

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ  
УЧРЕЖДЕНИЕ «ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР  
ИНСТИТУТ ЦИТОЛОГИИ И ГЕНЕТИКИ СИБИРСКОГО ОТДЕЛЕНИЯ  
РОССИЙСКОЙ АКАДЕМИИ НАУК»

*На правах рукописи*

**ИВАНОВ РОМАН АРТЕМОВИЧ**

**ФИЛОСТРАТИГРАФИЧЕСКИЙ И  
ФИЛОТРАНСКРИПТОМНЫЙ АНАЛИЗ ГЕНОВ,  
СВЯЗАННЫХ С РАЗВИТИЕМ ОНКОЛОГИЧЕСКИХ  
ЗАБОЛЕВАНИЙ**

1.5.8. - математическая биология, биоинформатика  
(биологические науки)

**ДИССЕРТАЦИЯ**

на соискание ученой степени  
кандидата биологических наук

Научный руководитель:

д.б.н. Лашин Сергей Александрович

## Новосибирск - 2025

## Оглавление

|                                                                                                                      |     |
|----------------------------------------------------------------------------------------------------------------------|-----|
| Список используемых сокращений и терминов .....                                                                      | 3   |
| Введение .....                                                                                                       | 4   |
| 1. Обзор литературы .....                                                                                            | 13  |
| 1.1 Применение концепций эволюционной биологии в исследовании заболеваний человека .....                             | 13  |
| 1.2 Методы определения режима отбора на основе сравнения последовательностей .....                                   | 19  |
| 1.3 Филостратиграфический анализ .....                                                                               | 26  |
| 1.4 Программные инструменты филостратиграфии .....                                                                   | 39  |
| 1.4.1 fagin .....                                                                                                    | 39  |
| 1.4.4 GenEra .....                                                                                                   | 41  |
| 1.4.5 oggmap .....                                                                                                   | 43  |
| 1.4.6 Orthoscape .....                                                                                               | 44  |
| 1.5 Базы данных и биоинформатические ресурсы онкологических заболеваний .....                                        | 45  |
| 1.6 Молекулярно-генетические аспекты онкологических заболеваний .....                                                | 52  |
| 1.6.1 Ключевые признаки злокачественных заболеваний .....                                                            | 53  |
| 1.6.2 Сигнальные пути и нарушения их работы в онкологических заболеваниях .....                                      | 56  |
| 1.7 Заключение к литературному обзору .....                                                                          | 59  |
| 2. Материалы и методы .....                                                                                          | 61  |
| 2.1 Материалы исследования .....                                                                                     | 61  |
| 2.2 Методы .....                                                                                                     | 63  |
| 3. Результаты .....                                                                                                  | 68  |
| 3.1 Разработка веб-сервиса Orthoweb 2 для филостратиграфического и филотранскриптомного анализа возраста генов ..... | 68  |
| 3.2 Конвейер sancEv для автоматизированного комплексного анализа данных экспрессии генов, из базы данных TCGA .....  | 73  |
| 3.3 Анализ дифференциальной экспрессии генов в злокачественных опухолях разных клинических стадий .....              | 76  |
| 3.4 Анализ обогащения терминами генной онтологии списков дифференциально экспрессирующихся генов .....               | 80  |
| 3.5 Филостратиграфический анализ дифференциально экспрессирующихся генов .....                                       | 82  |
| 3.6 Филотранскриптомный анализ клинических стадий карцином человека .....                                            | 90  |
| 3.7 Анализ генных сетей злокачественных заболеваний .....                                                            | 97  |
| Заключение .....                                                                                                     | 113 |
| Выводы .....                                                                                                         | 116 |
| Список используемой литературы .....                                                                                 | 117 |
| Приложение .....                                                                                                     | 141 |

## **Список используемых сокращений и терминов**

### **Сокращения:**

DI – Divergence Index (Индекс Изменчивости)

PAI – Phylostratigraphic Age Index (Филостратиграфический Индекс Возраста)

TAI – Transcriptome Age Index (Индекс Возраста Транскриптома)

TDI – Transcriptome Divergence Index (Индекс Изменчивости Транскриптома)

BRCA – Breast Invasive Carcinoma (Карцинома Молочной Железы)

KIRC – Kidney Renal Clear Cell Carcinoma (Светлоклеточная Карцинома Почки)

COAD – Colon Adenocarcinoma (Аденокарцинома Кишечника)

LIHC – Liver Hepatocellular Carcinoma (Гепатоцеллюлярная Карцинома Печени)

LUAD – Lung Adenocarcinoma (Аденокарцинома Лёгкого)

PRAD – Prostate Adenocarcinoma (Аденокарцинома Предстательной Железы)

UCEC – Uterine Carcinosarcoma (Карцинома Эндометрия)

CPM – Count Per Million (Число фрагментов на миллион картированных прочтений)

ГО – Генная Онтология (Gene Ontology)

ДЭГ – дифференциально экспрессирующиеся гены

KEGG – Kyoto Encyclopedia of Genes and Genomes (Киотская энциклопедия генов и геномов)

КО – KEGG Orthology (Ортологи KEGG)

RNA-seq – секвенирование РНК

LUCA – Last Universal Common Ancestor – последний универсальный общий предок

LCA – Last Common Ancestor – последний общий предок

## Введение

### Актуальность проблемы

Злокачественные новообразования продолжают оставаться одной из основных причин смертности в мире, что обуславливает непреходящую актуальность фундаментальных исследований механизмов онкогенеза. Наряду с интенсивным изучением молекулярных и клинических аспектов этих заболеваний, в последние годы существенно возрос интерес к пониманию роли эволюции генов в возникновении и прогрессировании злокачественных опухолей [Greaves, Maley, 2012; Merlo и др., 2006; Pepper и др., 2009; Dujon et al., 2021; Gillies, Verduzco, Gatenby, 2012; Maley et al., 2017; Mina et al., 2017]. Анализ происхождения генов и метаболических путей, ассоциированных с онкопатологиями, открывает новые возможности для объяснения причин приобретения генами онкогенных свойств и особенностей поведения опухолевых клеток [Davies, Lineweaver, 2011]. В этом контексте эволюционный подход, опирающийся на принципы молекулярной филогенетики и сравнительной геномики, позволяет выявлять фундаментальные механизмы, лежащие в основе формирования отличительных черт злокачественных клеток.

Одним из относительно новых подходов к анализу молекулярной эволюции генов является филостратиграфический анализ, в ходе которого каждому гену сопоставляется наиболее ранний идентифицируемый предок на таксономическом древе. Это позволяет отнести ген к определённой филострате, описываемой *индексом филостратиграфического возраста* (Phylostratigraphic Age Index – PAI), в которой они предположительно возникли [Domazet-Lošo, Brajković, Tautz, 2007c; Domazet-Lošo, Tautz, 2008b]. Этот подход показал, что гены, вовлеченные в фундаментальные клеточные

процессы, часто восходят к ранним эволюционным эпохам, отражая консервативность ключевых биологических функций.

На основе филостратиграфического анализа в работах Domazet-Lošo и соавторов [Domazet-Lošo, Brajković, Tautz, 2007c], а также Drost, Grosse и Quint [Quint и др., 2012b] были разработаны специальные метрики, которые позволяют количественно охарактеризовать вклад генов разного «возраста» в формирование транскриптомного профиля (с учётом уровней их экспрессии). Первым таким показателем является *индекс возраста транскриптома* (Transcriptome Age Index – TAI), который отражает средний эволюционный возраст экспрессируемых в определённых условиях генов. Вторым показателем, *индекс дивергенции транскриптома* (Transcriptome Divergence Index – TDI), описывает среднюю степень дивергенции транскриптома, указывая на то, насколько увеличивается активность более консервативных или, напротив, менее консервативных генов [Quint и др., 2012b].

Филостратиграфический анализ был применен и в изучении патологических процессов, в том числе онкологических заболеваний [Domazet-Lošo, Tautz, 2008b]. При этом было показано, что некоторые гены, вовлеченные в развитие онкологических заболеваний, имеют древнее эволюционное происхождение. Дополнительные доказательства различий в эволюционных характеристиках генов, вовлеченных в онкогенез, были представлены в работе [Zhang et al., 2019b], где с помощью филостратиграфического анализа сетей ко-экспрессии генов в клетках карциномы яичника было показано, что функциональные модули, связанные с опухолевой прогрессией, не являются равномерно распределёнными по индексам возрастов генов. Авторы обнаружили, что кластеры «древних» генов определяют ключевые метаболические и сигнальные пути, необходимые для поддержания базовых клеточных функций, тогда как более «молодые» гены способствуют тонкой настройке взаимодействий с микроокружением опухоли.

В контексте изучения опухолей, индексы TAI и TDI позволяют выявлять характерные паттерны экспрессии генов на разных стадиях патологического процесса. В частности, они отражают смещение равновесия между активностью «древних» и «молодых» генов при онкологических заболеваниях. Так, было показано, что несмотря на высокую гетерогенность, разные опухоли демонстрируют повторяемый сдвиг транскриптомов к сходным режимам экспрессии, что приводит к ограниченному набору устойчивых клеточных состояний. Филогенетический анализ в данном случае указывает, что этот сдвиг обусловлен усиленной активностью древних генетических программ в сочетании с подавлением функций генов, появившихся в ходе эволюции многоклеточности, что согласуется с идеей о функциональной специализации генов разного возраста в онкогенезе [Chen, He, 2016a].

Эта гипотеза также была подтверждена филотранскриптомным анализом широкого ряда солидных опухолей, который показал, что переход нормального клеточного состояния к злокачественному сопровождается сдвигом в экспрессии генов: активируются эволюционно «древние» гены, участвующие в процессах, возникших у одноклеточных организмов, в то время как экспрессия генов, связанных с многоклеточными регуляторными процессами, подавляется [Trigos и др., 2017].

Таким образом, филотранскриптомный анализ дополняет филогенетический анализ, предоставляя количественные метрики для исследования не только времени происхождения генов, но и их реальной функциональной активности в контексте эволюции и патологии.

Однако, несмотря на значительный прогресс, в данной области сохраняются существенные пробелы. В частности, остается недостаточно изученной динамика филотранскриптомных показателей (TAI/TDI) на различных стадиях клинического развития опухолей. Большинство существующих работ анализируют опухоли без детализации по стадиям, что

ограничивает понимание эволюционной перестройки транскриптома в процессе прогрессирования заболевания.

Исходя из вышеописанного, была сформулирована цель данной работы и ряд задач, позволяющие ее достичь.

**Целью** работы является филогенетический и филотранскриптомный анализ генов, вовлечённых в патогенез онкологических заболеваний человека.

Для достижения этой цели были поставлены следующие **задачи**:

1. Разработка программных средств для филогенетического и филотранскриптомного анализа генных сетей, а также для автоматизированной обработки и анализа данных экспрессии генов из опухолевых образцов, полученных на различных клинических стадиях;
2. Выявление белок-кодирующих генов, дифференциально экспрессирующихся на разных стадиях развития злокачественных опухолей;
3. Исследование макро- (PAI) и микроэволюционных (DI) характеристик генов, дифференциально экспрессирующихся в тканях опухолей разных органов человека, с использованием филогенетического анализа;
4. Исследование филотранскриптомных характеристик (TAI и TDI) транскриптомов на разных стадиях развития опухолей человека;
5. Анализ эволюционных характеристик генов, входящих в генные сети онкологических заболеваний человека из баз данных KEGG PATHWAY и WikiPathways.

### **Научная новизна**

В рамках данной кандидатской диссертации с использованием методов филогенетики и филотранскриптомного анализа был проведён комплексный анализ эволюционных характеристик генов, дифференциально

экспрессирующихся в различных клинических стадиях девяти распространенных карцином человека. Впервые определены индексы возраста генов, которые проявляют дифференциальную экспрессию на различных стадиях патологического развития ряда карцином человека. Выявлено статистически значимое преобладание как эволюционно «молодых» генов (Vertebrata, PAI = 6), так и очень «древних» (Cellular Organism, PAI = 1) среди дифференциально экспрессирующихся генов. Была разработана методика филотранскриптомного анализа, позволяющая оценивать эволюционную динамику экспрессии генов в развитии онкологических заболеваний. Филотранскриптомный анализ показал, что в здоровых тканях средний возраст транскриптома существенно выше, чем в опухолевых образцах, а для карциномы печени обнаружено особое повышение TAI, связанное с активностью генов метаболизма лекарственных средств. Нелинейный паттерн изменения TAI при прогрессировании опухоли, статистически подтверждённый для карцином лёгкого, молочной железы, печени и мочевого пузыря, впервые демонстрирует увеличение активации молодых генов на ранних и поздних стадиях при одновременном росте экспрессии «древних» генов на промежуточных стадиях. Филостратиграфический анализ ключевых сигнальных путей, вовлечённых в развитие злокачественных новообразований, показал различие значений PAI между генами рецепторов и лигандных молекул с генами, обеспечивающими передачу сигнала внутри клетки.

С использованием современных технологий была разработана новая версия веб-сервиса – Orthoweb 2, обеспечивающая расчёт эволюционных индексов на основе ортологов и гомологии последовательностей, а также интеграцию и визуализацию генных сетей из внешних источников. Программный комплекс Orthoweb 2 является единственным в мире на сегодняшний день программным комплексом, позволяющим проводить филостратиграфический анализ как отдельных генов, так и генных сетей, не требуя ни знания языков программирования, ни установки ресурсоёмких

программ типа Cytoscape. Важным усовершенствованием является реализованная способность к интеграции в процесс анализа эволюционного анализа генных сетей дополнительных данных, включая функциональную аннотацию генов (GO), информацию об однонуклеотидных полиморфизмах (SNP) и данные об экспрессии генов.

### **Теоретическая и практическая значимость**

Обнаруженные закономерности распределений индексов возрастов транскриптомов расширяют представления о динамике эволюционного давления в ходе прогрессии опухолей и может служить концептуальной основой для разработки стадий-специфичных биомаркеров. Филостратиграфический анализ сигнальных путей, связанных с злокачественными заболеваниями, позволяет различать консервативные мишени, жизненно важные для клеточной выживаемости, и более молодые регуляторы, потенциально уязвимые для таргетной терапии, тем самым открывая новые направления персонализированного лечения.

Программный комплекс Orthoweb 2 обеспечивает общедоступную платформу для интегрального эволюционного анализа генов и генных сетей, что существенно упрощает внедрение филостратиграфических и филотранскриптомных метрик в рутинные исследования молекулярной онкологии. Возможность объединять данные о возрасте генов с профилями экспрессии, SNP и функциональной аннотацией повышает точность интерпретации результатов и способствует формированию эволюционно обоснованных стратегий диагностики и терапии онкологических заболеваний.

### **Вклад автора**

Основная часть работы выполнена автором самостоятельно. Автор принимал участие в постановке задач, разработке конвейера программ, проведении вычислительных экспериментов, анализе данных, обсуждении полученных результатов.

### **Положения, выносимые на защиту**

1. Разработан программный комплекс Orthoweb 2, предназначенный для расчета эволюционных характеристик генов – филостратиграфических и филотранскриптомных индексов на основе транскриптомных данных, примененный для анализа генов дифференциально экспрессирующихся на различных стадиях развития карцином человека.
2. Эволюционный возраст дифференциально экспрессирующихся генов, вовлеченных у человека в развитие карциномы лёгких, печени, простаты, мочевого пузыря и молочной железы, различается на разных стадиях развития опухоли – на ранних и поздних стадиях преобладают эволюционно «молодые» гены, на промежуточных – эволюционно «древние».

### **Апробация работы**

Материалы работы вошли в отчёты по гранту РФФИ (№20-04-00885). Результаты диссертации были доложены на научных конференциях: 13-я международная конференция Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022), 4-8 July 2022. Novosibirsk, Russia; Системная биология и биоинформатика (SBB-2023); 14-я международная школа молодых ученых, 22-26 мая 2023 г., Новосибирск, Россия.; 14-я международная конференция Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2024), 5-10 August 2024; Марчуковские научные чтения, 4-8 октября 2021 г., Новосибирск, Россия.

### **Публикации**

По материалам диссертации опубликовано 7 работ, из них 3 статьи в рецензируемых научных журналах, входящих в перечень ВАК, 4 тезиса конференций, а также одно авторское свидетельство.

Статьи в рецензируемых журналах, входящих в перечень ВАК:

1. **Иванов Р.А.**, Лашин С.А. Внутриопухолевая гетерогенность: модели возникновения и эволюции злокачественных опухолей // Вавиловский журнал генетики и селекции. – 2023. – Т. 27, № 7. – С. 815–819.
2. **Иванов Р.А.**, Мухин А.М., Казанцев Ф.В., Мустафин З.С., Афонников Д.А., Матушкин Ю.Г., Лашин С.А. Orthoweb: программный комплекс для эволюционного анализа генных сетей // Вавиловский журнал генетики и селекции. – 2024. – Т. 28, № 8. – С. 874–881.
3. **Ivanov R.A.**, Afonnikov D.A., Matushkin Y.G., Lashin S.A. Evolutionary transcriptomics of cancer development // International Journal of Molecular Sciences. – 2025. – Vol. 26, № 11. – P. 5041.

Авторские свидетельства:

1. Лашин С.А., Мустафин З.С., Казанцев Ф.В., Мухин А.М., **Иванов Р.А.**, Матушкин Ю.Г., Афонников Д.А. Программа для расчёта филостратиграфических индексов (Ортовеб), Свидетельство № 2024618826, 17 апреля 2024

Тезисы:

1. **Иванов Р.А.**, Лашин С.А. Филостратиграфический анализ генов онкологических заболеваний человека // 14th International young scientists school «System Biology and Bioinformatics» (SBB-2023) (22–26 мая 2023 г., Новосибирск, Россия): Тезисы докладов. – Новосибирск, 2023. – С. 12.
2. **Ivanov R.**, Mustafin Z., Lashin S. Phylostratigraphic analysis of human cancers transcriptomic data // BGRS/SB-2022: 13th International Multiconference on “Bioinformatics of Genome Regulation and Structure/Systems Biology” (4–8 July 2022, Novosibirsk, Russia): Abstracts. – Novosibirsk, 2022. – P. 792.

3. **Ivanov R.**, Lashin S., Afonnikov D., Matushkin Yu. Integration of phylostratigraphic and phylotranscriptomic analysis in the study of differentially expressed genes in cancer tissues // BGRS/SB-2024: 14th International Multiconference on “Bioinformatics of Genome Regulation and Structure/Systems Biology” (5–10 August 2024, Novosibirsk, Russia): Abstracts. – Novosibirsk, 2024. – P. 801.
4. **Иванов Р.А.**, Афонников Д.А., Матушкин Ю.Г., Лашин С.А. Эволюционный подход к изучению дифференциальной экспрессии в тканях карцином различных стадий // VI Международная конференция «ПОСТГЕНОМ–2024» (29 октября – 2 ноября 2024 г., Москва, Россия): Сборник тезисов докладов. – Москва, 2024. – С. 17.

### **Структура и объем работы**

Работа состоит из введения, обзора литературы, описания материалов и методов, результатов, заключения, выводов и списка литературы. Работа изложена на 144 страницах, содержит 25 иллюстрации, 6 таблиц и 3 приложения. Список литературы включает 277 ссылок.

### **Благодарности.**

Автор выражает благодарность и глубокую признательность научному руководителю д.б.н. Лашину С.А., а также д.б.н. Афонникову Д.А., к.б.н. Землянской Е.В., к.б.н. Клименко А.И. и к.б.н. Казанцеву Ф.В. за консультации и плодотворные научные дискуссии, а также академику Колчанову Н.А. и всем сотрудникам Отдела Системной Биологии ИЦиГ СО РАН.

## 1. Обзор литературы

### 1.1 Применение концепций эволюционной биологии в исследовании заболеваний человека

Многофакторные заболевания (МФЗ), такие как диабет, болезнь Крона, ишемическая болезнь сердца, инсульт, различные злокачественные новообразования (включая новообразования лёгкого, толстой кишки, простаты и молочной железы), шизофрения и биполярное расстройство, представляют собой значительную проблему для современной медицины. Комплексность этих заболеваний обусловлена взаимодействием множества генетических факторов, а также влиянием окружающей среды. В отличие от моногенных заболеваний, таких как болезнь Хантингтона [Li, Friedman, Li, 2007; Tabrizi и др., 2020], муковисцидоз [O’Neal, Knowles, 2018] или наследственный гемохроматоз [Chokoshvili, Vears, Borry, 2018; Edwards и др., 2015], где генетическая основа хорошо изучена, многофакторные заболевания все еще не полностью поняты с точки зрения их генетических механизмов.

Классические подходы к изучению МФЗ, как правило, сосредоточены на анализе молекулярно-биологических и клинических аспектов. Тем не менее, применение эволюционных концепций в анализ МФЗ предоставляет возможность для более глубокого понимания этих заболеваний, расширяя представления о причинах их появления, распространения и устойчивости среди человеческих популяций [Benton и др., 2021].

Важным направлением исследований заболеваний является изучение адаптивных изменений в человеческом геноме, зафиксировавшихся под влиянием изменений окружающей среды [Carroll и др., 2014; Timpson и др., 2018]. Это позволяет не только лучше понять процессы формирования генетического разнообразия в современных популяциях и оценить влияние этих изменений на здоровье и развитие заболеваний, но также может помочь выявить «упущенную наследуемость» при многофакторных заболеваниях.

Такой подход способствует глубокому пониманию природы МФЗ и может сыграть важную роль в разработке новых стратегий диагностики, профилактики и лечения [Степанов, 2016].

Область медицины, изучающая заболевания через призму эволюции, называется *эволюционной медициной*. Эволюционный подход к медицине основан на идее о том, что болезни человека возникают в результате ограничений, компромиссов, несоответствий и конфликтов, присущих сложным биологическим системам, которые взаимодействуют с изменчивой окружающей средой посредством естественного отбора. Эволюционная медицина выделяет несколько ключевых принципов для объяснения возникновения сложных генетических заболеваний [Benton и др., 2021].

Первый принцип основывается на понимании того, что естественный отбор не стремится создать идеальные организмы. Вместо этого, он направлен на увеличение репродуктивной пригодности в рамках ограничений, накладываемых законами физики, а также доступностью и взаимодействием уже существующих биологических вариаций. Эти факторы формируют или ограничивают дальнейшее развитие видов [Futuyma, 2010].

Второй принцип касается несоответствия между генетическим наследием и условиями современной жизни. Изменения в образе жизни, такие как недостаток физической активности и неправильное питание, приводят к распространению заболеваний, таких как ожирение, диабет и сердечно-сосудистые заболевания. Примером может служить адаптация к условиям, бедным калориями, в прошлом, которая способствовала развитию генетических вариантов, эффективно использующих метаболические ресурсы. В условиях современного изобилия калорий это может привести к ожирению [Fay, 2013; Speakman, 2006].

Третий принцип тесно связан с концепцией биологических компромиссов, которая глубоко укоренена в дилемме Холдейна о невозможности одновременной оптимизации всех характеристик из-за ограничений, налагаемых естественным отбором. Суть дилеммы заключается в том, что для

каждой адаптивной черты существует своя цена или компромисс. Например, разнообразие физических характеристик организма напрямую зависит от его общих энергетических ресурсов, и улучшение одного аспекта может происходить за счёт другого [Stearns, 1989]. Это можно увидеть на примере плейотропных генетических вариантов, которые оказывают влияние на несколько физиологических систем одновременно, способствуя возникновению подобных компромиссов. Примечательно, что некоторые симптомы, традиционно воспринимаемые как проявления болезни, могут на самом деле быть адаптивными реакциями организма. Примерами таких реакций могут служить: повышенная температура [Evans, Repasky, Fisher, 2015] или утреннее недомогание при беременности [Almond и др., 2016]. Лихорадка, традиционно считающаяся симптомом инфекции или болезни, на самом деле является адаптивной реакцией, призванной сделать среду организма менее благоприятной для патогенных микроорганизмов. Повышение температуры тела может усилить иммунную функцию и подавить рост некоторых бактерий и вирусов. Утренняя же тошнота, часто встречающаяся на ранних сроках беременности, является адаптивной реакцией, защищающей плод от токсинов и патогенов в критические периоды развития. Это подтверждается тем, что утренняя тошнота связана с меньшим риском выкидыша [Almond и др., 2016].

Работы в области эволюционной медицины можно разделить на два основных направления: изучение макроэволюционного влияния на болезни человека и изучение микроэволюционных адаптаций и типа естественного отбора.

Исследование влияния макроэволюции основано на понимании того, что ключевые биологические системы и процессы, необходимые для жизнедеятельности клеток – такие как репликация ДНК, транскрипция и трансляция – являются результатом ранних эволюционных инноваций, которые присущи всем живым организмам. Несмотря на своё

фундаментальное значение, эти инновации создали предпосылки для возникновения множества современных заболеваний.

Эволюционный переход к многоклеточности, вероятно происходивший неоднократно на протяжении истории живой природы [Brunet, King, 2017; Knoll, 2011], ярко демонстрирует связь между эволюционными новшествами и появлением болезней [Rokas, 2008]. Появление многоклеточных организмов позволило сформировать сложные структуры тела из множества клеток, что стало возможным благодаря развитию способности клеток регулировать клеточные циклы, контролировать рост и образовывать сложные коммуникационные сети. Однако, вместе с этим, многоклеточность подготовила почву для возникновения онкологических заболеваний [Albuquerque и др., 2018]. Например, бесконтрольная пролиферация – адаптация, критически важная для выживания простейших, – в контексте многоклеточности требует строгой регуляции, нарушение которой напрямую ведет к неоплазии.

Гены, отвечающие за контроль над клеточным циклом, обычно классифицируют на две категории: «гены-смотрители» и «гены-привратники» [Michor, Iwasa, Nowak, 2004]. Гены-смотрители играют ключевую роль в базовом контроле над клеточным циклом и репарацией ДНК. Мутации в этих генах часто приводят к повышению частоты других мутаций или нестабильности генома, что, в свою очередь, увеличивает риск развития опухоли. Функции генов-смотрителей уходят корнями в самое начало появления клеточной жизни [Domazet-Lošo, Tautz, 2010b]. Гены-привратники, появившиеся позже в ходе эволюции многоклеточных организмов [Domazet-Lošo, Tautz, 2010b], напрямую ассоциированы с процессами опухолеобразования благодаря их роли в регуляции роста, апоптоза (программируемой гибели клеток) и взаимодействия клеток.

Понимание процессов роста и созревания опухолей у отдельных пациентов также требует эволюционного подхода. Разработка стратегий лечения, которые учитывают эволюцию устойчивости к лекарствам и

гетерогенность опухолей, стала одним из основополагающих принципов современной онкотерапии. Это подчёркивает важность эволюционного изучения биологии злокачественных новообразований для создания более эффективных методов лечения [Enriquez-Navas, Wojtkowiak, Gatenby, 2015; Gerlinger и др., 2014; Gerstung и др., 2020].

Макроэволюционные процессы, о которых шла речь выше, заложили основу для возникновения генетических нарушений и заболеваний. Однако, для глубокого понимания болезней, необходимо уделить внимание и более поздним изменениям, произошедшим в ходе эволюции человека. Если сравнить людей с их ближайшими эволюционными родственниками по линии приматов, например, с шимпанзе, мы обнаруживаем заболевания, в частности болезнь Альцгеймера [Berbée и др., 2011], микроцефалия [Mefford, Eichler, 2009] или мышечная дистрофия Дюшенна [Varki, 2010], которые либо отсутствуют у других видов, либо имеют существенно отличное течение [O'Bleness и др., 2012]. Анализ типов естественного отбора, воздействующего на определённые биологические процессы и связанные с ними гены, позволяет глубже понять, как недавние эволюционные изменения в человеке повлияли на развитие заболеваний.

Эти типы естественного отбора включают:

1. Движущий отбор, который способствует распространению одного варианта гена в популяции. Этот тип отбора способствует распространению адаптивных мутаций, которые увеличивают приспособленность организмов к их окружающей среде. Эти мутации могут улучшить способность организма выживать или размножаться, что приводит к их быстрой фиксации в популяции [Rasmussen и др., 2014].

2. Стабилизирующий отбор, который поддерживает определенные генетические вариации на уровне, близком к среднему для данной популяции, устраняя мутации из популяции. Это способствует сохранению стабильности определенных белков и генома в целом [Huang, Gulko, Siepel, 2017].

3. Балансирующий отбор – это тип естественного отбора, при котором генетическое разнообразие поддерживается в популяции. В отличие от движущего или стабилизирующего отборов, которые способствуют доминированию одного аллеля за счёт других, балансирующий отбор сохраняет несколько различных аллелей у определенного гена [Siewert, Voight, 2017].

Работы по анализу влияния и определению типа отбора для генов проводилось и в контексте онкологических заболеваний. Примером гена, ассоциированного с опухолевыми процессами и находящегося под действием стабилизирующего отбора, является ген *BRCA1*, мутации в котором связаны с ранним началом развития опухолей молочной железы [Pavard, Metcalf, 2007]. В нормальных условиях белок, кодируемый геном *BRCA1*, функционирует как часть комплекса, который помогает в точной и эффективной репарации ДНК, тем самым защищая клетки от генетических мутаций, способных инициировать злокачественную трансформацию.

Мутации в гене *BRCA1* могут нарушить эту защитную функцию, снижая способность клеток к репарации повреждений ДНК. Это приводит к накоплению мутаций в других критически важных генах, которые контролируют клеточный цикл, апоптоз и другие процессы, что может инициировать опухолевую трансформацию. Мутации в гене *BRCA1* значительно повышают риск злокачественных новообразований молочной железы и ассоциированы с увеличением риска опухолей фаллопиевых труб и поджелудочной железы [Al-Sukhni и др., 2008; Cass и др., 2005; Lee, Boyer, 2001; Li и др., 2022; Tai и др., 2007; Takaoka, Miki, 2018]. В то же время было показано, что ген *TRPV6*, ассоциированный с аденокарциномой простаты, подвергался действию движущего отбора в европейской популяции, что привело к появлению нескольких несинонимичных замен у людей с разным генетическим происхождением [Stajich, 2004]. Ген *UGT2B4*, участвующий в метаболизме стероидных гормонов, мутации в котором связаны с повышенным риском развития карциномы молочной железы, демонстрирует

признаки недавнего движущего и балансирующего отбора в африканской популяции [Sun и др., 2011]. Движущий отбор также был обнаружен для гена *PPP2R5E* [Grochola и др., 2009], который участвует в ингибировании клеточного роста и деления. Этот ген кодирует регуляторную субъединицу опухолеподавляющей протеинфосфатазы 2A и у него были обнаружены следы движущего отбора в некоторых популяциях [Grochola и др., 2009]. Наблюдаемый движущий отбор этого гена способствует уменьшению восприимчивости к саркоме мягких тканей у европеоидной популяции. Исследование эволюционных характеристик данного гена может способствовать выявлению вариантов гена, ассоциированных с повышенной восприимчивостью к данному заболеванию.

Подводя итог, можно сказать, что приведенные данные свидетельствуют о том, что заболевания и ассоциированные с ними гены подвергались воздействию эволюционных процессов и естественного отбора в ходе эволюции человечества.

## **1.2 Методы определения режима отбора на основе сравнения последовательностей**

Естественный отбор является ключевым механизмом эволюции, определяющим адаптацию организмов к изменяющимся условиям среды и формирующим генетическую структуру популяций. Его режимы играют различную роль в поддержании или трансформации генетического разнообразия. Сравнительный анализ нуклеотидных последовательностей – ключевой инструмент для выявления режимов естественного отбора, действующего на геномные элементы [Yang, 2007; Kryazhimskiy, Plotkin, 2008; Nei, 2000; Nielsen, Yang, 1998]. Основная задача таких методов заключается в дифференциации нейтральных эволюционных процессов от тех, на которые влияют стабилизирующий или движущий отбор [Hurst, 2002; Yang, 2006]. Современные подходы к решению этой задачи опираются на анализ замен в

кодирующих белки регионах, а также на сопоставление внутривидовой и межвидовой изменчивости [Eyre-Walker, Keightley, 2007; Sullivan, 1998; Codon evolution: mechanisms and models, 2012].

Одним из наиболее распространённых методов оценки селективного давления является расчёт отношения частот несинонимичных замен ( $K_a$  или  $d_N$ ) к частотам синонимичных замен ( $K_s$  или  $d_S$ ) на сайт [Masatoshi, Takashi, 1986; Yang, Nielsen, 2000; Li, 1993]. Этот подход основывается на структурных особенностях генетического кода: синонимичные замены, как правило, не влияют на аминокислотную последовательность белков в силу вырожденности генетического кода и считаются нейтральными, тогда как несинонимичные замены изменяют белок и подвержены действию отбора [Kimura, 1983; Pál, Papp, Hurst, 2001; Zhang, 2000]. Значение  $d_N/d_S < 1$  указывает на преобладание стабилизирующего отбора, который снижает адаптивный потенциал, но повышает устойчивость популяций [Masatoshi, Takashi, 1986; Bustamante et al., 2005; Stoletzki, Eyre-Walker, 2011]. Этот режим доминирует в стабильных экологических условиях, обеспечивая устойчивость белковых структур и метаболических путей [Krylov et al., 2003]. Например, гены, кодирующие рибосомальные белки [Bustamante et al., 2005; Drummond, Wilke, 2008; Subramanian, Kumar, 2004] или ферменты центрального метаболизма [Wall et al., 2005; Zhang, 2005], часто подвергаются стабилизирующему отбору, что проявляется в низких значениях  $d_N/d_S$ . В то же время отношение  $d_N/d_S$  значимо больше единицы свидетельствует о движущем отборе, способствующем фиксации адаптивных изменений [Yang, 2007; Nielsen, Yang, 1998; Smith, Eyre-Walker, 2002]. Этот отбор часто наблюдается у генов, участвующих во взаимодействии с окружающей средой и, в частности, в регуляции иммунной системы, (например, гены, кодирующие белки главного комплекса гистосовместимости – МНС – Major histocompatibility complex у млекопитающих или дефензины у растений) [Bustamante et al., 2005; Jacquet et al., 2019; Sawyer et al., 2005].

Отношение  $dN/dS$ , близкое к 1, свидетельствует о нейтральной эволюции гена, то есть скорость несинонимичных замен примерно равна скорости синонимичных замен. Это отношение означает, что на ген в меньшей степени действует давление отбора. Гены, подвергающиеся нейтральной эволюции, часто связаны с функциями, которые не являются критически важными для выживания или размножения организмов, вследствие чего они могут быть менее ограничены отбором. К примерам таких генов можно отнести псевдогены, гены, возникшие после недавних генных дупликаций или гены, которые утратили свою биологическую роль в результате изменений в среде обитания или метаболизме организма, как, например, гены обонятельных рецепторов человека [Gilad et al., 2004; Gilad, Man, Glusman, 2005].

Существует несколько различных методов для расчёта значений этого отношения, отличающихся разными способами учета позиций триплетов или их весовых коэффициентов. Ранние подходы к оценке основывались на попарном сравнении последовательностей:

**Метод LWL85** при расчёте отношения  $d_N/d_S$  использует упрощенную модель для оценки числа несинонимичных ( $D_N$ ) и синонимичных ( $D_S$ ) замен, предполагая равномерную скорость замен между сайтами. Коррекция множественных замен выполняется через линейную аппроксимацию [Li, Wu, Luo, 1985]:

$$d_N = \frac{D_N}{n} * \frac{1}{1 - \frac{4}{3}p_S}, d_S = \frac{D_S}{s} * \frac{1}{1 - \frac{4}{3}p_S},$$

где  $n$  и  $s$  – число несинонимичных и синонимичных сайтов. Метод недооценивает  $d_S$  при высокой дивергенции ( $d_S > 1$ ) из-за насыщения синонимичных замен

**Метод LWLm** является модификацией LWL85. Он заменяет линейную коррекцию на нелинейную модель, учитывающую вероятности множественных замен для каждого типа сайта. Это повышает точность анализа при высоких значениях  $d_S > 2$  [Li, 1993].

$$d_S = \frac{(L_2A_2 + L_4A_4)}{L_2 + L_4} + B_4, d_n = A_0 + \frac{(L_0B_0 + L_2B_2)}{L_0 + L_2}$$

где  $L_i$  – это число сайтов с вероятностью множественных замен:  $L_0$  – все замены меняют аминокислоту,  $L_2$  – одна из замен сохраняет аминокислоту (двукратное вырождение) и  $L_4$  все замены сохраняют аминокислоту (четырёхкратное вырождение).  $A_i$  – число транзиций, а  $B_i$  – число трансверсий на сайт  $i$ .

Транзиции и трансверсии – это типы мутаций, заключающиеся в замене нуклеотидных оснований в молекуле ДНК: при транзиции одно пуриновое основание заменяется на другое пуриновое, а пиримидиновое на другое пиримидиновое, тогда как при трансверсии – пуриновое основание заменяется на пиримидиновое и наоборот.

**Метод LPB93** объединяет идеи LWL85 с учетом вариабельности замен между кодонными позициями. Метод разделяет оценку  $d_N$  и  $d_S$  для каждой позиции триплета, что снижает систематическую ошибку при анализе генов со смещением частот кодонов [Li, 1993; Pamilo, Bianchi, 1993].

**Метод Нея-Годжобори (Nei-Gojobori)** предполагает ряд допущений относительно равной и постоянной вероятности нуклеотидных замен на протяжении всей последовательности гена, следуя модели Джукса-Кантора о равновероятности нуклеотидных замен [Masatoshi, Takashi, 1986].

$$d_N = -\frac{3}{4} \ln \left( 1 - \frac{4}{3} p_N \right), d_S = -\frac{3}{4} \ln \left( 1 - \frac{4}{3} p_S \right),$$

где  $p_N$  и  $p_S$  – соотношение предсказанного числа синонимичных и несинонимичных замен в последовательностях со средним числом синонимичных и несинонимичных замен.

Также этот метод подразумевает анализ только близкородственных последовательностей, в которых множественные замены в одном локусе маловероятны. В силу описанных выше допущений, и в силу того, что в реальных биологических системах эти допущения часто не соблюдаются, этот метод может давать менее точные оценки во многих ситуациях, таких как: неравная вероятность нуклеотидных замен [Kosakovsky Pond, Frost, 2005], непостоянная скорость замен во времени, зависимость скорости замен от

положения сайта в кодоне [Yang et al., 2000], а также эпистаз [Pollock, Thiltgen, Goldstein, 2012] и гетерогенное селективное давление внутри гена [Yang, 1998].

Ключевое преимущество методов попарного сравнения – вычислительная простота и минимальные требования к данным: для анализа достаточно двух последовательностей. Однако эти методы игнорируют филогенетические связи и предполагают равномерную скорость замен между сайтами. Это приводит к систематическим ошибкам при высокой дивергенции (например, из-за насыщения синонимичных сайтов) и делает их неприменимыми для анализа сложных сценариев отбора, таких как случаи эпизодического положительного отбора [Zhang, Yang, Nielsen, 2005].

Для решения указанных ограничений были разработаны кодонные и филогенетические модели. Кодонные модели преодолевают ограничения попарных методов, рассматривая эволюцию на уровне кодонов и используя методы максимального правдоподобия:

**Голдман и Янг (Goldman and Yang)** разработали метод максимального правдоподобия для оценки отношения  $d_N/d_S$  между двумя последовательностями, с учетом поправки на разные вероятности трансверсии и транзиции [Goldman, Ziheng, 1994]. Учет различий в вероятности этих мутаций позволяет методу Голдмана-Янга более точно оценивать отношение  $d_N/d_S$ , отражая адаптивные и консервативные силы, действующие на гены.

**Янг и Нильсон (Yang and Nielson)** разработали модификацию метода Голдмана-Янга, в которой учитывалась поправка на разные вероятности мутаций в различных нуклеотидных позициях в кодонах [Yang, Nielsen, 2000].

**Мьюз и Гаут (Muse & Gaut)** разработали модель которая упрощает расчёты через предположение о том, что синонимичные замены происходят с одинаковой скоростью для всех кодонов. Однако при сильной предвзятости кодонов модель недооценивает  $d_N/d_S$ , так как не учитывает вариации в скоростях синонимичных замен [Muse, Gaut, 1994].

Для расчёта отношения  $d_N/d_S$  при помощи попарных и кодонных моделей было разработано большое число программных инструментов. В частности стоит упомянуть программные комплексы *MEGA* [Kumar et al., 2018] и *KaKs Calculator* [Zhang, 2022]. *MEGA* (Molecular Evolutionary Genetics Analysis) предоставляет графический интерфейс для расчёта  $d_N/d_S$  с использованием попарных методов, таких как метод *Нея-Годжобори* и в то же время инструменты для реконструкции филогенетических деревьев. *KaKs Calculator* реализует все выше перечисленные методы попарного сравнения и кодонных моделей, что позволяет сравнивать результаты различных моделей на одних данных, что полезно для проверки надёжности полученных результатов.

Филогенетические методы, реализованные в программных инструментах *codeml* (PAML) и *HyPhy*, расширяют кодонные подходы, интегрируя информацию о родстве видов через деревья. Это позволяет учитывать историю эволюции через нахождение общих предковых последовательностей и обнаруживать режим отбора, действующий на отдельные ветви филогенетических деревьев.

PAML, разработанный Ziheng Yang, реализует кодонные модели для анализа отношения  $d_N/d_S$  с использованием методов максимального правдоподобия [Álvarez-Carretero, Kapli, Yang, 2023; Yang, 2007]. Основной компонент пакета, связанный с оценкой режима отбора – *codeml*, позволяет оценивать параметр  $d_N/d_S$  в контексте филогенетических деревьев. Программа поддерживает несколько типов моделей, включая «модели сайтов», такие как *M7* и *M8*, которые учитывают вариабельность  $d_N/d_S$  между кодонами, где для работы *M7* предполагается, что все сайты подвержены нейтральному или стабилизирующему отбору, в то время как для работы *M8* делается предположение, что ряд кодонов может быть подвержен движущему отбору. Помимо этого, в PAML входят «модели ветвей», предназначенные для проверки гипотез о различиях в  $d_N/d_S$  в отдельных ветвях филогенетического дерева, а также «модели ветвей и сайтов», которые позволяют выявить

кодона, подвергающиеся движущему отбору в определенных ветвях филогенетических деревьев [Álvarez-Carretero, Kapli, Yang, 2023; Yang, 2007].

Работа с PAML требует ручной подготовки входных данных, включая задание топологии филогенетических деревьев, и сопряжена с высокими требованиями к вычислительным ресурсам.

HyPhy (Hypothesis Testing Using Phylogenies) представляет собой открытый программный пакет, разработанный для анализа молекулярной эволюции с фокусом на изучение естественного отбора [Pond, Frost, Muse, 2005]. Его ключевыми особенностями являются поддержка пользовательских моделей замен, интеграция с филогенетическими деревьями и возможность тестирования сложных гипотез о режимах отбора [Kosakovsky Pond et al., 2020].

Методы оценки режимов естественного отбора играют важную роль в понимании эволюционных процессов, связывая генетические изменения с адаптацией организмов. Их развитие от простых попарных сравнений к сложным филогенетическим моделям отражает стремление учесть биологическую реальность, где мутации, отбор и филогенетические связи видов взаимодействуют нелинейно. Каждый подход имеет свои ограничения: ранние методы, несмотря на вычислительную эффективность, упрощают ключевые аспекты эволюции генов, такие как вариация скорости замен по сайтам и филогенетическим линиям, наличие эпистаза или эпизодического отбора, тогда как современные филогенетические модели требуют значительных ресурсов и экспертизы в интерпретации результатов. Выбор метода оценки режима естественного отбора зависит от целей исследования, качества данных и гипотез, что подчёркивает важность их всесторонней оценки. Эти методы не только углубляют понимание прошлых адаптаций, но и служат инструментами для прогнозирования эволюционных ответов на глобальные изменения среды.

### 1.3 Филостратиграфический анализ

Развитие высокопроизводительных методов секвенирования и компьютерных технологий в XXI веке привели к качественному скачку в возможностях компьютерного анализа эволюционных процессов, открыв новые направления исследований во всех областях биологии. Одним из методов эволюционного анализа, ставшим возможным благодаря этим техническим достижениям, стал предложенный в середине 2000-х годов Томиславом Домазетом-Лошо (Tomislav Domazet-Lošo) и Дитхардом Таутцем (Diethard Tautz) филостратиграфический анализ. Филостратиграфия является важным инструментом в области эволюционной биологии, позволяющим выяснить происхождение и разнообразие семейств генов. Истоки этого метода восходят к пионерской работе Цукеркандля и Полинга (Zuckerkandl and Pauling) [Zuckerkandl, Pauling, 1965], предложивших гипотезу молекулярных часов, согласно которой накопление мутаций в ДНК и соответствующие изменения аминокислотных последовательностей в белках происходят с относительно постоянной скоростью. Это позволяет приблизительно «датировать» момент появления гена или время расхождения филогенетических линий, исходя из степени различий между сходными участками геномов у разных видов. Эта концепция расчёта временных интервалов на основе постоянной скорости эволюционных изменений легла в основу филостратиграфии, которая впоследствии была формализована Т. Домазетом-Лошо и Д. Таутцем [Domazet-Lošo, 2003]. Они разработали систематический подход, позволяющий проследить приблизительное время появления генов по их появлению в филогенетических линиях, что дало основу для понимания эволюционной истории генов.

Целью филостратиграфического анализа является определение времени возникновения гена-основателя на основе оценки распределения ортологичных ему генов в геномах организмов, принадлежащих к различным таксономическим группам. Под ортологами в контексте филостратиграфии понимают гены в разных видах, которые произошли от общего предкового

гена в результате видообразования. Они часто сохраняют сходную функцию в ходе эволюции и могут выполнять одни и те же биологические роли в разных организмах [Gabaldón, Koonin, 2013; Koonin, 2005]. Паралоги – это гены, возникшие в результате дупликации исходного гена в пределах одного генома, которые в дальнейшем эволюционируют независимо в рамках одного вида или близкородственных линий. В отличие от ортологов, которые разделяются вследствие видообразования и часто сохраняют сходные функции, паралоги могут приобретать новые биологические роли или специализироваться в выполнении отдельных задач, что связано с отсутствием эволюционного давления, поддерживающего их исходную структуру [Gabaldón, Koonin, 2013; Koonin, 2005]. В контексте филостратиграфического анализа паралоги представляют собой важный объект исследования, поскольку их наличие требует тщательного разграничения с ортологами для корректной интерпретации эволюционной истории генов. Это особенно актуально при идентификации генов-основателей, которые, в отличие от паралогов, не являются результатом дупликаций, а отражают появление принципиально новых функциональных элементов в геноме [Domazet-Lošo, 2003; Domazet-Lošo, Brajković, Tautz, 2007a]. То есть, их образование — это не просто результат дупликации существующих генов или генов с реорганизованными функциональными доменами. Скорее, гены-основатели представляют собой новые функциональные белки или белковые домены, приобретающие новое функциональное значение. После появления такого нового функционального домена можно ожидать, что он сохранит свою первичную белковую последовательность в такой степени, что его можно будет отследить с помощью поиска по ортологии.

Значительная часть белок-кодирующих генов современных организмов восходят к кладе возникновения клеточной жизни, иначе говоря, происходят от ограниченного набора генов, уже существовавших у последнего универсального общего предка (LUCA – Last Universal Common Ancestor) всех живых организмов. В контексте эволюционной биологии клада (от греч.

Κλάδος – ветвь) представляет собой группу организмов, состоящую из общего предка и всех его потомков, живых или вымерших [Futuyma, Kirkpatrick, 2017]. Предполагается, что эволюционная новизна на молекулярном уровне в значительной степени обусловлена дупликацией и неофункционализацией ранее существующей генетической информации [Dohmen и др., 2020; Tautz, Domazet-Lošo, 2011]. Однако геномные исследования последних десятилетий выявили, что происхождение часть генов не может быть прослежено до их предковых форм у последнего универсального общего предка [Tautz, Domazet-Lošo, 2011; Palmieri, Kosiol, Schlötterer, 2014; Sun, Zhao, Zhang, 2015; Zhang et al., 2019c]. Происхождение таких генов удается установить только в рамках определенных эволюционных линий, а их ортологи обнаруживаются только у ограниченного числа видов. В литературе такие гены<sup>1</sup> принято называть орфанными (или генами сиротами) в случае полного отсутствия их ортологов у других видов, и таксономически ограниченными генами или же кладо-специфичными генами в случае нахождения ортологов исследуемого гена лишь в пределах определенной кланды таксономического дерева. Появление кладо-специфичных генов обычно связывают с событиями основания генных семейств, то есть с появлением гена-основателя [Domazet-Lošo, Brajković, Tautz, 2007a]. Поиск и изучение кладо-специфичных генов и их функций представляет особый интерес, поскольку в ряде исследований было показано, что новые гены могут играть значительную роль в появлении новых эволюционных признаков и быть связаны с появлением новых морфологических признаков, как, например, у наземных растений [Bowles, Bechtold, Paps, 2020] и многоклеточных животных [Paps, Holland, 2018]. Кроме этого, было продемонстрировано, что эволюционно «молодые» гены связаны с каскадами развития органов, в частности, в развитии тканей головного мозга

---

<sup>1</sup> В частности, термин орфанных генов (orphan genes) широко распространен в литературе, но подразумевает под собой разные определения: как гены, не имеющие известных ортологов [Arendsee, Li, Wurtele, 2014]; как синоним таксономически ограниченных генов [Barrera-Redondo и др., 2023]; или же любые генетические последовательности с открытой рамкой считывания, уникальные для одного вида [Yu, Stoltzfus, 2012]. Однако в настоящее время преимущественно используется определение таксономически ограниченных (TRG) или кладо-специфичных генов, как более корректное.

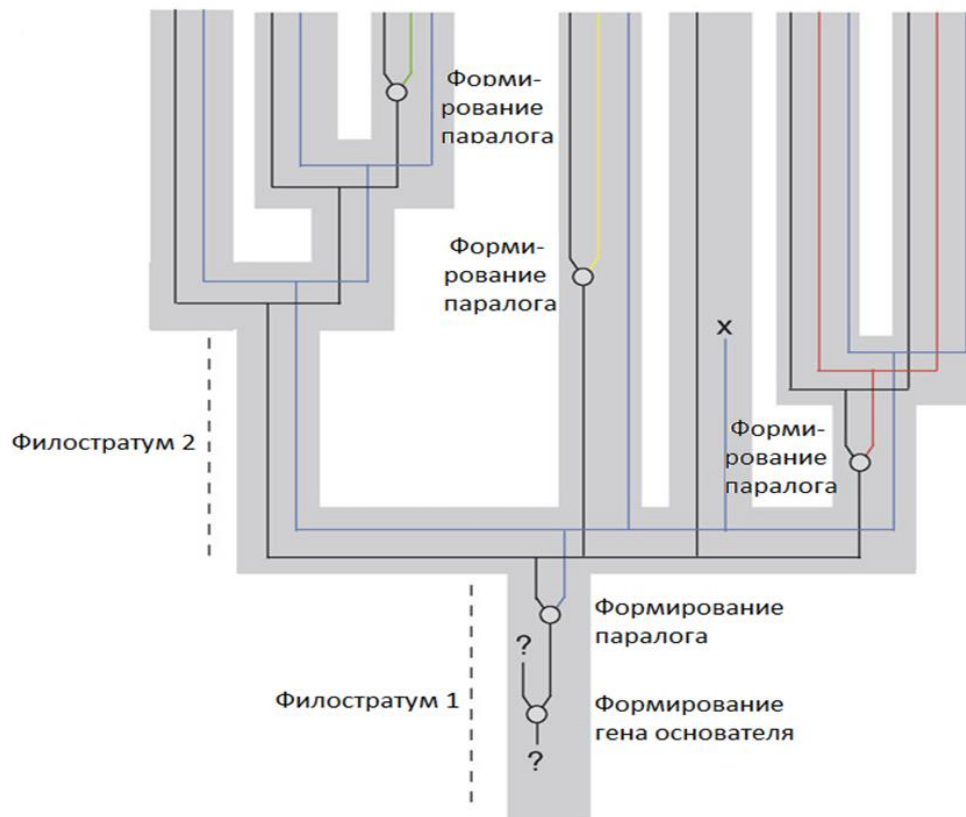
[An и др., 2023], а также было продемонстрировано, что кладо-специфичные гены перепредставлены в системах борьбы со стрессом и иммунной системе [Dornburg, Yoder, 2022]. Ряд исследователей также предполагает, что кладо-специфичные гены связаны с экологической специализацией в различных таксонах [Baalsrud и др., 2018].

Филостратиграфический анализ помогает решать эволюционные вопросы, такие как возможная связь между появлением кладо-специфичных генов и специфическими для линии эволюционными новшествами во время крупных эволюционных радиаций [Mendoza de и др., 2013], изучение эволюции онтогенетических транскрипционных паттернов [Domazet-Lošo, Tautz, 2010a; Drost и др., 2017], изучение скорости эволюции «молодых» генов по сравнению с «древними» [Moutinho, Eyre-Walker, Dutheil, 2022], или выявление появления полностью новых белков через события *de novo* возникновения генов [Carvunis и др., 2012].

Однако интерпретация кладо-специфичных генов как возникших *de novo* (т.е. в результате событий основания новых генов/семейств генов) является дискуссионной, поскольку их отсутствие за пределами изучаемой линии может объясняться недостаточной чувствительностью методов обнаружения ортологов [Moyers, Zhang, 2017, 2016; Weisman, Murray, Eddy, 2020].

Методология, предложенная в первых работах по филостратиграфическому анализу [Domazet-Lošo, Brajković, Tautz, 2007b], включает несколько ключевых этапов. В общем случае, на первом этапе создается референсная филогения, отражающая эволюционные взаимоотношения между определенным набором видов. Эта филогения делится на отдельные филостраты, каждая из которых отражает определенный эволюционный период (Рис.1). Поиск ортологов в методе Домазета-Лошо

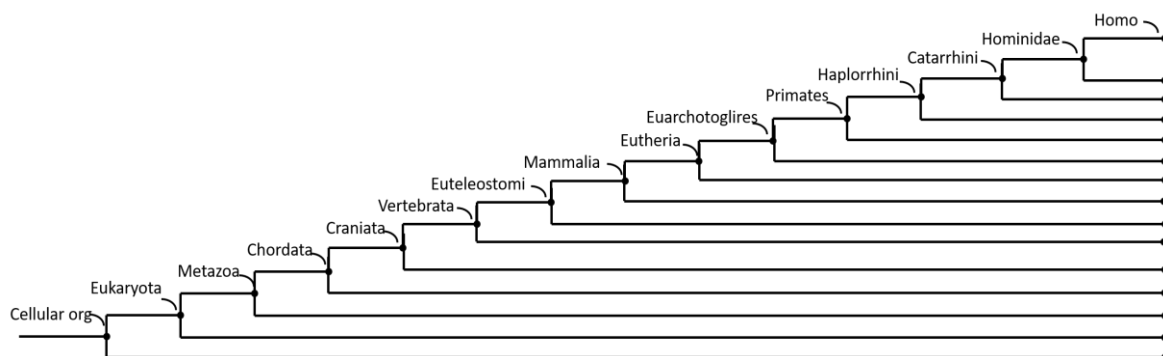
проводится на основании результатов BLAST выравнивания [Domazet-Lošo, Brajković, Tautz, 2007b].



**Рисунок 1** Модель возникновения семейств белков, адаптировано из [Domazet-Lošo, Brajković, Tautz, 2007a]. Серое дерево описывает филогенетические отношения между таксонами, внутренние деревья (разноцветные линии) описывают эволюцию семейств после возникновения гена.

В качестве единицы возраста (филостраты) определяется таксономический ранг наиболее ранней клады, в который входит предполагаемый ортолог. Числовой индекс от корня древа видов, присваиваемый этой филострате, называется филостратиграфическим индексом возраста (Phylostratigraphic Age Index, PAI). Ряд филострат зависит от выбранного древа видов для

исследования<sup>2</sup>. Например, филогенетический ряд человека, взятый из базы данных KEGG, представляет собой 15 филогенетрат – начиная от филогенстраты возникновения клеточных организмов и эукариот и заканчивая филогенстратой возникновения вида *Homo sapiens* (Рис.2). Другими словами, филогенстратический ряд можно рассматривать как эволюционную родословную конкретного вида



| PAI | Таксон             | Возраст (млн) | PAI | Таксон              | Возраст (млн) |
|-----|--------------------|---------------|-----|---------------------|---------------|
| 1   | Cellular Organisms | 4100          | 9   | Eutheria            | 160           |
| 2   | Eukaryota          | 1850          | 10  | Euarchontoglires    | 65            |
| 3   | Metazoa            | 665           | 11  | Primates            | 55            |
| 4   | Chordata           | 541           | 12  | Haplorrhini         | 50            |
| 5   | Craniata           | 535           | 13  | Catarrhini          | 44            |
| 6   | Vertebrata         | 525           | 14  | Hominidae           | 17            |
| 7   | Euteleostomi       | 420           | 15  | <i>Homo</i>         | 2.8           |
| 8   | Mammalia           | 225           | 16  | <i>Homo sapiens</i> | 0.35          |

**Рисунок 2** Филогенстратическое представление таксономического дерева *Homo sapiens* на основании родословной человека базы данных KEGG и таблица соотнесения индексов PAI с филогенстратами и оценками эволюционного возраста, выявленными при помощи метода молекулярных часов.

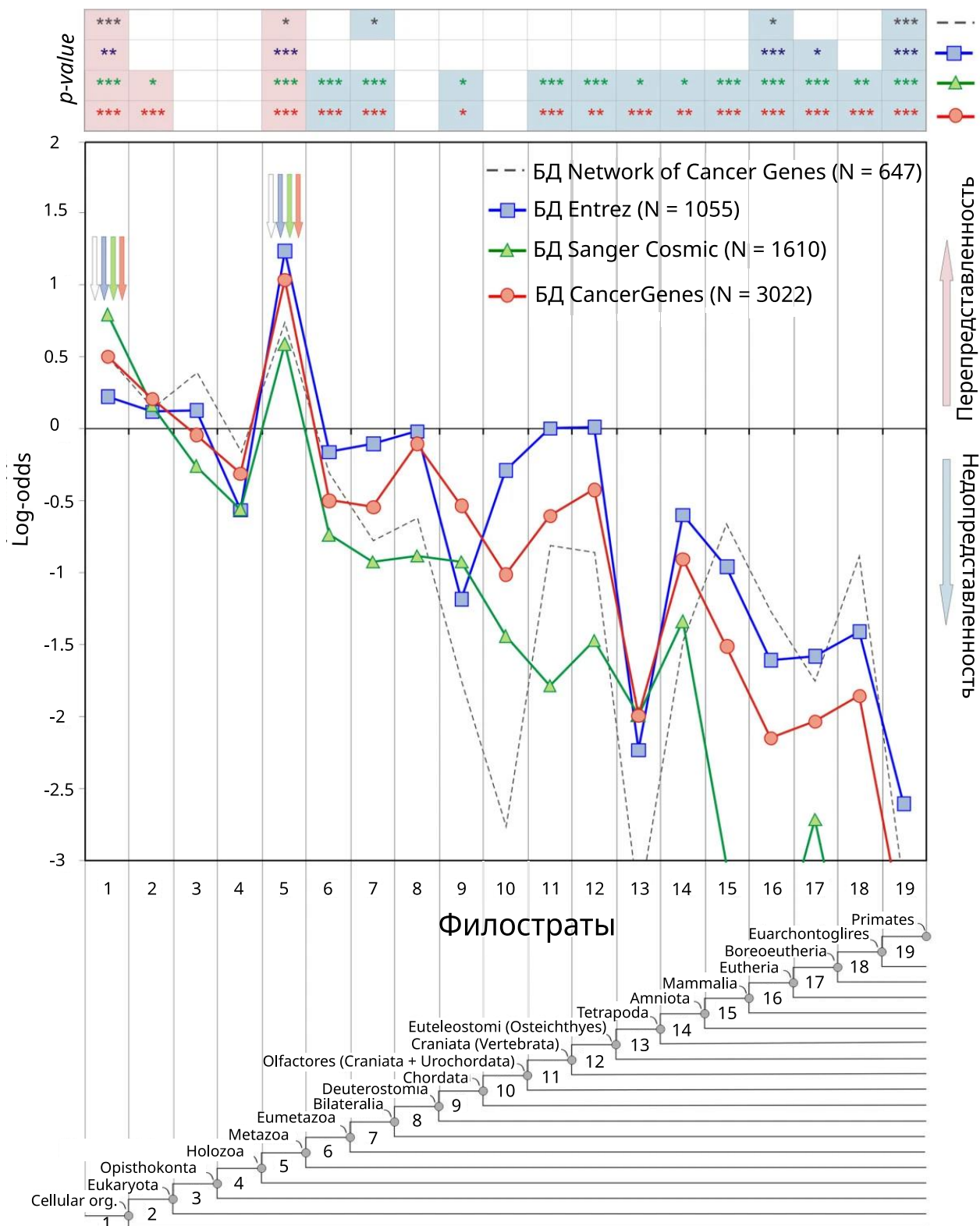
<sup>2</sup> Некоторые программные инструменты, реализующие филогенстратический анализ, такие как phylostrati и fagin, определяют ряд филогенстрат на основании анализа филогенетических деревьев [Arendsee и др., 2019a, 2019b].

Домазетом-Лошо и Таутцем был выполнен ряд работ по анализу индекса филогенетического возраста генов.

В одной из ранних работ [Domazet-Lošo, Tautz, 2008a] по данным базы COSMIC [Bamford и др., 2004] был проведён филогенетический анализ генов, ассоциированных с онкологическими заболеваниями. Этот анализ выявил важную закономерность: онкогены преимущественно относятся к филогенетам, соответствующим происхождению клеточной жизни (Cellular Organisms, PAI=1) и многоклеточных организмов (Metazoa, PAI=3). Данный факт позволяет предположить тесную связь между формированием злокачественных процессов и этими ключевыми эволюционными переходами. Примечательно, что исследование также показало выраженное снижение частоты появления новых генов, ассоциированных с онкологическими заболеваниями, в более поздних филогенетах происхождения Эуметазои (Eumetazoa, PAI=6) и Двустороннесимметричных организмов (Bilateria, PAI=7) (Рис.3).

Ключевой вывод этой работы состоит в том, что появление онкогенов в эволюционной истории носит неравномерный характер, причем гены, ассоциированные с опухолевой прогрессией, восходят к ранним этапам эволюции. Это указывает на прямую связь между эволюцией многоклеточности и уязвимостью к онкологическим заболеваниям, что указывает на фундаментальный эволюционный характер механизмов канцерогенеза. Авторы отмечают ограничения филогенетического подхода, включая человеко-центричную реконструкцию филогенетического дерева и возможную потерю ортологов вследствие ограниченного числа секвенированных геномов. Тем не менее, они подчеркивают, что метод

остаётся информативным для анализа эволюционных этапов, связанных с формированием молекулярных основ опухолевого процесса.



**Рисунок 3** Результаты филостратиграфического анализа генов, ассоциированных с злокачественными новообразованиями (переведено и адаптировано из [Domazet-Lošo, Tautz, 2008a]).

Дополнительные свидетельства эволюционных различий генов, вовлеченных в онкогенез, были представлены в работе, где с помощью филогенетического анализа сетей ко-экспрессии генов в клетках карциномы яичников было показано, что функциональные модули, связанные с опухолевой прогрессией, не являются равномерно распределёнными по индексам филогенетического возраста генов. Авторы обнаружили, что кластеры «древних» генов определяют ключевые метаболические и сигнальные пути, необходимые для поддержания базовых клеточных функций, тогда как более «молодые» гены способствуют тонкой настройке взаимодействий с микроокружением опухоли [Zhang et al., 2019b].

Описанная методика представляет собой обобщённый подход, и несмотря на кажущуюся концептуальную простоту филогенетики, на практике определение филогенетического кластера оказывается достаточно сложной задачей. Поскольку метод филогенетики основывается на идентификации самого раннего предполагаемого ортолога [Domazet-Lošo, Brajković, Tautz, 2007b], он чувствителен к ошибкам первого рода, например, к неверной идентификации ортологов из-за повторяющихся последовательностей [Weisman, Murray, Eddy, 2020]. Гены, уникальные для одного вида и не имеющие ортологов у других видов, часто могут быть либо не аннотированы, либо аннотированы некорректно [Moyers, Zhang, 2016; Tautz, Domazet-Lošo, 2011]. Распознавание и выделение таких генов из "шума" данных является сложной задачей [McLysaght, Hurst, 2016; Vakirlis и др., 2018]. Кроме того, филогенетические деревья могут содержать ошибки или быть недостаточно определёнными, имея множество неразрешённых ветвей. Также данные могут содержать систематические искажения: отдельные филогенетические кластеры могут быть представлены избыточно или недостаточно в различных анализах, а методы аннотации генов могут значительно различаться в зависимости от времени проведения исследования и изучаемых видов.

Кроме того, первые исследования в области филогенетики [Domazet-Lošo, Tautz, 2010a, 2008a; Šestak и др., 2013] подверглись критике из-за ошибок

в выявлении ортологов быстро эволюционирующих генов при использовании алгоритма BLAST [Moyers, Zhang, 2017, 2016]. В этих работах были сгенерированы списки случайных генетических последовательностей так, чтобы их филогенетические индексы соответствовали таксономическому рангу последнего универсального общего предка. В последующей проверке классическим методом филогенетики, было выявлено, что филогенетический анализ в связи с ошибками поиска гомологии алгоритма BLAST недооценивает значения PAI для значительной части сгенерированных генов (13,85%), в особенности для последовательностей, с высокой изменчивостью, малой длиной и/или короткими консервативными регионами. Эти гены в работах Домазета-Лошо и соавторов относились к категории очень эволюционно «молодых». Следует отметить, что проблема некорректной оценки гомологии при использовании алгоритма BLAST была известна ещё до публикации первых исследований по филогенетики. Однако в этих работах сообщалось о значительно меньшей доле ошибочно определенных ортологов [Sierk, Pearson, 2004; Soding, 2005].

В связи с указанными проблемами в современной эволюционной биологии возникла потребность в разработке и применении более продвинутых методов анализа ортологов для изучения индексов филогенетического возраста генов. В разделе 1.4 представлен обзор нескольких инструментов и алгоритмов филогенетики, учитывающих критику оригинального метода и вносящих в него новые идеи и усовершенствования, направленные на повышение точности и достоверности анализа.

Для комплексного понимания эволюции генов, объединяющего их происхождение и текущую функциональную активность, филогенетический подход был расширен за счёт включения анализа уровня экспрессии генов. Это привело к формированию *филотранскриптного анализа* – метода, позволяющего исследовать взаимосвязь между таксономическим возрастом генов и изменениями их активности. Используя филотранскриптный анализ, можно выявить, как

эволюционные характеристики генома связаны с транскрипционной регуляцией и функциональной динамикой генов в различных биологических контекстах. К филотранскриптомным индексам относятся два эволюционных индекса: TAI [Domazet-Lošo, Tautz, 2010a] и TDI [Quint и др., 2012a] – индекс возраста транскриптома и индекс дивергенции транскриптома, соответственно, которые рассчитывают связь возраста генов с уровнями их экспрессии:

$$TAI_s = \frac{\sum_{i=1}^n ps_i e_{is}}{\sum_{i=1}^n e_{is}}$$

где  $ps_i$  – целое число, отражающее PAI для гена  $i$ ,  $e_{is}$  – значение уровня экспрессии, полученного с транскриптомных данных по анализу экспрессии гена  $i$  в образце  $s$ ,  $n$  – общее число генов. TAI отражает средний PAI генов, с наибольшим уровнем экспрессии в выбранном биологическом процессе.

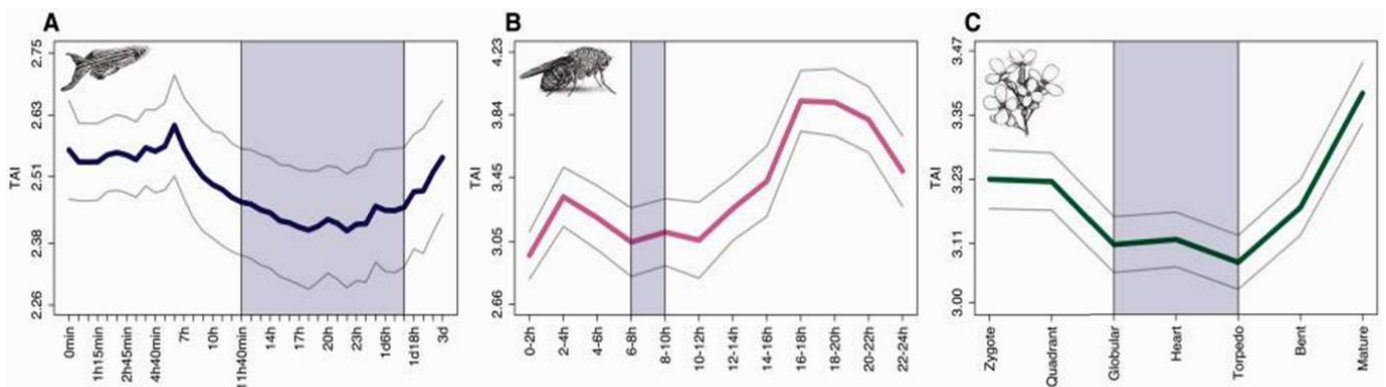
$$TDI_s = \frac{\sum (e_{is} * ds_i)}{\sum e_{is}}$$

где  $e_{is}$  обозначает уровень экспрессии гена  $i$  в образце  $s$ , а  $ds_i$  обозначает соответствующий индекс дивергенции гена  $i$ . Значения TDI количественно отражают консервативность транскриптома при определенных условиях и помогают выявить биологические процессы или стадии развития организма, в которых наблюдается повышенная экспрессия более или менее консервативных генов.

Одним из наиболее показательных направлений применения филотранскриптомного анализа является изучение эволюционных особенностей эмбриогенеза. Этот метод позволяет объединять экспериментальные данные об экспрессии генов с филостратиграфическими индексами их возраста, что дает возможность выявить паттерны активности генов с разными значениями PAI в различных условиях, в данном случае – на разных стадиях эмбриогенеза.

Сравнивая значения TDI на разных стадиях развития или в разных условиях среды, можно выявить закономерности консервативности или

дивергенции признаков, которые могут быть связаны с определенным эволюционным давлением или селективными преимуществами. Эта информация представляет интерес для понимания эволюционного значения паттернов экспрессии генов. Подобный анализ позволил выявить закономерности между уровнем экспрессии и индексами филостратиграфического возраста генов в процессе эмбриогенеза растений и животных. Результаты анализа показали, что в ранней и поздней стадии эмбриогенеза *Arabidopsis thaliana*, *Danio rerio* и *Drosophila melanogaster* наблюдаются повышенные значения TAI и TDI, что указывает на более высокую активность эволюционно «молодых» и менее консервативных генов. Напротив, на средних стадиях эмбриогенеза фиксируются пониженные значения обоих индексов, отражающие преобладание экспрессии эволюционно «древних» и высоко консервативных генов [Cridge, Dearden, Brownfield, 2019; Drost и др., 2015; Quint и др., 2012a]. Эта закономерность совпадает с ранее предложенной молекулярной моделью «*паттерна песочных часов*» эмбриогенеза [Irie, Kuratani, 2014].



**Рисунок 4** Паттерн «песочных часов» эмбриогенеза в развитии *D. rerio* и *D. melanogaster* и *A. thaliana*, полученный путем анализа филотранскриптомных индексов возраста транскриптома на разных стадиях развития этих организмов [Drost и др., 2015]

Стоит отметить, что несмотря на то, что подобный паттерн прослеживается у многих исследованных организмов, в ряде моллюсков и полихет наблюдается ровно противоположная картина, в которой в средней

стадии эмбриогенеза активны менее консервативные гены, а в ранней и поздней – более консервативные [Wu, Ferger, David Lambert, 2019]. Кроме того, филотранскриптомный анализ позволяет оценить индекс возраста транскриптома в отдельных органах. Исследование образцов корня, листьев, стебля и цветка *Arabidopsis thaliana* выявило значительные различия в значениях ТАИ их активных транскриптомов. Особенно заметным был контраст между «древним» транскриптомом корня и «молодым» транскриптомом цветка [Lei и др., 2017].

В контексте филотранскриптомного анализа онкологических заболеваний, было показано, что несмотря на высокую гетерогенность злокачественных опухолей, они проявляют конвергентную эволюционную тенденцию, приводящую к ограниченному набору клеточных состояний. Филостратиграфический анализ в данном случае указывает, что этот сдвиг обусловлен усиленной активностью древних генетических программ в сочетании с подавлением функций генов, появившихся в ходе эволюции многоклеточности, что согласуется с концепцией функциональной специализации генов различных возрастных групп в онкогенезе [Chen, He, 2016a].

Эта гипотеза также была подтверждена филотранскриптомным анализом широкого ряда солидных опухолей, который показал, что переход нормального клеточного состояния к злокачественному сопровождается сдвигом в экспрессии генов: происходит активация эволюционно «древних» генов, характерных для процессов, возникших у одноклеточных организмов, в то время как экспрессия генов, связанных с многоклеточными регуляторными процессами, подавляется [Trigos и др., 2017].

Таким образом, филотранскриптомный анализ дополняет филостратиграфический, предоставляя количественные показатели для исследования не только происхождения генов, но и их реальной функциональной активности в контексте эволюции и патологии.

## 1.4 Программные инструменты филогенетики

Развитие методов секвенирования и рост объема геномных данных выявили ограничения филогенетического подхода, связанные, в частности, с несовершенством алгоритма BLASTP для определения ортологов, применявшегося на ранних этапах развития метода (см. раздел 1.3). Помимо этого, в силу вычислительной сложности, филогенетический анализ полногеномных данных с применением алгоритмов BLASTP может занимать до нескольких недель [Buchfink, Reuter, Drost, 2021]. С развитием вычислительных технологий и накоплением геномных данных возникла потребность в разработке специализированных программных инструментов, определяющих филогенетические индексы с большей вычислительной точностью и скоростью. В данной главе рассматриваются современные программные инструменты в этой области: fagin, GenEra, oggmap и Orthoscape.

### 1.4.1 fagin

Программа fagin представляет собой инструмент, реализованный на языке R, который используется для анализа филогенетического индекса возраста генов [Arendsee и др., 2019a] (<https://github.com/arendsee/fagin>). Подход, применяемый в ней, значительно отличается от метода филогенетики, предложенного в первых работах Домазета-Лошо и Таутца, как концептуально, так и методически. Этот метод сначала определяет синтенные геномные интервалы в целевом геноме – регионы генома, в которых сохраняется взаимное расположение генов у разных видов, где, как ожидается, находятся ортологи каждого целевого гена. Затем он ищет в этих интервалах любые следы сходства, учитывая, как аминокислотные, так и нуклеотидные последовательности. Этот подход более чувствителен и специфичен по сравнению с классическим методом филогенетики. Благодаря этому, он позволяет точнее определять быстро эволюционирующие и «молодые» гены.

Однако концептуально ключевое различие заключается в акценте на положительных доказательствах. Метод филостратиграфии, предложенный Домазетом-Лошо и Таутцем, определяет возраст генов на основе негативных доказательств – отсутствия обнаруживаемых гомологов за пределами клады. В отличие от нее, метод *fagin* использует анализ на основе синтении для поиска позитивных доказательств наличия некодирующих последовательностей у близкородственных видов. Это позволяет *fagin* различать гены, возникшие *de novo*, и гены, подвергшиеся быстрым эволюционным изменениям. Ниже приведены основные этапы работы *fagin*.

Предварительная обработка исходных данных: Первый этап работы *fagin* включает в себя очистку, проверку и обобщение всех исходных данных. К ним относятся филогенетическое дерево, связывающее исследуемый вид с одним или несколькими целевыми видами, последовательности геномов исследуемого и целевого видов, файлы GFF, с аннотацией генов для каждого вида, и список генов. Для *fagin* также требуются парные карты синтении между геномом исследуемого вида и геномами каждого целевого вида.

Определение классов ортологии: На втором этапе каждый запрашиваемый ген относится к классу ортологии на основе предполагаемых интервалов поиска в целевых геномах. *fagin* рассматривает три общих случая, которые составляют классификацию ортологов «UNA»: AAic (аминокислотная идентичность), NTic (нуклеотидная идентичность) и Unknown. AAic относится к генам, имеющим ортологию по аминокислотным последовательностям с уже известными кладо-специфическими генами в пределах поискового интервала. NTic относится к генам с ортологией по нуклеотидным последовательностям с любой последовательностью в пределах интервала поиска. Если ортологии не обнаружено, *fagin* пытается определить биологическую или техническую причину отсутствия сходства. Из биологических причин такого поведения могут быть вставки и делеции последовательности гена, слишком сильные и частые хромосомные

транслокации, которые могут привести к слишком далекому разделению участков хромосом, чтобы между ними можно было проследить синтению.

Определение филогенетического индекса возраста: На третьем этапе оценки, полученные на этапе классов ортологов, используются для определения филогенетического индекса для каждого гена и оценки степени достоверности этих оценок. Помимо определения PAI, каждому гену в итоговом файле присписывается потенциальное биологическое происхождение, которое определяется с помощью классификации «UNA», которая сводит информацию по всему филогенетическому дереву в единый вектор меток, представляющих достоверность существования генетического или негенетического ортолога в каждой аутгруппе. Этот этап помогает отличить совпадения, пропущенные из-за некачественных данных, от совпадений, пропущенных из-за отсутствия гомолога гена в целевых таксонах.

#### 1.4.4 GenEra

GenEra [Barrera-Redondo и др., 2023] – это программный комплекс, предназначенный для оценки возраста последнего общего предка (LCA – Last Common Ancestor) семейств генов.

В этапы работы GenEra входят:

Первоначальное выравнивание последовательностей: отличительная особенность GenEra заключается в замене традиционных методов поиска BLASTP [Camacho и др., 2009], PSI-BLAST или HMMER – сверхбыстрым алгоритмом выравнивания белков DIAMOND v2. Выбор DIAMOND обосновывается тем, что BLAST ограничивает число лучших совпадений последовательностей, потенциально упуская информацию о таксонах, находящихся в отдаленном родстве. DIAMOND же, в свою очередь, позволяет настраивать список парных выравниваний по всей нередуцированной базе данных белков NCBI, при этом предоставляя результаты гораздо быстрее, чем подходы на основе BLASTP.

Установка порогового значения e-Value: устанавливается порог e-value ниже  $10^{-5}$ , чтобы результаты совпадения последовательностей считались надежными. Этот порог выбран, поскольку считается общепринято используемым в оценке ортологов. Полученные в результате поиска значения e-value преобразуются и используются для кластерного анализа, чтобы предсказать семейства генов с помощью MCL (Markov Clustering Algorithm).

Создание базы данных родословных: GenEra делает запросы в *NCBI tax2lin* для создания базы данных родословных, чтобы связать идентификаторы NCBI Taxonomy из парных выравниваний DIAMOND с их таксономической принадлежностью. Эта база данных необходима для определения эволюционных отношений между найденными генами и интересующими видами. В процессе присваивания таксономического уровня могут возникать сложности, связанные с неоднородностью данных секвенирования у организмов, не являющихся модельными видами: недостаточное покрытие геномов или секвенирование только ряда генов у немодельных видов. Это может привести к ошибкам в предсказании индексов возраста генов, в частности к ложноположительным ошибкам возникновения генов-основателей генных семейств. Разработчики GenEra решают эту проблему, осуществляя анализ на основе всех доступных совпадений по последовательности и сохраняя для анализа только те таксономические уровни, на которых у как минимум одного вида наблюдается совпадение более чем с 10% белков искомого вида. Это позволяет отбросить из базы данных последовательностей NCBI те виды, которые представлены лишь несколькими просеквенированными генами.

Кластеризация и оценка возраста: после выравнивания генов GenEra ищет наиболее отдаленный ортолог для оценки возраста генов для каждого кластера, чтобы оценить количество событий возникновения генных семейств на протяжении эволюционной истории искомого вида. Этот этап кластеризации при помощи многочленной логистической регрессии позволяет

отнести гены к наиболее древнему возрасту среди генов одного генного семейства.

Поиск по принципу «все против всех» (All-vs-All): После присвоения возраста всем генам запрашиваемого вида GenEra выполняет выравнивание по принципу «все против всех» в DIAMOND для всех генов исследуемого вида против них самих же. Этот шаг позволяет определить паралоги среди генов искомого вида.

Оценка таксономической репрезентативности: После таксономического анализа GenEra рассчитывает показатель таксономической репрезентативности, чтобы оценить надежность присвоения возраста гена на основании совпадения последовательностей. Этот показатель учитывает наличие ортологов гена хотя бы у одного репрезентативного вида для каждого промежуточного таксономического уровня между наиболее отдаленно родственным родом и искомым видом.

Оценка ошибки при обнаружении ортологии: GenEra оценивает, могут ли ошибки в обнаружении гомологии (HDF) объяснить неверное присвоение возраста генам запроса. Он рассчитывает вероятности HDF для каждого гена, не имеющего прослеживаемого ортолога у исследуемого вида, и проверяет нулевую гипотезу о непрослеживаемой ортологии для каждого гена. Возможность проверки HDF для каждого таксономического уровня зависит от таксономической выборки, заданной пользователем.

#### 1.4.5 oggmap

oggmap [Ullrich, Glytnasi, 2023] – это пакет Python, предназначенный для получения карт ортологий (ортокарт) или, другими словами, значений филостратиграфического индекса возраста заданных групп ортологов из результатов, полученных с помощью таких инструментов, как OrthoFinder [Emms, Kelly, 2019] и eggNOG [Huerta-Cepas et al., 2019]. В отличие метода филостратиграфии, предложенного Домазетом-Лошо и Таутцем, этот подход

не подразумевает шага поиска ортологов при помощи инструментов выравнивания, а использование готовых результатов поиска ортологов в виде ортокарт и дальнейшее использование их в расчёте возраста генов. Эти ортокарты содержат информацию об индексах возраста генов в каждой группе ортологов, которая затем, в рамках работы oggmap, может быть использована для расчёта взвешенных данных об экспрессии (Transcriptome Age Index или TAI) из данных одноклеточного секвенирования РНК (scRNA-seq) или же анализироваться отдельно.

Этапы работы oggmap:

Идентификация ортогрупп: oggmap начинает с получения ортогрупп из результатов работы инструментов предсказания ортологов, таких как OrthoFinder, eggNOG или PLAZA. Эти инструменты определяют группы генов, которые, как предполагается, произошли от общего предкового гена (LCA) в результате видообразования. Помимо этого, oggmap поддерживает использование предрасчитанных данных по поиску ортогрупп

Присвоение филогенетической группы на основе LCA (последний общий предок): После идентификации ортологичных групп oggmap присваивает этим группам классы возраста генов. Это делается путем определения последнего общего предка (LCA) видов, представленных в ортологичной группе. Этот LCA считается самым последним общим предком видов в группе

#### 1.4.6 Orthoscape

Orthoscape [Mustafin и др., 2017] – это плагин к программному комплексу Cytoscape, предназначенный для комплексного анализа эволюционных характеристик генов и генных сетей. В процессе работы Orthoscape обращается к базе данных KEGG с помощью REST API, запрашивая такую информацию, как списки ортологов, доменный состав белков, кодируемых анализируемыми генами, таксономический ряд организма (ген которого подвергается анализу), нуклеотидные/аминокислотные последовательности

гена/белка на основании которых и проводится анализ эволюционных характеристик.

Помимо анализа индексов PAI генов, входящих в генную сеть, Orthoscape также позволяет рассчитывать отношения синонимичных и несинонимичных замен. Orthoscape предлагает пользователям гибкие настройки для параметров анализа, что позволяет проводить исследования с высокой степенью спецификации и адаптации под конкретные научные задачи. В Orthoscape пользователи могут выбирать из различных методов для расчёта эволюционного возраста генов. Один из методов представляет собой использование таблиц идентичности последовательностей белков согласно алгоритму Смита-Ватермана, что позволяет определять ортологичные отношения между генами на основе выравнивания последовательностей. При этом пользователи могут самостоятельно настраивать порог идентичности последовательностей, что обеспечивает большую гибкость в выборе ортологичных групп для дальнейшего анализа. Кроме того, доступен метод отбора ортологов на основе степени сходства доменного состава белков, что расширяет возможности исследований в области эволюционной биологии.

### **1.5 Базы данных и биоинформатические ресурсы онкологических заболеваний**

Современные исследования в области онкологии всё чаще опираются на биоинформатические ресурсы, которые систематизируют и интегрируют данные о генетических, транскриптомных и эпигенетических изменениях, ассоциированных со злокачественными трансформациями. Развитие технологий высокопроизводительного секвенирования (например, next-generation sequencing, NGS) привело к экспоненциальному росту объёмов информации о молекулярных механизмах онкогенеза. В этой связи базы данных играют ключевую роль, предоставляя инфраструктуру для хранения, анализа и интерпретации сложных наборов данных. Они позволяют не только

идентифицировать биомаркеры и терапевтические мишени, но и изучать гетерогенность опухолей, прогнозировать клинические исходы и разрабатывать персонализированные подходы к лечению.

**The Cancer Genome Atlas (TCGA)** – это масштабный исследовательский проект, посвящённый изучению геномных, транскриптомных и эпигеномных изменений при онкологических заболеваниях. Запущенный в 2006 году как совместная инициатива Национального института рака (NCI) и Национального института исследования генома человека (NHGRI) при поддержке правительства США, TCGA стал фундаментом для понимания молекулярных механизмов злокачественных процессов. Проект завершился в 2017 году, но его материалы остаются ключевым ресурсом для научного сообщества. В рамках инициативы проанализировано свыше 11 000 образцов опухолей и здоровых тканей более чем 33 типов, что делает его одним из наиболее полных источников информации о генетических аномалиях при злокачественной трансформации. Образцы нормальных тканей, которые называются нормальными, прилегающими к опухоли тканями (NAT), берут из гистологически нормальных тканей, прилегающих к краям опухоли. Материалы проекта TCGA на настоящий момент хранятся и поддерживаются в рамках платформы Genomic Data Commons (GDC), которая обеспечивает их унификацию, обновление и доступ для исследователей [Heath et al., 2021]. В состав базы данных входят несколько категорий молекулярных данных, каждая из которых охватывает определенный аспект биологии опухолей (Табл.1). Структура GDC организована в иерархическом формате, охватывающем несколько уровней. На первичном уровне хранятся «сырые», необработанные данные, полученные в результате высокопроизводительного секвенирования и других экспериментальных методов. Это включает данные полногеномного секвенирования (WGS), экзомного секвенирования (WXS) и транскриптомного профилирования (RNA-Seq). На вторичном уровне представлены обработанные данные: идентифицированные геномные варианты (мутации, структурные перестройки, вариации числа копий), уровни

экспрессии генов, а также эпигенетические маркеры, такие как метилирование ДНК.

**Таблица 1.** Типы данных, полученных в проекте TCGA

| Тип                | Подтип                           | Описание                                                                                                                          | Формат         |
|--------------------|----------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|----------------|
| Клинические данные | Клинические данные               | Доступная клиническая информация (может включать демографическую информацию, информацию о лечении, данные о выживаемости и т. д.) | XML, TXT       |
|                    | Данные биообразцов               | Информация об обработке образцов Biospecimen Core Resource Center                                                                 | XML, TXT       |
|                    | Патологоанатомические заключения | Патологические заключения (для отдельных случаев)                                                                                 | PDF            |
| Число копий генов  | SNP микрочипы                    |                                                                                                                                   | CEL, IDAT, TXT |
|                    | Микрочип копий генов             |                                                                                                                                   | TXT, TSV, MAT  |
|                    | Низкопороговое секвенирование    | Низкопороговое секвенирование всего генома опухолевых и нормальных сопоставленных образцов                                        |                |
| ДНК                | Полноэкзомное секвенирование     | Сравнение покрытия секвенирования опухоли и нормы                                                                                 | BAM, VCF, TSV  |
|                    | Полногеномное секвенирование     | Секвенирование экзона опухолевых и нормальных образцов                                                                            | BAM, VCF, MAF  |
|                    | SNP микрочип                     |                                                                                                                                   | CEL, IDAT, TXT |
|                    | Следы последовательности         | Необработанные данные капиллярного секвенирования                                                                                 | SCF, TR        |
| Изображения        | Диагностические изображения      | Изображения всего среза ткани, используемые для диагностики опухоли                                                               | SVS            |

|                                 |                                      |                                                                           |               |
|---------------------------------|--------------------------------------|---------------------------------------------------------------------------|---------------|
|                                 | Тканевые изображения                 | Изображения всего среза ткани, используемые для диагностики опухоли       | SVS           |
|                                 | Рентгенологические изображения       | Предоперационные радиологические снимки (МРТ, КТ, ПЭТ)                    | DCM           |
| Метилирование                   | Бисульфитное секвенирование          | Полногеномное секвенирование после обработки бисульфитом                  | BAM, VCF, BED |
|                                 | Bead array                           |                                                                           | TXT, IDAT     |
| Микросателлитная нестабильность |                                      | Маркеры микросателлитной нестабильности, гомо- и гетерозиготность аллелей | FSA, TXT      |
| миРНК                           | миРНК секвенирование                 | Секвенирование миРНК опухоли                                              | BAM, TXT      |
|                                 | Микрочип                             |                                                                           | TXT           |
| Экспрессия мРНК                 | мРНК секвенирование                  | Секвенирование mRNA с поли(А)-обогащением                                 | BAM, TXT      |
|                                 | Полнотранскриптомное секвенирование  | Секвенирование mRNA с удалением рибосомной РНК                            | BAM, TXT      |
|                                 | Микрочип                             |                                                                           | CEL, TXT      |
| Концентрации белков             | Белковые микрочипы в обращенной фазе | Изображения белковых биочипов и сырые сигналы                             | TIFF, TXT     |

Помимо молекулярно-генетических данных, для всех образцов предоставляются клиническая и фенотипическая информация. Они включают данные о демографии и диагнозе пациента по классификации Всемирной Организации Здравоохранения, клинической стадии, тип опухоли и о назначенном лечении. Стадии опухолей оцениваются по системе классификации TNM (Tumor-Node-Metastasis) на основе трех ключевых критериев:

1. T (Tumor) – оценивает размер первичной опухоли и степень ее распространения в окружающие ткани.
  - T0 – опухоль не обнаружена

- T1 – опухоль  $\leq 2$  см, не распространяется за пределы органа
- T2 – опухоль  $> 2$  см или прорастает в соседние структуры
- T3-T4 – более глубокая инвазия или вовлечение отдалённых тканей

2. N (Node) – характеризует поражение лимфоузлов метастазами.

- N0 – метастазов в лимфоузлах нет
- N1-N3 – указывают на количество или локализацию пораженных узлов (например, N1 – 1-3 узла, N2 – 4-9 узлов)

3. M (Metastasis) – описывает наличие отдаленных метастазов.

- M0 – метастазы отсутствуют
- M1 – выявлены отдаленные метастазы
- MX – наличие метастаз неизвестно

Стадии опухолей от I до IV определяются комбинацией этих показателей (Табл. 2) [Cserni et al., 2018]

**Таблица 2.** Система определения стадий развития опухоли по системе TNM

|                   | <b>T</b> | <b>N</b> | <b>M</b> |
|-------------------|----------|----------|----------|
| <b>Стадия I</b>   | 1,2      | 0        | 0        |
| <b>Стадия II</b>  | 3,4      | 0        | 0        |
| <b>Стадия III</b> | Любая    | 1,2      | 0        |
| <b>Стадия IV</b>  | Любая    | Любая    | 1        |

Эти сведения позволяют связывать молекулярные профили с клиническими характеристиками, прогнозом и эффективностью терапии.

Контроль качества в GDC реализован через унифицированные пайплайны (GATK [Auwera, O'Connor, 2020], STAR [Dobin et al., 2013], SeSAMe [Mauland, Winther, Ruan, 2024]), что устраняет технические артефакты и обеспечивает сопоставимость. Доступ к данным регулируется: открытые данные, такие как мутации, клинические аннотации, уровни экспрессии доступны свободно, а необработанные файлы мультиомных данных требуют отдельного допуска для защиты персональной информации.

При помощи TCGA было обнаружено несколько десятков новых генов, ассоциированных с онкогенезом. В частности, в аденокарциномах лёгкого были выявлены мутации в сигнальном пути RTK-Ras-Raf и их связь с устойчивостью к терапии [Cancer Genome Atlas Research Network et al., 2016]. Другая работа обнаружила ряд мутаций в сигнальных рецепторах и транскрипционных регуляторов в глиобластомах [Brennan et al., 2013].

Международный консорциум по изучению генома рака (ICGC) – другой масштабный проект в этой области. Его цель – систематическое исследование геномных, транскриптомных и клинических данных пациентов с онкологическими заболеваниями [Zhang et al., 2019a]. Данные консорциума охватывают более 37 000 геномов, охватывающих более 30 типов опухолей. Собранные материалы включают результаты полногеномного и экзомного секвенирования опухолевых и нормальных тканей, профили экспрессии генов, эпигенетические модификации, такие как метилирование ДНК и клинические характеристики пациентов (стадия заболевания, тип опухоли, выживаемость). Ключевая особенность ICGC – децентрализованный подход: данные поступают из 26 национальных проектов по анализу онкологических заболеваний, а затем унифицируются в соответствии со стандартами консорциума. Такой подход сохраняет гетерогенность источников, но требует строгой валидации и аннотации вариантов. Доступ к данным организован через портал ICGC Data Portal, где часть информации открыта, а чувствительные данные (например, сырые данные РНК-Seq) требуют специального разрешения.

**COSMIC (Catalogue Of Somatic Mutations In Cancer)** – одна из ключевых баз данных в онкологии, созданная для систематизации информации о генетических изменениях, связанных с онкологическими процессами. Она создана Институтом Сэнгера и систематизирует информацию о соматических мутациях и их клинических последствиях. Процесс сбора и интеграции данных в COSMIC основан на систематическом анализе научной литературы,

экспериментальных исследований и больших проектов по секвенированию геномов, таких как TCGA и ICGC.

База содержит детальную информацию об однонуклеотидных мутациях, делециях, инсерциях, вариации числа копий и химерных генах, ассоциированных с онкологическими заболеваниями. Для каждого гена указаны его роль в опухолевом процессе – например, отнесение к онкогенам или супрессорам опухолей – типы мутаций в нем и их расположение в геноме. Описание каждой мутации сопровождается информацией о типе опухоли, в котором эта мутация была обнаружена.

Данные об образцах опухолей разделены на три взаимосвязанных блока: индивидуальный, опухолевый/тканевый и характеристики образца. Индивидуальные особенности включают в себя возраст, пол, этническую принадлежность, факторы среды, такие как курение или инфицирование вирусом папилломы человека, семейный анамнез, историю болезни и предшествующую терапию. Характеристики опухоли/ткани включают: источник среза опухоли (первичная опухоль или метастаза), локализацию метастазирования, стадию заболевания, реакцию на лекарственные препараты и цитогенетические данные. Описание образца включает: источник образца (хирургически зафиксированный, взятый при вскрытии, клеточная линия), влияние терапии и информация по микросателлитной нестабильности.

Биоинформатические ресурсы и базы данных стали неотъемлемой частью современных онкологических исследований, обеспечивая систематизацию, интеграцию и анализ больших объемов молекулярных и клинических данных. Это особенно критично для исследований, охватывающих различные типы опухолей, где сравнение геномных, транскриптомных и эпигенетических профилей опухолей выявляет как общие молекулярные паттерны, так и уникальные особенности отдельных типов опухолей.

## 1.6 Молекулярно-генетические аспекты онкологических заболеваний

Злокачественные опухоли – это сложный комплекс заболеваний, возникающий в результате многоэтапного накопления генетических и эпигенетических изменений в нормальных клетках, что приводит к неконтролируемому росту клеток и образованию опухолей. Центральную роль в этом играют два класса генов: онкогены, активация которых стимулирует неконтролируемую пролиферацию клеток, и гены-супрессоры опухолей, инактивация которых устраняет естественные барьеры для канцерогенеза. Дисбаланс между этими группами генов приводит к утрате контроля над клеточным циклом, апоптозом и репарацией ДНК, формируя основу опухолевой прогрессии [Hanahan, Weinberg, 2011].

Онкогены формируются в результате мутаций протоонкогенов, чаще всего через точечные мутации, амплификацию или хромосомные перестройки. Примером служат мутации генов GTP-аз *Ras*, приводящие к постоянной активации сигнального пути MAPK [Prior, Lewis, Mattos, 2012]. Амплификация *c-myc* усиливает экспрессию регуляторов клеточного цикла, способствуя опухолевому росту карцином молочной железы, яичников и некоторых плоскоклеточных карцином [Brison, 1993; Dang, 2012]. Хромосомные транслокации, такие как образование филадельфийской хромосомы, приводят к синтезу химерных белков с онкогенной активностью [Falini, Mason, 2002; Heerema, 1998; Mitelman, 2000].

Гены-супрессоры опухолей поддерживают геномную стабильность, регулируют клеточный цикл, подавляют избыточную пролиферацию и инициируют апоптоз. Их инактивация требует двух независимых событий – по модели Кнудсона: наследуемой и соматической мутаций [Jia, Zhao, 2019; Knudson, 1971]. Основные механизмы инактивации включают делеции, эпигенетическое подавление экспрессии и мутации с потерей функции.

Делеции участков хромосом, содержащих гены-супрессоры, часто приводят к их инактивации. Например, ген *RB1*, регулирующий переход клетки из G1 в S-фазу цикла [Dyson, 2016], часто инактивируется делециями

при ретинобластоме, остеосаркомах и аденокарциномах лёгкого. Потеря *RB1* (13q14) приводит к неконтролируемой активации транскрипционного фактора E2F и ускоренной пролиферации опухолевых клеток [Engeland, 2022]. Делеция хромосомного локуса 17p13.1, содержащего ген *TP53*, наблюдается в 50% случаев колоректальной аденокарциномы и ассоциирована с потерей контроля над апоптозом и геномной стабильностью [Vogelstein, Lane, Levine, 2000]

Эпигенетическое подавление генов-супрессоров опухолей происходит главным образом через гиперметилирование ДНК и модификации гистонов, что приводит к компактизации хроматина и ингибированию экспрессии генов-супрессоров. Гиперметилирование промоторов (напр., генов *Rb1*, *p16*, *MLH1*, *BRCA1*) эквивалентно их инактивации и нарушает контроль клеточного цикла, репарацию ДНК, апоптоз и адгезию [Baylin, 2005; Jones, Baylin, 2007, 2002].

#### 1.6.1 Ключевые признаки злокачественных заболеваний

Концепция ключевых признаков злокачественных заболеваний (hallmarks of cancer), предложенная Дугласом Ханаханом (Douglas Hanahan) и Робертом Вайнбергом (Robert A. Weinberg) [Hanahan, Weinberg, 2000], систематизирует молекулярные механизмы и процессы, протекающие в клетках в процессе злокачественной трансформации. Одним из первых признаков опухолевых клеток является **способность к автономной пролиферации**, реализуемая через постоянную активацию сигнальных путей, контролирующих клеточный цикл. [Sever, Brugge, 2015]. Параллельно формируется **резистентность к антипролиферативным сигналам** за счёт инактивации генов-супрессоров (*TP53*, *RB1*, *PTEN*) [Burkhardt, Sage, 2008; Lee, Chen, Pandolfi, 2018; Olivier, Hollstein, Hainaut, 2010] и гиперактивации циклин-зависимых киназ [Malumbres, Barbacid, 2009]. **Уклонение от запрограммированной клеточной гибели** позволяет опухолевым клеткам выживать в условиях стресса (повреждение ДНК, гипоксия) [Singh, Letai, Sarosiek, 2019]. Этот эффект связан с нарушениями в митохондриальном и рецепторном путях апоптоза, включая гиперэкспрессию антиапоптотических генов семейства Bcl-2 [Degenhardt et al., 2002; Delbridge, Strasser, 2015; Karlberg et al., 2010; Ren et

al., 2010], инактивацию p53, который мутирует более чем в половине опухолей и в норме является мощным индуктором транскрипции проапоптотических генов (например, *BAX*, *PUMA*) [Hafner et al., 2019], а также каспаз [Garg et al., 2016; Gyrd-Hansen, Meier, 2010; Lavrik, 2011] и подавление альтернативных форм гибели, таких как некроптоз [Gong et al., 2019; Zhang et al., 2022]. Аутофагия, в свою очередь, может либо способствовать выживанию клеток, либо ограничивать рост опухоли на ранних этапах [Morselli et al., 2009; White, 2015].

**Неограниченный репликативный потенциал** опухолевых клеток обеспечивается реактивацией теломеразы – фермента, достраивающего теломерные повторы и предотвращающего их критическое укорочение. В норме экспрессия генов каталитической субъединицы теломеразы (hTERT) минимальна, но в большинстве опухолей она значительно увеличена, что способствует клеточному "бессмертию" [Kyo et al., 2008; Poremba et al., 2002].

Дополнительно к поддержанию длины теломер, опухолевые клетки обходят контрольные точки клеточного цикла [Engeland, 2022]. Ключевыми механизмами служат инактивация ретинобластного белка (pRB) – регулятора перехода G1/S – и гиперфосфорилирование его CDK4/6-циклинов D комплексами [Malumbres, Barbacid, 2009; Massagué, 2004]. Параллельная инактивация p53 лишает клетку механизма остановки цикла в ответ на стресс или повреждение ДНК, позволяя ей делиться даже при наличии мутаций. Таким образом, опухолевые клетки приобретают способность к бесконтрольной пролиферации, игнорируя антипролиферативные сигналы [Massagué, 2008].

**Индукция ангиогенеза**, необходимая для снабжения опухоли кислородом и питательными веществами, запускается усиленной экспрессией *VEGF* и *FGF* на фоне снижения ангиостатических факторов [Carmeliet, Jain, 2011].

Новообразованные сосуды при этом морфологически и функционально отличаются от нормальных, однако критически важны для роста и метастазирования опухоли, делая сам процесс ангиогенеза и его регуляторы

потенциальными мишенями для противоопухолевой терапии [Rajabi, Mousa, 2017]. **Инвазия и метастазирование** реализуются через эпителиально-мезенхимальный переход, инициируемый сигналами микроокружения (например, TGF- $\beta$ ) [Dongre, Weinberg, 2019; Lee, Hwang, Choi, 2017], потерю адгезивных свойств [Loh et al., 2019], активацию металлопротеиназ [Castro-Castro et al., 2016] и ремоделирование внеклеточного матрикса [Liu, Cao, 2016].

В 2011 году Hanahan и Weinberg дополнили концепцию четырьмя новыми признаками [Hanahan, Weinberg, 2011]. Среди них – **репрограммирование клеточного метаболизма**, при котором опухолевые клетки переходят на аэробный гликолиз (эффект Варбурга) для удовлетворения потребностей в биосинтезе [Liberti, Locasale, 2016]. Это сопровождается активацией гликолитических ферментов и снижением митохондриального окислительного фосфорилирования [Shiratori et al., 2019; Weinberg et al., 2010]. Опухолевые клетки также проявляют способность к **уклонению от иммунного ответа**, подавляя активность Т- и NK-клеток [Kim, Emi, Tanabe, 2007]. Это достигается снижением экспрессии *MHC I*, секрецией иммуносупрессивных цитокинов и экспрессией *PD-L1*, что приводит к функциональному истощению цитотоксических лимфоцитов [Dhatchinamoorthy, Colbert, Rock, 2021].

Еще один признак злокачественности – **геномная нестабильность** выражается в повышенной частоте мутаций в ДНК опухолевых клеток по сравнению с нормальными соматическими клетками и охватывает широкий спектр генетических нарушений, включая точечные мутации, вставки, делеции, амплификации генов, хромосомные перестройки (транслокации, инверсии) и изменения числа хромосом (анеуплоидия) [Hanahan, Weinberg, 2011].

Центральную роль в геномной нестабильности играют дефекты в системах репарации ДНК. Нарушение работы путей, ответственных за исправление ошибок репликации (неспаренные основания – mismatch repair, MMR), устранение двунитевых разрывов (гомологичная рекомбинация – HR,

негомологичное соединение концов – NHEJ) или репарацию повреждений, вызванных экзогенными факторами (например, нуклеотид-экцизионная репарация – NER), ведет к закреплению мутаций в геноме. Последствия геномной нестабильности носят каталитический характер для развития опухолей. Она значительно увеличивает вероятность возникновения и накопления "драйверных" мутаций – изменений в генах, непосредственно способствующих неконтролируемой пролиферации, выживанию, инвазии и метастазированию (онкогены, гены-супрессоры опухолей, гены, связанные с апоптозом, ангиогенезом).

**Хроническое воспаление, сопровождающее опухоль,** активирует онкогенные транскрипционные факторы, включая NF-κB и STAT3 [Landskron et al., 2014; O'Shea, Plenge, 2012]. Их конститутивная активация, индуцируемая провоспалительными цитокинами (например, TNF-α, IL-1β), бактериальными продуктами или онкогенным стрессом, приводит к экспрессии антиапоптотических факторов (например, *BCL-2*, *BCL-XL*), провоспалительных цитокинов (*IL-6*, *IL-8*) и активных форм кислорода [Landskron et al., 2014; Zhao et al., 2021], усиливающих геномную нестабильность и поддерживающих опухолевое микроокружение [Landskron et al., 2014; O'Byrne, Dalglish, 2001].

Таким образом, злокачественная трансформация представляет собой результат кумулятивного нарушения регуляции ключевых генов – активации онкогенов и инактивации супрессоров опухолей посредством мутаций, хромосомных перестроек и эпигенетических механизмов. Эти первичные повреждения конвергируют в приобретение клетками набора ключевых признаков злокачественности (по Hanahan & Weinberg). Фенотипическая реализация этих дефектов и признаков злокачественности осуществляется преимущественно через дисрегуляцию внутриклеточных сигнальных путей, что детально рассматривается в следующей подглаве.

#### 1.6.2 Сигнальные пути и нарушения их работы в онкологических заболеваниях

Сигнальные пути клеток – это сложные молекулярные цепочки событий, передающие внешние сигналы к внутриклеточным мишеням и регулирующие ключевые процессы, включая рост, дифференцировку и выживание [Sever, Brugge, 2015]. Нарушения в этих системах, вызванные мутациями или эпигенетической инактивацией, играют центральную роль в онкогенезе.

**Сигнальный путь RAS/MAPK** регулирует пролиферацию и дифференцировку клеток, а его активация инициируется рецепторами тирозинкиназ в ответ на факторы роста [Simanshu, Nissley, McCormick, 2017]. Белки RAS действуют как молекулярные переключатели, активируя каскад RAF–MEK–ERK, результатом чего становится транскрипция генов, контролирующих клеточный цикл [Samatar, Poulikakos, 2014]. Активность пути строго контролируется механизмами отрицательной обратной связи, включая дефосфорилирование компонентов пути и деградацию сигнальных молекул [Lake, Corrêa, Müller, 2016].

Однако мутации в *RAS* (*KRAS*, *NRAS*, *HRAS*) или *BRAF* нарушают контроль, вызывая конститутивную активацию сигнала, что ведёт к неконтролируемой пролиферации [Prior, Hood, Hartley, 2020; Schubbert, Shannon, Bollag, 2007]. Примером служит мутация *BRAF* V600E, характерная для половины случаев меланомы и являющаяся мишенью таргетной терапии [Chapman et al., 2011; Davies et al., 2002; Platz et al., 2008].

**Сигнальный путь PI3K/AKT/mTOR** является центральным регулятором клеточного роста, метаболизма, выживаемости и пролиферации [Fruman et al., 2017]. Он активируется через рецепторы тирозинкиназ и далее задействует PI3K, образующий PIP3 – ключевой мессенджер для активации AKT и последующей стимуляции mTOR [Glaviano et al., 2023; Laplante, Sabatini, 2012]. Активность пути в норме контролируется негативными регуляторами, включая фосфатазу PTEN [Lee, Chen, Pandolfi, 2018].

Гиперактивация пути PI3K/AKT/mTOR – одно из наиболее частых молекулярных нарушений в опухолях, ассоциированное с резистентностью к апоптозу, усиленным ангиогенезом и метастазированием [Fruman, Rommel,

2014; Porta, Paglino, Mosca, 2014; Glaviano et al., 2023]. При протоковой карциноме молочной железы мутации гена *PIK3CA* (кодирующего каталитическую субъединицу PI3K), выявляемые в 30-40% случаев, приводят к конститутивной активации пути независимо от внешних стимулов [Samuels et al., 2004].

**Сигнальный путь Wnt/ $\beta$ -катенин**, в норме регулирующий эмбриогенез и тканевой гомеостаз, также критичен для канцерогенеза. В отсутствие сигнала  $\beta$ -катенин деградирует в составе комплекса APC/Axin/GSK3 $\beta$  [Nusse, Clevers, 2017]. Связывание Wnt с рецепторами Frizzled/LRP5/6 ингибирует этот комплекс, позволяя  $\beta$ -катенину накапливаться в цитоплазме и транспортироваться в ядро [MacDonald, Tamai, He, 2009]. В ядре  $\beta$ -катенин взаимодействует с транскрипционными факторами TCF/LEF, стимулируя экспрессию генов, контролирующих пролиферацию (например, *c-Myc*, *Cyclin D1*) и судьбу клеток [Anastas, Moon, 2013].

**Сигнальный путь JAK-STAT** обеспечивает передачу сигналов от цитокинов и факторов роста, регулируя процессы клеточной пролиферации, дифференцировки и иммунного ответа [Rawlings, Rosler, Harrison, 2004]. Активация начинается с связывания лиганда с рецептором, что запускает фосфорилирование тирозинкиназами JAK белков STAT [O'Shea, Plenge, 2012]. В норме активность пути ограничена супрессорами, такими как SOCS (Suppressors of Cytokine Signaling) [Alexander, Hilton, 2004].

При онкогенезе этот путь часто подвергается гиперактивации, особенно в гематологических злокачественных новообразованиях. Так, мутация *JAK2* V617F обуславливает автономную пролиферацию миелоидных клеток, а постоянная активация STAT3 и STAT5 в лимфомах поддерживает выживание опухолевых клеток. [Baxter et al., 2005]. Кроме того, STAT3 модулирует противоопухолевый иммунный ответ, подавляя активность дендритных клеток и Т-лимфоцитов, что способствует формированию иммуносупрессивного микроокружения. [Yu et al., 2014].

## 1.7 Заключение к литературному обзору

Филостратиграфический анализ представляет собой инструмент для изучения эволюционной истории генов и их функций. Прослеживая линии генов до их первого общего предка, можно получить представление о происхождении и эволюции различных заболеваний, в том числе – злокачественных новообразований.

В частности, этот подход показал значимые результаты в изучении онкологических заболеваний. Было установлено, что многие гены, вовлечённые в процессы онкогенеза, являются «древними», то есть они впервые появились у ранних форм жизни и сохранялись на протяжении всей эволюционной истории. Это позволяет предположить, что многие молекулярные механизмы, лежащие в основе злокачественной трансформации клеток, глубоко укоренены в эволюционном прошлом.

На первый взгляд, эволюционно ранее происхождение генов, участвующих в онкогенезе, может оказаться неожиданным, поскольку онкологические заболевания у человека часто рассматриваются как следствия возрастных изменений или воздействий окружающей среды. Однако это представляется обоснованным, если принять во внимание, что эти гены играют фундаментальную роль в процессах роста, деления и выживания клеток – процессах, которые являются ключевыми для жизни.

Например, многие гены, участвующие в делении клеток и репарации ДНК – в процессах, нарушения которых лежат в основе развития злокачественных новообразований, – имеют высокий уровень консервативности у разных видов. Это свидетельствует о том, что они выполняют важнейшие функции, которые сохранялись на протяжении сотен миллионов лет эволюции.

Понимание эволюционной истории генов, вовлечённых в онкогенез, может также способствовать определению новых терапевтических мишеней. Если ген сохранялся в ходе эволюции, это указывает на его ключевое значение в функционировании клеток и потенциальную ценность как мишени для терапии онкологических заболеваний. Кроме того, филостратиграфический

подход позволяет глубже понять молекулярные основы злокачественной трансформации. Анализ эволюции генов, вовлечённых в онкогенез, и их функций у различных организмов, способствует более глубокому пониманию механизмов, лежащих в основе развития и прогрессирования опухолей.

В заключение следует отметить, что применение филогенетики в исследованиях онкологических заболеваний позволяет получить ценные сведения о происхождении и эволюционной истории генов, вовлечённых в онкогенез. На основе проведённого литературного обзора и обсуждения ключевых аспектов эволюционного анализа возраста генов, мы переходим к определению конкретных целей нашего исследования. В представленной работе наша цель заключается в изучении эволюционных индексов генов, ассоциированных с разными стадиями онкологических заболеваний.

## 2. Материалы и методы

### 2.1 Материалы исследования

Данные по экспрессии генов в тканях злокачественных опухолей были получены из базы данных The Cancer Genome Atlas (TCGA – <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>). Эти данные представляют собой количество транскриптов, картированных на каждый ген, полученные в результате секвенирования транскриптомов из образцов тканей опухолей и здоровых тканей. В этой работе мы использовали данные по экспрессии генов в тканях печёночной карциномы, светлоклеточной карциномы почек, аденокарциномы кишечника, карциномы молочной железы, аденокарциномы простаты, карциномы щитовидной железы, карциномы уротелия мочевого пузыря, эндометриоидной карциномы тела матки и из соответствующих им образцов здоровых тканей от тех же пациентов (Табл. 3). Помимо данных по экспрессии генов, из базы данных TCGA также были использованы данные, содержащие клинические наблюдения о состоянии пациентов и опухолей.

Образцы тканей из TCGA классифицируются по типам опухолей, а также по типам образцов (например, первичная опухоль, нормальная ткань). Образцы нормальных тканей (называемые NAT – Normal Adjacent Tissues или нормальные, прилегающие к опухоли ткани) берут из гистологически нормальных тканей, прилегающих к краям опухоли, и часто используют в качестве контроля для сравнения с образцами опухоли. Несмотря на то, что было показано, что уровни экспрессии ряда генов в них могут несколько отличаться от истинно здоровых тканей [Aran и др., 2017; Chandran и др., 2005; Huang, Stern, Zhao, 2016; Sanz-Pamplona и др., 2014], было принято решение использовать их в качестве контроля, чтобы избежать возможного влияния

групповых эффектов при использовании данных из тканей здоровых людей, взятых в рамках других протоколов.

**Таблица 3.** Количество образцов, для которых были получены данные по экспрессии генов в базе данных TCGA.

| TCGA_id | Опухоль                                    | I<br>стадия | II<br>стадия | III<br>стадия | IV<br>стадия | NAT |
|---------|--------------------------------------------|-------------|--------------|---------------|--------------|-----|
| LIRC    | Гепатоцеллюлярная<br>карцинома печени      | 255         | 3            | 79            | 3            | 50  |
| UCEC    | Эндометриоидная<br>карцинома тела<br>матки | 338         | 53           | 84            | 29           | 35  |
| COAD    | Аденокарцинома<br>кишечника                | 168         | 185          | 116           | 22           | 41  |
| BRCA    | Карцинома<br>молочной железы               | 215         | 639          | 249           | 13           | 112 |
| PRAD    | Аденокарцинома<br>простаты                 | 176         | 174          | 53            | 2            | 52  |
| KIRC    | Светлоклеточная<br>карцинома почки         | 417         | 123          | 124           | 84           | 72  |
| LUAD    | Аденокарцинома<br>лёгкого                  | 298         | 127          | 84            | 27           | 59  |
| BLCA    | Карцинома<br>уротелия мочевого<br>пузыря   | 4           | 130          | 142           | 138          | 18  |
| THCA    | Карцинома<br>щитовидной железы             | 284         | 52           | 112           | 55           | 59  |

## 2.2 Методы

В данной главе описаны программные средства и методы, использованные для реализации вычислительного конвейера `cancEv` для анализа транскриптомных данных. Основными инструментами, применяемыми в ходе работы, были языки программирования Python и R, а также `bash`-скрипты. Схема вычислительного конвейера `cancEv` представлена на Рис. 7.

### Идентификация дифференциально экспрессирующихся генов

Поиск дифференциально экспрессирующихся генов проводился при помощи программного пакета *DESeq2* [Ritchie и др., 2015]. *DESeq2* – это программный пакет, разработанный для анализа данных секвенирования, который позволяет проводить количественную оценку экспрессии генов и выявлять дифференциально экспрессирующиеся гены между различными экспериментальными условиями. *DESeq2* использует статистический подход, основанный на модели отрицательного биномиального распределения, для моделирования подсчётов чтений и оценки вариабельности данных. В рамках экспериментального дизайна, целью было выявить гены, демонстрирующие дифференциальную экспрессию между образцами, взятыми от пациентов на определённой стадии развития опухоли, и образцами нормальной ткани, полученными от здоровых участков, соседствующих с опухолью.

В процессе проведения экспериментов по секвенированию РНК (RNA-seq) для анализа дифференциальной экспрессии генов часто возникают систематические ошибки, связанные с различиями в условиях проведения экспериментов, обработке образцов или платформах секвенирования. Эти ошибки известны как групповой эффект (*batch effects*). Неучтенные групповые эффекты могут привести к значительным искажениям в данных, что затрудняет интерпретацию результатов. Для снижения систематических различий в уровнях экспрессии, связанных с техническими различиями, в нашей работе проводилась корректировка группового эффекта путем

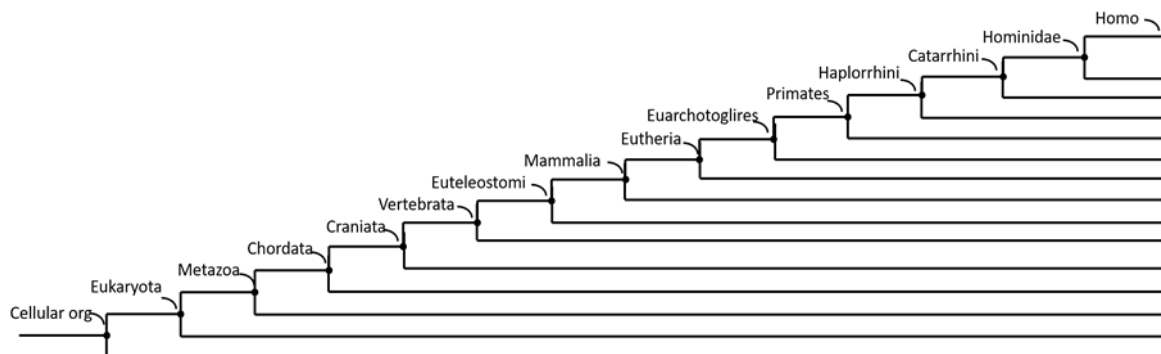
включения данных о номере эксперимента из столбца *tcga.cgc\_case\_batch\_number* в дизайн модели *DESeq2*.

Для отбора дифференциально экспрессирующихся генов (ДЭГ) был использован порог степени изменения экспрессии  $\log FC > 1$  и  $< -1$ , что соответствует изменению экспрессии в два раза, и при учете уровня ложных срабатываний менее 5%. Оценка ложных срабатываний проводилась по *p-value* с поправкой на множественное тестирование Бенджамини-Хохберга ( $\text{adj.p.val} < 0.05$ ).

### Определение эволюционных характеристик генов

Для изучения эволюционных процессов и филогенетического анализа генетических данных использовалось веб-приложение Orthoweb 2. В основе анализа лежит общепринятая таксономия NCBI, представленная в базах данных KEGG в виде 16 таксономических рангов (Табл.4).

**Таблица 4.** Значения PAI и соответствующие им таксономические ранги из таксономии NCBI



| PAI | Таксон             | Возраст (млн) | PAI | Таксон              | Возраст (млн) |
|-----|--------------------|---------------|-----|---------------------|---------------|
| 1   | Cellular Organisms | 4100          | 9   | Eutheria            | 160           |
| 2   | Eukaryota          | 1850          | 10  | Euarchontoglires    | 65            |
| 3   | Metazoa            | 665           | 11  | Primates            | 55            |
| 4   | Chordata           | 541           | 12  | Haplorrhini         | 50            |
| 5   | Craniata           | 535           | 13  | Catarrhini          | 44            |
| 6   | Vertebrata         | 525           | 14  | Hominidae           | 17            |
| 7   | Euteleostomi       | 420           | 15  | <i>Homo</i>         | 2.8           |
| 8   | Mammalia           | 225           | 16  | <i>Homo sapiens</i> | 0.35          |

Для выбранных на основе анализа дифференциальной экспрессии белок-кодирующих генов на этапе предобработки данных были рассчитаны индексы эволюционных характеристик генов.

Ниже представлено детальное описание настроек и процесса анализа, используемых для каждого из инструментов веб-приложения.

### **Разработка программного комплекса Orthoweb 2**

Orthoweb 2 разработан на языке программирования Java с использованием фреймворка *Spring* для реализации серверной части приложения и фреймворков *vue/webix* – для реализации клиентской части. Для визуализации сетей используется группа библиотек *cytoscape.js*. Документоориентированная СУБД используется для хранения данных из базы KEGG (таксоны, список ортологов, белок-кодирующие последовательности и т.д.) и промежуточных результатов анализа. Это значительно увеличивает скорость дальнейшей работы с этими же данными. Примеры структуры запросов REST API к базе данных приведены в Приложении А.

Для хранения результатов анализа была создана база данных с использованием PostgreSQL. База данных состоит из 5 таблиц: организмы, гены, Индексы PAI, Индексы DI, термины генной онтологии (идентификаторы и наименования), SNP, Индексы PAI, определенные на основе КО-групп. Таким образом мы имеем доступ к имеющейся информации по всем посчитанным PAI и DI индексам для генов конкретных организмов. Программный интерфейс доступа к этим данным реализован как REST API приложение в рамках технологии FLASK (<https://flask.palletsprojects.com/en/1.1.x/>).

Веб-приложение Orthoweb 2 использовалось для расчёта эволюционного возраста генов, рассчитанных в единицах PAI [Mustafin, Lashin, Matushkin, 2021]. Для расчёта PAI использовался метод расчёта по таблицам идентичности, при этом использовался порог идентичности

последовательностей в 60% [Pevsner, 2015]. Индекс дивергенции (DI) рассчитывался на основе значений отношения несинонимичных замен к синонимичным ( $d_N/d_S$ ), полученных при попарном сравнении белок-кодирующих последовательностей ортологов *Homo sapiens* с ортологами видов *Gorilla gorilla gorilla*, *Pongo abelii* и *Pan troglodytes*.

### Филотранскриптомный анализ

Анализ филотранскриптомных индексов производился при помощи программного пакета *myTAI* [Drost и др., 2018]. Этот инструмент позволяет вычислять индекс возраста транскриптома (TAI) и индекс дивергенции транскриптома (TDI), за счёт использования внешней информации о индексе филогенетического возраста (PAI) и дивергенции (DI), полученной из сторонних программ, а также оценивать статистическую значимость наблюдаемых паттернов. Для корректной работы пакета значения индексов дивергенции  $d_N/d_S$  были ранжированы соответственно их децилю (10% квантилю) следующим образом: Первый дециль (Divergence Stratum 1) включает в себя 10% наименьших значений  $d_N/d_S$ , а 10 дециль (= Divergence Stratum 10) – наибольшие значения  $d_N/d_S$  между 91-100% квантилем.

Данные о возрасте генов, индексах дивергенции, полученные из Orthoweb 2 и значения экспрессии генов для образцов разных клинических стадий опухолей объединялись в два набора данных для расчёта филотранскриптомных индексов. Профили TAI и TDI для каждого типа злокачественных заболеваний были визуализированы и оценены с использованием теста FlatLineTest на отсутствие значимых паттернов. Для типов опухолей, у которых профиль TAI/TDI статистически значимо отличался от равномерного распределения (FlatLineTest  $p < 0.05$ ), был применен *редукционный тест песочных часов* (ReductiveHourglassTest) и *обратный тест песочных часов* (Reverse Hourglass Test) [Drost и др., 2015], для проверки гипотезы о наличии характерного паттерна «песочных часов» в онтогенезе опухоли, предполагающего максимальную консервативность

транскриптома на ранних и поздних стадиях опухолевого процесса по сравнению со средними стадиями (или наоборот, для обратного теста)."

### **Функциональный анализ генов**

FoldGO – это веб-инструмент для анализа функционального обогащения транскриптома в зависимости от степени изменения экспрессии (доступен по адресу: <http://webfsgor.sysbio.cytogen.ru/>) [Wiebe et al., 2020].

FoldGO был использован для выявления терминов генной онтологии (GO – Gene Ontology), обогащенных в наборах генов с согласованными изменениями экспрессии в различных стадиях клинического развития опухолей. Для корректировки порогов значимости обогащения при множественном тестировании применялась поправка Бенджамини-Хохберга ( $FDR < 0,05$ ).

Результаты анализа FoldGO были визуализированы в виде тепловой карты (heatmap), где по оси X представлены четыре стадии каждого из девяти типов злокачественных заболеваний, а по оси Y – перечисленные выше термины ГО. Каждая ячейка отображает медианное значение  $\log FC$ , отражающее направление и величину сдвига экспрессии в соответствующей функциональной категории на конкретной стадии опухоли. Визуализация выполнена с использованием Python-пакета *seaborn* [Waskom и др., 2017].

### 3. Результаты

#### 3.1 Разработка веб-сервиса Orthoweb 2 для филогенетического и транскриптомного анализа возраста генов

На основе ранее разработанного приложения Orthoscape, представляющего собой расширение программной платформы Cytoscape, в 2020 году в нашей лаборатории было создано независимое веб-приложение Orthoweb 2 (<https://orthoweb.sysbio.cytogen.ru/run.html>), позволяющее определять макро- и микроэволюционные характеристики генов в формате эволюционных индексов PAI и DI.

В рамках диссертационной работы был разработан веб-сервис Orthoweb 2, предназначенный для проведения филогенетического анализа возраста генов. В новой версии реализована расширенная функциональность, включающая поддержку дополнительных типов данных и аналитических методик, направленных на повышение точности и информативности проводимого анализа.

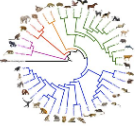
Ключевым нововведением стало внедрение нескольких методов отбора ортологичных генов. Первый метод позволяет выбирать ортологичные гены на основе данных групп ортологов KEGG (KEGG Orthology, KO), которые уже предварительно отобраны в базе и не требуют дополнительных настроек от пользователя. Однако стоит отметить, что данные доступны не для всех генов; например, для человека примерно две трети генов, представленных в KEGG, ассоциированы с KO-группой.

Если пользователь отбирает ортологи с помощью KO групп, то он может добавить фильтрацию по наименованию гена, чтобы исключить из KO-групп паралогичные гены. Тогда входящие в KO группу гены будут также проверяться на совпадение наименования генов.

Вторым способом отбора ортологов является использование таблицы сходства последовательностей (Best Similarity Table), доступной для

подавляющего большинства генов, представленных в KEGG. По данной таблице пользователь может отобрать ортологичные гены на основе таких параметров, как сходство аминокислотных последовательностей кодируемых генами белков и результат работы алгоритма Смита-Ватермана по локальному выравниванию последовательностей, однако пользователю необходимо самостоятельно предоставить Orthoweb 2 пороговые значения этих параметров для отбора генов.

Кроме того, для списков генов в Orthoweb 2 были реализованы дополнительные режимы анализа, такие как анализ ассоциированных терминов генной онтологии (ГО) или обогащения SNP (однонуклеотидными полиморфизмами). Эти функции можно активировать в соответствующих полях интерфейса (Рис. 5). Также была реализована опция использования онлайн баз данных, что может замедлить работу программы, но обеспечит использование актуальных данных, что особенно важно, если данные по исследуемым генам могли недавно измениться. Для получения информации по терминам генной онтологии используется ресурс <http://geneontology.org/> [Ashburner и др., 2000; Carbon и др., 2021]. Информация загружается через предоставляемый API (Application Programming Interface, интерфейс программного доступа к ресурсам). Например, запрос для гена TBP имеет вид <http://api.geneontology.org/api/bioentity/gene/NCBIGene:6908/function>, в нем указана база данных и идентификатор гена в ней. Orthoweb 2 предоставляет эту информацию самостоятельно, для большинства модельных организмов используются связанные с ними базы данных (например, TAIR для *Arabidopsis thaliana* [Berardini и др., 2015], FlyBase для *Drosophila melanogaster* [Gramates и др., 2022], а для остальных организмов – база UniProt.



Welcome to OrthoWeb. On this page you can launch the evolutionary analysis of gene sets.

---

Work ID:  [Use ID converter](#) ?

---

Setup parameters or use the defaults

|                                                                                                                                                 |                                                                                                                             |
|-------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|
| <p><i>The type of orthology</i> ?</p> <p><input checked="" type="radio"/> KEGG Orthology groups <input type="radio"/> Best Similarity Table</p> | <p><i>KO groups filtering</i> ?</p> <p><input checked="" type="radio"/> All genes <input type="radio"/> Only same label</p> |
|-------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|

---

|                                                                                                                                                              |                                                                                                                       |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|
| <p><i>The thresholds to filter orthologous genes</i> ?</p> <p>Identity: <input type="text" value="0,5"/></p> <p>SW Score: <input type="text" value="0"/></p> | <p><i>dN/dS setup</i> ?</p> <p>dN/dS level: <input type="text" value="2"/></p> <p>Organisms: <input type="text"/></p> |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|

---

|                                                                                                                                                                                                                             |                                                                                                                                                                                                                         |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><i>Additional parameters</i> ?</p> <p><input checked="" type="checkbox"/> DI analysis <input type="checkbox"/> GO analysis</p> <p><input type="checkbox"/> SNP analysis <input type="checkbox"/> Use online database</p> | <p><i>Choose the type of input data</i> ?</p> <p><input checked="" type="radio"/> Form <input type="radio"/> Gene list file <input type="radio"/> Network file</p> <p>Genes: <input type="text" value="hsa:57018"/></p> |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

**Рисунок 5** Страница Orthoweb 2 (<https://orthoweb.sysbio.cytogen.ru/>), на которой можно настроить параметры и провести анализ списков генов.

Если в базе Gene Ontology есть данные по исследуемому гену и в KEGG имеется необходимый идентификатор, что выполняется практически для всех хорошо изученных генов, то будут загружены идентификаторы терминов ГО, ассоциированных с геном и их наименования. Если данных все-таки не нашлось, будет получено значение «No data».

Данные об общем количестве однонуклеотидных полиморфизмов в последовательности исследуемого гена загружаются с ресурса NCBI, где содержится обширная база SNP [Sherry, 2001]. Запрос конструируется на основании идентификатора гена и, например, для гена ТВР с идентификатором hsa:6908, имеет следующий вид: [https://www.ncbi.nlm.nih.gov/snp/?term=6908\[Gene\\_ID\]](https://www.ncbi.nlm.nih.gov/snp/?term=6908[Gene_ID]). В результате запроса пользователю предоставляется количество найденных SNP. Если данных не было найдено, то пользователь получит ответ «No data». Следует отметить, что поиск SNP осуществляется только для генов человека.

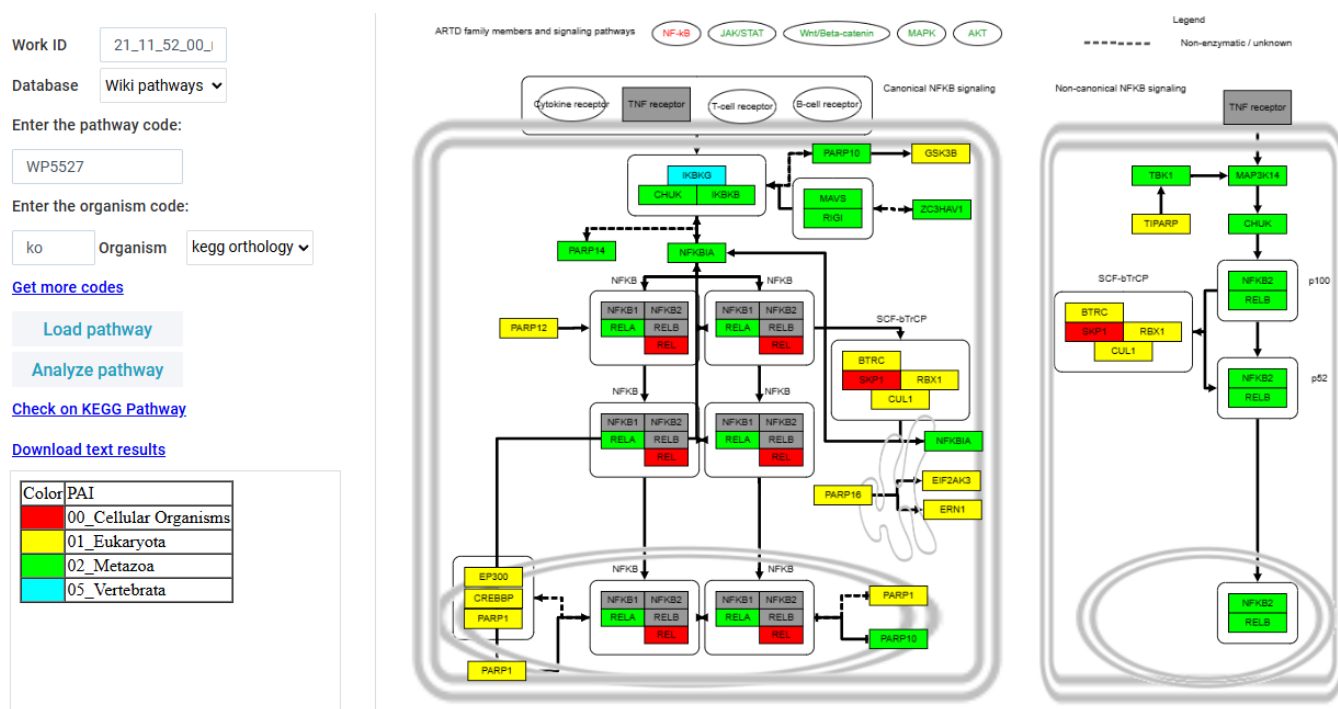
В рамках новых возможностей добавлена функция анализа эволюционных характеристик транскриптомов (TAI, TDI) на основе данных об экспрессии генов. В отличие от данных об ассоциациях с терминами ГО, обогащении SNP или нуклеотидных последовательностях, которые могут быть получены автоматически из внешних баз данных, для расчёта индексов TAI/TDI обязательно требуется предоставление файла с данными экспрессии генов. Данные по экспрессии генов должны быть представлены в файле табличного вида. Первая колонка содержит идентификаторы генов. Каждая последующая колонка соответствует одному образцу или условию (в заголовке колонки) и содержит значения уровня экспрессии генов (в строках) для этого образца/условия.

Кроме того, в Orthoweb 2 был реализован модуль для эволюционного анализа генных сетей. Доступен как анализ импортированных сетей из баз данных KEGG PATHWAY и WikiPathways, так и анализ сетей, импортированных из текстового файла.

Анализ сетей из баз данных доступен по ссылке (<https://orthoweb.sysbio.cytogen.ru/pathway.html>). Реализован импорт и анализ сетей из двух баз данных. Первая из них – KEGG Pathway, содержит множество генных сетей и метаболических путей, разбитых на группы (описывающие метаболизм, функционирование различных систем организма, заболевания человека и т.д.). Для начала анализа пользователю необходимо указать код метаболического пути и организм, для которого будет импортирована сеть. Сети в KEGG PATHWAY строятся на основе КО групп и могут быть адаптированы на любой организм, чьи гены представлены в КО группе, показанной на сети. В результате анализа пользователь получит генную сеть, раскрашенную в соответствии со значением PAI, определенным на основании представленных в сети КО групп.

Второй базой данных, для которой реализован импорт сетей, является WikiPathways. Сети, представленные в WikiPathways, содержат значительно больше деталей, сущностей и вариантов взаимодействий, чем в KEGG, в связи

с чем полноценный их импорт затруднён, в сравнении с импортом сетей из KEGG Pathway, а анализ требует учёта варианта идентификаторов из множества различных баз данных. Импорт и анализ сетей WikiPathways в Orthoweb 2 может отображаться менее наглядно из-за дополнительных этапов обработки при импорте. На рисунке 6 представлен пример импорта сети из WikiPathways, на которой отражена большая часть информации, включая ключевые генные продукты (белки) и их взаимодействия, импортированные из WikiPathways. Дополнительные элементы визуализации исходной сети, такие как фоновые изображения клеточных структур или компартментов, при импорте в Orthoweb 2 могут не отображаться.



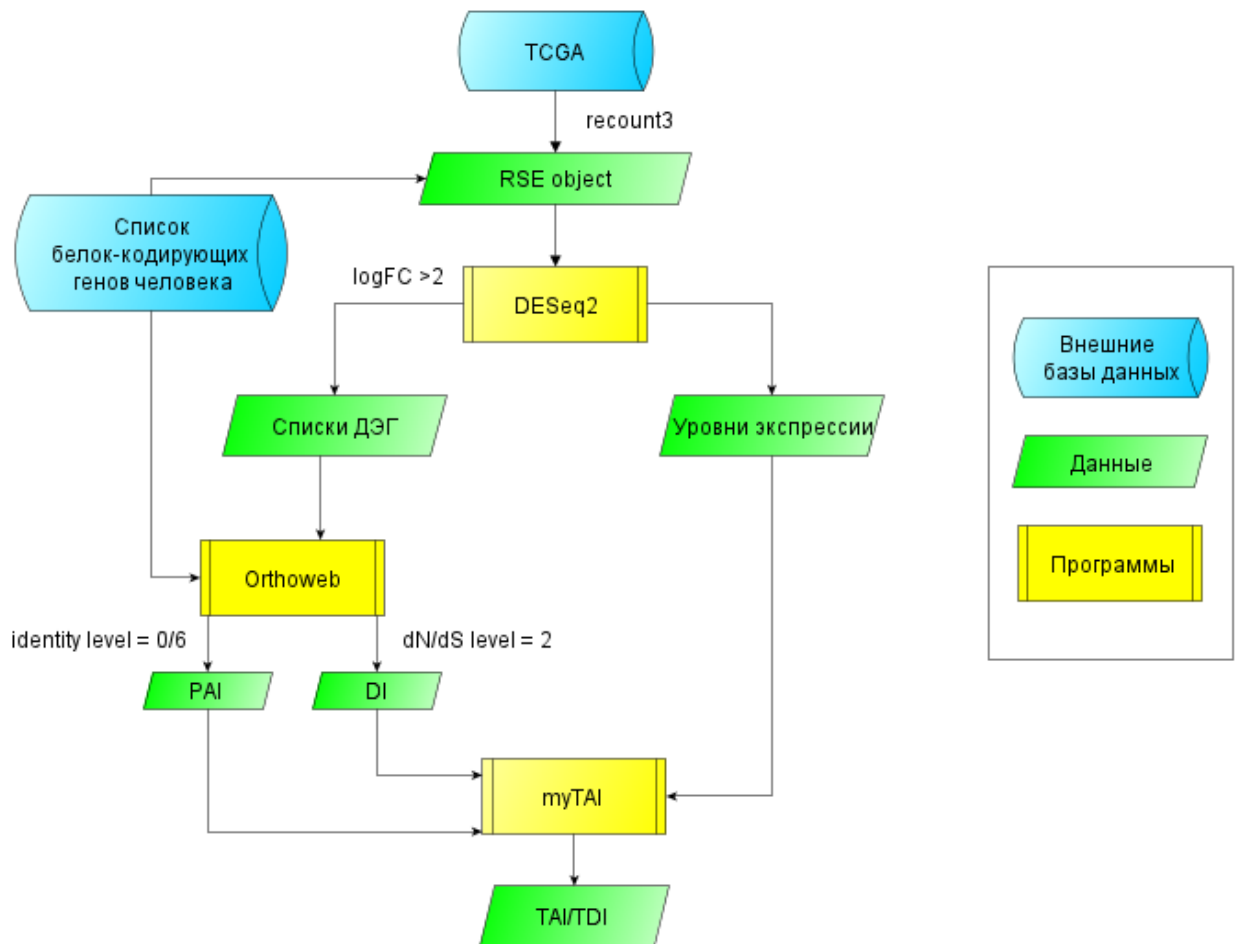
**Рисунок 6** Пример импорта и анализа генной сети сигнального пути NF-κB signaling, вовлечённого в регуляцию молекулярных механизмов канцерогенеза (WP5527) из WikiPathways. Узлы окрашены в соответствии с их PAI.

Также была реализована функция эволюционного анализа генных сетей, подготовленных пользователем. Процедуры импорта и анализа в Orthoweb 2 идут отдельными шагами, можно сначала импортировать сеть и взаимодействовать с ней, например, перемещать элементы, а уже потом запускать анализ. Для импорта сети пользователю необходимо подготовить

файл в формате tsv. Файл должен содержать три столбца с данными. В первом столбце содержатся узлы, из которых выходят рёбра, во втором – в которые эти рёбра входят, и в третьем – вес ребра между ними. Формат файла и название столбцов для данных были по аналогии с ресурсом для реконструкции генных сетей – STRING [Mering von и др., 2005]. Таким образом, пользователь может не только визуализировать самостоятельно подготовленный набор генов и взаимодействий между ними, но и, имея на руках только список генов, воспользоваться средством STRING для поиска взаимодействий между ними. Благодаря ориентации на STRING, данные, полученные с его помощью, могут быть импортированы в Orthoweb 2 без дополнительной обработки пользователем. Если пользователь решит самостоятельно подготовить такой файл, то для совместимости с выходными данными STRING и упрощения импорта, рекомендуется использовать названия столбцов #node1 (содержит узлы, из которых выходят рёбра), node2 (содержит узлы, в которые эти рёбра входят) и combined\_score (содержит интересующую пользователя метрику оценки ребра). В случае сети, импортированной из STRING, колонка combined\_score содержит надежность найденного взаимодействия с весом от 0 до 1. После анализа цвет генов изменится в соответствии с показателем PAI, а размер в соответствии с показателем DI.

### **3.2 Конвейер cancEv для автоматизированного комплексного анализа данных экспрессии генов, из базы данных TCGA**

В данном разделе описывается разработанный и реализованный программный конвейер для комплексного анализа данных экспрессии генов, полученных из проектов базы данных TCGA (The Cancer Genome Atlas). Схема вычислительного конвейера представлена в Рис. 7.



**Рисунок 7** Схема вычислительного конвейера canceEv, реализующего автоматизированный анализ данных экспрессии генов из базы данных TCGA, с последующим филогенетическим и филотранскриптомным анализом. Входные и выходные данные обозначены параллелограммами, внешние базы данных – цилиндрами, отдельные шаги алгоритма – прямоугольниками. Скрипты, позволяющие связать разные этапы работы, написаны на языке программирования Python.

В конвейере реализовано несколько этапов:

Первый этап – выгрузка и фильтрация данных об экспрессии генов из базы данных TCGA. Запросы к базе данных осуществляются при помощи пакета R `recount3` по названию проекта TCGA, посвященному интересующей ткани [Wilks и др., 2021]. На этом этапе данные представлены в формате RSE (RangedSummarizedExperiment), который обеспечивает структуру для хранения данных экспрессии генов и клинических характеристик образцов.

На следующем этапе проводится фильтрация некодирующих участков генома из исследуемых данных с использованием списка белок-кодирующих генов из NCBI Gene Database [Piovesan и др., 2019]. Далее все наборы данных проходили фильтрацию для исключения генов с низким уровнем экспрессии (менее 10 прочтений на ген). В среднем, таким образом отфильтровывалось около 20% генов – например, для анализа образцов кишечника, 15532 из 19116 генов были сохранены для анализа дифференциальной экспрессии после фильтрации.

Далее происходит фильтрация образцов в данных, для выделения образцов, относящихся к образцам типа Primary Tumor и Solid Normal Tissue. Также удаляются все образцы без диагностированной клинической стадии в мета-данных RSE. Это позволяет сфокусироваться на сравнении опухолевых и нормальных тканей. На этом же этапе выполняются скрипты проверки целостности данных после фильтрации. На выходе этапа данные сохраняются в формате RSE.

Второй этап – анализ дифференциальной экспрессии. Обработанные данные по экспрессии передаются в программный пакет DESeq2 вместе с дизайном эксперимента (Solid Normal Tissue против образцов опухолевой ткани определённой клинической стадии) для проведения анализа дифференциальной экспрессии генов. В процессе этапа данные преобразуются в формат DESeqDataSet, а на выходе формируются 4 файла со списками дифференциально экспрессирующихся генов в формате .csv, 4 файла с матрицами уровней экспрессии генов в формате .csv и визуализация результатов анализа в виде volcano plot в формате .png.

Третий этап – расчёт эволюционных индексов веб-сервисом Orthoweb 2.

Для корректной работы Orthoweb 2 идентификаторы генов конвертируются из идентификаторов Ensembl в KEGG Gene ID. Пороговые значения задаются в интерфейсе веб-сервиса вручную. На выходе формируются таблицы со значениями эволюционных индексов PAI и DI в формате .txt.

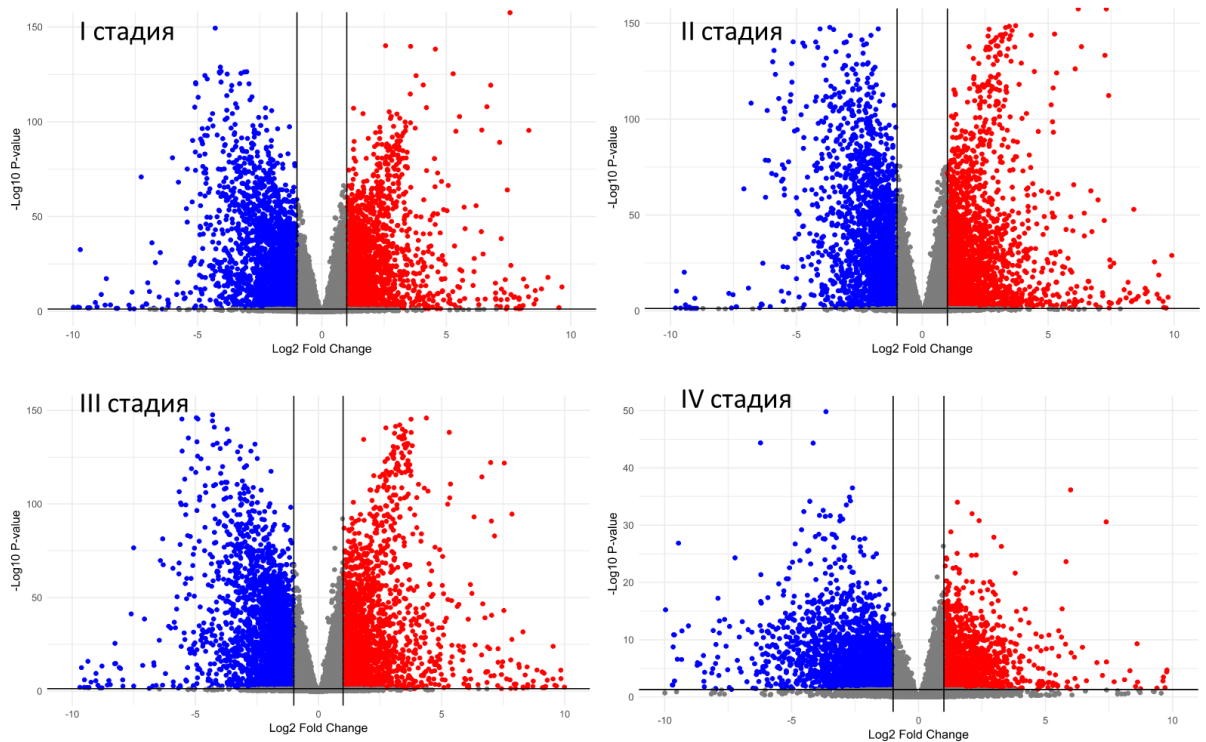
Четвертый этап – анализ филотранскриптомных индексов и редукционные тесты паттерна песочных часов. На вход пакета `myTAI` подаются таблицы со значениями эволюционных индексов PAI и DI из вывода Orthoweb 2 в формате .txt и матрицы уровней экспрессии в формате .csv со средними значениями уровней экспрессии в образцах каждой из четырех клинических стадий карцином. На выходе выдаются таблицы со значениями индексов TAI и TDI для каждого типа злокачественных заболеваний, а также изображения с визуализацией паттернов этих индексов в формате .png и результатами редукционного теста песочных часов.

### **3.3 Анализ дифференциальной экспрессии генов в злокачественных опухолях разных клинических стадий**

Представлены результаты исследования дифференциальной экспрессии генов, ассоциированных с онкологическими заболеваниями, на различных стадиях развития опухоли. Исследование базировалось на анализе данных о дифференциальной экспрессии генов, полученных с использованием высокопроизводительного секвенирования РНК (RNA-seq) из образцов тканей, поражённых опухолью, на разных стадиях её прогрессирования (подробнее см. «Материалы и методы»). Данные по экспрессии генов были получены из базы данных онкологических заболеваний TCGA. Образцы тканей из TCGA классифицируются по типам опухолей, а также по типам образцов (например, первичная опухоль, нормальная ткань). Образцы нормальных тканей (называемые NAT – нормальные, прилегающие к опухоли) берут из гистологически нормальных тканей, прилегающих к краям опухоли, и часто используют в качестве контрольных образцов для сравнения с образцами опухоли.

Полученные в ходе этого этапа списки дифференциально экспрессирующихся генов и матрицы их экспрессии для каждого типа злокачественной опухоли сохранялись для последующего анализа.

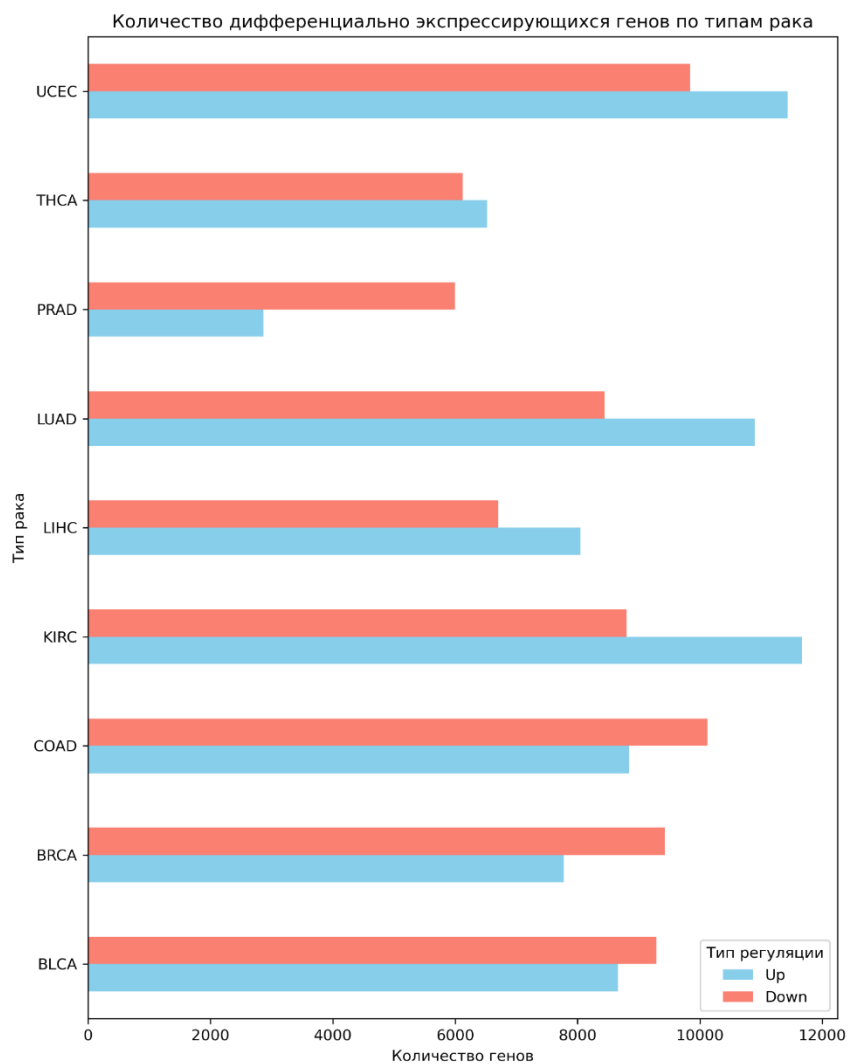
Дифференциально экспрессирующиеся гены (ДЭГ) для каждой стадии были визуализированы в виде Volcano-диаграммы (Рис. 8).



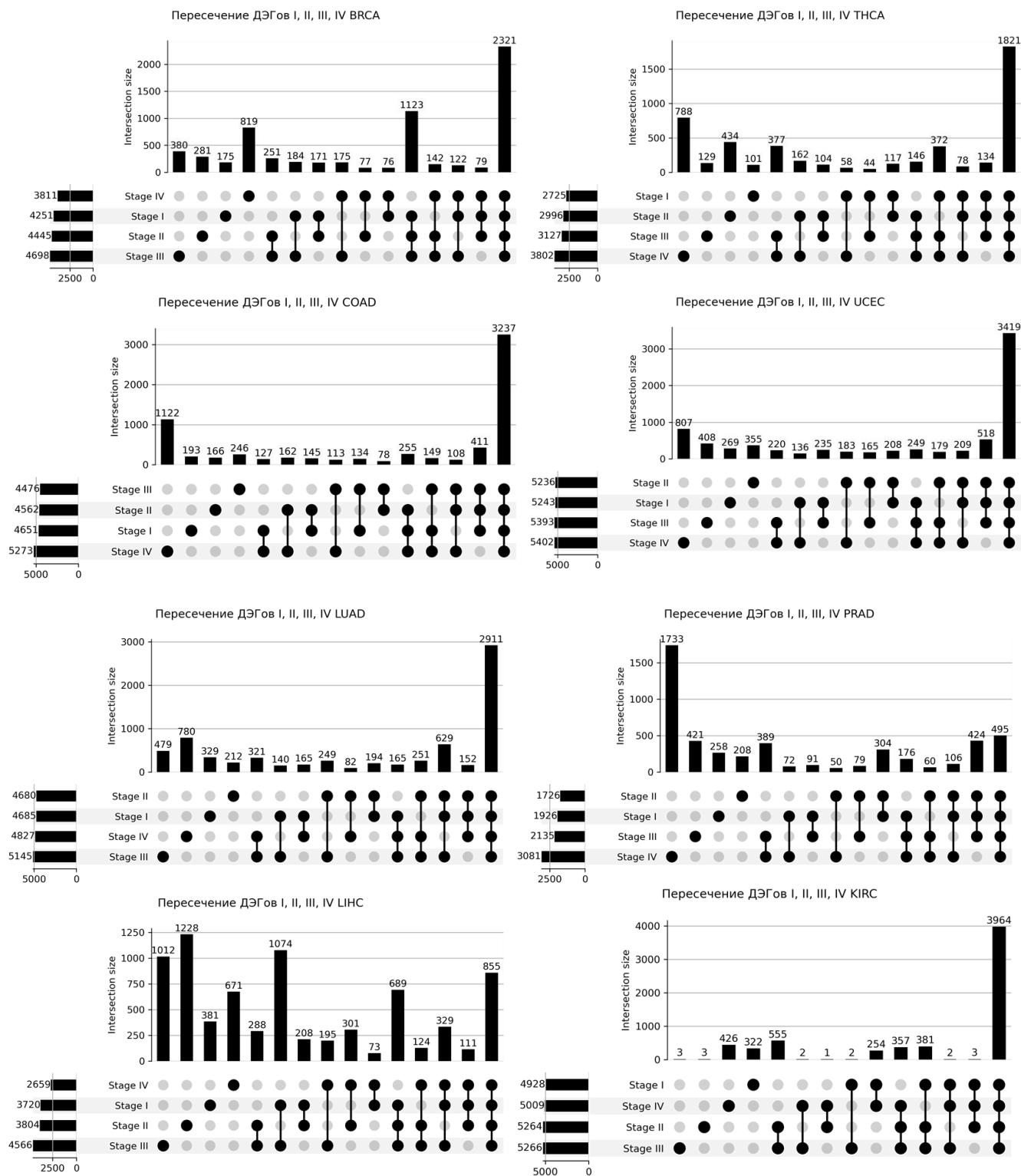
**Рисунок 8** Volcano-диаграмма, отражающая дифференциально экспрессирующиеся гены при сравнении образцов 4 стадий карциномы молочной железы с образцами NAT; Пороги значимости:  $|\log FC| > 1$ ,  $FDR < 0.05$ .

В ходе исследования были сформированы списки генов, демонстрирующих дифференциальную экспрессию на различных клинических стадиях каждого типа злокачественных новообразований (Приложение Б). Для каждого типа злокачественных заболеваний было подсчитано суммарное количество дифференциально экспрессирующихся генов с повышенной и пониженной экспрессией (объединённых по всем стадиям) (Рис. 9).

Далее были идентифицированы пересечения дифференциально экспрессирующихся генов между различными стадиями опухолей (Рис. 10). Этот этап позволил выделить гены, дифференциально экспрессирующиеся на определённой стадии развития опухоли, а также определить набор общих генов.



**Рисунок 9** Количество дифференциально экспрессирующихся генов по типам опухолей с повышенной и пониженной экспрессией. Списки ДЭГ с повышенной или пониженной экспрессией на разных стадиях одного онкологического заболевания были объединены в один список; Пороги значимости:  $|\log FC| > 1$ ,  $FDR < 0.05$ .

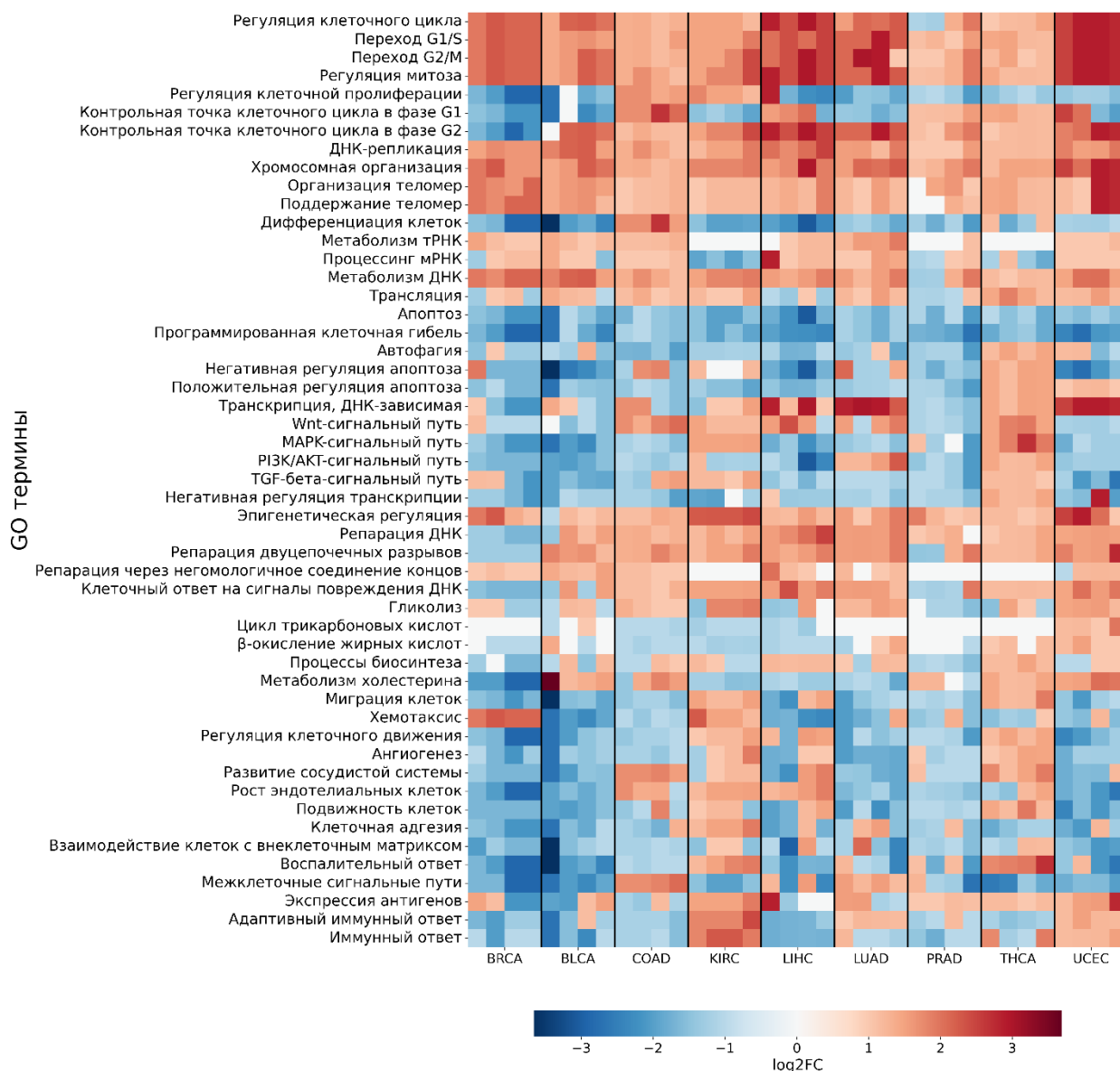


**Рисунок 10** Пересечения списков дифференциально экспрессирующихся генов между клиническими стадиями злокачественных заболеваний; Пороги значимости:  $|\log FC| > 1$ ,  $FDR < 0.05$ .

### 3.4 Анализ обогащения терминами генной онтологии списков дифференциально экспрессирующихся генов

С помощью инструмента FoldGO проведён анализ с целью выявления специализированных функциональных сдвигов в экспрессии генов, происходящих в процессе опухолевой прогрессии [Wiebe et al., 2020]. Данный метод позволяет определить термины генной онтологии (GO), по которым были обогащены наборы генов, демонстрирующие идентичные сдвиги экспрессии (fold change, FC) на разных стадиях развития опухоли. Из всех полученных результатов выделяются категории, отражающие ключевые признаки злокачественности (Приложение, Табл. B1)

Анализ генной онтологии при помощи инструмента FoldGO (Рис. 11) показал, что, вне зависимости от стадии прогрессии или органа происхождения опухоли, наблюдается повсеместное увеличение экспрессии генов, ассоциированных с регуляцией клеточного цикла (в том числе регуляцией переходов G1/S и G2/M и митозом), что согласуется с известным фактом ускоренного деления злокачественных клеток [Evan, Vousden, 2001; Hanahan, Weinberg, 2011; Malumbres, 2014; Otto, Sicinski, 2017; Vogelstein, Kinzler, 2004]. Аналогично была выявлена повышенная экспрессия генов, участвующих в путях, ответственных за поддержание целостности ДНК, в частности систем репарации двуцепочечных разрывов ДНК негомологичным соединением концов, которые характеризуются повышенной вероятностью совершения ошибок и накопления мутаций, например: *PRKDC* [Chen et al., 2020, 2021; Tan et al., 2020], *MAD2L2* [Li et al., 2018; Xu et al., 2024] и *POLQ* [Schrempf, Slysikova, Loizou, 2021; Wood, Doublié, 2016]. Также в большинстве случаев наблюдается повышение экспрессии генов, вовлечённых в процессы трансляции, транскрипции и эпигенетической регуляции экспрессии. Такие изменения часто сопровождают повышенную метаболическую активность опухолей и отражают потребность в активном синтезе белков и РНК, необходимых для быстрого роста клеток [Baylin, Jones, 2011; Hanahan, Weinberg, 2011; Morgan, Shilatifard, 2015; Truitt, Ruggero, 2016].



**Рисунок 11** Изменение уровней экспрессии генов, обогащенных терминами ГО биологических процессов, ассоциированных с ключевыми признаками злокачественности (hallmarks of cancer).

В то же время отмечается снижение уровней экспрессии генов, связанных с терминами ГО апоптоза, а также других форм запрограммированной клеточной гибели во всех образцах. Подавление апоптоза представляет собой один из классических механизмов уклонения злокачественной клетки от программируемой гибели, позволяя мутировавшим клеткам продолжать деление [Carneiro, El-Deiry, 2020; Evan, Vousden, 2001; Wong, 2011]. Снижение экспрессии также наблюдалось у генов, ассоциированных с межклеточной адгезией и подвижностью клеток [Janiszewska, Primi, Izard, 2020; Le Bras,

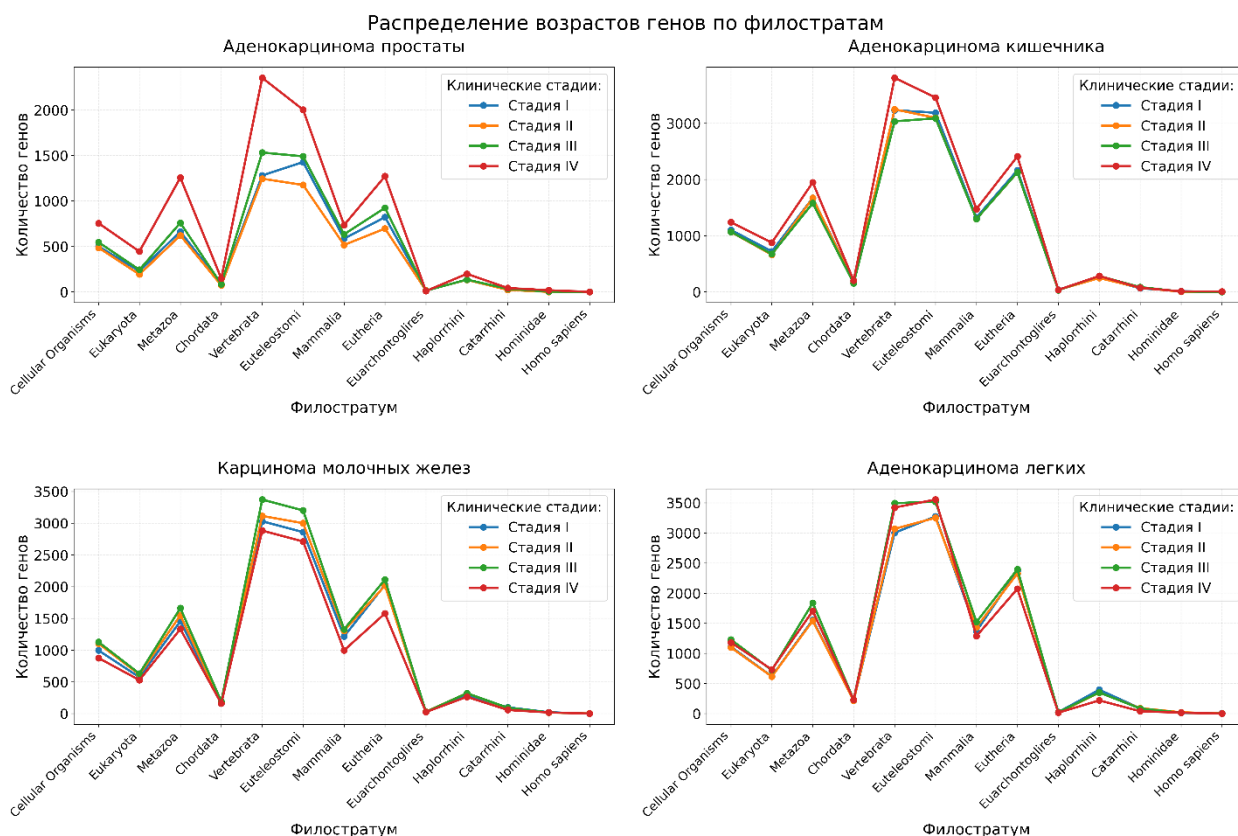
Taubenslag, Andl, 2012; Paul, Mistriotis, Konstantopoulos, 2017; Stuelten, Parent, Montell, 2018].

### **3.5 Филостратиграфический анализ дифференциально экспрессирующихся генов**

Списки дифференциально экспрессирующихся генов проанализировали с использованием веб-сервиса Orthoweb 2.

Анализ выявил пик появления генов, ассоциированных с онкологическими процессами, в ранних этапах эволюции (Рис. 12), а именно в период перехода от одноклеточных к многоклеточным организмам (Metazoa, PAI = 3). Это может быть связано с увеличением сложности и разнообразия клеточных функций в эти периоды, что могло привести к развитию более сложных механизмов регуляции роста и деления клеток. В частности, к генам этого возраста относятся гены, активно участвующие в процессах межклеточных взаимодействий, клеточного хемотаксиса, сигнального пути MAP киназы и передачи сигналов, такие как *TNFSF11*, *NOS1*, *TP53I3*, *CDX2* и *BRAF*.

Дополнительно выявлены пики более эволюционно «молодых» генов, принадлежащих к кладам позвоночных (Vertebrata, PAI = 6), костных позвоночных (Eutoleostomi, PAI = 7) и плацентарных млекопитающих (Eutheria, PAI = 9). Это указывает на потенциальное возникновение регуляторных генов, способствующих защите от злокачественных заболеваний. Среди генов позднего эволюционного происхождения выделяются те, что участвуют в регуляции клеточного цикла, роста клеток, воспалительных процессов и ангиогенеза. К генам, возникшим в филострате позвоночных, относятся гены *FZD3*, участвующий в регуляции аденокарциномы кишечника, *BMP4*, влияющий на развитие и регенерацию тканей, а также на процессы апоптоза, *CALB2*, *EGFR*, *NR2E1* и гены семейства факторов роста фибробластов *FGF*, известных за свой вклад в рост клеток, морфогенез, восстановление тканей, рост опухолей и их инвазию.



**Рисунок 12** Распределение дифференциально экспрессирующихся генов (ДЭГ) по филогруппам на примере различных стадий патологического развития опухолей аденокарциномы кишечника, лёгкого, предстательной железы и карциномы молочной железы.

К генам с индексом  $PAI = 6$  относят такие гены регуляторов процессов опухолеобразования, как *GPX3* (глутатионпероксидаза 3), обладающая антиоксидантной активностью и потенциально участвующая в предотвращении окислительного стресса, связанного с онкогенезом, *NPM1*, *KIF14* и *FAP* (фиброаденинпептидаза), участвующие в модификации микроокружения опухоли и способствующие его росту и распространению, *MYC* – ключевой онкоген, регулирующий экспрессию множества генов, связанных с клеточным циклом и метаболизмом, и *BCL2*, известный своей ролью в подавлении апоптоза, что может способствовать выживанию опухолевых клеток. К генам с индексом  $PAI = 9$ , возникшим в филогруппе плацентарных, также относятся гены *CASP5*, *CXCL2*, *PEG3*, *CAMP*,

принимающие участие в регуляции клеточного деления, апоптоза, воспалительного ответа и онкогенеза.

Мы также проанализировали распределения эволюционных возрастов генов, дифференциально экспрессирующихся в различных типах опухолей по сравнению с распределениями возрастов всех белок-кодирующих генов (Рис. 13). Можно отметить, что доля эволюционно «молодых генов» ( $PAI = 4-9$ ) выше по сравнению с полным списком генов.

Гены из всех списков ДЭГ, а также полный набор белок-кодирующих генов, были разбиты на две группы: 1) группа эволюционно «древних» генов с  $PAI \leq 3$ , что соответствует филогенетике перехода к многоклеточным; 2) группа эволюционно «молодых» генов с  $PAI > 3$ . Для анализа были составлены таблицы сопряженности, которые отражают распределение эволюционно «молодых» и «древних» генов в различных стадиях развития опухолей. В качестве статистического метода был применён точный тест Фишера, что позволило оценить наличие статистически значимых различий в распределении этих двух категорий генов между выборками ДЭГ и общим списком белок-кодирующих генов человека (Табл. 5).

**Таблица 5.** Результаты точного теста Фишера по сравнению распределения по группам эволюционно «древних» и эволюционно «молодых» генов среди всех белок-кодирующих генов человека, и среди списков дифференциально экспрессирующихся генов каждой клинической стадии опухоли. В колонках указаны: Проект TCGA, клиническая стадия, p-value теста, количество эволюционно «древних» генов, количество эволюционно «молодых» генов, среднее значение PAI в выборке

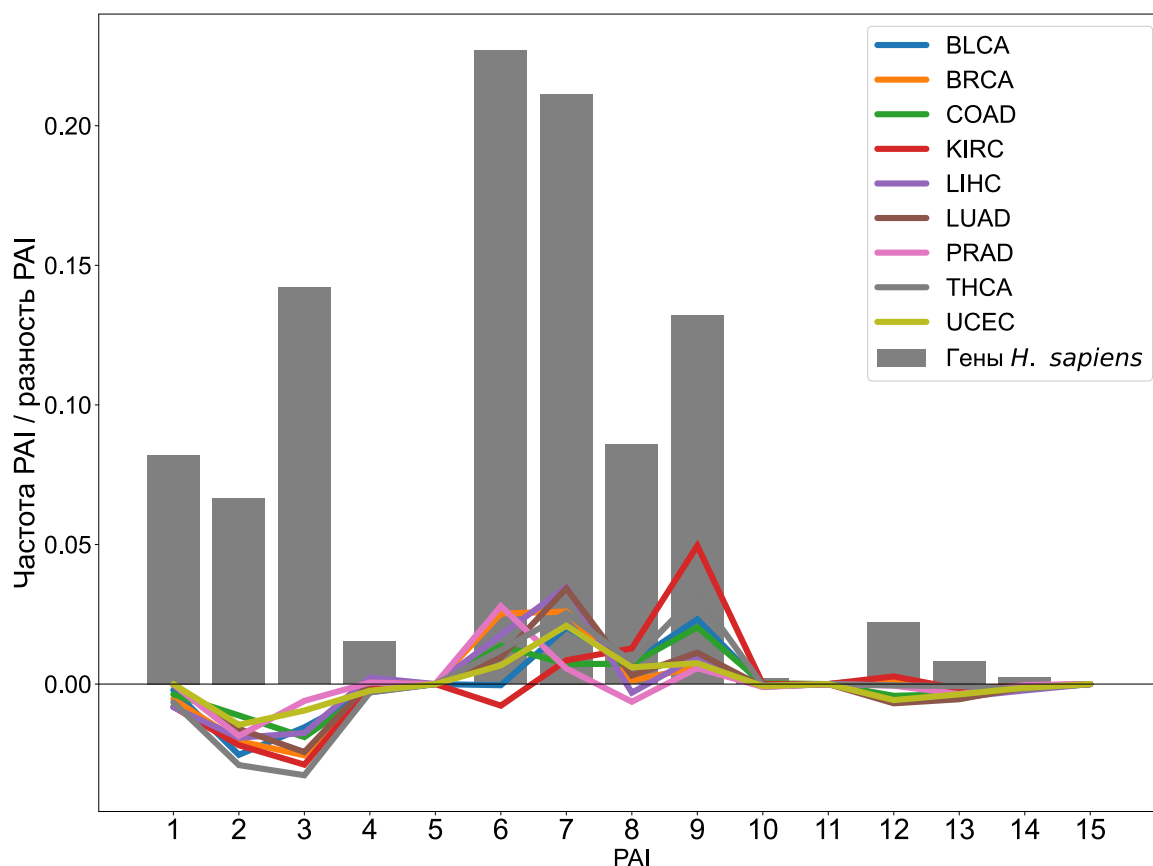
| Проект TCGA | Стадия | P-value                | Эв. «древних» | Эв. «молодых» | PAI  |
|-------------|--------|------------------------|---------------|---------------|------|
| BLCA        | I      | $6.82 \times 10^{-5}$  | 1803          | 4809          | 5.88 |
| BLCA        | II     | $6.54 \times 10^{-10}$ | 826           | 2517          | 6.08 |
| BLCA        | III    | $3.09 \times 10^{-11}$ | 987           | 2985          | 6.11 |

|      |     |                        |      |      |      |
|------|-----|------------------------|------|------|------|
| BLCA | IV  | $9.88 \times 10^{-12}$ | 996  | 3024 | 6.12 |
| BRCA | I   | $1.38 \times 10^{-18}$ | 1008 | 3241 | 6.16 |
| BRCA | II  | $1.39 \times 10^{-14}$ | 1089 | 3353 | 6.11 |
| BRCA | III | $4.45 \times 10^{-17}$ | 1140 | 3555 | 6.11 |
| BRCA | IV  | $3.15 \times 10^{-15}$ | 912  | 2897 | 6.07 |
| COAD | I   | $2.54 \times 10^{-12}$ | 1164 | 3484 | 6.05 |
| COAD | II  | $3.10 \times 10^{-13}$ | 1132 | 3427 | 6.06 |
| COAD | III | $1.85 \times 10^{-13}$ | 1106 | 3366 | 6.10 |
| COAD | IV  | $1.79 \times 10^{-10}$ | 1354 | 3917 | 6.01 |
| KIRC | I   | $3.03 \times 10^{-37}$ | 1087 | 3839 | 6.31 |
| KIRC | II  | $1.62 \times 10^{-29}$ | 1217 | 4045 | 6.25 |
| KIRC | III | $1.33 \times 10^{-29}$ | 1217 | 4047 | 6.26 |
| KIRC | IV  | $3.66 \times 10^{-27}$ | 1161 | 3846 | 6.26 |
| LIHC | I   | $2.83 \times 10^{-15}$ | 887  | 2831 | 6.18 |
| LIHC | II  | $5.64 \times 10^{-20}$ | 879  | 2923 | 6.16 |
| LIHC | III | $2.09 \times 10^{-10}$ | 1158 | 3407 | 6.07 |
| LIHC | IV  | $2.60 \times 10^{-08}$ | 653  | 2004 | 6.00 |
| LUAD | I   | $2.15 \times 10^{-23}$ | 1095 | 3587 | 6.24 |
| LUAD | II  | $6.82 \times 10^{-25}$ | 1085 | 3592 | 6.22 |
| LUAD | III | $5.13 \times 10^{-17}$ | 1263 | 3879 | 6.10 |
| LUAD | IV  | $4.39 \times 10^{-13}$ | 1206 | 3618 | 5.98 |
| PRAD | I   | $3.25 \times 10^{-07}$ | 464  | 1460 | 6.10 |
| PRAD | II  | 0.00014                | 433  | 1291 | 6.02 |
| PRAD | III | $1.09 \times 10^{-07}$ | 516  | 1616 | 6.08 |
| PRAD | IV  | 0.000936               | 819  | 2260 | 5.93 |
| THCA | I   | $5.25 \times 10^{-23}$ | 580  | 2144 | 6.27 |
| THCA | II  | $1.69 \times 10^{-31}$ | 611  | 2383 | 6.37 |
| THCA | III | $7.55 \times 10^{-29}$ | 656  | 2470 | 6.31 |
| THCA | IV  | $7.53 \times 10^{-26}$ | 846  | 2955 | 6.23 |

|                     |     |                        |      |      |      |
|---------------------|-----|------------------------|------|------|------|
| UCEC                | I   | $4.54 \times 10^{-11}$ | 1340 | 3901 | 6.03 |
| UCEC                | II  | $2.92 \times 10^{-10}$ | 1346 | 3889 | 6.08 |
| UCEC                | III | $5.95 \times 10^{-10}$ | 1393 | 3998 | 6.02 |
| UCEC                | IV  | $4.57 \times 10^{-06}$ | 1440 | 3959 | 5.93 |
| <i>Общий список</i> | -   |                        | 5664 | 9848 | 5.68 |

Во всех рассмотренных типах опухолей и на всех клинических стадиях наблюдалось статистически значимое ( $p$ -value Фишера  $< 0.001$ ) увеличение доли эволюционно «молодых» генов ( $PAI > 3$ ) среди дифференциально экспрессирующихся генов (ДЭГ) по сравнению с их представленностью в общем списке белок-кодирующих генов человека.

Для оценки статистической значимости различий между распределениями индекса  $PAI$  в выборках ДЭГ, ассоциированных с различными типами злокачественных опухолей, и общим распределением возраста всех генов человека был применен непараметрический критерий Колмогорова-Смирнова, который позволяет оценить степень различия между двумя эмпирическими распределениями без предположения о конкретной форме этих распределений (Табл. 6).



**Рисунок 13** Серыми столбцами показано распределение частот встречаемости филостратиграфических индексов всех белок-кодирующих генов человека. Цветными линиями – отклонение частот филострат в списке дифференциально экспрессирующихся генов соответствующего типа злокачественного новообразования от распределения среди всех белок-кодирующих генов человека.

Для дополнительной проверки, чтобы определить отличие распределения PAI в выборке от ожидаемого распределения, основанного на общем наборе генов человека, для каждого распределения PAI списков ДЭГ было рассчитано значение хи-квадрат. Затем, чтобы оценить достоверность этих различий, была выполнена процедура бутстрепа, в ходе которой было сгенерировано 100,000 случайных выборок из общего списка генов. Каждая случайная выборка имела такой же размер, как и оригинальная выборка ДЭГ опухолей.

Для каждой случайной выборки также было рассчитано значение хи-квадрат (Табл. 6).

**Таблица 6.** Сравнение распределений PAI списков в строках указаны: Тип опухоли и клиническая стадия, значения критерия хи-квадрат, p-value теста Колмогорова-Смирнова, p-value критерия хи-квадрата выборки ДЭГ этой стадии.

| Тип опухоли                  | BLCA_I                | BLCA_II               | BLCA_III              | BLCA_IV                | BRCA_I                 | BRCA_II                | BRCA_III               | BRCA_IV                |
|------------------------------|-----------------------|-----------------------|-----------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| Критерий $\chi^2$            | 67.27162              | 64.44491              | 60.25682              | 82.25087               | 89.8055                | 80.00764               | 83.28187               | 74.6598                |
| р-значение стат. К-С         | $2.96 \times 10^{-6}$ | $8.56 \times 10^{-6}$ | $1.19 \times 10^{-6}$ | 0.031405               | $1.49 \times 10^{-8}$  | $8.65 \times 10^{-8}$  | $3.74 \times 10^{-08}$ | $4.31 \times 10^{-10}$ |
| р-значение критерия $\chi^2$ | $1.03 \times 10^{-9}$ | $3.45 \times 10^{-9}$ | $2.03 \times 10^{-8}$ | $1.53 \times 10^{-12}$ | $5.38 \times 10^{-14}$ | $4.11 \times 10^{-12}$ | $9.73 \times 10^{-13}$ | $4.26 \times 10^{-11}$ |

| Тип опухоли                  | COAD_I                 | COAD_II                | COAD_III               | COAD_IV                | KIRC_I                 | KIRC_II                | KIRC_III               | KIRC_IV                |
|------------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| Критерий $\chi^2$            | 81.16956               | 88.31734               | 87.64031               | 70.99851               | 186.4799               | 138.2144               | 138.3501               | 157.3173               |
| р-значение стат. К-С         | $7.95 \times 10^{-8}$  | $3.17 \times 10^{-7}$  | $3.02 \times 10^{-05}$ | $1.82 \times 10^{-6}$  | $2.01 \times 10^{-15}$ | $2.45 \times 10^{-15}$ | $1.88 \times 10^{-16}$ | $1.31 \times 10^{-20}$ |
| р-значение критерия $\chi^2$ | $2.47 \times 10^{-12}$ | $1.04 \times 10^{-13}$ | $1.41 \times 10^{-13}$ | $2.08 \times 10^{-10}$ | $1.99 \times 10^{-33}$ | $1.37 \times 10^{-23}$ | $1.29 \times 10^{-23}$ | $1.85 \times 10^{-27}$ |

| Тип опухоли                  | LIHC_I                 | LIHC_II                | LIHC_III               | LIHC_IV                | LUAD_I                 | LUAD_II                | LUAD_III               | LUAD_IV                |
|------------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| Критерий $\chi^2$            | 96.58888               | 97.99785               | 74.84199               | 63.81845               | 135.5847               | 130.5432               | 95.15559               | 104.3661               |
| р-значение стат. К-С         | $5.45 \times 10^{-6}$  | $1.22 \times 10^{-10}$ | 0.000149               | $7.10 \times 10^{-10}$ | $6.55 \times 10^{-8}$  | $3.57 \times 10^{-15}$ | $5.07 \times 10^{-6}$  | $1.43 \times 10^{-16}$ |
| р-значение критерия $\chi^2$ | $2.59 \times 10^{-15}$ | $1.37 \times 10^{-15}$ | $3.94 \times 10^{-11}$ | $4.50 \times 10^{-9}$  | $4.65 \times 10^{-23}$ | $4.80 \times 10^{-22}$ | $4.92 \times 10^{-15}$ | $7.74 \times 10^{-17}$ |

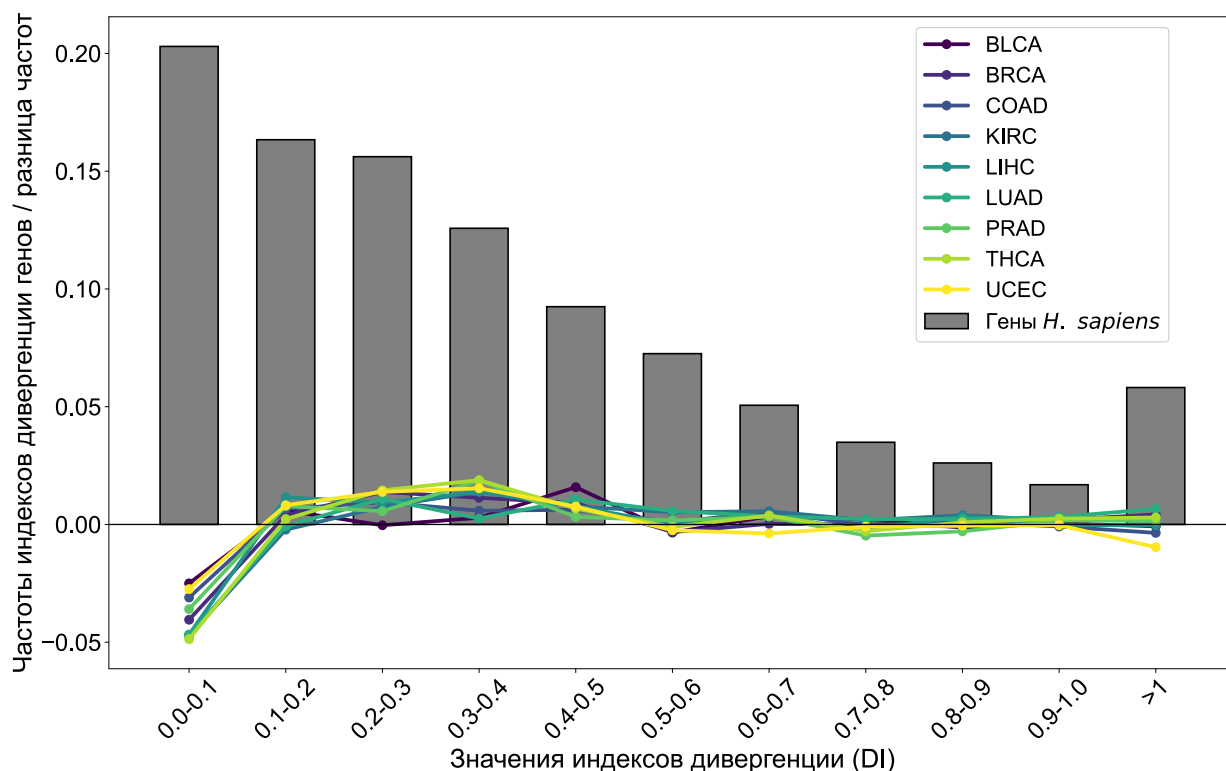
| Тип опухоли | PRAD_I | PRAD_II | PRAD_III | PRAD_IV | THCA_I | THCA_II | THCA_III | THCA_IV |
|-------------|--------|---------|----------|---------|--------|---------|----------|---------|
|-------------|--------|---------|----------|---------|--------|---------|----------|---------|

|                              |                       |                       |                       |                        |                        |                        |                        |                        |
|------------------------------|-----------------------|-----------------------|-----------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| Критерий $\chi^2$            | 57.40895              | 50.98069              | 56.85088              | 42.37884               | 113.7603               | 160.5797               | 133.2819               | 127.2402               |
| р-значение стат. К-С         | $8.87 \times 10^{-5}$ | 0.009352              | 0.075102              | $3.98 \times 10^{-05}$ | $1.18 \times 10^{-16}$ | $2.41 \times 10^{-18}$ | $1.86 \times 10^{-14}$ | $1.67 \times 10^{-13}$ |
| р-значение критерия $\chi^2$ | $6.67 \times 10^{-8}$ | $9.39 \times 10^{-7}$ | $8.41 \times 10^{-8}$ | $2.88 \times 10^{-05}$ | $1.08 \times 10^{-18}$ | $4.00 \times 10^{-28}$ | $1.35 \times 10^{-22}$ | $2.21 \times 10^{-21}$ |

Основываясь на полученных результатах, была оценена доля случайных выборок, в которых значение хи-квадрат превышало значение, полученное для исходной выборки ДЭГ. На основании этого определялось р-значение для каждой выборки, что позволило оценить статистическую значимость наблюдаемых различий. Результаты анализа показали, что для большинства рассмотренных типов злокачественных новообразований значения хи-квадрат в исходных выборках ДЭГ значительно превышали соответствующие значения в подавляющем большинстве случайных выборок, что свидетельствует о высокой статистической значимости различий в распределении PAI между выборками ДЭГ и общим списком генов.

Аналогичный анализ был также проведён и для распределений индексов дивергенции. По распределению значений индекса дивергенции большая часть генов идентифицировалась как эволюционно консервативная (Рис. 14).

Таким образом, полученные результаты указывают на значительную роль эволюционно «молодых» генов в развитии злокачественных заболеваний через различные механизмы, включая регуляцию клеточного цикла (*MYC*, *CDKN*), апоптоза (*CALB2*, *EGFR*, *NR2E1*), ангиогенеза (*VEGFA*, *TGFβ*) и воспалительных процессов (*NFKB2*, *RELA*, *CXCL8*, *CTLA4*). Большинство дифференциально экспрессирующихся генов характеризовалось высоким уровнем консервативности, что также означает, что их последовательности и функции остаются относительно неизменными на протяжении эволюции. Эта консервативность может быть результатом их основополагающей роли в жизнедеятельности клеток и организмов.



**Рисунок 14** Серыми столбцами показано распределение частот встречаемости индексов дивергенции генов среди всех белок-кодирующих генов человека, участвующих в анализе. Цветными линиями показана разница между частотой DI генов списка дифференциально-экспрессирующихся генов в соответствующем типе опухоли и частотой встречаемости аналогичного DI среди белок-кодирующих генов человека.

### 3.6 Филотранскриптомный анализ клинических стадий карцином человека

Для оценки динамики эволюционного возраста транскриптомов в процессе опухолевой прогрессии был проведён комплексный филотранскриптомный анализ. Анализ объединил данные индексов филостратиграфического возраста генов (рассчитанные с использованием Orthoweb 2, 60% идентичности) и уровней их экспрессии на клинических стадиях опухоли. В результате были получены филотранскриптомные индексы для каждой стадии.

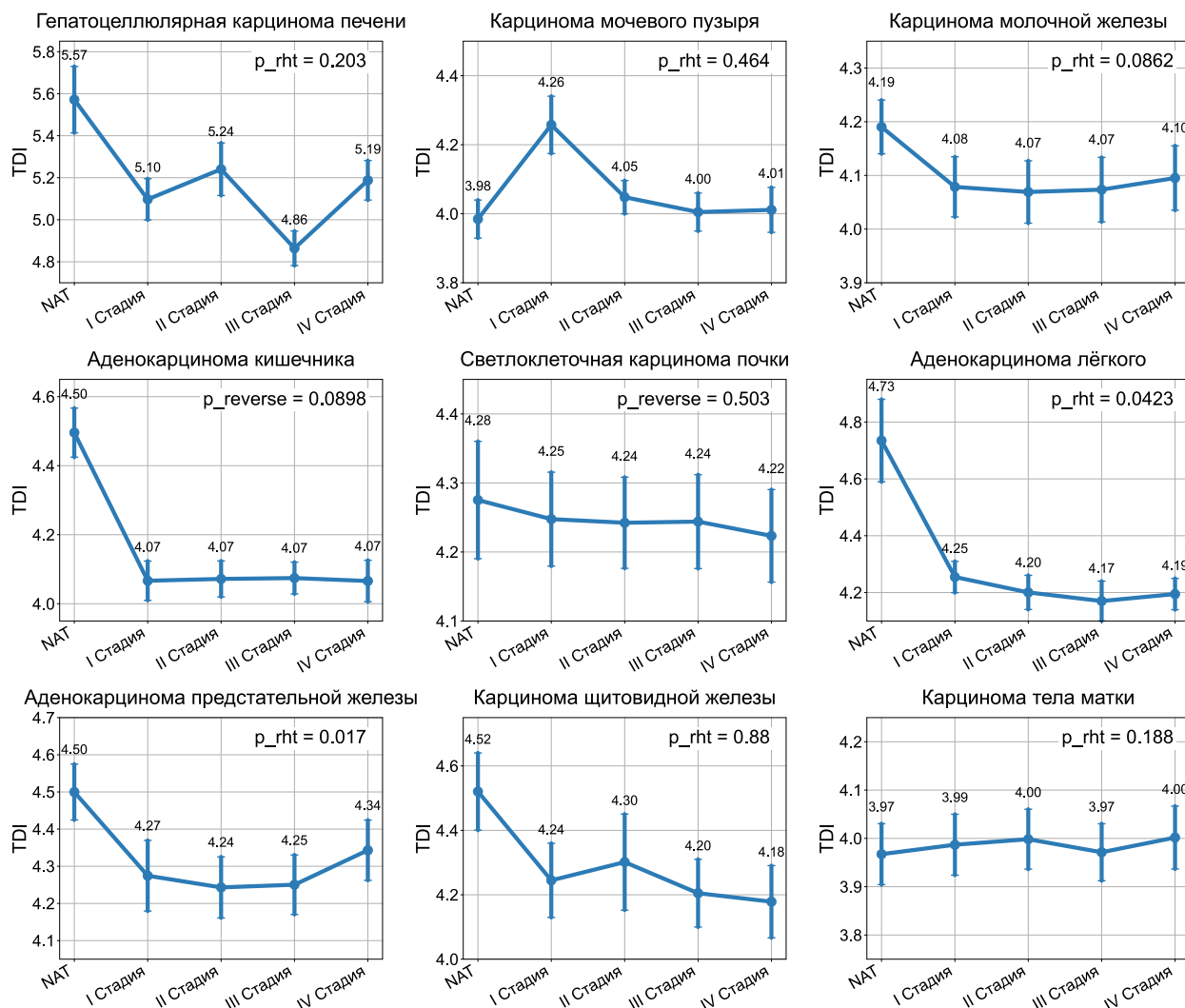
В качестве основных аналитических инструментов использовались индекс возраста транскриптома (TAI) и индекс дивергенции транскриптома (TDI). Статистическую значимость паттернов TAI и TDI оценивали с использованием тестов редукционных и обратных песочных часов [Drost и др., 2017, 2018]. Анализ охватил 5057 образцов карцином различных типов на разных клинических стадиях и 498 образцов здоровых тканей из базы данных TCGA. Ранняя стадия развития опухоли определялась как стадия I, средние стадии – II и III, поздняя стадия – IV. Абсолютные значения филотранскриптомных индексов каждой стадии представлены в Таблице B2 и B3 (см. Приложение B).

Анализ выявил последовательно более низкие значения TAI в карциномах по сравнению со здоровыми тканями. Исключение составили ткани светлоклеточной карциномы почек, где TAI контроля был ниже. Повышенный TAI в контрольных тканях может указывать на активацию путей, ассоциированных с адаптивным ответом на онкогенные сигналы и повреждения ДНК. Повышенный TAI в контроле также может объясняться активацией эволюционно «молодых» систем иммунного ответа в нормальных прилежащих к опухоли тканях (NAT) под влиянием сигналов из опухоли. Относительно низкий TAI в опухолях, таким образом, отражает активацию консервативных онкогенных путей, регулирующих базовые процессы клеточного деления и метаболизма. Примечательно, что среди самих карцином карцинома печени выделялась заметно более высокими значениями TAI по сравнению с другими типами опухолей. Эта особенность, вероятно, связана со специфической физиологической ролью печени в детоксикации и метаболизме ксенобиотиков, требующей повышенной экспрессии эволюционно «молодых» генов, ассоциированных с этими функциями (например, гены *CYP3A4* и *UGT1*). Важно отметить, что потенциальное влияние предшествующей противоопухолевой терапии на экспрессионные профили пациентов в исследуемой когорте является фактором, который также нельзя исключать при интерпретации наблюдаемых различий в TAI.

Снижение значений TAI в опухолевых образцах согласуется с атавистической гипотезой злокачественных заболеваний [Chen, He, 2016a; Davies, Lineweaver, 2011; Bussey, Davies, 2021; Chen et al., 2015; Thomas et al., 2017; Vincent, 2012]. Эта гипотеза постулирует, что злокачественная трансформация связана с реактивацией эволюционно консервативных программ, приводящих к возврату к «древним» молекулярным фенотипам [Merlo и др., 2006; Trigos и др., 2017; Lambert et al., 2011], которые приоритизируют пролиферацию и выживание, подобно одноклеточным организмам. Согласно ей, опухоли демонстрируют популяционную динамику, сходную с одноклеточными формами жизни.

Однако тщательное изучение динамики TAI и TDI по стадиям выявило более сложную картину (Рис. 15, 16, Табл. A1). Для аденокарцином лёгкого ( $p\_rht = 0.0377$ ) и предстательной железы ( $p\_rht = 0.00531$ ) был выявлен статистически значимый паттерн «песочных часов» в распределении TDI, характеризующийся снижением консервативности транскриптома (повышением TDI) на стадиях I и IV и повышением консервативности (снижением TDI) на стадиях II-III (Рис. 15, Табл. A2).

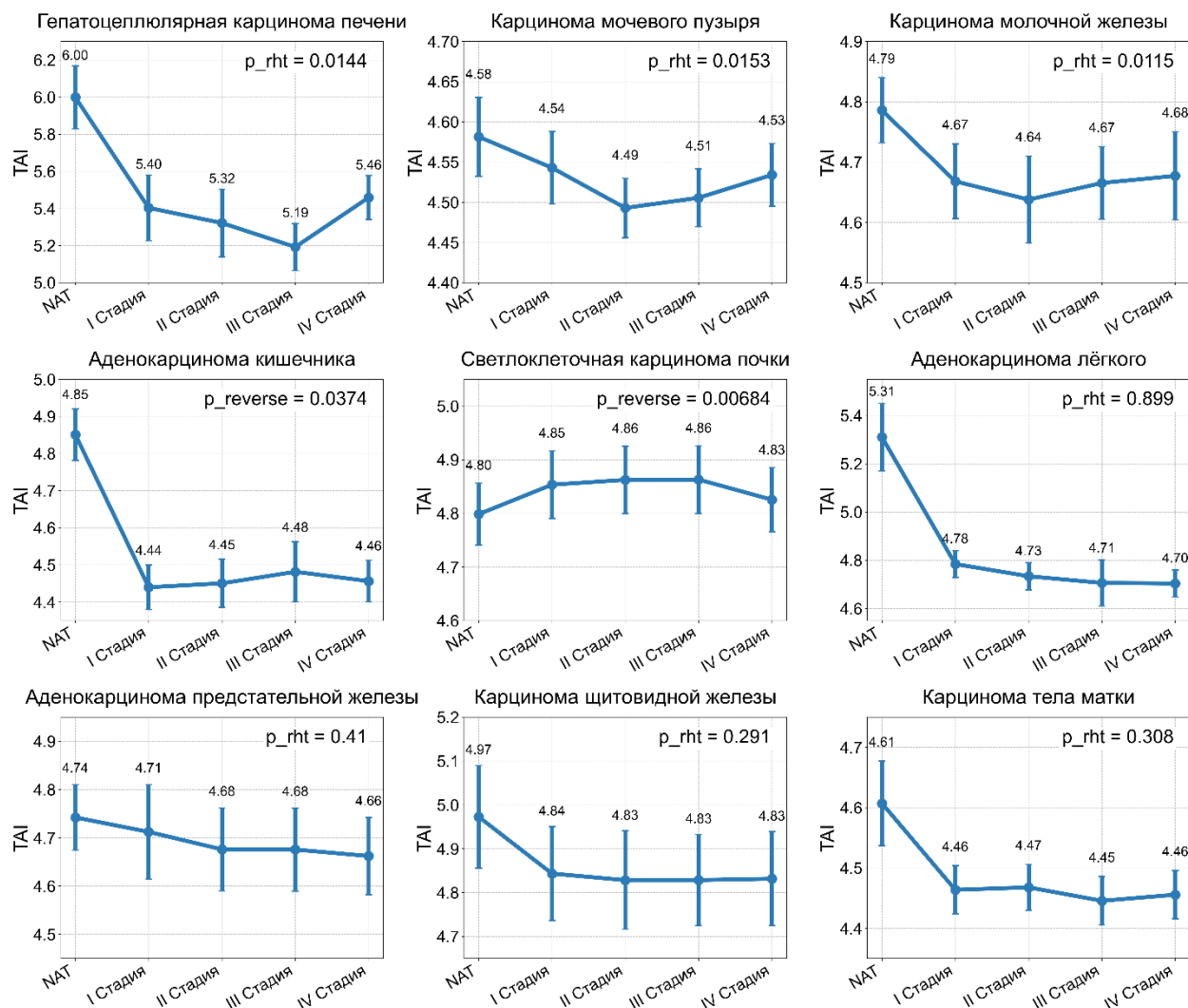
Аналогично, паттерн «песочных часов» в распределении TAI был выявлен для карциномы печени ( $p\_rht = 0.00942$ ), протоковой карциномы молочных желез ( $p\_rht = 0.0107$ ) и карциномы мочевого пузыря ( $p\_rht = 0.0337$ ) (Рис. 16, Табл. A1). Выявленные закономерности сходны с моделью «песочных часов» в эмбриональном развитии, где средние стадии характеризуются преобладанием «древних», высококонсервативных генов. Это свидетельствует о том, что опухолевые клетки временно реактивируют программы раннего эволюционного происхождения перед привлечением более «молодых» генов для развития злокачественности, что ставит под сомнение парадигму однонаправленной регрессии к одноклеточности.



**Рисунок 15** Распределение значений индексов дивергенции транскриптома (TDI). В распределениях TDI аденокарцином лёгкого и предстательной железы обнаружены статистически значимые паттерны «песочных часов».

Наблюдаемые эволюционные паттерны целесообразно рассматривать в контексте возможных процессов каждой стадии:

На начальной стадии канцерогенеза наблюдается активация эволюционно «молодых» генов. Ряд этих генов ассоциирован с механизмами регуляции роста и выживания, которые необходимы злокачественным клеткам для начального развития. В частности, зафиксировано изменение экспрессии регуляторных генов-привратников, контролирующих критические точки клеточного цикла, ответа на повреждение ДНК и межклеточной адгезии.



**Рисунок 16** Распределение значений индексов возраста транскриптома (TAI). В распределениях TAI карциномы мочевого пузыря, молочной железы и печени выявлены статистически значимые паттерны «песочных часов».

Их PAI, как правило, выше, чем у генов-смотрителей, непосредственно участвующих в процессах клеточного деления и апоптоза.

Повышение экспрессии эволюционно «молодых» генов может быть стратегией, позволяющей злокачественным клеткам выживать и размножаться в начальной фазе опухолевого процесса. Например, на первых стадиях развития опухолей, в сравнении с промежуточными стадиями наблюдается значительное повышение экспрессии следующих генов-привратников: *CDKN2A*, ген регулятора остановки клеточного цикла в G1 фазе (Eutoleostomi, PAI = 7), *CHEK2* (Vertebrata, PAI = 6), кодирующий серин-треониновую киназу, активируемую в ответ на двунитевые разрывы ДНК и

координирующую остановку цикла, репарацию и апоптоз через взаимодействие с ATM, p53 и BRCA1, и *CDH1* (Eutoleostomi, PAI = 7) кодирующий E-кадгерин – основной белок адгезивных контактов, подавляющий инвазию и метастазирование, потеря которого представляет собой раннее событие эпителиально-мезенхимального перехода (EMT).

На промежуточных клинических стадиях опухоли (II, III), характеризующихся активным ростом опухоли и усиливающейся внутриопухолевой конкуренцией, происходит выраженный сдвиг в сторону активации более эволюционно «древних» генов. Эти гены поддерживают фундаментальные клеточные процессы: энергетический метаболизм, репликацию ДНК, репарацию повреждений и базовый клеточный цикл. Активация этих путей обеспечивает опухолевым клеткам необходимую стабильность, устойчивость к стрессовым условиям микроокружения (таким как гипоксия и дефицит питательных веществ) и высокую пролиферативную активность, что критически важно для их конкуренции и клональной экспансии. Примерами генов, демонстрирующих значительное увеличение экспрессии ( $\log_2 FC > 2$ ) на этих стадиях, являются: *AMPK* (Eukaryota, PAI = 2), ключевая киназа каскадов АТФ-фосфорилирования, регулирующая катаболические процессы для восстановления энергетического гомеостаза, *MSH2* (Eukaryota, PAI = 2) – ген, ответственный за репарацию ошибочно спаренных нуклеотидов, *PKM2* – пируваткиназа, катализирующая аэробный гликолиз и *IDH2* (Cellular Organism, PAI = 1) – изоцитратдегидрогеназа, приводящая к накоплению онкометаболита 2-гидроксиглутарата, нарушающий эпигенетическую регуляцию опухолеобразования.

На поздних стадиях ряда опухолей, связанных с диссеминацией и метастазированием, наблюдается реактивация эволюционно более «молодых» генов. Эти гены кодируют белки, необходимые для преодоления тканевых барьеров, ремоделирования внеклеточного матрикса, стимуляции ангиогенеза, подавления противоопухолевого иммунитета и адаптации к метастатическим нишам – функций, эволюционно появившихся у

многоклеточных организмов и критичных для инвазивного роста. Яркими примерами являются гены, у которых отмечено значительное повышение экспрессии ( $\log_2 FC > 3.5$ ) на стадии IV в аденокарциномах лёгкого, простаты и протоковой карциноме молочной железы: *STC1* (Vertebrata, PAI = 6), кодирующий станиокальцин-1 – многофункциональный гормон, участвующий в регуляции ангиогенеза, клеточной миграции и защиты от апоптоза, уровень которого повышен в метастатических очагах; *VCAN* (Euteleostomi, PAI = 7), кодирующий версикан – крупный протеогликан внеклеточного матрикса, способствующий пролиферации, миграции, инвазии и ангиогенезу опухолевых клеток через взаимодействие с рецепторами (CD44, TLR2) и модуляцию сигнальных путей (EGFR, Wnt/ $\beta$ -catenin), что ассоциировано с плохим прогнозом; и *SPP1* (Eutheria, PAI = 9), кодирующий остеопонтин (Secreted Phosphoprotein 1) – многофункциональный цитокин/матриксный белок, который играет важную роль в инвазии, метастазировании, ангиогенезе, ускользании от иммунного ответа (подавление Т-клеток, поляризация макрофагов в M2-фенотип) и выживании в отдаленных нишах, а также является сильным прогностическим маркером агрессивности.

Примечательно, что в опухолях кишечника наблюдалась противоположная картина, в которой был выявлен обратный паттерн «песочных часов» в индексе возраста транскриптома ( $p\_reverse\_hourglass = 2.41 \cdot 10^{-6}$ ) (Рис. 16, Табл. A1). Это подчёркивает разнообразие в поведении различных типов опухолей и соответствующих им наборов генов.

Результаты филотранскриптомного анализа демонстрируют, что прогрессирование опухоли при различных типах злокачественных заболеваний нельзя рассматривать как простую регрессию к одноклеточному состоянию. Вместо этого оно отражает сложное динамическое переключение между эволюционно древними и более молодыми генетическими программами. Опухоли эксплуатируют глубоко консервативные пути для регуляции базовых метаболических и пролиферативных потребностей,

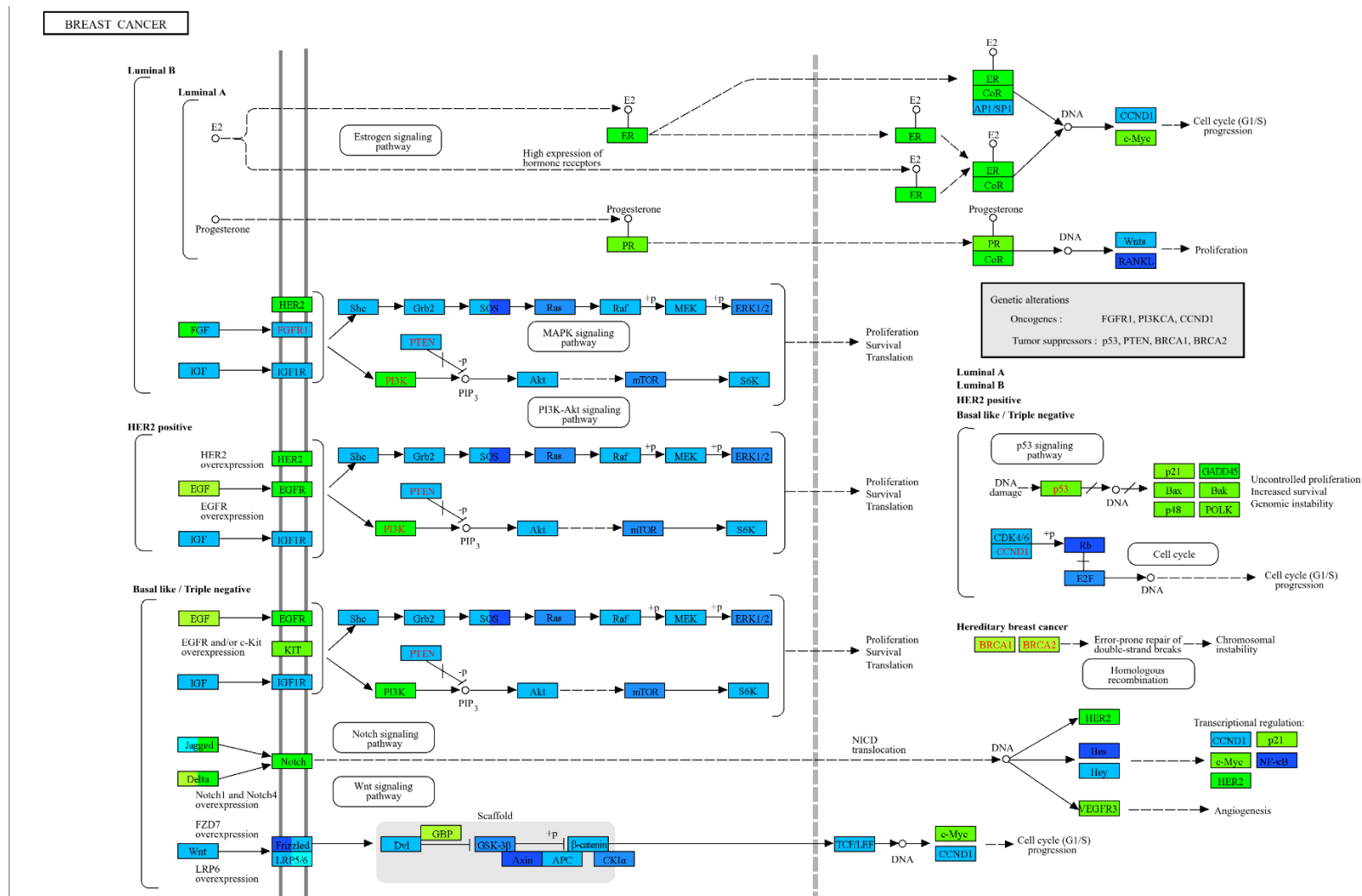
особенно на средних стадиях, что согласуется с атавистической гипотезой. Одновременно они подключают более «молодые», поздно возникшие в эволюции гены, необходимые для преодоления таких задач, как инвазия, уклонение от иммунитета и метастазирование на поздних стадиях, формируя характерный паттерн «песочных часов». Такое динамическое взаимодействие подчёркивает эволюционную пластичность злокачественных новообразований и выявляет потенциальные уязвимости, связанные как с консервативными путями, так и с недавно приобретенными функциями, что открывает новые перспективы для таргетной терапии.

### **3.7 Анализ генных сетей злокачественных заболеваний**

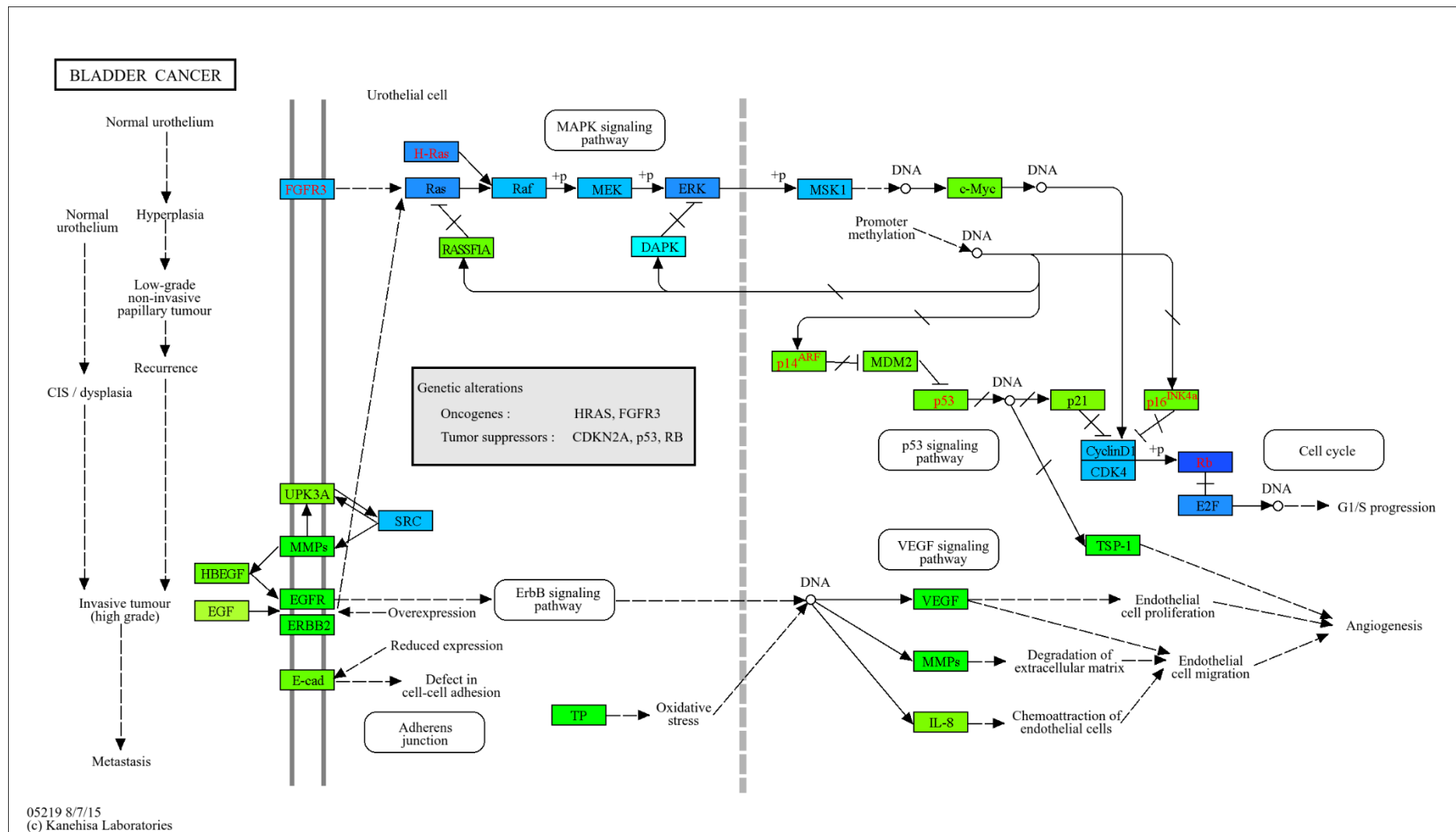
В данном разделе представлены результаты филогенетического анализа эволюционных индексов генов, входящих в состав различных генных сетей злокачественных заболеваний. Целью исследования являлось выявление закономерностей эволюции состава генных сетей опухолей, обусловленных вовлечением новых генов.

Для анализа были выбраны генные сети процессов развития злокачественных опухолей в различных тканях человека, представленных в базах данных KEGG PATHWAYS (Рис. 17, 18) и WikiPathways (Рис.19).





**Рисунок 18** Результаты филогенетического анализа генов сети карциномы молочной железы из базы данных KEGG PATHWAYS hsa05224: Breast cancer – *Homo sapiens* (human).



**Рисунок 19** Результаты филогенетического анализа генов сети карциномы стенки мочевого пузыря из базы данных WikiPathways WP2828: Bladder cancer – *Homo sapiens*.

Наряду с изучением тканеспецифичных сетей онкологических заболеваний, целью нашей работы был анализ процессов, универсальных для различных типов опухолей. Поскольку сигнальные пути являются ключевыми регуляторами клеточного цикла, апоптоза, метаболизма и других фундаментальных клеточных функций, их дисрегуляция служит частой причиной развития и прогрессирования онкологических заболеваний. В частности, это связано с активацией ангиогенеза, неконтролируемой пролиферации и подавления апоптоза. Для выявления таких универсальных механизмов мы провели анализ генной сети ключевых сигнальных путей, связанных с метаболизмом злокачественных клеток, используя базу данных KEGG Pathways (Рис. 20).

На сети можно выделить несколько ключевых сигнальных путей:

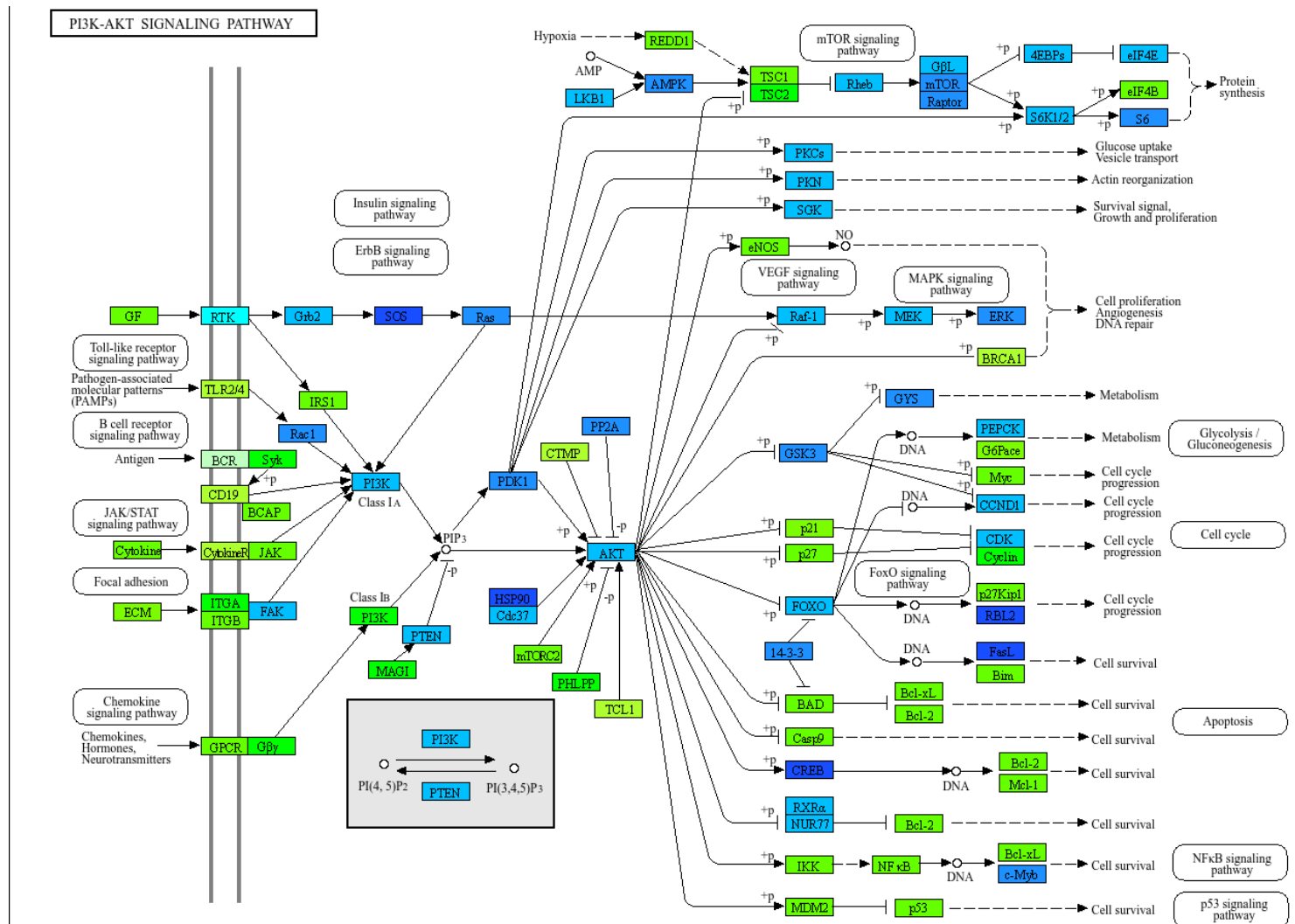
ERK-MAPK сигнальный путь является одним из основных механизмов, регулирующих клеточный рост, деление и дифференцировку клеток, а также участвует в регуляции процессов ангиогенеза [Guo и др., 2020; Zhou, Yang, Song, 2018]. Этот путь активируется рядом внешних стимулов, включая ростовые факторы, цитокины и другие сигналы окружающей среды. В его основе лежит каскад фосфорилирования: активация рецептора на поверхности клетки передает сигнал через серию киназ (Raf, MEK, ERK), которые в конечном итоге переносят сигнал в ядро для регуляции экспрессии генов. Дисрегуляция этого пути часто связана с онкогенезом. Из филостратиграфического анализа генов этого сигнального пути можно отметить, что внешние стимулы и ростовые факторы кодируются преимущественно группой эволюционно «молодых» генов со значениями эволюционного возраста  $PAI = 6, 7$  и  $8$  (Vertebrata, Euteleostomi и Mammalia соответственно), к которым относятся такие факторы и гены, кодирующие их рецепторы, такие как *TGFA*, *EGFR*, *EGF*, *PDGF*, *PDGFR*, *KITLG*, *FLT3LG* и *HGF* (Рис. 21).





Но в то же время в инициации пути также присутствуют инсулиноподобный фактор роста IGF и фактор роста фибробластов FGF с эволюционным возрастом происхождения в филострате многоклеточных организмов. Серия киназ, через которую сигнал от рецепторов переходит в ядро, состоит из белков, кодируемых эволюционно «древними» генами со значениями PAI = 1, 2, 3, продукты которых являются известными онкогенами, к которым относятся семейство малых ГТФаз RAS (HRAS, KRAS, NRAS), семейство серин-треониновых протеин-киназ RAF (BRAF, ARAF, RAF1) и семейство MAP киназ (MAP2K1/2, MAPK1, MAPK3). Дальнейшая передача сигнала в ядро приводит к активации генов, регулирующих ангиогенез, посредством семейства факторов роста эндотелия сосудов VEGF, которые обладают значением PAI = 6 (Vertebrata).

PI3K/Akt путь играет ключевую роль в регуляции процессов выживания клеток, апоптоза, роста и метаболизма [Fruman, Rommel, 2014]. Активация этого пути начинается с связывания ростовых факторов с их рецепторами, что приводит к активации ключевого фактора этого сигнального пути PI3K (фосфатидилинозитол-3-киназы). Примечательно, что в комплексе генов, кодирующих субъединицы PI3K, присутствуют как эволюционно «древние» *PIK3CA*, *PIK3R2* и *PIK3R3* (Metazoa, PAI = 3), так и эволюционно «молодые» гены *PIK3CB* и *PIK3CD* (Vertebrata, PAI = 6) (Рис. 22). Активированная PI3K катализирует превращение фосфатидилинозитола в PIP3, который служит вторичным посредником, активирующим Akt. Akt может активировать или подавлять различные молекулярные мишени, влияя на процессы выживания и деления клеток через фосфорилирование mTOR. Гены, кодирующие оба этих белка, прослеживают своё происхождение до первых Эукариот (Eukaryota, PAI=2). Нарушение регуляции этого пути часто ассоциируется со злокачественными процессами [Fruman et al., 2017; Jiang et al., 2025].

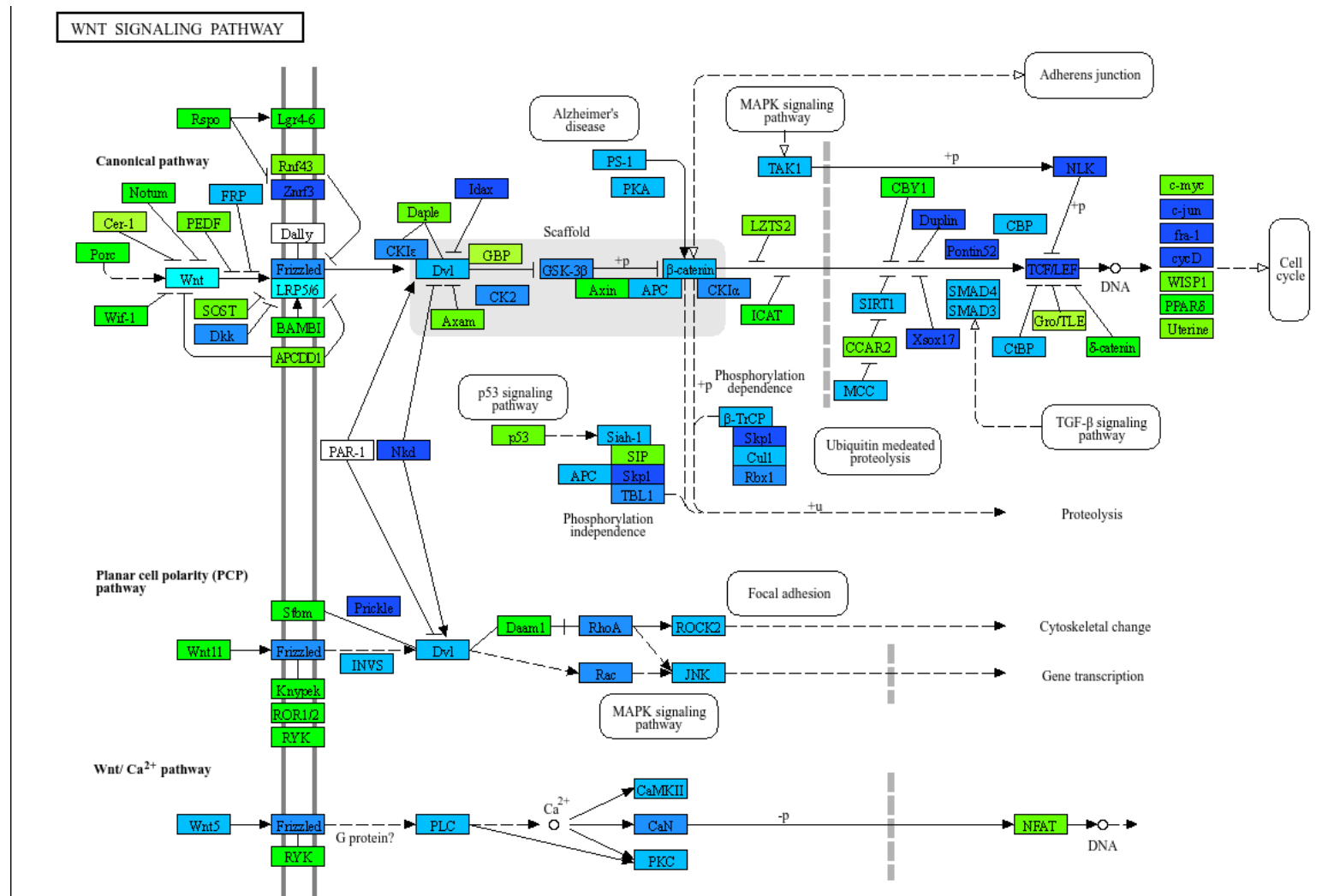


**Рисунок 22** Результаты филогенетического анализа генов сети PI3K-Akt сигнального пути из базы данных KEGG PATHWAY hsa: 04151.

WNT/ $\beta$ -catenin путь (Рис. 23) участвует в контроле клеточного цикла, адгезии, миграции и дифференцировки клеток. Активация пути начинается со связывания лигандов WNT с рецепторами Frizzled и LRP на поверхности клетки. Это приводит к стабилизации и накоплению  $\beta$ -catenin в цитоплазме и его последующему перемещению в ядро, где он взаимодействует с факторами транскрипции и стимулирует экспрессию целевых генов. Дисрегуляция этого пути связана с развитием множества типов опухолей [Zhan, Rindtorff, Boutros, 2017]. Этот сигнальный путь является одним из самых древних, и включает в себя преимущественно гены, возникшие на этапе происхождения многоклеточных организмов и Эукариот (PAI=2, 3).

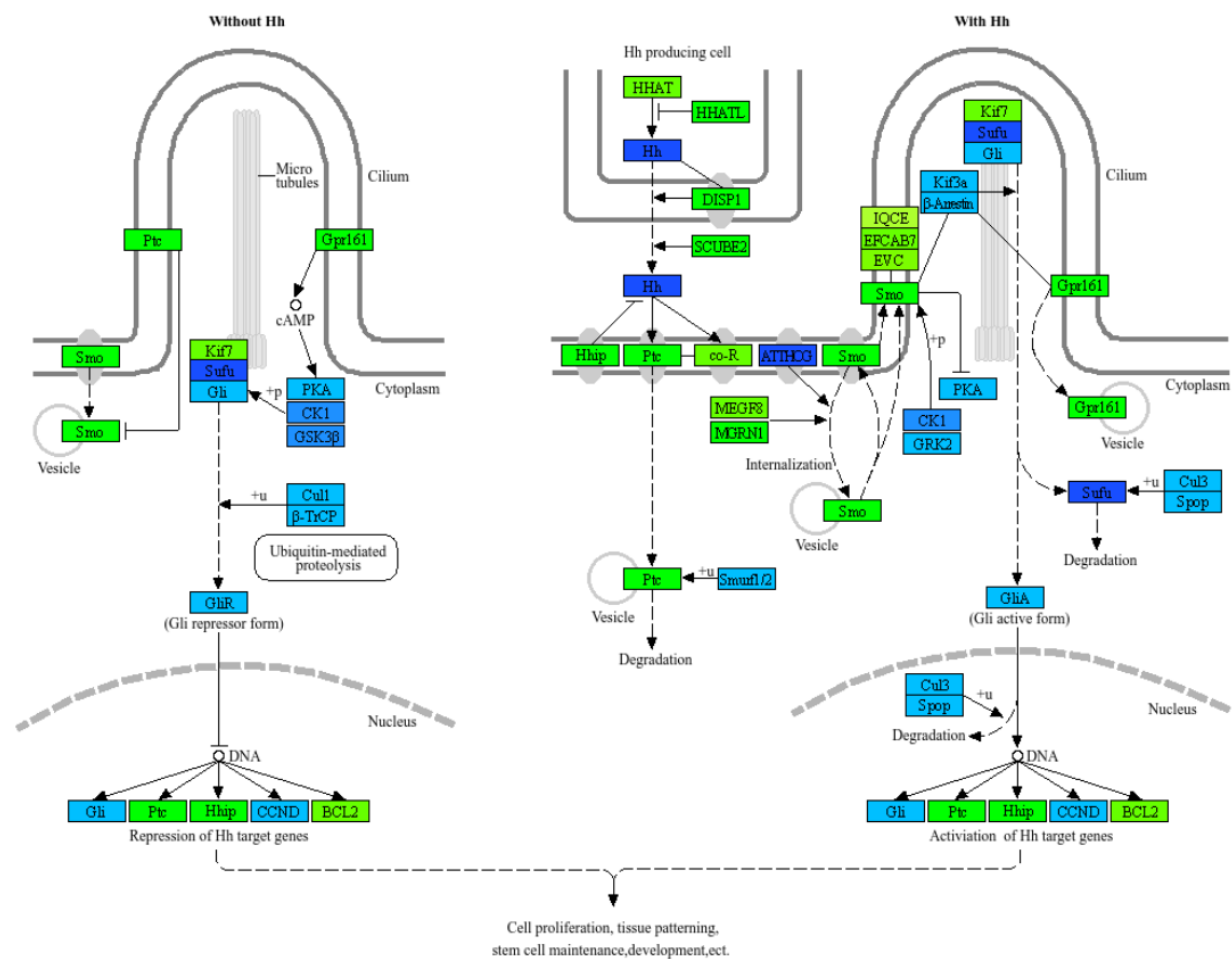
Сигнальный путь Hedgehog (Hh) (Рис. 24) участвует в эмбриональном развитии, а также участвует в поддержании тканевого гомеостаза. Нарушения регуляции компонентов этого пути ассоциированы с рядом злокачественных новообразований [Skoda и др., 2018]. Путь активируется при связывании одного из лигандов семейства Hh (Cellular Organisms, PAI = 1) с рецептором PTCH1 (Vertebrata, PAI = 6), что приводит к активации SMO (Vertebrata, PAI = 6). Активированный SMO запускает каскад внутриклеточных взаимодействий, в том числе ингибирует супрессор SUFU (Cellular Organisms, PAI = 1), который в норме удерживает транскрипционные факторы GLI (Metazoa, PAI = 3) в цитоплазме в неактивной форме.

Дополнительная регуляция пути осуществляется через белки CUL3 (Metazoa, PAI = 3) и SPOP, участвующие в убиквитин-зависимом протеолизе и деградации активных форм GLI, а также через циклин D (CCND, Metazoa, PAI = 3), являющийся мишенью GLI-зависимой транскрипции и обеспечивающий прогрессирование клеточного цикла. Белок GRK2 регулирует фосфорилирование SMO, модулируя его активность и внутреннюю локализацию.



**Рисунок 23** Результаты филогенетического анализа генов сети Wnt сигнального пути из базы данных KEGG PATHWAY hsa: 04310.

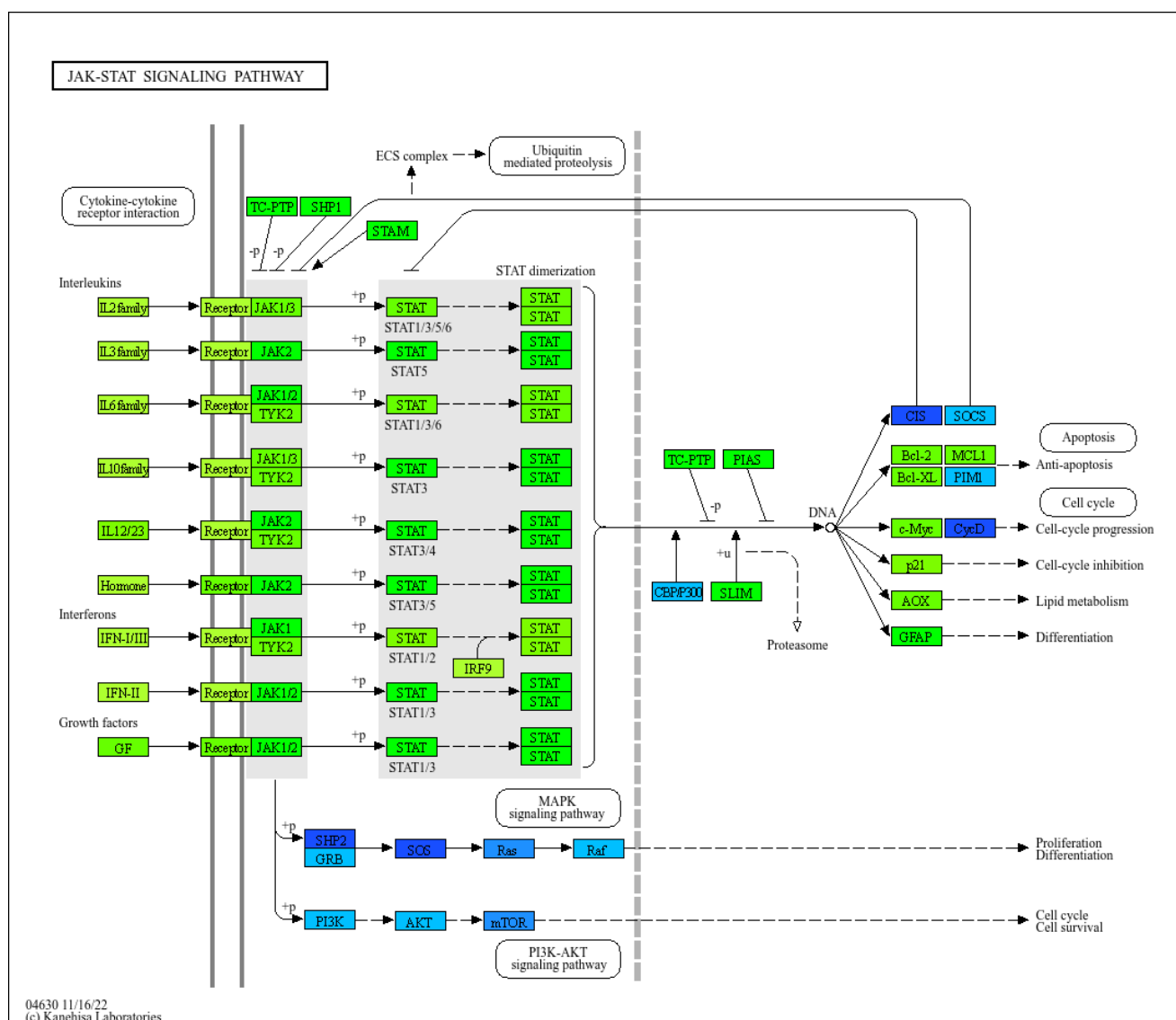
# HEDGEHOG SIGNALING PATHWAY



**Рисунок 24** Результаты филостратиграфического анализа генов сети Hedgehog сигнального пути из базы данных KEGG PATHWAY hsa: 04340

JAK-STAT путь (Рис. 25) передаёт сигналы от цитокинов и факторов роста в ядро клетки, где регулируется экспрессия генов. После связывания лиганда с рецептором происходит активация тирозинкиназ JAK (Janus kinases): JAK1, JAK2, JAK3 – принадлежащие к филострате Позвоночных (Vertebrata, PAI=6). Активированные киназы JAK фосфорилируют тирозиновые остатки на рецепторе. Фосфорилированные тирозиновые остатки на рецепторе служат местами присоединения для белков STAT (Signal Transducers and Activators of Transcription). Белки семейства STAT кодируются генами, чье происхождение также прослеживается до филостраты

происхождения позвоночных (Vertebrata, PAI=6). Фосфорилированные STAT белки (STAT1, STAT3 и STAT5) диссоциируют от рецептора, димеризуются и перемещаются в ядро клетки для связи с последовательностями ДНК для регуляции их транскрипции. JAK-STAT путь чрезвычайно важен для нормального функционирования иммунной системы, а его дисрегуляция связана с аутоиммунными нарушениями, иммунодефицитными состояниями и злокачественными новообразованиями [Brooks, Putoczki, 2020].



**Рисунок 25** Результаты филогенетического анализа генов сети JAK-STAT сигнального пути из базы данных KEGG PATHWAY hsa: 04630.

Пути сигналов клеточного цикла представляют собой процессы, регулирующие последовательность событий, приводящих к делению клетки и

её репликации. В нормальных условиях эти механизмы строго контролируются, что предотвращает нарушения клеточного цикла и связанные с ними злокачественные трансформации. В исследуемой генной сети в сигнальных путях регуляции клеточного цикла указан ряд ключевых белков этих процессов, которые по их возрастам можно выделить в две группы: эволюционно «молодых» (Vertebrata, PAI = 6) регуляторов клеточного цикла, к которым относятся p53, MDM2 и белки семейства ингибиторов циклин-зависимых киназ; и эволюционно «древних» регуляторов (Cellular Organism, Metazoa, PAI = 1, 3), в которую входят Rb1, MYC, и циклин-зависимые киназы:

- p53 – это белок-супрессор опухолей, который играет ключевую роль в предотвращении развития новообразований. Он активируется в ответ на различные стрессовые сигналы, такие как повреждение ДНК, и может запускать процессы, направленные на остановку клеточного цикла для ремонта ДНК, активацию апоптоза (программируемой клеточной смерти) или стабилизацию генома.
- MDM2 – это белок, который регулирует активность p53, связываясь с ним и способствуя его деградации. При определённых условиях это позволяет клетке продолжать деление. Однако при злокачественных заболеваниях чрезмерная активность MDM2 может привести к недостаточной активации p53, что способствует выживанию и размножению опухолевых клеток.
- Ингибиторы циклин-зависимых киназ p15, p16, p21 и p27 участвуют в регуляции клеточного цикла блокируя активность циклин-зависимых киназ (ЦЗК, CDK) путем прямого связывания с ними или с комплексами циклин-CDK, предотвращая фосфорилирование целевых белков и останавливая прогрессирование клеточного цикла.
- Ретинобластосома Rb1 и ЦЗК (CDK4, CDK6) – ключевые регуляторы клеточного цикла, которые предотвращают переход из фазы G1 в S за

счёт ингибирования факторов транскрипции, необходимых для этого перехода. Активация CDK4/6 фосфорилирует Rb1, что приводит к его инактивации и позволяет клетке перейти к следующей фазе цикла. Ортологи гена ретинабластомы прослеживаются до филостраты возникновения клеточных организмов, тогда как ортологи циклин-зависимых киназ – до филостраты возникновения многоклеточных организмов.

Проведённый филостратиграфический анализ ключевых сигнальных путей, ассоциированных с онкологическими процессами, выявил существенные различия в значениях PAI их генов, что позволяет сделать ряд обобщений. Сравнительный анализ путей показал контрастные эволюционные паттерны: сигнальный путь Wnt/ $\beta$ -catenin характеризуется исключительно консервативным молекулярным ядром, где основные компоненты (лиганды, рецепторы, эффекторные белки) возникли на этапе становления многоклеточных организмов (Metazoa) или даже раньше (Eukaryota). Напротив, JAK-STAT путь представляет собой эволюционно позднюю систему, поскольку как рецепторный комплекс (JAK киназы), так и ядерные эффекторы (STAT факторы) появились лишь на этапе возникновения позвоночных (Vertebrata). Это согласуется с его специализированной ролью в регуляции сложных межклеточных взаимодействий, характерных для развитой иммунной системы высших организмов.

Наиболее показательными с точки зрения эволюционной динамики оказались пути ERK-MAPK и PI3K/Akt, демонстрирующие смешанный возраст элементов сетей. В обоих случаях наблюдается четкая стратификация: рецепторный аппарат на клеточной поверхности, воспринимающий внешние сигналы (ростовые факторы, цитокины), кодируется преимущественно эволюционно «молодыми» генами (PAI=6-8, Vertebrata-Mammalia). В то же время, внутриклеточные сигнальные пути, представленные серинами/треониновыми киназами (Raf, MEK, ERK в MAPK пути; Akt, mTOR

в R3K пути), кодируются «древними» генами (PAI=1-3, Eukaryota-Metazoa). Подобное распределение индексов филостратиграфического возраста генов, при котором новые системы восприятия внеклеточных сигналов через специфичные рецепторы интегрируются в древние, консервативные сигнальные подсети, выполняющие базовые клеточные функции в ходе эволюции позвоночных и млекопитающих, обеспечивает эволюционную пластичность, позволяя тонко регулировать ключевые процессы в ответ на изменчивую среду многоклеточного организма.

## Заключение

Данная работа посвящена комплексному филостратиграфическому анализу молекулярных механизмов канцерогенеза, интегрирующему методы биоинформатики, транскриптомики и эволюционной биологии. Результаты исследования позволяют сделать выводы о фундаментальных закономерностях эволюционной динамики транскриптомов в процессе опухолевой прогрессии.

Разработанный веб-сервис Orthoweb 2 и автоматизированный конвейер для обработки данных TCGA – *cancEv* позволили провести масштабный филостратиграфический и филотранскриптомный анализ 9 карцином человека. Установлено статистически значимое преобладание эволюционно «молодых» генов (Vertebrata, PAI = 6) и эволюционно «древних» (Cellular Organism, PAI = 1) среди дифференциально экспрессирующихся генов во всех изученных типах опухолей, что указывает на их функциональную значимость в регуляции опухолеспецифических процессов. Данное наблюдение согласуется с атавистической теорией канцерогенеза, согласно которой злокачественная трансформация сопровождается реактивацией глубоко консервативных генетических программ [Chen, He, 2016b; Davies, Lineweaver, 2011; Bussey, Davies, 2021; Chen et al., 2015; Thomas et al., 2017; Vincent, 2012]. Анализ индексов дивергенции дифференциально экспрессирующихся генов показал высокую консервативность их последовательностей.

Эти закономерности прослеживаются также в результатах филотранскриптомного анализа: были обнаружены статистически значимо более высокие индексы возраста транскриптома здоровых тканей по сравнению с возрастами транскриптома злокачественных тканей, что может быть объяснено как общим увеличением активности эволюционно «древних» генов при онкологических заболеваниях, так и влиянием опухолей на окружающие ткани, из которых отбирались образцы здоровых тканей.

Также было выявлено значительно более высокое значение индекса возраста транскриптома карциномы печени, что соотносится с повышением активности генов, связанных с метаболизмом лекарственных препаратов, применяемых в терапии онкологических заболеваний.

Одновременно с этим был выявлен нелинейный паттерн «песочных часов» в индексах возраста транскриптомов (TAI) при прогрессировании опухолей в ряде карцином. Установлено, что усиление экспрессии эволюционно «молодых» генов на ранних и поздних стадиях сочетается с повышенной экспрессией эволюционно «древних» генов на промежуточных стадиях, что противоречит гипотезу однонаправленной регрессии к одноклеточному состоянию. Данный паттерн, статистически значимый для карцином лёгкого, печени, молочной железы и мочевого пузыря, отражает сложную адаптационную логику опухолевой прогрессии. Однако в аденокарциномах кишечника наблюдается противоположная картина – на начальных и поздних стадиях опухоли возраст транскриптома снижается, а на промежуточных стадиях – увеличивается.

Анализ распределения филогенетических индексов (PAI) генов, входящих в ключевые сигнальные пути, выявил стратификацию их эволюционного возраста. Было обнаружено, что некоторые сигнальные пути, такие как сигнальные пути WNT/ $\beta$ -catenin и Hedgehog, состоят из генов с очень ранним происхождением, что подчёркивает их фундаментальную роль в развитии и функционировании злокачественных клеток. В то же время, другие ключевые сигнальные пути онкологических процессов включали цитоплазматические белки раннего эволюционного происхождения, тогда как их регуляторы и активирующие рецепторы относились к генам более позднего происхождения. Такая архитектура отражает интеграцию новоприобретённых регуляторных элементов с консервативными молекулярными механизмами в ходе эволюции многоклеточных организмов, обеспечивая адаптацию к специфическим условиям тканевого микроокружения.

Данное исследование формирует методологическую основу для эволюционной транскриптомики онкологических заболеваний, подчёркивая значимость филогенетического контекста при анализе молекулярных механизмов злокачественной трансформации.

## Выводы

1. Разработаны программные комплексы *Orthoweb 2* для филостратиграфического и филотранскриптомного анализа генных сетей и *cancEv* для анализа транскриптомных данных из базы The Cancer Genome Atlas (TCGA) на разных клинических стадиях опухолей.
2. Выявлено, что в карциномах лёгкого, печени, мочевого пузыря, простаты и молочной железы, независимо от клинической стадии опухоли, повышена экспрессия генов, ассоциированных с клеточной пролиферацией и репарацией ДНК, и снижена активность проапоптотических генов.
3. Показано, что в тканях карцином доля дифференциально экспрессирующихся генов, относящихся к «эволюционно молодым» (PAI = 6, 7, 9), достоверно выше, чем их доля в общем составе генома человека.
4. Показано, что в карциномах лёгкого, печени, мочевого пузыря и простаты, а также в протоковой карциноме молочной железы на ранних и поздних стадиях преобладает экспрессия эволюционно «молодых» генов, тогда как на промежуточных – эволюционно «древних».
5. Установлено, что ключевые сигнальные пути, вовлечённые в онкогенез, различаются по эволюционному возрасту своих компонентов: пути WNT/ $\beta$ -catenin и Hedgehog содержат преимущественно эволюционно «древние» гены (PAI = 1–2), путь JAK–STAT представлен в основном «молодыми» генами (PAI = 6), тогда как в ERK–MAPK и PI3K/Akt присутствуют как эволюционно «древние» гены, кодирующие цитоплазматические белки, так и более «молодые», кодирующие регуляторные и рецепторные белки.

### Список используемой литературы

1. Степанов, В. А. ЭВОЛЮЦИЯ ГЕНЕТИЧЕСКОГО РАЗНООБРАЗИЯ И БОЛЕЗНИ ЧЕЛОВЕКА / В. А. Степанов // Генетика. -- 2016. -- Т. 52. -- № 7. - - С. 852-864.
2. Albuquerque, T. A. F. From humans to hydra: patterns of cancer across the tree of life / T. A. F. Albuquerque, L. Drummond do Val, A. Doherty, J. P. de Magalhães // Biological Reviews. -- 2018. -- Т. 93. -- № 3. -- С. 1715-1734.
3. Almond, D. An adaptive significance of morning sickness? Trivers–Willard and Hyperemesis Gravidarum / D. Almond, L. Edlund, M. Joffe, M. Palme // Economics & Human Biology. -- 2016. -- Т. 21. -- С. 167-171.
4. Al-Sukhni, W. Germline BRCA1 mutations predispose to pancreatic adenocarcinoma / W. Al-Sukhni, H. Rothenmund, A. Eppel Borgida и др. // Human Genetics. -- 2008. -- Т. 124. -- № 3. -- С. 271-278.
5. Álvarez-Carretero, S. Beginner’s Guide on the Use of PAML to Detect Positive Selection / S. Álvarez-Carretero, P. Kapli, Z. Yang // Molecular Biology and Evolution. -- 2023. -- Т. 40. -- № 4.
6. An, N. A. De novo genes with an lncRNA origin encode unique human brain developmental functionality / N. A. An, J. Zhang, F. Mo и др. // Nature Ecology & Evolution. -- 2023.
7. Aran, D. Comprehensive analysis of normal adjacent to tumor transcriptomes / D. Aran, R. Camarda, J. Odegaard и др. // Nature Communications. -- 2017. -- Т. 8. -- № 1. -- С. 1077.
8. Arendsee, Z. fagin: synteny-based phylostratigraphy and finer classification of young genes / Z. Arendsee, J. Li, U. Singh и др. // BMC Bioinformatics. -- 2019a. -- Т. 20. -- № 1. -- С. 440.
9. Arendsee, Z. phylostratr : a framework for phylostratigraphy / Z. Arendsee, J. Li, U. Singh и др. // Bioinformatics. -- 2019b. -- Т. 35. -- № 19. -- С. 3617-3627.
10. Arendsee, Z. W. Coming of age: orphan genes in plants / Z. W. Arendsee, L. Li, E. S. Wurtele // Trends in Plant Science. -- 2014. -- Т. 19. -- № 11. -- С. 698-708.
11. Ashburner, M. Gene Ontology: tool for the unification of biology / M. Ashburner, C. A. Ball, J. A. Blake и др. // Nature Genetics. -- 2000. -- Т. 25. -- № 1. -- С. 25-29.
12. Baalsrud, H. T. De Novo Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data / H. T. Baalsrud, O. K.

- Tørresen, M. H. Solbakken и др. // *Molecular Biology and Evolution*. -- 2018. -- Т. 35. -- № 3. -- С. 593-606.
13. Bamford, S. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website / S. Bamford, E. Dawson, S. Forbes и др. // *British Journal of Cancer*. -- 2004. -- Т. 91. -- № 2. -- С. 355-358.
  14. Barrera-Redondo, J. Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using GenEra / J. Barrera-Redondo, J. S. Lotharukpong, H.-G. Drost, S. M. Coelho // *Genome Biology*. -- 2023. -- Т. 24. -- № 1. -- С. 54.
  15. Benton, M. L. The influence of evolutionary history on human health and disease / M. L. Benton, A. Abraham, A. L. LaBella и др. // *Nature Reviews Genetics*. -- 2021. -- Т. 22. -- № 5. -- С. 269-283.
  16. Berardini, T. Z. The arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome / T. Z. Berardini, L. Reiser, D. Li и др. // *genesis*. -- 2015. -- Т. 53. -- № 8. -- С. 474-485.
  17. Berbée, J. F. P. Apolipoprotein CI Knock-Out Mice Display Impaired Memory Functions / J. F. P. Berbée, T. Vanmierlo, K. Abildayeva и др. // *Journal of Alzheimer’s Disease*. -- 2011. -- Т. 23. -- № 4. -- С. 737-747.
  18. Bowles, A. M. C. The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty / A. M. C. Bowles, U. Bechtold, J. Paps // *Current Biology*. -- 2020. -- Т. 30. -- № 3. -- С. 530-536.e2.
  19. Brooks, A. J. JAK-STAT Signalling Pathway in Cancer / A. J. Brooks, T. Putoczki // *Cancers*. -- 2020. -- Т. 12. -- № 7. -- С. 1971.
  20. Brunet, T. The Origin of Animal Multicellularity and Cell Differentiation / T. Brunet, N. King // *Developmental Cell*. -- 2017. -- Т. 43. -- № 2. -- С. 124-140.
  21. Buchfink, B. Sensitive protein alignments at tree-of-life scale using DIAMOND / B. Buchfink, K. Reuter, H.-G. Drost // *Nature Methods*. -- 2021. -- Т. 18. -- № 4. -- С. 366-368.
  22. Camacho, C. BLAST+: architecture and applications / C. Camacho, G. Coulouris, V. Avagyan и др. // *BMC Bioinformatics*. -- 2009. -- Т. 10. -- № 1. -- С. 421.
  23. Cancer Risks Associated With BRCA1 and BRCA2 Pathogenic Variants / S. Li, V. Silvestri, G. Leslie и др. // *Journal of Clinical Oncology*. -- 2022. -- Т. 40. -- № 14. -- С. 1529-1541.

24. Carroll, S. P. Applying evolutionary biology to address global challenges / S. P. Carroll, P. S. Jørgensen, M. T. Kinnison и др. // *Science*. -- 2014. -- Т. 346. -- № 6207.
25. Carvunis, A.-R. Proto-genes and de novo gene birth / A.-R. Carvunis, T. Rolland, I. Wapinski и др. // *Nature*. -- 2012. -- Т. 487. -- № 7407. -- С. 370-374.
26. Cass, I. BRCA-Mutation–Associated Fallopian Tube Carcinoma / I. Cass, C. Holschneider, N. Datta и др. // *Obstetrics & Gynecology*. -- 2005. -- Т. 106. -- № 6. -- С. 1327-1334.
27. Chandran, U. R. Differences in gene expression in prostate cancer, normal appearing prostate tissue adjacent to cancer and prostate tissue from cancer free organ donors / U. R. Chandran, R. Dhir, C. Ma и др. // *BMC Cancer*. -- 2005. -- Т. 5. -- № 1. -- С. 45.
28. Chen, H. The Convergent Cancer Evolution toward a Single Cellular Destination / H. Chen, X. He // *Molecular Biology and Evolution*. -- 2016a. -- Т. 33. -- № 1. -- С. 4-12.
29. Chen, H. The Convergent Cancer Evolution toward a Single Cellular Destination / H. Chen, X. He // *Molecular Biology and Evolution*. -- 2016b. -- Т. 33. -- № 1. -- С. 4-12.
30. Chokoshvili, D. Expanded carrier screening for monogenic disorders: where are we now? / D. Chokoshvili, D. Vears, P. Borry // *Prenatal Diagnosis*. -- 2018. -- Т. 38. -- № 1. -- С. 59-66.
31. Cridge, A. G. The Developmental Hourglass in the Evolution of Embryogenesis / A. G. Cridge, P. K. Dearden, L. R. Brownfield // *Evolutionary Developmental Biology*. – Cham : Springer International Publishing, 2019. – С. 1-10.
32. Davies, P. C. W. Cancer tumors as Metazoa 1.0: tapping genes of ancient ancestors / P. C. W. Davies, C. H. Lineweaver // *Physical Biology*. -- 2011. -- Т. 8. -- № 1. -- С. 015001.
33. Dhatchinamoorthy, K. Cancer Immune Evasion Through Loss of MHC Class I Antigen Presentation / K. Dhatchinamoorthy, J. D. Colbert, K. L. Rock // *Frontiers in Immunology*. -- 2021. -- Т. 12. -- С. 636568.
34. Dohmen, E. The modular nature of protein evolution: domain rearrangement rates across eukaryotic life / E. Dohmen, S. Klasberg, E. Bornberg-Bauer и др. // *BMC Evolutionary Biology*. -- 2020. -- Т. 20. -- № 1. -- С. 30.

35. Domazet-Lošo, T. An Evolutionary Analysis of Orphan Genes in *Drosophila* / T. Domazet-Lošo // *Genome Research*. -- 2003. -- T. 13. -- № 10. -- C. 2213-2219.
36. Domazet-Lošo, T. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages / T. Domazet-Lošo, J. Brajković, D. Tautz // *Trends in Genetics*. -- 2007a. -- T. 23. -- № 11. -- C. 533-539.
37. Domazet-Lošo, T. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages / T. Domazet-Lošo, J. Brajković, D. Tautz // *Trends in Genetics*. -- 2007b. -- T. 23. -- № 11. -- C. 533-539.
38. Domazet-Lošo, T. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns / T. Domazet-Lošo, D. Tautz // *Nature*. -- 2010a. -- T. 468. -- № 7325. -- C. 815-819.
39. Domazet-Lošo, T. An ancient evolutionary origin of genes associated with human genetic diseases / T. Domazet-Lošo, D. Tautz // *Molecular Biology and Evolution*. -- 2008a. -- T. 25. -- № 12. -- C. 2699-2707.
40. Domazet-Lošo, T. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa / T. Domazet-Lošo, D. Tautz // *BMC Biology*. -- 2010b. -- T. 8. -- № 1. -- C. 66.
41. Dornburg, A. On the relationship between extant innate immune receptors and the evolutionary origins of jawed vertebrate adaptive immunity / A. Dornburg, J. A. Yoder // *Immunogenetics*. -- 2022. -- T. 74. -- № 1. -- C. 111-128.
42. Drost, H.-G. Evidence for Active Maintenance of Phylotranscriptomic Hourglass Patterns in Animal and Plant Embryogenesis / H.-G. Drost, A. Gabel, I. Grosse, M. Quint // *Molecular Biology and Evolution*. -- 2015. -- T. 32. -- № 5. -- C. 1221-1231.
43. Drost, H.-G. myTAI: evolutionary transcriptomics with R / H.-G. Drost, A. Gabel, J. Liu и др. // *Bioinformatics*. -- 2018. -- T. 34. -- № 9. -- C. 1589-1590.
44. Drost, H.-G. Cross-kingdom comparison of the developmental hourglass / H.-G. Drost, P. Janitza, I. Grosse, M. Quint // *Current Opinion in Genetics & Development*. -- 2017. -- T. 45. -- C. 69-75.
45. Edwards, J. G. Expanded Carrier Screening in Reproductive Medicine—Points to Consider / J. G. Edwards, G. Feldman, J. Goldberg и др. // *Obstetrics & Gynecology*. -- 2015. -- T. 125. -- № 3. -- C. 653-662.
46. Emms, D. M. OrthoFinder: phylogenetic orthology inference for comparative genomics / D. M. Emms, S. Kelly // *Genome Biology*. -- 2019. -- T. 20. -- № 1. -- C. 238.

47. Enriquez-Navas, P. M. Application of Evolutionary Principles to Cancer Therapy / P. M. Enriquez-Navas, J. W. Wojtkowiak, R. A. Gatenby // *Cancer Research*. -- 2015. -- T. 75. -- № 22. -- C. 4675-4680.
48. Evans, S. S. Fever and the thermal regulation of immunity: the immune system feels the heat / S. S. Evans, E. A. Repasky, D. T. Fisher // *Nature Reviews Immunology*. -- 2015. -- T. 15. -- № 6. -- C. 335-349.
49. Fay, J. C. Disease consequences of human adaptation / J. C. Fay // *Applied & Translational Genomics*. -- 2013. -- T. 2. -- C. 42-47.
50. Fruman, D. A. PI3K and cancer: lessons, challenges and opportunities / D. A. Fruman, C. Rommel // *Nature Reviews Drug Discovery*. -- 2014. -- T. 13. -- № 2. -- C. 140-156.
51. Futuyma, D. J. EVOLUTIONARY CONSTRAINT AND ECOLOGICAL CONSEQUENCES / D. J. Futuyma // *Evolution*. -- 2010. -- T. 64. -- № 7. -- C. 1865-1884.
52. Gerlinger, M. Cancer: Evolution Within a Lifetime / M. Gerlinger, N. McGranahan, S. M. Dewhurst и др. // *Annual Review of Genetics*. -- 2014. -- T. 48. -- № 1. -- C. 215-236.
53. Goldman, N. A codon-based model of nucleotide substitution for protein-coding DNA sequences. / N. Goldman, Y. Ziheng // *Molecular Biology and Evolution*. -- 1994.
54. Gramates, L. S. FlyBase: a guided tour of highlighted features / L. S. Gramates, J. Agapite, H. Attrill и др. // *Genetics*. -- 2022. -- T. 220. -- № 4.
55. Greaves, M. Clonal evolution in cancer / M. Greaves, C. C. Maley // *Nature*. -- 2012. -- T. 481. -- № 7381. -- C. 306-313.
56. Grochola, L. F. Recent Natural Selection Identifies a Genetic Variant in a Regulatory Subunit of Protein Phosphatase 2A that Associates with Altered Cancer Risk and Survival / L. F. Grochola, A. Vazquez, E. E. Bond и др. // *Clinical Cancer Research*. -- 2009. -- T. 15. -- № 19. -- C. 6301-6308.
57. Guo, Y. ERK/MAPK signalling pathway and tumorigenesis (Review) / Y. Guo, W. Pan, S. Liu и др. // *Experimental and Therapeutic Medicine*. -- 2020.
58. Huang, X. Transcriptional Profiles from Paired Normal Samples Offer Complementary Information on Cancer Patient Survival – Evidence from TCGA Pan-Cancer Data / X. Huang, D. F. Stern, H. Zhao // *Scientific Reports*. -- 2016. -- T. 6. -- № 1. -- C. 20567.

59. Huang, Y.-F. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data / Y.-F. Huang, B. Gulko, A. Siepel // *Nature Genetics*. -- 2017. -- T. 49. -- № 4. -- C. 618-624.
60. Irie, N. The developmental hourglass model: a predictor of the basic body plan? / N. Irie, S. Kuratani // *Development*. -- 2014. -- T. 141. -- № 24. -- C. 4649-4655.
61. Kimura, M. The neutral theory of molecular evolution / M. Kimura. – Cambridge; New York : Cambridge University Press, 1983. – 367 c.
62. Knoll, A. H. The Multiple Origins of Complex Multicellularity / A. H. Knoll // *Annual Review of Earth and Planetary Sciences*. -- 2011. -- T. 39. -- № 1. -- C. 217-239.
63. Lee, W.-H. BRCA1 and BRCA2 in breast cancer / W.-H. Lee, T. G. Boyer // *The Lancet*. -- 2001. -- T. 358. -- C. S5.
64. Lei, L. Plant organ evolution revealed by phylotranscriptomics in *Arabidopsis thaliana* / L. Lei, J. G. Steffen, E. J. Osborne, C. Toomajian // *Scientific Reports*. -- 2017. -- T. 7. -- № 1. -- C. 7567.
65. Li, W.-H. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. / W.-H. Li, C.-I. Wu, C.-C. Luo // *Molecular Biology and Evolution*. -- 1985.
66. Li, X. J. Interacting proteins as genetic modifiers of Huntington disease / X. J. Li, M. Friedman, S. Li // *Trends in Genetics*. -- 2007. -- T. 23. -- № 11. -- C. 531-533.
67. Masatoshi, N. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. / N. Masatoshi, G. Takashi // *Molecular Biology and Evolution*. -- 1986.
68. Mauland, R. Sesame: A power spectrum emulator pipeline for beyond- $\Lambda$ CDM models / R. Mauland, H. A. Winther, C.-Z. Ruan // *Astronomy & Astrophysics*. -- 2024. -- T. 685. -- Sesame. -- C. 156.
69. McLysaght, A. Open questions in the study of de novo genes: what, how and why / A. McLysaght, L. D. Hurst // *Nature Reviews Genetics*. -- 2016. -- T. 17. -- № 9. -- C. 567-578.
70. Mefford, H. C. Duplication hotspots, rare genomic disorders, and common disease / H. C. Mefford, E. E. Eichler // *Current Opinion in Genetics & Development*. -- 2009. -- T. 19. -- № 3. -- C. 196-204.

71. Mendoza, A. de. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages / A. de Mendoza, A. Seb -Pedr s, M. S.  estak и др. // *Proceedings of the National Academy of Sciences*. -- 2013. -- T. 110. -- № 50.
72. Mering, C. von. STRING: Known and predicted protein-protein associations, integrated and transferred across organisms / C. von Mering, L. J. Jensen, B. Snel и др. // *Nucleic Acids Research*. -- 2005. -- T. 33. -- № DATABASE ISS. -- C. 433-437.
73. Merlo, L. M. F. Cancer as an evolutionary and ecological process / L. M. F. Merlo, J. W. Pepper, B. J. Reid, C. C. Maley // *Nature Reviews Cancer*. -- 2006. -- T. 6. -- № 12. -- C. 924-935.
74. Michor, F. Dynamics of cancer progression / F. Michor, Y. Iwasa, M. A. Nowak // *Nature Reviews Cancer*. -- 2004. -- T. 4. -- № 3. -- C. 197-205.
75. Moutinho, A. F. Strong evidence for the adaptive walk model of gene evolution in *Drosophila* and *Arabidopsis* / A. F. Moutinho, A. Eyre-Walker, J. Y. Dutheil // *PLOS Biology*. -- 2022. -- T. 20. -- № 9. -- C. e3001775.
76. Moyers, B. A. Further Simulations and Analyses Demonstrate Open Problems of Phylostratigraphy / B. A. Moyers, J. Zhang // *Genome Biology and Evolution*. -- 2017. -- T. 9. -- № 6. -- C. 1519-1527.
77. Moyers, B. Phylostratigraphic Bias Creates Spurious Patterns of Genome Evolution / B. Moyers, J. Zhang // *Molecular Biology and Evolution*. -- 2016. -- T. 33. -- № 11. -- C. 3031-3031.
78. Mustafin, Z. S. Orthoscape: A cytoscape application for grouping and visualization KEGG based gene networks by taxonomy and homology principles / Z. S. Mustafin, S. A. Lashin, Y. G. Matushkin и др. // *BMC Bioinformatics*. -- 2017.
79. Mustafin, Z. S. Phylostratigraphic analysis of gene networks of human diseases / Z. S. Mustafin, S. A. Lashin, Yu. G. Matushkin // *Vavilov Journal of Genetics and Breeding*. -- 2021. -- T. 25. -- № 1. -- C. 46-56.
80. O'Bleness, M. Evolution of genetic and genomic features unique to the human lineage / M. O'Bleness, V. B. Searles, A. Varki и др. // *Nature Reviews Genetics*. -- 2012. -- T. 13. -- № 12. -- C. 853-866.
81. O'Neal, W. K. Cystic Fibrosis Disease Modifiers: Complex Genetics Defines the Phenotypic Diversity in a Monogenic Disease / W. K. O'Neal, M. R. Knowles // *Annual Review of Genomics and Human Genetics*. -- 2018. -- T. 19. -- № 1. -- C. 201-222.

82. Paps, J. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty / J. Paps, P. W. H. Holland // *Nature Communications*. -- 2018. -- T. 9. -- № 1. -- C. 1730.
83. Pavard, S. Negative Selection on BRCA1 Susceptibility Alleles Sheds Light on the Population Genetics of Late-Onset Diseases and Aging Theory / S. Pavard, C. J. E. Metcalf // *PLoS ONE*. -- 2007. -- T. 2. -- № 11. -- C. e1206.
84. Pepper, J. W. SYNTHESIS: Cancer research meets evolutionary biology / J. W. Pepper, C. Scott Findlay, R. Kassen и др. // *Evolutionary Applications*. -- 2009. -- T. 2. -- № 1. -- C. 62-70.
85. Piovesan, A. Human protein-coding genes and gene feature statistics in 2019 / A. Piovesan, F. Antonaros, L. Vitale и др. // *BMC Research Notes*. -- 2019. -- T. 12. -- № 1. -- C. 315.
86. Porta, C. Targeting PI3K/Akt/mTOR Signaling in Cancer / C. Porta, C. Paglino, A. Mosca // *Frontiers in Oncology*. -- 2014. -- T. 4. -- № 64. -- C. 10.
87. Quint, M. A transcriptomic hourglass in plant embryogenesis / M. Quint, H. G. Drost, A. Gabel и др. // *Nature*. -- 2012a. -- T. 490. -- № 7418. -- C. 98-101.
88. Quint, M. A transcriptomic hourglass in plant embryogenesis / M. Quint, H. G. Drost, A. Gabel и др. // *Nature*. -- 2012b. -- T. 490. -- № 7418. -- C. 98-101.
89. Rasmussen, M. D. Genome-Wide Inference of Ancestral Recombination Graphs / M. D. Rasmussen, M. J. Hubisz, I. Gronau, A. Siepel // *PLoS Genetics*. -- 2014. -- T. 10. -- № 5. -- C. e1004342.
90. Ritchie, M. E. limma powers differential expression analyses for RNA-sequencing and microarray studies / M. E. Ritchie, B. Phipson, D. Wu и др. // *Nucleic Acids Research*. -- 2015. -- T. 43. -- № 7. -- C. e47-e47.
91. Rokas, A. The Origins of Multicellularity and the Early History of the Genetic Toolkit For Animal Development / A. Rokas // *Annual Review of Genetics*. -- 2008. -- T. 42. -- № 1. -- C. 235-251.
92. Sanz-Pamplona, R. Aberrant gene expression in mucosa adjacent to tumor reveals a molecular crosstalk in colon cancer / R. Sanz-Pamplona, A. Berenguer, D. Cordero и др. // *Molecular Cancer*. -- 2014. -- T. 13. -- № 1. -- C. 46.
93. Šestak, M. S. Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems / M. S. Šestak, V. Božičević, R. Bakarić и др. // *Frontiers in Zoology*. -- 2013. -- T. 10. -- № 1. -- C. 18.
94. Sherry, S. T. dbSNP: the NCBI database of genetic variation / S. T. Sherry // *Nucleic Acids Research*. -- 2001. -- T. 29. -- № 1. -- C. 308-311.

95. Siewert, K. M. Detecting Long-Term Balancing Selection Using Allele Frequency Correlation / K. M. Siewert, B. F. Voight // *Molecular Biology and Evolution*. -- 2017. -- T. 34. -- № 11. -- C. 2996-3005.
96. Skoda, A. M. The role of the Hedgehog signaling pathway in cancer: A comprehensive review / A. M. Skoda, D. Simovic, V. Karin и др. // *Bosnian Journal of Basic Medical Sciences*. -- 2018. -- T. 18. -- The role of the Hedgehog signaling pathway in cancer. -- № 1. -- C. 8-20.
97. Speakman, J. R. Thrifty genes for obesity and the metabolic syndrome — time to call off the search? / J. R. Speakman // *Diabetes and Vascular Disease Research*. -- 2006. -- T. 3. -- № 1. -- C. 7-11.
98. Stajich, J. E. Disentangling the Effects of Demography and Selection in Human History / J. E. Stajich // *Molecular Biology and Evolution*. -- 2004. -- T. 22. -- № 1. -- C. 63-73.
99. Stearns, S. C. Trade-Offs in Life-History Evolution / S. C. Stearns // *Functional Ecology*. -- 1989. -- T. 3. -- № 3. -- C. 259.
100. Sun, C. A signature of balancing selection in the region upstream to the human UGT2B4 gene and implications for breast cancer risk / C. Sun, D. Huo, C. Southard и др. // *Human Genetics*. -- 2011. -- T. 130. -- № 6. -- C. 767-775.
101. Tabrizi, S. J. Huntington disease: new insights into molecular pathogenesis and therapeutic opportunities / S. J. Tabrizi, M. D. Flower, C. A. Ross, E. J. Wild // *Nature Reviews Neurology*. -- 2020. -- T. 16. -- № 10. -- C. 529-546.
102. Tai, Y. C. Breast Cancer Risk Among Male BRCA1 and BRCA2 Mutation Carriers / Y. C. Tai, S. Domchek, G. Parmigiani, S. Chen // *JNCI Journal of the National Cancer Institute*. -- 2007. -- T. 99. -- № 23. -- C. 1811-1814.
103. Takaoka, M. BRCA1 gene: function and deficiency / M. Takaoka, Y. Miki // *International Journal of Clinical Oncology*. -- 2018. -- T. 23. -- № 1. -- C. 36-44.
104. Tautz, D. The evolutionary origin of orphan genes / D. Tautz, T. Domazet-Lošo // *Nature Reviews Genetics*. -- 2011. -- T. 12. -- № 10. -- C. 692-702.
105. The evolutionary history of 2,658 cancers / M. Gerstung, C. Jolly, I. Leshchiner и др. // *Nature*. -- 2020. -- T. 578. -- № 7793. -- C. 122-128.
106. The Gene Ontology resource: enriching a Gold mine / S. Carbon, E. Douglass, B. M. Good и др. // *Nucleic Acids Research*. -- 2021. -- T. 49. -- № D1. -- C. D325-D334.

107. Timpson, N. J. Genetic architecture: the shape of the genetic contribution to human traits and disease / N. J. Timpson, C. M. T. Greenwood, N. Soranzo и др. // *Nature Reviews Genetics*. -- 2018. -- T. 19. -- № 2. -- С. 110-124.
108. Trigos, A. S. Altered interactions between unicellular and multicellular genes drive hallmarks of transformation in a diverse range of solid tumors / A. S. Trigos, R. B. Pearson, A. T. Papenfuss, D. L. Goode // *Proceedings of the National Academy of Sciences*. -- 2017. -- T. 114. -- № 24. -- С. 6406-6411.
109. Ullrich, K. K. oggmap: a Python package to extract gene ages per orthogroup and link them with single-cell RNA data / K. K. Ullrich, N. E. Glynnasi // *Bioinformatics*. -- 2023. -- T. 39. -- № 11.
110. Vakirlis, N. A Molecular Portrait of De Novo Genes in Yeasts / N. Vakirlis, A. S. Hebert, D. A. Opulente и др. // *Molecular Biology and Evolution*. -- 2018. -- T. 35. -- № 3. -- С. 631-645.
111. Varki, A. Uniquely human evolution of sialic acid genetics and biology / A. Varki // *Proceedings of the National Academy of Sciences*. -- 2010. -- T. 107. -- № supplement\_2. -- С. 8939-8946.
112. Waskom, M. mwaskom/seaborn: v0.8.1 (September 2017).  
mwaskom/seaborn / M. Waskom, O. Botvinnik, D. O’Kane, и др. – Zenodo, 2017.
113. Weisman, C. M. Many, but not all, lineage-specific genes can be explained by homology detection failure / C. M. Weisman, A. W. Murray, S. R. Eddy // *PLOS Biology*. -- 2020. -- T. 18. -- № 11. -- С. e3000862.
114. Wilks, C. recount3: summaries and queries for large-scale RNA-seq expression and splicing / C. Wilks, S. C. Zheng, F. Y. Chen и др. // *Genome Biology*. -- 2021. -- T. 22. -- № 1. -- С. 323.
115. Wu, L. Gene expression does not support the developmental hourglass model in three animals with Spiralian development / L. Wu, K. E. Ferger, J. David Lambert // *Molecular Biology and Evolution*. -- 2019. -- T. 36. -- № 7. -- С. 1373-1383.
116. Yang, Z. Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models / Z. Yang, R. Nielsen // *Molecular Biology and Evolution*. -- 2000. -- T. 17. -- № 1. -- С. 32-43.
117. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood / Z. Yang // *Molecular Biology and Evolution*. -- 2007. -- T. 24. -- № 8. -- С. 1586-1591.

118. Yu, G. Population Diversity of ORFan Genes in *Escherichia coli* / G. Yu, A. Stoltzfus // *Genome Biology and Evolution*. -- 2012. -- T. 4. -- № 11. -- C. 1176-1187.
119. Zhan, T. Wnt signaling in cancer / T. Zhan, N. Rindtorff, M. Boutros // *Oncogene*. -- 2017. -- T. 36. -- № 11. -- C. 1461-1473.
120. Zhang, Z. KaKs\_Calculator 3.0: Calculating Selective Pressure on Coding and Non-coding Sequences / Z. Zhang // *Genomics, Proteomics & Bioinformatics*. -- 2022. -- T. 20. -- № 3. -- C. 536-540.
121. Zhou, G. Correlation of ERK/MAPK signaling pathway with proliferation and apoptosis of colon cancer cells / G. Zhou, J. Yang, P. Song // *Oncology Letters*. -- 2018.
122. Zuckerkandl, E. Molecules as documents of evolutionary history / E. Zuckerkandl, L. Pauling // *Journal of Theoretical Biology*. -- 1965. -- T. 8. -- № 2. -- C. 357-366.
123. Alexander, W. S. The Role of Suppressors of Cytokine Signaling (SOCS) Proteins in Regulation of the Immune Response / W. S. Alexander, D. J. Hilton // *Annual Review of Immunology*. -- 2004. -- Vol. 22. -- № 1. -- P. 503-529.
124. Anastas, J. N. WNT signalling pathways as therapeutic targets in cancer / J. N. Anastas, R. T. Moon // *Nature Reviews Cancer*. -- 2013. -- Vol. 13. -- № 1. -- P. 11-26.
125. Auwera, G. V. der. Genomics in the cloud: using Docker, GATK, and WDL in Terra. Genomics in the cloud / G. V. der Auwera, B. D. O'Connor. – First edition. – Beijing Boston Farnham Sebastopol Tokyo : O'Reilly, 2020. – 1 p.
126. Baxter, E. J. Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders / E. J. Baxter, L. M. Scott, P. J. Campbell et al. // *The Lancet*. -- 2005. -- Vol. 365. -- № 9464. -- P. 1054-1061.
127. Baylin, S. B. DNA methylation and gene silencing in cancer / S. B. Baylin // *Nature Clinical Practice Oncology*. -- 2005. -- Vol. 2. -- № S1. -- P. 4-11.
128. Baylin, S. B. A decade of exploring the cancer epigenome — biological and translational implications / S. B. Baylin, P. A. Jones // *Nature Reviews Cancer*. -- 2011. -- Vol. 11. -- № 10. -- P. 726-734.
129. Brison, O. Gene amplification and tumor progression / O. Brison // *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*. -- 1993. -- Vol. 1155. -- № 1. -- P. 25-41.

130. Burkhart, D. L. Cellular mechanisms of tumour suppression by the retinoblastoma gene / D. L. Burkhart, J. Sage // *Nature Reviews Cancer*. -- 2008. -- Vol. 8. -- № 9. -- P. 671-682.
131. Bussey, K. J. Reverting to single-cell biology: The predictions of the atavism theory of cancer / K. J. Bussey, P. C. W. Davies // *Progress in Biophysics and Molecular Biology*. -- 2021. -- Vol. 165. -- Reverting to single-cell biology. -- P. 49-55.
132. Bustamante, C. D. Natural selection on protein-coding genes in the human genome / C. D. Bustamante, A. Fledel-Alon, S. Williamson et al. // *Nature*. -- 2005. -- Vol. 437. -- № 7062. -- P. 1153-1157.
133. Cancer Genome Atlas Research Network. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas / Cancer Genome Atlas Research Network, J. D. Campbell, A. Alexandrov et al. // *Nature Genetics*. -- 2016. -- Vol. 48. -- № 6. -- P. 607-616.
134. Carmeliet, P. Principles and mechanisms of vessel normalization for cancer and other angiogenic diseases / P. Carmeliet, R. K. Jain // *Nature Reviews Drug Discovery*. -- 2011. -- Vol. 10. -- № 6. -- P. 417-427.
135. Carneiro, B. A. Targeting apoptosis in cancer therapy / B. A. Carneiro, W. S. El-Deiry // *Nature Reviews Clinical Oncology*. -- 2020. -- Vol. 17. -- № 7. -- P. 395-417.
136. Castro-Castro, A. Cellular and Molecular Mechanisms of MT1-MMP-Dependent Cancer Cell Invasion / A. Castro-Castro, V. Marchesin, P. Monteiro et al. // *Annual Review of Cell and Developmental Biology*. -- 2016. -- Vol. 32. -- № 1. -- P. 555-576.
137. Chapman, P. B. Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation / P. B. Chapman, A. Hauschild, C. Robert et al. // *New England Journal of Medicine*. -- 2011. -- Vol. 364. -- № 26. -- P. 2507-2516.
138. Chen, H. The reverse evolution from multicellularity to unicellularity during carcinogenesis / H. Chen, F. Lin, K. Xing, X. He // *Nature Communications*. -- 2015. -- Vol. 6. -- № 1. -- P. 6367.
139. Chen, Y. Prevalence of *PRKDC* mutations and association with response to immune checkpoint inhibitors in solid tumors / Y. Chen, Y. Li, Y. Guan et al. // *Molecular Oncology*. -- 2020. -- Vol. 14. -- № 9. -- P. 2096-2110.
140. Chen, Y. Role of *PRKDC* in cancer initiation, progression, and treatment / Y. Chen, Y. Li, J. Xiong et al. // *Cancer Cell International*. -- 2021. -- Vol. 21. -- № 1. -- P. 563.

141. Cserni, G. The new TNM-based staging of breast cancer / G. Cserni, E. Chmielik, B. Cserni, T. Tot // *Virchows Archiv.* -- 2018. -- Vol. 472. -- № 5. -- P. 697-703.
142. Dang, C. V. MYC on the Path to Cancer / C. V. Dang // *Cell.* -- 2012. -- Vol. 149. -- № 1. -- P. 22-35.
143. Davies, H. Mutations of the BRAF gene in human cancer / H. Davies, G. R. Bignell, C. Cox et al. // *Nature.* -- 2002. -- Vol. 417. -- № 6892. -- P. 949-954.
144. Degenhardt, K. BAX and BAK mediate p53-independent suppression of tumorigenesis / K. Degenhardt, G. Chen, T. Lindsten, E. White // *Cancer Cell.* -- 2002. -- Vol. 2. -- № 3. -- P. 193-203.
145. Delbridge, A. R. D. The BCL-2 protein family, BH3-mimetics and cancer therapy / A. R. D. Delbridge, A. Strasser // *Cell Death & Differentiation.* -- 2015. -- Vol. 22. -- № 7. -- P. 1071-1080.
146. Dobin, A. STAR: ultrafast universal RNA-seq aligner / A. Dobin, C. A. Davis, F. Schlesinger et al. // *Bioinformatics.* -- 2013. -- Vol. 29. -- STAR. -- № 1. -- P. 15-21.
147. Domazet-Lošo, T. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages / T. Domazet-Lošo, J. Brajković, D. Tautz // *Trends in Genetics.* -- 2007c. -- Vol. 23. -- № 11. -- P. 533-539.
148. Domazet-Lošo, T. An ancient evolutionary origin of genes associated with human genetic diseases / T. Domazet-Lošo, D. Tautz // *Molecular Biology and Evolution.* -- 2008b. -- Vol. 25. -- № 12. -- P. 2699-2707.
149. Dongre, A. New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer / A. Dongre, R. A. Weinberg // *Nature Reviews Molecular Cell Biology.* -- 2019. -- Vol. 20. -- № 2. -- P. 69-84.
150. Drummond, D. A. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution / D. A. Drummond, C. O. Wilke // *Cell.* -- 2008. -- Vol. 134. -- № 2. -- P. 341-352.
151. Dujon, A. M. Identifying key questions in the ecology and evolution of cancer / A. M. Dujon, A. Aktipis, C. Alix- Panabières et al. // *Evolutionary Applications.* -- 2021. -- Vol. 14. -- № 4. -- P. 877-892.
152. Dyson, N. J. *RB1* : a prototype tumor suppressor and an enigma / N. J. Dyson // *Genes & Development.* -- 2016. -- Vol. 30. -- *RB1*. -- № 13. -- P. 1492-1502.

153. Engeland, K. Cell cycle regulation: p53-p21-RB signaling / K. Engeland // Cell Death & Differentiation. -- 2022. -- Vol. 29. -- Cell cycle regulation. -- № 5. -- P. 946-960.
154. Evan, G. I. Proliferation, cell cycle and apoptosis in cancer / G. I. Evan, K. H. Vousden // Nature. -- 2001. -- Vol. 411. -- № 6835. -- P. 342-348.
155. Eyre-Walker, A. The distribution of fitness effects of new mutations / A. Eyre-Walker, P. D. Keightley // Nature Reviews Genetics. -- 2007. -- Vol. 8. -- № 8. -- P. 610-618.
156. Falini, B. Proteins encoded by genes involved in chromosomal alterations in lymphoma and leukemia: clinical value of their detection by immunocytochemistry / B. Falini, D. Y. Mason // Blood. -- 2002. -- Vol. 99. -- Proteins encoded by genes involved in chromosomal alterations in lymphoma and leukemia. -- № 2. -- P. 409-426.
157. Fruman, D. A. The PI3K Pathway in Human Disease / D. A. Fruman, H. Chiu, B. D. Hopkins et al. // Cell. -- 2017. -- Vol. 170. -- № 4. -- P. 605-635.
158. Futuyma, D. J. Evolution / D. J. Futuyma, M. Kirkpatrick. – Fourth edition. – Sunderland, Massachusetts : Sinauer Associates, Inc., Publishers, 2017. – 599 p.
159. Gabaldón, T. Functional and evolutionary implications of gene orthology / T. Gabaldón, E. V. Koonin // Nature Reviews Genetics. -- 2013. -- Vol. 14. -- № 5. -- P. 360-366.
160. Garg, H. Survivin: a unique target for tumor therapy / H. Garg, P. Suri, J. C. Gupta et al. // Cancer Cell International. -- 2016. -- Vol. 16. -- Survivin. -- № 1. -- P. 49.
161. Gilad, Y. A comparison of the human and chimpanzee olfactory receptor gene repertoires / Y. Gilad, O. Man, G. Glusman // Genome Research. -- 2005. -- Vol. 15. -- № 2. -- P. 224-230.
162. Gilad, Y. Loss of Olfactory Receptor Genes Coincides with the Acquisition of Full Trichromatic Vision in Primates / Y. Gilad, V. Wiebe, M. Przeworski et al. // PLoS Biology. -- 2004. -- Vol. 2. -- № 1. -- P. e5.
163. Gillies, R. J. Evolutionary dynamics of carcinogenesis and why targeted therapy does not work / R. J. Gillies, D. Verduzco, R. A. Gatenby // Nature Reviews Cancer. -- 2012. -- Vol. 12. -- № 7. -- P. 487-493.
164. Glaviano, A. PI3K/AKT/mTOR signaling transduction pathway and targeted therapies in cancer / A. Glaviano, A. S. C. Foo, H. Y. Lam et al. // Molecular Cancer. -- 2023. -- Vol. 22. -- № 1. -- P. 138.

165. Gong, Y. The role of necroptosis in cancer biology and therapy / Y. Gong, Z. Fan, G. Luo et al. // *Molecular Cancer*. -- 2019. -- Vol. 18. -- № 1. -- P. 100.
166. Gyrd-Hansen, M. IAPs: from caspase inhibitors to modulators of NF- $\kappa$ B, inflammation and cancer / M. Gyrd-Hansen, P. Meier // *Nature Reviews Cancer*. -- 2010. -- Vol. 10. -- IAPs. -- № 8. -- P. 561-574.
167. Hafner, A. The multiple mechanisms that regulate p53 activity and cell fate / A. Hafner, M. L. Bulyk, A. Jambhekar, G. Lahav // *Nature Reviews Molecular Cell Biology*. -- 2019. -- Vol. 20. -- № 4. -- P. 199-210.
168. Hanahan, D. Hallmarks of Cancer: The Next Generation / D. Hanahan, R. A. Weinberg // *Cell*. -- 2011. -- Vol. 144. -- Hallmarks of Cancer. -- № 5. -- P. 646-674.
169. Hanahan, D. The Hallmarks of Cancer / D. Hanahan, R. A. Weinberg // *Cell*. -- 2000. -- Vol. 100. -- № 1. -- P. 57-70.
170. Heath, A. P. The NCI Genomic Data Commons / A. P. Heath, V. Ferretti, S. Agrawal et al. // *Nature Genetics*. -- 2021. -- Vol. 53. -- № 3. -- P. 257-262.
171. Heerema, N. A. Chromosomes in Lymphomas and Solid Tumors / N. A. Heerema // *Cancer Investigation*. -- 1998. -- Vol. 16. -- № 3. -- P. 183-187.
172. Huerta-Cepas, J. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses / J. Huerta-Cepas, D. Szklarczyk, D. Heller et al. // *Nucleic Acids Research*. -- 2019. -- Vol. 47. -- № D1. -- P. D309-D314.
173. Hurst, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution / L. D. Hurst // *Trends in Genetics*. -- 2002. -- Vol. 18. -- The Ka/Ks ratio. -- № 9. -- P. 486-487.
174. Jacquet, S. Evolution of Hepatitis B Virus Receptor NTCP Reveals Differential Pathogenicities and Species Specificities of Hepadnaviruses in Primates, Rodents, and Bats / S. Jacquet, J.-B. Pons, A. De Bernardo et al. // *Journal of Virology*. -- 2019. -- Vol. 93. -- № 5. -- P. e01738-18.
175. Janiszewska, M. Cell adhesion in cancer: Beyond the migration of single cells / M. Janiszewska, M. C. Primi, T. Izard // *Journal of Biological Chemistry*. -- 2020. -- Vol. 295. -- Cell adhesion in cancer. -- № 8. -- P. 2495-2505.
176. Jia, P. Characterization of Tumor-Suppressor Gene Inactivation Events in 33 Cancer Types / P. Jia, Z. Zhao // *Cell Reports*. -- 2019. -- Vol. 26. -- № 2. -- P. 496-506.

177. Jiang, M. PI3K/AKT/mTOR Axis in Cancer: From Pathogenesis to Treatment / M. Jiang, K. Zhang, Z. Zhang et al. // *MedComm*. -- 2025. -- Vol. 6. -- PI3K/AKT/mTOR Axis in Cancer. -- № 8. -- P. e70295.
178. Jones, P. A. The Epigenomics of Cancer / P. A. Jones, S. B. Baylin // *Cell*. -- 2007. -- Vol. 128. -- № 4. -- P. 683-692.
179. Jones, P. A. The fundamental role of epigenetic events in cancer / P. A. Jones, S. B. Baylin // *Nature Reviews Genetics*. -- 2002. -- Vol. 3. -- № 6. -- P. 415-428.
180. Karlberg, M. Pro-apoptotic Bax is the major and Bak an auxiliary effector in cytokine deprivation-induced mast cell apoptosis / M. Karlberg, M. Ekoff, V. Labi et al. // *Cell Death & Disease*. -- 2010. -- Vol. 1. -- № 5. -- P. e43-e43.
181. Kim, R. Cancer immunoediting from immune surveillance to immune escape / R. Kim, M. Emi, K. Tanabe // *Immunology*. -- 2007. -- Vol. 121. -- № 1. -- P. 1-14.
182. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma / A. G. Knudson // *Proceedings of the National Academy of Sciences*. -- 1971. -- Vol. 68. -- Mutation and Cancer. -- № 4. -- P. 820-823.
183. Koonin, E. V. Orthologs, Paralogs, and Evolutionary Genomics / E. V. Koonin // *Annual Review of Genetics*. -- 2005. -- Vol. 39. -- № 1. -- P. 309-338.
184. Kosakovsky Pond, S. L. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection / S. L. Kosakovsky Pond, S. D. W. Frost // *Molecular Biology and Evolution*. -- 2005. -- Vol. 22. -- Not So Different After All. -- № 5. -- P. 1208-1222.
185. Kosakovsky Pond, S. L. HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies / S. L. Kosakovsky Pond, A. F. Y. Poon, R. Velazquez et al. // *Molecular Biology and Evolution*. -- 2020. -- Vol. 37. -- № 1. -- P. 295-299.
186. Kryazhimskiy, S. The Population Genetics of dN/dS / S. Kryazhimskiy, J. B. Plotkin // *PLoS Genetics*. -- 2008. -- Vol. 4. -- № 12. -- P. e1000304.
187. Krylov, D. M. Gene Loss, Protein Sequence Divergence, Gene Dispensability, Expression Level, and Interactivity Are Correlated in Eukaryotic Evolution / D. M. Krylov, Y. I. Wolf, I. B. Rogozin, E. V. Koonin // *Genome Research*. -- 2003. -- Vol. 13. -- № 10. -- P. 2229-2235.
188. Kumar, S. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms / S. Kumar, G. Stecher, M. Li et al. // *Molecular Biology and Evolution*. -- 2018. -- Vol. 35. -- MEGA X. -- № 6. -- P. 1547-1549.

189. Kyo, S. Understanding and exploiting *hTERT* promoter regulation for diagnosis and treatment of human cancers / S. Kyo, M. Takakura, T. Fujiwara, M. Inoue // *Cancer Science*. -- 2008. -- Vol. 99. -- № 8. -- P. 1528-1538.
190. Lake, D. Negative feedback regulation of the ERK1/2 MAPK pathway / D. Lake, S. A. L. Corrêa, J. Müller // *Cellular and Molecular Life Sciences*. -- 2016. -- Vol. 73. -- № 23. -- P. 4397-4413.
191. Lambert, G. An analogy between the evolution of drug resistance in bacterial communities and malignant tissues / G. Lambert, L. Estévez-Salmeron, S. Oh et al. // *Nature Reviews Cancer*. -- 2011. -- Vol. 11. -- № 5. -- P. 375-382.
192. Landskron, G. Chronic Inflammation and Cytokines in the Tumor Microenvironment / G. Landskron, M. De La Fuente, P. Thuwajit et al. // *Journal of Immunology Research*. -- 2014. -- Vol. 2014. -- P. 1-19.
193. Laplante, M. mTOR Signaling in Growth Control and Disease / M. Laplante, D. M. Sabatini // *Cell*. -- 2012. -- Vol. 149. -- № 2. -- P. 274-293.
194. Lavrik, I. N. Regulation of death receptor-induced apoptosis induced via CD95/Fas and other death receptors / I. N. Lavrik // *Molecular Biology*. -- 2011. -- Vol. 45. -- № 1. -- P. 150-155.
195. Le Bras, G. F. The regulation of cell-cell adhesion during epithelial-mesenchymal transition, motility and tumor progression / G. F. Le Bras, K. J. Taubenslag, C. D. Andl // *Cell Adhesion & Migration*. -- 2012. -- Vol. 6. -- № 4. -- P. 365-373.
196. Lee, H.-M. Diverse pathways of epithelial mesenchymal transition related with cancer progression and metastasis and potential effects of endocrine disrupting chemicals on epithelial mesenchymal transition process / H.-M. Lee, K.-A. Hwang, K.-C. Choi // *Molecular and Cellular Endocrinology*. -- 2017. -- Vol. 457. -- P. 103-113.
197. Lee, Y.-R. The functions and regulation of the PTEN tumour suppressor: new modes and prospects / Y.-R. Lee, M. Chen, P. P. Pandolfi // *Nature Reviews Molecular Cell Biology*. -- 2018. -- Vol. 19. -- № 9. -- P. 547-562.
198. Li, W.-H. Unbiased estimation of the rates of synonymous and nonsynonymous substitution / W.-H. Li // *Journal of Molecular Evolution*. -- 1993. -- Vol. 36. -- № 1. -- P. 96-99.
199. Li, Y. MAD2L2 inhibits colorectal cancer growth by promoting NCOA3 ubiquitination and degradation / Y. Li, L. Li, M. Chen et al. // *Molecular Oncology*. -- 2018. -- Vol. 12. -- № 3. -- P. 391-405.

200. Liberti, M. V. The Warburg Effect: How Does it Benefit Cancer Cells? / M. V. Liberti, J. W. Locasale // Trends in Biochemical Sciences. -- 2016. -- Vol. 41. -- № 3. -- P. 211-218.
201. Liu, Y. Characteristics and Significance of the Pre-metastatic Niche / Y. Liu, X. Cao // Cancer Cell. -- 2016. -- Vol. 30. -- № 5. -- P. 668-681.
202. Loh, C.-Y. The E-Cadherin and N-Cadherin Switch in Epithelial-to-Mesenchymal Transition: Signaling, Therapeutic Implications, and Challenges / C.-Y. Loh, J. Chai, T. Tang et al. // Cells. -- 2019. -- Vol. 8. -- № 10. -- P. 1118.
203. MacDonald, B. T. Wnt/ $\beta$ -Catenin Signaling: Components, Mechanisms, and Diseases / B. T. MacDonald, K. Tamai, X. He // Developmental Cell. -- 2009. -- Vol. 17. -- Wnt/ $\beta$ -Catenin Signaling. -- № 1. -- P. 9-26.
204. Maley, C. C. Classifying the evolutionary and ecological features of neoplasms / C. C. Maley, A. Aktipis, T. A. Graham et al. // Nature Reviews Cancer. -- 2017. -- Vol. 17. -- № 10. -- P. 605-619.
205. Malumbres, M. Cell cycle, CDKs and cancer: a changing paradigm / M. Malumbres, M. Barbacid // Nature Reviews Cancer. -- 2009. -- Vol. 9. -- Cell cycle, CDKs and cancer. -- № 3. -- P. 153-166.
206. Malumbres, M. Cyclin-dependent kinases / M. Malumbres // Genome Biology. -- 2014. -- Vol. 15. -- № 6. -- P. 122.
207. Massagué, J. G1 cell-cycle control and cancer / J. Massagué // Nature. -- 2004. -- Vol. 432. -- № 7015. -- P. 298-306.
208. Mina, M. Conditional Selection of Genomic Alterations Dictates Cancer Evolution and Oncogenic Dependencies / M. Mina, F. Raynaud, D. Tavernari et al. // Cancer Cell. -- 2017. -- Vol. 32. -- № 2. -- P. 155-168.
209. Mitelman, F. Recurrent chromosome aberrations in cancer / F. Mitelman // Mutation Research/Reviews in Mutation Research. -- 2000. -- Vol. 462. -- № 2-3. -- P. 247-253.
210. Morgan, M. A. Chromatin signatures of cancer / M. A. Morgan, A. Shilatifard // Genes & Development. -- 2015. -- Vol. 29. -- № 3. -- P. 238-249.
211. Morselli, E. Anti- and pro-tumor functions of autophagy / E. Morselli, L. Galluzzi, O. Kepp et al. // Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. -- 2009. -- Vol. 1793. -- № 9. -- P. 1524-1532.
212. Muse, S. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast

genome. / S. Muse, B. Gaut // *Molecular Biology and Evolution*. -- 1994. -- Vol. 11. -- № 5. -- P. 715-724.

213. Nei, M. *Molecular Evolution and Phylogenetics* / M. Nei. – Cary : Oxford University Press, Incorporated, 2000. – 1 p.

214. Nielsen, R. Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene / R. Nielsen, Z. Yang // *Genetics*. -- 1998. -- Vol. 148. -- № 3. -- P. 929-936.

215. Nusse, R. Wnt/ $\beta$ -Catenin Signaling, Disease, and Emerging Therapeutic Modalities / R. Nusse, H. Clevers // *Cell*. -- 2017. -- Vol. 169. -- № 6. -- P. 985-999.

216. O'Byrne, K. J. Chronic immune activation and inflammation as the cause of malignancy / K. J. O'Byrne, A. G. Dalgleish // *British Journal of Cancer*. -- 2001. -- Vol. 85. -- № 4. -- P. 473-483.

217. Olivier, M. TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use / M. Olivier, M. Hollstein, P. Hainaut // *Cold Spring Harbor Perspectives in Biology*. -- 2010. -- Vol. 2. -- № 1. -- P. a001008.

218. O'Shea, J. J. JAK and STAT Signaling Molecules in Immunoregulation and Immune-Mediated Disease / J. J. O'Shea, R. Plenge // *Immunity*. -- 2012. -- Vol. 36. -- № 4. -- P. 542-550.

219. Otto, T. Cell cycle proteins as promising targets in cancer therapy / T. Otto, P. Sicinski // *Nature Reviews Cancer*. -- 2017. -- Vol. 17. -- № 2. -- P. 93-115.

220. Pál, C. Highly Expressed Genes in Yeast Evolve Slowly / C. Pál, B. Papp, L. D. Hurst // *Genetics*. -- 2001. -- Vol. 158. -- № 2. -- P. 927-931.

221. Palmieri, N. The life cycle of Drosophila orphan genes / N. Palmieri, C. Kosiol, C. Schlötterer // *eLife*. -- 2014. -- Vol. 3. -- P. e01311.

222. Pamilo, P. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. / P. Pamilo, N. O. Bianchi // *Molecular Biology and Evolution*. -- 1993. -- Vol. 10. -- № 2. -- P. 271-281.

223. Paul, C. D. Cancer cell motility: lessons from migration in confined spaces / C. D. Paul, P. Mistriotis, K. Konstantopoulos // *Nature Reviews Cancer*. -- 2017. -- Vol. 17. -- Cancer cell motility. -- № 2. -- P. 131-140.

224. Pevsner, J. *Bioinformatics and Functional Genomics* : New York Academy of Sciences Series / J. Pevsner. – 3rd ed. – Newark : John Wiley & Sons, Incorporated, 2015. – 1 p.

225. Platz, A. Human cutaneous melanoma; a review of *NRAS* and *BRAF* mutation frequencies in relation to histogenetic subclass and body site / A. Platz, S. Egyhazi, U. Ringborg, J. Hansson // *Molecular Oncology*. -- 2008. -- Vol. 1. -- № 4. -- P. 395-405.
226. Pollock, D. D. Amino acid coevolution induces an evolutionary Stokes shift / D. D. Pollock, G. Thiltgen, R. A. Goldstein // *Proceedings of the National Academy of Sciences*. -- 2012. -- Vol. 109. -- № 21. -- P. 1352-1359.
227. Pond, S. L. K. HyPhy: hypothesis testing using phylogenies / S. L. K. Pond, S. D. W. Frost, S. V. Muse // *Bioinformatics*. -- 2005. -- Vol. 21. -- HyPhy. -- № 5. -- P. 676-679.
228. Poremba, C. Telomerase as a prognostic marker in breast cancer: high-throughput tissue microarray analysis of hTERT and hTR / C. Poremba, B. Heine, R. Diallo et al. // *The Journal of Pathology*. -- 2002. -- Vol. 198. -- № 2. -- P. 181-189.
229. Prior, I. A. The Frequency of Ras Mutations in Cancer / I. A. Prior, F. E. Hood, J. L. Hartley // *Cancer Research*. -- 2020. -- Vol. 80. -- № 14. -- P. 2969-2974.
230. Prior, I. A. A Comprehensive Survey of Ras Mutations in Cancer / I. A. Prior, P. D. Lewis, C. Mattos // *Cancer Research*. -- 2012. -- Vol. 72. -- № 10. -- P. 2457-2467.
231. Rajabi, M. The Role of Angiogenesis in Cancer Treatment / M. Rajabi, S. Mousa // *Biomedicines*. -- 2017. -- Vol. 5. -- № 2. -- P. 34.
232. Rawlings, J. S. The JAK/STAT signaling pathway / J. S. Rawlings, K. M. Rosler, D. A. Harrison // *Journal of Cell Science*. -- 2004. -- Vol. 117. -- № 8. -- P. 1281-1283.
233. Ren, D. BID, BIM, and PUMA Are Essential for Activation of the BAX- and BAK-Dependent Cell Death Program / D. Ren, H.-C. Tu, H. Kim et al. // *Science*. -- 2010. -- Vol. 330. -- № 6009. -- P. 1390-1393.
234. Samatar, A. A. Targeting RAS–ERK signalling in cancer: promises and challenges / A. A. Samatar, P. I. Poulikakos // *Nature Reviews Drug Discovery*. -- 2014. -- Vol. 13. -- № 12. -- P. 928-942.
235. Samuels, Y. High Frequency of Mutations of the *PIK3CA* Gene in Human Cancers / Y. Samuels, Z. Wang, A. Bardelli et al. // *Science*. -- 2004. -- Vol. 304. -- № 5670. -- P. 554-554.
236. Sawyer, S. L. Positive selection of primate *TRIM5*  $\alpha$  identifies a critical species-specific retroviral restriction domain / S. L. Sawyer, L. I. Wu, M.

Emerman, H. S. Malik // Proceedings of the National Academy of Sciences. -- 2005. -- Vol. 102. -- № 8. -- P. 2832-2837.

237. Schrempf, A. Targeting the DNA Repair Enzyme Polymerase  $\theta$  in Cancer Therapy / A. Schrempf, J. Slysokova, J. I. Loizou // Trends in Cancer. -- 2021. -- Vol. 7. -- № 2. -- P. 98-111.

238. Schubbert, S. Hyperactive Ras in developmental disorders and cancer / S. Schubbert, K. Shannon, G. Bollag // Nature Reviews Cancer. -- 2007. -- Vol. 7. -- № 4. -- P. 295-308.

239. Sever, R. Signal Transduction in Cancer / R. Sever, J. S. Brugge // Cold Spring Harbor Perspectives in Medicine. -- 2015. -- Vol. 5. -- № 4. -- P. a006098-a006098.

240. Shiratori, R. Glycolytic suppression dramatically changes the intracellular metabolic profile of multiple cancer cell lines in a mitochondrial metabolism-dependent manner / R. Shiratori, K. Furuichi, M. Yamaguchi et al. // Scientific Reports. -- 2019. -- Vol. 9. -- № 1. -- P. 18699.

241. Sierk, M. L. Sensitivity and selectivity in protein structure comparison / M. L. Sierk, W. R. Pearson // Protein Science. -- 2004. -- Vol. 13. -- № 3. -- P. 773-785.

242. Simanshu, D. K. RAS Proteins and Their Regulators in Human Disease / D. K. Simanshu, D. V. Nissley, F. McCormick // Cell. -- 2017. -- Vol. 170. -- № 1. -- P. 17-33.

243. Singh, R. Regulation of apoptosis in health and disease: the balancing act of BCL-2 family proteins / R. Singh, A. Letai, K. Sarosiek // Nature Reviews Molecular Cell Biology. -- 2019. -- Vol. 20. -- № 3. -- P. 175-193.

244. Smith, N. G. C. Adaptive protein evolution in Drosophila / N. G. C. Smith, A. Eyre-Walker // Nature. -- 2002. -- Vol. 415. -- № 6875. -- P. 1022-1024.

245. Soding, J. Protein homology detection by HMM–HMM comparison / J. Soding // Bioinformatics. -- 2005. -- Vol. 21. -- № 7. -- P. 951-960.

246. Stoletzki, N. Estimation of the Neutrality Index / N. Stoletzki, A. Eyre-Walker // Molecular Biology and Evolution. -- 2011. -- Vol. 28. -- № 1. -- P. 63-70.

247. Stuelten, C. H. Cell motility in cancer invasion and metastasis: insights from simple model organisms / C. H. Stuelten, C. A. Parent, D. J. Montell // Nature Reviews Cancer. -- 2018. -- Vol. 18. -- № 5. -- P. 296-312.

248. Subramanian, S. Gene Expression Intensity Shapes Evolutionary Rates of the Proteins Encoded by the Vertebrate Genome / S. Subramanian, S. Kumar // *Genetics*. -- 2004. -- Vol. 168. -- № 1. -- P. 373-381.
249. Sullivan, J. Molecular Evolution.—W.-H. Li. / J. Sullivan // *Systematic Biology*. -- 1998. -- Vol. 47. -- № 1. -- P. 174-176.
250. Sun, W. Identification and evolution of the orphan genes in the domestic silkworm, *Bombyx mori* / W. Sun, X.-W. Zhao, Z. Zhang // *FEBS Letters*. -- 2015. -- Vol. 589. -- № 19PartB. -- P. 2731-2738.
251. Tan, K. T. *PRKDC* : new biomarker and drug target for checkpoint blockade immunotherapy / K. T. Tan, C.-N. Yeh, Y.-C. Chang et al. // *Journal for ImmunoTherapy of Cancer*. -- 2020. -- Vol. 8. -- № 1. -- P. e000485.
252. The Somatic Genomic Landscape of Glioblastoma / C. W. Brennan, R. G. W. Verhaak, A. McKenna et al. // *Cell*. -- 2013. -- Vol. 155. -- № 2. -- P. 462-477.
253. Thomas, F. Cancer adaptations: Atavism, de novo selection, or something in between? / F. Thomas, B. Ujvari, F. Renaud, M. Vincent // *BioEssays*. -- 2017. -- Vol. 39. -- Cancer adaptations. -- № 8. -- P. 1700039.
254. Truitt, M. L. New frontiers in translational control of the cancer genome / M. L. Truitt, D. Ruggero // *Nature Reviews Cancer*. -- 2016. -- Vol. 16. -- № 5. -- P. 288-304.
255. Vincent, M. Cancer: A de- repression of a default survival program common to all cells?: A life- history perspective on the nature of cancer / M. Vincent // *BioEssays*. -- 2012. -- Vol. 34. -- Cancer. -- № 1. -- P. 72-82.
256. Vogelstein, B. Cancer genes and the pathways they control / B. Vogelstein, K. W. Kinzler // *Nature Medicine*. -- 2004. -- Vol. 10. -- № 8. -- P. 789-799.
257. Vogelstein, B. Surfing the p53 network / B. Vogelstein, D. Lane, A. J. Levine // *Nature*. -- 2000. -- Vol. 408. -- № 6810. -- P. 307-310.
258. Wall, D. P. Functional genomic analysis of the rates of protein evolution / D. P. Wall, A. E. Hirsh, H. B. Fraser et al. // *Proceedings of the National Academy of Sciences*. -- 2005. -- Vol. 102. -- № 15. -- P. 5483-5488.
259. Weinberg, F. Mitochondrial metabolism and ROS generation are essential for Kras-mediated tumorigenicity / F. Weinberg, R. Hamanaka, W. W. Wheaton et al. // *Proceedings of the National Academy of Sciences*. -- 2010. -- Vol. 107. -- № 19. -- P. 8788-8793.
260. White, E. The role for autophagy in cancer / E. White // *Journal of Clinical Investigation*. -- 2015. -- Vol. 125. -- № 1. -- P. 42-46.

261. Wiebe, D. S. Fold-Change-Specific Enrichment Analysis (FSEA): Quantification of Transcriptional Response Magnitude for Functional Gene Groups / D. S. Wiebe, N. A. Omelyanchuk, A. M. Mukhin et al. // *Genes*. -- 2020. -- Vol. 11. -- № 4. -- P. 434.
262. Wong, R. S. Apoptosis in cancer: from pathogenesis to treatment / R. S. Wong // *Journal of Experimental & Clinical Cancer Research*. -- 2011. -- Vol. 30. -- № 1. -- P. 87.
263. Wood, R. D. DNA polymerase  $\theta$  (POLQ), double-strand break repair, and cancer / R. D. Wood, S. Doublié // *DNA Repair*. -- 2016. -- Vol. 44. -- P. 22-32.
264. Xu, K. MAD2L2, a key regulator in ovarian cancer and promoting tumor progression / K. Xu, X. Zheng, H. Shi et al. // *Scientific Reports*. -- 2024. -- Vol. 14. -- № 1. -- P. 130.
265. Yang, Z. Computational molecular evolution : Oxford series in ecology and evolution / Z. Yang. – Oxford : Oxford University Press, 2006. – 357 p.
266. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution / Z. Yang // *Molecular Biology and Evolution*. -- 1998. -- Vol. 15. -- № 5. -- P. 568-573.
267. Yang, Z. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites / Z. Yang, R. Nielsen, N. Goldman, A.-M. K. Pedersen // *Genetics*. -- 2000. -- Vol. 155. -- № 1. -- P. 431-449.
268. Yu, H. Revisiting STAT3 signalling in cancer: new and unexpected biological functions / H. Yu, H. Lee, A. Herrmann et al. // *Nature Reviews Cancer*. -- 2014. -- Vol. 14. -- № 11. -- P. 736-746.
269. Zhang, J. The International Cancer Genome Consortium Data Portal / J. Zhang, R. Bajari, D. Andric et al. // *Nature Biotechnology*. -- 2019a. -- Vol. 37. -- № 4. -- P. 367-369.
270. Zhang, J. Rates of Conservative and Radical Nonsynonymous Nucleotide Substitutions in Mammalian Nuclear Genes / J. Zhang // *Journal of Molecular Evolution*. -- 2000. -- Vol. 50. -- № 1. -- P. 56-68.
271. Zhang, J. Significant Impact of Protein Dispensability on the Instantaneous Rate of Protein Evolution / J. Zhang // *Molecular Biology and Evolution*. -- 2005. -- Vol. 22. -- № 4. -- P. 1147-1155.
272. Zhang, J. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level / J. Zhang, Z. Yang, R. Nielsen // *Molecular Biology and Evolution*. -- 2005. -- Vol. 22. -- № 12. -- P. 2472-2479.

273. Zhang, L. Phylostratigraphic analysis of gene co-expression network reveals the evolution of functional modules for ovarian cancer / L. Zhang, Y. Tan, S. Fan et al. // *Scientific Reports*. -- 2019b. -- Vol. 9. -- № 1. -- P. 2623-2623.
274. Zhang, T. Necroptosis pathways in tumorigenesis / T. Zhang, Y. Wang, H. Inuzuka, W. Wei // *Seminars in Cancer Biology*. -- 2022. -- Vol. 86. -- P. 32-40.
275. Zhang, W. Origination and evolution of orphan genes and de novo genes in the genome of *Caenorhabditis elegans* / W. Zhang, Y. Gao, M. Long, B. Shen // *Science China Life Sciences*. -- 2019c. -- Vol. 62. -- № 4. -- P. 579-593.
276. Zhao, H. Inflammation and tumor progression: signaling pathways and targeted intervention / H. Zhao, L. Wu, G. Yan et al. // *Signal Transduction and Targeted Therapy*. -- 2021. -- Vol. 6. -- № 1. -- P. 263.
277. Codon evolution: mechanisms and models / G. M. Cannarozzi, A. Schneider ред. . – Oxford : Oxford University Press, 2012. – 1 с.

## Приложение

### *Приложение А.*

Примеры запросов REST API к базе данных Orthoweb 2

- <https://orthoweb.sysbio.cytogen.ru/api-db> - в дальнейшем используется в запросах как <server\_name>
- <server\_name>/organisms - Доступ к списку имеющихся организмов в базе.
- <server\_name>/organisms/<name> - получить информацию по организму с названием <name> по количеству собранных данных.
- <server\_name>/pai/organism/<name> - получить список имеющихся значений филогенетического индекса генов (PAI) в базе для организма <name>.
- <server\_name>/ko\_pai/organism/<name> - получить список имеющихся значений филогенетического индекса генов (PAI), посчитанного с помощью КО групп, из базы для организма <name>.
- <server\_name>/go\_id/organism/<name> - получить список имеющихся идентификаторов терминов генной онтологии в базе для организма <name>.
- <server\_name>/go\_name/organism/<name> - получить список имеющихся наименований терминов генной онтологии в базе для организма <name>.
- <server\_name>/snp/organism/hsa - получить список имеющихся значений количества найденных SNP для генов человека.
- <server\_name>/pai/organism/<name>/identity/<value> - получить список имеющихся значений филогенетического индекса генов (PAI) в базе для организма с параметром идентичности равным <name>.
- <server\_name>/di/organism/<name> - получить список имеющихся значений индекса эволюционной изменчивости генов (DI) в базе для организма <name>.

- `<server_name>/di/organism/<name>/identity/<value>` - получить список имеющихся значений индекса эволюционной изменчивости генов (DI) в базе для организма `<name>` с параметром идентичности равным `<value>`.

### ***Приложение Б. Списки дифференциально экспрессирующихся генов***

В ходе выполнения данного исследования были сформированы списки генов, демонстрирующих статистически значимую дифференциальную экспрессию при сравнении опухолевых образцов разных стадий развития с соответствующими нормальными тканями (NAT) по данным RNA-seq. Анализ охватывал несколько типов злокачественных новообразований, представленных в базе данных TCGA.

Полные таблицы дифференциально экспрессирующихся генов для каждого типа опухоли и каждой стадии представлены по следующей ссылке:

<https://disk.yandex.ru/d/IZU5L1dAPRFPhQ>

Списки включают:

- **GeneID** - идентификатор гена базы данных KEGG;
- **baseMean** - среднее нормализованное значение экспрессии;
- **log2FoldChange** - логарифм отношения изменения экспрессии (по основанию 2);
- **lfcSE** - стандартная ошибка оценки log2FoldChange;
- **stat** - отношение log2FoldChange к его стандартной ошибке;
- **pvalue** - р-значение для нулевой гипотезы о том, что экспрессия гена не различается между сравниваемыми условиями.;
- **padj** - скорректированное значение р с использованием метода Бенджамини-Хохберга (FDR);
- **PAI** - филостратиграфический индекс возраста гена (Phylostratigraphic Age Index);
- **DI** - индекс дивергенции гена (Divergence Index).

**Приложение В.****Таблица В1.** Термины Генной Онтологии, использованные в анализе

| <b>GO id</b> | <b>Название термина</b>                                               |
|--------------|-----------------------------------------------------------------------|
| GO:0051726   | Регуляция клеточного цикла                                            |
| GO:0000082   | Переход G1/S                                                          |
| GO:0000086   | Переход G2/M                                                          |
| GO:0007088   | Регуляция митоза                                                      |
| GO:0042127   | Регуляция клеточной пролиферации                                      |
| GO:0071774   | Контрольная точка клеточного цикла в фазе G1                          |
| GO:0031577   | Контрольная точка клеточного цикла в фазе G2                          |
| GO:0006260   | ДНК-репликация                                                        |
| GO:0051276   | Хромосомная организация                                               |
| GO:0032200   | Организация теломер                                                   |
| GO:0000723   | Поддержание теломер                                                   |
| GO:0030154   | Дифференциация клеток                                                 |
| GO:0006399   | Метаболизм тРНК                                                       |
| GO:0006397   | Процессинг мРНК                                                       |
| GO:0006259   | Метаболизм ДНК                                                        |
| GO:0006412   | Трансляция                                                            |
| GO:0006915   | Апоптоз                                                               |
| GO:0012501   | Программированная клеточная гибель                                    |
| GO:0006914   | Автофагия                                                             |
| GO:0043066   | Негативная регуляция апоптоза                                         |
| GO:0043065   | Положительная регуляция апоптоза                                      |
| GO:0006351   | Транскрипция, ДНК-зависимая                                           |
| GO:0016055   | Wnt-сигнальный путь                                                   |
| GO:0000165   | МАРК-сигнальный путь                                                  |
| GO:0014065   | PI3K/AKT-сигнальный путь                                              |
| GO:0007179   | TGF-бета-сигнальный путь                                              |
| GO:0010629   | Негативная регуляция транскрипции                                     |
| GO:0040029   | Эпигенетическая регуляция                                             |
| GO:0006281   | Репарация ДНК                                                         |
| GO:0006302   | Репарация двуцепочечных разрывов                                      |
| GO:0006303   | Ремонт двуцепочечных разрывов путём нехомологичного соединения концов |
| GO:0006974   | Клеточный отчет на сигналы повреждения ДНК                            |
| GO:0006096   | Гликолиз                                                              |
| GO:0006099   | Цикл трикарбоновых кислот                                             |
| GO:0006635   | $\beta$ -окисление жирных кислот                                      |
| GO:0009058   | Процессы биосинтеза                                                   |
| GO:0008203   | Метаболизм холистерола                                                |
| GO:0016477   | Миграция клеток                                                       |
| GO:0060326   | Хемотаксис                                                            |
| GO:2000145   | Регуляция клеточного движения                                         |

|                   |                                                |
|-------------------|------------------------------------------------|
| <b>GO:0001525</b> | Ангиогенез                                     |
| <b>GO:0001568</b> | Развитие сосудистой системы                    |
| <b>GO:0001935</b> | Рост эндотелиальных клеток                     |
| <b>GO:0048870</b> | Подвижность клеток                             |
| <b>GO:0007155</b> | Клеточная адгезия                              |
| <b>GO:0007160</b> | Взаимодействие клеток с внеклеточным матриксом |
| <b>GO:0006954</b> | Воспалительный ответ                           |
| <b>GO:0007267</b> | Межклеточные сигнальные пути                   |
| <b>GO:0019882</b> | Экспрессия антигенов                           |
| <b>GO:0002250</b> | Адаптивный иммунный ответ                      |
| <b>GO:0006955</b> | Иммунный ответ                                 |

**Таблица В2.** Значения TAI стадий развития онкологических заболеваний.  
Столбец NAT – Normal Adjacent Tissues

| <b>TAI</b>  | <b>NAT</b> | <b>I Стадия</b> | <b>II стадия</b> | <b>III стадия</b> | <b>IV стадия</b> |
|-------------|------------|-----------------|------------------|-------------------|------------------|
| <b>UCEC</b> | 4.607078   | 4.463704        | 4.467779         | 4.445552          | 4.455624         |
| <b>BLCA</b> | 4.981481   | 4.752902        | 4.492966         | 4.505613          | 4.534112         |
| <b>BRCA</b> | 4.785891   | 4.668032        | 4.637815         | 4.665382          | 4.677226         |
| <b>COAD</b> | 4.850936   | 4.439789        | 4.45038          | 4.481497          | 4.456311         |
| <b>KIRC</b> | 4.798397   | 4.853467        | 4.862405         | 4.863008          | 4.825348         |
| <b>LIHC</b> | 5.99863    | 5.403328        | 5.32169          | 5.192634          | 5.459084         |
| <b>LUAD</b> | 5.310213   | 4.784176        | 4.733831         | 4.706017          | 4.703828         |
| <b>PRAD</b> | 4.742203   | 4.712207        | 4.675865         | 4.675532          | 4.662134         |
| <b>THCA</b> | 4.972783   | 4.843374        | 4.828472         | 4.828682          | 4.831671         |

**Таблица В3.** Значения TDI стадий развития онкологических заболеваний.  
Столбец NAT – Normal Adjacent Tissues

| <b>TDI</b>  | <b>NAT</b> | <b>I Стадия</b> | <b>II стадия</b> | <b>III стадия</b> | <b>IV стадия</b> |
|-------------|------------|-----------------|------------------|-------------------|------------------|
| <b>UCEC</b> | 3.967398   | 3.986911        | 3.998328         | 3.971333          | 4.001767         |
| <b>BLCA</b> | 3.984667   | 4.257497        | 4.048051         | 4.004898          | 4.011184         |
| <b>BRCA</b> | 4.190287   | 4.078578        | 4.069145         | 4.073497          | 4.095158         |
| <b>COAD</b> | 4.495807   | 4.066575        | 4.071877         | 4.074385          | 4.065866         |
| <b>KIRC</b> | 4.275209   | 4.247558        | 4.242292         | 4.243964          | 4.223329         |
| <b>LIHC</b> | 5.571663   | 5.097211        | 5.24053          | 4.864015          | 5.187719         |
| <b>LUAD</b> | 4.734863   | 4.254778        | 4.200492         | 4.170093          | 4.194738         |
| <b>PRAD</b> | 4.499775   | 4.274551        | 4.243194         | 4.250399          | 4.343177         |
| <b>THCA</b> | 4.520441   | 4.244759        | 4.301416         | 4.204916          | 4.178473         |