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

Английский текст https://vavilov.elpub.ru/jour

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.

Для цитирования: Мустафин З.С., Лашин С.А., Матушкин Ю.Г. Филостратиграфический анализ генных сетей заболеваний человека. *Вавиловский журнал генетики и селекции*. 2021;25(1):46-56. DOI 10.18699/VJ21.006

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

© Мустафин З.С., Лашин С.А., Матушкин Ю.Г., 2021
Контент доступен под лицензией Creative Commons Attribution 4.0

Введение

Исследование ключевых факторов, влияющих на развитие и протекание заболеваний, – одно из важнейших направлений как для медицины, так и для биологии (Степанов, 2016). Как известно, формирование фенотипических признаков, обеспечивающих адаптацию организмов к условиям окружающей среды, контролируется не отдельными генами, а генными сетями – группами координированно функционирующих генов и продуктов их работы (РНК, белками, метаболитами и др.) (Колчанов и др., 2013). Возникает задача выделения ключевых структурных особенностей сетей, элементов сетей, а также их численного описания. Одной из важных характеристик является возраст гена. Возраст гена, принадлежащего к ортологической группе, определяется как этап возникновения наиболее недавнего предка всех видов, представленных в этой группе (Liebeskind et al., 2016).

Современные методы анализа дают возможность оценить эволюционные характеристики генов, в частности филостратиграфический анализ – методология, предложенная в 2007 г. T. Domazet-Lošo, – позволяет определить возраст гена с помощью специального индекса, получаемого в результате анализа ортологичных генов и сравнения положения организмов, чьи гены рассматриваются в анализе на филогенетическом дереве (Domazet-Lošo et al., 2007).

Для работы с генными сетями существует множество программных средств. Одни сконцентрированы на реконструкции сетей на основании данных из биологических баз, например String (Szklarczyk et al., 2019), GeneMANIA (Montojo et al., 2010). Другие имеют обширный функционал по визуализации элементов сети, выявлению ее структурных особенностей: Cytoscape (Shannon et al., 2003), yEd (https://www.yworks.com/products/yed). Программный комплекс Cytoscape выгодно отличается от других средств тем, что, помимо обширных возможностей по построению сети, компоновке и покраске ее элементов, анализу структурных особенностей, он позволяет пользователям писать собственные приложения на языке Java и встраивать их в Cytoscape в качестве плагинов. Это открывает сообществу возможность реализовывать весь интересующий функционал и добавлять его в Cytoscape. Например, такие известные средства, как String и GeneMANIA, способные реконструировать сеть по списку генов на основании извлечения взаимодействий из баз биологических данных, имеют свои собственные плагины в Cytoscape и позволяют пользоваться своей функциональностью, сочетая ее с возможностями Cytoscape и других его плагинов. Пользователю также становится доступным импорт готовых сетей, например из баз Pathway Commons (Cerami et al., 2011) или KEGG Pathway (Kanehisa et al., 2017), без необходимости разбора форматов представления сети в этих базах. Наконец, с учетом всех имеющихся возможностей любой пользователь может написать собственное приложение под свои задачи и поделиться им с сообществом.

В настоящей работе представлены результаты анализа генных сетей одним из таких плагинов, Orthoscape (Mustafin et al., 2017), способным проанализировать эволюционные особенности генов в генной сети. Продемонстрировано, что большинство генов, описанных в генных сетях, подвержено стабилизирующему отбору, и обнаружена высокая достоверная корреляция между временем возникновения гена и наблюдаемым уровнем генетической изменчивости – чем моложе ген, тем выше уровень генетической изменчивости. Показано, что среди проанализированных генных сетей наибольшая доля эволюционно молодых генов выявлена в сетях, связанных с заболеваниями иммунной системы (65 %), а эволюционно древних – 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.

Материалы и методы

Исходные данные для анализа. В работе использовали генные сети, представленные в базе KEGG Pathway, раздел Human Diseases. Сети в этом разделе разбиты на категории, всего таких категорий 11 (суммарно включающих 80 сетей): нейродегенеративные заболевания (neurodegenerative diseases, 5 сетей), сердечно-сосудистые заболевания (cardiovascular diseases, 5 сетей), заболевания, связанные с иммунной системой (immune diseases, 8 сетей), эндокринные заболевания и нарушения метаболизма (endocrine and metabolic diseases, 6 сетей), инфекционные заболевания, вызванные бактериями (infectious diseases: bacterial, 10 сетей), инфекционные заболевания, вызванные вирусами (infectious diseases: viral, 9 сетей), инфекционные заболевания, вызванные паразитами (infectious diseases: parasitic, 6 сетей), лекарственная устойчивость к противоопухолевым препаратам (drug resistance: antineoplastic, 4 сети), рак: обобщение (cancers: overview, 7 сетей), специфические типы рака (cancers: specific types, 15 сетей), зависимость от химических соединений, вызывающих привыкание (substance dependence, 5 сетей).

Необходимые данные для проведения анализа: списки ортологичных генов, нуклеотидные последовательности генов и аминокислотные последовательности кодируемых ими белков, доменный состав, информация о таксономических рядах организмов, чьи гены рассматривали в анализе, – также были взяты из базы KEGG.

Используемое программное обеспечение. Анализ проводили на базе программного комплекса Cytoscape (Shannon et al., 2003) – многофункционального средства для визуализации и анализа сетей. Для импорта сетей из KEGG Pathway в работе использовали плагин CyKEGGParser (Nersisyan et al., 2014). Для выполнения филостратиграфического анализа и анализа индекса эволюционной изменчивости был взят плагин Orthoscape (Mustafin et al., 2017).

Методы оценки эволюционных характеристик генов. Orthoscape позволяет оценить две эволюционные характеристики генов. Первая характеристика, вычисляемая с помощью Orthoscape, – *филостратиграфический индекс гена* (phylostratigraphic age index, PAI). Он показывает, в какой степени отдален от корня филогенетического дерева таксон, отражающий возраст гена, т. е. такой таксон, на

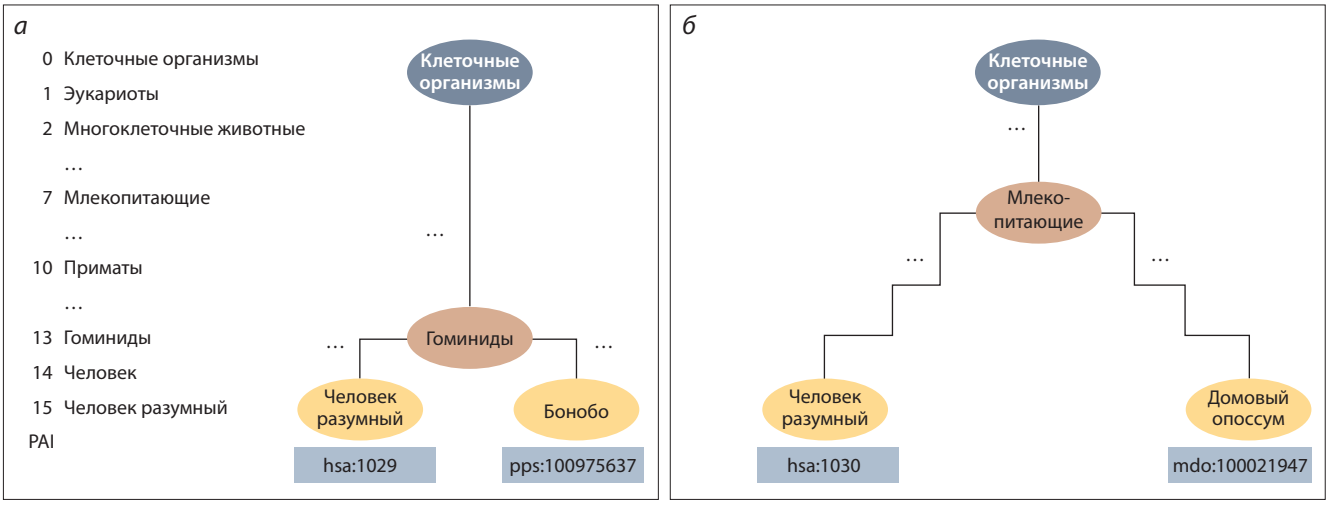

Рис. 1. Пример определения PAI для двух генов *Homo sapiens* (человек).

а – пример эволюционно молодого гена hsa:1029, наиболее отдаленным от исследуемого организмом, у которого был найден ортолог этого гена, является *Pan paniscus* (шимпанзе бонобо); *б* – пример эволюционно более древнего гена hsa:1030, наиболее отдаленный от исследуемого организм, у которого был найден ортолог этого гена, – *Monodelphis domestica* (домовый опоссум). Можно заключить, что ген на примере (*а*) эволюционно моложе гена на примере (*б*). Шкала слева показывает индекс PAI, который соответствует глубине узла таксономического дерева (подробнее см. табл. 1).

котором произошло расхождение исследуемого вида с наиболее отдаленным родственным таксоном, в котором обнаружен ортолог рассматриваемого гена. Таким образом, чем больше PAI исследуемого гена, тем он моложе (рис. 1). Для расчета PAI в Orthoscape применяется сервис KEGG Orthology, что дает возможность учитывать среди всех гомологов гена именно ортологичные.

Таблица 1. Список таксонов, выделенных для филостратиграфического анализа генов *H. sapiens*

PAI	Таксон	Возраст, млн лет
0	Cellular organism (клеточные организмы, корень дерева)	4100 (Bell et al., 2015)
1	Eukaryota (эукариоты)	1850 (Leander, 2020)
2	Metazoa (многоклеточные животные)	665 (Maloof et al., 2010a)
3	Chordata (хордовые)	541 (Maloof et al., 2010b)
4	Craniata (плеченогие)	535 (Maloof et al., 2010b)
5	Vertebrata (позвоночные)	525 (Shu et al., 1999)
6	Euteleostomi (костные позвоночные)	420 (Diogo, 2007)
7	Mammalia (млекопитающие)	225 (Datta, 2005)
8	Eutheria (плацентарные)	160 (Luo et al., 2011)
9	Euarchontoglires (грызунообразные + эуархонты)	65 (Kumar et al., 2013)
10	Primates (приматы)	55 (Chatterjee et al., 2009)
11	Haplorrhini (обезьяны)	50 (Dunn et al., 2016)
12	Catarrhini (узконосые обезьяны)	44 (Harrison, 2013)
13	Hominidae (гоминиды)	17 (Hey, 2005)
14	Homo (люди)	2.8 (Schrenk et al., 2014)
15	Homo sapiens (человек разумный)	0.35 (Scerri et al., 2018)

Важная характеристика для филостратиграфического анализа – список таксономических единиц, описывающих этапы расхождения на эволюционном дереве организма, чьи гены исследуются с другими организмами, ортологи которых могут быть найдены. Полный список таксонов, использованный в анализе для определения филостратиграфического индекса генов *H. sapiens*, а также примерный эволюционный возраст этих таксонов в млн лет от нашего времени приведены в табл. 1. Следует отметить, что дискуссии на эту тему ведутся, в разных источниках указаны разные показатели образования того или иного таксона; значения в табл. 1 отражают примерные оценки.

Программа Orthoscape также позволяет оценить *индекс эволюционной изменчивости гена* (divergence index, DI). Он показывает тип отбора, которому подвержен ген. Индекс DI вычисляется на основании отношения *dN*/*dS*, где *dN* – доля несинонимичных замен в последовательностях исследуемого гена и его ортолога, т. е. таких замен, которые приводят к смене кодируемой данным триплетом аминокислоты; *dS* – доля синонимичных замен, т. е. не приводящих к замене кодируемой аминокислоты. Значение индекса в диапазоне от 0 до 1 свидетельствует о том, что ген подвержен стабилизирующему отбору, 1 – нейтральной эволюции, а больше 1 – движущему отбору. Анализ данного индекса имеет смысл только при сравнении близкородственных организмов, поскольку методика не дает учесть многократные замены в одной и той же позиции, которые неизбежно будут накоплены при сравнении с организмами, эволюционно отдаленными от исследуемого. Вычисление *dN/dS* проходит в два этапа:

1. Выравнивание исходных последовательностей рассматриваемого гена и ортологичного гена. Оно осуществляется с помощью алгоритма Нидлмана–Вунша, выравниваются аминокислотные последовательности с сохранением нуклеотидных триплетов, кодирующих аминокислоты. Затем позиции с разрывами удаляются.

2. Выравненные последовательности подаются на вход средству PAML (phylogenetic analysis by maximum likelihood) (Yang, 2007). Для вычисления *dN/dS* применяются методы, по-разному учитывающие позиции триплетов, их частоту встречаемости и прочие факторы. В PAML реализованы методы: Nei–Gojobori (Nei, Gojobori, 1986), Yang & Nielsen (Yang, Nielsen, 2000), LWL85 (Li, 1985), LWLm (Li, 1993), LPB93 (Pamilo, Bianchi, 1993). Для расчета DI мы использовали значение *dN/dS*, вычисленное по методу LPB93. Значение *dN/dS* рассчитывается для каждой пары ген-ортолог, итоговое значение DI определяется по формуле

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

где $dnds_i$ – значение *dN/dS* отношения для последовательности гена и ортолога *i*; *n* – число ортологов, попавших в анализ.

Результаты и обсуждение

Анализ эволюционных характеристик генных сетей

С помощью Orthoscape были посчитаны индексы PAI и DI для всех генов, представленных в 80 проанализированных генных сетях из KEGG Pathway, Human Diseases. На основании этих данных были вычислены значения PAI для каждой генной сети (табл. 2) как среднее значение PAI всех генов, задействованных в сети, и PAI категории как среднее значение PAI всех сетей из этой категории. Аналогичным образом для каждой генной сети был определен индекс DI.

Среди проанализированных 80 сетей наблюдается варьирование PAI от 0.44 (т. е. большая часть генов эволюционно древняя, генная сеть «Никотиновая зависимость») до 6.38 (т. е. большая часть генов эволюционно молодая, генная сеть «Астма»). Изменение DI гена, как правило, происходит в пределах DI < 1, т. е. в пределах стабилизирующего отбора, тем не менее уровень изменчивости генов, задействованных в разных сетях, также сильно варьирует: от 0.16 до 0.64. Наиболее выделяются по индексам PAI и DI сети «Астма» и «Никотиновая зависимость». В сети «Астма» преобладают эволюционно молодые и изменчивые гены, а в сети «Никотиновая зависимость» – эволюционно древние и консервативные. Результат анализа PAI для сетей «Астма» и «Никотиновая зависимость» приведен на рис. 2; результаты анализа DI этих же сетей – на рис. 3.

Большинство генов в сети «Астма» – эволюционно молодые, появившиеся на уровне позвоночных (см. рис. 2, *а*, окрашены зеленым и желтым цветами). Напротив, в сети «Никотиновая зависимость» (см. рис. 2, *а*) все гены были определены как эволюционно древние, возникшие на этапах образования клеточной формы жизни (Cellular organisms) до многоклеточных животных (Metazoa).

Анализ индекса DI свидетельствует о том, что практически все гены в сети «Астма» (см. рис. 3, *а*) являются более эволюционно изменчивыми, чем гены, вовлеченные в сеть «Никотиновая зависимость» (см. рис. 3, *б*), гены которой очень консервативны.

Рассмотрим полученные оценки величин PAI для 11 категорий заболеваний (см. табл. 2). Наиболее выделяются из них 4: по высокому показателю PAI и DI – это болезни, связанные с иммунной системой (immune diseases, 8 сетей), и инфекционные заболевания, вызванные паразитами (infectious diseases: parasitic, 6 сетей). Низкий показатель PAI и DI характерен для специфических типов рака (cancers: specific types, 15 сетей) и зависимостей от химических соединений, вызывающих привыкание (substance dependence, 5 сетей).

Гены из рассмотренных выше категорий, а также полный набор 1436 генов были разбиты на две группы: 1) группа эволюционно древних генов с PAI < 5 (возраст генов соответствует периоду эволюции от формирования одноклеточных организмов (Cellular organisms) до хордовых (Chordata)); 2) группа эволюционно молодых генов с PAI ≥ 5 (возраст генов соответствует периоду эволюции от плеченогих (Craniata) до современного человека). Далее были составлены таблицы сопряженности и с помощью точного теста Фишера проведена оценка, является ли достоверным отличие в разбиении генов на группы в категории от разбиения в полном списке генов (табл. 3).

Среднее значение PAI всех 1436 исследованных генов составило 2.49. По результатам табл. 3 видно, что генные сети, связанные с заболеваниями иммунной системы, обладают не только самым высоким значением филостратиграфического индекса (5.21), но и достоверно отличным распределением доли молодых и древних генов от аналогичной доли среди всех проанализированных генов (см. в последней строке табл. 3).

Доля молодых генов в категории заболеваний, связанных с иммунной системой (immune diseases), составила 65 %. При этом наибольшая доля генов приходится на позвоночных (Vertebrata) и костных позвоночных (Euteleostomi), что соответствует современным представлениям о развитии специфического иммунитета: он существует у хрящевых рыб (акул и скатов) и, следовательно, появился по крайней мере 400–500 млн лет назад. У этих рыб есть гены, родственные генам вариабельной области *Ig* (*IgV*) или генам рецепторов Т-клеток (*TκP*). При этом еще более примитивные позвоночные – круглоротые (миксины и миноги) – не имеют системы приобретенного иммунитета, у них нет ни *IgV*-, ни *TκP*-генов (Галактионов, 2004). Анализ выявил также и некоторую долю эволюционно древних генов в категории заболеваний, связанных с иммунной системой. Это соответствует сложившемуся представлению о том, что некоторые функции иммунной системы возникали еще у одноклеточных, например способность к фагоцитозу; клетки, имеющие маркер Т-лимфоцита, впервые обнаруженные у кольчатых червей, система гистосовместимости, – у губок (Хаитов, 2016). С другой стороны, наибольшая доля эволюционно древних генов характерна для категории зависимостей от химических соединений, вызывающих привыкание (substance dependence), а именно 88 %. Большинство рассмотренных генов вовлечены в функционирование нервной системы, включая нейротрансмиттерные функции.

Достоверным отличием доли эволюционно древних и эволюционно молодых генов от аналогичной среди всех проанализированных генов обладает категория инфек-

Таблица 2. Средние значения индексов PAI и DI для генов, вовлеченных в генные сети заболеваний человека из базы данных KEGG Pathway, Human Diseases

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

* Категория: 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.

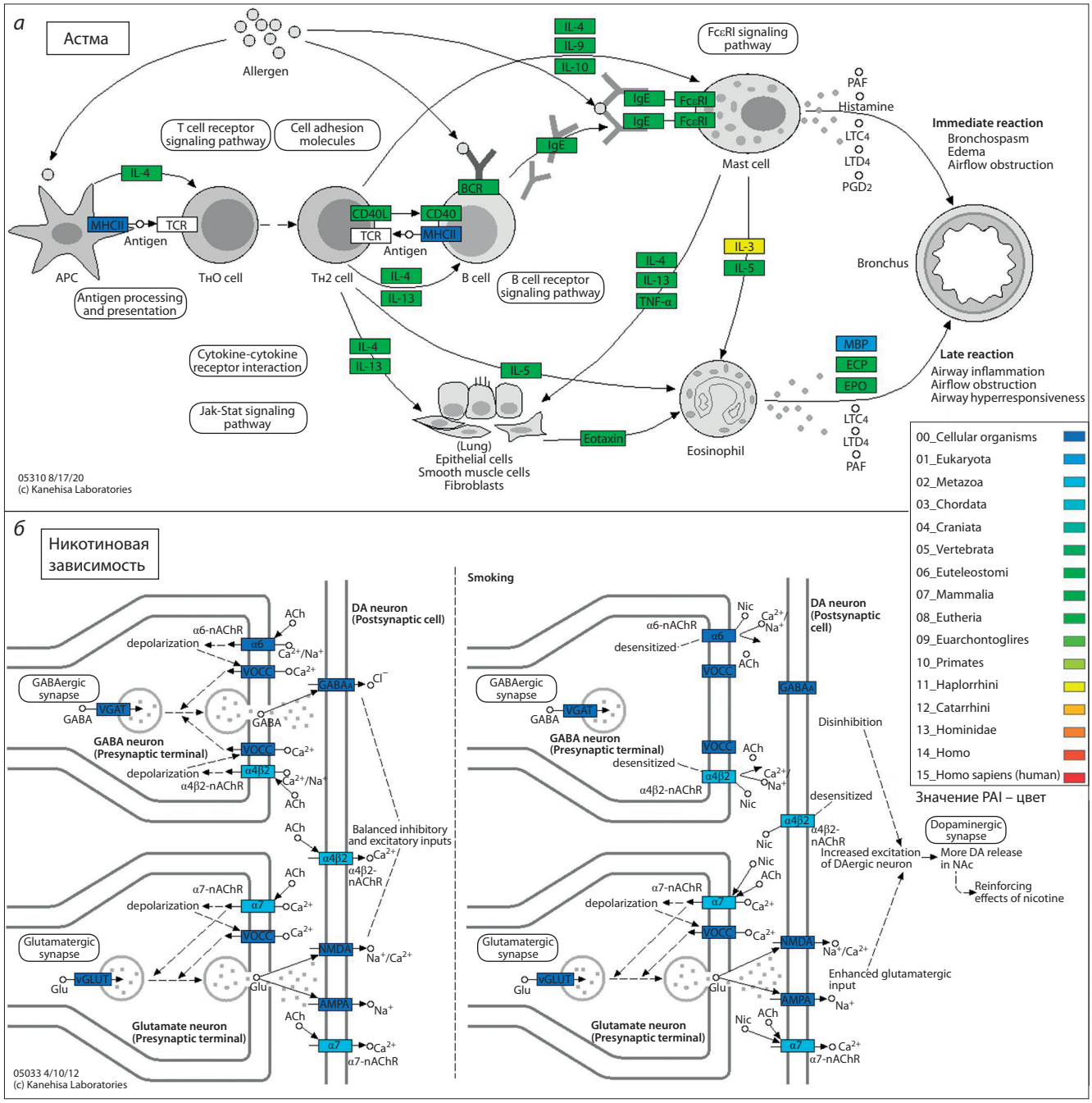

Рис. 2. Схемы генных сетей заболеваний «Астма» (*а*) и «Никотиновая зависимость» (*б*) из базы данных KEGG Pathway, Human Diseases с вычисленными значениями PAI.

Гены, кодирующие белки в этих сетях, показаны прямоугольниками с названием гена; цвет прямоугольника соответствует возрасту гена. Схема соответствия цвета и возраста гена приведена справа. Окрашенные в синий и голубой цвета гены относятся к наиболее эволюционно древним таксонам, в зеленый и желтый – к более эволюционно молодым относительно обозначенных голубым.

ционных заболеваний, вызванных паразитами (infectious diseases parasitic), 53 % эволюционно молодых генов. В этом случае высокая доля эволюционно молодых генов может быть напрямую связана с высокой долей эволюционно молодых генов и высокой эволюционной изменчивостью генов, найденной в категории заболеваний, связанных с иммунной системой. Именно инфекционные заболевания – один из важнейших движущих факторов эволюции иммунной системы. При этом инфекционные заболевания различной природы и иммунная система коэволюционируют в процессе формирования механизмов борьбы друг с другом (Sasaki et al., 2000; Khakoo, 2004; Zheleznikova, 2014).

Отметим также категорию специфических типов рака (cancers specific types), включающую гены, ассоциированные с канцерогенезом. Для нее наблюдается достоверное превышение доли древних генов над молодыми в сравнении с их распределением (древние/молодые) в полной выборке проанализированных генов. Этот результат соответствует современным представлениям о том,

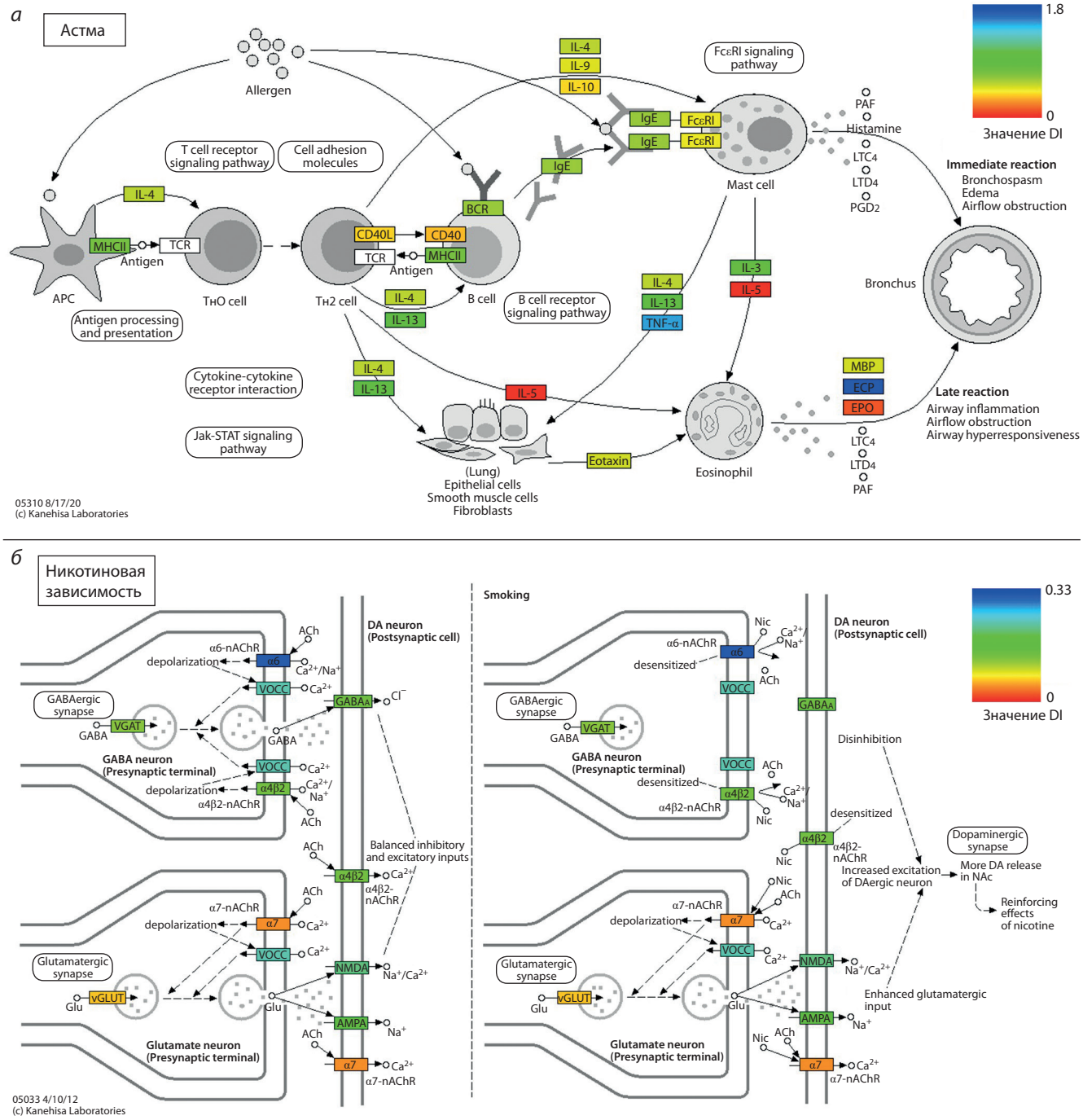

Рис. 3. Схемы генных сетей заболеваний «Астма» (*а*) и «Никотиновая зависимость» (*б*) из базы данных KEGG Pathway, Human Diseases с вычисленными значениями DI.

Гены, кодирующие белки в этих сетях, показаны прямоугольниками с названием гена; цвет прямоугольника соответствует уровню изменчивости гена. В правой верхней части графика для каждой сети приведена цветовая схема соотношения цветов и индекса DI. Шкала для каждой сети индивидуальна, и даже наиболее изменчивые гены, задействованные в сети «Никотиновая зависимость», обладают минимальной изменчивостью по сравнению с генами, вовлеченными в сеть «Астма».

что генные сети, вовлеченные в процессы развития рака, формировались на этапах возникновения многоклеточных организмов (Domazet-Lošo, Tautz, 2010).

Рассмотрим более подробно две категории заболеваний: 1) связанных с иммунной системой и обладающих наибольшей долей эволюционно молодых генов и 2) связанных с формированием зависимостей от химических веществ, вызывающих привыкание, обладающих наибольшей долей эволюционно древних генов (рис. 4). Нижняя и верхняя точки каждого графика показывают минимальное и максимальное значения PAI, оранжевая звезда – медиану значений PAI, ширина графика для каждой позиции по оси ординат (т. е. для каждого PAI) – долю генов с этим конкретным PAI (см. рис. 4). Можно видеть, что в случае

Таблица 3. Результаты точного теста Фишера по сравнению распределения по группам эволюционно древних и эволюционно молодых генов среди всех генов, описанных в генных сетях заболеваний человека из KEGG Pathway, Human Diseases, и среди генов в рамках одной категории

Категория KEGG Pathway, Human Diseases	Гены		PAI	*p*-value-теста
	эволюционно древние	эволюционно молодые		
Заболевания, связанные с иммунной системой	56	106	5.21	8.84×10^{-15}
Инфекционные заболевания, вызванные паразитами	74	84	4.08	2.79×10^{-6}
Специфические типы рака	187	54	1.77	4.41×10^{-4}
Зависимость от химических соединений, вызывающих привыкание	69	9	1.03	1.75×10^{-5}
Всего из 1436 генов	952	484	2.49	–

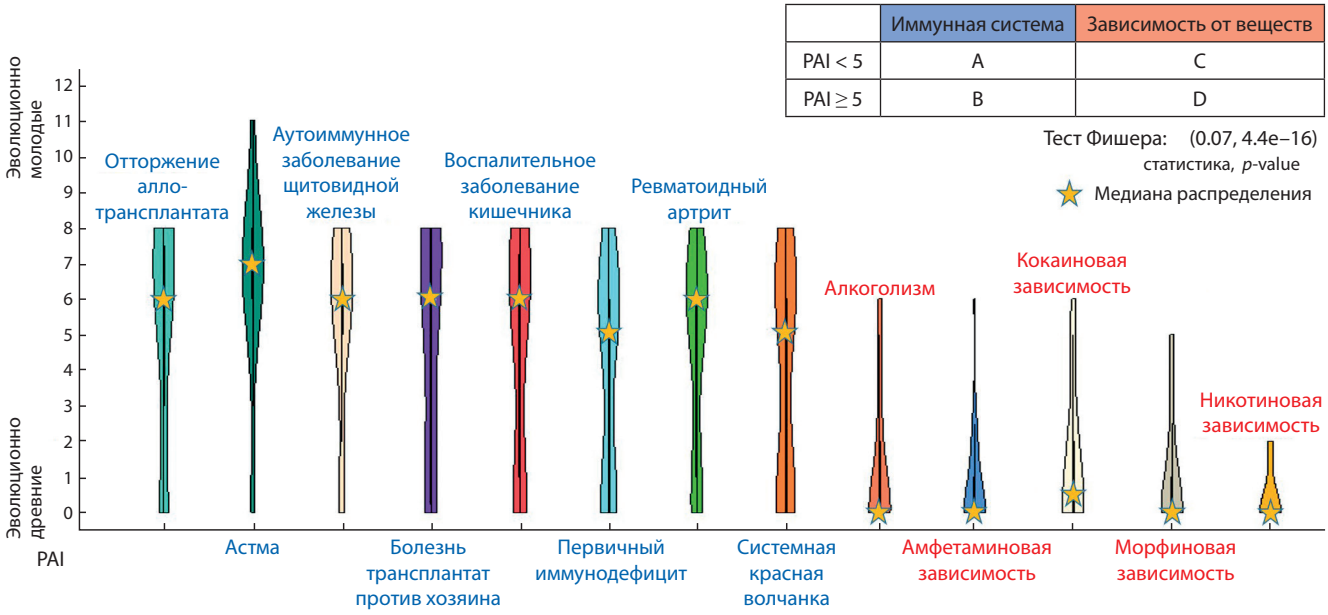

Рис. 4. Распределение PAI среди восьми сетей заболеваний, связанных с иммунной системой (подписаны синим) и пяти сетей заболеваний, связанных с зависимостями от веществ, вызывающих привыкание к химическим соединениям (подписаны красным).

Графики визуализированы с помощью R пакета vioplot, скрипт подготовлен Orthoscape.

заболеваний, связанных с иммунной системой, медиана распределений PAI колеблется в диапазоне (5, 7) (от позвоночных (Vertebrata) до млекопитающих (Mammalia)), а сами распределения имеют характер, выражающийся в уменьшении числа генов с соответствующим значением PAI при уменьшении PAI. В случае заболеваний, связанных с зависимостями от веществ, вызывающих привыкание к химическим соединениям, медиана находится в диапазоне (0, 1) – клеточные организмы (Cellular organisms) и эукариоты (Eukaryota), сами распределения имеют характер, выражающийся в увеличении числа генов с соответствующим значением PAI при уменьшении PAI. Распределения носят принципиально разный характер, если сравнивать в них долю эволюционно древних и эволюционно молодых генов, что показал также и точный тест Фишера с достоверностью p-value = 4.4×10^{-16}.

Распределение PAI среди всех генов, задействованных в 80 рассмотренных генных сетях из KEGG Pathway, Human Diseases, представлено на рис. 5. Это распределение

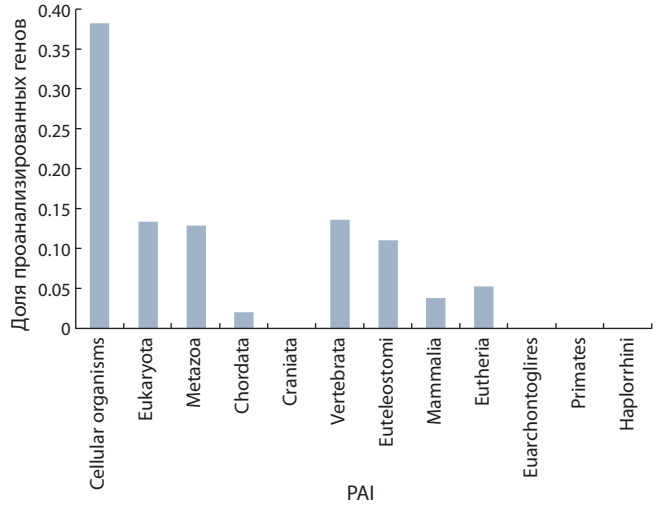

Рис. 5. Распределение PAI среди всех генов, задействованных в генных сетях из KEGG Pathway, Human Diseases.

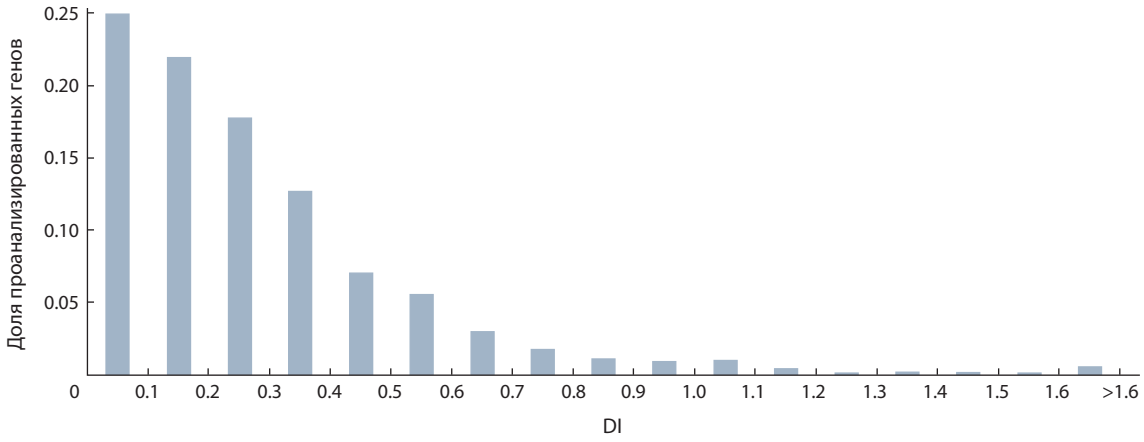

Рис. 6. Распределение DI среди всех генов, задействованных в генных сетях из KEGG Pathway, Human Diseases.

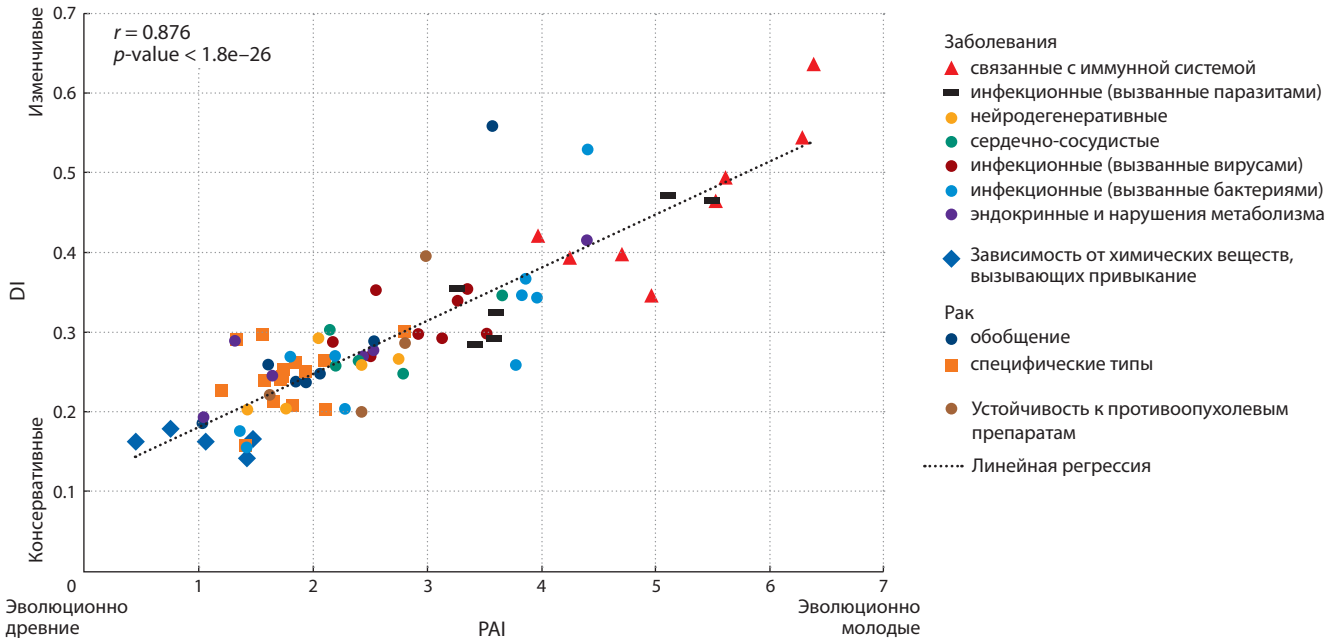

Рис. 7. Диаграмма рассеяния для средних значений индексов PAI и DI для 80 генных сетей заболеваний человека, описанных в базе KEGG Pathway, Human Diseases.

Фигурами разных цветов и размеров отмечены различные категории заболеваний.

имеет два пика. Левый пик включает гены, сформировавшиеся на раннем этапе эволюции (от возникновения клеточной организации жизни до хордовых), а правый – гены, сформировавшиеся на последующих этапах эволюции (от позвоночных до плацентарных). При этом эволюционно древних генов оказалось больше, чем эволюционно молодых.

Распределение DI среди всех генов, задействованных в рассмотренных генных сетях из KEGG Pathway, Human Diseases, приведено на рис. 6. Анализ DI позволяет оценить, какому типу отбора подвержены гены. При этом он корректно интерпретируется только в случае сравнения последовательностей анализируемых генов с ортологичными генами эволюционно близких организмов. Для вычисления *dN/dS* последовательности генов человека сравнивали с последовательностями ортологичных генов у других гоминид; если ортологов было несколько, то в качестве DI использовали среднее значение *dN/dS*. Лишь для 38 из 1436 изученных нами генов были получены значения DI > 1 (девять из них приходятся на одну категорию – заболеваний, связанных с иммунной системой). Из данного распределения следует, что большинство генов, входящих в состав исследованных генных сетей, эволюционировало в режиме стабилизирующего отбора (DI < 1).

Представлялось интересным изучить взаимоотношение между PAI и DI для исследованных нами 80 генных сетей. Результаты этого анализа показаны на рис. 7 на одном графике, с учетом разбиения заболеваний по категориям.

Анализ показал, что между PAI и DI имеется большая и высокодостоверная корреляция ($r = 0.876$, $p\text{-value} < 1.8 \times 10^{-26}$), т. е. наблюдается зависимость между средним эволюционным возрастом генов в генных сетях

и уровнем их генетической изменчивости: чем меньше эволюционный возраст генов, тем больше уровень их генетической изменчивости. Это хорошо согласуется с тем, что эволюционно древние гены вовлечены в ключевые для функционирования организма процессы, на них наложено множество ограничений со стороны других генов, особенностей организации молекулярно-генетических систем и им не свойственна высокая изменчивость. Эволюционно молодые гены, напротив, обеспечивают адаптацию к современным условиям жизни, и у них более высокая изменчивость.

Заключение

Филостратиграфический анализ – современная методология, позволяющая на основании данных о сходстве генетических последовательностей и происхождении организмов оценить возраст генов в масштабе всего генома. Вместе с информацией о том, какому типу отбора подвержен ген как единица наследственности, результаты анализа дают возможность судить о роли тех или иных генов в эволюции генных сетей организма.

При анализе генных сетей из базы данных KEGG Pathway, Human Diseases выявлено несколько тенденций. Большинство генов, задействованных в исследованных генных сетях, эволюционировали в режиме стабилизирующего отбора (DI < 1). Обнаружена достоверная зависимость ($r = 0.876$, p-value$< 1.8 \times 10^{-26}$) между средним эволюционным возрастом генов в генных сетях и уровнем их генетической изменчивости: чем меньше эволюционный возраст генов, тем больше их уровень генетической изменчивости. Некоторые категории генных сетей значительно выделяются по доле эволюционно молодых и эволюционно древних генов. Наибольшая доля эволюционно молодых генов (65 %) отмечена в генных сетях, связанных с заболеваниями иммунной системы. Наибольшая доля эволюционно древних генов (88 %) обнаружена в генных сетях, описывающих формирование зависимостей человека от химических соединений, вызывающих привыкание.

Показано, что генные сети, ответственные за развитие инфекционных заболеваний, вызванных паразитами, достоверно обогащены эволюционно молодыми генами, а генные сети, ответственные за развитие специфических типов рака, – эволюционно древними генами. Такие результаты говорят об активном процессе адаптации иммунной системы человека к возникающим угрозам. Кроме того, гены, задействованные в заболеваниях, вызывающих привыкание к химическим соединениям, обладают минимальным числом замен, т. е. такие гены максимально консервативны. В этом направлении можно провести отдельную работу с расширением исходных сетей с помощью доступных на сегодняшний день классификаторов и баз данных.

Список литературы / References

Галактионов В.Г. Иммунология: учебник для студентов вузов, обучающихся по направлению 510600 «Биология» и биол. специальностям. М.: Академия, 2004.
[Galaktionov V.G. Immunology: a Guide for University Students Studying in Track 510600 “Biology” and Biological Specialties. Moscow: Academia Publ., 2004. (in Russian)]

Колчанов Н.А., Игнатьева Е.В., Подколодная О.А., Лихошвай В.А., Матушкин Ю.Г. Генные сети. *Вавиловский журнал генетики и селекции.* 2013;17(4/2):833-850.
[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)]

Степанов В.А. Эволюция генетического разнообразия и болезни человека. *Генетика*. 2016;52(7):852-864.
[Stepanov V.A. Evolution of genetic diversity and human diseases. *Russ. J. Genet*. 2016;52(7):746-756.]

Хаитов Р.М. Иммунология: учебник для студентов медицинских вузов. М.: ГЭОТАР-Медиа, 2016.
[Khaitov R.M. Immunology: a Guide for Students of Medical Universities. Moscow, 2016. (in Russian)]

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.

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.

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

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.

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. https://doi.org/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

Благодарности. Работа поддержана грантом РФФИ № 20-04-00885 А и бюджетным проектом № 0259-2021-0009.

Прозрачность финансовой деятельности. Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 14.01.2021. После доработки 20.01.2021. Принята к публикации 20.01.2021.