1. Аннотация

Координированная экспрессия генов — один из важнейших механизмов, обеспечивающих стабильную работу биологических систем. В нормальных и стрессовых условиях клетки многие гены должны активироваться или подавляться синхронно, чтобы обеспечивать баланс белкового синтеза, снижать вариабельность экспрессии и формировать скоординированные адаптивные ответы. Это явление играет ключевую роль в развитии, поддержании гомеостаза и реакции на внешние стимулы — и опирается как на регуляторные взаимодействия (например, общие энхансеры и факторы транскрипции), так и на пространственную организацию хроматина, включая разнообразные формы доменных структур.

У эукариот пространственная организация генома приобретает особое значение: она позволяет изолировать генетически активные участки друг от друга и формировать модульные, полунезависимые единицы регуляции. Тем не менее, несмотря на признание значимости доменов и связанных с ними процессов, на практике выявление таких структур остаётся технически трудоёмкой задачей, требующей экспериментов вроде Hi-C — ресурсоёмких, дорогих и ограниченных в применении.

При этом существует альтернативный путь: обратить внимание не на структуру, а на её функциональные проявления. Предполагается, что пространственная изоляция может отражаться на уровне экспрессии — в виде локальной координации между генами. То есть можно попытаться оценить не структуру напрямую, а её след — согласованную регуляцию.

Целью данной работы была разработка и тестирование универсального, лёгкого в применении и статистически обоснованного метода выявления участков с координированной экспрессией генов — как возможных индикаторов доменных структур или других функциональных модулей регуляции.

Для достижения поставленной цели были решены следующие задачи:

- Разработка статистического теста, оценивающего уровень координации экспрессии в пределах заданной хромосомной области.
- Создание системы контроля качества метода: положительный и отрицательный контроль, проверка чувствительности и специфичности, моделирование шумов.
- Оптимизация вычислений, позволяющая масштабировать метод под реальные объёмы данных (до целых хромосом и всех образцов).
- Адаптация метода под разные варианты интерпретации “попадания” гена в домен (по старту, концу, центру, пропорционально длине).
- Построение алгоритма решения обратной задачи — локализации возможных доменов по картам экспрессии без использования структурных данных.
- Формирование базы под создание публичного онлайн-инструмента, в котором любой исследователь сможет загрузить свои данные и получить результат.

2. Обзор литературы

2.1 Роль координированной экспрессии в геномной регуляции.

В живых организмах большинство клеточных процессов требует участия не одного, а сразу многих белков, РНК и других молекул, закодированных в разных участках генома. Поэтому возникает необходимость в согласованной — координированной — экспрессии множества генов, обеспечивающей функционирование сложных биологических систем. Такая потребность обусловлена как функциональной взаимосвязью белков, так и необходимостью синхронного реагирования на изменения окружающей среды.

Координированно экспрессией генов принято называть слаженное включение и выключение их множеств в ответ на внутренние или внешние сигналы. (https://doi.org/10.1016/j.tig.2012.07.003) Она играет ключевую роль в регуляции клеточной активности, позволяя организму быстро адаптироваться к внешним условиям, запускать программы дифференцировки, иммунного ответа, метаболической перестройки и других фундаментальных процессов. Поддержание такой скоординированной работы требует интеграции регуляторных сигналов на нескольких уровнях.

Эпигенетический уровень представляет собой первый слой контроля. Химические модификации ДНК (например, метилирование) и гистонов (ацетилирование, метилирование и др.) влияют на доступность участков хроматина для транскрипционной машинерии. (https://doi.org/10.1016/B978-0-12-800140-0.00001-7) Эти метки формируют "эпигенетический ландшафт", способный избирательно активировать или подавлять транскрипцию групп функционально связанных генов. Показано, что благодаря своей

устойчивости и способности наследоваться при митозе, эпигенетические модификации обеспечивают долговременную координацию экспрессии (https://doi.org/10.1016/B978-0-12-800140-0.00001-7).

Пространственная организация хроматина считается следующим уровнем регуляции (https://doi.org/10.1016/j.jmb.2014.09.013). Геном структурирован в участки, внутри которых регуляторные элементы (например, энхансеры) и соответствующие им гены физически сближаются в ядре, несмотря на значительное линейное расстояние на ДНК. Такая трёхмерная архитектура генома позволяет эффективно наводить регуляторную активность и предотвращает "перекрёстные" воздействия на гены из соседних доменов. Именно этот уровень — структурный, хроматиновый — и является фокусом изучения в данной работе.

Транскрипционный уровень включает регуляцию за счёт транскрипционных факторов — белков, распознающих специфические ДНК-последовательности и модулирующих активность промоторов. Один транскрипционный фактор может контролировать десятки или сотни генов, активируя целые регуляторные сети, как например демонстрируется на примере (https://doi.org/10.1105/tpc.108.065250). Особую роль здесь играют суперэнхансеры — кластеры регуляторных элементов, отвечающие за экспрессию ключевых генов, определяющих клеточную идентичность (https://www.cell.com/trends/cancer/abstract/S2405-8033(17)30061-4?sf183699197=1&dgcid=twitter_social_trecan-backlist-7).

Посттранскрипционный уровень охватывает механизмы контроля стабильности мРНК, их локализации и трансляции. Основными регуляторами здесь выступают микроРНК и РНК-связывающие белки (RBPs). Благодаря способности одной микроРНК

одновременно подавлять множество мРНК, объединённых общими мотивами, достигается тонкая настройка и координация экспрессии на уровне посттранскриптов. Множество работ посвящено в том числе тому, чтобы установить связи между работой RBPs и итоговым состоянием клетки (https://www.cell.com/trends/cancer/abstract/S2405-8033(17)30089-4?sf175071588=1, https://doi.org/10.3390/cancers12092699)

Посттрансляционные модификации представляют собой дополнительный механизм регуляции активности уже синтезированных белков. Такие модификации, как фосфорилирование, ацетилирование и убиквитинирование, позволяют клетке быстро переключать функции белковых комплексов в ответ на сигналы, без необходимости повторного синтеза компонентов (https://doi.org/10.1371/journal.pcbi.1002933).

Сигнальные пути интегрируют различные уровни регуляции, передавая информацию от рецепторов на поверхности клетки к ядру. Классические каскады, такие как MAPK, Wnt, Notch, PI3K-Akt и NF-κB, координируют экспрессию обширных генных программ за счёт активации транскрипционных факторов и других элементов регуляторной сети. С одной стороны этот уровень регуляции уже находится существенно выше изучаемого в данной работе, с другой – благодаря достижениям протеомики именно он изучен черезвычайно подробно(https://doi.org/10.3390/ijms21124507, DOI: 10.1146/annurev.physiol.67.040103.152647).

В процессе эволюции у разных организмов сложились различные стратегии организации генома, направленные на эффективную координацию экспрессии. У прокариот таким решением стали опероны — группы функционально связанных генов, транскрибируемых в составе одной мРНК с общего промотора. Это обеспечивает быстрый

и слаженный синтез необходимых белков, однако ограничивает гибкость регуляции (https://doi.org/10.1016/j.mib.2008.02.005).

У эукариот, чей геном значительно более компактен в пространстве, но растянут линейно, эволюция привела к другим стратегиям. Вместо кластеров генов здесь используется совместное использование регуляторных элементов, в частности энхансеров, которые могут активировать несколько генов одновременно, независимо от их положения на ДНК. Такое взаимодействие обеспечивается трёхмерной организацией хроматина, позволяющей сближать гены и регуляторные участки в ядре (https://link.springer.com/article/10.1007/s00412-015-0538-5). Ключевую роль играют домены — области, внутри которых взаимодействия между регуляторами и генами происходят чаще, чем с соседними участками. Эти обособленные регионы формируют архитектурный каркас координированной экспрессии и служат важным звеном в сложной сети регуляции. (https://doi.org/10.1016/j.gde.2020.02.015)

2.2Доменные структуры в геноме эукариотических клеток

Терминологическое замечание. В настоящем обзоре под доменными структурами понимаются пространственно и функционально организованные участки хроматина в ядре клетки, такие как топологически ассоциированные домены (TADs), ламина-ассоциированные домены (LADs), компартменты и другие элементы трёхмерной архитектуры генома. Следует отличать эти структуры от белковых доменов — консервативных участков в последовательности белков, обладающих специфическими функциями и структурой. Несмотря на совпадение терминов, речь идёт о принципиально

различных уровнях организации: геномном и протеиновом соответственно. Далее кратко опишем классификацию, функции и свойства основных видов доменных структур.

Первые исследования доменных структур датируются 1970-ми годами (https://symposium.cshlp.org/content/42/109.short), и несмотря на большое количество предложенных с тех пор моделей их устройства (https://www.cell.com/molecular-cell/fulltext/S1097-2765(16)30181-2#), некоторые вопросы, в частности механизм формирования доменов, их наследования и эволюции остаются открытыми (https://www.cell.com/cell/fulltext/S0092-8674(16)30073-3). Тем не менее, остаётся возможным классифицировать доменные структуры по их “механическому устройству”.

Топологически ассоциированные домены (TADs) — это структурные единицы пространственной организации эукариотического генома, представляющие собой области ДНК, внутри которых участки хроматина взаимодействуют друг с другом значительно чаще, чем с участками за пределами этого домена. Иными словами, TAD — это замкнутая область, внутри которой происходят интенсивные физические контакты между генами и их регуляторными элементами, а границы домена ограничивают распространение таких взаимодействий.

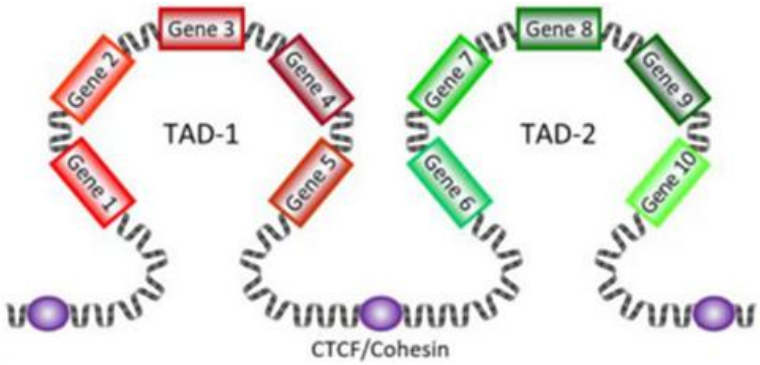

Рис.N: Иллюстрация связи между пространственной укладкой генов в TAD и точками связывания коэзина и CTCF.

Открытие TAD-доменов стало возможным благодаря развитию методов хроматин-конформационного анализа, прежде всего Hi-C — технологии, позволяющей выявлять пары участков ДНК, находящихся в тесной пространственной близости внутри ядра. Впервые TADs были описаны в 2012 году в работе Dixon et al., где авторы продемонстрировали, что геном млекопитающих разделён на такие домены размером от сотен тысяч пар оснований до нескольких миллионов пар оснований (https://www.nature.com/articles/nature11082). Границы TAD-доменов, как правило, обогащены связыванием специфических белков, прежде всего CTCF (CCCTC-binding factor) и комплекса коэзина. Эти белки участвуют в формировании так называемых хроматиновых петель, которые и замыкают TAD в петлевидную структуру. Нарушение границ TAD, например, в результате мутаций или структурных перестроек генома, может приводить к неправильной активации генов — феномену, известному как "enhancer hijacking" — и ассоциирован с различными патологиями, включая онкологические заболевания (https://pubmed.ncbi.nlm.nih.gov/25959774/).

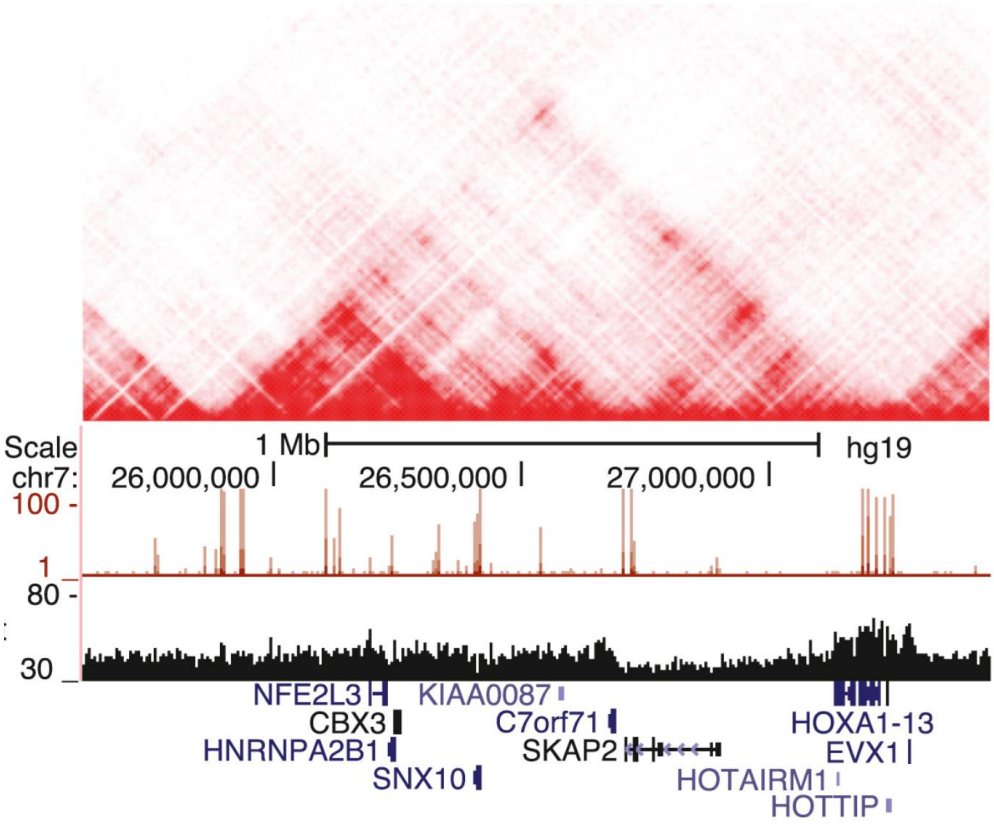

Рис.N. Тепловая карта взаимодействий топологически ассоциированного домена, расположенного вблизи локуса HOXA в лимфобластоидных клетках GM12878 от Rao et al. (2014) при разрешении 10 кб. Ниже также показаны треки Chip-seq из Encode. Заметим также, что внутри одного TAD существует несколько точек с локально обогащённым сигналом, что указывает на существование субпетель внутри одного домена.

Ламино-ассоциированные домены (LADs) называют участки генома, находящиеся в тесной пространственной связи с ядерной ламиной — сетчатой структурой из белков ламинов, выстилающей внутреннюю сторону ядерной оболочки. Эти домены представляют собой обширные регионы, размером от сотен п.о. до нескольких миллионов п.о., которые имеют тенденцию к репрессированному, транскрипционно неактивному состоянию. LADs были впервые систематически охарактеризованы с использованием технологии DamID —

метода, позволяющего определить участки ДНК, находящиеся в непосредственной близости к определённым белкам, в том числе к ламинам (Pickersgill et al., Genome Research, 2006; Guelen et al., Nature, 2008). Анализ показал, что LAD-домены покрывают значительную часть генома — до 30–40% — и имеют характерные эпигенетические признаки, включая обогащение метками гетерохроматина (например, H3K9me2/3), низкую плотность активных промоторов и низкий уровень экспрессии генов. Функционально LADs рассматриваются как репрессивные домены, обеспечивающие пространственную изоляцию "молчаливых" генов от активных регуляторных элементов. Их привязка к ядерной ламине способствует поддержанию стабильного неактивного состояния и может играть роль в длительной транскрипционной репрессии, например, во время дифференцировки клеток.

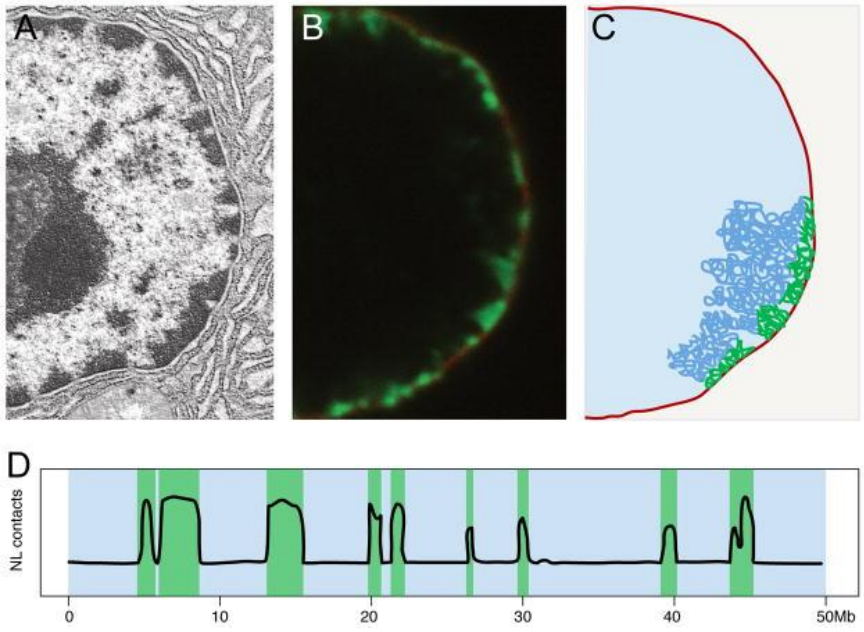

Рис.N: Иллюстрация укладки с хромосомы (синяя) с помощью гетерохроматина (зелёный) на поверхности ламины (красный). A) Электронная микрофотография, B) Конфокальная микроскопия с окрашиванием гетерохроматина, C) Схематичное изображение, D) Оценочный диапазон размеров и расстояний между LAD.

Геномы разных типов клеток имеют как конститутивные LADs (cLADs) — консервативные и стабильные по всем клеткам домены, так и вариабельные LADs (vLADs), меняющие своё положение в зависимости от клеточного типа или физиологического состояния. Такая изменчивость свидетельствует о том, что контакт с ламиной может быть динамическим механизмом регуляции и участвовать в перепрограммировании клеточной идентичности (https://doi.org/10.1186/s13059-022-02662-6).

Кроме того, нарушение взаимодействия между геномом и ядерной ламиной связано с рядом патологий, включая ламинопатии — наследственные заболевания, вызванные мутациями в генах ламинов, а также с изменениями в экспрессии при старении и канцерогенезе (van Steensel & Belmont, Cell, 2017).

Pc-домены (Polycomb group domains) — это участки хроматина, обогащённые белками Polycomb-группы, которые играют ключевую роль в поддержании транскрипционной репрессии генов, особенно во время эмбрионального развития и дифференцировки клеток. Эти домены формируются за счёт действия двух основных многофункциональных белковых комплексов: PRC1 и PRC2 (Polycomb Repressive Complexes 1 и 2). PRC2 катализирует триметилирование гистона H3 (маркер H3K27me3), что служит сигналом для привлечения PRC1, компактизующего хроматин и блокирующего транскрипцию (https://onlinelibrary.wiley.com/doi/full/10.1002/med.21358?saml_referrer=). Pc-домены, как правило, стабильны в пределах клеточной линии и характеризуются высоким уровнем H3K27me3, а также часто находятся в регионах, содержащих ключевые гены развития, включая Hox-кластеры.

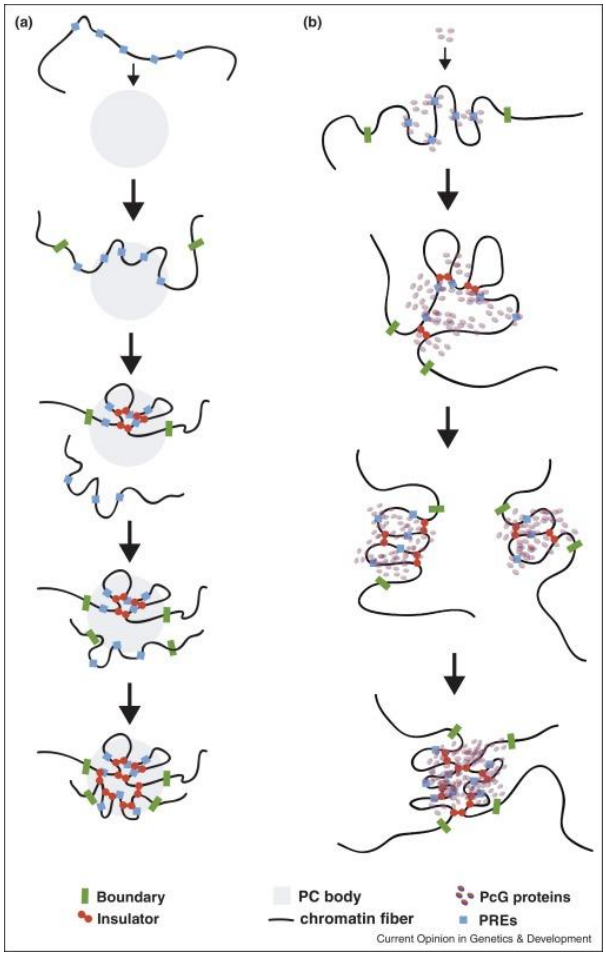

Рис.N: 2 способа (a/b) складывания петель на хромосоме с помощью PcG. В первом случае наблюдается предварительное формирование комплекса хроматина с поликомб-белками, во втором – сборка комплекса происходит непосредственно на целевой хромосоме. Обе модели на настоящий момент считаются равновероятно верными. (https://www.sciencedirect.com/science/article/pii/S0959437X13001834)

Эти домены способны формировать изолированные трехмерные субструктуры внутри ядра — так называемые Polycomb bodies — за счёт взаимодействий между отдалёнными участками Pc-хроматина, что поддерживает устойчивую репрессию генов даже при изменениях в окружающей транскрипционной среде. Pc-домены могут существовать как внутри топологически ассоциированных доменов, так и на их границах, но обладают собственной специфической организацией, отличной от TADs или LADs.

Нарушение структуры и функции Polycomb-доменов ассоциировано с аномальной экспрессией генов, потерей клеточной идентичности и развитием различных патологий, включая рак (https://doi.org/10.4161/cc.5.16.3222). Благодаря способности наследоваться при делении клеток, Pc-домены служат основой для эпигенетической памяти в соматических клетках многоклеточных организмов.

Форум-домены — это структурные участки генома, первоначально описанные на основе анализа участков, подверженных разрушению ДНК в процессе апоптоза. Термин возник из исследований, где было обнаружено, что при программируемом клеточном распаде хроматин разрезается не хаотично, а по вполне определённым границам (https://doi.org/10.1186/1475-4924-1-5), формируя относительно устойчивые и воспроизводимые по размерам доменные единицы, которые кроме того можно получить, подвергнув участок ДНК распаду in vitro (https://academic.oup.com/nar/article/39/9/3667/1252080). Позднее выяснилось, что эти границы совпадают с функционально и структурно обособленными блоками хроматина, обладающими определённой регуляторной автономией. Форум-домены, как правило, имеют размеры от нескольких десятков до сотен тысяч пар основания и содержат кластеры функционально связанных генов или регуляторных элементов.

Считается, что границы форум-доменов определяются специфическими последовательностями ДНК, а также связью с белками, обеспечивающими структурную организацию хроматина, такими как CTCF и компоненты коэзина. В отличие от более строго организованных TAD-доменов, форум-домены могут варьироваться по составу и быть чувствительными к клеточному состоянию. Они проявляют корреляцию с профилями генной экспрессии и, по-видимому, отражают модульную архитектуру генома,

необходимую для тонкой настройки координированной регуляции генов (https://doi.org/10.1073/pnas.89.15.6751).

В некоторых исследованиях (https://doi.org/10.1142/S1094406021500116, https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0060245) также упоминается, что форум-домены могут соответствовать единицам репликации и участвовать в структурной сегрегации генома при переходах между фазами клеточного цикла. Хотя их природа и механизмы формирования изучены не так подробно, как в случае TADs или LADs, форум-домены представляют интерес как функционально обособленные блоки, участвующие в пространственной организации хроматина и регуляции транскрипции.

Помимо рассмотренных выше топологически ассоциированных доменов (TADs), ламино-ассоциированных доменов (LADs), Pc-доменов и форум-доменов, в геноме эукариот описано и множество других типов доменных структур, отражающих разные аспекты пространственной и функциональной организации хроматина. Среди них — компартменты A и B, выявленные при Hi-C анализе как чередующиеся крупномасштабные участки хроматина с разной транскрипционной активностью и плотностью упаковки (https://doi.org/10.1016/j.ceb.2024.102406); репликационные домены, отражающие синхронные блоки ДНК-репликации; а также энхансерные кластеры (https://www.nature.com/articles/ng.3539) и суперэнхансер-домены (https://www.nature.com/articles/s41467-018-03279-9), объединяющие регуляторные элементы, участвующие в контроле ключевых программ экспрессии.

Каждый из этих типов доменов фокусируется на определённом уровне регуляции — от взаимодействий с ядерным ламином до координации транскрипции и репликации. Современная геномика продолжает уточнять классификацию и взаимосвязи этих доменов,

что позволяет всё более точно описывать архитектуру ядра и её роль в регуляции генетической информации.

2.3Молекулярно – биологические (экспериментальные) методы детекции и изучения доменных структур

Изучение трёхмерной организации генома стало возможным благодаря развитию ряда молекулярно-биологических методов, позволяющих выявлять пространственные взаимодействия между участками ДНК, их привязанность к ядерным структурам и эпигенетическое состояние. Каждый из описанных выше типов доменных структур был открыт (или подтверждён) с использованием специфических экспериментальных подходов, многие из которых требуют сложной пробоподготовки, дорогостоящего оборудования и глубокой биоинформатической обработки. В данной главе мы рассмотрим ключевые технологии, позволившие идентифицировать TAD-, LAD-, Pc- и форум-домены, с акцентом на их принцип действия, применимость и ограничения.

Метод Hi-C (от англ. High-throughput Chromosome Conformation Capture) представляет собой одну из наиболее широко применяемых технологий для изучения трёхмерной организации генома, в частности — для идентификации TADs. Принцип метода основан на фиксации хроматиновых взаимодействий с помощью формальдегида, расщеплении ДНК рестриктазами, последующем лигировании пространственно сближенных фрагментов и высокопроизводительном секвенировании полученных химерных молекул Illumina методом (https://www.sciencedirect.com/science/article/abs/pii/S1046202312001168). Области с

высокой плотностью внутридоменных контактов и пониженной частотой взаимодействия с соседними участками интерпретируются как TAD-домены. Ключевая особенность Hi-C заключается в том, что он фиксирует все возможные взаимодействия между всеми участками генома, создавая матрицу контактов, в которой каждая ячейка отражает частоту взаимодействия пары локусов – именно этим методом было получено изображение, показанное в качестве иллюстрации в параграфе 1.1. Это делает метод крайне ресурсоёмким: чтобы получить матрицу с разрешением в 5–10 тыс. п.о. (необходимую для надёжной детекции TAD-доменов), требуется порядка 200–500 миллионов парных чтений на один образец, что уже на уровне секвенирования может стоить от нескольких сотен до нескольких тысяч долларов, в зависимости от глубины и платформы (https://www.tandfonline.com/doi/abs/10.2217/14622416.6.7.777). Дополнительную сложность представляет пробоподготовка, требующая высокой степени воспроизводимости, а также последующий анализ данных, включающий алгоритмы нормализации, фильтрации и доменной сегментации (например, DI score, insulation score, или HMM). Стоит отметить также, что алгоритмы подоменной сегментации данных, полученных этим методом, часто требуют предтренировки модели (https://dl.acm.org/doi/abs/10.1145/383952.384021) и потому не могут быть использованы для широкого круга экспериментов. Эти факторы делают Hi-C малопригодным для рутинного применения в небольших лабораториях, но при этом незаменимым в крупномасштабных проектах, таких как ENCODE или 4D Nucleome (https://www.nature.com/articles/nature23884).

Следующий важный метод - DamID (DNA adenine methyltransferase identification), благодаря которому были проведены исследования ламино-ассоциированных доменов

(LADs). В основе метода лежит использование фермента Dam метилтрансферазы, который добавляет метильные группы на аденины в ДНК в непосредственной близости от специфической белковой мишени (https://doi.org/10.1242/dev.173666). В случае с LAD, метилтрансфераза привязывается к белкам ядерной оболочки, таким как ламин A, и метилирует участки ДНК, которые физически взаимодействуют с этой оболочкой. После этого можно извлечь и секвенировать метилированную ДНК, чтобы выяснить, какие участки генома находятся рядом с ламином и составить карту таких взаимодействий, как например в (https://www.embopress.org/doi/full/10.15252/embr.202050636).

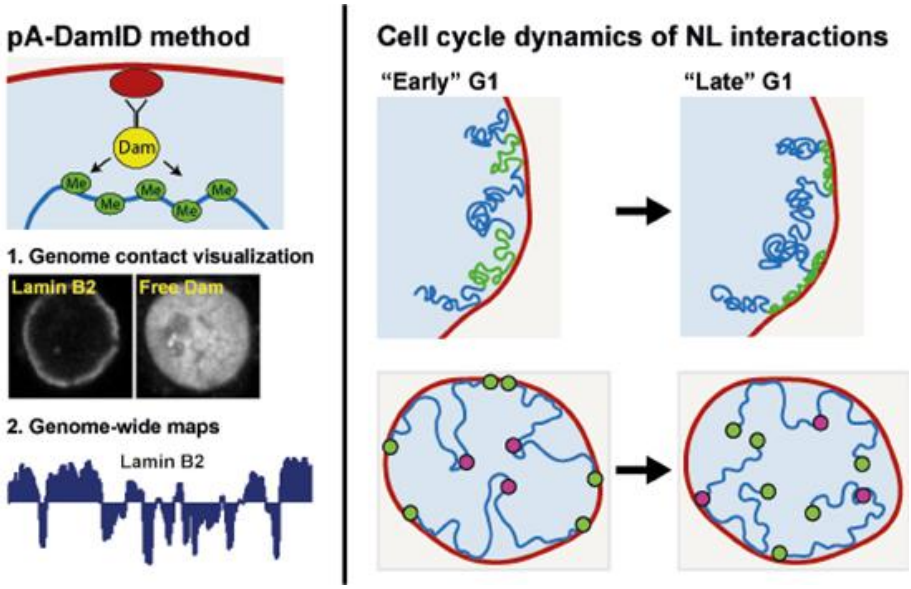

Рис. N: Пример использования DamID метода для исследования динамики LAD в различных фазах клеточного цикла. Заметим, что химерный белок Dam+aLam требуется создавать отдельно под каждый вид всилу возможных изменений структуры ламины.

Этот метод имеет несколько преимуществ: он позволяет изучать взаимодействия генома с ядерной мембраной в живых клетках, без необходимости в экстенсивной фиксации, как это требуется в методе Hi-C. Однако он также имеет ограничения — его

разрешение ограничено, и точность предсказания ламино-ассоциированных доменов может варьироваться в зависимости от используемого материала и условий. Кроме того, хотя DamID значительно дешевле, чем Hi-C, он всё равно требует большого объёма данных и мощных вычислительных ресурсов для обработки (https://www.nature.com/articles/nprot.2016.084). Оценки затрат варьируются, но в зависимости от задачи цена секвенирования и анализа может составить несколько тысяч долларов (https://doi.org/10.3389/fimmu.2014.00531).

Следующий популярный метод - Chromatin Immunoprecipitation followed by sequencing (или кратко ChIPSeq) позволяет определить участки генома, с которыми взаимодействуют конкретные белки или белковые комплексы, связанные с хроматином. Принцип метода заключается в фиксации белков к ДНК (обычно формальдегидом), фрагментации хроматина, иммунопреципитации с использованием специфических антител к белку интереса, обратной кросс-линковке и последующем высокопроизводительном секвенировании осаждённых фрагментов. Результатом является карта участков ДНК, ассоциированных с данным белком (https://link.springer.com/article/10.1186/1471-2164-12-134).

В контексте изучения пространственной организации генома метод ChIP-seq широко применяется для анализа распределения белков, участвующих в формировании и поддержании различных типов доменных структур. В частности, он используется в следующих направлениях:

1) TAD-граничные элементы: Так как многие границы TAD совпадают с участками, на которых обогащены архитектурные белки CTCF (CCCTC-binding factor) и коэзином, то при помощи ChIP-seq, направленного на CTCF и компоненты коэзина (например, RAD21), можно определить позиции этих белков на геноме и тем самым локализовать потенциальные границы TAD (https://www.nature.com/articles/s41467-019-10725-9).
2) Для изучения Pc-доменов используется ChIP-seq, направленный на специфические модификации, характерные для молчаливого хроматина, прежде всего H3K27me3 (триметилирование 27-го лизина гистона H3). Таким образом можно картировать протяжённые Pc-домены, (https://academic.oup.com/nar/article/39/17/7415/2411300) характеризующиеся высокой плотностью этой эпигенетической метки и сниженной транскрипционной активностью.
3) В некоторых исследованиях для выявления LAD используется ChIP-seq по белкам ядерной ламины, например Lamin B1. Отметим однако, что такой подход лишь косвенно отражает распределение LAD и может иметь меньшую чувствительность и специфичность по сравнению с методами прямой визуализации, такими как DamID. (https://link.springer.com/article/10.1186/s13059-022-02662-6) Тем не менее, при грамотном контроле и перекрёстной валидации результаты ChIP-seq по Lamin B1 могут быть интерпретированы как карта потенциальных LAD.

Преимущество ChIP-seq — его относительная простота и доступность: он требует гораздо меньше клеточного материала по сравнению с методами типа Hi-C и не предполагает всеобъемлющего секвенирования всего набора взаимодействий, как в методах на базе 3C. Однако метод критически зависит от качества и специфичности

используемых антител. Ошибки в выборе или характеристиках антител могут привести к ложным положительным или отрицательным результатам (https://link.springer.com/article/10.1186/s13072-016-0100-6). Кроме того, ChIP-seq не позволяет напрямую идентифицировать пространственные контакты между участками ДНК, а лишь указывает на потенциальные регуляторные и структурные "якоря", которые могут быть вовлечены в формирование доменов. Поэтому он часто используется в комбинации с пространственно-ориентированными методами (Hi-C, 4C) и транскриптомными подходами.

Завершая главу, стоит кратко упомянуть и другие экспериментальные методы, применяемые для изучения пространственной организации генома и обнаружения доменных структур. Ранние подходы, такие как 3C (Chromosome Conformation Capture) и его производные — 4C и 5C, были предшественниками Hi-C и позволяли выявлять взаимодействия между двумя или множеством заданных участков ДНК (https://genesdev.cshlp.org/content/26/1/11.short). Хотя они обладают меньшим охватом по сравнению с Hi-C, они остаются востребованными для таргетированного анализа и менее ресурсоёмки.

Интерес представляет и SPRITE (Split-Pool Recognition of Interactions by Tag Extension) — метод, позволяющий выявлять множественные взаимодействия без стадии лигирования, что особенно полезно при изучении высокоуровневых кластеров, таких как Polycomb-компартменты (https://thesis.library.caltech.edu/16147/).

Методы визуализации, в частности DNA-FISH (fluorescence in situ hybridization), позволяют непосредственно наблюдать пространственное положение определённых геномных локусов, включая, например, LAD (https://www.nature.com/articles/s41586-019-1233-0) или Pc-домены, в ядре клетки. Хотя этот подход не даёт "глобальной" картины взаимодействий, он остаётся незаменимым для качественной валидации результатов высокопроизводительных методов.

Наконец, ATAC-seq и DNase-seq, направленные на картирование открытого хроматина, применяются для уточнения положений доменных границ (https://link.springer.com/article/10.1186/s43556-020-00009-w), так как в этих регионах часто локализуются сайты связывания регуляторных белков, таких как CTCF или гистон-модифицирующие комплексы.

Нельзя не отметить, что несмотря на всё более тонкие экспериментальные методы, математические и вычислительные подходы не отстают по значимости. Выделим два основных направления развития: с одной стороны, они позволяют проверить, связаны ли наблюдаемые доменные структуры с координированной экспрессией генов или другими функциональными признаками; с другой — наоборот, попытаться выявить доменные границы исходя из характерных паттернов экспрессии, без прямого учёта пространственной информации. Именно этим возможностям посвящена следующая глава.

2.5 Вычислительные методы валидации или предсказания доменных структур

Одним из наиболее распространённых подходов к валидации функциональной значимости доменных структур, таких как TAD или Pc-домены, является анализ координированной экспрессии генов, расположенных внутри этих доменов. Предполагается, что если домен представляет собой функционально целостную единицу регуляции, то гены внутри него должны демонстрировать более высокую степень коррелированной экспрессии по сравнению с генами, разделёнными доменными границами (bioRxiv. 2019;771402). Для проверки этой гипотезы применяются различные статистические методы.

Первой группой статистических валидационных тестов являются permutation-тесты: в них случайным образом перемешиваются координаты доменных границ или расположение генов, после чего пересчитывается средняя внутридоменная корреляция. Это позволяет сформировать "нулевое распределение", соответствующее ожиданиям при отсутствии реальной связи между структурой и экспрессией, и оценить статистическую значимость наблюдаемого эффекта, как например было сделано для доказательства связи между топологией ДНК и экспрессией в (https://www.nature.com/articles/s41588-019-0462-3). Такие подходы многократно подтверждали, что гены внутри доменных структур демонстрируют более высокую внутригрупповую корреляцию уровней экспрессии, особенно при сравнении различных клеточных состояний или условий, например, в норме и патологии(https://doi.org/10.1186/s13072-019-0317-2). Это служит косвенным, но воспроизводимым доказательством того, что пространственная организация хроматина

оказывает влияние на транскрипционную регуляцию, а доменные структуры представляют собой не только структурные, но и функциональные единицы генома.

Помимо permutation-тестов, валидация координированной экспрессии может проводиться вторым способом - с использованием методов Монте-Карло (ММК). Хотя оба подхода основаны на генерации случайных распределений, между ними существует принципиальное различие. Permutation-тесты работают с фиксированным набором значений, производя перестановки наблюдаемых данных (например, значений экспрессии генов) для оценки вероятности получения наблюдаемого результата случайно. В отличие от этого, ММК предполагают генерацию данных из заранее заданного вероятностного распределения (например, нормального или Пуассона), что позволяет учитывать вариативность, присущую биологическим системам, но требует более строгих предположений о природе этих данных. В частности, с помощью такого подхода в (https://www.cell.com/biophysj/fulltext/S0006-3495(18)34283-8) было показано как TAD могут быть связаны с межхромосомным взаимодействием. Таким образом, ММК обеспечивают большую гибкость, но требуют аккуратного выбора модели распределения. Более того, в рамках таких подходов часто напрямую вычисляются статистические метрики (например, корреляция или взаимная информация между генами внутри домена), не всегда сопровождаясь строгой проверкой на статистическую устойчивость или значимость. Это упрощает первичную оценку, но накладывает ограничения на интерпретацию полученных результатов без последующего тестирования гипотез.

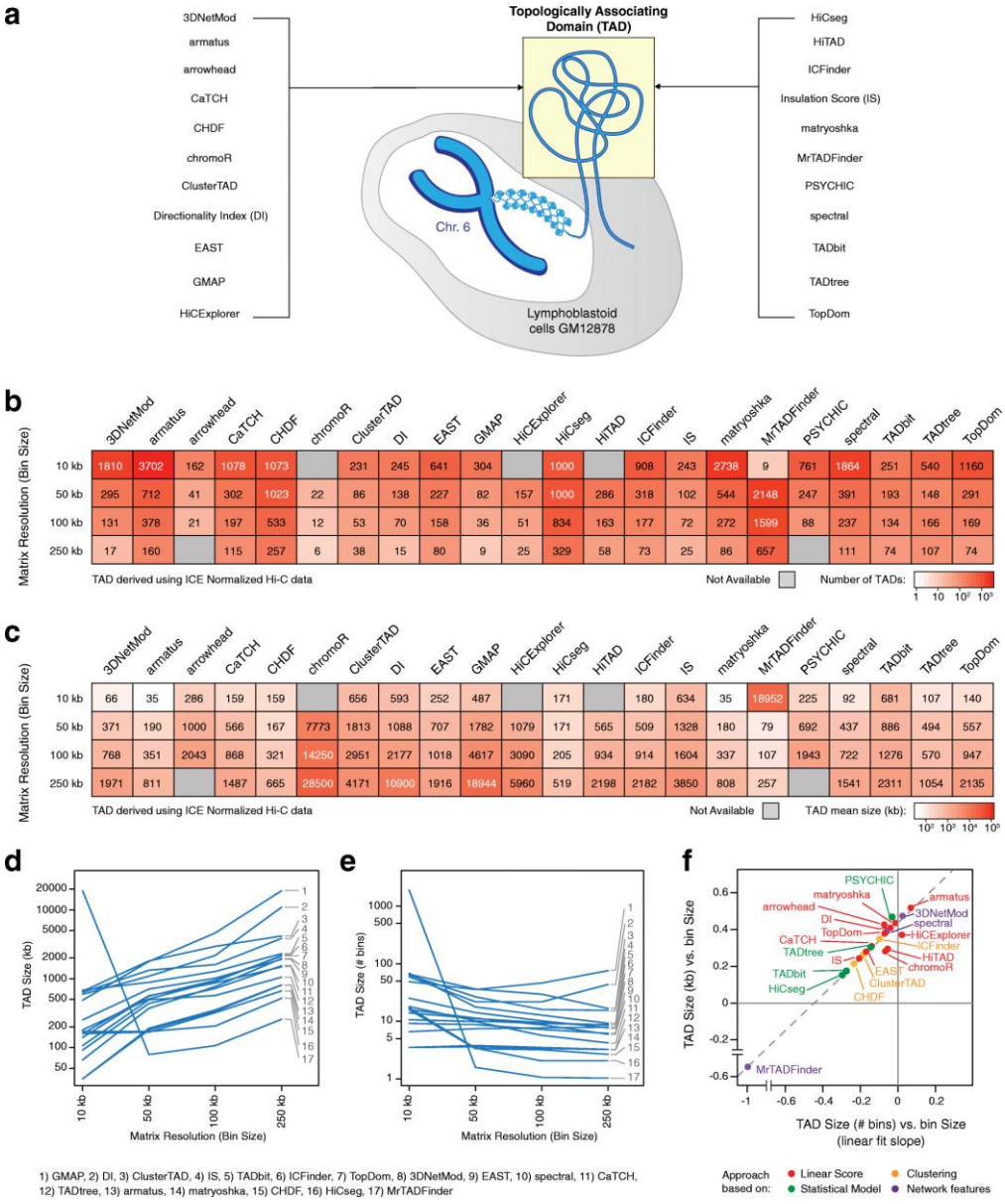

Рис.N. Сравнительная характеристика 17 моделей для валидации TAD по данным Hi-C. Заметим, что некоторые модели показываю наилучшие результаты на малых доменах, некоторые – на более крупных, и как правило их области наибольшей продуктивности не пересекаются.

(https://link.springer.com/article/10.1186/s13059-018-1596-9?fromPaywallRec=false)

Важно отметить, что обе группы валидационных тестов не существуют в вакууме и требуют предварительных данных о доменной структуре генома. Чаще всего такие данные получают с использованием Hi-C или ChIP-seq, и, следовательно, несмотря на кажущуюся

простоту математической оценки, сам по себе анализ возможен лишь при наличии априорной информации о границах доменов, полученной экспериментально. Это ставит определённые ограничения на масштабируемость таких исследований.

Далее стоит упомянуть, что помимо математических моделей, исследующих зависимости между доменной организацией генома и координированной экспрессией, существуют также и решения обратной задачи – попытки предсказать расположения доменов по данным экспрессии генов, сиквенсам и другим (https://link.springer.com/article/10.1186/s12864-019-6303-z) признакам. Прежде всего, за последние несколько лет появилось множество (https://link.springer.com/protocol/10.1007/978-1-0716-4136-1_22) описаний моделей машинного и глубокого обучения (ML и DL соответственно) - класс стохастических алгоритмов, в которых способность к предсказанию расположения доменных структур тренируется на больших объёмах заранее известных данных. Например, уже существуют модели для классификации или предсказания расположения TAD с точностью менее 1000 пар оснований на основе данных Hi-C (https://academic.oup.com/nar/article/47/13/e78/5485073). Говоря именно об ML-моделях, одной из ключевых проблем является качество данных, как и в случае с методами валидации. Результаты предсказаний зависят от того, насколько репрезентативными и точными являются исходные данные, такие как секвенирование (например, Hi-C или ChIP-seq). Даже небольшие ошибки в таких данных могут привести к ошибочным выводам о структуре доменов. Другим вызовом является интерпретируемость моделей. В то время как машинное обучение может давать точные результаты, часто бывает сложно объяснить, какие именно признаки или взаимодействия между генами привели к предсказанию

конкретного домена. Особенно сложной интерпретация результатов становится при переходе от более простых ML моделей к более многослойным (глубоким) способам.

Кроме ML/DL моделей, существует также несколько алгоритмов для предсказания расположения доменов (в основном TAD) на основе гибридных данных (https://www.science.org/doi/full/10.1126/sciadv.aaw1668) – как по корреляции значений экспрессии, так и по структурным данным о маркерах связывания хроматина. Тем не менее, придётся повторить, что такие алгоритмы тем не менее используют заранее измеренные данные о хромосоме (https://www.sciencedirect.com/science/article/abs/pii/S1084952115300112), на которой предполагается поиск доменов. Следовательно, при попытке изучить новый организм, ткань или патологию, придётся оптимизировать метод под задачу или проводить измерения Hi-C или ChIP-Seq методом.

Таким образом, возникает естественная потребность в математическом подходе, который, с одной стороны, не требовал бы большого объёма априорных данных (например, Hi-C или ChIP-seq), а с другой — сохранял бы простоту и устойчивость permutation-подходов, широко применяемых в валидации доменных структур. Особенно актуально это в контексте ограниченного доступа к дорогим методам прямого анализа пространственной архитектуры хроматина: стоимость Hi-C и ChIP-seq по-прежнему значительно превышает расходы на получение данных транскриптомики, таких как RNA-seq или CAGE. Следовательно, подход, основанный исключительно на информации об экспрессии, представляется не только методологически привлекательным, но и практичным. Такой метод должен позволять не только статистически обоснованно проверять гипотезу о связи пространственной организации генома с координированной экспрессией генов, но и, по

возможности, приближённо локализовать границы функциональных доменов исключительно на основе транскриптомных данных. Это открыло бы путь к предварительной аннотации доменов в условиях, когда экспериментальные данные пространственной хроматинной архитектуры недоступны или слишком затратны для получения, и стало бы полезным инструментом для планирования таргетных исследований.

3.Материалы и методы

3.1 Используемые экспериментальные и модельные данные

Для анализа пространственной организации транскрипционной активности в геноме и оценки связей между доменными структурами и координированной экспрессией мы использовали как экспериментальные, так и сгенерированные модельные данные.

В качестве основного источника транскриптомных данных были выбраны общедоступные репозитории NCBI Gene Expression Omnibus (GEO) (https://pmc.ncbi.nlm.nih.gov/articles/PMC99122/) и ресурс GeneHancer (https://www.genecards.org/Guide/GeneHancer) . Эти базы предоставляют нормализованные карты экспрессии, измеренные в TPM (Transcripts Per Million) — стандартизированной метрике, которая учитывает длину гена и общую глубину секвенирования. Наиболее широко используемым методом получения таких данных остаётся RNA-seq (doi: https://doi.org/10.1101/836973) — высокопроизводительное секвенирование транскриптов, позволяющее в масштабе всего генома оценить уровни экспрессии с высокой чувствительностью и разрешением.

Следует, однако, отметить, что несмотря на широкую распространённость RNA-seq и значительные успехи в стандартизации методов обработки, точность количественной

оценки уровней транскрипции может варьировать в зависимости от качества подготовки библиотек, глубины секвенирования, используемых алгоритмов выравнивания и нормализации, а также биологической вариабельности образцов (https://link.springer.com/article/10.1007/s11427-011-4255-x). В частности, уровни TPM могут быть менее надёжны при сравнении низкоэкспрессируемых генов или в регионах с высокой плотностью транскриптов. Эти ограничения необходимо учитывать при построении моделей, опирающихся на экспрессионные профили.

В рамках настоящей работы мы сосредоточились на третьей хромосоме человека (hg38), выбор которой был продиктован стремлением к снижению объёма обрабатываемых данных без существенной потери общности результатов. На основе полученных срезов реальных карт транскрипции (https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr3%3A10183319-10195354&hgsid=2545994700_PXfsBA9P132p4afkF8a8RaprZABG) были построены эмпирические распределения ключевых параметров: длины генов, уровней экспрессии в TPM (transcripts per 1 million reads) и расстояний между соседними граничными элементами доменов. Эти распределения затем использовались для генерации модельных выборок с реалистичной структурой: гены и их экспрессия синтетически размещались на хромосоме с сохранением статистических свойств, наблюдаемых в эксперименте, а границы доменов закладывались согласно аппроксимированным закономерностям доменной организации.

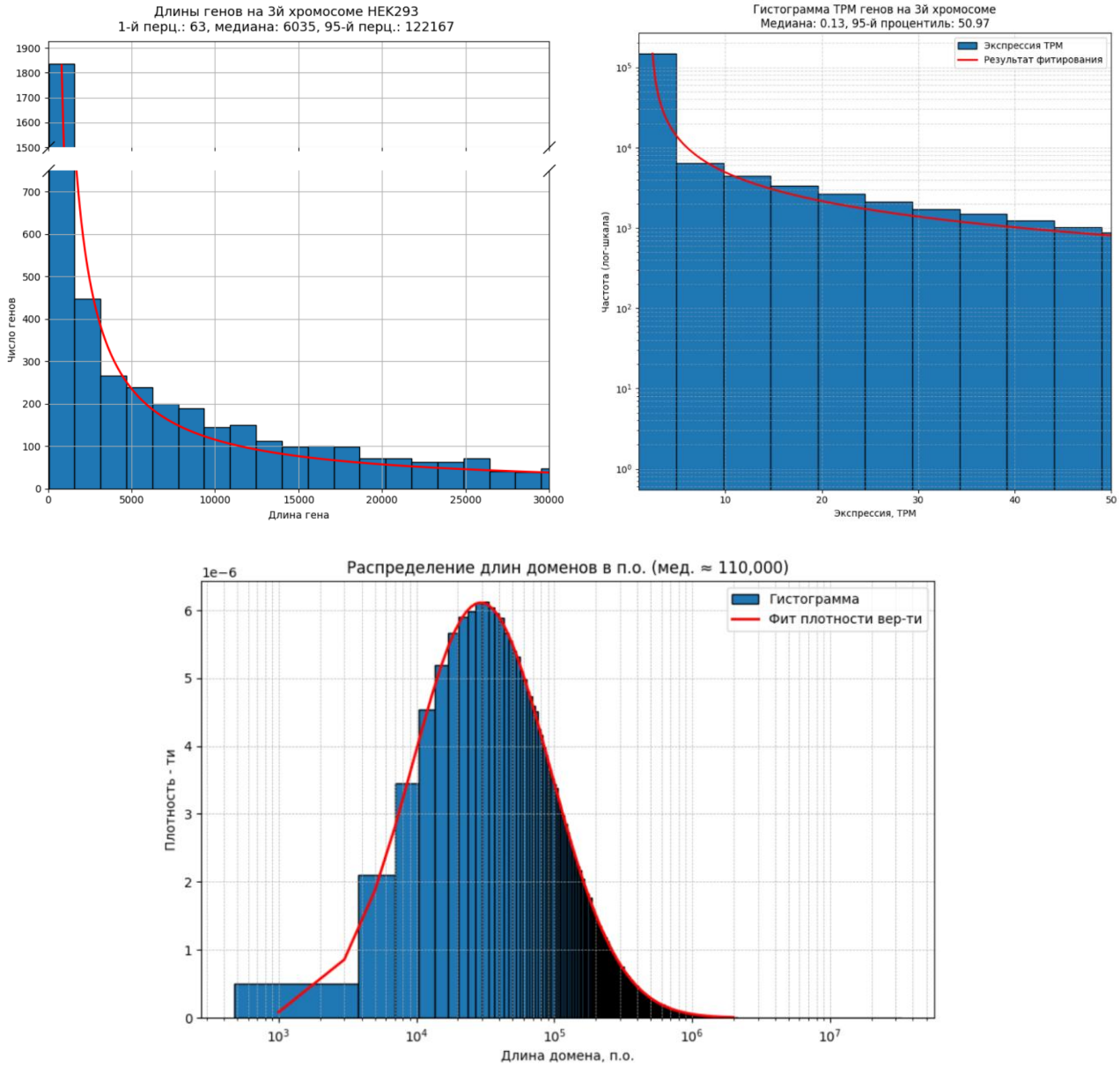

Рис.N: Распределения длин генов, уровней экпрессии и длин доменов на hg38 HEK293 и кривые фитирования, по которым далее строятся модельные данные

Таким образом, сгенерированные данные сохраняют статистическую правдоподобность и являются валидными для тестирования методики.

3.2 Программные пакеты

Разработка и валидация моделей, описанных в данной работе, проводились с использованием языка программирования Python версии 3.10. Основная часть вычислений реализована с применением стандартных модулей Python (`subprocess`, `threading`, `multiprocessing`, `glob`, `random`) и внешних библиотек `numpy`, `scipy`, а также `tqdm` — для визуализации прогресса выполнения операций. Для ускорения отдельных участков кода использовалась библиотека `numba` (через декоратор `jit`). При работе с биологическими последовательностями применялись инструменты пакета Biopython.

Следует отметить, что точный набор библиотек и их версии могут меняться при дальнейшем развитии проекта. В будущем вопросы совместимости программного окружения могут повлиять на воспроизводимость результатов при повторном запуске. Поэтому при практическом использовании рекомендуется зафиксировать рабочее окружение, например, средствами `pip freeze`, `conda` или `poetry`. Ниже приведена таблица используемых модулей Python, которые необходимы для корректной работы скриптов и воспроизводимости результатов, описанных в данной работе.

Библиотека	**Версия**	**Назначение**
Python interpreter	3.10.8	Основной интерпретатор python
numpy	2.2.5	Численные вычисления и работа с большими данными
scipy	1.14.1	Статистические вычисления
tqdm	4.66.5	Визуальное оформление для удобства потоковых запусков
numba	0.61.2	Ускорение вычислений с помощью JIT - компилляции
biopython	1.79	Работа с биологическими последовательностями, парсинг результатов секвенирования

3.3 Статистический анализ, математические методы

Описание, доказуемость:

Разработанный метод основан на статистически доказуемой процедуре перестановочного (пермутационного) тестирования. В основе алгоритма лежит проверка гипотезы о наличии координированной экспрессии генов внутри заранее определённых (или предполагаемых, как в нашем случае) доменных структур.

Пусть заданы:

- $D = \{D_1, D_2, \ldots, D_n\}$— множество доменов (непересекающихся диапазонов на хромосоме),
- $G = \{G_1, G_2, \ldots, G_m\}$ — множество генов с измеренными значениями экспрессии,
- $E = \{e_1, e_2, \ldots, e_m\}$— вектор уровней экспрессии для каждого гена.

Проверяется нулевая гипотеза **H_0:** *координация экспрессий внутри доменов не отличается от случайной*, то есть наблюдаемая структура может быть получена случайной перестановкой экспрессий между генами. Для оценки статистической значимости исследуется эмпирическое распределение **T** среднего уровня экспрессии в доменах под нулевой гипотезой. Это достигается через многократную генерацию перестановок экспрессий:

- При каждой итерации производится циклический сдвиг (ротация) вектора E, не нарушающий его глобального распределения,

- Производится повторное распределение экспрессий по доменам, и вычисляется значение статистики $\mathbf{T^{(i)}}$:

$$\boldsymbol{\mu_i = \frac{1}{i}\sum_{j=1}^{i} T^{(j)}} \qquad \boldsymbol{\sigma_i^2 = \frac{1}{i}\sum_{j=1}^{i}(T^{(j)} - \mu i)}$$

- Полученное множество $\{\mathbf{T^{(1)}, T^{(2)}, \ldots, T^{(R)}}\}$, формирует эмпирическое нулевое распределение (Здесь **R** – выбранное для анализа количество ротаций).

Оригинальное значение статистики $\mathbf{T^{(0)}}$, полученное без перестановок, сравнивается с этим распределением, что позволяет оценить p-value или z-оценку:

$$\boldsymbol{z = \frac{T^{(0)} - \mu(T^{(1)}, T^{(2)}, \ldots, T^{(R)})}{\sigma_T}}$$

где $\boldsymbol{\mu_T}$ и $\boldsymbol{\sigma_T}$ — среднее и стандартное отклонение статистики T по перестановкам.

В целях ясности в приложении N приводится алгоритмическое описание основной функции, осуществляющей оценку наличия координированной экспрессии генов в пределах заданных доменных структур. Алгоритм представлен в абстрактной форме (псевдокод), понятной вне зависимости от используемого языка программирования. Полная версия доступна для скачивания и анализа в открытом репозитории проекта.

Доказательство корректности:

Рассмотрим корректность перестановочного алгоритма оценки координированной экспрессии классическим способом через инвариант цикла. Предположим, что исходное распределение **E** содержит структурированную (негомогенную) компоненту, то есть

существуют блоки генов (в рамках доменов), где экспрессия координирована: среднее **μ** и дисперсия **σ** внутри доменов отличаются от того, что ожидается при случайном распределении (0 и 1 соответственно так как это соответствует нормального распределения последовательности случайных чисел). На каждом шаге $\mathbf{i \in \{1,\ldots,R\}}$ алгоритм производит **ротацию** вектора экспрессий: **E(i) = roll(E, si)** где $\mathbf{si \sim rand\_int(-m, m)}$— случайный циклический сдвиг. Таким образом, структура экспрессий сохраняется глобально (маргинальное распределение инвариантно), но локальная координация между соседними генами нарушается. Обозначим статистику интереса (среднее и дисперсию в доменах) как $\mathbf{T^{(i)} = T(E^{(i)})}$ а исходную как $\mathbf{T^{(0)}}$. Пусть также $\mathbf{\mu(T_i)}$ и $\mathbf{\sigma(T_i)}$ — среднее и дисперсия по i ротациям.

Инвариант цикла заключается в следующем утверждении: "*На каждой итерации i, множество значений* $\{T^{(1)}, T^{(2)}, \ldots, T^{(i)}\}$ *приближается к эмпирическому распределению (нормальное распределение с средним в 0 и дисперсией равной 1) при нулевой гипотезе* H_0*: экспрессия в доменах случайна."*

(i = 1): Первая ротация генерирует вектор **E(1)=roll(E, s1)** — случайный сдвиг. Этот вектор является перестановкой исходного, с сохранением глобального распределения, но со случайной локальной структурой. Следовательно, **T(1)**— реализация **T** при $\mathbf{H_0}$. Инвариант выполнен.

Пусть после i шагов инвариант выполнен: $\mathbf{\{T(1),\ldots,T(i)\} \subset T}$ **при** $\mathbf{H_0}$ Тогда на i+1 шаге создаётся новый вектор $\mathbf{E(i+1) = roll(E, s_{i+1}),}$ причём $\mathbf{s_{i+1} \sim random(-m, m),}$ независимо от предыдущих. Следовательно, **T(i+1)**— независимый элемент выборки из того же распределения, а поскольку элемент независим, его добавление снижает дисперсию оценки

среднего, повышая её точность и значение z – критерия сходится к 0, т.к. было взято случайное число, пусть и из ограниченного набора (случайная выборка и другая случайная выборка из одного и того же исходного набора практически неотличимы). Инвариант сохраняется.

По завершении R итераций мы получаем R независимых реализаций статистики при H. Эти значения используются для оценки математического ожидания и дисперсии, и затем — для вычисления z-критерия, где — наблюдаемая статистика в исходных данных. Таким образом, оценка справедлива при выполнении инварианта на каждом шаге $\mathbf{T}^{(0)}$.

Оценка сложности и терминируемость алгоритма

Пусть D — число доменов, G — максимальное число генов в одном домене, R — число итераций (циклических ротаций), N — общее число генов во всей хромосоме.

На каждой итерации алгоритм выполняет следующие действия:

1. Циклический сдвиг экспрессионного массива**:** Операция np.roll имеет амортизированную сложность O(N).
2. Применение весов к подмассивам (восстановление структуры доменов)**:** Для каждого домена производится срез и поэлементное умножение длиной не более G, всего таких операций D, итоговая сложность O(D·G)
3. Вычисление статистик по доменам: Для каждого домена используется векторизованное вычисление. Это даёт O(D·N).
4. Накопление вертикальной суммы, суммы квадратов и количества элементов**:** Также O(N), поскольку каждый ген участвует ровно в одной такой операции.

Итого, весь алгоритм требует: **O(R·G·D)**

При этом память используется для хранения:

1. одного массива экспрессий длины N,
2. массива весов (разбитого по доменам, также суммарно длины N),
3. трёх вертикальных массивов длины D,
4. накопленных сумм и счётчиков — также O(D).

То есть, алгоритм использует память O(N+D) что на самом деле условиях N < D дает O(N)

Поскольку число итераций R задаётся явно, и в каждой итерации не выполняется рекурсия или алгебра с неопределённой длиной, алгоритм **гарантированно терминируется** за R итераций и является **полиномиально ограниченным** по времени и объёму.

4.Результаты и обсуждение

4.0 Разработка и отладка центральных функций метода

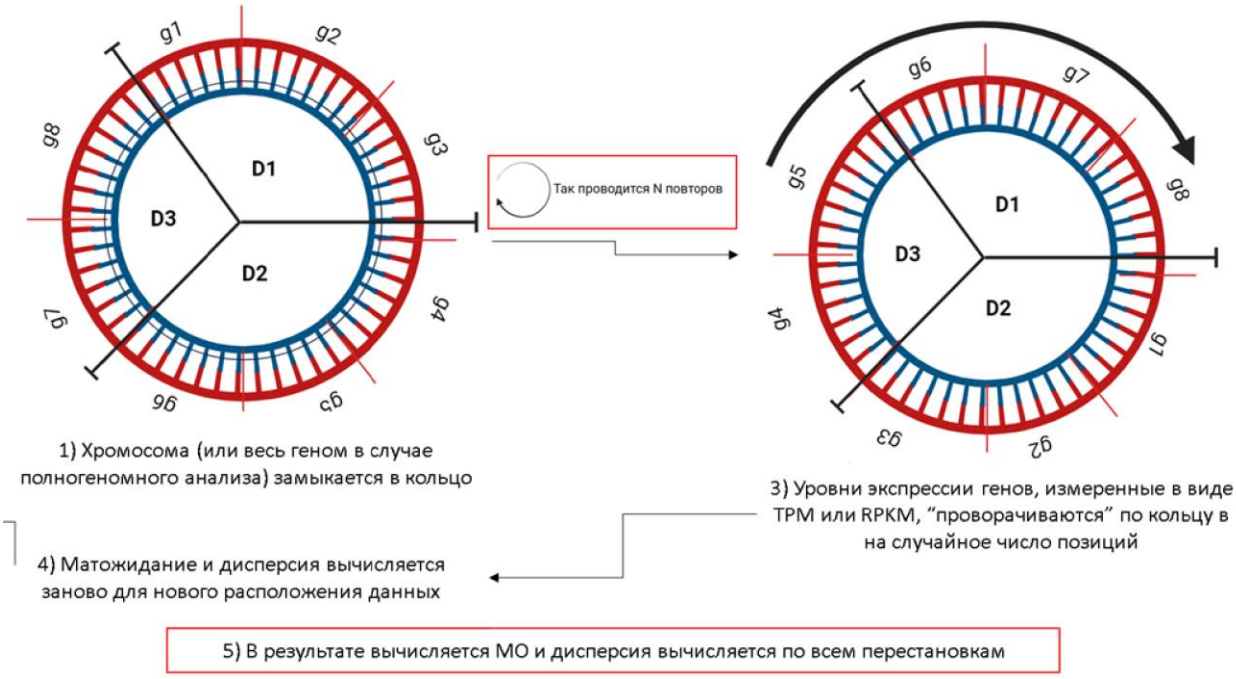

Рис.N: Иллюстрация принципа работы пермутативного алгоритма

На первом этапе работы была реализована вычислительная основа для анализа координированной экспрессии генов в пределах хромосомных доменов. Мы разработали и протестировали все ключевые модули, включая:

1. Генерацию случайных и синтетических хромосом с различными уровнями упорядоченности,
2. Расчет z-критерия, отражающего уровень координированности экспрессии в доменах

3. Пермутационные тесты, позволяющие контролировать ложно положительные результаты (FDR)
4. Встроенные проверки через статистические тесты (в частности, тест Андерсона-Дарлинга) на соответствие теоретически предполагаемому распределению

Таким образом, на первом этапе работы была создана базовая вычислительная платформа, способная как на численный анализ, так и на структурные биоинформатические эксперименты, включая положительные и отрицательные контроли, работу с реальными и синтетическими данными, и масштабирование задач от отдельных хромосом до целых геномов.

Хотя наша задача связана с большими объёмами данных и потенциально требовательными по времени вычислениями (особенно при использовании десятков тысяч пермутаций), мы сознательно выбрали язык Python в качестве основного инструмента разработки. Причин этому несколько:

1. Python позволяет очень быстро встраивать новые функции, переписывать логику, масштабировать проект или адаптировать его под разные входные форматы данных — что особенно важно в научной разработке, когда задачи меняются по мере получения новых идей и результатов.
2. Большое количество удобных библиотек (например, numpy, scipy, matplotlib, pandas) делают Python идеальной средой для быстрой разработки и визуальной проверки гипотез.

3. Поддержка сообщества и наличие биоинформатических библиотек. Многие популярные биоинформатические проекты, такие как Biopython, HTSeq, Scanpy, также ориентированы на Python, что делает интеграцию с внешними источниками данных и аннотациями (например, TAD/LAD картами) максимально бесшовной.

Безусловно, языки типа C/C++, Rust или Julia могли бы дать прирост в скорости, однако на данном этапе приоритетом была не абсолютная производительность, а скорость разработки, гибкость и адаптивность. В будущем, по мере стабилизации логики алгоритма, возможна частичная оптимизация критически важных участков кода через numba, cython или даже переписывание отдельных модулей на низкоуровневом языке — но уже на более зрелой стадии проекта.

4.1 Отрицательный контроль (проверка центальной гипотезы)

Центральной гипотезой предлагаемого подхода является утверждение о том, что в случае отсутствия структурной связи между пространственной организацией генома и уровнями экспрессии (т.е. при случайной организации доменов и случайных значениях транскрипционной активности), результирующее распределение z-оценок, полученных в ходе пермутационного тестирования, должно соответствовать стандартному нормальному закону с математическим ожиданием 0 и дисперсией 1. Ложноположительными результатами в данной парадигме считаются значения $|z| > 1.96$.

Для верификации этой гипотезы в качестве отрицательного контроля была проведена серия испытаний на 10 000 искусственно синтезированных транскриптомных

картах, построенных на базе структуры 3-й хромосомы человека. При этом реальные значения экспрессии были заменены на случайные, равномерно распределённые значения в интервале от 0 до 95-го процентиля эмпирического распределения. Сами доменные структуры также генерировались случайным образом с сохранением статистических параметров (число доменов, распределение длин и т.д.).

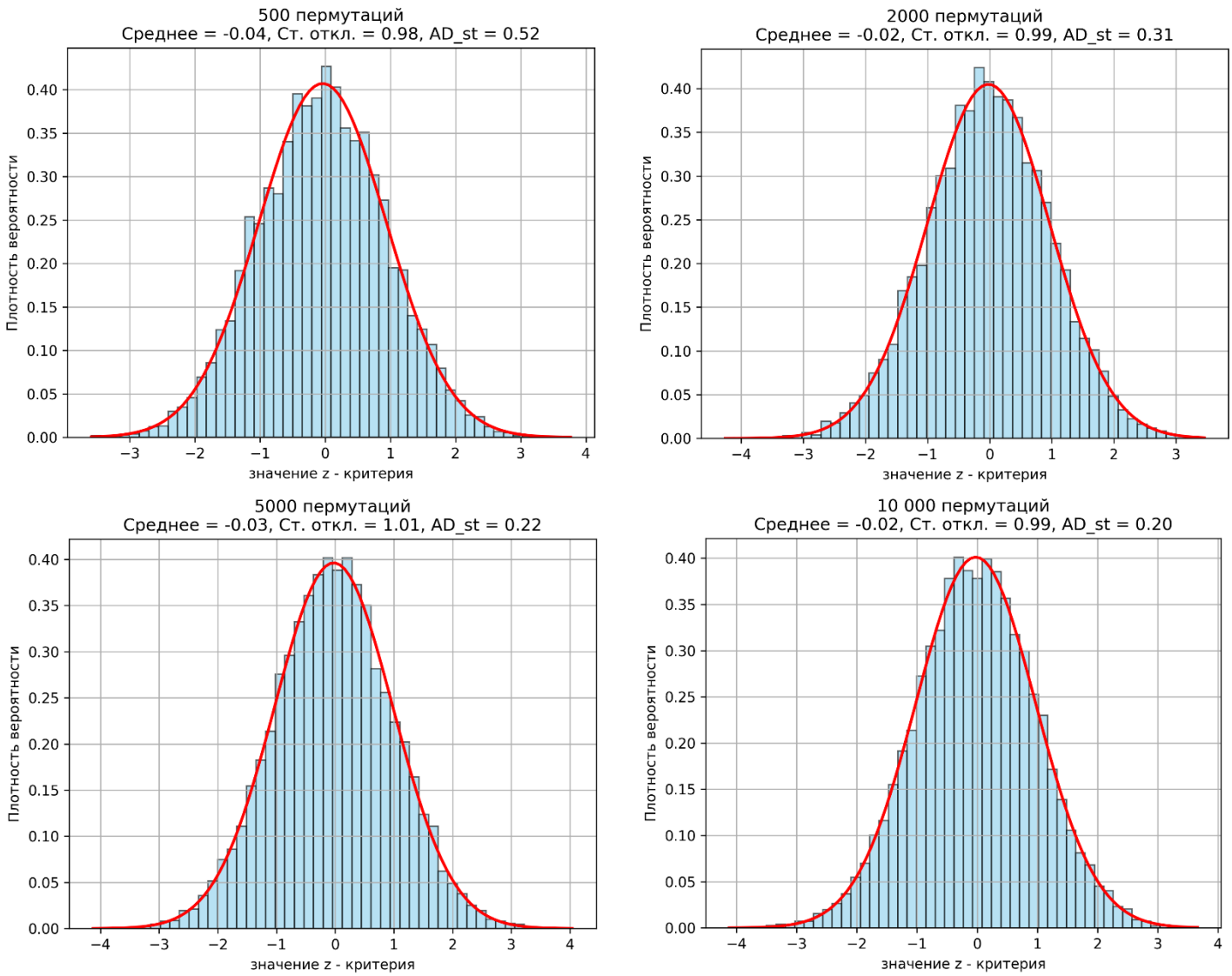

Рис.N: Гистограммы значений результатов теста при различном количестве итераций. Среднее значение во всех трех случаях остаётся близким к 0, тогда как критерий Андерсона-Дарлинга выполняется тем более точно, чем больше производится пермутаций.

Полученные распределения z-значений для разного количества пермутаций были протестированы на соответствие нормальному закону с помощью критерия Андерсона–Дарлинга. Во всех сериях экспериментов критерий не выявил значимых отклонений от нормальности. Среднее значение z стремилось к нулю, а увеличение числа итераций пермутаций приводило к лучшему соответствию теоретической нормальной кривой, что

подтверждает, как корректность самой гипотезы, так и устойчивость алгоритма к ложноположительным результатам при корректной параметризации.

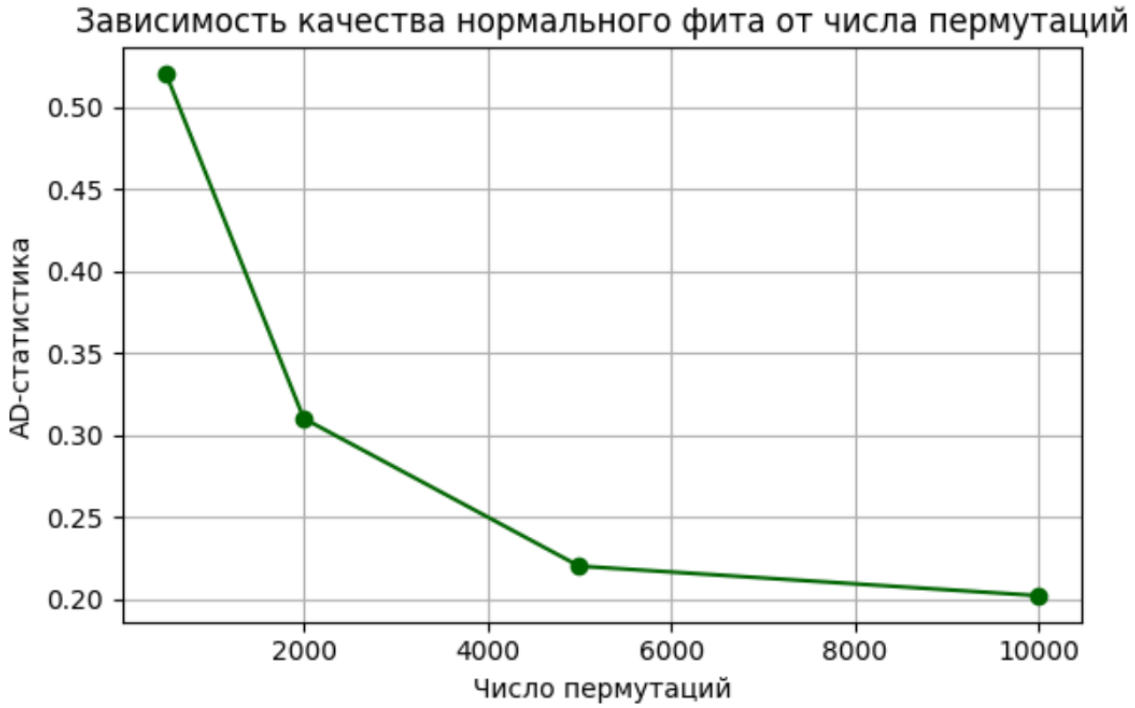

Рис.N: Зависимость качества нормального распределения от числа пермутаций.

По результатам этого теста были обнаружены 2 свойства методики:

Во – первых, улучшение качества фитирования гистограммы гауссианй с ростом числа пермутаций объясняется увеличением размера выборке, изображающей в тесте генеральную совокупность – в теоретическом пределе для этого требуется сделать поворотов не меньше, чем всего генов на хромосоме (если использовать именно повороты а не перемешивания). Однако важно отметить, что прирост качества распределения имеет убывающий характер: при переходе от нескольких сотен к тысячам пермутаций наблюдается резкое улучшение согласия с нормальным распределением, но после порога 5 – 7.5 тыс. поворотов (для одной хромосомы) дальнейшее увеличение числа пермутаций даёт лишь незначительное улучшение. Таким образом, бесконечное увеличение числа

пермутаций нецелесообразно с точки зрения вычислительной эффективности, и на практике далее будеет использоваться компромиссное значение в 10 тысяч пермутаций.

Во – вторых, визуально на гистограммах сохраняются незначительные отклонения эмпирических распределений от теоретической гауссианы, особенно в области малых значений z. Такие отклонения обусловлены, во-первых, структурой входных данных: существенная доля доменов содержит крайне малое количество генов (вплоть до одного или нуля), что делает невозможным какую-либо интерпретацию или выявление координированной экспрессии в таких регионах. К тому же, малые значения z не имеют практического значения в контексте гипотезы, поскольку лежат в зоне статистической незначимости ($|z| < 1.96$) и потому не влияют на расчёт доли ложноположительных находок. Таким образом, подобные отклонения в центральной части распределения z-критерия не подрывают корректность теста и соответствие результатов центральной гипотезе.

4.2 Положительный контроль

Для проведения положительного контроля была смоделирована ситуация с заведомо координированной экспрессией, собрав искусственную хромосому, состоящую исключительно из участков, соответствующих доменам TAD линии GM12878 https://www.hubrecht.eu/app/uploads/2018/02/Kind_Key_2015_Kind_Genome-wide-maps-of-nuclear-lamina-interactions-in-single-human-cells.pdf , аннотированных в реальных данных. Экспрессии генов в этих доменах были взяты без изменений из оригинальных данных (https://github.com/mdozmorov/HiC_data https://gtexportal.org/home/downloads/egtex/methylation), после чего такие фрагменты были сшиты в одну искусственную хромосому. Анализ показал, что z-распределение для 10 000

таких синтетических хромосом смещено в сторону положительных значений и хорошо аппроксимируется нормальным распределением с центром около 2.6 и дисперсией ~0.2.

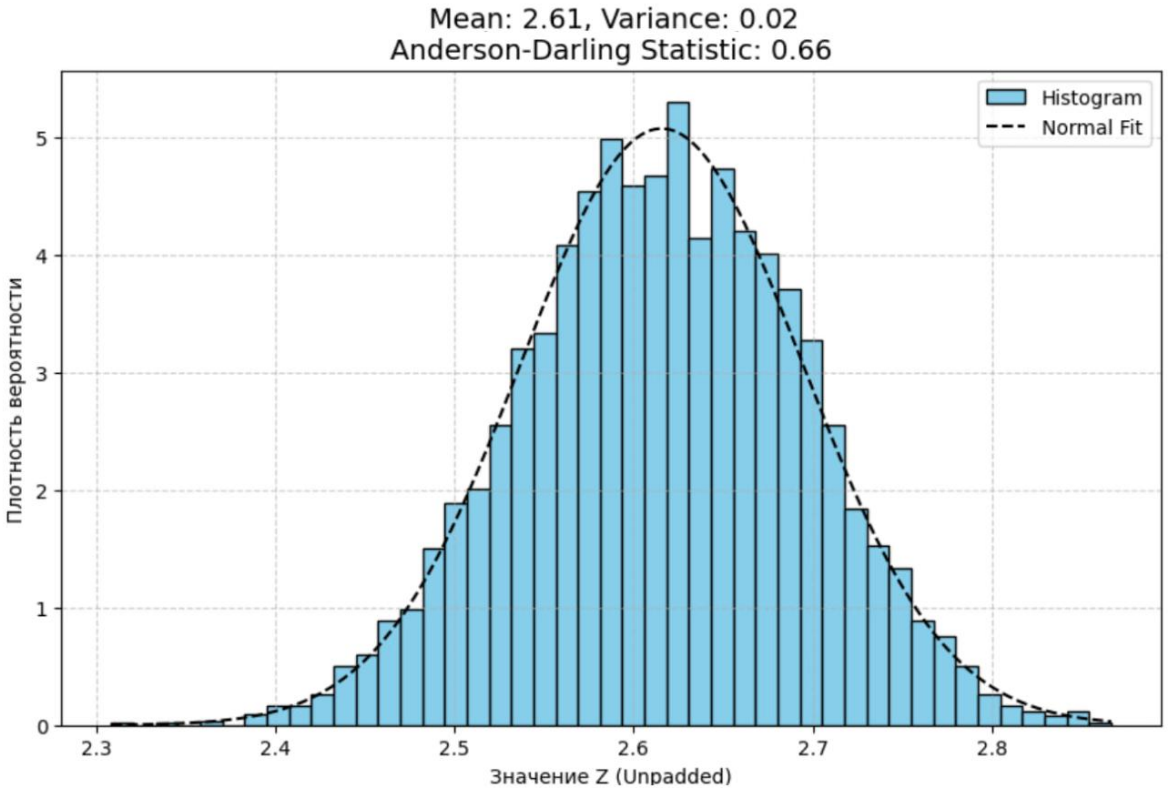

Рис.N: Гистограмма распределения результатов теста на хромосоме, составленной из топологически – ассоциированных доменов. Существенное отклонение среднего значения от 0 сигнализирует о способности теста отличать координированные участки от случайных.

Интерпретация конкретного значения смещения не является целью данного анализа, поскольку основная задача положительного контроля — убедиться, что в случае наличия реальной координации экспрессии тест реагирует отличимо от нулевой модели. Это подтверждает его чувствительность. Примечательно, что как активные (с высокими уровнями экспрессии, вплоть до 50), так и молчащие (почти нулевые) домены корректно классифицируются как координированные. Это важно, поскольку координированная экспрессия подразумевает согласованную регуляцию, которая может выражаться как в активации, так и в синхронном подавлении работы генов.

Несмотря на то что значение статистики Андерсона-Дарлинга (0.66) для положительного контроля оказалось чуть выше, чем для отрицательного (например, AD ≈ 0.2), это не является тревожным сигналом. Во-первых, тест AD оценивает согласие эмпирического распределения с теоретическим нормальным, и более высокое значение не обязательно означает "хуже" в смысле биологической интерпретации — оно всего лишь отражает, что данные более плотно сгруппированы около определённого значения (в данном случае $z \approx 2.6$), что как раз ожидаемо для положительного контроля. Во-вторых, синтетическая хромосома, составленная только из координированных доменов, нарушает предпосылку независимости наблюдений и, следовательно, ожидать точного соответствия результатов точному нормальному распределению нет причины.

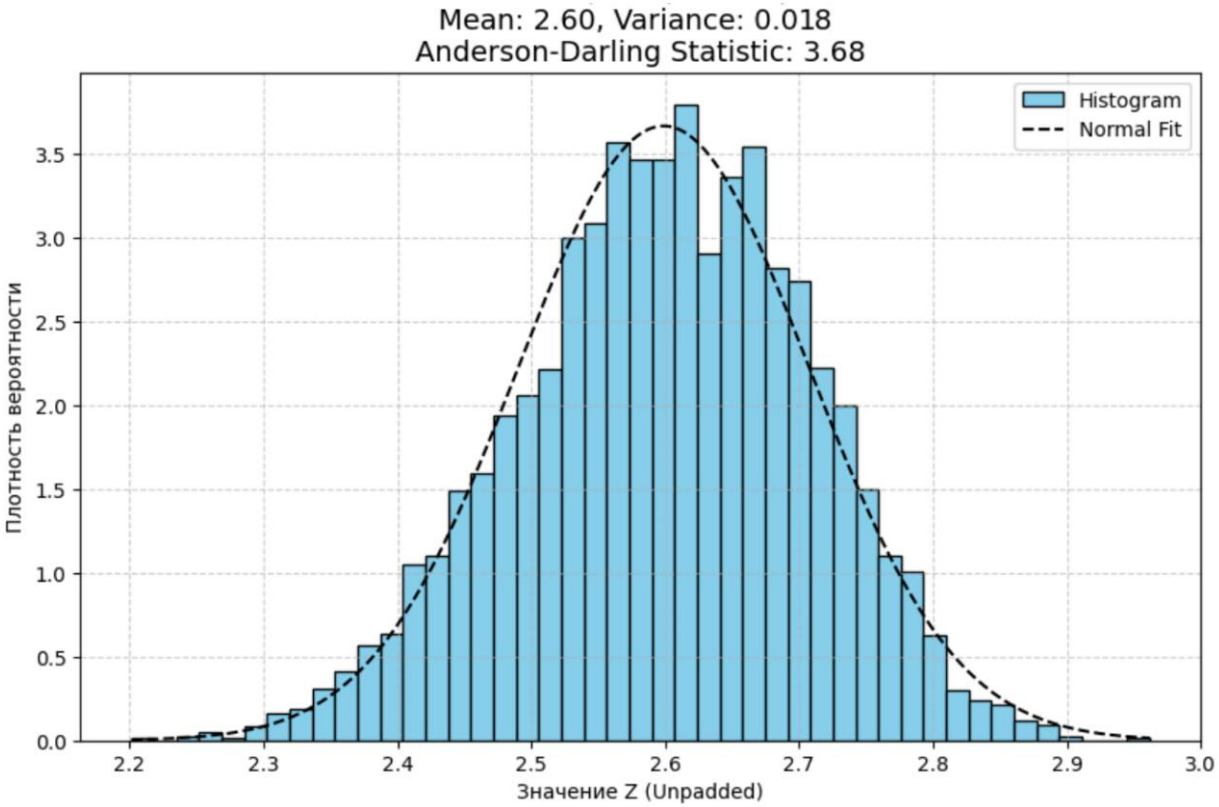

Рис.N: Гистограмма распределения результатов теста на модели полного генома, составленного из топологически – ассоциированных доменов.

Дополнительно мы протестировали работу алгоритма на ещё одном варианте положительного контроля — когда вместо одной синтетической хромосомы, составленной из TAD-доменов, был сгенерирован целый "геном", набрав последовательность из фрагментов, соответствующих TAD-доменам, до достижения общей длины, сопоставимой

с размером реального генома. В этом случае распределение z-критериев сохраняет отчётливое смещение в положительную сторону, но становится более шумным. Это абсолютно ожидаемо, так как увеличение длины приводит к большему разнообразию участков и выраженности индивидуальных флуктуаций. Однако ключевой результат сохраняется: z-значения стабильно смещены, а значит, тест продолжает успешно отличать координированные данные от случайных, даже в более сложных условиях.

3.3 Оценка динамического диапазона

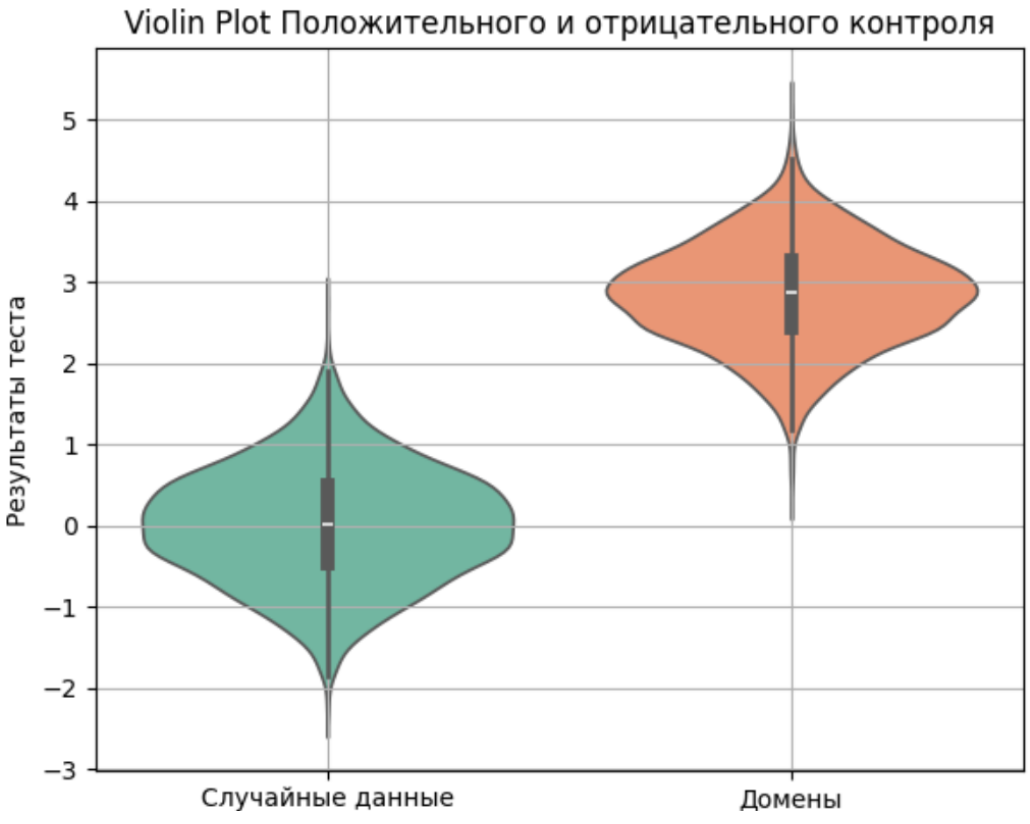

Рис.N: Violin plot сравнения положительного и отрицательного контроля. Малая область перекрытия гистограмм показывает, что относительное количество ложноположительных и ложноотрицательных результатов мало.

Для оценки динамического диапазона чувствительности теста была проведена визуализация и количественное сравнение двух распределений z-значений, соответствующих положительному и отрицательному контролям. Violin plot на Рис. N демонстрирует, что в подавляющем большинстве случаев значения статистики для доменов (положительный контроль) и случайных данных (отрицательный контроль) формируют два чётко разделённых пика. По результатам анализа, зона перекрытия между 95%-перцентильными интервалами двух распределений составляет менее 10% от общего количества данных. Это означает, что в подавляющем большинстве случаев тест способен уверенно отличить упорядоченные (координированные) домены от случайных, что подтверждает его высокую специфичность и пригодность для дальнейших биологических

интерпретаций.

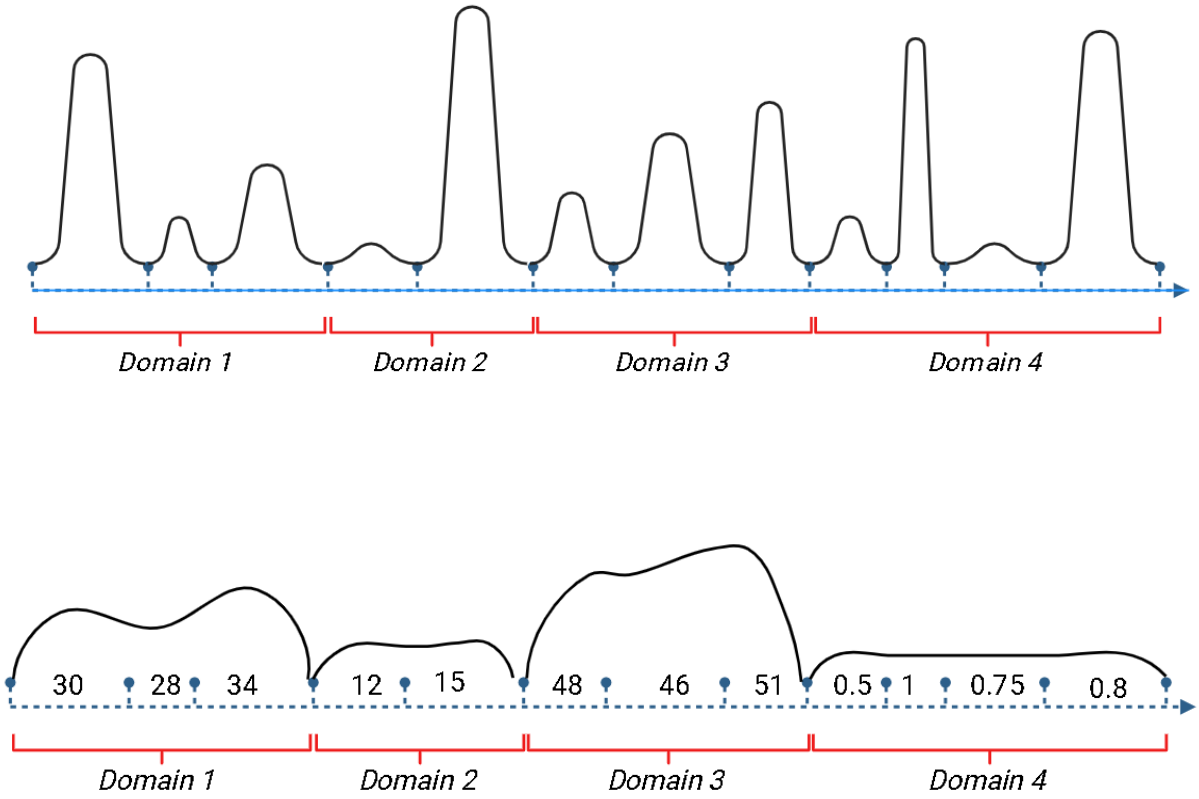

Рис.N: Иллюстрация: Модельные хромосомы с случайными (1) и псевдо координированными (2) значения экспрессии

4.3Оценка статистической устойчивости

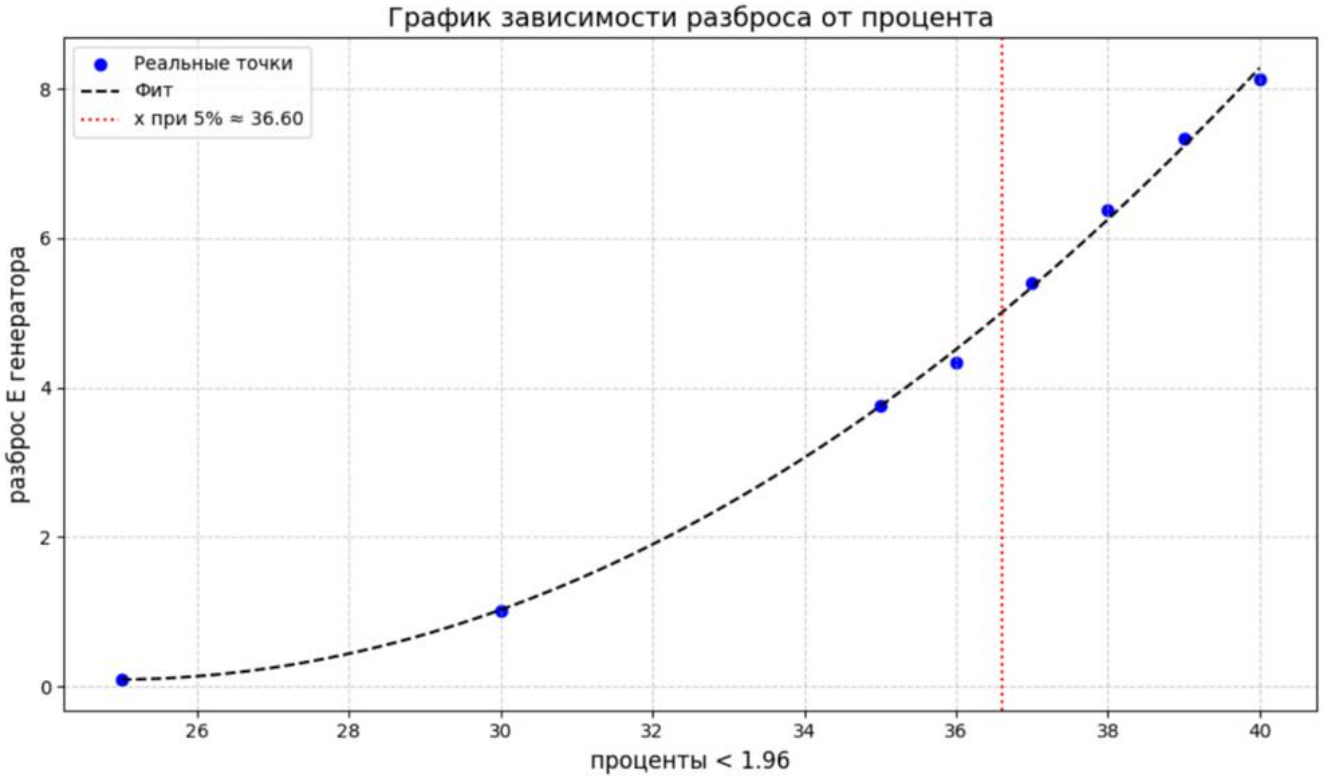

Рис.N: График зависимости количества ложноотрицательных результатов от амплитуды вносимого шума

Одним из ключевых требований к надёжности разработанного статистического теста является его устойчивость к экспериментальному шуму. Для количественной оценки этой устойчивости нами была проведена серия контрольных испытаний, в которых использовались заведомо координированные (положительные) хромосомы, синтезированные из аннотированных TAD/LAD. На следующем этапе в эти данные последовательно вносился аддитивный шум с равномерным распределением, и

фиксировалась доля доменов, которые переставали определяться тестом как статистически значимые. Таким образом строилась кривая зависимости количества ложноотрицательных результатов от амплитуды шума.

Результаты анализа показали, что при амплитуде шумов до ±17.8% от среднего значения TPM (что эквивалентно 36% от полного разброса значений в типичных доменах), точность теста остаётся выше 95%. Это означает, что методология демонстрирует высокую толерантность к погрешностям измерения экспрессии, что критично для работы с биологическими данными, неизбежно содержащими технологические и биологические вариации.

На основании этой устойчивости можно численно оценить минимальную необходимую глубину секвенирования, при которой сохраняется работоспособность алгоритма. Погрешность в ±17.8% означает, что дисперсия измерений TPM должна быть не более ~0.03×TPM². Для RNA-Seq на практике известно (https://link.springer.com/article/10.1186/s13059-016-0881-8), что при глубине секвенирования порядка 20–30 миллионов ридов на образец медианная погрешность определения TPM для умеренно экспрессирующихся генов составляет около 10–15%, (https://link.springer.com/article/10.1007/s11427-011-4255-x) то есть находится в пределах, допустимых для нашего теста. Более того, с учётом усреднения значений по доменам, которое само по себе снижает шум за счёт агрегации, становится очевидно, что тест способен работать при стандартных протоколах RNA-Seq, без необходимости экстремального увеличения глубины секвенирования. 20–30 миллионов ридов на образец в контексте bulk RNA-Seq — это нормальная

(https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0190152), стандартная глубина секвенирования, что сигнализирует о том, что для работы нашего алгоритма не обязательно тратить ресурсы на сверх глубокое секвенирование > 50 mr, что позволят не только технически упростить, но и несколько удешевить анализ доменных структур. Это делает его практически применимым даже в условиях ограниченных экспериментальных ресурсов и повышает воспроизводимость анализа.

4.6 Оптимизация и профилирование метода

На этапе оптимизации особое внимание уделялось наиболее ресурсоёмким участкам кода, связанным с оценкой координированной экспрессии по множеству доменов в хромосоме. Ключевым шагом стало отказ от использования стандартных питоновских контейнеров (таких как списки и словари) в пользу структурированных массивов NumPy (https://doi.org/10.1038/s41586-020-2649-2). Это решение позволило не только сократить объём потребляемой памяти, но и радикально ускорить вычисления: за счёт векторизации операций, доступа к данным по предсказуемым адресам и устранения накладных расходов, связанных с динамической типизацией Python-объектов. Особенно важным это оказалось в циклических операциях и при множественной фильтрации данных.

1. Для главной ресурсозатратной функции в математическом модуле, отвечающей за перебор доменов и расчёт статистик по ним, удалось в 9 раз улучшить соотношение “время/точность”. Хотя изменения асимптотического поведения добиться не удалось (осталась прежняя сложность по числу доменов), переход на более низкоуровневые подходы дал значительный

выигрыш в скорости. Одновременно с этим, модуль стал более гибким: ключевые параметры анализа (тип распределения, пороги, метод фильтрации, тип нормализации) теперь задаются извне, через конфигурационные файлы или аргументы командной строки, что повысило адаптивность к новым данным и сценариям.

2. Также была произведена адаптация всего кода под стандарты POSIX и GNU Coding Standards (https://doi.org/10.1016/j.infsof.2023.107299), чтобы упростить развёртывание и интеграцию в системах с UNIX-ориентированным пайплайном. Благодаря этому стало проще масштабировать расчёты на кластерах и в рамках HPC-задач.
3. Для распараллеливания задач реализована псевдопараллельная модель с использованием multiprocessing.Pool.map, при которой задачи делятся на независимые подзадачи по числу доступных ядер процессора. Это позволило многократно ускорить анализ больших наборов хромосом или симуляций.
4. Наконец, в планах — перевод кода на Python 3.14, после выхода стабильного релиза. Этот шаг откроет возможность использовать истинный многопоточный параллелизм, так как новая версия языка снимает ограничение GIL (Global Interpreter Lock) (https://docs.python.org/3.14/whatsnew/3.14.html), которое ранее мешало масштабировать задачи, где узким местом становится именно CPU, а не ввод-вывод.

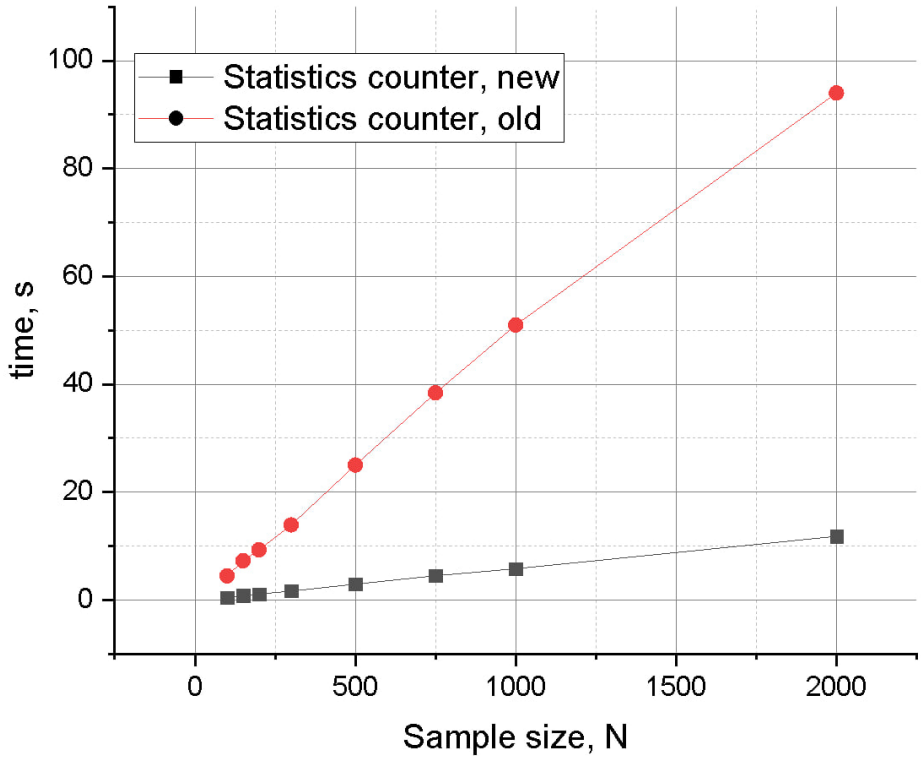

Рис.N: Графики зависимости времени расчёта координированности экспрессии от объёма выборки до (красн.) и после (черн.) оптимизации

Работы по оптимизации были проведены с расчётом на будущее развертывание инструмента как полноценного веб-ресурса, доступного каждому специалисту в области молекулярной биологии. В перспективе планируется, что любой исследователь сможет загрузить на сайт собственные данные по экспрессии и быстро получить оценку координированной экспрессии внутри доменов. Поэтому уже на этом этапе особое внимание уделяется масштабируемости, надёжности, стандартизации и производительности.

4.7 Проблема параметризации и соотнесения генов с доменами

Один из ключевых вопросов, с которым мы столкнулись при разработке алгоритма — это неоднозначность в том, как именно "приписывать" экспрессию гена конкретному домену. В литературе и практике пока нет устоявшегося консенсуса, и биологический смысл этого соответствия может варьироваться в зависимости от гипотезы, которую тестирует исследователь. Поэтому в системе реализованы четыре альтернативных подхода, каждый из которых можно выбрать как параметр анализа:

1. По началу транскрипции (TSS) — ген относится к тому домену, в котором находится его стартовая точка. Эта модель логична, если считать, что регуляция экспрессии в первую очередь происходит на этапе инициации (https://doi.org/10.1016/j.ydbio.2009.08.009), и что доменная структура влияет на доступность промоторов.
2. По концу транскрипции (TTS) — предполагается, что экспрессия определяется положением терминатора. Это может иметь смысл, если предположить, что доменная архитектура влияет на завершение транскрипции или на посттранскрипционные процессы (https://doi.org/10.1093/bib/bbaa389), включая стабильность РНК.
3. По середине гена — простой эвристический подход, который может быть оправдан, если распределение влияния домена на экспрессию равномерно по всей длине гена, и мы хотим избежать смещения в сторону длинных генов.
4. Пропорциональное распределение TPM — наиболее гибкий и биологически правдоподобный способ. В нём экспрессия гена дробится между доменами

пропорционально степени перекрытия, то есть сколько процентов длины гена лежит в пределах каждого домена. Этот подход особенно важен, если ген располагается на границе доменов или его регуляция может зависеть от нескольких структурных зон.

Каждый из этих способов имеет как биологическое, так и вычислительное обоснование, и поэтому мы предоставили пользователю возможность выбирать подход в зависимости от своих данных и задач. Пока что опытным путём было замечено, что пропорциональный метод даёт наиболее гладкие и устойчивые распределения выходных значений теста, особенно на реальных данных. Это может указывать на то, что эффект доменов на экспрессию действительно является непрерывным и размытым, а не дискретным и точечным.

Тем не менее, мы считаем важным предоставить исследователю свободу выбора, так как в разных типах клеток, организмах и при разных биологических задачах один и тот же способ может быть как уместным, так и вводящим в заблуждение.

4.8 Обратная задача: Поиск (предсказание) TAD по значениям экспрессии

Одним из самых интересных и перспективных направлений в нашей работе стал поиск границ TAD-доменов на основе только данных по экспрессии, без явной структуры 3D-хроматина. Это особенно актуально для тех случаев, когда отсутствуют экспериментальные Hi-C данные или они технически затруднительны в получении. Мы попытались подойти к задаче обратного анализа: можно ли по результату теста о координации экспрессии оценить, где потенциально находится структурный домен?

Для этого мы провели следующий моделирующий эксперимент:

- Берётся модельная хромосома — линейный массив значений TPM (экспрессии), сгенерированных случайным образом, без какой-либо пространственной координации между генами. Это имитирует фон в отсутствии доменной структуры.
- В произвольное место такой хромосомы вставляется искусственный TAD-домен — участок, в котором значения TPM сгенерированы так, чтобы быть скоррелированными между собой. Это можно интерпретировать как "вшитую" координированную экспрессию на фоне случайного шума.
- Далее, по этой хромосоме начинает двигаться рамка анализа — участок фиксированной длины, имитирующий окно, в пределах которого запускается тест на координированность экспрессии. Такое сканирование позволяет определить, как локальное значение теста изменяется при движении по хромосоме.

Результаты оказались весьма показательными. Когда рамка анализирует фоновую область, значения теста, как и ожидалось, колеблются около 0, так как координации нет. Однако, как только окно начинает "наползать" на границу TAD, значения теста начинают постепенно расти. При достижении максимального перекрытия с доменом — т.е. когда почти вся рамка лежит внутри TAD — значение теста достигает пика. А затем, когда рамка выходит с другой стороны, значения снова падают.

Если нанести эти результаты в виде графика "позиция рамки → значение теста", то на выходе получается характерный "колокол", в центре которого находится область с наибольшей координацией. Ширина этого колокола (оцененная по FWHM — ширине на половине высоты) грубо соответствует реальной длине вставленного домена. Таким образом, локальный максимум на этом графике указывает, где с высокой вероятностью

расположен TAD, несмотря на то, что ни одна явная граница нам изначально не была известна. Однако следует подчеркнуть ограничения этого подхода:

- - Поскольку размер окна анализа фиксирован, а значения теста отражают агрегированную информацию, метод не позволяет точно локализовать короткие домены (меньше половины ширины окна) — в этом случае пик получается слишком сглаженным или даже незаметным.
- - Аналогично, очень длинные домены могут быть недооценены, если окно не охватывает их целиком.

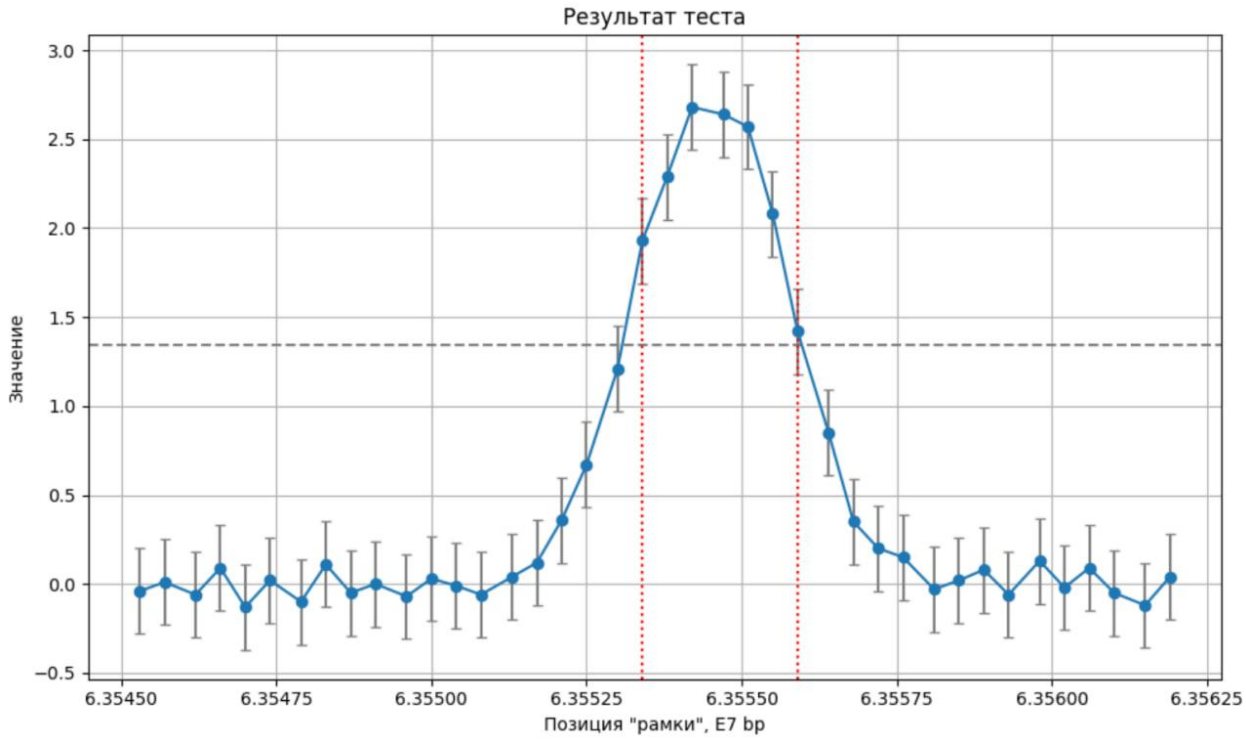

Рис.N: Масштабированный график результатов анализа в области хромосомы, оцениваемой как потенциально имеющая доменную структуру. Именно в этом регионе перед анализом были помещены значения экспрессии из TAD.

Тем не менее, даже в этих условиях метод даёт ценные результаты: он позволяет обнаружить "территории с подозрением на структурированность", без необходимости проводить дорогостоящее Hi-C секвенирование. На практике это даёт возможность:

- сначала провести первичное сканирование экспрессионных данных и отметить регионы с аномальной координацией;
- а затем прицельно проверить эти регионы другими методами (например, Capture Hi-C или ChIP-seq), сэкономив ресурсы и значительно сократив область интереса.

В перспективе такой подход может стать основой для алгоритмического предсказания TAD-доменов в плохо аннотированных геномах, особенно в нетипичных клеточных состояниях (например, при опухолевой трансформации), где структура хроматина нарушена, но при этом измерения экспрессии доступны.

4.9 Сравнение результатов предсказания с картой TAD на примере 15й хромосомы человека

Для демонстрации применимости нашего подхода на реальных биологических данных был проведён дополнительный эксперимент. В качестве базы структурной информации мы использовали публично доступные карты TAD-доменов, полученные методом Hi-C и аннотированные через TAD Map (https://cb.csail.mit.edu/tadmap/). Для экспрессионных данных была загружена таблица медианных значений TPM экспрессии из проекта GTEx (https://www.gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression), в частности, выборка по ткани префронтальной коры головного мозга. Эта ткань была выбрана исходя из предположения о потенциально высокой степени регуляторной специфичности и сложности координации экспрессии.

Для уменьшения объёма вычислений и лучшей визуальной интерпретации, анализ был сфокусирован на 15-й хромосоме человека. Эта хромосома характеризуется достаточно хорошим покрытием TAD-доменами (около 75% длины), при этом её компактный размер позволяет проводить быстрый анализ и строить наглядные графики.

Был выбран один усреднённый образец ткани (префронтальная кора), из которого извлечены значения экспрессии всех генов, расположенных на 15-й хромосоме. Параллельно были выделены соответствующие TAD-домены из Hi-C карты.

Затем для всей хромосомы был проведён анализ с помощью разработанного теста на наличие участков координированной экспрессии. Алгоритм скользящим окном оценивает уровень координации между генами в пределах окна и возвращает численное значение, соответствующее статистической значимости — чем выше значение, тем более

выраженной является гипотеза о наличии согласованной экспрессии в данной области. График ниже отображает фрагмент хромосомы (примерно 10% от полной длины), чтобы можно было рассмотреть визуальные детали соизмеримо с масштабом доменов.

На итоговом графике (см. ниже) были совместно отображены:

- оригинальные TAD-домены по Hi-C;

- результаты нашего анализа, нанесённые как кривая по координатам хромосомы.

Рис.N: График результатов анализа 15й хромосомы человека в сопоставлении с картой доменных структур из проекта TADMap.

Можно отметить, что во многих случаях пики значения теста хорошо совпадают с границами известных TAD-доменов. Особенно выраженное соответствие наблюдается в областях, где домены относительно широкие (средняя длина в этих случаях составляет более 600 кб) и содержат достаточно большое количество генов (в среднем ≥9). В этих

участках алгоритм демонстрирует положительные сигналы, что позволяет предположить наличие функционально согласованной экспрессии.

Однако метод пока хуже справляется с предсказанием мелких или плотно расположенных доменов: если два TAD идут вплотную друг к другу, часто возникает эффект "слипания" — метод объединяет их в один более широкий сигнал. Это объяснимо тем, что наш анализ оценивает не топологические границы, а функциональный след координации, который не всегда резко обрывается.

4.10 Неожиданное наблюдение о влиянии ткани

Во время проведения предыдущего эксперимента было обнаружено неожиданное, но потенциально значимое поведение алгоритма. При попытке провести аналогичный анализ, но уже на данных по кроветворным клеткам костного мозга, часть доменов, выявленных в нейронах, не была обнаружена. Это наблюдение пока не до конца понято и может объясняться:

1) либо действительно выраженными тканеспецифичными различиями в экспрессии генов, входящих в домены;
2) либо статистическими ограничениями, шумами или вариабельностью между выборками GTEx

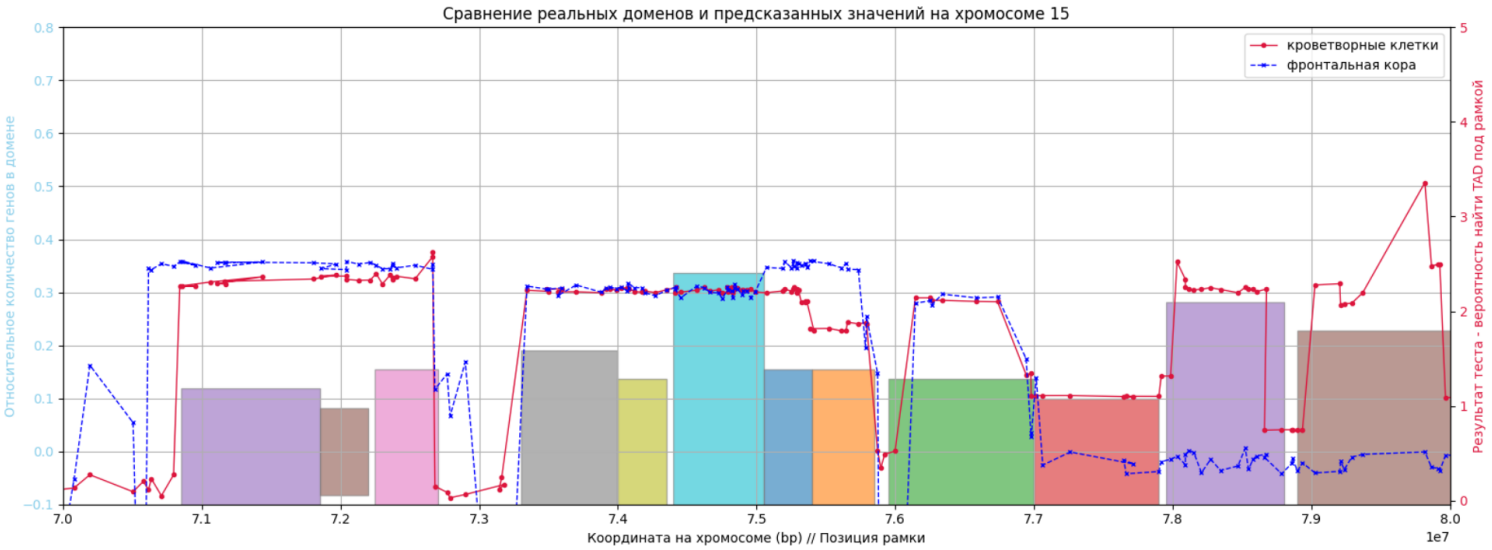

Рис.NСравнительный график результатов анализа кроветворной ткани (красн.) и нейронов (син.) на 15 хромосоме человека.

Тем не менее, данный факт указывает на возможность использования метода для сравнительного анализа разных тканей, а в перспективе — и для выявления тканеспецифичных функциональных доменов без привлечения структурных данных. Это может оказаться важным для поиска регуляторных элементов, изменяющих своё поведение между типами клеток или в ответ на внешние воздействия.

5. Выводы

В рамках данной работы была разработана, реализована и протестирована система анализа координированной экспрессии генов как косвенного индикатора пространственной организации хроматина. Ниже перечислены основные достижения и выводы по этапам исследования:

Был разработан статистический тест, оценивающий уровень координации экспрессии между генами, находящимися в пределах заданного доменного окна. Этот тест не требует знаний о пространственной структуре хромосомы и работает только с данными экспрессии. Его цель — выявить области, в которых гены "ведут себя согласованно", что может отражать влияние пространственной изоляции.

Сравнение заведомо некоординированных и координированных областей на синтетических данных с имитацией реальных распределений показало, что распределения почти не перекрываются, что говорит о высоком разделяющем потенциале теста. По доле перекрытия были количественно оцены чувствительность и специфичность теста.

Путём постепенного внесения шума в координированные данные было продемонстрировано, что точность метода сохраняется на уровне ≥95% до тех пор, пока разброс значений TPM не превышает ~±17.8%. Это позволило сделать оценку минимальной необходимой глубины секвенирования — около 20–30 млн ридов на образец — для устойчивой работы алгоритма, что делает метод практически применимым в реальных условиях.

В ходе разработки была проведена серьёзная оптимизация конечного исследовательского иснтрумента, что обеспечило значительный прирост скорости анализа.

Также реализована поддержка мультипроцессинга с ориентацией на последующую адаптацию под true multithreading после выхода Python 3.14. Это особенно важно, поскольку в будущем из инструмента планируется сделать веб-сервис, доступный для молекулярных биологов без навыков программирования — с возможностью загрузки собственных измерений и получения интерпретируемых результатов.

Ввиду отсутствия общепринятого консенсуса относительно того, как относить ген к домену (по старту, по концу, по центру или пропорционально перекрытию), мы реализовали все четыре способа как настраиваемые параметры. Каждый из них имеет собственные молекулярные обоснования (напр., транскрипционные стартовые сайты или транскрипционные терминаторы как ключевые регуляторные точки). Это делает наш инструмент гибким и пригодным для разных исследовательских задач. В практическом применении наиболее стабильные результаты дал пропорциональный способ, обеспечивающий гладкость распределений и меньшую чувствительность к крайним значениям.

Мы показали, что наш метод можно использовать не только для подтверждения наличия координации, но и для поиска предполагаемых TAD-доменов. Путём вставки искусственного домена в фоновую хромосому и сканирования её рамкой анализа мы обнаружили, что карта значений теста формирует "колокол" в области домена. Ширина пика примерно соответствует длине домена, позволяя грубо локализовать области с потенциальной топологической изоляцией. Это открывает возможность неинвазивного и недорогого предсказания TAD в условиях, где нет данных Hi-C.

Полученный инструмент демонстрирует как высокую чувствительность к координации экспрессии, так и потенциал в структурной биоинформатике — от фильтрации TAD-кандидатов до оптимизации экспериментов. Его высокая настраиваемость, устойчивость к шуму и техническая адаптация под большие объёмы данных делают его перспективной основой для онлайн-сервиса в помощь экспериментальным молекулярным биологам.

Наша работа является шагом к "экспресс-диагностике" 3D-хроматина по экспрессионным данным и, как мы надеемся, поможет в ускорении и удешевлении эпигеномных исследований.

6.Список литературы

7.Приложения

П.1 Описательный псевдокод пермутативного алгоритма

```
ALGORITHM Evaluate_Coordinated_Expression(Domens, Genes, Expressions, Iterations):

    MAPPING_METHOD ← get_mapping_method()

    IF MAPPING_METHOD ∉ {middle, proportional, start, stop} THEN

        RAISE error

    GENE_LENGTHS ← array of gene end - start for each gene in Genes

    MEAN_GENE_LENGTH ← average(GENE_LENGTHS)=

    ORIGINAL_DOMAINS ← map_genes_to_domains(Domens, Genes, Expressions, MAPPING_METHOD)

    DOMAIN_LENGTHS ← array of lengths of each domain in ORIGINAL_DOMAINS


    (μ₁, σ₁²) ← compute_mean_variance(ORIGINAL_DOMAINS)

    AVERAGE_LENGTH ← average(DOMAIN_LENGTHS)

    INITIALIZE:

        μ_sum ← 0

        σ²_sum ← 0

        valid_count ← 0

        N ← number of domains

        vertical_sum[N] ← 0

        vertical_sumsq[N] ← 0

        vertical_count[N] ← 0

    ROTATED_DOMAINS ← ORIGINAL_DOMAINS

    FOR i FROM 1 TO Iterations DO:

        ROTATED_DOMAINS ← rotate_gene_expression(ROTATED_DOMAINS)

        (μᵢ, σᵢ²) ← compute_mean_variance(ROTATED_DOMAINS)

        IF μᵢ IS DEFINED THEN

            μ_sum ← μ_sum + μᵢ
```

```
            σ²_sum ← σ²_sum + σᵢ²
            valid_count ← valid_count + 1
        END IF
        FOR j FROM 1 TO N DO:
            CLEAN_EXPR ← remove_NaN(ROTATED_DOMAINS[j])
            n ← length(CLEAN_EXPR)
            IF n > 0 THEN
                vertical_sum[j] ← vertical_sum[j] + sum(CLEAN_EXPR)
                vertical_sumsq[j] ← vertical_sumsq[j] + sum_of_squares(CLEAN_EXPR)
                vertical_count[j] ← vertical_count[j] + n
            END IF
        END FOR


    END FOR
    μ₂ ← μ_sum / valid_count
    σ₂² ← σ²_sum / valid_count
    z ← Z_CRITERION(μ₁, μ₂, σ₁², σ₂², total_gene_count = length(Genes)
    FOR j FROM 1 TO N DO:
        IF vertical_count[j] > 0 THEN
            mean[j] ← vertical_sum[j] / vertical_count[j]
            var[j] ← (vertical_sumsq[j] / vertical_count[j]) - mean[j]^2
        ELSE
            mean[j] ← 0; var[j] ← 0
        END IF
    END FOR
    μ_vert ← average(mean[1..N])
    σ_vert² ← average(var[1..N])
    RETURN [z, μ₁, σ₁², μ₂, σ₂², μ_vert, σ_vert²]
END ALGORITHM
```