Virus Genes (2006) 33:15–26
DOI 10.1007/s11262-005-0035-7

ORIGINAL ARTICLE

Optimization of codon usage of poxvirus genes allows for improved transient expression in mammalian cells

John W. Barrett · Yunming Sun · Steven H. Nazarian · Tara A. Belsito · Craig R. Brunetti · Grant McFadden

Received: 5 August 2005 / Accepted: 3 October 2005

Abstract Transient expression of viral genes from certain poxviruses in uninfected mammalian cells can sometimes be unexpectedly inefficient. The reasons for poor expression levels can be due to a number of features of the gene cassette, such as cryptic splice sites, polymerase II termination sequences or motifs that lead to mRNA instability. Here we suggest that in some cases the problem of low protein expression in transfected mammalian cells may be due to inefficient codon usage. We have observed that for many poxvirus genes from the yatapoxvirus genus this deficiency can be overcome by synthesis of the gene with codon sequences optimized for expression in primate cells. This led us to examine colon usage across 2-dozen sequenced members of the *Poxviridae*. We conclude that codon usage is surprisingly divergent across the different *Poxviridae* genera but is much more conserved within a single genus. Thus, *Poxviridae* genera can be divided into distinct groups based on their observed codon bias. When viewed in this context, successful transient expression of transfected poxvirus genes in uninfected mammalian cells can be more accurately predicted based on codon bias. As a corollary, for specific poxvirus genes with less favorable codon usage, codon optimization can result in profoundly increased transient expression levels following transfection of uninfected mammalian cell lines.

Keywords Yatapoxvirus · Codon usage · Codon optimization · Effective codon number · Poxvirus expression

C. R. Brunetti
Department of Biology, Trent University, K9J 7B8 Peterborough, ON, Canada

J. W. Barrett · Y. Sun · S. H. Nazarian · T. A. Belsito · G. McFadden
Biotherapeutics Research Group, Robarts Research Institute, N6G 2V4 London, ON, Canada

J. W. Barrett · Y. Sun · S. H. Nazarian · T. A. Belsito · G. McFadden (✉)
Department of Microbiology and Immunology, University of Western Ontario, N6G 2V4 London, ON, Canada
e-mail: mcfadden@robarts.ca

Introduction

Our lab is interested in the dissection of poxvirus gene function, particularly those genes with a predicted immunomodulatory function [1]. Towards this goal we routinely attempt to express specific poxvirus open reading frames (ORFs) from uninfected mammalian expression vectors for further study in the absence of other competing viral proteins. As well, expression vectors often allow the fusion of the viral protein in-frame with epitope tags that permit detection of the fused, expressed protein. This strategy has generally been successful for the transient expression of leporipoxvirus genes, however we have consistently experienced difficulty expressing many yatapoxvirus genes from mammalian expression vectors. To date we have cloned several dozen viral genes from both tanapox virus (TPV) and yaba monkey tumor virus (YMTV) into the expression vector pcDNA3.1myc/his (Invitrogen), and have routinely observed little or no protein expression following transfection into human or primate cells. This poor transient expression could be due to the presence of cryptic splice sites, polymerase II termination sites or mRNA instability motifs within the

ORF resulting in truncated, incomplete or unstable transcripts. However another explanation is that inefficient colon usage could restrict the amount of translated product from mammalian cells [2–4]. To probe this issue, we have employed the baculovirus expression system (BES) to over-express yatapoxvirus genes of interest, usually with great success [5, 6]. To date, all of the yatapoxvirus genes we have cloned into AcNPV are expressed efficiently. Although the BES has numerous advantages and allows production of moderate quantities of poxvirus protein, there are still advantages to being able to transiently express a poxvirus gene in an uninfected mammalian cell. As well, we have frequently mutated any predicted cryptic splice sites without altering the encoded amino acid sequence. Although such predicted splice sequences could be altered by site directed mutagenesis, we were still not ever able to transiently express yatapoxvirus proteins with efficiencies comparable to genes derived from the lepori- or orthopoxviruses (unpublished). However, for several yatapoxvirus genes of interest we chemically synthesized versions with codon sequences optimized for the human translation machinery. These optimized viral gene sequences were then cloned into pcDNA3.1 myc/his and shown to now express at high efficiency in both human (HEK293) and non-human primate cells (Cos7). These results are consistent with codon optimization of genes from other viruses, including HIV and HPV [2–4]. This observation led us to examine codon usage bias in members of the *Poxviridae* family.

Poxvirus members belong to the family *Poxviridae* which is divided into two sub-families: the Entomopoxviriane, which are invertebrate poxviruses and can be further subdivided into three ''Types'' that are restricted to several insect families, and the Chordopoxvirinae, which is subdivided into eight genera that infect vertebrates. Complete genomic sequences are now available for representatives of all chordopox genera comprising over 2-dozen representative members (www.poxvirus.org). Here we examine the codon usage profiles of these selected poxviruses and try to derive some general principles regarding the ability to predict efficiencies of translation and transient expression of poxvirus genes in mammalian cells.

Materials and methods

Sequences

Poxvirus genomes were identified from NCBI and the open reading frames saved as fasta files using the ''viewing coding regions'' option of Entrez. Lists of the nucleotide coding sequences were loaded into the online version of CodonW [7]; http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html) and the effective codon number and percent GC at the third position was measured. All data was compiled into Excel:MAC v2004 (Microsoft), manipulations were performed and the numbers were plotted against each other. These plots indicate the codon bias on the *y*-axis so that the more biased (i.e. non-random) the codon usage is, the closer it will be to a value of 20. The more unbiased (i.e. random) the codon usage, the closer the plot shifts towards 61 (the maximum effective number) where each codon has an equal opportunity to encode an amino acid.

Transfections and immunoblotting

HEK293 and CoS7 cells were transfected using Lipofectamine 2000 (Invitrogen Inc.) according to manufacturer's specifications. Two micrograms of plasmid DNA was transfected into each well of a six-well dish. Expression was detected with anti-myc (Invitrogen) at 1:5,000 dilution, anti-His (Qiagen) at 1:10,000 or anti-gp38 [5] at 1:10,000.

RT-PCR

Total RNA was extracted from transfected cells at 48 h post transfection using a Qiagen RNeasy mini kit (Qiagen). First strand synthesis was achieved with Superscript II reverse transcriptase (Invitrogen) in a 20 μl reaction volume using oligo-(dT) as a primer. The cDNA was used as a template for PCR amplification. Primers used for PCR amplifying native and mutated 2L were 5′-ccc*aagctt*catggataagttactattatttagcac (forward primer, *Hind*III site italicized) and 5′-cc*gctcgag*ggtttccgtcttcttcatcctcttc (reverse primer, *Xho*I site italicized). Primers used for PCR amplifying optimized 2L were 5′-atg aac aaa ctg atc ctg ttc agc (forward primer) and 5′-gcc aag tct tcc tcg tcc tct tcg (reverse primer). The reaction mix was incubated for one cycle at 95°C for 3 min and then 30 cycles at 95°C for 30 s, 53 °C for 1 min and 72 °C for 2 min. Products were amplified with platinum TAQ (Invitrogen) and resolved on 1% agarose.

Results

Transient expression of individual viral proteins allows for the study of the specific gene products without the complications of the background contributions from

the other viral proteins. Towards this goal, we have cloned several dozen yatapoxvirus genes into mammalian expression vectors to analyze their function. Unfortunately, we have been unable to detect any expression of yatapoxvirus genes from mammalian expression vectors. For example, the 2L gene from Tanapox virus (T2L), an inhibitor of human tumor necrosis factor (huTNF, [5] can be readily detected by immunoblotting from TPV-infected primate cells however it is not detectable when transiently expressed in uninfected primate cells (Fig. 1a, compare lanes 1 and 4). In contrast, the same T2L open reading frame is well expressed in the baculovirus expression system (Fig. 1b, lane 2). The scenario where certain viral genes, cloned from mammalian viruses, are not expressed following transfection in mammalian cells but are well expressed from baculovirus promoters in insect cells led us to examine further the problem. Processing the sequences through software (http://www.friutfly.org/cgi-bin/seq_tools/splice.pl) that searches for cryptic splice sites predicted several potential splice sites (Table 1) When we examined native T2L transcript levels we found that the T2L transcript was indeed truncated in transfected Cos7 cells (Fig. 1c, lane 3). We attempted to first correct for the cryptic splice sites by site directed mutagenesis which solved the issue of truncated transcripts (Fig. 1c, lane 4), however that did not solve the lack of protein expression (Fig. 1a, lane 5). To overcome this problem we have synthesized several yatapoxvirus genes with codon optimized sequences that favour translation in human cells (TopGene, Montreal, QC). Because poxvirus genes are normally transcribed in the cytoplasm and never encounter the nucleus there has not been any selection pressure exerted by host cell nucleus-resident pathways. Codon optimization of T2L led to detection of transcripts of the correct size (Fig. 1c, lane 5), and T2L protein from transient expression was now readily detectable with our antibodies (Fig. 1a, lane 6). In the case of T2L, codon optimization, by adjusting the proportion of AT in the third codon position to a higher proportion of GC (Fig. 2) resulted in the switch from undetectable to significant protein expression. Such a dramatic change through codon optimization led us to examine the codon usage patterns in poxvirus family members.

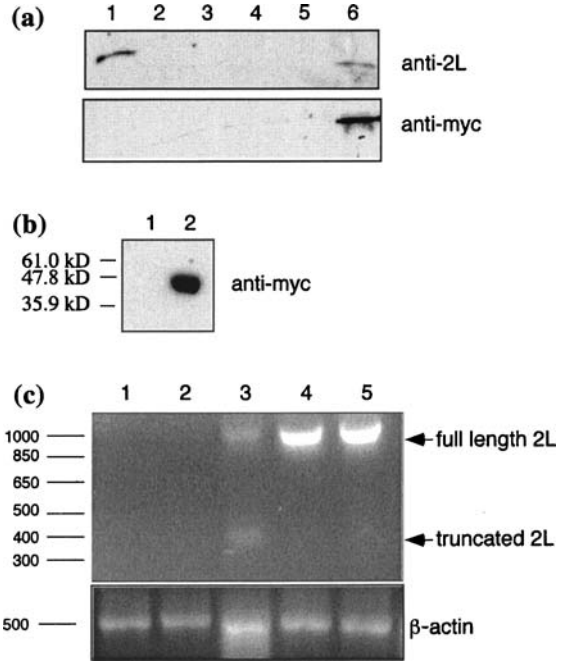

Fig. 1 Expression of native and codon optimized T2L. (**a**) top panel. Comparison of TPV2L expression by immunoblotting with anti-2L [5]. Lane 1 represents supernatant from TPV infected OMK cells [8] collected 48 hpi. Lane 2 is uninfected control supernatant. Lanes 3–6 are supernatants collected from transfected COS7 cells. Lane 3 is transfection control with pcDNA3.1 M/H vector alone. Lane 4 is pcDNA3.1 M/H + native T2L. Lane 5 is pcDNA3.1 M/H + T2L following correction of the predicted splice sites and Lane 6 is pcDNA3.1 M/H + optimized T2L. Bottom panel are the same samples detected following immunoblotting with anti-myc (1:5,000). All supernatant were collected 48 h post transfection. (**b**) Detection of T2L from Sf-21 cell supernatants following infection by AcNPV control (lane 1) or AcT2L (lane 2) and detected with anti-myc (1:5,000). Markers indicate protein mass in kilodaltons (kD). (**c**) top panel. RT-PCR of total RNA from untransfected COS7 cells (lane 1), or COS7 cells transfected with the empty vector (lane 2), pcDNA3.1 M/H + native T2L (lane 3), pcDNA3.1 M/H + corrected T2L (lane 4) or pcDNA3.1 M/H + optimized T2L (lane 5). Bottom panel. β-Actin control samples. The ladder indicates DNA sizes in base pairs

Effective codon number in the poxvirus genome

The twenty amino acids utilized by the universal translational machinery are encoded by 61 codons. The redundancy of codon specificity, and the particular preference of codon selection within a given species, can be informative about its genetic structure and organization. The range of codon usage bias was therefore examined for the *Poxviridae*. Complete genomic sequences for two entomopoxvirus species and 19 representative chordopoxvirus genomes are available in Genbank (Table 2).

To measure the codon bias within a gene, it is first necessary to determine the actual codon usage and compare it to the possible codon options available for each amino acid. This calculation is considered the effective codon number (Nc) and this statistic has been developed for comparative studies and evolutionary divergence analyses [9]. The effective codon number estimates the average number of codons actually used

Table 1 Predicted TPV-2L cryptic splice sites* and corrected sequences

Start	End	Score	Original sequence	Mutated sequence
365	379	0.99	gttatggGtatgtag	gttatggCtatgtag
223	237	0.97	acgccAggTaacgat	acgccGggAaacgat
192	232	0.90	tgattggtttaatatttctAggagtccccacacgccaggta	tgattggtttaatatttctCggagtcccc acacgccaggta
675	689	0.86	tcctgaCgtAattac	tcctgaTgtTattac
610	624	0.81	gtaacgggTaatgag	gtaacgggCaatgag
712	726	0.74	tttaaaggTgaatat	tttaaaggCgaatat
171	185	0.62	CggaaaCgtAaattt	cggaaaTgtGaattt
628	642	0.58	gaagatggTaacatg	gaagatgGCaacatg
846	860	0.53	AgatttaGtacgtac	agatttaAtacgtac
453	467	0.50	TcccaagGttggaat	tcccaagAttggaat
568	582	0.44	gctaaaggTgaaata	gctaaaggCgaaata
879	893	0.36	TcaactaGtatgtgt	tcaactaCtatgtgt
533	573	0.31	gatgttcattagctattttAgattaccaaa aaatggctaaa	gatgttcattagctattttGgattaccaaa aaatggctaaa
920	960	0.20	aagtttatacttgttctgaAggatgcaat ggagagctatac	aagtttatacttgttctgaGggatgcaat ggagagctatac

*Predicted splice sites were determined using the following website: (http://www.friutfly.org/cgi-bin/seq_tools/splice.pl)

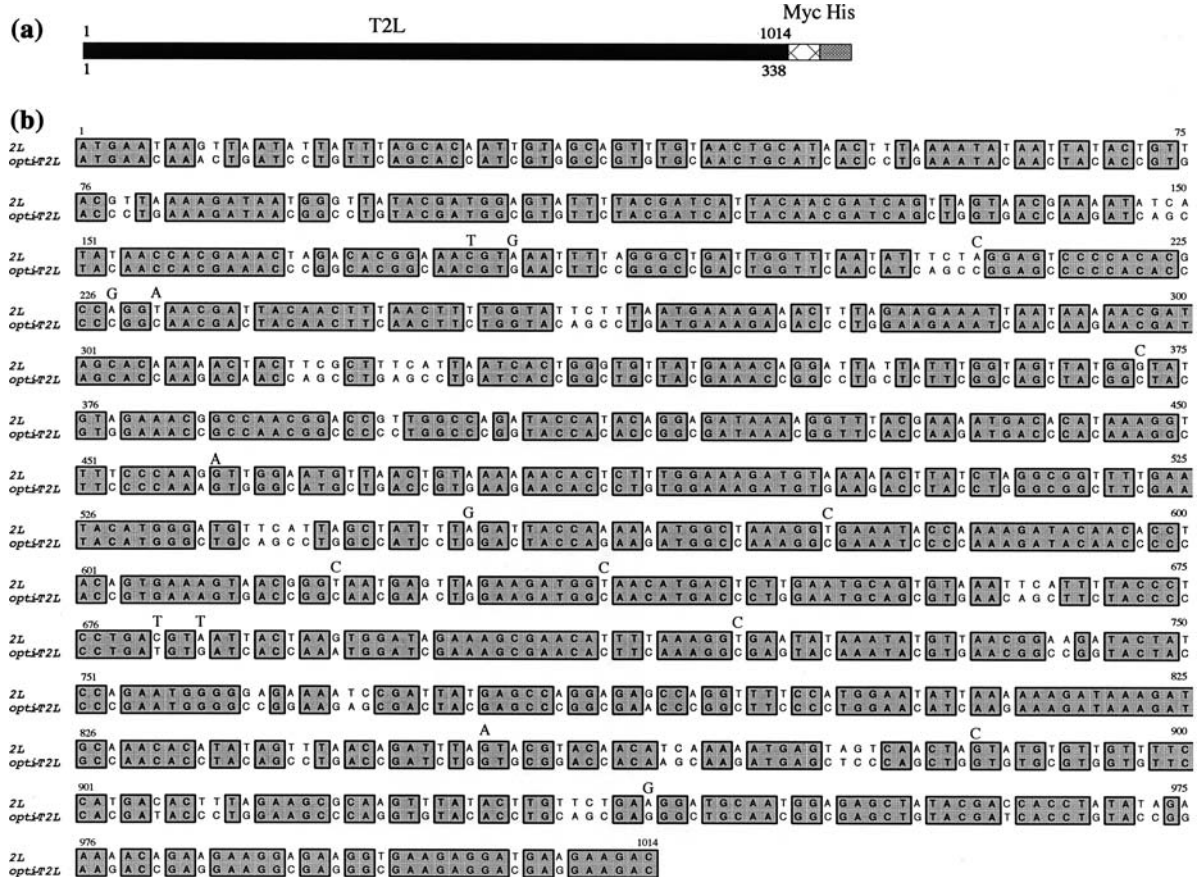

Fig. 2 Codon optimization of T2L. (**a**) Cartoon of the fusion of T2L with the myc/his epitopes. The numbers above the cartoon indicate the nucleotide position. The numbers below the cartoon indicate the amino acid number. (**b**) The nucleotide sequence of the native T2L gene (2L) is aligned to the optimized version of the gene (opti-T2L). Identical nucleotides are shaded and boxed. Above the native T2L sequence, in bold text, are the 17 nucleotides that were altered by site directed mutagenesis to alter the cryptic splice sites and correspond to the sites described in Table 1

in a given gene. This statistic can take a value ranging from 20, indicating extreme bias where every amino acid is only encoded by a single codon up to a maximum value of 61 indicating that all codons are exploited equally without bias to encode the amino acids. We have 21 poxvirus genomes in Genbank and have

Table 2 *Poxviridae* members examined in this study and their effective codon number (Nc)

Subfamily	Genus	Species	Accession #	Total GC (%)	GC1 (%)	GC2 (%)	GC3 (%)	Nc
Entompoxvirinae								
	Type B	Amsacta moorei AmEPV	NC_002520	18.4	24.524	20.93	9.73	26.99
		Melanoplus sanguinipes MsEPV	NC_001993	18.8	26.35	21.95	8.09	31.45
Chordopoxvirinae								
	Leporipoxvirus	Myxoma MV	NC_001132	44.42	44.79	34.62	53.84	52.88
		Rabbit fibroma SFV	NC_001266	39.96	41.84	32.71	45.34	52.9
	Parapoxvirus	Orf virus orf	NC_005336	65.14	62.29	44.82	88.3	35.41
		Bovine papular stomatitis BSPV	NC_005337	65.07	61.42	43.77	90.01	35.86
	Molluscipoxvirus	Molluscum Contagiosum MCV	NC_001731	63.91	63.71	46.07	81.94	38.99
	Capripoxvirus	Lumpy skin disease LSDV	NC_003027	26.29	32.09	27.42	19.35	39.58
	Suipoxvirus	Swinepox SPV	NC_003389	27.87	33.72	29.55	20.34	40.43
	Orthopoxvirus	Ectromelia ECTV	NC_004105	34.79	40.06	34.06	30.23	46.79
		Variola Major VARV	NC_001611	33.41	39.16	31.99	29.08	46.85
		Camelpox CMLV	NC_003391	33.82	39.25	32.36	29.84	46.92
		Monkeypox MPV	NC_003310	34.98	40.05	34.63	30.24	46.99
		Vaccinia VACV	NC_001559	34.23	39.59	32.72	30.39	47.1
		Cowpox CPV	NC_003663	34.85	40.17	33.68	30.71	47.35
		Rabbitpox RPV	NC_005858	33.96	39.49	32.32	30.06	47.5
	Avipoxvirus	Canarypox CNYV	NC_005309	30.96	37.45	30.71	24.72	41.87
		Fowlpox FPV	NC_002188	31.81	37.38	31.33	26.72	44.08
	Yatapoxvirus	Yaba monkey tumor YMTV	NC_005179	30.22	34.19	28.7	27.77	42.99
		Yaba like disease YLDV	NC_002642	27.29	32.91	27.83	21.11	39.37
	Unclassifed	Muledeer pox	NC_006966	26.8	N/A	N/A	16.4	38.6
	Average							42.42

estimated the Nc for all ORFs using CodonW [7]. The effective number of codons (Nc) used by the *Poxviridae* was on average 42.4 and ranged from a very biased Nc of 26.99 (Amsacta moorei entomopoxvirus) to a more random Nc of 52.9 (Shope fibroma virus) (Table 2). Two viruses had Nc values in the range of 50–61, 11 in the range 40–50, 7 between 30 and 40 and a single species between 20 and 30. Although the *Poxviridae*, as a whole, exhibited a range of codon bias, approximately five species displayed extensive bias while the rest exhibited only minor codon usage bias.

Poxvirus genera can be separated into distinct classes based on codon bias and GC content

Twenty-one poxvirus genomes were compared by plotting the effective codon number (Nc) against the proportion GC in the third position (GC_3) (Fig. 3). Each plot presents the complete complement of ORFs from each genome. There is wide variation in effective codon number (Nc) and GC_3% among the species, however several trends are apparent. Generally, all the ORFs within a species exhibit a similar GC_3% and a codon bias that results in a clustering of the ORFs. The exception is that the entomopoxviruses, which encode a subset of 6–10 genes which appear to deviate from the majority. While the majority of the entomopoxvirus ORFs appear to be extremely AT rich in the third position, these ''outliers'' have a higher GC_3 content. The parapoxvirus, and to a lesser extent the molluscipoxvirus genomes, also exhibit a subgroup of outlier genes that deviate from the main group (Fig. 3). In these genera, there are 16 genes, which have a lower percent of GC (less than 70%) in the third position and these ORFs exhibit much less codon bias. Where available it also appears that members within a specific genus maintain a conserved codon bias reflected in the effective codon number. We plotted the theoretical effective codon number (line) estimated solely on GC concentration. This suggested that for most poxvirus members the actual codon bias was close to the predicted value based on GC content.

Based on the plots of the 21 genomes we can group all poxviruses into one of four classes (Fig. 3). Class one represents genomes with a highly biased codon usage and with a very low GC percentage in the 3rd position. This class includes the two entomopoxviruses only. We would predict that future EPV sequences would also reflect this trend. Class two has a more random codon usage with a 50% GC in 3rd position and is exclusive to the leporipoxviruses, represented here by myxoma virus and rabbit fibroma virus and appears to encode ORFs that exhibit almost random codon usage. Class three includes those genomes, which are highly biased in their codon usage but, in contrast to the entomopoxviruses, these species are

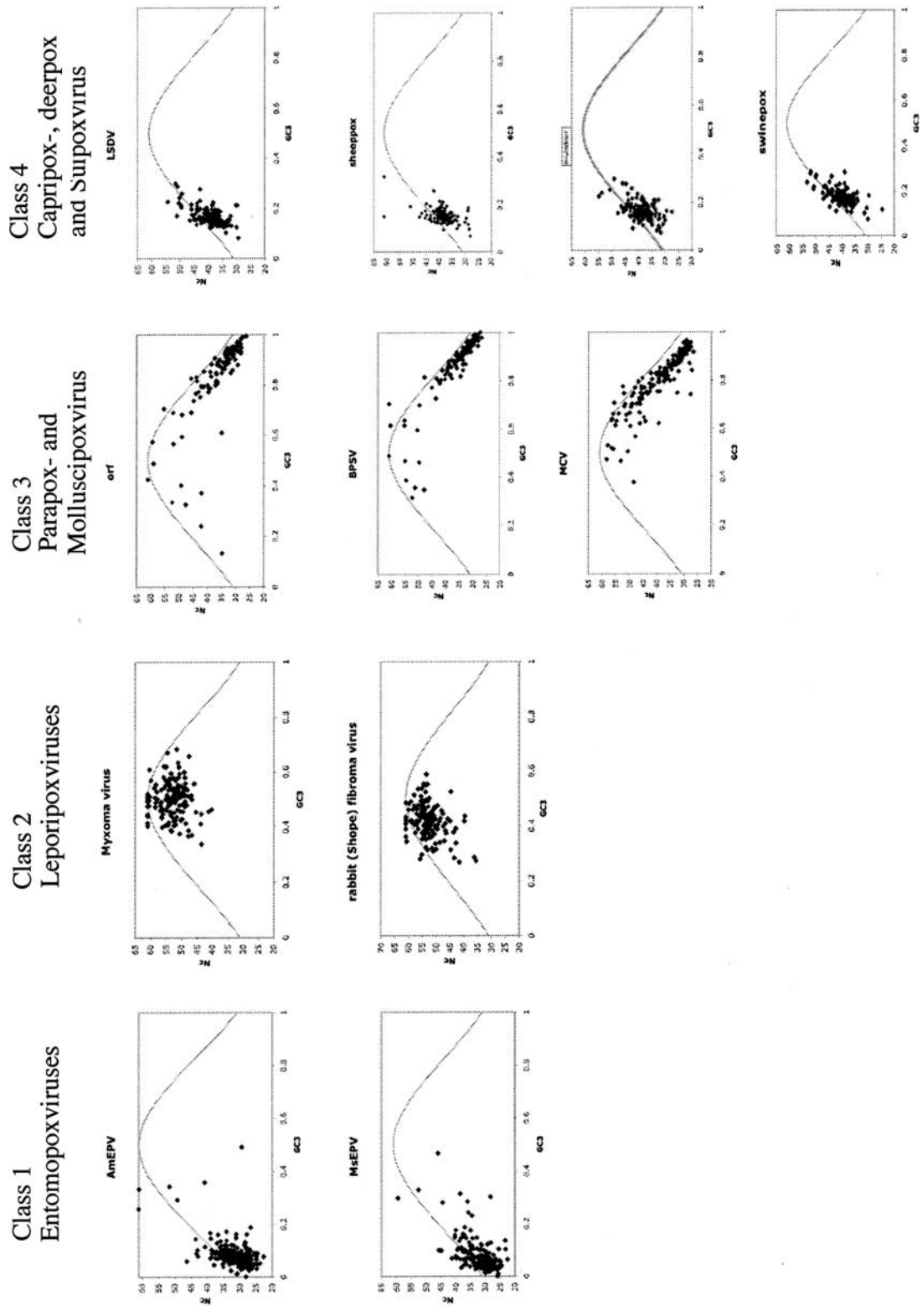

Fig. 3 Graphic presentation of the effective codon number (Nc) and proportion GC in the third codon position (GC3) for each poxviral species. The 21 members have been clustered into groups based on their effective codon numbers. Each open reading frame is represented by a dot and the theoretical Nc at any given proportion of GC_3 is drawn as a line

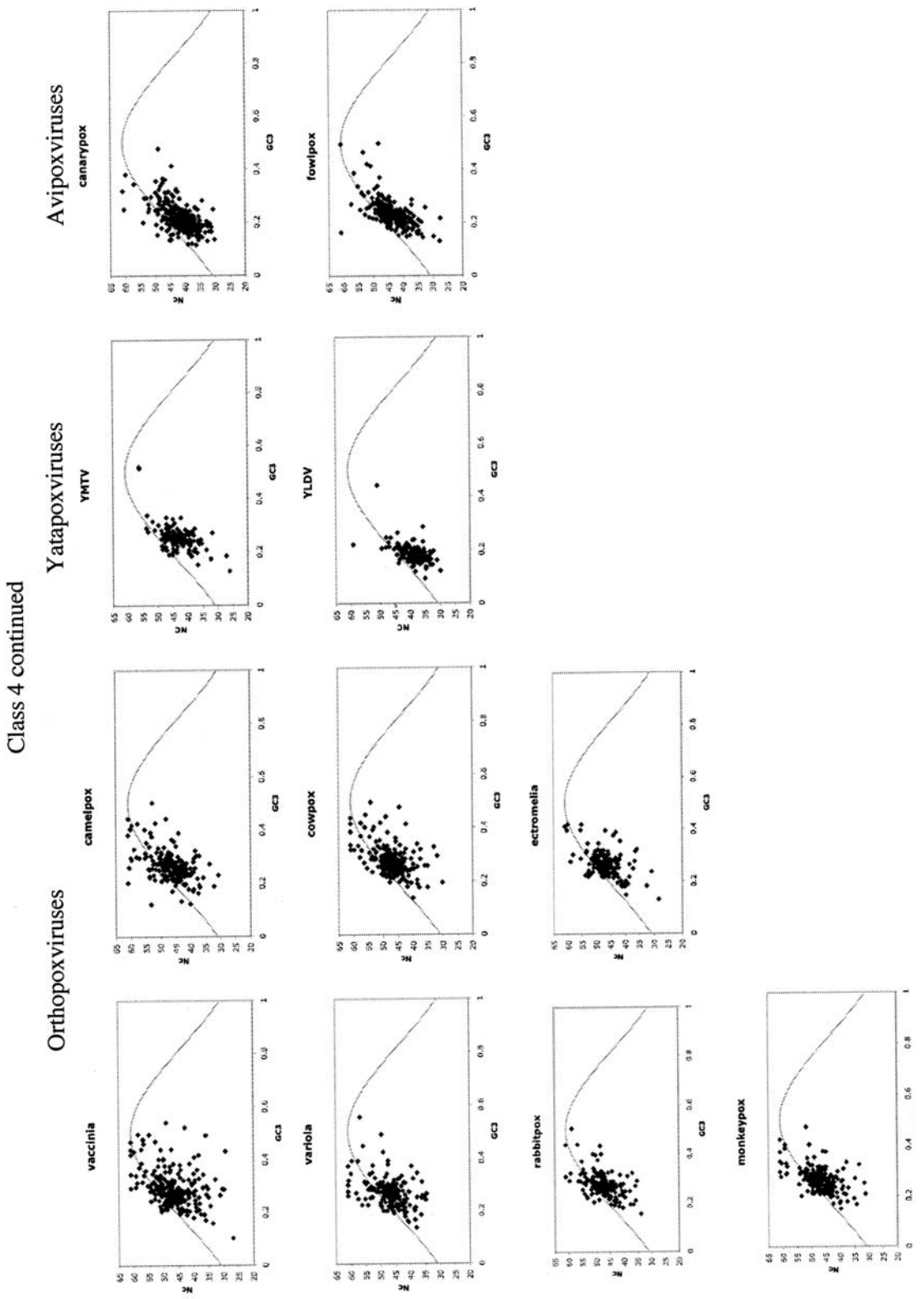

Fig. 3 continued

highly GC rich in the 3rd position. Members of the parapoxvirus (orf and BPVS) and the single molluscipox are members of this class.

The final class is the largest and contains the majority of poxvirus genera. Once more genomes are sequenced and analyzed this final group may break down into two distinct classes, however for now this final class includes the capripoxviruses, the single member of the suipoxvirus and deerpox, which is unclassified, and these genes are characterized by mild codon bias (Nc_{avg} = 39.42) and between 10% and 20% GC_3 (Fig. 3). The remaining members of this class exhibit a more random codon usage pattern (Nc_{avg} = 45.2), similar to class two, however, in contrast, the 3rd position GC% is much lower, on average between 25% and 40% and include all published genomes of the orthopoxviruses, avipoxviruses, and yatapoxviruses. Overall, poxvirus species exhibit a range of codon bias usage, however members within a genus have evolved a codon usage bias consistent with other members of their genus. This conservation of the codon usage appears to be GC concentration specific, rather than dependent on host requirements. For example, when we plot the percent GC_3 for all coding regions against the GC content of the first two codon positions (GC_{1+2}), for each genome we find a high correlation between GC in position 1 and 2 and maintenance of GC in the 3rd position. The grouping of the members into the four defined groups is easily visualized (Fig. 4).

Highly conserved genes do not share codon bias across species

Unexpectedly, conservation of codon bias for orthologous genes across the multiple poxvirus genera was not observed. For example, when we examine three highly conserved genes found in all published poxviruses, including DNA polymerase, P4a (the major core protein) and uracil DNA glycosidase, we find that the codon bias is conserved only within the particular genus (Fig. 5). The entomopoxviruses, parapoxviruses and molluscipoxvirus are all highly biased in the codon

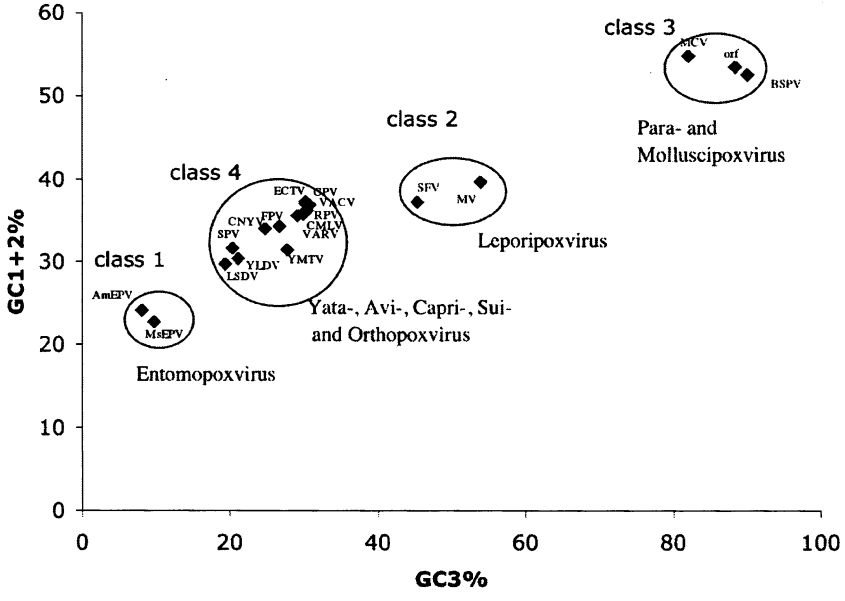

Fig. 4 There is a strong correlation between the GC content of the third synonymous codon position (GC3) and the GC content of the first and second codon positions (GC_{1+2}). Each data point represents the average values calculated for each poxvirus species. The groupings described for Fig. 2 are circled and each species is identified by an abbreviated name. The abbreviations are taken from Table 1

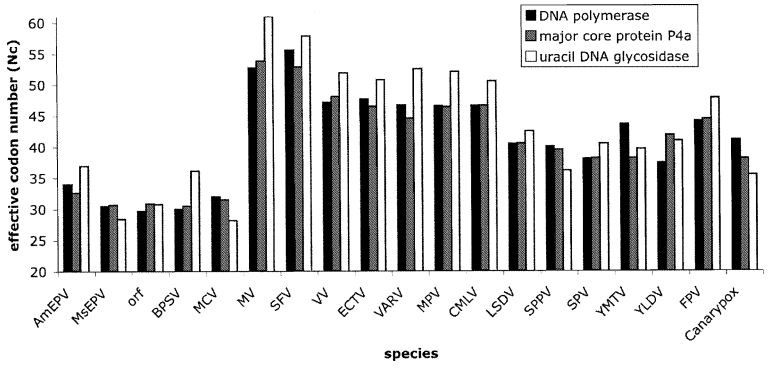

Fig. 5 Codon bias of conserved genes is not maintained amongst the poxvirus species. The effective codon number as an estimate of codon usage for DNA polymerase, P4a and uracil DNA glycosidase was calculated for each species and compared within and between species

usage of these three genes and this is reflected in the low effective codon number for these three groups. In contrast, the leporipoxviruses are essentially random in the codon selection and the rest of the species fall somewhere in between. Therefore it appears that viral genes that are thought to have evolved from a common ancestor have further adapted to the host genetic environment in which the individual poxviruses have invaded. This is true of genes that possess a cellular homolog (DNA polymerase) and those that are of strictly viral origin (major core protein).

Codon bias does not reflect an ability to infect humans

Eight poxvirus members from four genera have the ability to infect and produce productive infections in humans, including members of the orthopoxviruses (vaccinia, variola, cowpox, monkeypox), the yatapoxviruses (tanapox, yaba monkey tumor virus), the parapoxviruses (orf, pseudocowpox) and molluscipoxviruses (molluscum contagiosum) [10, 11]. It might be predicted that the ability to infect humans would require a codon usage profile that matches codon usage in humans, or possibly a conserved codon bias shared amongst species able to infect humans. However this is not borne out by analyses of the actual codon bias of these members. The genomic sequences of seven of the eight members with the ability to replicate in humans are available and there does not seem to be any relationship between the codon usage and ability to infect humans. In fact, variola and MCV infections are restricted to human hosts but variola exhibits less codon usage bias (Nc = 46.8) than does molluscum contagiosum (Nc = 39) despite a dramatic difference in their GC_3 content. The ORFs of variola are generally AT rich in the GC_3 position (GC_3 = 29%) versus molluscum contagiosum ORFs which are very GC rich (GC_3 = 82%; Table 2). A plot of the effective codon number against %GC_3 for 135 human cellular genes [9] looks more similar to the profiles for molluscum contagiosum and orf virus (Fig. 3) than for variola virus. The profiles for the orthopoxviruses and other members of class 4 appear most similar to effective codon plot profiles for the amoeba, *Dictyostelium discoideum* [9].

Codon optimization dramatically improves transient expression of yatapoxvirus genes

We have assumed that low expression levels following transfection of certain yatapoxvirus genes were the result of cryptic splice sites that were being processed in the nucleus leading to truncated transcripts. Of the yatapox ORFs we have tested none has been adequately expressed transiently from pcDNA3. 1myc/his in mammalian cells. Recently we have had three ORFs synthesized to optimize codon usage for human cells. The expression of the modified yatapox genes was dramatic. The codon optimization resulted in excellent expression levels from both transfected human (HEK293) and non-human primate (Cos7) cells (Fig. 6). Comparison between the natural codon usage and third position GC levels with the optimized ORFs indicate that there are some striking differences (Fig. 7). The three native ORFs have mild codon bias however they are strikingly AT rich in the 3rd position of the codon. In contrast the optimized versions of the same genes are now strongly biased and are extremely GC rich in the 3rd position of the codon. Based on these results and our earlier work it may be possible to predict which pox genomes encode genes which would be resistant to transient expression, using common mammalian expression vectors in mammalian cells (Table 3). Basically those genomes that are AT rich, including the entomopox-, the yatapox-, the orthopox-, capripox-, and suipoxviruses would be predicted to be resistant to transient expression in human and non-human primate cells. In contrast we would predict that genes from parapox- and molluscipoxviruses should be well expressed in transient mammalian systems. As well that might also explain why YMTV and TPV

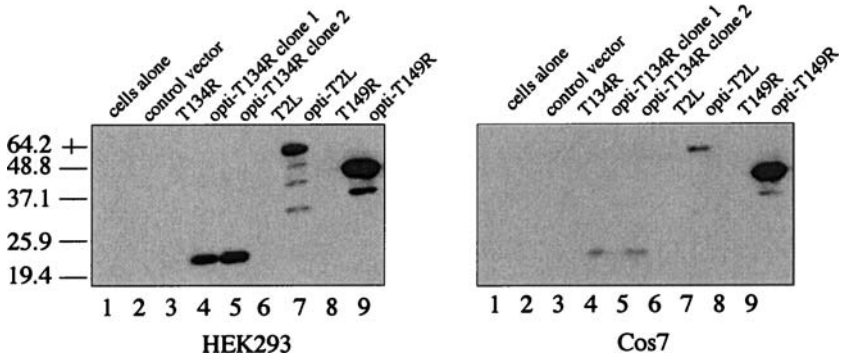

Fig. 6 Codon optimization of several tanapox virus genes improves the transient expression. The left panel illustrates the expression results following transfection of HEK293 cells with various plasmids encoding the native or optimized genes from tanapoxvirus. The right panel shows the same samples transfected into Cos7 cells

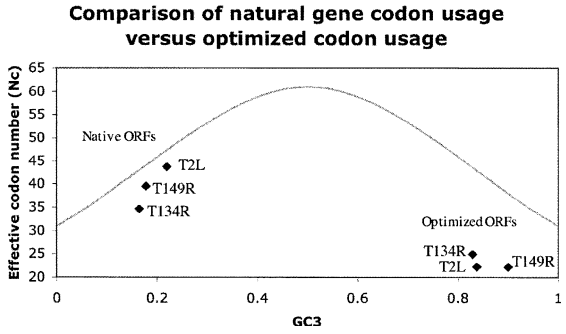

Fig. 7 Comparison of natural gene codon usage versus optimized codon usage. The effective codon numbers for three native TPV genes are compared against the effective codon numbers of the same genes following codon optimization

Table 3 Predicted expression of genes from various poxviruses

Poxvirus genomes whose genes are predicted to be resistant or difficult to express transiently in mammalian cells	Poxvirus genomes whose genes are predicted, or have been demonstrated to be well expressed in transient mammalian systems
Yatapoxviruses	Molluscipoxviruses
Entomopoxviruses	Parapoxviruses
Avipoxviruses	Leporipoxviruses
Capripoxviruses	
Suipoxviruses	
Orthopoxviruses	

genes are so well expressed in the baculovirus expression system. The high proportion of AT in the 3rd position is already adapted to the insect cells environment and may reflect an evolutionary history that involves replication within an insect host.

Examination of all poxvirus genomes and the proportion of GC content at each position of the codon indicate that all the genomes have a decreasing proportion GC at each successive position except for the leporipoxviruses, the parapoxviruses and the molluscipoxvirus (Table 2). For the five species within these three genera the highest proportion of GC occurs in the 3rd position whereas all other species the highest GC content occurs in the first position. The relationship between genomic GC content and GC_3 indicates that all pox genomes except for MV, SFV, MCV, orf and BPSV contain an overall GC content between 18% and 30% with a smaller proportion of GC at the 3rd position (Table 2). In contrast the other five genomes have an overall GC content that ranges between 40% (MV, SFV) to 65% (orf, BPSV, MCV) and in each case the GC content of the 3rd position is even higher at about 50% for MV and above 80% GC for MCV, orf and BPSV (Table 2).

Codon usage bias in the *Poxviridae* is related to GC content

The total GC content of the *Poxviridae* genomes range from 18% (AmEPV) to 65% (orf virus) GC (Table 2). However the GC_3% ranges from 8% (MsEPV) to a staggering 90% (BSPV). Poxviruses contain very little non-coding DNA within their genomes and since the first two codon positions are constrained by codon specificity requirements it would be predicted that the 3rd position would exhibit the most variation. We compared overall GC content to the GC content at each position of the codon (GC_1, GC_2 GC_3) calculated from the complete coding complement. The assumption was that the third or synonymous position of the codon would be under less selection pressure because of the redundancy of the amino acid coding. However we found that the highest correlation was between overall GC content and GC_1 ($r^2 = 0.98$) and GC_3 ($r^2 = 0.98$) (Fig. 8). All members of the *Poxviridae* maintained this strong correlation. And this relationship indicates that codon usage is tightly linked to individual GC content.

Discussion

Undetectable levels of transient gene expression of yatapoxvirus genes prompted us to examine codon usage in the family *Poxviridae*. Our results indicate that there are high-GC content (parapox- and molluscipoxviruses) and low-GC content (entomopoxviruses) poxviruses and it is those genomes with the largest GC extremes that exhibit the largest bias in codon usage. Codon usage bias in poxviruses is skewed in the direction of the overall GC content. We found that optimizing for codon usage resulted in dramatic improvement of the expression signal of transiently

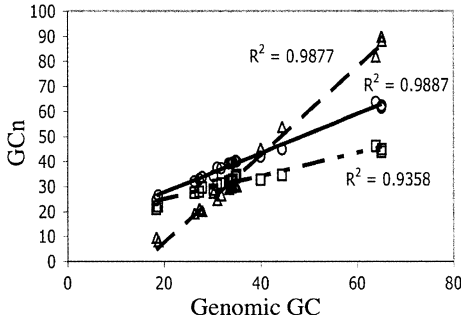

Fig. 8 G + C content at the three colon positions against the total GC content within each poxvirus member. The open circles represent GC_1 versus GC_{total}, the open squares indicate GC_2 versus GC_{total} and the open triangles are GC_3 versus GC_{total}. The lines represent best fit

expressed genes. In the case described here, the native sequence of tanapox T2L, which is AT rich and exhibits some codon bias was synthesized to increase the codon usage bias by increasing the GC concentration at the third codon position. The percent GC_3 in the native form of the gene was altered from less than 20% to greater than 80% GC in the codon optimized version (Fig. 2). Because poxviruses replicate exclusively within the cytoplasm the viral transcripts have not evolved under nuclear splicing or processing selection and this may explain the wide variation in GC content within the family *Poxviridae*. The 21 poxvirus members included in our study infect a wide range of hosts however they show similar trends between their genomic GC content and amino acid composition, and therefore the codon bias employed. Even members whose life cycle is restricted to infection of a single species, such as variola virus and molluscum contagiosum, which only infect humans, maintain amino acid composition related to their own specific GC content.

The observation that modification of the GC_3 in the optimized codons led to dramatic expression is not surprising because codon usage in other virus families is also related to GC content [12]. However poxviruses have two distinct features that make them unique. First they encode all the necessary transcription machinery within their virus factories in the cytoplasm and therefore do not rely on cellular components [13]. Second, although members of the *Poxviridae* infect a wide range of hosts including insects, birds, reptiles and mammals, with a few exceptions, individual poxvirus species have a narrow host range [11]. Therefore we can expect that individual poxviruses have adapted to the molecular features of their particular host. The genomes are well conserved suggesting a common ancestor and they have been incredibly successful. This is likely due to the fact that they do not require residency within the nucleus but rather construct their own virus factories within the cytoplasm.

There is a plasticity to the codon usage found in the poxviruses that does not necessarily reflect the common evolutionary history. We have examined three conserved poxvirus orthologs, which are predicted to function in a similar manner in members of the *Poxviridae* including DNA polymerase, P4a (the major core protein) and DNA uracil glycosidase and which all pox members encode however there is variable codon usage between the pox species for the same genes. The biases appear related to genomic GC content.

It has been suggested that codon bias reflects the level of gene expression and/or length of gene [14] however this does not seem to be supported in the *Poxviridae* because the codon usage for the same orthologs are different depending on the poxviral member (Fig. 5). It has also been suggested that codon bias could have evolved based on host requirements. However this does not hold for MCV, which has a GC rich genome (63.9% GC, Table 2) and variola virus which is more AT rich (33.4% GC, Table 2) however both replicate exclusively in human tissues. Perhaps the difference in GC concentration may be explained by the cell type or tissue in which the virus is resident. MCV is found exclusively in the keratinocytes of the dermis while variola virus can be found through out the body including in the lymphatic system, respiratory system and blood [10]. The incongruence between the %AT of a poxvirus genome and the AT concentration of its host genomic DNA has been noted before [15]. Capripox- and parapoxviruses both infect ungulates (sheep, goats, antelopes) however following the sequencing of 2.5 Kb of capripoxvirus DNA it was noted that the high AT concentration (72.4%) of the capripoxvirus DNA did not reflect the AT concentration of the evolutionary hosts [15]. Selected analyses have suggested that sheep and goats had AT concentrations around 50%. As well parapoxviruses, which share the same host range have a viral genomic content of around 37% AT [16]. Complete sequence data now confirms these earlier estimates. The parapoxviruses genomes (orf and BSPV) are 34.9% AT rich, while the capripoxvirus member (LSDV) is 73.3% AT rich (Table 2).

Unlike the situation in poxvirus genomes, analysis of the SARS coronavirus and other members of the Nidovirales indicated significant variation in codon usage bias among different genes within a species [17]. It was concluded that GC composition was the primary determinant of synonymous codon usage among these virus genes but the bias was manifested at the gene level rather than at the genome level [17].

A study on the codon usage in nucleopolyhedroviruses (NPV), another family of large dsDNA viruses, concluded that there was significant variation in codon usage by genes within the same virus. Again this is different from what we are reporting for the poxviruses. However the NPV study was based on six genes and we examined the complete complement of ORFs. Individual variation might be lost in the overall picture for the poxviruses. As well significant variation in codon usage in homologous genes encoded by different NPVs was observed. This is similar to our observations with poxviruses. Finally there was no correlation between level of gene expression and codon bias in NPV or between gene length and codon bias, and patterns of codon usage appeared to be a direct function of GC content of the virus encoded genes [12]. This is consistent with our observations reported here.

There are now examples from several virus families that indicate that alteration of the native codons will result in dramatically improved expression. In most cases the expression problem seems to be inappropriate codon usage. Native human papillomavirus (HPV)-16 E5 utilized infrequently used codons in 33 of its 83 amino acids and was undetectable following transient transfection however once the sequence was optimized for more common codons, used in mammalian genes, expression increased 6 to 9-fold [2]. Another HPV gene, L1, hampered by codon usage bias different from the host was corrected by codon optimization resulting in a 100-fold increase in expression levels [3].

In conclusion the members of the *Poxviridae* have genomes with a wide range of GC content and this appears to regulate their codon usage bias. The codon bias does not seem to be related to the size of the genes or their expression level because the codon bias seems to be maintained within genomes but not between genera. Optimizing codon usage has improved the transient expression of several pox genes in mammalian cells. Based on the calculation of the effective codon number for all ORFs from all complete genomes we would predict that the best species to study by transient expression of native genes should from the parapox-, mollusci- and leporipoxviruses genera. However those poxvirus members that are resistant to transient transfection and expression in human or non-human primate cells will likely benefit from codon optimization.

Acknowledgements GM is a Canada Research Chair in Molecular Virology. This research was supported by CIHR and NCIC. We thank T. Irvine for technical assistance and D. Hall for administrative support.

References

1. B.T. Seet, J.B. Johnston, C.R. Brunetti, J.W. Barrett, H. Everett, C. Cameron, J. Sypula, S.H. Nazarian, A. Lucas, G. McFadden, Annu. Rev. Immunol. **21**, 377–423 (2003)
2. G.L. Disbrow, I. Sunitha, C.C. Baker, J. Hanover, R. Schlegel, Virology **311**(1), 105–114 (2003)
3. N. Mossadegh, L. Gissmann, M. Muller, H. Zentgraf, A. Alonso, P. Tomakidi, Virology **326**(1), 57–66 (2004)
4. K.L. Nguyen, M. Ilano, H. Akari, E. Miyagi, E.M. Poeschla, K. Strebel, S. Bour, Virology **319**(2), 163–175 (2004)
5. C.R. Brunetti, M. Paulose-Murphy, R. Singh, J. Qin, J.W. Barrett, A. Tardivel, P. Schneider, K. Essani, G. McFadden, Proc. Natl. Acad. Sci. USA **100**(8), 4831–4836 (2003)
6. D.R. O'Reilly, L.K. Miller, V.A. Lucknow, *Baculovirus Expression Vectors: A Laboratory Manual*. (W.H. Freeman and Company, New York, NY 1992)
7. J.F. Peden, Analysis of codon usage, University of Nottingham (2000)
8. S. Mediratta, K. Essani, Can. J. Microbiol. **45**(1), 92–96 (1999)
9. F. Wright, Gene **87**(1), 23–29 (1990)
10. J.J. Esposito, F. Fenner, in *Fields Virology*, vol. 2, 4th edn. eds. by D.M. Knipe, P.M. Howley (Lippincott Williams & Wilkins, Philadelphia 2001), pp. 2885–2921
11. G. McFadden, Nat. Rev. Immunol. **3**, 201–213 (2005)
12. D.B. Levin, B. Whittome, J. Gen. Virol. **81**(Pt 9), 2313–2325 (2000)
13. B. Moss, in *Fields Virology*, vol. 2, 4th edn. eds. by D.M. Knipe, P.M. Howley (Lippincott Williams & Wilkins, Philadelphia 2001) pp. 2849–2883
14. J.R. Powell, E.N. Moriyama, Proc. Natl. Acad. Sci. USA **94**(15), 7784–7790 (1997)
15. P.D. Gershon, D.N. Black, J. Gen. Virol. **70**(Pt 3), 525–533 (1989)
16. R. Wittek, C.C. Kuenzle, R. Wyler, J. Gen. Virol. **43**(1), 231–234 (1979)
17. W. Gu, T. Zhou, J. Ma, X. Sun, Z. Lu, Virus Res. **101**(2), 155–161 (2004)