고 있는 것을 확인할 수가 있었고 이는 기존 HIV1에 대한 연구결과와 일치하였다. CAI 및 similarity D(A, B) 분석으로부터 HIV는 gag, pol, env 세 유전자에서 모두 인간의 코돈 사용과 높은 유사성이 측정되었지만 번역압력을 크게 받고 있지 않는 것이 확인되었다. 또한 HIV1에 비해 HIV2의 코돈 사용이 인간과 더 비슷하였지만 전반적인 코돈 사용 편향은 오히려 낮았다. 특이적인 결과는 HIV의 기원 바이러스인 SIVcpz_gor, SIVsm 그룹이 gag 유전자에서 인간의 코돈 사용에 대한 유사성이 HIV보다 높게 관찰되었고 인간의 코돈 사용에 유의한 적응을 하고 있는 것이 확인되었다. 이는 HIV1와 HIV2가 SIV로부터 인간으로 종간전파 후 gag 유전자에서는 선택압력에 의해 인간의 코돈 사용을 피하는 방향으로 진화된 것일 수 있다는 의미이다.

본 연구는 영장류 렌티바이러스의 전반적인 코돈 사용 패턴과

그룹별 차이를 확인하고, 질병의 양상과 바이러스의 기원에 따라 HIV1, HIV2, SIV에서 공통적 또는 특이적으로 발현되는 진화적·유전적 특성을 탐색하는 것을 주 목적으로 하여 연구를 설계하고 진행하였다. 이에 따라 gag, pol, env 유전자의 계통수 토폴러지와 SIV 숙주의 분류체계를 근거로 코돈 사용 패턴 분석을 위한 서열데이터의 그룹핑을 수행하였기 때문에 그룹 내의 패턴 차이, 보다 세밀한 수준의 HIV lineage 정보 및 지리적 차이는 고려되지 않았다는 한계점을 지닌다. 따라서 이를 고려한 후속 연구를 통해 숙주 면역 반응이나 이에 대한 바이러스의 감수성(susceptibility)을 포함하여 인간 숙주에 보다 특이적인 바이러스의 특성들을 확인할 수 있다면, 본 연구의 결과와 함께 다양한 관점을 제시할 수 있을 것이다. 코돈의 역최적화(deoptimization), 디뉴클리오티드 사용 및 코돈 페어 사용과 같은 정보는 새로운 약독화 백신의 생성을 위해 여러 RNA 바이러스 유전체에 적용될 수 있다. 약독화 백신의 안전성과 안정성의 문제로 인해 아직까지도 약독화 HIV1 백신의 성공적인 개발은 어려운 상황이기 때문에, 코돈 분석 결과 기반의 유용 정보를 백신개발에 직접 활용하기까지는 앞으로 많은 시간이 필요할지도 모른다. 그러나 재코딩된 HIV1 변이체는 다양한 미생물에 대한 면역화를 위한 백신 벡터로써 유전자 치료에 사용되거나, 또는 특정 질병 상태를 치료하기 위한 특이적인 선천 면역 반응을 유도하기 위한 면역 요법에 사용될 수 있다. 따라서 향후 HIV1에 대한 추가적인 연구를 통해 디뉴클리오티드 사용 및 코돈 페어 사용 분석 결과를 확보하여 체내에서의 병원성, 숙주와의 상호작용, 면역저해 또는 항바이러스 작용 등의 기전과 관련된 유용 정보를 구명한다면 에이즈 치료에 있어 새로운 방향을 제시할 수 있을 것이다.

# 제 5 장. 결론

현재 HIV에 대한 코돈 분석은 HIV1의 구조 유전자 및 보조 유전자에 대한 코돈 사용 분석, 시간에 따른 코돈 사용 패턴의 변화(연도 또는 환자 기준 전기/후기), 다른 질병양상을 띠고 있는 HIV2와의 비교 분석, 지리적 분포에 따른 코돈 사용 패턴 분석(Ahn & Son, 2006) 등으로 이루어지고 있다. HIV의 기원 바이러스인 SIV에 대한 코돈 분석 및 HIV와의 비교 분석이 미비하고 기존 연구에서의 제한적인 코돈 인덱스 사용이 파악되어, 본 연구에서는 코돈 인덱스 산출 프로그램 개발을 기반으로 영장류 렌티바이러스에 대한 코돈 사용 비교분석을 진행하였다. 개발된 CUIs 프로그램은 사용성 및 웹 서버 기반 프로그램의 불안정성 요소를 고려하여 GUI 기반 스탠드얼론 프로그램으로 개발하였고, FASTA 파일로부터 코돈 분석에 이용될 수 있는 11가지의 데이터 산출과 클리닝 기능을 지원하고 있다. 영장류 렌티바이러스에 대한 코돈 분석은 gag, pol, env 유전자의 뉴클리오티드 조성, 전반적인 코돈 사용 편향, 진화압력을 형성하는 돌연변이/선택 압력의 상대적 크기, 아미노산별 사용 편향, 인간의 코돈 사용에 대한 적응정도 등에 대하여 비교분석을 진행하였다. 구조 유전자별로 현재 이용가능한 모든 서열 데이터를 포함시킴으로써 결과값의 신뢰성을 높였고, 코돈 사용면에서 진화적으로 받고 있는 주요압력과 그 작용방향의 확인이 가능하였다. 또한 영장류 렌티바이러스 전체에서 보존적인 코돈 사용 패턴이 보이는 아미노산을 제시하고, 코돈 사용의

측면에서 HIV1과 HIV2의 감염양상 차이에 기여 가능성이 있는 부분을 확인할 수 있었다.

개발된 CUIs 프로그램은 JAVA swing 을 사용하여 개발되었으며 크게 서열자체만으로 계산 가능한 인덱스 와 레퍼런스 셋이 필요한 인덱스 계산으로 두 흐름을 갖고 있다. 레퍼런스 셋은 CUD 와 HIVE-CUTs 에서 35,799 가지 생물 및 52 가지 인간 조직(tissue)의 코돈 사용 테이블을 프로그램에 내재하여 레퍼런스 셋이 필요한 인덱스 계산의 활용성을 넓혔다. CUIs 프로그램은 RSCU, ENC, ICDI, CAI, GRAVY, AROM 등의 기본 인덱스 이외에 기존 인덱스 계산 툴을 제공하고 있지 않는 ICDI, RCA, similarity D(A, B) 인덱스를 통합하여 제공하도록 설계 및 프로그래밍하여 관련 분석 시 필요한 모든 인덱스의 산출과 활용이 가능하게 하였다. 또한 GUI 기반 스탠드얼론 시스템으로 사용성 및 안정성을 보장하였고 서열 체크로 입력된 데이터의 클리닝 작업을 지원함으로써 연구자의 편리성을 높였다.

개발된 프로그램을 활용한 영장류 렌티 바이러스 코돈 분석 연구에서는 먼저 계통수 구축을 통해 SIV 의 유전적 근연관계를 확인하였고 이것을 기반으로 SIV 의 분류체계를 함께 고려하여 SIV 의 그룹핑 작업을 진행하였다. 최종적으로 총 11 개 그룹이 코돈 비교 분석에 사용되었다. 뉴클레오티드 조성, PR2 및 RSCU 분석으로부터 세 유전자(gag, pol, env)에서 모두 극도로 심한 A 편향을 확인할 수 있었고 CpG 디뉴클리오티드의 사용을 피하는 패턴을 확인할 수가 있었다. 4 코돈 아미노산 중 Pro 와 Ala 및 6 코돈 아미노산 Arg 에서의 코돈 사용이 이런 특징을 특히 잘 보여주고 있다. 세 유전자에서 모두 CCA, ACA, GCA, AGA, AGG, GGA 코돈의 RSCU 값이 1.6 이상으로 과표현되어 해당 아미노산의 코돈 사용 패턴이 영장류 그룹 전체에서 보존적인 양상을 보이고 있다. 이는

기존 gp41 유전자 구역에서 코돈 사용 편집으로 확인된 AGG→CGU 치환이 바이러스의 기능 및 env 유전자의 발현을 중지한 결과와 함께 Pro, Thr, Ala, Arg, Gly 아미노산의 과표현 코돈이 바이러스의 생활사에 중요한 역할을 수행하고 있다는 것을 시사하는 근거가 될 수 있다. 또한 gag 와 pol 유전자에서 HIV1 의 TGT(Cys) 코돈 사용이 기타 영장류 렌티바이러스 그룹 및 인간의 코돈 사용과 달리 특이적으로 과표현된 부분에서 HIV1 의 고병원성으로 작용하는 잠재적 기작과의 연관성을 유추해볼 수 있다. ENC 분석에서 HIV2 에 비해 HIV1 의 코돈 사용 편향이 높은 결과를 관찰할 수 있었고 ENC-GC3s plot 및 중립진화 분석으로부터 영장류 렌티바이러스 전체가 돌연변이 압력보다 선택압력을 훨씬 많이 받고 있다는 것을 확인할 수가 있었다. CAI 및 similarity D(A, B) 분석에서 HIV1 보다 HIV2 의 코돈 사용이 인간과 더 유사하였고 gag 유전자에서 HIV 의 기원 바이러스인 SIVcpz_gor 와 SIVsm 그룹이 인간의 코돈 사용과 더 높은 유사성이 확인됨으로써 HIV 는 선택압력에 의해 gag 유전자에서 인간의 코돈 사용을 피하는 방향으로 진화하였다는 것을 확인할 수 있었다. 이상의 결과는 HIV 코돈 편집의 참고자료로 활용되어 새로운 치료 타겟 발견에 기여할 수 있으며, 연구에서 개발된 프로그램 또한 향후 바이러스 코돈 연구에 유용하게 사용될 수 있을 것으로 기대된다.

# 참 고 문 헌


Ahn, I., & Son, H. S. (2006). Epidemiological comparisons of codon usage patterns among HIV-1 isolates from Asia, Europe, Africa and the Americas. Experimental & molecular medicine, 38(6), 643-651.

Apetrei, C., Kaur, A., Lerche, N. W., Metzger, M., Pandrea, I., Hardcastle, J., ... & Marx, P. A. (2005). Molecular epidemiology of simian immunodeficiency virus SIVsm in US primate centers unravels the origin of SIVmac and SIVstm. Journal of virology, 79(14), 8991-9005.

Athey, J., Alexaki, A., Osipova, E., Rostovtsev, A., Santana-Quintero, L. V., Katneni, U., ... & Kimchi-Sarfaty, C. (2017). A new and updated resource for codon usage tables. BMC bioinformatics, 18(1), 1-10.

Deml, L., Bojak, A., Steck, S., Graf, M., Wild, J., Schirmbeck, R., ... & Wagner, R. (2001). Multiple effects of codon usage optimization on expression and immunogenicity of DNA candidate vaccines encoding the human immunodeficiency virus type 1 Gag protein. Journal of virology, 75(22), 10991-11001.

Desimmie, B. A., Delviks-Frankenberrry, K. A., Burdick, R. C., Qi, D., Izumi, T., & Pathak, V. K. (2014). Multiple APOBEC3 restriction factors for HIV-1 and one Vif to rule them all. Journal of molecular biology, 426(6), 1220-1245.

Drummond, D. A., & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell, 134(2), 341-352.

Esbjörnsson, J., Månsson, F., Kvist, A., Isberg, P. E., Nowroozalizadeh, S., Biague, A. J., ... & Medstrand, P. (2012). Inhibition of HIV-1 disease progression by contemporaneous HIV-2 infection. New England Journal of Medicine, 367(3), 224-232.

Ficarelli, M., Antzin-Anduetza, I., Hugh-White, R., Firth, A. E., Sertkaya, H., Wilson, H., ... & Swanson, C. M. (2020). CpG dinucleotides inhibit HIV-1 replication through zinc finger antiviral protein (ZAP)-dependent and-independent mechanisms. Journal of virology, 94(6).

Foley, B. T. (2000). An overview of the molecular phylogeny of lentiviruses. HIV sequence compendium, 2000, 35-43.

Fox, J. M., & Erill, I. (2010). Relative codon adaptation: a generic codon bias index for prediction of gene expression. DNA research, 17(3), 185-196.

Freire-Picos, M. A., Gonzalez-Siso, M. I., Rodríguez-Belmonte, E., Rodríguez-Torres, A. M., Ramil, E., & Cerdan, M. E. (1994). Codon usage in Kluyveromyces lactis and in yeast cytochrome c-encoding genes. Gene, 139(1), 43-49.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., ... & Hahn, B. H. (1999). Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature, 397(6718), 436-441.

Gao, F., Li, Y., Decker, J. M., Peyerl, F. W., Bibollet-Ruche, F., Rodenburg, C. M., ... & Hahn, B. H. (2003). Codon usage optimization of HIV type 1 subtype C gag, pol, env, and nef genes: in vitro expression and immune responses in DNA-vaccinated mice. AIDS research and human retroviruses, 19(9), 817-823.

Gardner, M. B., & Luciw, P. A. (2008). Macaque models of human infectious disease. ILAR journal, 49(2), 220-255.

Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., ... & Korber, B. (2002). Diversity considerations in HIV-1 vaccine selection. Science, 296(5577), 2354-2360.

Goodman, D. B., Church, G. M., & Kosuri, S. (2013). Causes and effects of N-terminal codon bias in bacterial genes. Science, 342(6157), 475-479.

Hemelaar, J. (2012). The origin and diversity of the HIV-1 pandemic. Trends in molecular medicine, 18(3), 182-192.

Hizi, A., & Herzig, E. (2015). dUTPase: the frequently overlooked enzyme encoded by many retroviruses. Retrovirology, 12(1), 1-15.

Holmes, E. C. (2001). On the origin and evolution of the human immunodeficiency virus (HIV). Biological reviews, 76(2), 239-254.

Jenkins, G. M., & Holmes, E. C. (2003). The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus research, 92(1), 1-7.

Jiang, C., Lian, X., Gao, C., Sun, X., Einkauf, K. B., Chevalier, J. M., ... & Xu, G. Y. (2020). Distinct viral reservoirs in individuals with spontaneous control of HIV-1. Nature, 585(7824), 261-267.

Jiang, Y., Deng, F., Wang, H., & Hu, Z. (2008). An extensive analysis on the global codon usage pattern of baculoviruses. Archives of virology, 153(12), 2273.

Jordan-Paiz, A., Franco, S., & Martínez, M. A. (2021). Impact of Synonymous Genome Recoding on the HIV Life Cycle. Frontiers in Microbiology, 12, 575.

Jordan-Paiz, A., Nevot, M., Lamkiewicz, K., Lataretu, M., Franco, S., Marz, M., & Martinez, M. A. (2020). HIV-1 Lethality and Loss of Env Protein Expression Induced by Single Synonymous Substitutions in the Virus Genome Intronic-Splicing Silencer. Journal of Virology, 94(21).

Keele, B. F., Jones, J. H., Terio, K. A., Estes, J. D., Rudicell, R. S., Wilson, M. L., ... & Hahn, B. H. (2009). Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz. Nature, 460(7254), 515-519.

Kishino, H., & Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. Journal of molecular evolution, 29(2), 170-179.

Kofman, A., Graf, M., Bojak, A., Deml, L., Bieler, K., Kharazova, A., ... & Wagner, R. (2003). HIV-1 gag expression is quantitatively dependent on the ratio of native and optimized codons. Tsitologiia, 45(1), 86-93.

Kofman, A., Graf, M., Deml, L., Wolf, H., & Wagner, R. (2003). Codon usage-mediated inhibition of HIV-1 gag expression in mammalian cells occurs independently of translation. Tsitologiia, 45(1), 94-100.

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. science, 324(5924), 255-258.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular biology and evolution, 35(6), 1547-1549.

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. Journal of molecular biology, 157(1), 105-132.

LI, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36:96–99.

Link, J. O., Rhee, M. S., Winston, C. T., Zheng, J., Somoza, J. R., Rowe, W., ... & Cihlar, T. (2020). Clinical targeting of HIV capsid protein with a long-acting small molecule. Nature, 584(7822), 614-618.

Lobry, J. R., & Gautier, C. (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic acids research, 22(15), 3174-3180.

Mancuso, P., Chen, C., Kaminski, R., Gordon, J., Liao, S., Robinson, J. A., ... & Khalili, K. (2020). CRISPR based editing of SIV proviral DNA in ART treated non-human primates. Nature communications, 11(1), 1-11.

Meintjes, P. L., & Rodrigo, A. G. (2005). Evolution of relative synonymous codon usage in Human Immunodeficiency Virus type-1. Journal of bioinformatics and computational biology, 3(01), 157-168.

Moore, J. P., Kitchen, S. G., Pugach, P., & Zack, J. A. (2004). The CCR5 and CXCR4 coreceptors—central to understanding the transmission and pathogenesis of human

immunodeficiency virus type 1 infection. AIDS research and human retroviruses, 20(1), 111-126.

Motomura, K., Chen, J., & Hu, W. S. (2008). Genetic recombination between human immunodeficiency virus type 1 (HIV-1) and HIV-2, two distinct human lentiviruses. Journal of virology, 82(4), 1923-1933.

Nakamura, Y., Gojobori, T., & Ikemura, T. (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucleic acids research, 28(1), 292-292.

Nasrullah, I., Butt, A. M., Tahir, S., Idrees, M., & Tong, Y. (2015). Genomic analysis of codon usage shows influence of mutation pressure, natural selection, and host features on Marburg virus evolution. BMC evolutionary biology, 15(1), 1-15.

Neel, C., Etienne, L., Li, Y., Takehisa, J., Rudicell, R. S., Bass, I. N., ... & Peeters, M. (2010). Molecular epidemiology of simian immunodeficiency virus infection in wild-living gorillas. Journal of virology, 84(3), 1464-1476.

Pandit, A., & Sinha, S. (2011). Differential trends in the codon usage patterns in HIV-1 genes. PloS one, 6(12), e28889.

Peden, J. F. (2000). Analysis of codon usage.

Peeters, M., & Courgnaud, V. (2002). Overview of primate lentiviruses and their evolution in non-human primates in Africa. HIV sequence compendium, 2.

Puigbò, P., Bravo, I. G., & Garcia-Vallve, S. (2008). CAIcal: a combined set of tools to assess codon usage adaptation. Biology direct, 3(1), 1-8.

Puigbò, P., Bravo, I. G., & Garcia-Vallvé, S. (2008). E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). BMC bioinformatics, 9(1), 1-7.

Reis, M. D., Savva, R., & Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic acids research, 32(17), 5036-5044.

Rey-Cuillé, M. A., Berthier, J. L., Bomsel-Demontoy, M. C., Chaduc, Y., Montagnier, L., Hovanessian, A. G., & Chakrabarti, L. A. (1998). Simian immunodeficiency virus replicates to high levels in sooty mangabeys without inducing disease. Journal of virology, 72(5), 3872-3886.

Robertson, D. L., Anderson, J. P., Bradac, J. A., Carr, J. K., Foley, B., Funkhouser, R. K., ... & Korber, B. (2000). HIV-1 nomenclature proposal. Science, 288(5463), 55-55.

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., & Sánchez-Gracia, A. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. Molecular biology and evolution, 34(12), 3299-3302.

Sabi, R., Volvovitch Daniel, R., & Tuller, T. (2017). stAIcalc: tRNA adaptation index calculator based on species-specific weights. Bioinformatics, 33(4), 589-591.

Shabalina, S. A., Ogurtsov, A. Y., & Spiridonov, N. A. (2006). A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic acids research, 34(8), 2428-2437.

Shackelton, L. A., Parrish, C. R., & Holmes, E. C. (2006). Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. Journal of molecular evolution, 62(5), 551-563.

Sharp, P. M., & Hahn, B. H. (2010). The evolution of HIV-1 and the origin of AIDS. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1552), 2487-2494.

Sharp, P. M., & Li, W. H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. Journal of molecular evolution, 24(1-2), 28-38.

Sharp, P. M., & Li, W. H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic acids research, 15(3), 1281-1295.

Sharp, P. M., Bailes, E., Robertson, D. L., Gao, F., & Hahn, B. H. (1999). Origins and evolution of AIDS viruses. The Biological Bulletin, 196(3), 338-342.

Sharp, P. M., Shaw, G. M., & Hahn, B. H. (2005). Simian immunodeficiency virus infection of chimpanzees. Journal of virology, 79(7), 3891-3902.

Sheehy, A. M., Gaddis, N. C., Choi, J. D., & Malim, M. H. (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature, 418(6898), 646-650.

Silverj, A., & Rota-Stabelli, O. (2020). On the correct interpretation of similarity index in codon usage studies: Comparison with four other metrics and implications for Zika and West Nile virus. Virus Research, 286, 198097.

Smith, J. M., Rao Amara, R., Campbell, D., Xu, Y., Patel, M., Sharma, S., ... & Robinson, H. L. (2004). DNA/MVA vaccine for HIV type 1: effects of codon-optimization and the expression of aggregates or virus-like particles on the immunogenicity of the DNA prime. AIDS Research & Human Retroviruses, 20(12), 1335-1347.

Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. Proceedings of the National Academy of Sciences, 85(8), 2653-2657.

Sueoka, N. (1995). Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. Journal of molecular evolution, 40(3), 318-325.

Sueoka, N. (1999). Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+ C content of third codon position. Gene, 238(1), 53-58.

Supek, F., & Vlahoviček, K. (2004). INCA: synonymous codon usage analysis and clustering by means of self-organizing map. Bioinformatics, 20(14), 2329-2330.

Takehisa, J., Kraus, M. H., Ayouba, A., Bailes, E., Van Heuverswyn, F., Decker, J. M., ... & Hahn, B. H. (2009). Origin and biology of simian immunodeficiency virus in wild-living western gorillas. Journal of virology, 83(4), 1635-1648.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research, 22(22), 4673-4680.

van der Kuyl, A. C., & Berkhout, B. (2012). The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. Retrovirology, 9(1), 1-14.

Van Hemert, F. J., & Berkhout, B. (1995). The tendency of lentiviral open reading frames to become A-rich: constraints imposed by viral genome organization and cellular tRNA availability. Journal of molecular evolution, 41(2), 132-140.

Van Weringh, A., Ragonnet-Cronin, M., Pranckeviciene, E., Pavon-Eternod, M., Kleiman, L., & Xia, X. (2011). HIV-1 modulates the tRNA pool to improve translation efficiency. Molecular biology and evolution, 28(6), 1827-1834.

Vartanian, J. P., Henry, M., & Wain-Hobson, S. (2002). Sustained G→ A hypermutation during reverse transcription of an entire human immunodeficiency virus type 1 strain Vau group O genome. Journal of General Virology, 83(4), 801-805.

Vidyavijayan, K. K., Hassan, S., Precilla, L. K., Ashokkumar, M., Chandrasekeran, P., Swaminathan, S., & Hanna, L. E. (2017). Biased Nucleotide Composition and Differential Codon Usage Pattern in HIV-1 and HIV-2. AIDS research and human retroviruses, 33(3), 298-307.

Visseaux, B., Damond, F., Matheron, S., Descamps, D., & Charpentier, C. (2016). Hiv-2 molecular epidemiology. Infection, Genetics and Evolution, 46, 233-240.

Wong, E. H., Smith, D. K., Rabadan, R., Peiris, M., & Poon, L. L. (2010). Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. BMC evolutionary biology, 10(1), 1-14.

Yan, N., O'Day, E., Wheeler, L. A., Engelman, A., & Lieberman, J. (2011). HIV DNA is heavily uracilated, which protects it from autointegration. Proceedings of the National Academy of Sciences, 108(22), 9244-9249.

Yu, Q., Chen, D., König, R., Mariani, R., Unutmaz, D., & Landau, N. R. (2004). APOBEC3B and APOBEC3C are potent inhibitors of simian immunodeficiency virus replication. Journal of Biological Chemistry, 279(51), 53379-53386.

Zhou, J. H., Zhang, J., Sun, D. J., Ma, Q., Chen, H. T., Ma, L. N., ... & Liu, Y. S. (2013). The distribution of synonymous codon choice in the translation initiation region of dengue virus. PloS one, 8(10), e77239.

# ABSTRACT

## GENETIC SEQUENCE ANALYSIS OF LENTIVIRUS USING BIOINFORMATICS METHODS

**Xianglan Min**
**Laboratory of Computational Biology and Bioinformatics**
**Graduate School of Public Health**
**Seoul National University**

HIV1 is the representative species of the lentivirus genus, leading to AIDS worldwide and remains an ongoing public health problem. Nearly 40 years after the first AIDS case was reported, it is still very difficult to develop the vaccine and cannot cure it. A deep understanding of the viral codon usage pattern can be important information for predicting the evolution direction of viruses or as reference data for codon editing, revealing the antiviral mechanism of hosts and unknown parts of the viral life cycle. At present, the codon research on SIV, the origin of HIV, is very insufficient. So in this study, the main purpose is to explore the evolutionary and genetic characteristics that are commonly/specifically expressed in HIV1, HIV2, and SIV. Before the analysis, I developed a codon usage analysis program CUIs using JAVA language. Considering usability and stability, the CUIs is developed as a GUI-based standalone system and supports the calculation of RSCU, ENC, ICDI, CAI, similarity D(A, B) indices. AT, GC bias values for each amino acid required by PR2 analysis and indices such as tAI, RCA, GRAVY, AROMO were also included. For primate lentivirus codon analysis, first, a phylogenetic tree was

constructed for each of the gag, pol, and env genes to confirm the evolutionary relationship, and based on phylogenetic topology the grouping work was conducted in consideration of the SIV host classification system. Codon usage analysis evaluated the nucleotide composition, overall codon usage bias, the relative magnitude of mutational and selectional pressure, codon usage bias pattern, and the degree of adaptation to the human codon usage. As a result, although similarity with human codon usage is not low, it was confirmed that HIV is not evolving toward similarity with human codon usage to increase translational efficiency. In addition, while the overall codon usage bias of primate lentivirus is low, HIV1 shows a higher codon usage bias compared to HIV2 and can withstand higher selection pressure. This could be the basis for HIV1 to have higher viability and expression, leading to a large epidemic. Significant adaptation to human codon usage was observed in the gag genes of SIVcpz_gor and SIVsm, which may also be evidence supporting the origin of HIV in terms of codon usage. In addition, it was possible to suggest the specific codon usage such as TGT of HIV1 and CTA and GTC of HIV2, which are inferred to be related to potential mechanisms acting on the difference in pathogenicity between HIV1 and HIV2. It was also identified that the codon usage of Pro, Thr, Ala, Arg, and Gly amino acids was extremely conservative throughout the primate lentivirus, which is likely to have a significant impact on virus survival and life cycle progression. In conclusion, the result of this study about codon usage of lentivirus can be used as reference data for codon editing and contribute to the discovery of new treatment targets for HIV, and the newly developed program is also expected to be useful for codon usage research.

……

keywords : Bioinformatics, HIV1, Lentivirus, Codon usage pattern

Student Number : 2019-28056