

УДК 577.21:612.67

АНАЛИЗ ПРЕДСТАВЛЕННОСТИ ГЕНОВ, СВЯЗАННЫХ СО СТАРЕНИЕМ, В ПУБЛИЧНЫХ ГЕНОМНЫХ БАЗАХ: ОБЗОР ИНСТРУМЕНТОВ И ПЕРВЫЕ РЕЗУЛЬТАТЫ

Аникина И.А.

аспирант кафедры вирусологии и молекулярной биологии Московского государственного университета им. М.В. Ломоносова

e-mail: anikina.ia@gmail.com

Аннотация

Рассматриваются особенности поиска и сопоставления генов, ассоциированных со старением, в специализированных и универсальных публичных ресурсах. Описаны структура базы GenAge, возможности NCBI Gene, Ensembl и UniProt, а также правила унификации идентификаторов и контроля происхождения аннотаций. Выполнена ограниченная документальная проверка пилотной выборки из восьми генов человека: APOE, FOXO3, SIRT1, SIRT6, MTOR, TP53, WRN и SOD2. Для каждого показана принадлежность к перечню человеческих генов GenAge и предложена схема получения геномных и белковых аннотаций из первичных ресурсов. Обосновано, почему наличие записи в базе не равно доказательству причинного влияния на продолжительность жизни. Показаны потенциальные ошибки при объединении разных видов, изоформ и критериев доказательности. Результаты представляют собой воспроизводимый пилотный аудит доступного списка GenAge, а не завершённую автоматизированную выгрузку записей всех перечисленных баз.

Ключевые слова: старение, гены долголетия, GenAge, NCBI Gene, Ensembl, UniProt, биоинформатика, аннотация генов, пилотный анализ.

REPRESENTATION OF AGEING-RELATED GENES IN PUBLIC GENOMIC DATABASES: TOOL REVIEW AND INITIAL RESULTS

Anikina I.A.

Postgraduate student at the Department of Virology and Molecular Biology, Lomonosov Moscow State University

e-mail: anikina.ia@gmail.com

Abstract

Public genomic resources are compared as sources of evidence on ageing-related genes. GenAge is considered alongside general-purpose gene and protein annotation resources, including NCBI Gene, Ensembl and UniProt. A documented pilot sample of eight human gene symbols—APOE, FOXO3, SIRT1, SIRT6, MTOR, TP53, WRN and SOD2—is checked against the publicly available GenAge human gene catalogue. A reproducible identifier-mapping workflow and evidence-quality criteria are proposed. The distinction between database presence, functional annotation, association with human longevity and causal effects on ageing is emphasized. The results are limited to a pilot verification of GenAge catalogue entries; no complete cross-database API extraction or experimental validation is claimed.

Keywords: ageing, longevity, GenAge, NCBI Gene, Ensembl, UniProt, gene annotation, bioinformatics, pilot study.

Введение

Биологическое старение сопровождается изменениями регуляции генома, поддержания протеостаза, обмена веществ, клеточной сенесценции и репарации ДНК. Однако не существует единого класса «генов старения», который одинаково определялся бы во всех исследованиях. В одних публикациях так называют гены, изменение активности которых продлевает жизнь модельному организму, в других — локусы, статистически связанные с

долголетием человека, а в-третьих – гены, транскрипционная активность которых меняется с возрастом. Смещение этих смыслов существенно затрудняет сопоставление результатов [1-3].

Расширение публичных геномных ресурсов делает возможной проверку списка кандидатов без собственных лабораторных экспериментов. Вместе с тем расширение объема аннотаций не гарантирует увеличения числа независимо подтвержденных причинных связей. Универсальные базы хорошо описывают локализацию и последовательность генов, но не обязаны определять их роль в старении. Специализированные ресурсы содержат экспертный отбор, однако зависят от принятых критериев и даты обновления [1, 2].

Цель работы – разработать воспроизводимую процедуру оценки представленности генов старения в открытых ресурсах и проверить ее применимость на небольшой пилотной выборке. Задачи включают сравнительный обзор инструментов, описание критериев отбора генов и правил сопоставления идентификаторов, фиксацию первичных наблюдений и выявление ограничений, значимых для последующего масштабирования анализа.

Методы и методики

GenAge входит в комплекс Human Ageing Genomic Resources (HAGR) и содержит отдельно курируемые списки генов человека и модельных организмов. Раздел человеческих генов объединяет как непосредственно связанные со старением человека объекты, так и кандидаты, отобранные по экспериментальным данным модельных систем. В отчете HAGR описаны эволюция базы и принципы ее обновления [1, 2]. На открытой странице статистики GenAge для версии от 16 декабря 2024 г. указано 307 человеческих генов [9]. Это численность кураторского каталога на указанную дату, а не число экспериментально доказанных «генов долголетия».

NCBI Gene представляет собой универсальный справочник генетических локусов, их номенклатуры, описаний, связей с последовательностями и научными записями. Программный доступ предоставляется через E-utilities; на уровне отдельного гена доступные разделы и объем информации зависят от вида и статуса аннотации [4]. Такой ресурс удобен для проверки официального символа и идентификатора, но само присутствие записи NCBI Gene ничего не говорит о специфической связи гена со старением.

Ensembl обеспечивает устойчивые идентификаторы генов, транскриптов и продуктов трансляции, аннотацию геномных координат и межвидовые связи. Документированный интерфейс lookup/symbol позволяет получать сущность по символу и виду, а пакетная версия операции – обрабатывать группы символов [5]. Для воспроизводимости следует хранить версию сборки генома, релиз аннотации и статус сопоставления: один символ не обязательно соответствует единственной транскрипционной форме.

UniProtKB ориентирован на белковые последовательности, функциональную аннотацию, изоформы, каталитические особенности и достоверность описаний. Сведения UniProtKB/Swiss-Prot и автоматизированной части базы имеют разный уровень кураторской проверки; при сопоставлении важно сохранять идентификатор белковой записи и отличать ген от кодируемых белковых продуктов [6]. Геномные и протеомные базы поэтому целесообразно использовать как взаимодополняющие, а не конкурирующие источники.

Результаты

Для пилотного аудита выбраны восемь хорошо известных генов человека, представляющих различные механизмистические направления: метаболизм липидов (APOE), регуляцию транскрипции и ответа на стресс (FOXO3), NAD-зависимую регуляцию (SIRT1 и SIRT6), клеточную сигнализацию и рост (MTOR), контроль повреждений и клеточного цикла (TP53), поддержание генома (WRN) и антиоксидантную защиту (SOD2). Это целевой демонстрационный набор, а не случайная репрезентативная выборка всех генов старения.

В качестве начальной точки используется список Human Genes на официальном сайте GenAge, обновленный 16 декабря 2024 г. Для каждого символа проверяется наличие отдельной строки в этом каталоге [8]. Результат кодируется бинарной переменной «найден / не найден», при этом фиксируются вид *Homo sapiens* и наименование используемого каталога.

Поиск в произвольных веб-страницах без ограничения конкретным разделом базы не считается равнозначной проверкой включения в GenAge.

Следующий этап протокола предполагает сопоставление с NCBI Gene и Ensembl по официальному символу и виду, после чего в отдельной таблице должны сохраняться постоянный Gene ID, Ensembl Gene ID, хромосома, сборка генома, дата загрузки и ссылки на исходные записи. Белковые продукты связываются с UniProt по явным перекрестным идентификаторам. Автоматически приписывать гену идентификатор белка только по сходству строк недопустимо из-за синонимов и изоформ.

Критерии контроля: исключение устаревших символов, отдельный учет гена и транскриптов, подтверждение вида, запрет интерпретировать непройденный API-запрос как биологическое отсутствие, а также сохранение оригинальной строки ответа и даты обращения. Для итоговой сравнительной матрицы можно использовать статусы «подтверждено», «не найдено в указанной версии» и «не проверено». Последняя категория предотвращает подмену отсутствующих наблюдений отрицательным результатом.

Непосредственный просмотр опубликованного списка человеческих генов GenAge позволяет подтвердить наличие всех восьми заранее выбранных символов [8]. Список содержит для каждого гена внутренний идентификатор HAGRID и столбцы, описывающие количество связанных публикаций и взаимодействий. Эти величины характеризуют наполнение базы, а не напрямую силу биологического эффекта. Пилотный расчет охвата применим только к данной целевой выборке и не может быть распространен на неотобранные гены.

Выборка неоднородна по типу доказательств. Например, связь WRN с преждевременным старением опирается на изучение синдрома Вернера; для FOXO3 важны генетические ассоциации долголетия и регуляция клеточных процессов; для TP53 обсуждаются компромиссы между противоопухолевой защитой и старением тканей. При агрегировании таких записей оценка «1» означает лишь включение в GenAge, но не одинаковую достоверность всех соответствующих биологических утверждений.

Поскольку NCBI Gene, Ensembl и UniProt являются базами разных типов, бессодержательно сравнивать доли найденных генов без определения того, что именно считается подтвержденным совпадением (табл. 1). Настоящий аудит количественно проверяет только каталог GenAge.

Таблица 1 – Анализ генов

Ген	HAGRID GenAge	Преобладающий контекст
APOE	138	Липидный обмен, возрастные риски
FOXO3	123	Транскрипционный стресс-ответ
SIRT1	150	NAD-зависимая регуляция
SIRT6	239	Хроматин, метаболизм, репарация
MTOR	221	Питание и клеточный рост
TP53	6	Контроль повреждений и клеточного цикла
WRN	13	Поддержание стабильности генома
SOD2	129	Антиоксидантная защита

Все приведенные идентификаторы соответствуют строкам официального списка GenAge [8]. Функциональный контекст указан для ориентации и не является самостоятельным ранжированием степени влияния генов на продолжительность жизни.

$$C = \frac{N_{found}}{N_{selected}} \times 100 \%, \quad (1)$$

Для выбранных восьми записей $C = 100 \%$. Этот показатель корректен лишь как факт охвата целевой пилотной выборки; он не описывает полноту GenAge по отношению ко всем генам, вовлеченным в старение.

При дальнейшем масштабировании целесообразно разделить этапы идентификации, объединения и биологической интерпретации. Первичный файл должен содержать исходный

символ, вид, источник включения, проверенную версию и обоснование связи со старением. Затем через официальные API формируется таблица нормализованных идентификаторов. Сопоставления «один ко многим» сохраняются как отдельные записи, а не принудительно сводятся к одному результату.

Качество уместно описывать тремя независимыми величинами: покрытием кандидатного списка, полнотой обязательных аннотационных полей и согласованностью идентификаторов между источниками. Эти метрики отвечают на разные вопросы. Высокое покрытие само по себе не подтверждает высокой биологической доказательности, тогда как полнота аннотации нередко выше у хорошо изученных генов – вследствие исследовательского внимания, а не большей роли в старении.

Если доступны временные снимки, необходимо проверять устойчивость результатов к дате обновления: отдельные записи со временем переименовываются, уточняются или исключаются. Для специализированного каталога существенны и критерии отбора. Доказательство причинного влияния на срок жизни животного, воспроизводимая связь с возрастным фенотипом, популяционная ассоциация и транскрипционная корреляция – это разные основания, поэтому их не следует сводить в одну неразличимую категорию.

Воспроизводимый процесс должен включать словарь полей, протокол обработки пропусков и перечень исключенных либо неоднозначно сопоставленных записей. Для публичных данных особенно значимо архивирование параметров обращения; спустя несколько месяцев один и тот же запрос может вернуть уже иную аннотацию. По этой причине итоговую выгрузку следует сопровождать файлом с датами и версиями - а не только перечнем символов.

$$A = \frac{N_{complete}}{N_{checked}} \times 100 \%, \quad (2)$$

Здесь A обозначает долю проверенных записей, в которых заполнены все заранее объявленные обязательные поля; этот показатель целесообразно вычислять только после фактической межбазовой выгрузки. Такая последовательность принципиальна для интерпретации результата, поскольку до реального сопоставления источников невозможно корректно установить, какие именно записи подлежат проверке и каков состав обязательных полей - уже в прикладной процедуре оценки.

$$J = |G_1 \cap G_2| / |G_1 \cup G_2|, \quad (3)$$

Коэффициент Жаккара позволяет сопоставлять два предварительно нормализованных списка при одинаковом виде и определении единицы учета. Но применимость этой меры зависит от границ сопоставления: без общей области сравнения данное число будет малоинформативно, поскольку значение коэффициента – даже при формальной вычислимости – не отражает содержательно интерпретируемую степень совпадения.

Таблица 2 – Анализ ошибок

Ошибка	Последствие	Предотвращение
Синоним вместо действующего символа	Ложное несовпадение	Проверка HGNC и постоянных ID
Смещение видов	Ошибочная переносимость вывода	Фильтр таксона
Смещение гена и белка	Завышенный счет объектов	Отдельные уровни ID
Разные версии баз	Неустойчивость итога	Фиксация релиза и даты

Перечисленные риски необходимо документировать еще до подсчета согласованности; в противном случае количественные метрики будут фиксировать особенности склейки таблиц, а не биологические различия. Иначе говоря, сама процедура объединения источников - при неописанных допущениях и неснятых неоднозначностях - начнет определять итоговую оценку сильнее, чем сравниваемое биологическое содержание записей.

Обсуждение

Результат 8 из 8 в GenAge обусловлен целенаправленным выбором общеизвестных генов, уже связанных со старением; поэтому он не может рассматриваться как независимая оценка чувствительности базы. Для оценки полноты ресурса требуется внешний эталон, сформированный независимо от GenAge, а также объективные правила отбора положительных и отрицательных примеров. Нельзя, кроме того, трактовать норму включения в несколько универсальных баз как внешнюю валидацию биологической гипотезы: подобные базы аннотируют почти все хорошо изученные человеческие гены – именно по этой причине совпадение по ним не удостоверяет специфическую связь со старением.

Существует принципиальное различие между ассоциацией с продолжительностью жизни, изменением экспрессии с возрастом, клеточной сенесценцией и вмешательствами, увеличивающими срок жизни модельного организма. Следует учитывать видовую специфичность, ткань, возраст, пол, генетический фон и контекст воздействия. Механизм, подтвержденный в дрожжах или нематодах, требует отдельной оценки применимости к человеку, даже если обнаружен гомолог соответствующего гена.

Еще одна причина различий между базами – публикационное смещение. Хорошо изученные TP53 или SIRT1 обладают богатым набором публикаций и аннотаций, тогда как менее известный регулятор может отсутствовать в экспертной подборке из-за недостаточности доступных исследований. Поэтому число ссылок или взаимодействий в каталоге следует трактовать как дескриптивный показатель представленности литературы, а не как меру силы эффекта.

Ограничением настоящей работы служит небольшой целевой характер выборки и отсутствие сохраненной массовой выгрузки текущих API-ответов NCBI Gene, Ensembl и UniProt. Пилотные результаты нельзя экстраполировать на весь человеческий геном. Однако именно открытая фиксация этих границ делает исследование проверяемым и позволяет последовательно расширить его до полноценной сравнительной оценки ресурсов.

Заключение

Сравнение публичных геномных ресурсов показывает следующее: специализированный GenAge решает задачу экспертного отбора кандидатов, тогда как NCBI Gene, Ensembl и UniProt обеспечивают уровни стандартизированной геномной и белковой аннотации. Их совместное применение требует явного сопоставления устойчивых идентификаторов, видов и версий. Кроме того, для корректной интерпретации необходимо сохранять происхождение каждого вывода – иначе итог сопоставления будет зависеть не только от содержания ресурсов, но и от того, как именно были объединены разнородные записи.

Пилотный документальный аудит подтвердил присутствие восьми заранее отобранных генов в открытом перечне GenAge. Полученная доля отражает только покрытие выбранной демонстрационной группы; как представляется, она не может рассматриваться ни как оценка чувствительности базы, ни как доказательство влияния каждого гена на человеческое долголетие. В этой связи предложена процедура расширения исследования - на независимую выборку с последующей межбазовой сверкой, которая позволит отделить демонстрационный результат от более строгой сравнительной оценки.

Практическая ценность состоит в разделении собственно найденных записей, функциональных аннотаций и экспериментальных доказательств. Такая схема полезна для интерпретации данных. Она помогает избегать ложных количественных выводов и, при этом, создает основу для последующего исследования генетических механизмов старения на масштабируемых открытых данных – с отдельным учетом наличия записи, характера аннотации и степени экспериментального подтверждения.

Литература

1. Magalhães J.P., Abidi Z., Arantes Dos Santos G. et al. Human Ageing Genomic Resources: updates on key databases in ageing research // Nucleic Acids Research. 2024. Vol. 52, № D1. PP. 900-908

2. Tacutu R., Thornton D., Johnson E. et al. Human Ageing Genomic Resources: new and updated databases // *Nucleic Acids Research*. 2018. Vol. 46, № D1. PP. D108-1090.
3. Magalhães J.P., Budovsky A., Lehmann G. et al. The Human Ageing Genomic Resources: online databases and tools for biogerontologists // *Aging Cell*. 2009. Vol. 8, № 1. PP. 65-72.
4. Sayers E.W. et al. Database resources of the National Center for Biotechnology Information in 2025 // *Nucleic Acids Research*. 2025. Vol. 53, № D1. PP. 20-29.
5. Ensembl REST API. Lookup by symbol: документация. URL: https://rest.ensembl.org/documentation/info/symbol_lookup (дата обращения: 21.10.2026).
6. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025 // *Nucleic Acids Research*. 2025. Vol. 53, № D1. PP. 609-617.
7. Magalhães J.P., Curado J., Church G.M. Meta-analysis of age-related gene expression profiles identifies common signatures of aging // *Bioinformatics*. 2009. Vol. 25, № 7. PP. 875-881.
8. GenAge Human Genes: List of Entries // Human Ageing Genomic Resources. URL: <https://genomics.senescence.info/genes/allgenes.php> (дата обращения: 21.10.2026).
9. GenAge Database Statistics // Human Ageing Genomic Resources. URL: <https://www.genomics.senescence.info/genes/stats.php> (дата обращения: 21.10.2026).
10. Ensembl REST API. Batch lookup by symbol: документация. URL: https://rest.ensembl.org/documentation/info/symbol_post (дата обращения: 21.10.2026).