Вавиловский журнал генетики и селекции. 2021;25(1):46-56
DOI 10.18699/VJ21.006

Original Russian text www.bionet.nsc.ru/vogis/

Phylostratigraphic analysis of gene networks of human diseases

Z.S. Mustafin[1] ✉, S.A. Lashin[1, 2], Yu.G. Matushkin[1]

[1] Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
✉ mustafinzs@bionet.nsc.ru

Abstract. Phylostratigraphic analysis is an approach to the study of gene evolution that makes it possible to determine the time of the origin of genes by analyzing their orthologous groups. The age of a gene belonging to an orthologous group is defined as the age of the most recent ancestor of all species represented in that group. Such an analysis can reveal important stages in the evolution of both the organism as a whole and groups of functionally related genes, in particular gene networks. In addition to investigating the time of origin of a gene, the level of its genetic variability and what type of selection the gene is subject to in relation to the most closely related organisms is studied. Using the Orthoscape application, gene networks from the KEGG Pathway, Human Diseases database describing various human diseases were analyzed. It was shown that the majority of genes described in gene networks are under stabilizing selection and a high reliable correlation was found between the time of gene origin and the level of genetic variability: the younger the gene, the higher the level of its variability is. It was also shown that among the gene networks analyzed, the highest proportion of evolutionarily young genes was found in the networks associated with diseases of the immune system (65 %), and the highest proportion of evolutionarily ancient genes was found in the networks responsible for the formation of human dependence on substances that cause addiction to chemical compounds (88 %); gene networks responsible for the development of infectious diseases caused by parasites are significantly enriched for evolutionarily young genes, and gene networks responsible for the development of specific types of cancer are significantly enriched for evolutionarily ancient genes.

Key words: evolution; phylostratigraphic analysis; ortholog; gene network; gene age.

For citation: Mustafin Z.S., Lashin S.A., Matushkin Yu.G. Phylostratigraphic analysis of gene networks of human diseases. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2021;25(1):46-56. DOI 10.18699/VJ21.006

Филостратиграфический анализ генных сетей заболеваний человека

З.С. Мустафин[1] ✉, С.А. Лашин[1, 2], Ю.Г. Матушкин[1]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
✉ mustafinzs@bionet.nsc.ru

Аннотация. Филостратиграфический анализ – это подход к исследованию эволюции генов, позволяющий определить время возникновения генов за счет анализа филогенетических деревьев организмов, обладающих ортологичными к исследуемому генами. Такой анализ может открыть важные этапы в эволюции как организма в целом, так и групп функционально связанных генов, в частности генных сетей. В дополнение к исследованию времени возникновения гена изучается уровень его генетической изменчивости и то, какому типу отбора подвержен ген по отношению к наиболее близкородственным организмам. С помощью приложения Orthoscape были проанализированы генные сети из базы данных KEGG Pathway, Human Diseases, ассоциированные с заболеваниями человека. Выявлено, что большинство генов, описанных в генных сетях, подвержены стабилизирующему отбору, обнаружена высокая достоверная корреляция между временем возникновения гена и уровнем генетической изменчивости, которой он подвержен, – чем моложе ген, тем выше уровень генетической изменчивости. Было также показано, что среди проанализированных генных сетей наибольшая доля эволюционно молодых генов обнаружена в сетях, связанных с заболеваниями иммунной системы (65 %), а эволюционно древних генов – в сетях, ответственных за формирование зависимостей человека от веществ, вызывающих привыкание к химическим соединениям (88 %); генные сети, связанные с развитием инфекционных заболеваний, вызванных паразитами, достоверно обогащены эволюционно молодыми генами, а генные сети, ответственные за развитие специфических типов рака, – эволюционно древними генами.

Ключевые слова: эволюция; филостратиграфия; ортолог; генная сеть; возраст гена.

Introduction

The study of key factors that influence to the development of diseases is one of the most important research areas in both medicine and biology (Stepanov, 2016). It is known that the formation of phenotypic traits that provide the adaptation of organisms to environmental conditions is controlled not by individual genes, but by gene networks – groups of coordinately functioning genes and their products (RNA, proteins, metabolites, etc.) (Kolchanov et al., 2013). The task of highlighting the key structural features of networks, network elements, and their numerical description arises. One of the important characteristics in evolutionary biology is the age of a gene. The age of a gene belonging to an orthologous group is defined as the age of the most recent ancestor of all species represented in this group (Liebeskind et al., 2016).

Modern methods of analysis make it possible to evaluate the evolutionary characteristics of genes, in particular, phylostratigraphic analysis, a methodology proposed in 2007 by T. Domazet-Lošo, which allows to determine the age of a gene using a special index. The index is derived from analysis of orthologous genes and comparison of the position of organisms whose genes are considered in the analysis on a phylogenetic tree (Domazet-Lošo et al., 2007).

There are many software tools to work with gene networks. Some of them focus on reconstructing networks based on data from biological databases: String (Szklarczyk et al., 2019), GeneMANIA (Montojo et al., 2010). The others have extensive functionality for visualizing network elements and identifying its structural features: Cytoscape (Shannon et al., 2003), yEd (https://www.yworks.com/products/yed). Cytoscape has an advantage that in addition to its extensive capabilities of constructing networks, layouting and painting their elements and analyzing structural features, it allows users to write their own applications in Java. It makes possible for community to implement any functionality and plug in to Cytoscape. For example, well-known tools String and GeneMANIA for networks reconstruction from the list of genes based on extracting interactions from biological databases have their own plugins in Cytoscape and allow to use their functionality by combining it with the capabilities of Cytoscape and its other plugins. Also, the plugins allow to import networks from databases, for example, Pathway Commons (Cerami et al., 2011) or KEGG Pathway (Kanehisa et al., 2017), without hard parsing the formats of network representation.

The results of gene network analysis by one of such applications – Orthoscape (Mustafin et al., 2017), are presented in this paper. Orthoscape can analyze the evolutionary features of genes in gene network. It has been shown that most of the genes described in gene networks are under influence of stabilizing selection. A high reliable correlation has been found between the time of occurrence of a gene and the level of its genetic variability – the younger the gene, the higher the level of genetic variability is. Among the gene networks analyzed, the highest proportion of evolutionary young genes was detected in the networks associated with immune diseases (65 %), and the highest proportion of evolutionary ancient genes was detected in the networks responsible for the substance dependencies (88 %); gene networks responsible for the development of infectious diseases caused by parasites are significantly enriched in evolutionary young genes, and gene networks responsible for the development of specific types of cancer are significantly enriched in evolutionary ancient genes.

Materials and methods

Input data. Gene networks from KEGG Pathway, Human Diseases are used in this work. These networks are divided into 11 categories (with total number of 80 networks): neurodegenerative diseases (5 networks), cardiovascular diseases (5 networks), immune diseases (8 networks), endocrine and metabolic diseases (6 networks), infectious diseases: bacterial (10 networks), infectious diseases: viral (9 networks), infectious diseases: parasitic (6 networks), drug resistance: antineoplastic (4 networks), cancers: overview (7 networks), cancers: specific types (15 networks), substance dependence (5 networks).

The data required for the analysis, such as lists of orthologous genes, nucleotide sequences of genes and amino acid sequences of the proteins they encode, protein domains, taxonomic information about organisms whose genes were considered in the analysis were also taken from the KEGG database.

Software used. The analysis was performed using the Cytoscape software package (Shannon et al., 2003). CyKEGG Parser plugin was used to import networks from the KEGG Pathway (Nersisyan et al., 2014). Orthoscape plugin was used to perform phylostratigraphic analysis and analysis of so called divergence index – the index of evolutionary variability (Mustafin et al., 2017).

Methods for estimation the evolutionary characteristics of genes. Orthoscape allows to estimate two evolutionary characteristics of genes. The first one is *phylostratigraphic age index* (PAI). This index shows how far from the root of the phylogenetic tree is the taxon reflecting the age of the gene, i. e., the taxon where the studied species diverged from the most distant related taxon in which the ortholog of the studied gene was found. Thus, the more PAI of the gene, the younger it is (Fig. 1). KEGG Orthology service is used in Orthoscape to calculate PAI, which makes it possible to consider orthologous genes among all homologs.

Figure 1 shows examples of determining the age of a gene and the phylostratigraphic index, using human genes. On the left (*a*) the case when the most distant organism in which the ortholog of the studied gene was found is the bonobo is shown. The node most distant from the root of the phylogenetic tree that is common to *H. sapiens* and *P. paniscus* (bonobos) is Hominidae. It corresponds to the phylostratigraphic index is equal to 13. On the right (*b*) is the gene whose ortholog was found in *M. domestica* (gray short-tailed opossum), the most distant node is Mammalia, and the phylostratigraphic index of the gene is equal to 7. Since the PAI in example (*a*) is larger than the PAI in example (*b*), we

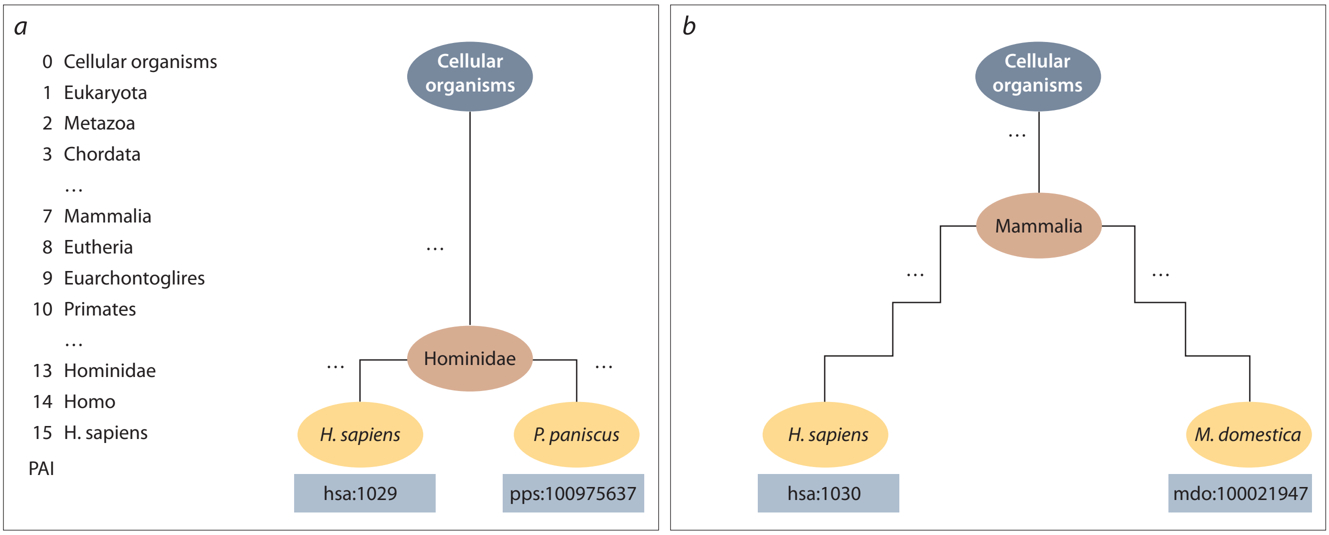


Fig. 1. The example of PAI determination for two *Homo sapiens* (human) genes.

a – the example of evolutionary young gene is hsa:1029, most distant from the studied organism in which the orthologous gene was found *Pan paniscus* (bonobo); *b* – the example of evolutionary older gene is hsa:1030, most distant from the studied organism in which the orthologous gene was found is *Monodelphis domestica* (gray short-tailed opossum). We can conclude that the gene on example (*a*) is evolutionary younger than the gene on example (*b*). The scale on the left shows the PAI index, which corresponds to the depth of a node in the phylogenetic tree (see Table 1 for details).

can conclude that the gene in example (*a*) is evolutionary younger than the gene in example (*b*).

An important characteristic for the phylostratigraphic analysis is the list of taxonomic units describing the stages of divergence on the phylogenetic tree. Table 1 shows the complete list of taxa used in the analysis to determine the phylostratigraphic age index of *H. sapiens* genes, as well as the approximate evolutionary age of these taxa in millions of years from our time. It should be noted that the discussions on this topic are ongoing and there are different data of the age; the values in the table reflect approximate estimates.

Table 1. The list of taxons used in phylostratigraphic analysis of *H. sapiens* genes

PAI	Taxon	Age (Mya)
0	Cellular organism (tree root)	4100 (Bell et al., 2015)
1	Eukaryota	1850 (Leander, 2020)
2	Metazoa	665 (Maloof et al., 2010a)
3	Chordata	541 (Maloof et al., 2010b)
4	*Craniata*	535 (Maloof et al., 2010b)
5	Vertebrata	525 (Shu et al., 1999)
6	Euteleostomi	420 (Diogo, 2007)
7	Mammalia	225 (Datta, 2005)
8	Eutheria	160 (Luo et al., 2011)
9	Euarchontoglires	65 (Kumar et al., 2013)
10	Primates	55 (Chatterjee et al., 2009)
11	Haplorrhini	50 (Dunn et al., 2016)
12	Catarrhini	44 (Harrison, 2013)
13	Hominidae	17 (Hey, 2005)
14	Homo	2.8 (Schrenk et al., 2014)
15	Homo sapiens	0.35 (Scerri et al., 2018)

Orthoscape also allows to estimate *divergence index* (DI). DI shows the type of selection to which the gene is influenced. This index is calculating based on the *dN/dS* ratio, where *dN* reflects the rate of nonsynonymous substitutions between the sequences of analyzed gene and its orthologous gene (the substitutions changing the amino acid encoded) and *dS* – reflects the rate of synonymous substitutions (the substitutions without changing the amino acid encoded). The index value from 0 to 1 indicates that the gene is under stabilizing selection, value is equal to 1 indicates neutral evolution, and greater than 1 indicates a driving selection. The analysis of this index makes sense only when comparing closely related organisms, because it can't take into account multiple substitutions in the same position, which will be inevitably accumulated when comparing the evolutionary distant organisms. Calculation of *dN/dS* takes place in two phases:

1. Alignment of original sequences. To align the sequences, the Needleman–Wunsch algorithm is used. The task is to align aminoacid sequences and nucleotide triplets correspond to them and remove the gaps from the result.
2. Aligned sequences are given as input to the PAML (phylogenetic analysis by maximum likelihood) (Yang, 2007) software. Various methods are used to calculate *dN/dS*. They take into account different positions of triplets, their frequency of occurrence, and other factors. There are Nei–Gojobori (Nei, Gojobori, 1986), Yang & Nielsen (Yang, Nielsen, 2000), LWL85 (Li, 1985), LWLm (Li, 1993), LPB93 (Pamilo, Bianchi, 1993) methods implemented in PAML. To count *DI*, Orthoscape uses LPB93

method. The formula to count *DI* is based on *dN/dS* for every gene-ortholog pair

$$DI = \frac{\sum_{i=1}^{n} dnds_i}{n},$$

where $dnds_i$ – *dN/dS* value for gene and *i*-th ortholog; *n* – number or orthologous genes.

Results and discussion

The analysis of evolutionary characteristics of gene networks

80 networks from KEGG Pathway, Human Diseases were analyzed using Orthoscape software. First of all, PAI and DI values for genes in network have been calculated. Based on these data, PAI values for every network have been calculated (Table 2) as an average PAI value of genes involved in network. Finally, PAI of the category of diseases has been calculated as an average of PAI value of networks from this category. The same metrics have been calculated for DI.

There are big PAI variations are observed among the analyzed 80 networks: from 0.44 (i. e., in "Nicotine addiction" gene network, the most of the genes are evolutionary ancient) to 6.38 (i. e., in "Asthma" gene network, the most of the genes are evolutionary young). The DI variation is usually within $0 < \text{DI} < 1$ interval, i. e., within stabilizing selection interval; however, the level of variability of genes involved in different networks also varies greatly, from 0.16 to 0.64. The diseases "Asthma" and "Nicotine addiction" are the most exuding according to the PAI and DI indices. In the "Asthma" network, evolutionary young and variable genes prevail, and in the "Nicotine addiction" network, evolutionary ancient and conservative genes prevail. Fig. 2 shows the result of PAI analysis for the "Asthma" and "Nicotine addiction" networks, and Fig. 3 shows the DI results of the same networks.

The most part of genes in the "Asthma" network (Fig. 2, *a*) are evolutionary young (colored in green and yellow), with origin on Vertebrata level. On the contrary, in the "Nicotine addiction" network (Fig. 2, *b*) all genes have been identified as evolutionary ancient, with origin from the cellular life form (Cellular organisms) to multicellular (Metazoa) stages.

Analysis of the DI indicates that almost all the genes involved in the "Asthma" network (see Fig. 3, *a*) are more evolutionary variable than the genes involved in the "Nicotine addiction" network (see Fig. 3, *b*), whose genes are very conservative.

Let's take a look at the estimations of PAI values for 11 disease categories (see Table 2). The most segregated networks are from 4 categories. High PAI and DI values is characteristic of Immune diseases (8 networks) and Infectious diseases: Parasitic (6 networks). Low PAI and DI value is characteristic of cancers: specific types (15 networks) and substance dependence (5 networks).

Genes from the categories above, as well as the complete set of 1436 genes, were divided into two groups: 1) a group of evolutionary ancient genes with PAI < 5 (the age of the genes corresponds to the period of evolution from Cellular organisms to Chordata); 2) a group of evolutionary young genes with PAI ≥ 5 (the age of the genes corresponds to the period of evolution from the Craniata to modern humans).

Contingency tables were created and Fisher's exact test was used to assess whether the difference in the partitioning of genes into groups in the category from the partitioning in the full list of genes was significant (Table 3).

The average PAI of all 1436 genes studied was equal to 2.49. The results from Table 3 show that gene networks from category Immune diseases have not only the highest value of the PAI (5.21), but also a significantly different distribution of the proportion of evolutionary young and ancient genes in comparison with such proportion among all genes analyzed (the last row of the Table 3).

The part of evolutionary young genes in Immune diseases category is 65 %. The most part of genes origin was at vertebrata stages (Vertebrata and Euteleostomi taxa), that corresponds to modern data about the development of specific immunity: it exists in cartilaginous fish (sharks and rays) and, therefore, appeared at least 400–500 million years ago. These fishes have genes related to the genes of the *Ig* variable region (*IgV*), or T-cell receptor (*TcR*) genes. At the same time, even more primitive vertebrates, the roundworms (hagfish and lampreys), do not have an acquired immunity system; they have neither *IgV* nor *TkR* genes (Galaktionov, 2015). The analysis also revealed a small fraction of evolutionary ancient genes in the Immune diseases category. This is consistent with the knowledge that some functions of the immune system originated as early as in unicellular organisms, such as the ability to phagocytose; cells with the T-lymphocyte marker first discovered in ringworms and the histocompatibility system – in sponges (Khaitov, 2009). On the contrary, the highest proportion of evolutionary ancient genes is characteristic of the "Substance dependence" diseases category, which includes genes responsible for addiction to chemicals (88 %). Most of the genes considered are involved in nervous system function, including neurotransmitter function.

The infectious diseases parasitic category, which includes genes associated with infectious diseases caused by parasites (53 % of the evolutionary young genes), has a significant difference in the proportion of evolutionary ancient and evolutionary young genes from that among all the genes analyzed. In the case of the infectious diseases parasitic category, the high proportion of evolutionary young genes can be directly related to the high proportion of evolutionary young genes and the high evolutionary variability of genes found in the Immune diseases category. It is infectious diseases that are one of the most important drivers of immune system evolution. At the same time, infectious diseases of different nature and the immune system co-evolve in the process of forming mechanisms to fight each other (Sasaki et al., 2000; Khakoo, 2004; Zheleznikova, 2014).

It should be noted, that there is a significant excess of the proportion of ancient genes over young genes compared to their distribution (ancient/young) in the full sample of

Table 2. Average values of PAI and DI indices for genes involved in gene networks from the KEGG Pathway, Human Diseases database

No.	Network*	PAI	DI
1	Asthma[1]	6.38	0.64
2	Graft-versus-host disease[1]	6.29	0.54
3	Autoimmune thyroid disease[1]	5.61	0.49
4	Allograft rejection[1]	5.53	0.46
5	Malaria[2]	5.49	0.46
6	African trypanosomiasis[2]	5.12	0.47
7	Inflammatory bowel disease (IBD)[1]	4.95	0.35
8	Rheumatoid arthritis[1]	4.70	0.40
9	Staphylococcus aureus infection[3]	4.41	0.53
10	Type I diabetes mellitus[9]	4.40	0.42
11	Primary immunodeficiency[1]	4.24	0.39
12	Systemic lupus erythematosus[1]	3.97	0.42
13	Tuberculosis[3]	3.96	0.34
14	Pertussis[3]	3.87	0.37
15	Legionellosis[3]	3.84	0.34
16	Salmonella infection[3]	3.77	0.26
17	Viral myocarditis[8]	3.66	0.35
18	Leishmaniasis[2]	3.60	0.33
19	Chagas disease (American trypanosomiasis)[2]	3.58	0.29
20	Chemical carcinogenesis[4]	3.56	0.56
21	Measles[5]	3.53	0.30
22	Toxoplasmosis[2]	3.42	0.28
23	Influenza A[5]	3.35	0.35
24	Amoebiasis[2]	3.26	0.36
25	Herpes simplex virus 1 infection[5]	3.26	0.34
26	Kaposi sarcoma-associated herpesvirus infection[5]	3.13	0.29
27	Antifolate resistance[6]	3.00	0.40
28	Hepatitis C[5]	2.92	0.30
29	Platinum drug resistance[6]	2.80	0.29
30	Acute myeloid leukemia[7]	2.80	0.30
31	Arrhythmogenic right ventricular cardiomyopathy[8]	2.79	0.25
32	Amyotrophic lateral sclerosis (ALS)[10]	2.75	0.27
33	Epstein-Barr virus infection[5]	2.54	0.35
34	Transcriptional misregulation in cancer[4]	2.53	0.29
35	AGE-RAGE signaling pathway in diabetic complications[9]	2.52	0.28
36	Hepatitis B[5]	2.50	0.27
37	Non-alcoholic fatty liver disease[9]	2.44	0.27
38	EGFR tyrosine kinase inhibitor resistance[6]	2.43	0.20
39	Alzheimer disease[10]	2.42	0.26
40	Fluid shear stress and atherosclerosis[8]	2.40	0.26
41	Epithelial cell signaling in Helicobacter pylori infection[3]	2.27	0.20
42	Dilated cardiomyopathy (DCM)[8]	2.19	0.26
43	Pathogenic Escherichia coli infection[3]	2.19	0.27
44	Human papillomavirus infection[5]	2.18	0.29
45	Human T-cell leukemia virus 1 infection[5]	2.16	0.29
46	Hypertrophic cardiomyopathy (HCM)[8]	2.14	0.30
47	Bladder cancer[7]	2.13	0.26
48	Pancreatic cancer[7]	2.10	0.20
49	Proteoglycans in cancer[4]	2.06	0.25
50	Prion diseases[10]	2.05	0.29
51	Viral carcinogenesis[4]	1.94	0.24
52	Non-small cell lung cancer[7]	1.93	0.25
53	Pathways in cancer[4]	1.86	0.24
54	Small cell lung cancer[7]	1.84	0.26
55	Chronic myeloid leukemia[7]	1.82	0.21
56	Shigellosis[3]	1.81	0.27
57	Parkinson disease[10]	1.76	0.20
58	Glioma[7]	1.74	0.25
59	Endometrial cancer[7]	1.72	0.24
60	Melanoma[7]	1.71	0.24
61	Colorectal cancer[7]	1.65	0.21
62	Insulin resistance[9]	1.64	0.25
63	Endocrine resistance[6]	1.62	0.22
64	Central carbon metabolism in cancer[4]	1.61	0.26
65	Thyroid cancer[7]	1.57	0.24
66	Breast cancer[7]	1.55	0.30
67	Alcoholism[11]	1.48	0.17
68	Cocaine addiction[11]	1.42	0.14
69	Bacterial invasion of epithelial cells[3]	1.42	0.15
70	Huntington disease[10]	1.42	0.20
71	Renal cell carcinoma[7]	1.41	0.16
72	Vibrio cholerae infection[3]	1.35	0.18
73	Prostate cancer[7]	1.33	0.29
74	Type II diabetes mellitus[9]	1.30	0.29
75	Basal cell carcinoma[7]	1.20	0.23
76	Morphine addiction[11]	1.06	0.16
77	Maturity onset diabetes of the young[9]	1.04	0.19
78	Choline metabolism in cancer[4]	1.03	0.19
79	Amphetamine addiction[11]	0.75	0.18
80	Nicotine addiction[11]	0.44	0.16

* Category 1 – immune diseases; 2 – infectious diseases parasitic; 3 – infectious diseases bacterial; 4 – cancers overview; 5 – infectious diseases viral; 6 – drug resistance antineoplastic; 7 – cancers specific types; 8 – cardiovascular diseases; 9 – endocrine and metabolic diseases; 10 – neurodegenerative diseases; 11 – substance dependence.

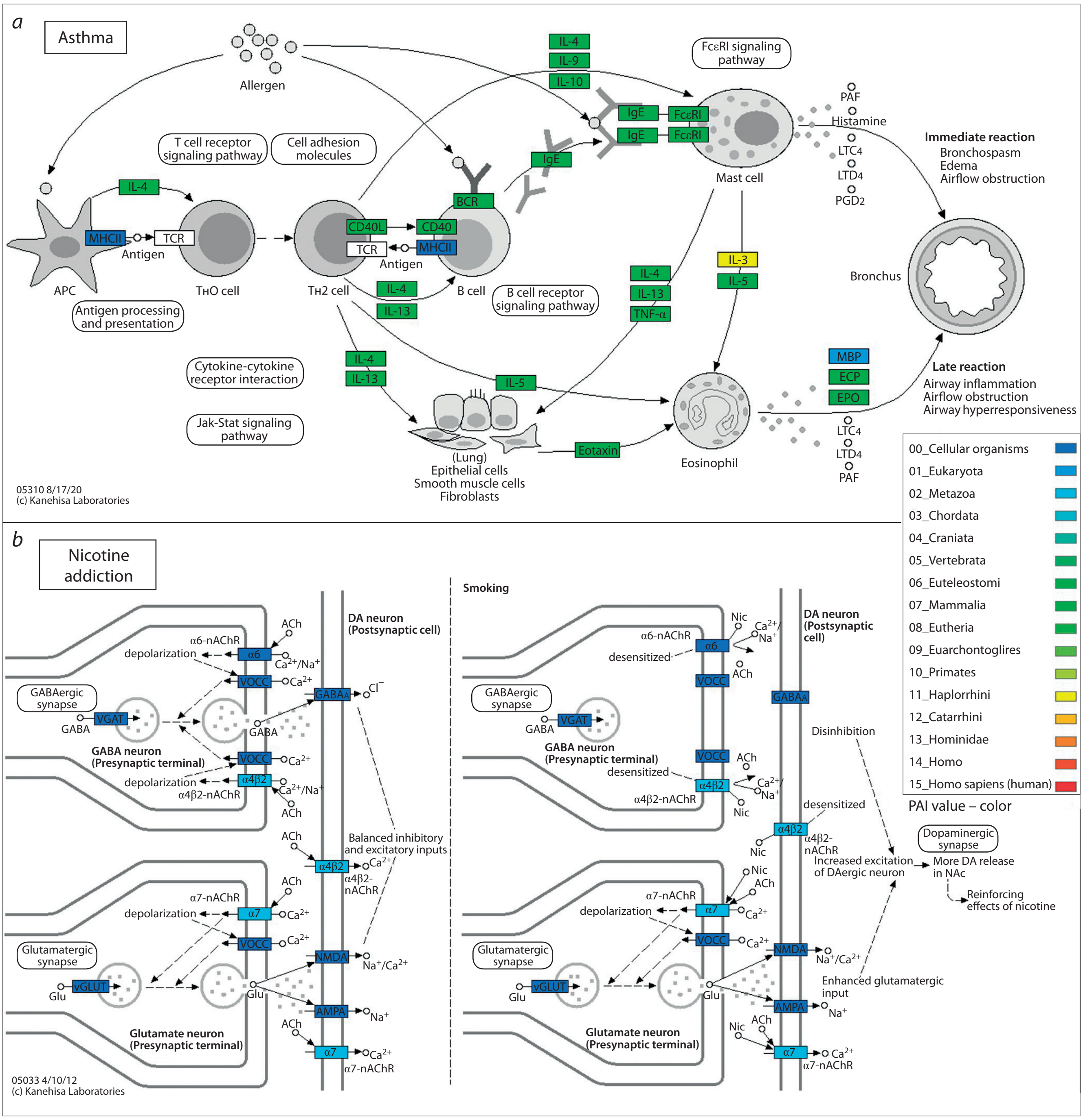


Fig. 2. Gene networks schemes of diseases "Asthma" (*a*) and "Nicotine addiction" (*b*) taken from the KEGG Pathway, Human Disease database with PAI values.

Gene coding the proteins in these networks are shown as rectangles with gene name, the color reflects the gene age. The genes colored in blue and cyan correspond to the most evolutionary ancient taxa, green and yellow correspond to evolutionary younger in compare with taxa colored in blue.

genes analyzed in cancers specific types category, which includes genes associated with carcinogenesis. This result is consistent with the current ideas that gene networks involved in cancer development processes were formed during the stages of multicellular organisms origin (Domazet-Lošo, Tautz, 2010).

Let us consider two categories of diseases in more details: (1) immune diseases with the highest proportion of evolutionary young genes and (2) substance dependence with the highest proportion of evolutionary ancient genes (Fig. 4).

Figure 4 shows the PAI distributions for 13 networks (8 immune diseases networks and 5 substance dependence networks) as "violin plot" graphs. The lower and upper points of each graph show the minimum and maximum PAI values, the orange star shows the median PAI values, and the width of the graph for each position on the ordinate

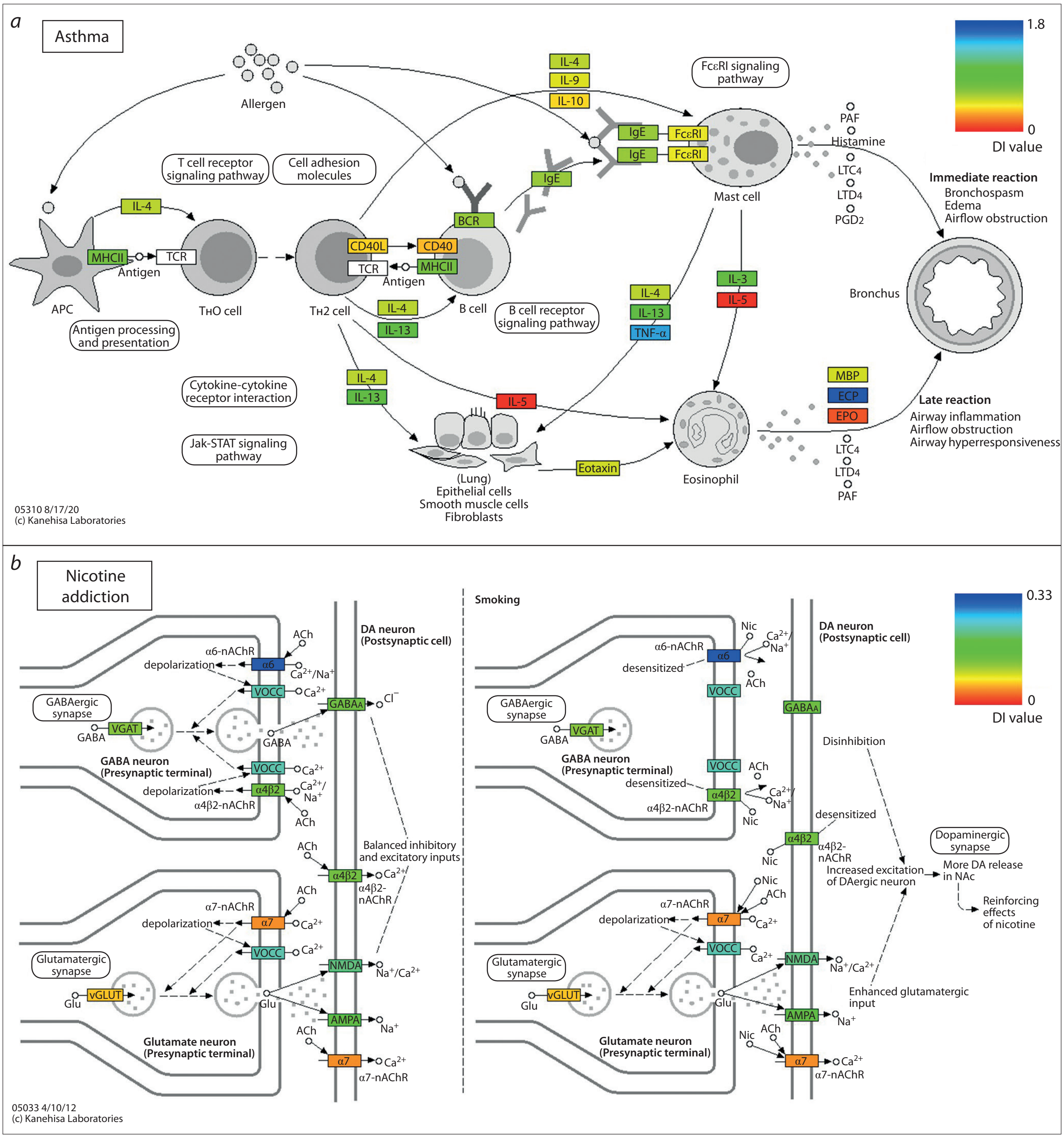


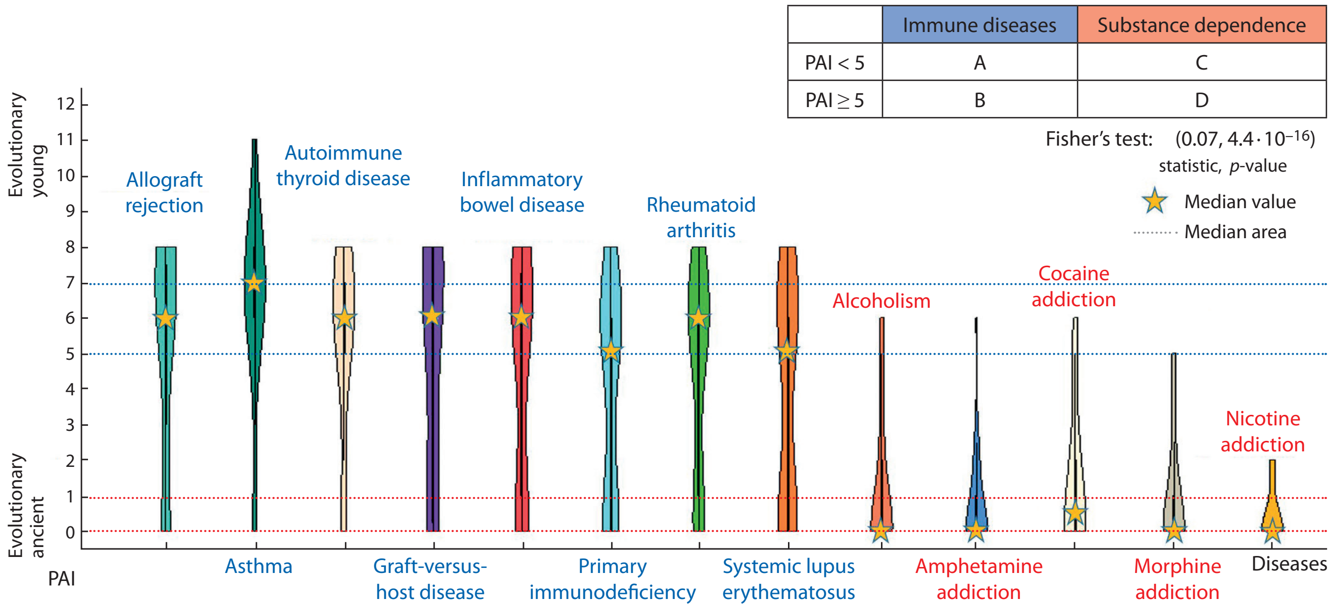
Fig. 3. Gene networks schemes of diseases "Asthma" (*a*) and "Nicotine addiction" (*b*) taken from KEGG Pathway, Human Disease database with DI values.

Gene coding the proteins in these networks are shown as rectangles with gene name, the color reflects the gene variability level. In the upper right part of the graph of each network placed the color scheme for DI. The scale for each network is individual, and even the most variable genes involved in the "Nicotine addiction" network have minimal variability compared to genes involved in the "Asthma" network.

axis (i. e. for each PAI) shows the proportion of genes with that particular PAI. The median for distributions of immune diseases varies in the range (5, 7) (from Vertebrata to Mammalia), and the distributions themselves have a character expressed in decreasing the number of genes with a corresponding PAI value as PAI decreases. The median for distributions of substance dependence varies in the range (0, 1) (Cellular organisms and Eukaryota), and the distributions themselves have a character expressed in increasing the number of genes with a corresponding PAI value as PAI decreases. These distributions are fundamentally different when comparing the proportion of evolutionary ancient and

Table 3. The results of the Fisher's exact test comparing the distribution of evolutionary ancient and evolutionary young genes among all genes described in the human disease gene networks from KEGG Pathway, Human Diseases, and among genes within the same category

KEGG Pathway, Human Diseases category	Genes		PAI	*p*-value
	evolutionary ancient	evolutionary young		
Immune diseases	56	106	5.21	8.84×10^{-15}
Infectious diseases parasitic	74	84	4.08	2.79×10^{-6}
Cancers specific types	187	54	1.77	4.41×10^{-4}
Substance dependence	69	9	1.03	1.75×10^{-5}
Total of 1436 genes	952	484	2.49	–



	Immune diseases	Substance dependence
PAI < 5	A	C
PAI ≥ 5	B	D

Fig. 4. Distribution of PAI among eight networks from category immune diseases (blue) and five networks from category substance dependence (marked in red).

Plots are visualized with R package vioplot, script is created by Orthoscape.

evolutionary young genes, as shown also by Fisher's exact test with p-value $= 4.4 \times 10^{-16}$.

Figure 5 shows the distribution of PAI among all the genes involved in the 80 gene networks considered from the KEGG Pathway, Human Diseases. This distribution has two peaks. The left peak includes genes formed early in evolution (from the emergence of cellular organization of life to chordates), and the right one includes genes formed at subsequent stages of evolution (vertebrates to placentals). There were more evolutionary ancient genes than evolutionary young ones.

Figure 6 shows the distribution of DI among all the genes involved in considered gene networks from the KEGG Pathway, Human Diseases. The DI analysis makes it possible to estimate what type of selection the genes are influenced. However, it only makes sense when the sequences of the analyzed genes are compared with the orthologous genes of evolutionary close organisms. To calculate *dN/dS*, human gene sequences were compared with the sequences of

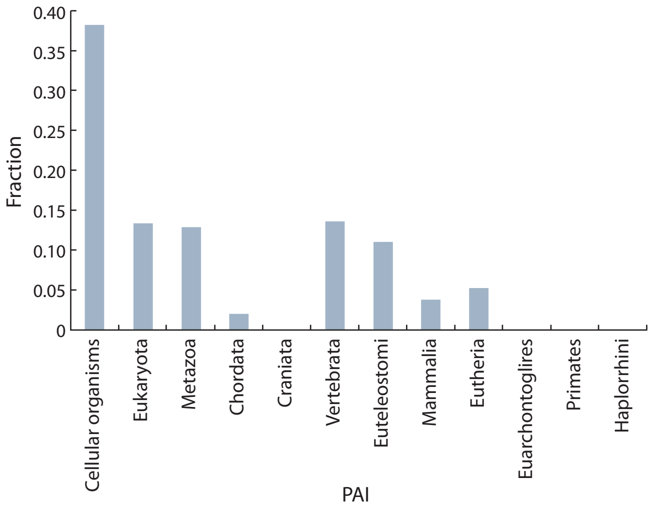


Fig. 5. Distribution of PAI among all genes involved in the considered gene networks from the KEGG Pathway, Human Diseases.

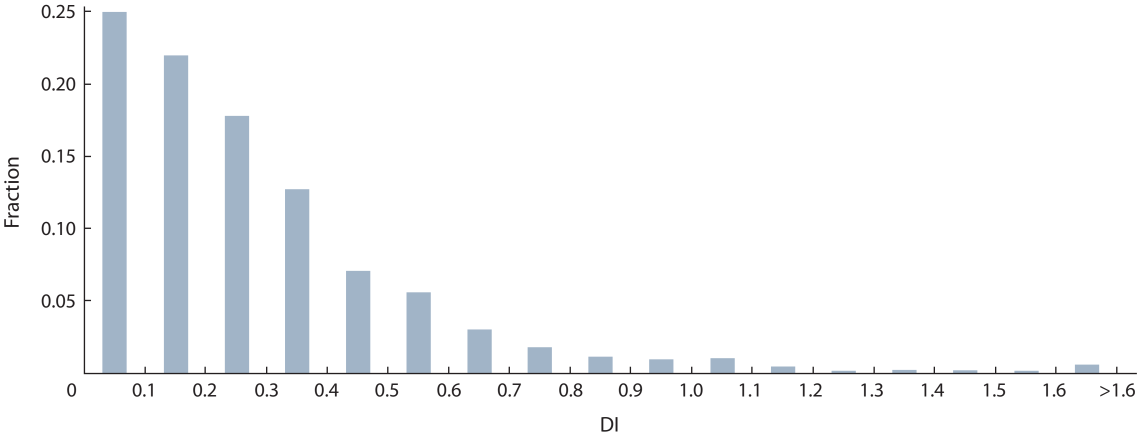


Fig. 6. Distribution of DI among all genes involved in the considered gene networks from the KEGG Pathway, Human Diseases.

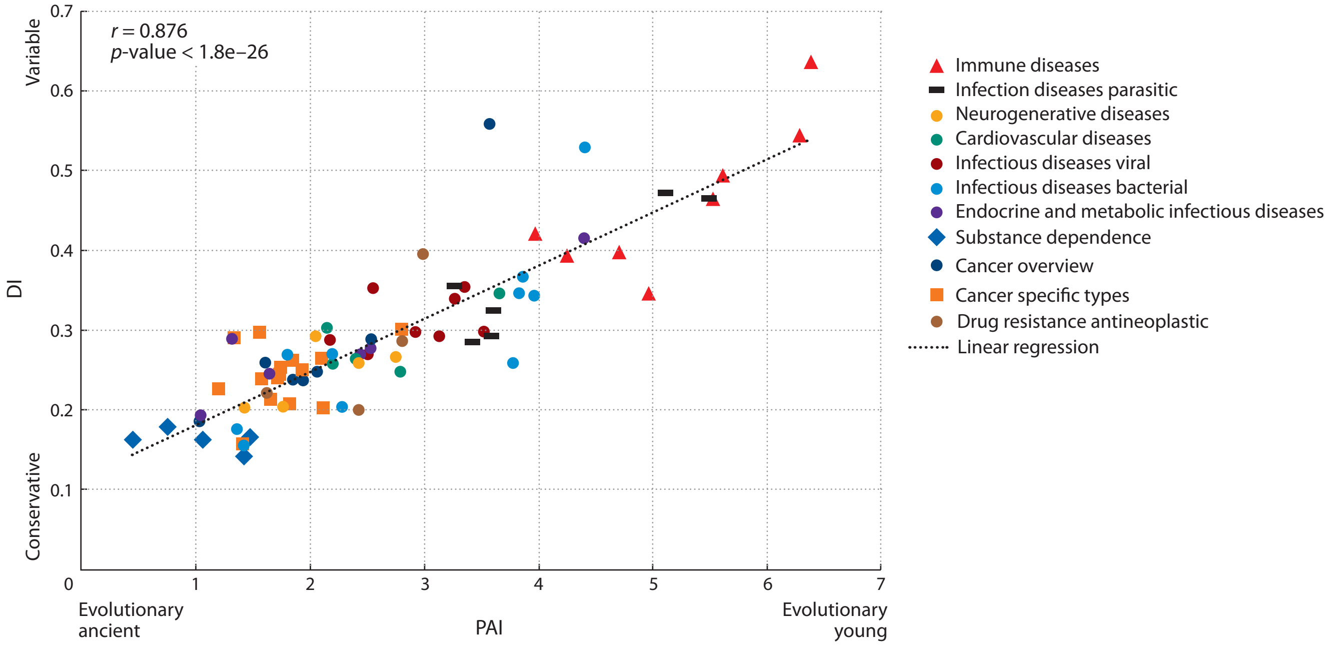


Fig. 7. Scatter plot for mean values of PAI and DI indices for 80 gene networks from KEGG Pathway, Human Diseases database. The figures of different forms and sizes show the different diseases categories.

orthologous genes of other hominids; if there were several orthologs, the average value of *dN/dS* was used as the DI. Only 38 of the 1,436 genes had DI values > 1 (nine of them fall into one category, immune diseases). The vast majority of genes included in studied gene networks evolved in the mode of stabilizing selection (DI < 1).

It was interesting to study the relationship between PAI and DI for the 80 gene networks we studied. Figure 7 presents the results of this analysis in a single graph, taking into account the categorization of diseases. Figures of different colors and sizes indicate different disease categories.

The analysis showed that there is a large significant correlation between PAI and DI with the value of the correlation coefficient ($r = 0.876$, p-value $< 1.8 \times 10^{-26}$). It means there is a relationship between the average evolutionary age of genes in gene networks and the level of their genetic variability: the less the evolutionary age of genes, the greater the level of their genetic variability. This agrees well with the fact that evolutionary ancient genes are involved in key processes for organism functionality; they are a subject to many restrictions by other genes and molecular-genetic systems organization peculiarities, so they are not characterized by high variability. On the contrary, evolutionary young genes provide adaptation to modern life conditions and are characterized by higher variability.

Conclusion

Phylostratigraphic analysis is a modern methodology that allows genome-wide estimation of gene ages based on data on the similarity of genetic sequences and the origin of organisms. Together with information on what type of selection a gene is subject to as a unit of heredity, the results of the analysis allow us to estimate the role of certain genes in the evolution of the gene networks of an organism.

Analysis of gene networks from the KEGG Pathway, Human Diseases database shows several trends. The vast majority of the genes involved in the gene networks studied evolved in the mode of stabilizing selection (DI < 1). There is significant ($r = 0.876$, p-value $< 1.8 \times 10^{-26}$) correlation between the average evolutionary age of genes in gene networks and their level of genetic variability: the lower the evolutionary age of genes, the greater the genetic variability is. Some categories of gene networks are especially distinguished by the proportion of evolutionary young and evolutionary ancient genes. The highest proportion of evolutionary young genes (65 %) was found in gene networks from immune diseases category. The highest proportion of evolutionary ancient genes (88 %) was found in gene networks from substance dependence category.

It is also shown that gene networks responsible for the functioning of infectious diseases caused by parasites are significantly enriched with evolutionary young genes, and gene networks responsible for the development of specific cancer types are significantly enriched with evolutionary ancient genes. Such results indicate an active process of adaptation of the human immune system to emerging threats. In addition, the genes involved in chemical addictive diseases have a minimum number of substitutions, i. e., such genes are as conservative as possible. Separate work can be carried out in this direction, with expansion of the original networks thanks to the classifiers and databases currently available.

References

Bell E.A., Boehnke P., Harrison T.M., Mao W.L. Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. *Proc. Natl. Acad. Sci. USA*. 2015;112:14518-14521. DOI 10.1073/pnas.1517557112.

Cerami E.G., Gross B.E., Demir E., Rodchenkov I., Babur Ö., Anwar N., Schultz N., Bader G.D., Sander C. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.* 2011;39: 685-690. DOI 10.1093/nar/gkq1039.

Chatterjee H.J., Ho S.Y., Barnes I., Groves C. Estimating the phylogeny and divergence times of primates using a supermatrix approach. *BMC Evol. Biol.* 2009;9:259. DOI 10.1186/1471-2148-9-259.

Datta P.M. Earliest mammal with transversely expanded upper molar from the Late Triassic (Carnian) Tiki Formation, South Rewa Gondwana Basin, India. *J. Vertebr. Paleontol.* 2005;25:200-207. DOI 10.1671/0272-4634(2005)025(0200:EMWTEU)2.0.CO;2.

Diogo R. The Origin of Higher Clades: Osteology, Myology, Phylogeny and Evolution of Bony Fishes and the Rise of Tetrapods. New York: CRC Press, 2007.

Domazet-Lošo T., Brajković J., Tautz D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* 2007;23:533-539. DOI 10.1016/j.tig.2007.08.014.

Domazet-Lošo T., Tautz D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol.* 2010;8:66.

Dunn R.H., Rose K.D., Rana R.S., Kumar K., Sahni A., Smith T. New euprimate postcrania from the early Eocene of Gujarat, India, and the strepsirrhine-haplorhine divergence. *J. Hum. Evol.* 2016;99: 25-51.

Galaktionov V.G. Immunology: a Guide for University Students Studying in Track 510600 “Biology” and Biological Specialties. Moscow: Academia Publ., 2004. (in Russian)

Harrison T. Catarrhine origins. In: A Companion to Paleoanthropology. New York: Blackwell Publ. Ltd., 2013;376-396.

Hey J. The ancestor’s tale A pilgrimage to the dawn of evolution. *J. Clin. Invest.* 2005;115:1680-1680.

Kanehisa M., Furumichi M., Tanabe M., Sato Y., Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017;45:D353-D361.

Khaitov R.M. Immunology: a Guide for Students of Medical Universities. Moscow, 2016. (in Russian)

Khakoo S.I. HLA and NK cell inhibitory receptor genes in resolving hepatitis C virus infection. *Science*. 2004;305(5685):872-874.

Kolchanov N.A., Ignat’eva E.V., Podkolodnaya O.A., Likhoshvay V.A., Matushkin Yu.G. Gene Networks. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding.* 2013;17(4/2):833-850. (in Russian)

Kumar V., Hallström B.M., Janke A. Coalescent-based genome analyses resolve the early branches of the euarchontoglires. *PLoS One.* 2013;8(4):e60019.

Leander B.S. Predatory protists. *Curr. Biol.* 2020;30:R510-R516.

Li W.-H. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* 1993;36(1):96-99.

Li W.H., Wu C.I., Luo C.C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 1985;2(2):150-174.

Liebeskind B.J., McWhite C.D., Marcotte E.M. Towards consensus gene ages. *Genome Biol. Evol.* 2016;8(6):1812-1823.

Luo Z.-X., Yuan C.-X., Meng Q.-J., Ji Q. A Jurassic eutherian mammal and divergence of marsupials and placentals. *Nature.* 2011;476: 442-445.

Maloof A.C., Porter S.M., Moore J.L., Dudas F.O., Bowring S.A., Higgins J.A., Fike D.A., Eddy M.P. The earliest Cambrian record of animals and ocean geochemical change. *Geol. Soc. Am. Bull.* 2010a; 122:1731-1774.

Maloof A.C., Rose C.V., Beach R., Samuels B.M., Calmet C.C., Erwin D.H., Poirier G.R., Yao N., Simons F.J. Possible animal-body fossils in pre-Marinoan limestones from South Australia. *Nat. Geosci.* 2010b;3:653-659.

Montojo J., Zuberi K., Rodriguez H., Kazi F., Wrig G., Donaldson S.L., Morris Q., Bader G.D. GeneMANIA cytoscape plugin: Fast gene function predictions on the desktop. *Bioinformatics.* 2010;26:2927-2928.

Mustafin Z.S., Lashin S.A., Matushkin Y.G., Gunbin K.V., Afonnikov D.A. Orthoscape: a cytoscape application for grouping and visualization KEGG based gene networks by taxonomy and homology principles. *BMC Bioinformatics.* 2017;18(S1):1-9.

Nei M., Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 1986;3:418-426.

Nersisyan L., Samsonyan R., Arakelyan A. CyKEGGParser: tailoring KEGG pathways to fit into systems biology analysis workflows. *F1000Res.* 2014;3:145.

Pamilo P., Bianchi N.O. Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. *Mol. Biol. Evol.* 1993;10(2): 271-281.

Sasaki K., Tsutsumi A., Wakamiya N. Mannose-binding lectin polymorphisms in patients with hepatitis C virus infection. *Scand. J. Gastroenterol.* 2000;35(9):960-965.

Scerri E.M.L., Thomas M.G., Manica A., Gunz P., Stock J.T., Stringer C., Grove M., Groucutt H.S., Timmermann A., Rightmire G.P., D’Errico F., Tryon C.A., Drake N.A., Brooks A.S., Dennell R.W., Durbin R., Henn B.M., Lee-Thorp J., DeMenocal P., Petraglia M.D., Thompson J.C., Scally A., Chikhi L. Did our species evolve in subdivided populations across Africa, and why does it matter? *Trends Ecol. Evol.* 2018;33(8):582-594.

Schrenk F., Kullmer O., Bromage T. The earliest putative homo fossils. In: Handbook of Paleoanthropology. Berlin; Heidelberg: Springer, 2014;1-19.

Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498-2504.

Shu D.-G., Luo H.-L., Conway Morris S., Zhang X.-L., Hu S.-X., Chen L., Han J., Zhu M., Li Y., Chen L.-Z. Lower Cambrian vertebrates from south China. *Nature.* 1999;402(6757):42-46.

Stepanov V.A. Evolution of genetic diversity and human diseases. *Russ. J. Genet*. 2016;52(7):746-756.

Szklarczyk D., Gable A.L., Lyon D., Junge A., Wyder S., Huerta-Cepas J., Simonovic M., Doncheva N.T., Morris J.H., Bork P., Jensen L.J., von Mering C. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019; 47(D1):D607-D613.

Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 2007;24(8):1586-1591.

Yang Z., Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 2000;17(1):32-43.

Zheleznikova G.F. Infection and immunity: strategies from both sides. *Med. Immunol.* 2014;8(5-6):597-614. DOI 10.15789/1563-0625-2006-5-6-597-614. (in Russian)

ORCID ID

Z.S. Mustafin orcid.org/0000-0003-2724-4497
S.A. Lashin orcid.org/0000-0003-3138-381X
Yu.G. Matushkin orcid.org/0000-0001-7754-8611

Acknowledgements. The work was supported by the Russian Foundation for Basic Research No. 20-04-00885 A and the budget project No. 0259-2021-0009.

Funding transparency. The authors do not hold financial interests in the presented materials or methods.

Conflict of interest. The authors declare no conflict of interest.

Received January 14, 2021. Revised January 20, 2021. Accepted January 20, 2021.