Tetranucleotide Profile of Herpesvirus DNA

Felix P. Filatov¹,², Alexander V. Shargunov¹

¹Mechnikov Federal Research Institute of Vaccines and Sera, 105064, Moscow, Russia;
²National Research Centre for Epidemiology and Microbiology named after the honorary academician N.F. Gamaleya,
123098, Moscow, Russia

Introduction. Herpesvirus DNAs (about 90% of the total genomic sequences of the Herpesvirales family presented in GenBank) contain at a minimum concentration one of the two tetranucleotides, CTAG or TCGA. The “underrepresentation” of CTAG was previously observed only in the DNA of some bacteria and phages. The aim of the study was the further analysis of the formal characteristics of herpesvirus DNA, as well as their comparison with the density of the virus/host DNA microhomology and with the genomic macrostructure of herpes viruses.

Materials and methods. Twenty strains and isolates of each of the five types of human herpes viruses (HHV1, HHV2, HHV3, HHV4, HHV5), 10 strains of HHV8, 5 strains of HHV6A, 4 strains of HHV6B and 3 strains of HHV7 were analyzed. GenBank tools were used to determine the frequency of tetranucleotides, and human DNA fragments with size matched herpesvirus DNA were used for comparison.

Results. Minimum CTAG concentration in DNA of herpes viruses is mainly characteristic of two- and single-segment genomes with direct or inverted terminal repeats (classes A, D, E), while the minimum TCGA density is characteristic mainly for DNA that is significantly less structured (classes B, C, F). By increasing CTAG density, human herpes viruses form a sequence close to the sequence of increasing the homology density of 20 nt with human DNA, which also correlates with the macrostructure of DNA. A parallel of this minimization with the DNA structure of herpes viruses or with their belonging to one or another subfamily — as well as the context of the “minimal” CpG (that is, TCGA) — is not noted in the literature. Although herpesvirus DNA is quite large (125–295 Kb), some of them (for example, HHV4, HHV5 and HHV7 DNA) show noticeable deviations from the second DNA parity rule, and can thus serve as a component of the molecular signature.

The Discussion suggests possible hypotheses for the origin of some of the observed phenomena.

Keywords: Herpesvirus DNA; tetranucleotide profile; CTAG/TCGA deficiency; Chargaff Second Parity Rule.

Acknowledgments. The study had no sponsorship.

Conflict of interest. The authors declare no apparent or potential conflicts of interest related to the publication of this article.

For citation: Filatov F.P., Shargunov A.V. Tetranucleotide profile of herpesvirus DNA. Journal of microbiology, epidemiology and immunobiology = Zhurnal mikrobiologii, epidemiologii i immunobiologii. 2020; 97(3): 216–226. (In Russ.).
DOI: https://doi.org/10.36233/0372-9311-2020-97-3-3

Received 22 April 2020
Accepted 25 May 2020

Тетрануклеотидный профиль герпесвирусных ДНК

Филатов Ф.П.¹,², Шаргунов А.В.¹

¹ФГБНУ «Научно-исследовательский институт вакцин и сывороток им. И.И. Мечникова», 105064, Москва, Россия;
²ФГБУ «Национальный исследовательский центр эпидемиологии и микробиологии имени почетного академика Н.Ф. Гамалеи», 123098, Москва, Россия

Введение. Герпесвирусные ДНК (около 90% всех полногеномных последовательностей семейства Herpesvirales, представленных в GenBank) содержат в минимальной концентрации один из двух тетranуклеотидов — СTAG или TCGA. «Недопредставленность» СTAG ранее наблюдалась только в ДНК некото рых бактерий и фагов. Ранее выявленная “недопредставленность” метилируемого димера СрГ находит своё выражение в низкой концентрации СTAG в ДНК герпесвирусов.

Материалы и методы. Проанализированы по 20 штаммов и изолятов каждого из пяти типов вирусов герпеса человека (HHV1, HHV2, HHV3, HHV4, HHV5), 10 штаммов HHV8, 5 штаммов HHV6A, 4 штамма HHV6B и 3 штамма HHV7. Для определения частоты тетranуклеотидов использовали инструменты GenBank, а для сравнения — фрагменты ДНК человека размером с ДНК герпесвирусов.

Результаты. Минимальная концентрация СTAG в ДНК герпесвирусов в основном характерна для двух- и односегментных геномов с прямыми или инвертированными концевыми повторами (классов А, Д и Е),
тогда как минимальная плотность TCGA — главным образом для значительно менее структурированной ДНК (классов B, C и F). По нарастанию плотности CTAG геномы герпесвирусов человека образуют после-
dовательность, близкую к последовательности 20 нт-гомологий ДНК герпесвирус/человек, организованной по нарастанию плотности, что также коррелирует с макроструктурой ДНК. Параллель этой минимизации со структурой ДНК вирусов герпеса или с их принадлежностью к тому или иному подсемейству в литера-
tуре не отмечена. Хотя герпесвирусные ДНК довольно велики (125–295 Кб), некоторые из них (например, ДНК HHV4, HHV5 и HHV7) демонстрируют заметные отклонения от второго правила четности ДНК и, таким образом, могут служить компонентом вирусных молекулярных сигнатур.

В Обсуждении предлагаются возможные гипотезы происхождения некоторых из отмеченных явлений.

Ключевые слова: герпесвирусная ДНК; тетрануклеотидный профиль; «недопредставленность» CTAG/TCGA; второе правило четности ДНК.

Источник финансирования. Авторы заявляют об отсутствии финансирования при проведении исследований.

Конфликт интересов. Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

Для цитирования: Филатов Ф.П., Шаргунов А.В. Тетрануклеотидный профиль герпесвирусных ДНК. Журнал микробиологии, эпидемиологии и иммунобиологии. 2020; 97(3): 216–226. DOI: https://doi.org/10.36233/0372-9311-2020-97-3-3

**Introduction**

Herpesviruses of the family *Herpesviridae*, including human HV, HHV, are divided into three sub-
families: alpha-HV, beta-HV and gamma-HV [1]. An-
other classification of HV is the classification by the
macrostructure of DNA (Fig. 1). It coincides not strict-
ly with the division into subfamilies and, in accordance
with generally accepted views, forms 6 classes from A
to F [2].

Alpha-HHV (HHV1, HHV2 and HHV3), as well
as beta one (HHV5) contain two-segment DNA; each
segment is bounded by mutually inverted monomeric
terminal repeats, TR₁ (DNA classes D and E). Class
A (HHV6A, HHV6B and HHV7) is an unsegmented,
unique linear sequence limited with direct monome-
ric terminal repeats containing two “islands” of telo-
mere-like hexanucleotides each. Gamma-HHV contain
DNA classes B and C, which have a unique sequence
limited to direct tandemly (non-monomeric) organized
short repeats, which number is not fixed (up to 45 in
case of HHV8), TR₂.

Class F DNAs are less structured or not structured
at all. There are herpesvirus DNAs with a more exo-
tic macrostructure (e.g., scutaviruses), but there are not
many. The data obtained by us in the proposed work
allowed us to combine classes A, D, E into one group
(DNA segments bounded by monomeric terminal re-
peats), and classes B, C, F into another (single DNA
segment bounded unfixed number of tandemly orga-
nized short terminal repeats).

Earlier, we noticed that the DNA molecules of
herpesvirus and its host contain short (20–29 nt) mu-
tually identical sequences, microhomologies, the con-
centration of which is not random and, we believe, is
explained by long (on an evolutionary scale) close in-
tergenomic relations between partners [3]. Later, we
found that such microhomologies have characteristic
distribution features in the herpesvirus genome, con-
centrating mainly in its terminal (direct or inverted)
repeats, especially in those regions of TR in which
there are no genes [4, 5]. But the most interesting thing
is that HHV species form the sequence of the virus/
host genomic microhomology by increasing the density,
which is consistent with the DNA macrostructure:
lower density in two-segment DNA, higher in single-
or non-segmented. We hypothesized that segments of

| DNA macrostructure | class | TR | HHV | subfamily |
|--------------------|-------|----|-----|----------|
|                   | D     | TR₁ | 1/2, 5 | alpha, beta |
|                   | E     | TR₁ | 3  | alpha    |
|                   | A     | TR₁ | 6A/Б, 7 | beta |
|                   | B     | TR₂ | 8  | gamma    |
|                   | C     | TR₂ | 4  | gamma    |
|                   | F     | TR₀ | –   | –        |

Fig. 1. Basic macrostructural classes of HHV DNA (the proportions of the lengths of the genome fragments are arbitrary).

TR₁ — monomeric terminal repeats (single rectangles), TR₂ — tandem organized repeats (non-fixed number of repeats),
TR₀ — no terminal repeats. BC[F] group of the HHV DNA (see text) is highlighted in gray.
viral genomes in which terminal repeats are mutually inverted may be blocked on themselves and interact less with host DNA. At the same time, single-segment ones having direct terminal repeats can be extended along the host DNA, which facilitates intergenomic interaction.

As an approach to analysis we used the comparison of nucleotide frequencies in DNA molecules and also the second Chargaff rule of parity, CPR2 [6], which becomes evident in DNA of more than 100,000 nt [7, 8]. CPR2 is formulated in the same way as the first one (CPR1), but applies only to one DNA strand. It has an approximate accuracy, which increases as the analyzed chain lengthens. It applies not only to mononucleotides but also to oligonucleotides up to 10–15 nt — with a decrease in accuracy as the analyzed oligonucleotide lengths [7, 9]. In metagenomics, tetranucleotide analysis is often used to form molecular signatures [10]. The frequency of tetranucleotides in the genomes of the herpes virus quite reliably corresponds to CPR2 and provides a more detailed characterization of DNA than mono-, di- and trinucleotides [7, 11]. In principle, the symmetries of the tetranucleotides of the herpes virus genome have been described previously [12], but they only confirmed the correspondence of CPR2. Our approach discovers other unusual properties of these genomes.

The aim of the study is to continue the analysis of the formal characteristics of herpesvirus DNA, as well as their comparison with the density of the virus/host DNA microhomology and with the genomic macrostructure of herpes viruses.

Materials and Methods

We analyzed ~ 90% of the nucleotide sequences of full-sized viral DNA molecules of each genus of all three families of vertebrate and invertebrate herpes viruses contained in GenBank. After analyzing 20 strains and isolates of each of the five types of human herpes viruses (HHV1, HHV2, HHV3, HHV4, HHV5), 10 strains of HHV8, all 5 strains of HHV6A, all 4 strains of HHV6B and all 3 strains of HHV7, we were convinced of the practical identity of intraspecific results (especially expressed in percent) and therefore, we present in the tables data only on the DNA of the reference strains of each type of herpes viruses.

For comparison, we used human DNA with a length of 1.5 megatons (5 fragments of 300,000 nt each):

- fragment Chr 03 163229646–163529646;
- fragment Chr 05 29372672–29672672;
- fragment Chr 14 64016329–64316329;
- fragments of Chr 21 15306102–15606102 and 33931862–34231862.

To determine the frequency of tetranucleotides, we used GenBank tools.

Results

We analyzed the tetranucleotide composition of the fully sequenced DNA of almost all herpes viruses of the Herpesvirales Order contained in the GenBank. The DNA type, that is GC (the predominance of G+C) or AT (the predominance of A+T) in one of its chains, does not give much in this regard, dividing all HHVs into two groups according to the types:

- type AT — HHV3 (alfa) and HHV6A, 6B, 7 (beta);
- type GC — HHV1,2 (alfa), 1, 6 (beta), and 4, 8 (gamma).

However, dinucleotide analysis illustrates well CPR2 [13], according to which A≈T, C≈G, C+T≈A+G and C+T≈A+G for one DNA strand. This is determined by the size of herpesvirus DNA — from 125 to 295 Kb.

The total number of tetranucleotides is 256 (4^4). To avoid the influence of a type of DNA on the results, which was shown previously [13, 14], we analyzed only tetranucleotides containing all four different bases, 4TNs. In HHV1 DNA (type GC), the smallest (“underrepresented”) is precisely such a tetranucleotide — CTAG (91 nt for the whole genome instead of ~600 nt in equal representation of all tetramers in the genome). In HHV6A DNA (type AT), the CTAG number is also close to the smallest (303 nt) among all tetramers and is the smallest of the tetranucleotides containing all four different bases. Only four tetramers, GGGC (245), ACCG (287), GGCC (288) and GGCT (296), are smaller — according to the type of DNA.

Of the 256 tetranucleotides, only 24 consist of all four nucleotides (P=4!=24). These 24, in turn, are divided into two groups: 8 of them (octet A) do not change during inversion, for example, CTAG|CTAG, the rest 16 (two octets B) are pairs B1 and B2 of mutually inverted non-identical tetranucleotides, for example, CTAG|CTAG. The tables and figures of octets A and B are shown separately. For correct comparison, the data are presented as a percentage of the sum of the frequencies of the tetranucleotides of each octet A and B. Table 1 compares the data for all 24 discussed tetranucleotides for all known types of control strains of human herpes virus. Table 1 shows that CTAG is “underrepresented” in the genomes of all HHVs, with the exception of HHV7. In HHV4 DNA, the TCAG tetramer is even more underrepresented (as in the human genome).

In accordance with the decrease in the “underrepresentation” of CTAG, the HHV genomes form a sequence that resembles the sequence of DNA microhomology virus/host by increasing their number (Fig. 2): the greatest “underrepresentation” of CTAG is characteristic of two-segment DNA, the smallest — for single-segment.

At the same time, the DNA of each HHV contains noticeably “overrepresented” tetranucleotides, which are also characteristic of the genomes of a certain mac-
rostructure: ACGT for two-segment DNA (classes D, E), TGCA for one-segment (class A, roseoloviruses) and CATG for one-segment (classes B, C). However, since the DNA of non-human herpes viruses is very poorly represented by host species in the GenBank, we did not further analyze the “maximum” tetranucleotides.

The columns of numbers related to the DNA of each virus are tetranucleotide DNA profiles, and they — in the case of HHV4, 8 and 7 — show a certain similarity with the profile of human DNA. In some cases (HHV4, 5, 7), octet B tetranucleotides presented in pairs (B1 and B2) demonstrate characteristic deviations from CPR2, which probably are associated with an insufficient DNA length of these viruses (HHV4,5) or with an insufficient number of strains in the GenBank, which do not provide sufficient reliability of the relevant data. The positive side of deviations from CPR2 is that they can be used as components of the molecular signatures of these viruses.

It is noteworthy that the difference between the maximum and minimum values in octet A is much larger, that is, more obvious than in octets B. In cases where the minimum density indices of tetranucleotides of octets B are less than octet A, their "underrepresentation" is directly related to type of DNA, that is, they have a drawback, which is the insufficient representation of data on these viruses in the GenBank.

### Table 1. Profile of identical (octet A) and non-identical (octets B1 and B2) tetramers containing four different nucleotides (4TN) of eight types of HHV (reference strains), expressed as a percentage of the total number of the octets A and B separately

| HHV | 1 | 2 | 3 | 5 | 6A | 6B | 7 | 4 | 8 | 4TN Human |
|-----|---|---|---|---|----|----|---|---|---|-----------|
| subfamily | alpha | beta | gamma |
| TR | 4TN octet A |
| CTAG | 2.8 | 2.6 | 4.9 | 5.0 | 7.5 | 7.3 | 13.9 | 9.2 | 8.3 | 12.2 |
| TCGA | 14.5 | 15.1 | 10.6 | 10.9 | 15.0 | 14.7 | 9.1 | 6.6 | 8.6 | 2.4 |
| AGCT | 12.7 | 13.7 | 8.2 | 11.2 | 10.3 | 11.2 | 17.7 | 18.4 | 15.1 | 18.9 |
| GATC | 13.2 | 13.5 | 12.9 | 11.6 | 13.6 | 14.0 | 10.8 | 10.3 | 10.0 | 11.4 |
| CATG | 15.1 | 13.1 | 15.7 | 13.7 | 13.8 | 14.4 | 14.1 | 19.6 | 16.2 | 21.6 |
| TGCA | 11.2 | 11.7 | 15.6 | 12.6 | 15.0 | 14.9 | 17.9 | 17.9 | 15.9 | 22.3 |
| ACGT | 16.9 | 17.5 | 18.7 | 20.9 | 14.8 | 14.1 | 8.1 | 9.6 | 13.7 | 3.0 |
| GTAC | 13.6 | 12.8 | 13.4 | 14.1 | 10.0 | 9.4 | 8.4 | 8.4 | 12.2 | 8.2 |
| 4TN octet B1 |
| CTGA | 5.3 | 5.4 | 3.9 | 5.9 | 6.9 | 7.1 | 7.6 | 8.9 | 7.4 | 11.2 |
| TACG | 7.2 | 7.0 | 8.3 | 7.6 | 5.9 | 5.3 | 3.4 | 2.6 | 4.8 | 1.0 |
| GCAT | 7.2 | 6.8 | 8.5 | 5.3 | 6.8 | 6.8 | 7.3 | 7.6 | 6.8 | 7.6 |
| AGTC | 5.5 | 5.4 | 4.3 | 5.9 | 5.7 | 5.7 | 5.9 | 7.9 | 5.8 | 6.3 |
| CATG | 5.2 | 4.6 | 6.1 | 5.9 | 6.0 | 6.3 | 7.8 | 8.9 | 8.0 | 9.6 |
| ATCG | 8.2 | 7.5 | 8.3 | 6.5 | 7.7 | 7.2 | 4.9 | 3.0 | 4.5 | 1.1 |
| GCTA | 3.4 | 3.2 | 4.4 | 4.4 | 4.1 | 4.0 | 4.9 | 4.8 | 5.1 | 6.3 |
| TGAC | 6.9 | 7.1 | 5.7 | 8.0 | 6.5 | 6.6 | 6.9 | 8.2 | 8.0 | 6.7 |

| 4TNs octet B2 |
| TCAG | 6.3 | 6.0 | 4.9 | 6.0 | 6.7 | 7.3 | 7.9 | 10.6 | 7.3 | 10.9 |
| CGTA | 7.5 | 8.0 | 7.9 | 7.7 | 6.1 | 6.0 | 4.0 | 3.3 | 4.7 | 1.0 |
| ATGC | 7.7 | 8.8 | 8.1 | 5.6 | 6.1 | 6.7 | 7.5 | 7.0 | 6.7 | 7.5 |
| GACT | 5.0 | 5.7 | 5.2 | 5.0 | 5.4 | 5.7 | 4.9 | 5.4 | 6.3 | 6.7 |
| ACTG | 5.0 | 4.5 | 6.3 | 5.9 | 6.9 | 6.7 | 7.3 | 6.2 | 8.7 | 10.0 |
| CGAT | 7.2 | 6.8 | 8.5 | 5.3 | 6.8 | 6.8 | 5.0 | 3.0 | 6.6 | 1.1 |
| TAGC | 3.7 | 3.1 | 4.7 | 5.2 | 5.3 | 5.1 | 7.8 | 4.8 | 4.4 | 6.3 |
| GTCA | 7.6 | 7.2 | 5.6 | 8.3 | 6.6 | 6.2 | 6.9 | 8.4 | 7.7 | 6.7 |

**Note.** Bold letters are tetramers CTAG and dimers Cpg in the tetramers discussed in text (corresponding cells of the both octets are highlighted in gray).
the format [TA/AT] [GC/CG]; the left and right tetramer pairs can be swapped, and the “/” means “or”. There are 8 such tetranucleotides, but for use — together with other tetramers — as molecular signatures, their involvement in the formation of a DNA type does not matter. Table 2 summarizes the data on the minimum (underrepresented) tetranucleotides of HHV DNA.

Next, we carried out a tetranucleotide analysis of completely sequenced DNAs of almost all other viruses of the Herpesvirales Superfamily of the GenBank NCBI. The data obtained are summarized in Table 3.

Table 3 shows that all herpesviruses are divided into two groups according two main underrepresented tetranucleotides — CTAG or TCGA. The difference between the two groups is parallel to their genomic macrostructure. Minimum CTAG (CTAGmin) is characteristic of structured DNA classes ADE, with large monomeric terminal repeats, TR1. TCGAmin is characteristic of less strictly structured DNA classes BCF with non-fixed tandem terminal repeats, TR2.

Discussion

The “underrepresentation” of CTAG tetranucleotide (CTAGmin) in the genomes of Escherichia and Salmonella, as well as some phages, has been known for quite some time [15] and continues to be studied [16]. For the first time, we systematically note here this feature for one of two large groups of herpesviruses and its parallel with their genomic structure. The larger group (ADE) is characterized by the presence of one or two segments bounded by TR1 monomeric terminal repeats, direct or mutually inverted. A smaller group of herpesviruses (BC[F]) contains a single-segment genome, limited by an undetermined number of tandem organized direct terminal repeats, TR2, and is characterized by the “underrepresentation” of another tetranucleotide, TCGA. The more general nature of the CTAGADE and TCGABCF parallels among animal viruses can be evidenced by the “underrepresentation” of CTAG beyond the Herpesvirales, in particular in the DNA of African swine fever viruses (Asfarviridae) and Shope rabbit fibromas (Poxviridae), structured similarly to the herpes viruses ADE. At the same time, the genomes of the smallpox and smallpox vaccines (Poxviridae) viruses are not structured in this way, and there is no “underrepresentation” of CTAG in them. These observations require serious expansion of the studies under discussion in other taxonomic groups of animal viruses.

An analysis of a series of strains (mainly up to 20) of the same type of HHV showed almost complete identity of the results, which, to a first approximation, allowed us to consider the results obtained quite reliable.

Table 2. A generalized version of the data on the human herpesviruses DNA (family Herpesviridae)

| Subfamily | Genus                | Species | Reference  | Size of the DNA, Kb | Class | Type | 4TNmin | Number of studied strains |
|-----------|----------------------|---------|------------|---------------------|-------|------|--------|--------------------------|
| alpha     | Herpes simplex virus | HHV1    | NC_001806  | 155                 | E     | GC   | CTAG   | 20                       |
|           | Herpes simplex virus | HHV2    | NC_001798  | 155                 | E     | GC   | CTAG   | 20                       |
|           | Varicella-Zoster virus | HHV3    | NC_001348  | 125                 | D     | AT   | CTAG   | 20                       |
| beta      | Cytomegalovirus      | HHV5    | NC_006273  | 236                 | E     | GC   | CTAG   | 20                       |
|           | Roseoloviruses       | HHV6A   | NC_001664  | 159                 | A     | AT   | CTAG   | 5                        |
|           |                      | HHV6B   | NC_000898  | 162                 | A     | AT   | CTAG   | 4                        |
|           |                      | HHV7    | NC_001716  | 153                 | A     | AT   | ACGT   | 3                        |
| gamma     | Lymphocryptovirus    | HHV4    | NC_007605  | 172                 | C     | GC   | TCGA   | 20                       |
|           | Rhadinovirus         | HHV8    | NC_009333  | 138                 | B     | GC   | CTAG   | 10                       |

Note. Non-CTAGmin DNAs are highlighted in italics (CpG — in bold italics). HHVs in the DNA with CpG>GpC are highlighted in gray.
Table 3. A generalized version of the data on the DNA of animal herpesviruses (families Herpesviridae, Alloherpesviridae and Malacoherpesviridae)

| Subfamily | Genus | Species | Reference | Size of the DNA, Kb | Class | Type | $4TN_{nm}$ |
|-----------|-------|---------|-----------|--------------------|-------|------|------------|
| alpha     | Iltovirus | Gallid AHV1 | NC_006623 | 149 | D | AT | CTAG |
|           | Psittacid AHV1 | NC_005264 | 163 | D | GC | CTAG |
|           | Anatidia | Anatid AHV1 | NC_013036 | 158 | F | AT | CTAG |
|           | Columbid AHV1 | NC_034266 | 204 | E | GC | CTAG |
|           | Falconid AHV1 | NC_024450 | 204 | E | GC | CTAG |
|           | Gallid AHV2 | NC_002229 | 178 | E | AT | CTAG |
|           | Gallid AHV2 | MF431495 | 178 | E | AT | CTAG |
|           | Gallid AHV3 | NC_002577 | 164 | E | GC | CTAG |
|           | Meleagrid AHV1 | NC_002641 | 159 | E | AT | CTAG |
|           | Sphenicid AHV1 | NC_033464 | 165 | D | AT | CTAG |
|           | Scutavirus | Testudinid HV3 | NC_002794 | 196 | D* | AT | TGCA |
| Simplex virus | Papiine AHV2 | NC_007453 | 156 | E | GC | CTAG |
|           | Saimiriine AHV1 | NC_014567 | 157 | D | GC | TGCA |
|           | Fruit bat AHV1 | NC_024306 | 149 | E | GC | GTAC |
| Varicella virus | Bovine AHV1 | NC_001847 | 135 | D | GC | CTAG |
|           | Bovine AHV5 | NC_005261 | 138 | F | GC | CTAG |
|           | Bubaline AHV1 | NC_043054 | 137 | F | GC | CTAG |
|           | Cercopithecine AHV9 | NC_002686 | 125 | D | AT | CTAG |
|           | Equid AHV3 | NC_024771 | 184 | E | GC | CTAG |
|           | Suid AHV1 | NC_006151 | 143 | D | GC | CTAG |
|           | Canid AHV1 | NC_030117 | 125 | D | AT | TCGA |
|           | Equid AHV4 | NC_001844 | 146 | D | GC | TCGA |
|           | Felid AHV1 | NC_013590 | 136 | D | AT | TCGA |
|           | Equid AHV1 | NC_001491 | 150 | D | GC | GATC |
|           | Equid AHV8 | NC_017826 | 149 | F | GC | GATC |
|           | Equid AHV9 | NC_011644 | 148 | D | GC | GATC |
| beta     | Cytomegalovirus | Aotine BHV1 | NC_016447 | 219 | E | GC | CTAG |
|           | Caviid BHV2 | NC_020231 | 234 | A | GC | CTAG |
|           | Cercopithecine BHV5 | NC_012783 | 226 | A | GC | CTAG |
|           | Papio ursinus CMV | NC_027016 | 226 | F | GC | CTAG |
|           | Cynomolgus CMV | NC_033176 | 224 | A | AT | CTAG |
|           | Macacine BHV3 | NC_006150 | 221 | F | AT | CTAG |
|           | Panine BHV2 | NC_003521 | 241 | D | GC | CTAG |
|           | Saimiriine BHV4 | NC_016448 | 197 | E | AT | CTAG |
| Subfamily | Genus | Species | Reference | Size of the DNA, Kb | Class | Type | 4TN<sub>min</sub> |
|-----------|-------|---------|-----------|--------------------|-------|------|------------------|
| Muromegalovirus | Murid BHV1 | NC_004065 | 230 | F | GC | CTAG |
| | Murid BHV8 | NC_019559 | 203 | F | AT | CTAG |
| | Rat CMV Maastricht | NC_002512 | 230 | A | GC | CTAG |
| Proboscivirus | Elephant BHV4 | NC_028379 | 206 | F | GC | CTAG |
| | Elephant BHV5 | NC_024696 | 181 | A | AT | CTAG |
| | Elephantid BHV1 | NC_020474 | 180 | A | AT | CTAG |
| Roseolovirus | Murine roseolovirus | NC_033620 | 174 | F | AT | CTAG |
| | Macaca nemestrina | NC_030200 | 137 | A | AT | CTAG |
| | Suid BHV2 | NC_022233 | 128 | A | AT | CTAG |
| gamma | Macavirus | Alcelaphine GHV1 | NC_002531 | 131 | F | AT | TCGA |
| | | Alcelaphine GHV2 | NC_024382 | 137 | F | AT | TCGA |
| | | Bovine GHV6 | NC_024303 | 145 | F | AT | TCGA |
| | | Ovine GHV2 | NC_007646 | 135 | F | AT | TCGA |
| Percavirus | Felis catus GHV1 | NC_028099 | 123 | F | AT | TCGA |
| | | Equid GHV5 | NC_026421 | 182 | B | GC | TCGA |
| | | Equid GHV2 | NC_001650 | 184 | A | GC | TCGA |
| Rhadinovirus | Ateline GHV3 | NC_001987 | 108 | F | AT | TCGA |
| | | Cricetid GHV2 | NC_015049 | 124 | F | AT | TCGA |
| | | Murid GHV4 | NC_001826 | 119 | F | AT | TCGA |
| | | Saiminiine GHV2 | NC_001350 | 113 | F | AT | TCGA |
| | | Dolphin GHV1 | NC_035117 | 167 | F | AT | CTAG |
| | | Macacine GHV5 | NC_003401 | 134 | F | GC | CTAG |
| Lymphocryptovirus | Callitrichine GHV3 | NC_004367 | 150 | F | AT | TCGA |
| | | Macacine GHV4 | NC_006146 | 171 | F | GC | TCGA |
| Unclassified gamma | Rhinolophus GHV1 | NC_040539 | 148 | A | AT | TCGA |
| Unclassified gamma | Eptesicus fuscus GHV | NC_040615 | 167 | F | GC | CTAG |

**Family: Alloherpesviridae (pisces and amphibia HV)**

| | | | | | | |
| | | | | | | |
| Cyprinivirus | Anguillid HV1 | NC_013668 | 249 | A | GC | TCGA |
| | | Cyprinid HV2 | NC_019495 | 290 | A | GC | TCGA |
| | | Cyprinid HV3 | NC_009127 | 295 | A | GC | TCGA |
| | | Cyprinid HV1 | NC_019491 | 291 | A | GC | TCGA |
| Ictalurivirus | Ictalurid HV1 | NC_001493 | 134 | F | GC | TCGA |
| | | Ictalurid HV2 | NC_036579 | 143 | A | GC | TCGA |
| Batrachovirus | Ranid HV1 | NC_008211 | 221 | A | GC | CTAG |
| | | Ranid HV2 | NC_008210 | 232 | A | GC | CTAG |

**Family: Malacoherpesviridae (invertebrates HV)**

| | | | | | | |
| | | | | | | |
| Aurivirus | Halilotid HV1 | NC_018874 | 212 | E | AT | CTAG |
| Ostreavirus | Ostreid HV1 | NC_005881 | 207 | F | AT | CTAG |

**Unclassified Herpesvirales**

| | | | | | | |
| | | | | | | |
| Unclassified | Bufonid HV1 | NC_040681 | 158 | F | AT | TCGA |

**Note.** Gray cells — BC [F] class DNA and “minimal” non-CTAG TN. Asterix in D* denotes the unusual macrostructure of the testudinid HV3 DNA, in which two approximately equal segments bounded by terminal repeats are separated by a short unique sequence.
The thermodynamic model of RNA shows that the tetramer CTAG(CUAG) violates the optimal structure of the stem loops of the molecule, which control the expression of genes, increasing their free energy. The authors of this hypothesis [15] suggested that the common ancestor of Salmonella and Escherichia had a significantly higher CTAG density, but evolutionary degeneration led to the replacement of CTAG in its descendants with a tetranucleotides neutral in this respect, and this trend is currently maintained. In a number of genes and in intergenic spaces in Escherichia and Salmonella, the indicated degeneracy led to the evolutionary replacement of CTAG, primarily with CTGG (to a lesser extent with ATAG, CTTT, CTTG).

In this regard, it is most appropriate to compare phylogenetically related (p.e. the same genus) human roseoloviruses HHV6 and HHV7. In the DNA of both viruses — compared with other herpes viruses — the frequencies of CTAG and CTGG are most different. The comparison shows that if in HHV7 (NC_001716) the frequency ratio CTAG/CTGG is 530:301, respectively, then in HHV6A (NC_001664) it is even the opposite and is 303:391 with close DNA sizes of both viruses. If the LeTang et al. [16] observation is also applicable to herpesviruses, then HHV7 is obviously closer to the evolutionary predecessor of both roseoloviruses than HHV6, in which many CTAG was replaced by CTGG. At the same time, HHV6 acquired the ability to integrate its genome into the host genome, which is not, as a rule, a prerequisite for closer relations with the host DNA, as evidenced by the similarity of the profile of the HHV7 tetranucleotide (not HHV6) and human DNA (Fig. 3), as well as a higher level of virus/host DNA microhomology in HHV7 than in HHV6, or a lower level of such microhomology in mardiviruses with pronounced telomeric islands in terminal repeats of DNA segments.

In Figure 3 some features of the analyzed 4TN profiles are additionally indicated. In accordance with the Second DNA Parity Rule, the similarity between B1 and B2 in human DNA is much greater (than in virus ones), since the fragments of human DNA we have analyzed are 300 Kb long, and the HPV genomes are much shorter. In HHV7, the differences between B1 and B2

Fig. 3. 4TN profile of HHV6A and HHV7 DNA compared to human DNA 4TN profile.
Octet A: human DNA is highlighted in gray, viral DNA — in black.
Octet B: human B1 is highlighted in light gray, human B2 is highlighted in dark gray, virus B1 is highlighted in bold black, B2 — in thin black.
are characteristic enough and may be used as an element of the molecular signature of the DNA of this virus (the same applies to the 4TN HHV4 profile, see Figure 4 below). The fact that GenBank represents the complete (almost complete) DNA sequences of only three HHV7 strains allows the use of statistical methods to validate the data presented here with great reservations. For this reason, we did not use these methods here, noting only that today it looks like a fact.

Figure 4 compares the 4TN DNA profile of another pair of viruses, HHV1 and HHV4. In the case of HHV1, a low content of CTAG allows the virus to cause an acute productive infection and accumulate in the cells of the entrance gate (fibroblasts), and then go into neurons, where it will remain for life — in particular, due to the inhibitory effect of host epigenetic mechanisms, one of which is methylation viral DNA. The concentration of CpG dinucleotides in the genome of HHV1 exceeds the average value, Table 1.

Low levels of CTAG can play a role in exacerbat-
ing latent infections. In the case of HHV4, the primary lytic infection is not characterized by a high level of viral syntheses, and after the transition to the chronic phase it is also regulated by epigenetic tools, including methylation C in CpG [17, 19]. At the same time, the obvious proximity of the 4TN profiles of the genome of HHV4 and the host indicate a similar response to these tools. The same can be said about HHV8 and the epigenetic regulation of its genes [20, 21]. Of the many epigenetic mechanisms that modify the expression of viral and host genes, we consider here only DNA methylation, more precisely, cytosine methylation in CpG, since this dimer is part of 4TN (TCGA), which allows it to be compared with another tetramer, CTAG, in proposed here aspect.

The hypothesis of a low density of CTAG tetramer due to its evolutionary degeneration does not explain the obvious limitations of its use and does not at all concern the reasons for the low density of another tetramer of octet A, TCGA, in the DNA of the members of the same superfamily. About 40% of CpG, the central pair of this tetramer, is located in the promoter zones of mammals [21, 22] and has a much lower density in complete sequences of vertebrate genomes than

![Fig. 4. 4TN profile of HHV1 and HHV4 DNA compared to human DNA 4TN profile. Octet A: human DNA is highlighted in gray, viral DNA — in black. Octet B: human B1 is highlighted in light gray, human B2 is highlighted in dark gray, virus B1 is highlighted in bold black, B2 — in thin black.](image-url)
might be expected [23, 24]. This underrepresentation is a consequence of the high frequency of mutations of methylated CpG sites in the genomes of hosts and their viruses, especially of those that interact closely with host DNA.

The reasons for the lowered CpG content were repeatedly discussed before [25], however, the issue is not the low density of CpG, but rather the context of this pair, i.e. in the TCGA, since this tetramer is present in herpesvirus DNA in a much lower concentration than ACGT.

Data from Le Tang et al. [16] show that the minimum content of CTAG (and TCGA) alone is not limited to herpesvirus DNA. We analyzed the tetranucleotide profile of large DNAs together with terminal repeats of some other viruses. CTAG was found to be “minimal” in African swine fever viruses (Asfarviridae family) and Shope fibroma virus (Poxviridae family), but not in smallpox and vaccinia viruses (also in the Poxviridae family), whose DNA does not have terminal repeats. This means that when constructing phylogenetic trees, it is necessary to take into account not only changes in genes and proteins, but also the evolution of the DNA molecule, including its characteristics discussed here.

In a first approximation, to analyze the density of potentially methylated cytosine in the herpesvirus genomes, it suffices to estimate the CpG:GpC ratio, which is not related to the genome type (AT or GC). This estimate is shown in Table 2: HHV DNA with CpG>GpC (darkened cells). In this case, the results presented here would concern only the concentration and the ratio CTAG/CpG in herpesvirus DNA, which may affect the level of viral synthesis. In its most general (non-strict) form, this ratio has a mirror character: the lowest concentration of CTAG is accompanied by the highest concentration of CpG (Table 1) — at least within the framework of the groups of classes ADE, BC[F]. Nevertheless, the ratio CTAG/CpG depletes our results, which indicate a difference in the tetranucleotide profile of herpesvirus DNA, specifying this ratio to CTAG/TCGA. In other words, the component of this ratio is the ratio of TCGA/ACGT, clearly expressed in the framework of the classes DE/A/BC[F], Table 1. In turn, this means the need to take into account the context that determines the functional value of the CpG dimer. Perhaps this context goes beyond the tetramer. But for reliable conclusions, it is necessary to expand the research beyond the scope of herpesviruses and seriously enrich GenBank with new complete viral DNA sequences. But in any case, the results demonstrated by us here indicate that the biological meaning of the macrostructure of herpesvirus DNA is much deeper than is commonly believed.

Л И Т Е Р А Т У Р А / REFERENCES

1. Whitley R., Kimberlin D., Prober C. Pathogenesis and disease. In: Arvin A., Campadelli-Fiume G., Mocarsky E., Moore P.S., Roizman B., Whitley R., eds. Human Herpesviruses: Biology, Therapy and Immunoprophylaxis. Chapter 32. Cambridge: Cambridge University Press; 2007.

2. Pellett P., Roizman B. Herpesviridae. In: Knipe D.M., Howley P.M., eds. Fields Virology. Philadelphia: Lippincott Williams & Wilkins; 2013: 1802-2.

3. Zabolotneva A., Tkachev V., Filatov F., Buzdin A. How many antiviral small interfering RNAs may be encoded by the mammalian genomes? Biol. Direct. 2010; 5: 62.

4. Filatov F., Shargunov A. Short nucleotide sequences in herpesviral genomes identical to the human DNA. J. Theor. Biol. 2015; 372: 12-21.

DOI: http://doi.org/10.1016/j.jtbi.2015.02.019

5. Filatov F., Shargunov A. Microhomology of Viral/Host DNAs and macrostructure of herpesviral genome. Int. J. Virol. AIDS. 2018; 5(1): 042.

DOI: http://doi.org/10.23937/2469-567X/1510042

6. Rudner R., Karkas J.D., Chargaff E. Separation of B. subtilis DNA into complementary strands, 3. Direct Analysis. Proc. Natl. Acad. Sci. USA. 1968; 60(3): 921-2.

DOI: http://doi.org/10.1073/pnas.60.3.921

7. Forsdyke D.R. Symmetry observations in long nucleotide sequences: a commentary on the discovery note of Qi and Cuticchia. Bioinformatics. 2002; 18(1): 215-7.

DOI: http://doi.org/10.1038/bioinformatics.18.1.215

8. Albrecht-Buehler G. Asymptotically increasing compliance of genomes with Chargaff’s second parity rules through inversions and inverted transpositions. Version 2. Proc. Natl. Acad. Sci. USA. 2006; 103(47): 17828-33.

DOI: http://doi.org/10.1073/pnas.0605553103

9. Bainsee P.F., Hampson S., Baldi P. Why are complementary strands symmetric? BioInformatics. 2002; 18(8): 1021-33.

DOI: http://doi.org/10.1093/bioinformatics/18.8.1021

10. Gori F., Mavroedis D., Jetten M.S.M., Marchiori E. The importance of Chargaff’s second parity rule for genomic signatures in metageneomics. Available at: http://www.biorxiv.org/content/biorxiv/early/2017/06/04/146001.full.pdf

11. Pride D.T., Blaser M.J. Identification of horizontally acquired genetic elements in Helicobacter pylori and other prokaryotes using oligonucleotide difference analysis. Genome Lett. 2002; 1(1): 2-15. DOI: http://doi.org/10.1166/gl.2002.003

12. Prabhu V.V. Symmetry observations in long nucleotide sequences. Nucleic Acids Res. 1993; 21(12): 2797-800.

DOI: http://doi.org/10.1093/nar/21.12.2797

13. Albrecht-Buehler G. The three classes of triplet profiles of natural genomes. Genomics. 2007; 89(5): 596-601.

DOI: http://doi.org/10.1016/j.ygeno.2006.12.009

14. Zhang S.H., Wang L. A novel common triplet profile for GC-rich prokaryotic genomes. Genomics. 2011; 97(5): 330-1.

DOI: http://doi.org/10.1016/j.ygeno.2011.02.005

15. Burge C., Campbell A.M., Karlin S. Over- and under-representation of short oligonucleotides in DNA sequences. Proc. Natl. Acad. Sci. USA. 1992; 89(4): 1358-62.

DOI: http://doi.org/10.1073/pnas.89.4.1358

16. Lang L., Zhu S., Mastriani E., Fang X., Zhou Y., Li Y., et al. Conserved intergenic sequences revealed by CTAG-profiling in Salmonella: thermodynamic modeling for function prediction. Sci. Rep. 2017; 7: 43565.

DOI: http://doi.org/10.1038/srep43565

17. Bhende P.M., Seaman W.T., Delecluse H.J., Kenney S.C. The EBV lytic switch protein, Z, preferentially binds to and activates the methylated viral genome. Nat. Genet. 2004; 36(10): 1099-104. DOI: http://doi.org/10.1038/ng1424

18. Kaufer B.B., Flamand L. Chromosomally integrated HHV-6: impact on virus, cell and organismal biology. Curr. Opin. Virol. 2014; 4: 111-8.

DOI: http://doi.org/10.1016/j.coviro.2014.09.010
19. Woellmer A., Hammerschmidt W. Epstein-Barr virus and host cell methylation: regulation of latency, replication and virus reactivation. *Curr. Opin. Virol.* 2013; 3(3): 260-5. DOI: http://doi.org/10.1016/j.coviro.2013.03.005

20. Lim C., Lee D., Seo T., Choi C., Choe J. Latency associated nuclear antigen of Kaposi’s sarcoma-associated herpesvirus functionally interacts with heterochromatin protein 1. *J. Biol. Chem.* 2003; 278(9): 7397-405. DOI: http://doi.org/10.1074/jbc.M211912200

21. Pantry S.N., Medveczky P.G. Epigenetic regulation of Kaposi's sarcoma associated herpesvirus replication. *Semin. Cancer Biol.* 2009; 19(3): 153-7. DOI: http://doi.org/10.1016/j.semcancer.2009.02.010

22. Fatemi M., Pao M.M., Jeong S., Gal-Yam E.N., Egger G., Weisenberger D.J., et al. Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. *Nucleic Acids Res.* 2005; 33(20): e176. DOI: http://doi.org/10.1093/nar/gni180

23. Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., et al. Initial sequencing and analysis of the human genome. *Nature.* 2001; 409(6822): 860-921. DOI: http://doi.org/10.1038/35057062

24. Stevens M., Cheng J., Li D., Xi M., Hong C., Maire C., et al. Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Res.* 2013; 23(9): 1541-53. DOI: http://doi.org/10.1101/gr.152231.112

25. Nicholas J. Evolutionary aspects of oncogenic herpesviruses. *Mol. Pathol.* 2000; 53(5): 222-37. DOI: http://doi.org/10.1136/mp.53.5.222

Information about the authors:
Felix P. Filatov— PhD (Med.), D. Sci. (Biol.), leading researcher, Laboratory of molecular biotechnology, Mechnikov Federal Research Institute of Vaccines and Sera, 105064, Moscow, Russia; leading researcher, National Research Centre for Epidemiology and Microbiology named after the honorary academician N.F. Gamaleya, 123098, Moscow, Russia. ORCID ID: https://orcid.org/0000-0001-6182-2241. E-mail: felix001@gmail.com

Alexander V. Shargunov — leading engeneer, Laboratory of DNA-containing viruses genetics, Mechnikov Federal Research Institute of Vaccines and Sera, 105064, Moscow, Russia. ORCID ID: https://orcid.org/0000-0001-5536-1557.

Contribution: the authors contributed equally to this article.

Информация об авторах:
Филатов Феликс Петрович — к.м.н., д.б.н., в.н.с. лаб. молекулярной биотехнологии ФГБНУ НИИВС им. И.И. Мечникова, 105064, Москва, Россия; ведущий научный сотрудник ФГБУ «НИЦ эпидемиологии и микробиологии им. почетного академика Н.Ф. Гамалеи», 123098, Москва, Россия. ORCID ID: https://orcid.org/0000-0001-6182-2241. E-mail: felix001@gmail.com

Шаргунов Александр Валерьевич — ведущий инженер, лаб. генетики ДНК-содержащих вирусов ФГБНУ НИИВС им. И.И. Мечникова, 105064, Москва, Россия. ORCID ID: https://orcid.org/0000-0001-5536-1557.

Участие авторов: все авторы сделали эквивалентный вклад в подготовку публикации.