Journal of Bacteriology and Virology 2013. Vol. 43, No. 4 p.317 – 327
http://dx.doi.org/10.4167/jbv.2013.43.4.317

Original Article

Codon Usage Bias of Human Cytomegalovirus Genes with Different Evolutionary Conservancy

Yu Young Kim and Chan Hee Lee*

Department of Microbiology, Chungbuk National University, Cheongju, Chungbuk, Korea

Human cytomegalovirus (HCMV) is a member of beta-herpesvirus and contains a double-stranded genome with longer than 230 Kbp. HCMV infection of human is mostly asymptomatic, but often causes fatal diseases in immunocompromised people. In this study, codon usages of HCMV genes were analyzed and attempted to correlate with evolutionary conservancy. Core genes are the most conserved genes common among herpesvirus family, β-herpes genes are common to β-herpesviruses, and CMV genes are the least conserved found only in CMVs. Core genes had higher codon adaptation index (CAI) and GC content of silent 3rd codon position (GC3s) values and lower effective number of codons (Nc) and Nc/GC3s values than CMV genes. The average length of core genes was statistically longer than CMV genes, and core genes were found to be less varied than CMV genes. β-herpes genes could be placed between core and CMV genes. Higher CAI and GC3s values along with lower Nc and Nc/GC3s values are suggestive of higher codon usage bias and more adaptation to host cells. Thus it is concluded that core genes of HCMV are more biased in codon usage and adapted to host cells compared to CMV genes.

Key Words: Human cytomegalovirus, Codon usage bias

서 론

인간거대세포바이러스(Human cytomegalovirus, HCMV)는 베타 허피스바이러스에 속하며, Human herpesvirus type5 (HHV-5)라고도 불린다. 이중 가닥의 선형 DNA를 genome으로 가지며, genome의 평균 길이는 230~240 Kbp 정도로, 사람에 감염하는 바이러스 중에서는 큰 편에 속한다 (1). HCMV는 전 세계에 널리 퍼져있는 바이러스로서 초기 감염 이후 숙주 내에서 오랜 시간 잠복감염하며 숙주의 일생 동안 감염을 유지한다. 건강한 사람에게서는 증상을 나타내지 않으나 면역력이 결핍된 환자에서는 심각한 질병을 유발하며, 신생아에게서 발생하는 감염의 주된 요인 중 하나이기도 하다. HCMV의 선천성 감염은 신생아에서 뇌성마비나 간질, 청력감소 등의 장애를 유발한다고 알려져 있다 (2). 이러한 이유로 백신 개발의 중요성이 인지되어 지난 수십 년간 백신개발 연구가 이어져왔으나 아직까지 성공적인 백신은 개발되지 않았다 (3). HCMV의 genome은 총 165개의 유전자를 암호화 하고 있는데 (4) 보유하는 유전자의 개수는 바이러스주에 따라 약간씩 다를 수는 있으나 대체로 동일하다. 이 유전자들은 기능과 기원에 따라 여러 가지 그룹으로 분류될 수 있다 (5, 6).

본 연구에서는 HCMV 게놈의 염기서열을 분석함으로써 HCMV에 대한 이해도를 높이고 특징을 파악하고자 하였으며, 이를 위하여 유전자별로 코돈 사용 편향성의

Received: October 29, 2013/ Revised: November 11, 2013/ Accepted: November 15, 2013
*Corresponding author: Chan Hee Lee. Department of Microbiology, Chungbuk National University, 52 Naesudong-Ro, Heungdeok-Gu, Cheongju, Chungbuk 361-763, Korea.
Phone: +82-43-261-2304, Fax: +82-43-273-2451, e-mail: chlee@chungbuk.ac.kr
**This work was supported by the research grant of the Chungbuk National University in 2011.

Table 1. HCMV strains used in this study

Strains	Origin	GenBank accession number	Isolation date
AD169-UK	USA	NC_001347.6	1956
AD169-UC	USA	FJ527563.1	1956
TOWNE	USA	FJ616285.1	1970
HAN38	Germany	GQ396662.1	2007
HAN20	Germany	GQ396663.1	2007
HAN13	Germany	GQ221973.1	2007
3157	United Kingdom	GQ221974.1	2001
3301	United Kingdom	GQ466044.1	2001
JP	United Kingdom	GQ221975.1	2001
TOLEDO	USA	GU937742.1	1984
MERLIN	United Kingdom	NC_006273.2	1999
U8	Italy	GU179288	2003
VR1814	Italy	GU179289	1996
U11	Italy	GU179290	2003
AF1	Italy	GU179291	2003
JHC	South Korea	HQ380895	2003

정도를 밝히고 비교 분석하였다. 모든 생명체는 염기3개가 한 세트인 코돈 단위로 유전정보를 저장하고 전달한다. 그러나 대부분의 아미노산은 하나 이상의 코돈에 의해 암호화된다. 동일한 아미노산을 암호화하는 서로 다른 종류의 코돈을 동의코돈이라고 하는데, 이 동의코돈들은 균등하게 사용되지 않으며 특정 코돈의 사용이 다른 코돈들에 비하여 선호되는 경향이 있다. 이를 코돈 사용의 편향성(codon usage bias)이라고 하며, 이는 생물종에 따라 다르게 나타나기도 하고 같은 게놈 내에서도 유전자에 따라 다르게 나타나기도 한다 (7, 8). 염기서열의 조성과 코돈 사용 편향성의 정도 및 원인에 대해 이해하는 것은 바이러스의 진화와 바이러스-숙주간의 상호작용, 또는 면역 반응을 연구하기 위해 필수적이다 (9, 10). 또한 코돈 사용 편향성에 대한 정보는 바이러스의 유전자 발현 조절과정을 이해하는 데에 도움을 준다. 이러한 정보를 이용하여 면역성을 생성시키는 바이러스 단백질의 효과적인 발현을 조절하는 방법으로 백신 디자인에도 응용이 될 수 있다 (11, 12).

2013년 9월 현재 NCBI의 GenBank에 등록되어 있는 전체 염기 서열이 밝혀진 HCMV 바이러스주는 모두 16개이다(http://www.ncbi.nlm.nih.gov/nuccore/?term=HHV-5+%22complete+genome%22). 해당 바이러스주들에 대한 염기서열 정보는 쉽게 얻을 수 있으나 바이러스 치고는 큰 크기의 게놈을 가진다는 점과, 바이러스주간 염기서열의 다양성이 크다는 점 때문에 HCMV 염기서열에 대한 분석은 다른 바이러스들에 비해 많이 이루어지지 않았다. 특히 염기서열의 패턴과 유전자의 기원 사이의 연관성을 보는 등의 후속 연구는 아직 시작단계에 있다. HCMV는 게놈의 크기가 큰 만큼 다른 바이러스들에 비하여 많은 수의 유전자를 가지고 있으며, 유전자들은 다양한 분류체계에 따라 쉽게 분류된다는 장점을 갖는다. 이러한 점을 이용하여 본 연구에서는 유전자의 특성 중 진화적인 보존성에 따라 코돈 사용 편향성의 경향이 어떻게 나타나는지에 대하여 중점적으로 분석하고자 하였다. 또한 상대적인 유전자의 발현량과 유전자의 길이, 유전자별 진화거리를 분석항목에 추가하여 유전자 그룹별로 나타내는 특징을 구체화하고자 하였다.

재료 및 방법

HCMV의 전체 게놈 염기서열

본 연구에서는 본 연구실에서 분리 보고한 JHC주를 비롯하여 NCBI(미국 국립생물공학정보센터, http://www.ncbi.nlm.nih.gov/)에서 제공하고 있는 16개 바이러스주의 전체 게놈 염기서열을 받아 분석의 재료로 사용하였다. 분석에 사용된 16개 바이러스주에 대한 정보는 Table 1에 나타낸 바와 같다.

전체 게놈 염기서열의 편집

16개 바이러스주의 전체 게놈 염기서열들은 분석에 사용하기 적합한 형태로 수정되었다. 먼저, NCBI 홈페이지로부터 각 바이러스주에 해당하는 유전정보를 얻은 뒤, 충북대학교 미생물학과 바이러스학 실험실에서 개발한 염기서열 분석 프로그램인 SeqMan ver.1.0을 이용하여 fasta 파일 형태의 전체 게놈 염기서열 파일과 text 파일 형태의 open reading frame (ORF) 별 주소파일을 추출하였다. 이렇게 얻어진 파일들을 다시 SeqMan 프로그램에 넣어, ORF별로 염기서열이 분리된 fasta 파일을 얻어낼 수 있었다. 이때 사용하는 주소파일이 ORF 전체의 염기서열에 대한 주소이면 유전자별로 ORF 전체의 DNA 염기서열을 가지는 파일이 생성되고, 주소파일이 유전자의

coding DNA sequence (CDS) 부분만을 지정한다면 생성되는 파일에는 유전자별로 CDS 부위의 염기서열만이 포함된다.

코돈 사용 분석

바이러스의 코돈 사용 경향을 분석하기 위해서 CodonW 프로그램(ver. 1.4.2, http://codonw.sourceforge.net)과 CAIcal 프로그램(ver. 1.4, http://genomes.urv.cat/CAIcal)을 사용하였다. 프로그램을 실행시키기 위한 input 파일로는, 전체 DNA 서열 중 실제로 단백질로 번역되는 부분인 CDS 부위의 염기서열만을 각 유전자 별로 분리한 fasta 파일을 사용하였다. CodonW 프로그램을 이용하여 코돈 사용의 경향을 나타내는 여러 가지 항목 중 사용된 코돈의 개수를 나타내는 'Effective number of codons' (ENc, 또는 Nc)와 코돈의 3번째 염기자리에 G나 C가 자리하는 비율을 나타내는 'GC content of silent 3rd codon position' (GC3s)를 얻었으며, 분석을 위한 기본 설정으로 Universal genetic code를 선택하였다. 또한, CAIcal 프로그램을 사용하여 'Codon adaptation index' (CAI) 값을 얻었으며 수치계산을 위하여 CAIcal 프로그램의 홈페이지에서 제공하는 *Homo sapiens*의 코돈 사용 표(http://genomes.urv.es/CAIcal/CU_human_nature)를 input host file로 사용하였다. 이때도 유전 코드의 기본값으로는 Universal genetic code를 지정하였다. 이상의 Nc와 GC3s, CAI와 같이 코돈 사용의 편향성의 정도와 경향을 수치화 한 항목들을 Codon usage bias index (CUI)라 한다.

유전자의 길이

NCBI 홈페이지로부터 다운로드받은 GenBank 파일을 SeqMan 프로그램에 넣어 바이러스주별로 각 ORF에 대한 시작 염기와 마지막 염기의 위치를 나타낸 주소파일을 추출해 낼 수 있었다. 이 ORF별 주소파일에서 ORF의 시작 위치와 마지막 위치의 주소까지의 차이에 1을 더해 해당 유전자의 길이를 구할 수 있었다.

유전자의 진화적인 변이거리

염기서열간의 진화적인 변이거리를 나타내는 유전적 거리를 구하기 위해서 PHYLIP 프로그램(ver.3.69, http://evolution.genetics.washington.edu/phylip.html)을 사용하였으며, 해당 프로그램을 실행시키기 위한 input 파일인 phylip 파일을 만들기 위하여 ClustalW2 프로그램(ver. 2.0.1, http://www.ebi.ac.uk)을 사용했다. 앞선 염기서열 편집과정의 결과 얻어진 HCMV 16개 바이러스주에 대한 ORF별 염기서열의 fasta 파일을 input 파일로 넣어 ClustalW2 프로그램으로부터 각 ORF에 대한 phylip 파일을 얻었다. 이렇게 얻어진 phylip 파일로 PHYLIP 프로그램의 DNAdist를 실행하였고 ORF별 distance 파일을 얻을 수 있었다. 진화적인 변이거리 계산을 위한 방법으로 Kimura 2-parameter를 지정하였고, ORF별로 얻어진 HCMV 16개 바이러스주간의 상호 진화적인 변이거리의 평균값을 분석에 사용하였다.

JHC주에 감염된 세포의 전사체 분석

전사체(transcriptome) 분석의 재료로서 JHC 바이러스에 감염된 세포로부터 추출된 RNA를 사용하였다. JHC 바이러스는 Human embryonic lung fibroblast (HEL) 세포를 숙주세포로 삼아 배양하였으며, 10%의 fetal bovine serum (FBS)을 첨가한 DMEM 배지를 첨가한 뒤 37℃, 5% CO_2의 환경에서 배양하였다. 바이러스를 제외한 비특이적인 효과를 최소화하기 위하여 20,000 rpm에서 30분 동안 원심분리를 하여 세포부유물질을 제거하였으며, 바이러스 침전물은 serum-free DMEM에 재현탁 된 후 0.22 μm의 필터를 통해 걸렀다. 사용된 JHC 바이러스는 실험실에서 5번의 계대를 거친 세대이다. 세포배양용 100 mm dish에서 4일간 배양한 HEL 세포에 1 plaque forming unit/cell로 JHC 바이러스를 감염시켰다. 한 시간 동안 세포에 바이러스를 접종한 후, 바이러스를 제거한 뒤 2%의 FBS를 첨가한 DMEM으로 배지를 교체해 주었다. 배지 교체 후 24시간 후에 세포로부터 RNA를 추출하였다. 바이러스의 mRNA에 대한 transcriptome 분석 결과는 (주)천랩에 RNA sequencing (RNA-seq)을 의뢰한 결과 얻을 수 있었다. JHC에 속하는 리드들을 각 ORF별로 나눈 뒤, 이를 상대적인 수치로 환산하여 나타냈는데 그 단위가 Reads per kilobase of exon per million mapped sequence reads (RPKM)이다. 이 RPKM값은 상대적인 mRNA의 양을 나타내 주는 수치이다. 본 연구에서는 ORF별로 환산된 RPKM값을 분석의 자료로서 사용하였다.

통계적 분석

실험으로써 얻어진 데이터들의 통계적인 의미를 알아보기 위해 여러 가지 통계 분석을 시행하였으며 이를 위해 통계 분석을 위한 프로그램인 SPSS 12.0 for windows

Table 2. Grouping of HCMV genes according to evolutionary conservation

CORE (n = 42)			β-HERPES (n = 27)		CMV (n = 69)			
UL44	UL45	UL46	UL23	UL24	RL1	RL6	RL10	RL11
UL47	UL48	UL48A	UL25	UL27	RL12	RL13	UL2	UL4
UL49	UL50	UL51	UL29	UL31	UL5	UL6	UL7	UL8
UL52	UL53	UL54	UL32	UL33	UL9	UL10	UL11	UL13
UL55	UL56	UL57	UL35	UL36	UL14	UL15A	UL16	UL17
UL69	UL70	UL71	UL38	UL43	UL18	UL19	UL20	UL21A
UL72	UL73	UL75	UL74	UL82	UL26	UL30	UL34	UL37
UL76	UL77	UL79	UL83	UL84	UL42	UL78	UL111A	UL116
UL80	UL85	UL86	UL88	UL91	UL119	UL120	UL121	UL123
UL87	UL89	UL93	UL92	UL96	UL124	UL130	UL132	UL146
UL94	UL95	UL97	UL112	UL117	UL147	TR1	IRS1	US1
UL98	UL99	UL100	UL122	US22	US2	US3	US6	US7
UL102	UL103	UL104	US23	US26	US8	US9	US10	US11
UL105	UL114	UL115	US28		US12	US13	US14	US15
					US16	US17	US18	US19
					US20	US21	US24	US27
					US29	US30	US31	US32
					US34			

(http://www-01.ibm.com/software/analytics/spss/)를 사용했다. Codon usage bias index (CUI)를 비롯한 데이터들의 그룹간 유의성과 그에 대한 신뢰도를 알아보기 위해 95% 신뢰수준에서 독립표본 *t*-test로 유의성을 검증하였다. CUI 및 유전자의 길이와 변이거리, RPKM값 각각에 대하여 그룹별로 데이터의 분포도를 알아보기 위하여 SigmaPlot (ver. 8.0) 프로그램을 이용하여 오차 막대가 있는 산점도 그래프를 그렸다.

결 과

본 연구에 사용된 HCMV 게놈들에 존재하는 ORF의 개수는 총 165개이며, 바이러스주에 따라 보유하는 ORF의 개수에는 약간씩 차이가 있으나 거의 동일하다. 165개의 ORF들은 다양한 기준에 따라 분류될 수 있다. 본 연구에서는 ORF를 분류하기 위한 방법으로 Dunn 등 (5)이 논문에서 제시한 group 분류방법을 사용하였다. ORF들은 진화적으로 보존되는 특성에 따라 허피스바이러스 전체에서 공통으로 나타나는 core gene (CORE), 베타 허피스바이러스에서 공통으로 나타나는 β-herpes gene (β-HERPES), 마지막으로 사이토메갈로바이러스에서만 공통적으로 발견되는 CMV gene (CMV)으로 구분할 수 있으며, 본 논문에서 실험에 사용한 총 165개의 HCMV ORF 중 138개의 ORF에 대한 분류기준을 Table 2에 정리하여 나타내었다.

유전자의 진화적인 보존성과 코돈 사용 특징

유전자의 진화적으로 보존되는 특성과 코돈 사용 경향간의 연관성을 알아보기 위하여, 유전자 그룹별로 코돈 사용의 편향 정도와 경향을 나타내는 codon usage bias index (CUI)를 비교하였다. ORF 그룹간 수치상에 의미 있는 차이가 있는지를 알아보기 위해 각 CUI 항목에 대하여 ORF의 그룹별로 오차막대가 있는 산점도를 그리고 두 그룹씩 *t*-test를 수행하였다. Fig. 1에 이를 나타내었으며, 이 그래프 상에는 *p*-value가 0.05 이하로 그룹간의 차이가 인정되는 것들에 한해서만 *p*-value값을 표시해 두

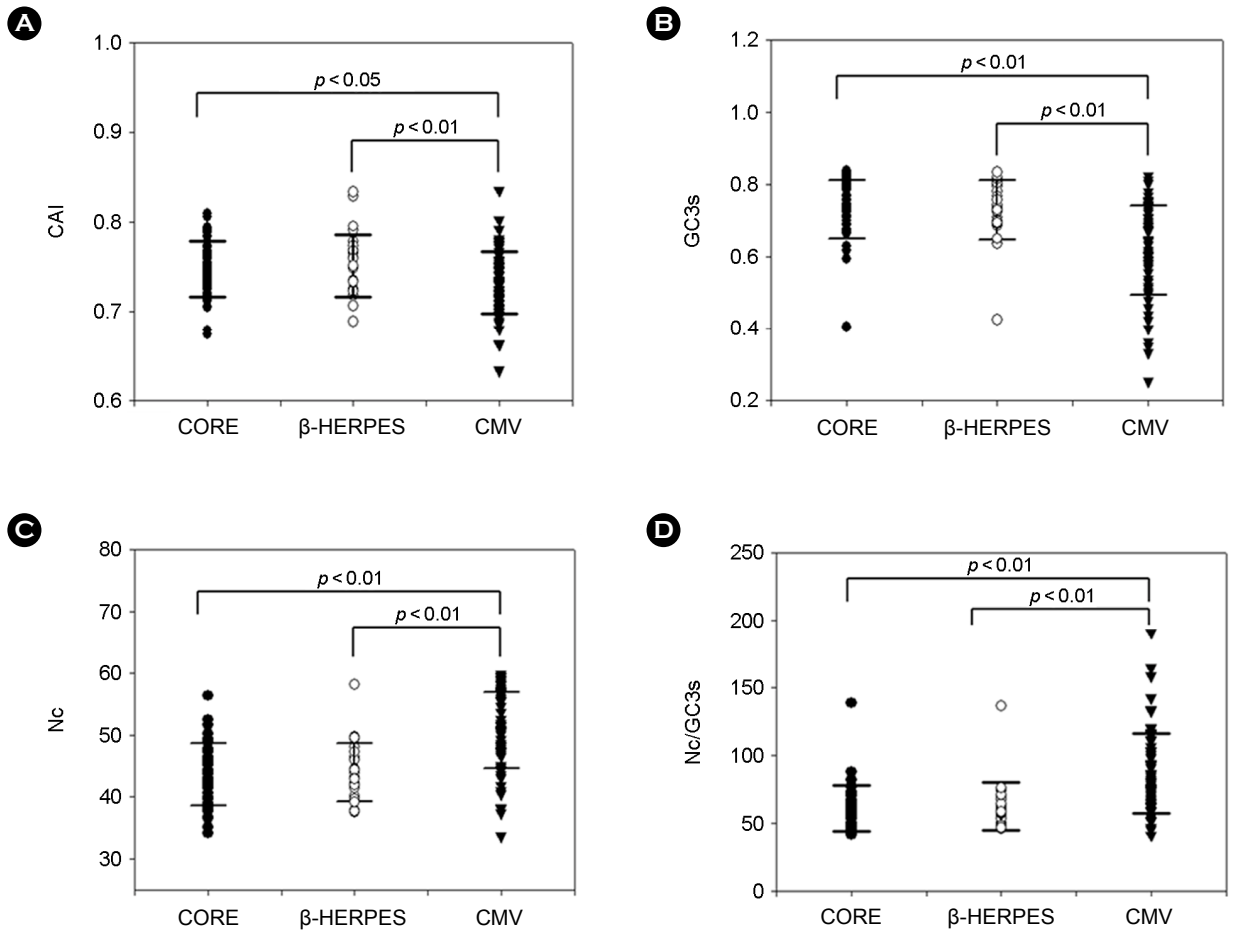

Figure 1. Codon usage bias index (CUI) values of HCMV genes. These graphs represent the distribution of CUI values of 3 groups. Each graph shows (A) CAI, (B) GC3s, (C) Nc, and (D) Nc/GC3s values. CORE genes are indicated by black circles, β-HERPES genes by white circles, and CMV genes by black triangles. Black bars indicate the standard error of the mean. Statistical significance of the difference in mean values was examined by student's *t*-test.

었다. 모든 항목에서 CORE와 β-HERPES 간의 *p*-value는 0.05 이상이었으며 이는 두 그룹간의 차이가 무시할 수 있을 정도로 작다는 사실을 말해준다. 반면 모든 항목에서 CORE와 CMV간, β-HERPES와 CMV간 *p*-value는 0.05 이하이므로, 그룹간의 차이가 유의하다는 사실을 알 수 있다.

CAI값은 해당 유전자의 코돈 사용 경향이 참고로 사용한 숙주의 코돈 사용 경향과 얼마나 유사한지를 보여주는 수치이다. CORE, β-HERPES, CMV의 세 그룹 모두에서 CAI의 평균값은 0.5 이상이었으며 이는 HCMV의 해당 유전자들이 숙주인 사람세포에 잘 적응해 있다는 것을 의미한다. CAI의 평균값은 그룹 간에 큰 차이를 보이지는 않으나, 대조적인 특성을 가지며 *t*-test 결과 차이가 유의한 것으로 나타난 두 그룹인 CORE와 CMV를 비교해 보았을 때, CORE에서 조금 더 높은 값을 보였다(Fig. 1A). 이는 CORE에 속하는 유전자들이 숙주의 시스템에 보다 더 잘 적응되어 있다는 것을 의미한다.

GC3s값은 코돈의 세 번째 자리인 wobble site에 guanine (G)이나 cytosine (C)이 있을 확률을 나타낸다. 이 값 또한 세 그룹 모두에서 0.5 이상의 높은 값을 가지며 adenine (A)이나 thymine (T)으로 끝나는 코돈보다는 G나 C로 끝나는 코돈의 빈도가 더 높다는 것을 의미한다. CAI에서와 마찬가지로 그룹간 비교를 통해 볼 때, CORE가 CMV에 비해 더 높은 GC3s값을 보였다(Fig. 1B). 즉 허피스바이러스에 공통으로 나타나는 유전자들이 여러 개의 동의코돈 중에서 GC-ending 코돈을 더욱 선호한다는 사실을 알

수 있다.

Nc값은 아미노산을 암호화하는 61개의 센스 코돈 중에서 몇 개의 코돈을 해당 유전자가 사용했는지를 나타내주는 값이다. 앞의 두 항목과는 달리 Nc값에서는 CMV의 평균값이 CORE보다 높게 나타남을 알 수 있다(Fig. 1C). 이는 허피스바이러스 공통의 유전자들이 코돈을 보다 편향적으로 사용한다는 사실을 말해준다. 이러한 경향은 Nc/GC3s값을 통해서도 볼 수 있으며 그 차이는 보다 확연하다. Nc값에서의 경향과 마찬가지로 Nc/GC3s값은 CMV에서 가장 높은데(Fig. 1D), 이는 GC3s값이 같을 때, CMV 유전자들이 더 많은 종류의 코돈을 사용한다는 것을 의미한다.

이러한 결과들을 종합해 볼 때, 유전자의 진화적인 보존성에 따라 코돈을 사용하는 경향이 다르게 나타난다는 것을 알 수 있으며, CORE와 CMV를 비교했을 때 CORE 유전자가 높은 CAI와 GC3s값을 가지며 반대로 낮은 Nc와 Nc/GC3s값을 가지는 경향을 관찰할 수 있었다.

유전자의 진화적인 보존성과 Nc plot

GC3s값의 변화에 따른 Nc값의 변화를 분산형 그래프로 나타낸 것을 Nc plot이라고 한다. 이 Nc plot은 게놈 내에서 각 유전자들의 코돈 사용 변화를 조사하는 효과적인 방법이라고 알려져 있다. Nc plot은 주로 'Wright's formula'라고 불리는 식인 $Nc = 2+GC3s+\{29/[(GC3s)^2+(1-GC3s)^2]\}$에 의해 그려지는 곡선 그래프와 함께 표시되며 이 곡선 위의 Nc값을 예상 Nc (expected Nc)라고 한다. 만약 유전자의 코돈 사용의 다양성을 나타내는 유일한 결정요인이 GC3s로서 나타나는 유전자 염기서열의 조성적인 제약이라면 GC3s에 대응하는 Nc값은 'Wright's formula'에 의해 그려진 곡선 위에 위치하게 된다 (13).

HCMV 유전자들의 코돈 사용 편향성을 결정하는 데에 조성적인 제약이 어느 정도의 영향을 미치고 있으며 그러한 영향의 정도가 유전자의 그룹에 따라 다르게 나타나는지 알아보기 위하여, 전체 ORF와 그룹별로 나눈 ORF들의 Nc plot을 Fig. 2에 나타내었다. ORF 전체에 대

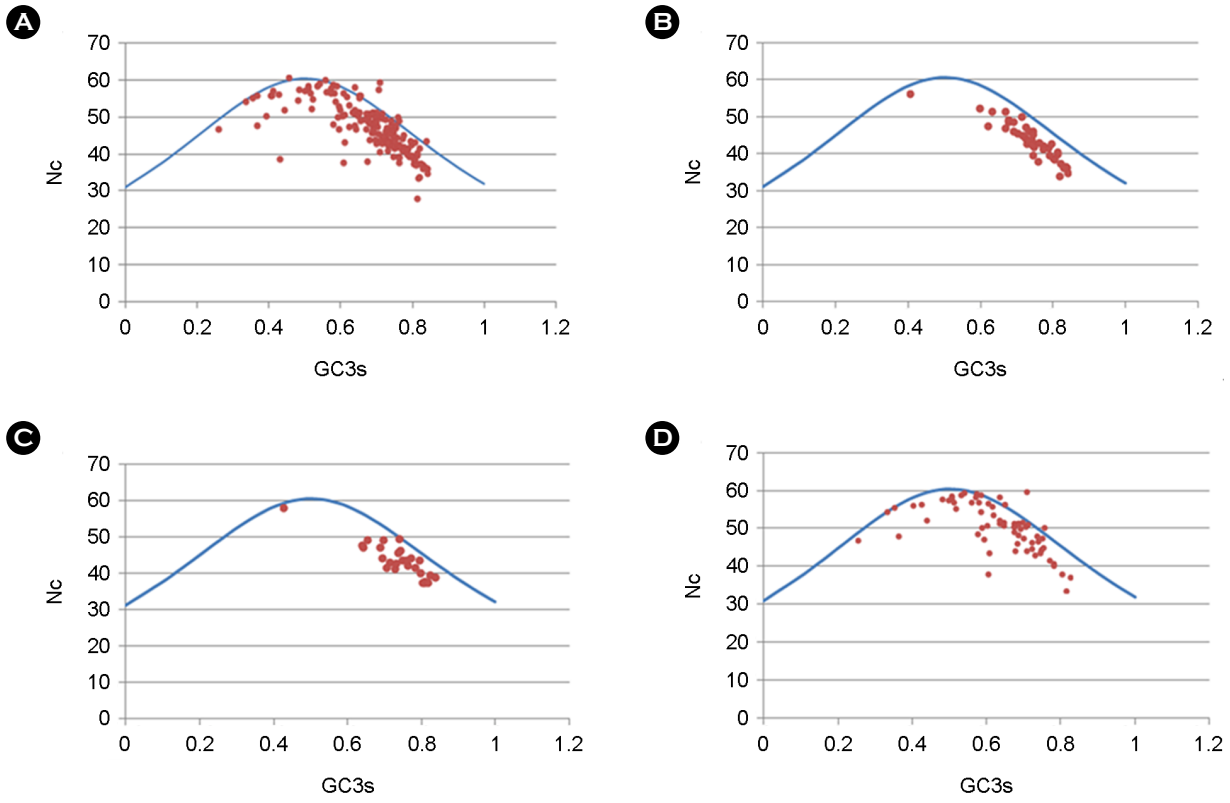

Figure 2. Nc versus GC3s plots for HCMV gene groups. (A) Scatter plot for all HCMV ORFs. (B) plot for CORE genes. (C) plot for β-HERPES genes. (D) plot for CMV genes. The continuous curve represents the expected Nc values by Wright's formula.

해 그린 Nc plot에서 점들은 대부분 예상 Nc 곡선 아래에 분포해 있으며, 곡선 위에는 소수의 점들만이 위치하고 있다(Fig. 2A). 이를 통해 HCMV의 ORF 대부분은 코돈 사용의 경향을 결정하는 데에 있어서 조성적인 제약을 주로 받고 있으며, 번역과정에서의 선택은 약한 영향을 준다는 것을 알 수 있다. 전체적으로 GC3s값은 0.2~0.8 사이의 넓은 범위를 가지나, 0.5 이상의 GC3s값을 갖는 ORF가 상대적으로 많으며, GC3s가 높은 ORF들에서 Nc값의 범위도 30에서 60에 이르기까지 넓게 나타났다. 이는 GC3s가 상대적으로 높은 유전자들이 다양한 형태로 코돈을 사용한다는 것을 알려준다. 또한, 대부분의 점들이 동일한 GC3s값을 가질 때, 예상된 Nc보다 낮은 값을 가지는 것을 통해 코돈 사용이 상대적으로 편향적이며 높은 GC3s값을 가진 ORF에서 이러한 경향이 더 심하다는 것도 알 수 있다 (14).

유전자의 진화적인 보존성에 따라 나눈 ORF 그룹인 CORE와 CMV의 Nc plot에서 대부분의 점들은 예상 Nc 곡선 아래에 분포해 있다. CORE 그룹에 속하는 ORF들의 Nc plot에서 점들은 대부분 GC3s값이 0.6~0.8 사이인 지점에 분포하며, 예상 Nc보다 낮은 Nc값을 보여 코돈 편향성이 상대적으로 높음을 알 수 있다(Fig. 2B). 반면 CMV에 속하는 ORF들은 GC3s값이 0.2~0.8인 부분에 넓게 분포해 있으며, CORE에 비해 다양한 GC3s값을 가지고 있다. 그러나 전체적으로 예상 Nc보다는 낮은 값을 가지면서 이 ORF들 또한 코돈을 편향적으로 사용한다는 사실을 알려주고 있다(Fig. 2D). β-HERPES에 속하는

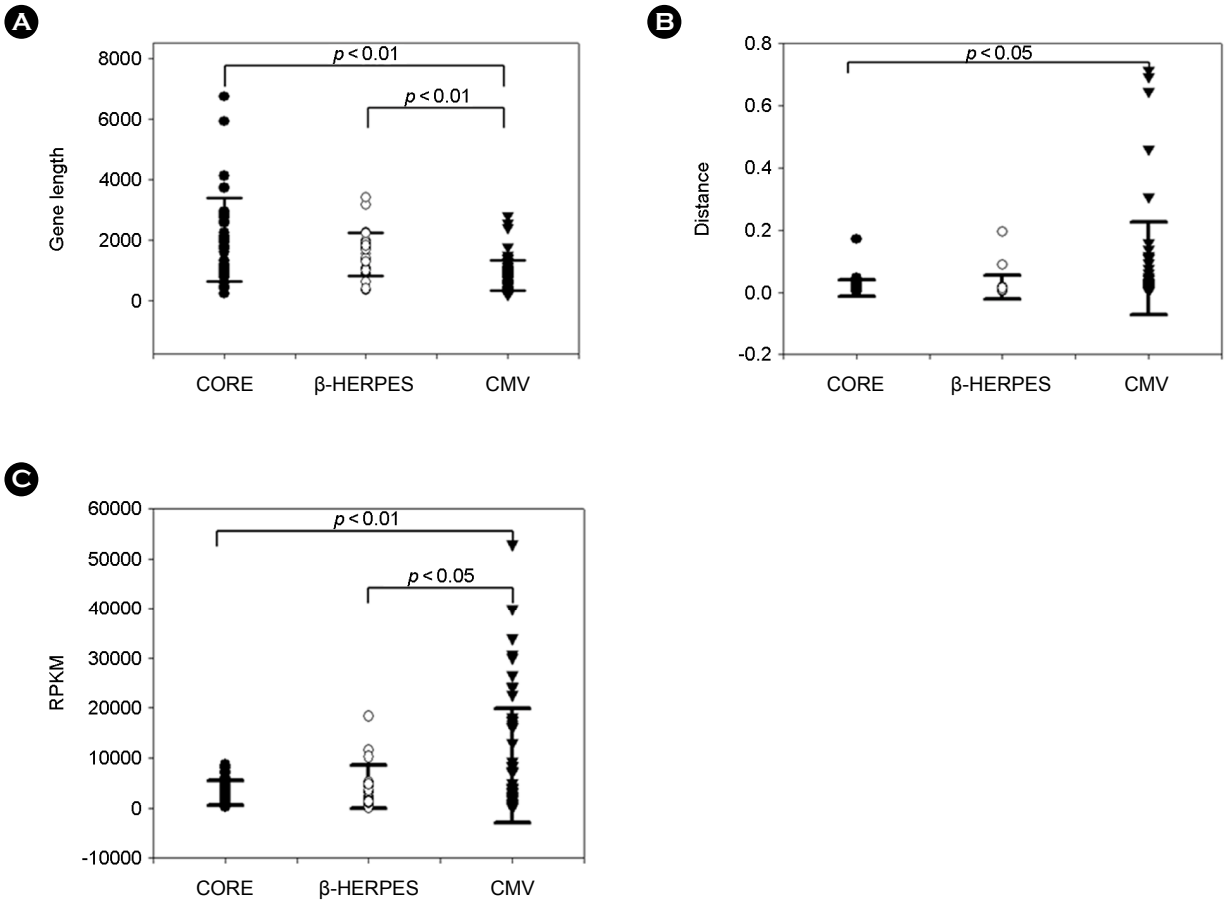

Figure 3. Genetic characteristics of HCMV genes. (A) Distribution of length of HCMV genes. (B) Genetic distance values of HCMV genes. (C) Relative amount of mRNA of HCMV genes. CORE genes are indicated by black circles, β-HERPES genes by white circles, and CMV genes by black triangles. Black bars indicate the standard error of the mean. Statistical significance of the difference in mean values was examined by student's *t*-test.

ORF들의 분포는 CORE 그룹과 매우 비슷한 경향을 보이며 대부분 GC3s값이 0.6~0.8인 지점에 분포하고 있다. 또한 Nc값도 예상 Nc보다 낮은 값을 보이고 있다(Fig. 2C).

유전자의 진화적인 보존성과 관련된 다양한 특징들

코돈 사용 편향성을 제외한 유전자의 길이와 진화적인 변이 정도, 상대적인 유전자의 발현량이 유전자의 진화적인 보존성과 어떠한 관련이 있는지 알아보기 위하여 유전자 그룹별로 해당 항목의 수치들을 구하여 비교하여 보았다. CUI 수치의 비교에서와 마찬가지로 각 항목에 대하여 ORF의 그룹별로 오차막대가 있는 산점도를 그리고 두 그룹씩 *t*-test를 수행하였고, 그 결과를 Fig. 3에 나타내었다. 코돈 사용 경향을 조사했을 때와 마찬가지로, 모든 항목에서 CORE와 β-HERPES간의 차이는 없었다($p > 0.05$). 반면 모든 항목에서 CORE와 CMV간 *p*-value는 0.05 이하로, 그룹간에 유의한 차이가 있었다. β-HERPES와 CMV는 유전자의 길이와 RPKM에 대해서는 의미 있는 차이를 보이고 있으나, 진화적인 변이거리에 대해서는 유의한 차이가 없었다.

분석에 사용한 3가지 항목 모두에서 ORF 그룹간 평균값의 차이를 확인할 수 있었다. 유전자의 길이의 평균값은 CORE에서 CMV로 갈수록 점점 작아지는 경향을 보이는데, 표준편차 또한 이와 같이 점차 감소하였다. 이는 CORE 유전자들의 길이가 CMV 유전자들에 비해 다양하게 나타나지만 평균적으로 더 길다는 것을 의미한다. 유전자의 평균적인 변이 정도를 나타내는 진화적인 거리(distance) 값에서는 앞선 결과와 대조적으로 CORE에서 CMV로 갈수록 평균값이 증가하는 양상을 보였다. 즉 CORE 유전자들의 변이가 가장 적고 CMV 유전자들의 변이 정도가 가장 크다는 것을 의미한다. 유전자의 상대적인 발현량을 나타내는 RPKM값은 distance와 마찬가지로 CORE에서 그 값이 가장 작고 CMV로 갈수록 점차 증가하는 양상을 보였다. 표준편차도 CMV에서 가장 크게 나타나며 CORE보다는 CMV 그룹으로 갈수록 유전자가 상대적으로 많이 발현되며, 유전자들 간에 발현량 차이가 다양하게 나타난다는 것을 보여주고 있다.

고 찰

본 연구에서는 HCMV ORF들의 진화적인 보존성에 따라 코돈을 사용하는 경향과 유전자의 발현량 등에 차이가 존재하는지 알아보았다. HCMV의 ORF들을 진화적인 보존성에 따라 CORE, β-HERPES, CMV의 세 그룹으로 나누어 비교한 결과, 유전자 그룹간에 수치상의 의미 있는 차이가 있음을 발견하였고 특히 CORE 그룹과 CMV 그룹이 모든 항목에 대하여 서로 반대의 경향을 나타낸다는 것을 알 수 있었다. CORE는 CMV에 비하여 높은 CAI와 GC3s값을 가지며 반대로 낮은 Nc와 Nc/GC3s값을 나타냈다. 또한 CORE에 속하는 유전자들의 길이가 더 길었으며 변이 정도는 낮았고 상대적인 발현량도 CMV에 비해 낮았다. β-HERPES는 모든 항목에 대하여 CORE와 CMV의 사이 값을 가지는데 CMV보다는 CORE와 비슷한 경향을 나타내었다.

HCMV의 ORF들을 분류하는 기준으로는 Dunn 등이 논문에서 제시한 분류기준을 사용하였다. Dunn 등은 논문에서 HCMV의 ORF들을 크게 진화적인 보존성과 단백질의 기능, 또는 성장에 대한 유전자의 요구성의 세 가지 기준에 따라 분류하였다 (5). 그 중 진화적인 보존성에 따른 분류에 의하면 ORF들 중 허피스바이러스 전체에 공통적으로 존재하며 알파, 베타, 감마 허피스바이러스의 공통 조상으로부터 유래한 것으로 추정되는 ORF들을 core gene (CORE) 그룹으로 분류하였다. 이 유전자들은 대체적으로 게놈의 중앙 지역에 위치해 있으며 이 중 대부분이 바이러스의 성장에 필수적이다. 반면 non-core 유전자들은 게놈의 끝부분에 위치해 있으며 일반적으로 높은 정도의 염기서열 다형성을 보인다. 이러한 경향은 HHV-6의 일부 아종에서도 관찰된 바 있다고 알려졌다 (15). 이 non-core 유전자들은 베타허피스 바이러스들에서만 공통적으로 발견되는 β-herpes 유전자 그룹(β-HERPES)과 사이토메갈로 바이러스들에서만 발견되는 cytomegalo 유전자 그룹(CMV)으로 나눌 수 있다 (5).

HCMV는 생존에 매우 중요한 역할을 하는 유전자들인 core 유전자만 가진 조상 게놈으로부터 진화해 오는 과정에서 감염이나 질병에 관련된 유전자들인 non-core 유전자들을 획득해 왔을 것으로 생각되고 있다 (5). 따라서 이는 core 유전자들은 모든 허피스바이러스에 존재하며 오랜 기간 유지되어 왔고 상대적으로 β-herpes 유전자들과 cytomegalo 유전자들은 존재해 온 기간이 짧을 것이라는 가설이다. 또한 이 중에서도 cytomegalo 유전자들의 역사가 가장 짧고 그만큼 다양성이 클 것이라 생각된다. Core 유전자들의 변이가 가장 적고 cytomegalo 유전자들의 변이 정도가 가장 크다는 우리의 실험결과는 이러한

가설을 잘 뒷받침해 준다.

CAI는 해당 유전자가 레퍼런스 유전자와 얼마나 비슷하게 코돈을 사용하고 있으며 잘 적응되었는지를 보여주는 지표이다. CAI는 0과 1사이의 값을 갖는데, 1의 값을 갖는 경우 숙주의 코돈 사용에 완전히 적응되어 강한 코돈 편향성을 나타내며 높은 발현수준을 보인다고 해석할 수 있다 (16). 이 CAI 수치는 CORE 그룹에서 더 높게 나오며 CORE 유전자들이 인간세포에 더 잘 적응되어 있음을 보여주고 있다. CORE 그룹의 유전자들은 상대적으로 긴 시간 동안 바이러스 내에 존재해 오면서 숙주세포인 사람의 번역체계에 더 잘 적응해 왔음을 나타내고 있다. 반면 CMV 유전자들은 보다 짧은 기간을 존재한 탓에 숙주세포에의 적응이 상대적으로 약하다는 것을 보여준다. 또한 유전자의 진화적인 변이 정도를 나타내는 유전적 거리 값에 있어서도 CORE의 값은 CMV의 값보다 낮았다. 이는 생존에 필수적인 CORE 유전자들이 오랜 시간이 지나도 거의 변이가 없이 유지되어 온 반면 진화과정에서 필요에 의해 획득한 CMV 유전자들은 보다 역동적인 변화의 과정을 거치며 개체의 기능과 생존력에 영향을 미치는 것으로 생각할 수 있다.

Nc는 각 ORF의 코돈 사용 편향성을 수치화하기 위해 사용되었다 (17). Nc는 20에서 61 사이의 값을 가지는데, 하나의 아미노산을 암호화하기 위해 오직 하나의 코돈만이 사용되는 경우 ORF가 가장 큰 코돈 평향성을 가지게 되고 이때의 Nc값은 20이 된다. 반면 아미노산을 암호화하기 위해 시작코돈과 종결코돈을 제외한 61가지의 모든 코돈이 동일하게 사용되는 경우 Nc값은 61이 되고 코돈의 편향성이 가장 적은 ORF의 상태를 나타낸다. Nc는 그 값이 35 이하인 경우 코돈 사용이 매우 편향적이라고 볼 수 있다 (13). HCMV의 유전자들은 대부분이 35 이상의 Nc값을 나타내며 코돈을 비교적 균일하게 사용하고 있음을 보여주었다. 그러나 그룹간의 비교를 통해 보았을 때 CMV 그룹에 비해 CORE 그룹의 Nc값이 더 낮아 비교적 편향적인 코돈 사용 경향을 볼 수 있었다. 즉 CORE 그룹의 유전자들이 특정한 코돈을 선호하는 경향이 더 두드러진다는 것을 의미한다. CORE 그룹의 유전자들은 바이러스의 생존과 성장에 필수적인 역할을 하는 유전자이므로 신속하고 정확하게 번역되어야 하므로 특정한 코돈을 주로 사용함으로써 번역과정에서의 오류를 줄이고 효율성을 높인 결과라고 보인다. 보통 높은 정도의 발현을 보이는 유전자들은 강한 코돈 편향성을 보이는데 이들은 세포 내에 풍부한 특정 tRNA를 사용함으로써 번역과정에서의 효율성을 높인다고 알려져 있다 (18). 이와 같은 이유로 CORE 그룹의 Nc값이 CMV 그룹보다 낮은 것을 설명할 수 있다.

그러나 상대적인 유전자의 발현량을 나타내는 RPKM 값에 대해서는 CORE 그룹의 상대적인 발현량이 CMV 그룹보다 낮음을 볼 수 있었다. 많은 경우 발현량이 많은 유전자에서 CAI값은 더 높고 Nc값은 더 낮다고 알려져 있으며, 특히 CAI값이 유전자의 발현량과 연관되어 있다고 보고된 바 있다 (19, 20). 이들은 일부 발현량이 많은 유전자의 예상에 사용되는 것이며 주로 박테리아의 게놈 연구로부터 얻은 결과이다. 그러나 일부 연구에서는 GC 함량이 높은 게놈에서 CAI 수치와 유전자 발현량 사이에 뚜렷한 상관관계가 없는 경우가 발견되었으며 (21, 22), 박테리아에서도 특정 범위의 CAI값을 가지는 유전자 그룹에서 CAI 수치와 유전자의 발현량 사이에 음의 상관관계가 존재함이 밝혀졌다 (23). 본 연구의 결과에서 CORE 그룹의 유전자들은 대부분 housekeeping 유전자로서 그 존재는 필수적이나 일정량만의 발현이 필요할 뿐 그 양은 크게 상관이 없는 기능들을 수행한다고 볼 수 있다. 반면 CMV 유전자들은 바이러스의 감염과 발병 등에 관련된 여러 기능적인 단백질을 암호화하고 있으므로 시간과 장소에 따라 그 발현량이 다양하게 나타날 수 있다.

앞선 많은 연구들에서 코돈 사용의 편향성을 결정하는 주된 두 가지 요소는 게놈의 염기서열 때문에 발생하는 돌연변이 압력(mutational pressure)과 세포 내에 존재하는 tRNA의 양 때문에 발생하는 전사선택 압력(translational selection pressure)이라고 알려져 왔다 (24, 25). 많은 세균 게놈들과 포유동물, 바이러스 게놈 등에 대한 연구들에서 코돈 사용 편향성을 일으키는 주된 요인은 번역과정에서의 압력보다는 변이에 대한 압력인 경우가 많다고 보고되었다 (9, 26~28). HCMV의 유전자들에 대해 그린 Nc-GC3s plot으로부터 우리는 HCMV의 유전자들 또한 염기조성에 의한 변이의 압력을 주로 받는다는 것을 알아낼 수 있었다. 즉 세포 내에 존재하는 특정 tRNA의 양에 따라 코돈을 결정하기 보다는 게놈의 염기서열 자체에 기인하여 코돈을 사용하는 경향이 결정된다는 것을 말한다. 이는 다양한 게놈들의 코돈 사용 경향 분석 결과에서 얻은 보편적인 결론들과 일치되는 내용을 보여주고 있다.

그룹간의 특징을 비교하기 위해 본 연구에서는 다양한

지표를 사용하였다. 본 연구에서 사용된 항목들간의 상관관계는 앞선 연구들에서도 찾아 볼 수가 있다. 보통 유전자의 길이는 발현량과 반비례하는데, 유전자의 길이가 증가할수록 유전자의 발현량은 감소하는 양상을 보이며 (18), Nc값과도 음의 상관관계를 보인다는 사실이 밝혀졌다 (29). 이러한 양상은 본 연구의 결과를 통해서도 동일하게 나타남을 관찰할 수 있었다. 또한, CAI 수치는 보통 Nc와 반대의 경향을 보이며 유전자의 편향성을 나타내며 GC3s와는 양의 상관관계를 보이는데 (30, 31), CORE 그룹의 유전자들에서도 높은 CAI값과 GC3s값, 그리고 낮은 Nc값을 가지는 것을 볼 수 있었다. 이처럼 HCMV 유전자들은 다른 게놈의 연구에서와 많은 부분에서 비슷한 양상을 나타내고 있다.

본 연구를 통해 우리는 HCMV의 유전자들을 그룹으로 나누어 코돈 사용의 편향성에 관련된 항목에 대한 수치들을 비교했고, 특히 CORE 그룹과 CMV 그룹간에 명확한 수치상의 차이가 존재함을 알 수 있었다. 이러한 수치상의 차이는 CORE 그룹과 CMV 그룹에 속하는 유전자들이 하나의 게놈 내에 존재하지만 여러 면에서 상이한 특징을 가진다는 것을 증명하는 결과이다. 이와 비슷하게 Karlin 등은 Epstein-Barr 바이러스에서 latent gene과 productive gene 그룹간의 코돈 사용의 차이를 보고한 바 있다 (11). 이러한 차이점에 기인하여 유전자들을 그룹으로 나누어 서로 비교하는 추가적인 연구를 수행한다면 HCMV의 어떤 유전자가 실제적으로 감염과 면역에 영향을 주는지에 대한 단서를 얻을 수 있을 것으로 생각한다.

REFERENCES

1) Landolfo S, Gariglio M, Gribaudo G, Lembo D. The human cytomegalovirus. Pharmacol Ther 2003;98:269-97.
2) Mocarski ES, Shenk T, Pass RF. Cytomegalovirus. Fields Virology; 2007. p.2702-72.
3) Khanna R, Diamond DJ. Human cytomegalovirus vaccine: time to look for alternative options. Trends Mol Med 2006; 12:26-33.
4) Dolan A, Cunningham C, Hector RD, Hassan-Walker AF, Lee L, Addison C, *et al*. Genetic content of wild-type human cytomegalovirus. J Gen Virol 2004;85:1301-12.
5) Dunn W, Chou C, Li H, Hai R, Patterson D, Stolc V, *et al*. Functional profiling of a human cytomegalovirus genome. Proc Natl Acad Sci U S A 2003;100:14223-8.
6) Rigoutsos I, Novotny J, Huynh T, Chin-Bow ST, Parida L, Platt D, *et al*. In silico pattern-based analysis of the human cytomegalovirus genome. J Virol 2003;77:4326-44.
7) Vetsigian K, Goldenfeld N. Genome rhetoric and the emergence of compositional bias. Proc Natl Acad Sci U S A 2009;106: 215-20.
8) Bouquet J, Cherel P, Pavio N. Genetic characterization and codon usage bias of full-length Hepatitis E virus sequences shed new lights on genotypic distribution, host restriction and genome evolution. Infect Genet Evol 2012;12:1842-53.
9) Shackelton LA, Parrish CR, Holmes EC. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J Mol Evol 2006;62:551-63.
10) Jenkins GM, Holmes EC. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 2003;92:1-7.
11) Karlin S, Blaisdell BE, Schachtel GA. Contrasts in codon usage of latent versus productive genes of Epstein-Barr virus:data and hypotheses. J Virol 1990;64:4264-73.
12) Fu M. Codon usage bias in herpesvirus. Arch Virol 2010;155: 391-6.
13) Wright F. The 'effective number of codons' used in a gene. Gene 1990;87:23-9.
14) Wu G, Nie L, Zhang W. Predicted highly expressed genes in *Nocardia farcinica* and the implication for its primary metabolism and nocardial virulence. Antonie Van Leeuwenhoek 2006;89:135-46.
15) Davison AJ, Dolan A, Akter P, Addison C, Dargan DJ, Alcendor DJ, *et al*. The human cytomegalovirus genome revisited: comparison with the chimpanzee cytomegalovirus genome. J Gen Virol 2003;84:17-28.
16) Sharp PM, Li WH. The codon adaptation index --a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 1987;15:1281-95.
17) Banerjee T, Gupta SK, Ghosh TC. Towards a resolution on the inherent methodological weakness of the "effective number of codons used by a gene". Biochem Biophys Res Commun 2005; 330:1015-8.
18) Roymondal U, Das S, Sahoo S. Predicting gene expression level from relative codon usage bias: An application to *Escherichia coli* genome. DNA Res 2009;16:13-30.
19) Waldman YY, Tuller T, Shlomi T, Sharan R, Ruppin E. Translation efficiency in human: tissue specificity, global

optimization and differences between developmental stages. Nucleic Acids Res 2010;38:2964-74.

20) Goetz RM, Fuglsang A. Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*. Biochem Biophys Res Commun 2005;327: 4-7.

21) Ohama T, Muto A, Osawa S. Role of GC-biased mutation pressure on synonymous codon choice *in Micrococcus luteus*, a bacterium with a high genomic GC-content. Nucleic Acids Res 1990;18:1565-9.

22) Lafay B, Atherton JC, Sharp PM. Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. Microbiology 2000;146:851-60.

23) dos Reis M, Wernisch L, Savva R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. Nucleic Acids Res 2003;31:6976-85.

24) Belalov IS, Lukashev AN. Causes and implications of codon usage bias in RNA viruses. PLoS ONE 2013;8:e56642.

25) Bulmer M. Coevolution of codon usage and transfer RNA abundance. Nature 1987;325:728-30.

26) Sharp PM, Stenico M, Peden JF, Lloyd AT. Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans 1993;21:835-41.

27) Jia R, Cheng A, Wang M, Xin H, Guo Y, Zhu D, *et al*. Analysis of synonymous codon usage in the UL24 gene of duck enteritis virus. Virus Genes 2009;38:96-103.

28) Jiang Y, Deng F, Wang H, Hu Z. An extensive analysis on the global codon usage pattern of baculoviruses. Arch Virol 2008; 153:2273-82.

29) Hassan S, Mahalingam V, Kumar V. Synonymous codon usage analysis of thirty two mycobacteriophage genomes. Adv Bioinformatics 2009;316936.

30) Prabha R, Singh DP, Gupta SK, Farooqi S, Rai A. Synonymous codon usage in *Thermosynechococcus elongatus* (cyanobacteria) identifies the factors shaping codon usage variation. Bioinformation 2012;8:622-8.

31) Zhao S, Zhang Q, Chen Z, Zhao Y, Zhong J. The factors shaping synonymous codon usage in the genome of *Burkholderia mallei*. J Genet Genomics 2007;34:362-72.