THE GENERATION OF CODING SEQUENCES OF CELLULAR GENOME THROUGH COOPTION OF VIRAL GENES

Popov N. N., Sklyar N.I., Kolotova T. Yu., Davydenko M. B., Voronkina I. A.

Mechnikov Institute of Microbiology and Immunology

Introduction

Now the symbiotic relationships between animals and plants from one side and microorganisms from the other are becoming increasingly evident. These relationships are formed not only between bacteria and protists, but between hosts and viruses as well. The awareness of the symbiotic relationships between the host organisms and viruses has emerged only lately, before viruses were regarded as parasites only. The reasons for such view on the role of viruses is that the symbiotic relationships are hidden and parasitic are evident. But as relationships between humans and viruses were being studied, the scientists have realized that the parasitic relationships leading to the development of pathologies are rather an exception, not a rule in the viral – host relationships. In the majority of cases, cooperative relationships are established between them [1].

The integration of the viral genomes into the genomes of the hosts, included the RNA viruses genomes, takes place with the unexpected high frequency [2]. During the studies of the vertebral genomes, the representatives of a number of RNA and DNA viruses were found, including the Ebola virus, filoviruses, bornaviruses, circoviruses, hepadnaviruses and parvoviruses [3-7]. In course of evolution, the exogenous retroviruses, by integrating into the genome, were transformed into one of the fractions of the mobile elements called endogenous retroviruses (ERV). RNA viruses integrate into the genome with the help of retroviral reverse transcriptase. Sometimes, the integration is accompanied by recombination with other viral sequences [8]. The integration of DNA viruses into the genome can take place during the reparation of the DNA double strand breaks with the help of the non-homologous ends connection mechanism [9].

Endogenous viruses change the genome both actively and passively. For instance, ERV retroviruses facilitate passive changes by homologous recombination that promotes the appearance of duplications, deletions or karyotype changes [10].

Endogenous retroviruses can actively change the genome with the help of the cis-regulatory elements they contain, that rewiring the regulatory networks. Besides regulatory networks building and rewiring, the viruses participate in acquisition of new genes by the genome of the host. This happens because of the cooption of the viral genes by the host genome, through the horizontal gene transfer, integration of the retrocopies of the genes into the host cellular genome and gene duplications. In the first part of the present review the processes of cooption by the host genome of the genes of viral descent.

Retroviral gene cooption

The genes of endogenous retroviruses can be coopted by the genome of the host organism and used by the host for its own purposes. One of the well-know examples of such cooption is the syncytin genes.

The first syncytin that was found in humans – syncitin-1 originated from the capsid protein gene that was coded by the *env* gene of the defective endogenous virus HERVW [11]. Syncitin promotes the merging of trophoblasts, which leads to the formation of syncitiotrophoblasts. Some syncitins, but not all, possess immune suppressive properties, suppressing immune response towards the fetus [12]. Besides the placenta arising syncytin–1 also participates in formation of multinuclear osteoclasts [13]. According to the latest data, the murine syncytin participates in the myocytes fusion [14].

Syncitin– 2 also causes the cellular fusion and originated from the *env* gene of the endogenous HERV-FRD retrovirus [15].

In humans, syncitin-1 causes the cellular fusion through interaction with the Na^+-dependent transporter of the neutral amino acids of the second type - hASCT2 [16]. Syncitin – 2 forms the receptor of MFSD2 that supposedly transports carbohydrates [17].

One more placenta-specific protein has originated from the *env* gene of the HERVF retrovirus - supressin. The protein suppresses the cell fusion by interacting with the hASCT2 receptor and prevents the syncitin-1 binding with the transport channel [18].

Two syncytin genes, syncytin–A and syncytin-B were found in mice. Both are expressed in placenta and cause the trophoblast fusion.

Besides humans and mice, syncytin-like genes were found in squirrel-like rodents, lagomorphs, guinea pigs, ruminants, predators, and other animals [19]. Expression of syncytin was found even in opossums that belong to the marsupial animals [20]. The most interesting aspects consist in the fact that all syncytin genes developed as a result of independent invasion of different exogenous retroviruses into the germ cells genomes. Therefore, in course of mammalian evolution, independent invasion of retroviruses and “taming” of the syncytin genes was taking place.

No precursor syncytin genes that would be present in the common ancestor of the placental mammals were found up to date. It leads the scientists to an idea that either the common ancestor did not have the syncitine genes, or such genes were present, but were replaced in different lineages by different retroviral genes.

Besides syncytins, a number of other genes that originated from the *env* gene of the HERV virus are expressed in the human genome in normal conditions. System search in the human genome led to the discovery of 18 genes that code Env proteins [21].

One of the gene that codes for the Env-like protein HEMO [human endogenous MER34 (medium-reiteration-frequency-family-34)], is one of the most ancient full-sized *env* gene identified in the human genome. The HEMO protein is released into the human blood by shedding. The HEMO protein is actively expressed in blood stem cells, as

DOI: 10.5281/zenodo.2547443

well as in placenta that leads to the increase in the protein concentration in pregnant women. The HEMO gene expression in embryogenesis starts on the 8-cell embryo stage and continues in the next stages of embryogenesis. HEMO can be a cytokine or a hormone [22]

Many eukaryotic retroelements have long terminal repeats (LTR) but do not contain the capsid gene *env*. Those are so-called viral retrotransposons that are most likely descended from the retroviruses that integrated into the genome [23, 24]. Retroviruses are able to translocate between cells, whereas LTR retrotransposons are not. This difference is due to the presence of the capsid gene *env* that is required for the invasion of the cell by the retrovirus. The loss or the acquisition of the *env* gene leads to the evolutionary switch between the LTR retrotransposons and retroviruses.

For instance, the LTR retrotransposon Gypsy, by capturing the gene coding for the capsid protein of the baculovirus, acquires the ability to form viral particles and becomes an infectious retrovirus [23].

The retroviral Gag protein consists of at least three functionally different domains: matrix domain that is involved in the attachment of the virus to the cellular plasma membrane; capsid domain that determines the protein-protein interactions in the viral particles formation; and nucleocapsid domain that contains a zinc finger-type motif that can interact specifically with the nucleic acids. Zinc finger-type motif promotes the attachment of the Gag protein to the RNA genome of the virus.

There are at least 85 genes in the human genome that form three families - Mart, Pnma and SCAN have originated from the Ty3/gypsy family of LTR retrotransposons that code for the capsid *gag* proteins. Many of these genes are conservative in the genomes of other mammals [19].

The Mart family ('Mammalian Retro Transposon') of the placental mammals consist of the 11 genes and originated from the retrotransposon Sushi that belongs to the Ty3/Gypsy retrotransposon family [25].

Most of the genes of the family are located on the mammalian X chromosome. Two autosomal Mart genes Peg10 (Mart2) and Peg11/Rtl1 (Mart1) undergo genome imprinting and are expressed from the paternal allele [26].

Among the 11 murine Mart genes, 8 are expressed in placenta [27]. At least three genes are required for the placenta development. Mice knocked-out in the Peg10/Mart2 gene, have serious placental defects and die at the embryonic development stage [28]. A deletion in the Mart7 genes disrupts the placental cells differentiation in mice [29]. In humans, overexpression of the Peg11/Mart1, caused by paternal disomy, leads to the development of an overly big placenta [30]. Maternal disomy of the 12 chromosome leads to the placental hypoplasia [31].

Both Peg10/Mart2 and Peg11/Mart1 are expressed not only in placenta, but also in the other embryonic tissues, as well as in the cells of the grown organism, including the brain [25]. Some of the genes codes for the ancestral proteins that contain the zinc finger-like motif, which points to the likely interaction of such proteins with DNA. The Peg10/Mart2 protein, most likely, is a transcription factor and regulates the transcription of the gene that codes the main myelin protein [32]. The destruction of the Mart4 gene causes the disruption of the cognition, memory and behavior processes in mice [33].

The next mammalian gene family, Ma/Pnma (for paraneoplastic Ma antigens) also originated from the Gypsy/Ty3 retrotransposons and consists of the 15 genes in the human genome. The murine genome contains 12 representatives if this family [19]. Most of them, as well as the Mart family genes, are localized in the X chromosome. The Ma/Pnma family genes influence the cellular proliferation, apoptosis and cancer development.

The SCAN domain of the protein is similar to the C-end domain of the capsid protein of the human immune deficiency virus [34], as well as the Gag protein of the Gmr1-like LTR retrotransposons that belong to the Ty3/gypsy retrotransposon family [35]. The protein family that contains the SCAN domain originated from the capsid protein of the Ty3/gypsy retrotransposons of the vertebrates.

The SCAN family is formed by transcription factors together with genes that code for proteins that contain zinc finger (ZF) amino acid motif [36]. In the mammalian genome, the SCAN domain consists of the 84 amino acid residues and is located on the N end of the ZF proteins. Genes that code for the SCAN proteins usually form small clusters that consist of 2-7 genes.

SCAN domains can often be found near the KRAB domain that contains the so-called Krüppel-associated box. The proteins that compose the KRAB domain suppress the expression of the exogenous and endogenous retroviruses. Usually genes that code for the KRAB domain also code for the ZF domains that recognize the DNA sequences. Therefore, the ZF domains recognize DNA domains specifically, including the integrated retroviruses, and KRAB domains promote the attachment of the protein complexes that epigenetically modify the chromatin, that in turn lead to the suppression of the locus. SCAN-KRAB-ZF genes suppress transcription [37].

Usually, SCAN, KRAB and ZF domains are located in a described above order and usually are coded by separate exons. But there are multiple SCAN-ZF genes that lack a KRAB domain.

According to the modern data, there is the following scenario of the emergence of the SCAN domain in the vertebrate's genome [35]. In course of the emergence of higher vertebrates (reptiles, birds, mammals) the gene coding capsid sequences of the Gmr1-l-like retroelement inserts upstream the gene coding two domains: KRAB domain and domain for the transcription factor containing the ZF motif.

After inserted sequence transcription splicing has allowed for a host ESCAN domain formation that transformed into a SCAN domain after N domain deletion. It was followed by amplification of the new gene. The genome of the *Anolis* lizard contains a lot of ZF genes with the ESCAN or SCAN domain [35].

50 human genes coding ZF proteins contain SCAN domain. The SCAN domain participated in the protein-protein interactions as it can form either multimers composed of homodimers or selectively interact with other proteins, forming multiheterodimers.

The ability of the SCAN domain to participate in the protein-protein interactions was transferred from the

DOI: 10.5281/zenodo.2547443

precursor gene that coded for the capsid protein that has multimerized in vivo and formed the core structure of the retroelement capsid. Inside the viral particle the capsid protein interacts with the nucleocapsid protein that contains the zinc finger motif and interacts with RNA.

The merging of the SCAN gene with the gene of the KRAB-ZF family could lead to the formation of a gene coding protein that can interact with the retrovirus due to the interaction between SCAN domain and capsid proteins, as well as the nucleocapsid proteins. This way, the organisms that contained a new gene obtained the ability to control the life cycle of the virus before its insertion in the genome by the cellular factors.

Indeed, the proteins that contain KRAB can repress the transcription of the retroviral genes that are located in the episome [38]. Moreover, the KRAB domain inhibits the integration of the HIV-1 virus into the genome [39]. Consequently, the proteins containing the KRAB domain can attach themselves to the viruses before their insertion in the cellular genome.

Possibly the mergence of the SCAN domain with the KRAB-ZF gene has allowed the proteins coded by the KRAB-ZF gene to attach to the Gmr1-like retrotransposons and possibly to other retroviruses and retrotransposons before their integration into the genome. But this means that the integration of the viruses into the genome can be controlled by the transcription factors of the cell.

One of the predictions of this hypothesis is the ability of the SCAN-ZF proteins to bond with the LTR retro element of the cell genome, especially with those that contain the Gmr1-like sequences. Indeed, at least 5 proteins that are coded by the SCAN-ZF genes of the *Anolis* lizard bond with different sites of the Gmr1-like retro elements of the genome**.** This way, the SCAN domain can play a role in the transcription silencing of the Gmr1-like retrotransposines in the *Anolis* lizard genome.

But multiple SCAN-ZF proteins of the Anolis lizard do not bond with the Gmr1-like retrotransposons; moreover, SCAN-ZF proteins can not interact with the Gmr1-like retro elements in the mammalian genomes because the latter are lost by the genomes in the early stages of evolution of mammals. It is possible that the SCAN domain interacts with the RNA molecules or the retro copies of other retroviruses inserted into the genome of the cell. But it has not been shown yet.

The protein can interact with the DNA of the cell with the help of the ZF domain. That is why the proteins that contain SCAN and ZF domain are functioning as transcription factors and bond with DNA in the genomes of amniotes. The transcription factors that contain SCAN and ZF domain participate in hemopoesis (MZF1), regulate the pluripotency of embryonic stem cells (ZNF20688), control the lipid metabolism (ZNF202), regulate the biosynthesis of cholesterol on the hippocampus (NRIF), chondrogenesis (ZNF449), as well as the behavior of the muscle stem cells, body temperature, maternal behavior and fat deposition (PW1/Peg3) [36].

At least three human genes code for the SCAN domain, but do not contain the tandem located sequences that code for ZF domains. It is interesting that two of the three genes contain N-terminal SCAN exon and C-terminal exon descended from the transposase of DNA transposons (Charlie in SCAND3 and PiggyBac in PGDB1) [35]. Moreover, the SCAND3 gene also possesses an additional exon descended from the retroviral integrase gene. The fact that both genes are conservative among placental mammals indicates the functionality of the coded proteins.

Based on the fact that two proteins gain an ability to specifically bond with DNA sequences upon addition of the SCAN domain, and the same domain makes the proteins transposase and integrase to be able to form double DNA strand breaks, one can make an intriguing hypothesis – these proteins could facilitate locus-specific rebuilding of the genome.

The gene that codes for the activity-regulated cytoskeleton-associated protein (Arc) is expressed in neurons only. The mammalian Arc gene descended from the gag gene of the Gypsy-26-I_DR retrotransposon at the point of divergence between mammals and amphibians [41]. The genes coding for Arc in flies and terrestrial mammals have emerged independently, but both have descended from the Gypsy-26-I_DR retrotransposon.

Arc gene belongs to the immediate-early genes and is a key mediator of the synaptic plasticity. The neuronal activity regulates the transcription and translation of the Arc gene.

The Arc protein participates in the destruction of the unnecessary dendrites in the lesser used synapses between neurons and in the decrease of the glutamate AMPA-receptors quantity on the nerve cell membranes. The Arc protein is absolutely irreplaceable component of the long-term memory and learning. The activity of the Arc gene in the mammalian nerve cells is necessary for the long-term information storage. The disruption on the gene expression is observed in a number of neurological disorders.

Arc proteins interact with each other and form virus-like particles that contain the mRNA of Arc gene [41]. After formation, the viral-like particles are encloses into membrane vesicles that are secreted from the neuron and fuse with other neurons. The more active the neuron is, the more membrane vesicles that contain Arc capsid are produced by the nerve cell. In the post-synaptic neurons the Arc mRNA molecules are translated.

It is interesting that in absence of mRNA no capsids from Arc molecules are formed. Experiments in bacteria have shown that the formation of the capsid does not require specific, Arc-coding mRNA. Moreover, it is not even necessary that it should be a ribonucleic acid. Capsids can form around other mRNA molecules or even single-strand DNA.

A question arises – why do capsids contain the Arc mRNA specifically? It can be supposed that it happens due to the high concentration of Arc mRNA in the area of the capsid formation, which increases the possibility of it being included into the capsid. But this is a hypothesis only. It may be proposed that there is a specific interaction between the capsid proteins and the gene mRNA, which makes the inclusion of Arc mRNA preferable.

Therefore, the descendants of retroviruses form virus-like particles in the neurons and participate in the memory formation in mammals. It is possible that Arc protein is a representative of a certain protein class that

DOI: 10.5281/zenodo.2547443

facilitates the target transfer of specific RNA from cell to cell.

Indeed, the virus-like particles are formed not only in the brain. Gag protein of the endogenous retrovirus HERVK is expressed in the human blastocysts pluripotent cells [42]. Gag protein forms virus-like particles in the blastocysts.

One more gene - Sirh11, descended from gag, is conservative in placental mammals and is expressed at a high level in the brain [43]. The deletion of the murine Sirh11 gene leads to changes in their behaviour which is possibly caused by the decrease in the level of extracellular noradrenaline in the prefrontal cortex [43]. In this way Sirh11 plays a role in the signal transfer, similar to Arc.

The genes of the Ens-1/Erni family descended from retroviral gag sequences are expressed specifically in the embryonic stem cells and in course of the early embryogenesis in chickens [43]. Proteins coded by the Ens-1/Erni genes participate in the formation of the neural plate in chicken in course of embryonic development through the control of Sox2 transcription factor expression [44].

Retroviral gene Pol codes for a protein that performs the reverse synthesis on the DNA on the RNA template. This protein also has RNAase activity that degrades RNA in the RNA/DNA heteroduplex, as well as integrase activity. The polypeptide that is synthesized on the Pol gene also contains an aspartylprotease domain that cleaves the Pol precursor polypeptide with formation of the mature proteins.

In Homo sapiens, two genes, DDI1 and NIX1 code for products similar to aspartylproteases of LTR retroelements [45]. Both proteins are related to the yeast protein Ddi1p and have homologs in α- and γ-proteobacteria [45]. In mice, NIX1 is expressed only in specific neurons of central nerve system. NIX1 connects the ligand-activated and constitutively active nuclear receptors and suppress the gene transcription [46].

In the vertebral genomes only two genes related to the retroviral integrase were found: Gin-1 and Gin-2 [47, 48]. Gin-2 gene was found in bony fishes, amphibians, birds and reptiles. This gene was lost on placental mammals, however. Practically nothing is known about these genes functions besides the fact that the pattern of expression on the zebrafish embryos allows to suppose that this protein is required for gastrulation [47]. The sources of those genes are not retroviruses, but the DNA transposon Gin. The Gin transposons have recruited the LTR retroviruses integrase and used it as a transposase [48].

Retroviruses integrated into the genome can be expressed, and the genes coded by them can be used by the cell for its own purposes. Hypomethylation of DNA and the transcription factor OCT4 activates the expression of the HERVK endogenous retroviruses on the early stages of embryonic development [42]. HERVK is being transcribed starting from the 8 cell stage of embryogenesis and the expressions stops at the stage of embryonic stem cells development. Both HERVK transcription and translation of open reading frames of the virus are activated, as well as the formation of viral particles consisting of Gag proteins.

Proteins synthesized on HERVK influence a number of cellular processes. Rec protein that is coded by one of the HERVK genes binds directly with cellular RNA. No specific RNA sequence for interaction with Rec protein was found. Most likely, the region of the RNA-protein interaction has a specific secondary structure. The interaction of Rec with LTR element of HERVK virus is based on the same principle – it is enabled by the secondary structure of RNA.

In the embryonic cells, Rec interacts with approximately 1600 mRNA molecules [42]. The Rec attachment to viral RNA leads to the exit RNA from the nucleus and subsequent translation. In a similar way cellular mRNA molecules bonded to the Rec interact with ribosomes more effectively. So Rec modulates the expression of a number of genes on the early stages of embryonic development.

It is supposed that the fine-tuning of the cellular programs carried out by the viral proteins determines the species specific and even individual characteristics of the early embryonic development. This happens because genomes of humans and primates have different HERVK retroviruses [49]. Moreover, human genomes also differ in the HERVK dissemination and content. In other words, HERVK elements not only form genetic polymorphism between primates, but are also responsible for polymorphism inside the species [49, 50].

Nonretroviral gene co-optation

Host genomes can co-opt not only retroviral genes, but also genes of other DNA and RNA viruses. In the case when viruses integrated into the germ cells genomes, viral genes get inherited and can be fixed in population.

In the animal genomes endogenous viral elements (EVE) were found that originated from such DNA-containing viruses as hepadnaviruses and herpesvirus 6 (HHV-6) [51, 52]. EVE herpesvirus element was found in human genome [42].

In genomes of different eukaryotes, including animals, sequences were found that originated from genes coding for the Rep protein that initiates replication in heminiviruses, nanoviruses, and circoviruses [53]. Some of the sequences are conservative and are expressed. Rep proteins have domains that bond to DNA, as well as endonuclease and nucleoside triphosphatase domains that are necessary for viral replication. Endogenous elements that code for Rep-like proteins can catalyze their own one-strand excision and insertion into the new genomic positions, behaving like mobile elements that able to reproduce in the genome in the same fashion as helitrones [54]. Most likely, helitrons that are descended from the one-strand DNA viruses and contain Rep proteins are related to the one-strand DNA viruses' proteins [53].

Genomes of different mammals, including elephants, have EVE elements that descended from the adeno associated DNA virus and contain full open reading frame of rep gene.

The rep gene of adeno-associated viruses codes for the proteins Rep78, Rep68, Rep52 and Rep40. Big Rep proteins Rep78 and Rep68 bond to the DNA specifically and have an endonuclease and helicase activity. They control replication, integration and transcription of the virus. Small Rep proteins have a nucleoside triphosphatase

DOI: 10.5281/zenodo.2547443

and helicase activity and are necessary for the virus packaging.

Rep proteins control and direct the viral life cycle. All viral promoters are unter the control of Rep proteins. Besides that, Rep proteins control promoters of the cellular genome both by bonding specifically to the cis-elements of DNA and through bonding to other proteins [55].

In South American rodent degu and elephants mRNA of the adenovirus rep gene is specifically expressed in liver [56]. Cooptation of the adenovirus gene and the specific expression in liver emerged in those two species independently [56]. Most likely, Rep proteins, as they are transcription factors, re-formatted the regulatory programs in the liver cells in these animals.

Hepatitis C virus is an RNA virus and belongs to the flavivirus family. There are EVE elements that have descended from the hepatitis C virus ancestors are present in the animal genoms, but no sequences related to the human HCV virus were found in the human genome. But the DNA, complementary to the RNA matrix of the virus (kDNA) was found in the infected patients [57]. EVE elements of hepatitis C virus were found in the genomes of rabbits and hares, which points to the possibility of integration of viral kDNA into the genome with the help of reverse transcriptase of the retrotransposones [58].

Endogenous bornavirus elements (EBL) emerged as a result of insertion of the bornavirus RNA, most likely mRNA, into the host cell genome. Most likely it happened with the help of the retrotransposase LINE1 of mobile elements [59, 60]. Integration of bornaviruses into the genomes of many species took place independently. EBL elements were found both in vertebrates and invertebrates [59, 60]. There were seven EBL elements found in the animal genome that code for the nucleoprotein N, matrix protein M, glycoprotein G and RNA-dependent RNA polymerase L. The corresponding EBL elements are labeled as EBLN, EBLM, EBLG и EBLL [61].

Endogenous bornavirus-like element L (EBLL) was found in the genomes of a number of eukaryotes. EBLL gene descended from the L gene of bornavirus and, respectively, codes for the RNA-dependent RNA polymerase (RdRp). In the genome of bats of the *Eptesicus spp*., the eEBLL-1 gene is actively transcribed. But it is not proven that the gene codes for functionally active RNA-dependent RNA polymerase [62].

EBL elements at present are single elements found in the human genomes that are descended from non – retroviral RNA viruses [52]. Some EBL elements contain open reading frame, which signifies the possibility of coding for proteins.

Endogenous bornavirus-like nucleoprotein element (EBLN) inserted into the human ancestor genome around 83.3 million years ago. At present, 7 EBLN elements were found in the human genome [63]. Based on its nucleotide sequence, the human EBLN1 gene coincides in 58% with N gene of bornavirus that codes for a nucleoprotein and codes for a protein that consists of 366 amino acids. Whether the protein is expressed is not yet known. But it is known for certain that this gene is translated into an mRNA [64]. At the same time in *Afroteria*, a special group of placental mammals evolved in Africa 83.3 million years ago, EBLN codes for a protein that can bond with the granular endoplasmic reticulum [65].

In the human body, EBLN1 is necessary for the prevention of endogenous DNA damage accumulation and for reparation of the exogenously induced DNA breaks [64]. The disruption of the cell cycle in the cells EBLN1-knocked out cells, possibly, is the consequence of the accumulation of the DNA damage, microtubules organization defects and the centrosomes disengagement.

Suppression of expression of EBLN1 in human oligodendroglia at 80% causes an G2/M arrest of the cell cycle and promotes the transition of cells into apoptosis [66]. The study of the gene expression profile has shown that the expression of 1067 genes is increased and expression of 2004 is decreased as a result of EBLN1 silencing in the oligodendrocytes.

Conclusion

In conclusion, it should be noted that viruses became a source of genes that code for a number of proteins that perform significant functions in the cell. Co-optation of the retroviral genes played a key role in course of evolution, especially in emergence of some aromorphoses. As was mentioned earlier, the genes of syncytin and Mart family are necessary for placenta formation. Arc gene is crucial for formation of memory.

In the process of evolution, viruses formed genes of transcription factors that control the activity of regulatory networks. In this way, viruses could format and re-format the regulatory networks not only by changing the cis-regulatory elements but also by participating in evolution of trans-regulatory proteins.

Besides formation of significant evolutional innovations, co-optation of the retroviral genes is the source of the evolutionary changes in different taxonomic groups and promoted intra-species diversification in vertebrates [49, 50].

Therefore, viruses could be regarded not only as parasitic selfish elements of the genome, but also as an instrument with the help of which the cell genomes acquire new genes.

References

1. Roossinck M. J. Move over bacteria! Viruses make their mark as mutualistic microbial symbionts // J. Virol. 2015. Vol. 89. P. 6532–6535.

2. Belyi V. A., Levine A. J., Skalka A. M. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes // PLoS Pathog. 2010. Vol. 6. P. e1001030.

3. Belyi V. A., Levine A. J., Skalka A. M. Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: The Parvoviridae and Circoviridae are more than 40 to 50 million years old // J. Virol. 2010. Vol. 84. P. 12458–12462.

4. Horie M., Honda T., Suzuki Y., Kobayashi Y., Daito T., Oshida T. [et al.] Endogenous non-retroviral RNA virus elements in mammalian genomes // Nature. 2010. Vol. 463. P. 84–87.

5. Gilbert C., Meik J. M., Dashevsky D., Card D. C., Castoe T. A., Schaack S. Endogenous hepadnaviruses,

DOI: 10.5281/zenodo.2547443

bornaviruses and circoviruses in snakes // Proc. Biol. Sci. 2014. Vol. 281. P. 20141122.
6. Taylor D. J., Leach R. W., Bruenn J. Filoviruses are ancient and integrated into mammalian genomes // BMC Evol. Biol. 2010. Vol. 10. P.193.
7. Feschotte C., Gilbert C. Endogenous viruses: Insights into viral evolution and impact on host biology // Nat. Rev. Genet. 2012. Vol. 13. P. 283–296.
8. Geuking M. B., Weber J., Dewannieux M., Gorelik E., Heidmann T., Hengartner H., Zinkernagel R. M., Hangartner L. Recombination of retrotransposon and exogenous RNA virus results in nonretroviralc DNA integration // Science. 2009. Vol. 323. P. 393–396.
9. Bill C. A., Summers J. Genomic DNA double-strand breaks are targets for hepadnaviral DNA integration // Proc. Natl. Acad. Sci. USA. 2004. Vol. 101. P. 11135–11140.
10. Shapiro J.A. Living Organisms Author Their Read-Write Genomes in Evolution // Biology (Basel). 2017. Vol.6(4). P.pii: E42
11. Mi S., Lee X,, Li X,, Veldman G.M., Finnerty H., Racie L., et al. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis // Nature. 2000. Vol.403. P.785–789.
12. Mangeney M., Renard M., Schlecht-Louf G., Bouallaga I., Heidmann O., Letzelter C., et al. Placental syncytins: genetic disjunction between the fusogenic and immunosuppressive activity of retroviral envelope proteins // Proc. Natl. Acad. Sci. USA. 2007. 104. P.20534–20539.
13. Søe K., Andersen T.L., Hobolt-Pedersen A.S., Bjerregaard B., Larsson L.I., Delaissé J.M. Involvement of human endogenous retroviral syncytin-1 in human osteoclast fusion // Bone. 2011. Vol.48. P.837–846.
14. Redelsperger F., Raddi N., Bacquin A., Vernochet C., Mariot V., Gache V., Blanchard-Gutton N., Charrin S., Tiret L., Dumonceaux J., Dupressoir A., Heidmann T. Genetic Evidence That Captured Retroviral Envelope syncytins Contribute to Myoblast Fusionand Muscle Sexual Dimorphism in Mice // PLoS Genet. 2016. Vol.12(9). e1006289.
15. Blaise S., de Parseval N., Bénit L., Heidmann T. Genome wide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. //Proc. Natl. Acad. Sci. USA. 2003. Vol.100. P.13013–13018.
16. Blond J.L., Lavillette D., Cheynet V., Bouton O., Oriol G., Chapel-Fernandes S., et al. An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor //J. Virol. 2000. Vol.74. P.3321–3329.
17. Esnault C., Priet S., Ribet D., Vernochet C., Bruls T., Lavialle C., et al. A placenta-specific receptor for the fusogenic, endogenous retrovirus-derived, human syncytin-2// Proc. Natl. Acad. Sci. USA. 2008. Vol.105. P.17532–17537.
18. Sugimoto J., Sugimoto M., Bernstein H., Jinno Y., Schust D. A novel human endogenous retroviral protein inhibits cell-cell fusion// Sci. Rep. 2013. Vol.3. P.1462.
19. Naville M., Warren I.A., Haftek-Terreau Z., Chalopin D., Brunet F., Levin P, Galiana D., Volff J.N. Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates // Clin Microbiol Infect. 2016. Vol.4. P.312-323.
20. Cornelis G., Vernochet C., Carradec Q., Souquere S., Mulot B., Catzeflis F., et al. Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials // Proc. Natl. Acad. Sci. USA. 2015. Vol.112. E487–96.
21. de Parseval N, Lazar V, Casella JF, Benit L, Heidmann T (2003) Survey of human genes of retroviral origin: Identification and transcriptome of the genes with coding capacity for complete envelope proteins// J Virol. Vol.77. P.10414–10422
22. Heidmann O., Béguin A., Paternina J., Berthier R., Deloger M., Bawa O., Heidmann T. HEMO, an ancestral endogenous retroviral envelope protein shed in the blood of pregnant women and expressed in pluripotent stem cells and tumors // Proc. Natl. Acad. Sci. USA. 2017. Vol.114(32). E6642-E6651.
23. Malik H.S., Henikoff S., Eickbush T.H. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses // Genome Res. 2000. Vol.10. P.1307–1325.
24. Ribet D., Harper F., Dupressoir A., Dewannieux M., Pierron G., Heidmann T. An infectious progenitor for the murine IAP retrotransposon: emergence of an intracellular genetic parasite from an ancient retrovirus // Genome Res. 2008. Vol.18. P.597–609.
25. Brandt J., Schrauth S., Veith A.M., Froschauer A., Haneke T., Schultheis C., et al. Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon derived neogenes in mammals // Gene. 2005; Vol.345. P. 101–111
26. Ono R., Kobayashi S., Wagatsuma H., Aisaka K., Kohda T., Kaneko-Ishino T., et al. A retrotransposon-derived gene, PEG10, is a novel imprinted gene located on human chromosome 7q21 // Genomics. 2001. Vol.73. P.232–237.
27. Henke C., Strissel P.L., Schubert M.T., Mitchell M., Stolt C.C., Faschingbauer F., et al. Selective expression of sense and antisense transcripts of the sushi-ichi-related retrotransposon-derived family during mouse placentogenesis // Retrovirology. 2015. Vol.12. P.9
28. Henke C., Ruebner M., Faschingbauer F., Stolt C.C., Schaefer N., Lang N., et al. Regulation of murine placentogenesis by the retroviral genes Syncytin-A, Syncytin-B and Peg10 // Differentiation. 2013. Vol.85. P.150–160.
29. Naruse M., Ono R., Irie M., Nakamura K., Furuse T., Hino T., et al. Sirh7/Ldoc1 knockout mice exhibit placental P4 overproduction and delayed parturition // Development. 2014. Vol.141. P.4763–4771
30. Kagami M., Yamazawa K., Matsubara K., Matsuo N., Ogata T. Placentomegaly in paternal uniparental disomy for human chromosome 14 // Placenta. 2008. Vol.29. P.760–781.
31. Georgiades P., Watkins M., Surani M.A., Ferguson-Smith A.C. Parental origin-specific developmental defects in mice with uniparental disomy for chromosome 12 // Development. 2000. Vol.127. P.4719–4728.
32. Steplewski A., Krynska B., Tretiakova A., Haas S., Khalili K., Amini S. MyEF-3, a developmentally controlled brain-derived nuclear protein which specifically

DOI: 10.5281/zenodo.2547443

interacts with myelin basic protein proximal regulatory sequences // Biochem. Biophys. Res. Commun. 1998. Vol.243. P.295–301.
33. Irie M., Yoshikawa M., Ono R., Iwafune H., Furuse T., Yamada I, et al. Cognitive function related to the Sirh11/Zcchc16 gene acquired from an LTR Retrotransposon in eutherians // PLoS Genet. 2015. Vol.11. e1005521.
34. Ivanov D., Stone J.R., Maki J.L., Collins T., Wagner G. Mammalian SCAN domain dimer is a domain-swapped homolog of the HIV capsid C-terminal domain // Mol. Cell. 2005. Vol.17. P.137–143.
35. Emerson R.O., Thomas J.H. Gypsy and the birth of the SCAN domain // J. Virol. 2011. Vol.85. P.12043-12052.
36. Edelstein L.C., Collins T. The SCAN domain family of zinc finger transcription factors // Gene. 2005. Vol.359. P.1–17
37. Itokawa Y., Yanagawa T., Yamakawa H., Watanabe N., Koga H., Nagase T. KAP1-independent transcriptional repression of SCAN-KRAB-containing zinc finger proteins // Biochem. Biophys. Res. Commun. 2009. Vol.388. P.689–694
38. Barde I., Laurenti E., Verp S., Groner A.C., Towne C., Padrun V., Aebischer P., Trumpp A., Trono D.. Regulation of episomal gene expression by KRAB/ KAP1-mediated histone modifications // J. Virol. 2009. Vol.83. P.5574–5580.
39. Allouch A., Di Primio C., Alpi E., Lusic M., Arosio D., Giacca M., Cereseto A. The TRIM family protein KAP1 inhibits HIV-1 integration // Cell Host Microbe. 2011. Vol.9. P.484–495.
40. Kaneko-Ishino T., Ishino F. The role of genes domesticated from LTR retrotransposons and retroviruses in mammals // Front. Microbiol. 2012. Vol.3. P.262.
41. Pastuzyn ED, Day CE, Kearns RB, Kyrke-Smith M, Taibi AV, McCormick J, Yoder N, Belnap DM, Erlendsson S, Morado DR, Briggs JAG, Feschotte C, Shepherd JD. The Neuronal Gene Arc Encodes a Repurposed Retrotransposon Gag Protein that Mediates Intercellular RNA Transfer // Cell. 2018. Vol.172(1-2). P.275-288.
42. Grow E.J., Flynn R.A., Chavez S.L., Bayless N.L., Wossidlo M., Wesche D.J., Martin L., Ware C.B., Blish C.A., Chang H.Y., Pera R.A., Wysocka J. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells // Nature. 2015. Vol.522. P.221-225.
43. Lerat E., Birot A.M., Samarut J., Mey A. Maintenance in the chicken genome of the retroviral-like cENS gene family specifically expressed in early embryos // J. Mol. Evol. 2007. Vol.65. P.215–227.
44. Papanayotou C., Mey A., Birot A.M., Saka Y., Boast S., Smith J.C., et al. A mechanism regulating the onset of Sox2 expression in the embryonic neural plate // PLoS Biol. 2008. Vol.6. e2.
45. Krylov D.M., Koonin E.V. A novel family of predicted retroviral-like aspartyl proteases with a possible key role in eukaryotic cell cycle control // Curr. Biol. 2001. Vol.11. R584–587.
46. Greiner E.F., Kirfel J., Greschik H., Huang D., Becker P., Kapfhammer J.P., et al. Differential ligand-dependent protein–protein interactions between nuclear receptors and a neuronal-specific cofactor // Proc. Natl. Acad. Sci. USA. 2000. Vol.97. P.7160–7165.
47. Chalopin D., Galiana D., Volff J.N. Genetic innovation in vertebrates: gypsy integrase genes and other genes derived from transposable elements // Int. J. Evol. Biol. 2012. Vol.2012. P.724519.
48. Marín I. GIN. transposons: genetic elements linking retrotransposons and genes // Mol. Biol. Evol. 2010. Vol.27. :P.903–911.
49. Shin W., Lee J., Son S.Y., Ahn K., Kim H.S., Han K. Human-Specific HERV-K Insertion Causes Genomic Variations in the Human Genome // PLoS ONE. 2013. Vol.8. e60605.
50. Böhne A., Brunet F., Galiana-Arnoux D., Schultheis C., Volff JN. Transposable elements as drivers of genomic and biological diversity in vertebrates // Chromosome Res. 2008. Vol.16. P.203–215.
51. Gravel A., Dubuc I., Morissette G., Sedlak R. H., Jerome K. R., Flamand L. Inherited chromosomally integrated human herpesvirus 6 as a predisposing risk factor for the development of angina pectoris // Proc. Natl. Acad. Sci. U.S.A. 2015. Vol.112. P.8058–8063.
52. Shen Z., Liu Y., Luo M., Wang W., Liu J., Liu W., et al. Nuclear factor Y regulates ancient budgerigar hepadnavirus core promoter activity // Biochem. Biophys. Res. Commun. 2016. Vol.478. P.825–830.
53. Liu H., Fu Y., Li B., Yu X., Xie J., Cheng J., et al. Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes // BMC Evol. Biol. 2011. Vol.11. P.276
54. Kapitonov V.V., Jurka J: Rolling-circle transposons in eukaryotes // Proc. Natl. Acad. Sci. USA. 2001. Vol.98. P.8714-8719.
55. Dutheil N., Smith S.C., Agúndez L., Vincent-Mistiaen Z.I., Escalante C.R., Linden R.M., Henckaerts E. Adeno-associated virus Rep represses the human integration site promoter by two pathways that are similar to those required for the regulation of the viral p5 promoter // J. Virol. 2014. Vol.88. p.8227-8241.
56. Kobayashi Y., Shimazu T., Murata K., Itou T., Suzuki Y. An endogenous adeno-associated virus element in elephants // Virus Res. 2018. pii: S0168-1702(18). P.30160-30166.
57. Zemer R., Kitay Cohen Y., Naftaly T., Klein A. Presence of hepatitis C virus DNA sequences in the DNA of infected patients // Eur. J. Clin. Invest. 2008. Vol.38. P.845–848.
58. Silva, E., Marques, S., Osório, H., Carvalheira, J., Thompson, G. Endogenous hepatitis C virus homolog fragments in European rabbit and hare genomes replicate in cell culture // Plos One. 2012. Vol.7. e49820.
59. Horie M., Honda T., Suzuki Y., Kobayashi Y., Daito T., Oshida T., et al. Endogenous non-retroviral RNA virus elements in mammalian genomes // Nature. 2010. Vol.463. P.84-87.
60. Belyi V.A., Levine A.J., Skalka A.M. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes // PLoS Pathog. 2010. Vol.6. e1001030.
61. Katzourakis A., Gifford R.J. Endogenous viral elements in animal genomes // PLoS Genet. 2010. Vol.6: e1001191
62. Horie M., Kobayashi Y., Honda T., Fujino K., Akasaka T., Kohl C., Wibbelt G., Mühldorfer K., Kurth A., Müller

DOI: 10.5281/zenodo.2547443

M.A., Corman V.M., Gillich N., Suzuki Y., Schwemmle M., Tomonaga K. An RNA-dependent RNA polymerase gene in bat genomes derived from an ancient negative-strand RNA virus // Sci. Rep. 2016. Vol.6. P.25873.
63. Sofuku K., Parrish N.F., Honda T., Tomonaga K. Transcription profiling demonstrates epigenetic control of non-retroviral RNA virus-derived elements in the human genome // Cell Rep. 2015. Vol.12. P.1548–1554.
64. Myers K.N., Barone G., Ganesh A., Staples C.J., Howard A.E., Beveridge R.D., Maslen S., Skehel J.M., Collis S.J. The bornavirus-derived human protein EBLN1 promotes efficient cell cycle transit, microtubule organisation and genome stability // Sci. Rep. 2016. Vol.6. P.35548.
65. Kobayashi Y., Horie M., Nakano A., Murata K., Itou T., Suzuki Y. Exaptation of Bornavirus-Like Nucleoprotein Elements in Afrotherians // PLoS Pathog. 2016. Vol.12. e1005785.
66. He P., Sun L., Zhu D., Zhang H., Zhang L., Guo Y.l., Liu S., Zhou J., Xu X., Xie P. Knock-Down of Endogenous Bornavirus-Like Nucleoprotein 1 Inhibits Cell Growth and Induces Apoptosis in Human Oligodendroglia Cells // Sci. Rep. 2016 Vol.6. P.25873.

THE GENERATION OF CODING SEQUENCES OF CELLULAR GENOME THROUGH COOPTION OF VIRAL GENES

Popov N. N., Sklyar N.I., Kolotova T. Yu., Davydenko M. B., Voronkina I. A.

This review attempts to summarize the available data concerning the influence of viruses on the generation of the cellular genome coding genes content.
For a long time endogenous retroviruses have been considered as selfish elements of the organism genome. But now there is growing evidence that endogenous retroviruses are more than genome junk and can serve as source for new coding sequences allowing organism evolution. Many genes derived from retroviruses have been identified in eukaryote through comparative genomics and functional analyses. In particular, genes derived from gag structural protein and envelope (env) genes, as well as from the integrase-coding and protease-coding sequences, have been identified in humans and other vertebrates. It has been proved that a number of these genes fulfill essential functions for the development and survival of their host. One of the best known co-opted retroviral genes encoded syncytin plays a key role in the placenta development. It is interesting that during mammalian evolution retroviral envelope genes have been domesticated several times independently to generate syncytin. The activity-regulated cytoskeletal protein Arc is important for cognitive functions and memory formation. Arc was one of over 100 human proteins that have been ‘‘domesticated’’ from the retrotransposon remains of ancient viruses. A number of genes that code the transcription factors have emerged as a result of “taming” the viral genes by the host organism. Now growing evidence reveals that not only retroviruses but other RNA viruses are reverse-transcribed and integrated into the genome of infected cells. It has been recently demonstrated that all Homo sapiens bornavirus like nucleoproteins (EBLN) are expressed in at least one tissue and consequently may have function.The co-option of the viral sequences not only can lead to the major evolutionary innovations, but also is able to create interspecies polymorphism.What it has been described here is probably only the tip of the iceberg, and future genome analyses will certainly uncover new virus-derived genes.

Keywords: endogenous retroviruses, Ty3/gypsy retrotransposon family, bornaviruses, adeno associated virus, SCAN domain, *arc* gene, syncytin.

DOI: 10.5281/zenodo.2547443