

ОБЗОРНЫЕ
И ТЕОРЕТИЧЕСКИЕ СТАТЬИ

УДК 57.088.1

МЕТОДЫ НОРМАЛИЗАЦИИ ДАННЫХ ChIP-seq И ИХ ПРИМЕНЕНИЕ
В ИССЛЕДОВАНИЯХ РЕГУЛЯТОРНЫХ УЧАСТКОВ
ГЕНОМА КЛЕТОК МОЗГА

© 2023 г. Ф. Е. Гусев^{1, 2, *}, Т. В. Андреева^{1, 2, **}, Е. И. Рогаев^{1, 2, ***}

¹Институт общей генетики им. Н.И. Вавилова Российской академии наук, Москва, 119991 Россия

²Научный центр генетики и науки о жизни, Университет “Сириус”, Краснодарский край, пгт. Сириус, 354340 Россия

*e-mail: gusev@vigg.ru

**e-mail: an_tati@vigg.ru

***e-mail: evivrog@gmail.com

Поступила в редакцию 14.12.2022 г.

После доработки 13.03.2023 г.

Принята к публикации 22.03.2023 г.

За последние годы метод иммунопреципитации хроматина с последующим глубоким секвенированием (ChIP-seq) стал одним из основных инструментов для исследования регуляции экспрессии генов. Как и другие способы молекулярного профилирования, ChIP-seq имеет ряд методических особенностей, которые могут оказывать нежелательный эффект на получаемые результаты, особенно в случаях, когда используются образцы клеток и тканей, качество которых сложно контролировать, например замороженные постмортальные образцы тканей мозга человека. Однако методы биоинформатического анализа совершенствуются с каждым годом и позволяют уменьшить эти эффекты на этапе анализа полученных данных секвенирования и позволяют сделать поправки (нормализовать данные) на неравномерность как технических особенностей ChIP-seq, так и в более общем смысле различных факторов исследований, например постмортальный интервал или гетерогенность клеточного состава исследуемых образцов. В этом обзоре мы рассмотрели широкий спектр предложенных методов нормализации данных ChIP-seq, особенности их применения и выбора в конкретном исследовании, в том числе в экспериментах с образцами клеток мозга человека. Представлены преимущества и недостатки существующих подходов к нормализации, а также примеры, свидетельствующие о перспективности использования методов ChIP-seq при исследовании мозга.

Ключевые слова: геномика, эпигеномика, хроматин, нормализация, иммунопреципитация, мозг.

DOI: 10.31857/S0016675823080088, **EDN:** XTNLER

Геном любого живого организма помимо самой генетической последовательности имеет также ряд факторов, которые характеризуют его работу в клетках. В клетках ДНК упакована в структуру, называемую хроматином, и при этом намотана на белковые комплексы, называемые нукleosомами, которые в свою очередь состоят из гистонов. Различные химические модификации гистонов, такие как метилирование или ацетилирование отдельных оснований, коррелируют с функцией соответствующих участков генома [1]. Суммарно такие изменения гистонов и другие химические изменения ДНК называются эпигеномом [2]. За последние годы подход, основанный на иммунопреципитации хроматина с последующим секвенированием (ChIP-seq), стал одним из основных методов исследования эпигенетического состояния генома. Он состоит из двух последовательных этапов: 1) выделение белков, связанных с ДНК, с

помощью специфических антител и 2) определение последовательностей выделенной ДНК. Картировав затем полученные последовательности на референсную последовательность генома исследуемого организма, можно определить связана ли ДНК каждого участка генома с исследуемым белком или нет, что можно использовать как для изучения транскрипционных факторов, так и для модификаций гистонов [3].

Как и многие другие методы количественных исследований в геномике, ChIP-seq используется в том числе и для сравнительного анализа нескольких образцов и сталкивается с теми же проблемами, что и, например, анализ экспрессии генов с помощью гибридизации на микрочип или секвенирования транскриптома (RNA-seq): наличие так называемой технической вариабельности при получении данных [4, 5]. Наблюдаемое в геномных профилях разнообразие условно делит-

ся на две части: биологическое и техническое. Биологическое разнообразие отражает действительные различия между особями (образцами) в живой природе, а техническое разнообразие отражает дополнительные эффекты, которые добавляются во время измерения. Самым очевидным таким эффектом является глубина секвенирования (т.е. суммарное количество прочтений, полученных для конкретного образца) [6]. Кроме этого, при подготовке библиотек секвенирования количество циклов амплификации может быть различным в зависимости от количества доступной ДНК, что в конечном итоге повлияет на измеренные уровни связанности белков с ДНК при выполнении ChIP-seq [7]. Существуют факторы, которые более сложно контролировать, например при подготовке партии образцов работы были разбиты на несколько дней, но в каждый из дней средняя температура воздуха в лаборатории была различной, что повлияло на скорость химических реакций и в конце концов внесло вклад в полученные данные [8]. При проведении ChIP-seq одним из основных таких факторов является неравномерная эффективность работы антител [9]. В той или иной степени эффективность всегда неполная, и это проявляется в появлении так называемого геномного шума, который выглядит как короткие прочтения, картированные не в локусы целевого обогащения (рис. 1,а). Например, для антител, специфичных к H3K4me3, который является маркером активных промоторов генов [10], почти всегда наблюдается картирование данных ChIP-seq в другие участки генома. Почти всегда неспецифические последовательности равномерно распределены по референсной последовательности [11] и эта равномерность является одним из главных условий возможности нормализации данных. Помимо технических аспектов применения различных методик получения и сравнения молекулярных профилей, исследования клеток мозга (особенно человека) имеют дополнительные аспекты, которые влияют на проведение достоверного анализа, в частности влияние на эпигенетический профиль различных аспектов жизнедеятельности человека (например, возраста или курения, этнической принадлежности и пр.), особенно наличие большого количества различных типов клеток в одном образце, которые опосредованно влияют на статистический анализ исследуемых фенотипов, уменьшая возможности получения достоверных результатов [5, 12].

НОРМАЛИЗАЦИЯ СПЕЦИФИЧНОСТИ СИГНАЛА ChIP-seq

Почти все методы для нормализации данных ChIP-seq, описанные ниже, опираются на первоначальное построение матрицы, описывающей полногеномное эпигенетическое состояние ис-

следуемых образцов. Для этого весь геном обычно разбивается последовательно на окна определенной длины (пересекающиеся друг с другом или нет). При этом длина окна зависит от изучаемого целевого белка [13]: для транскрипционных факторов, для посадки которых на ДНК обычно требуется 10–20 нуклеотидов, используются окна, длиной 50–100 нуклеотидов; для анализа модификаций гистонов используют окна, длина которых соответствует размеру ДНК, намотанной на нуклеосому (т.е. около 150 нуклеотидов). Для некоторых модификаций гистонов характерна более широкая локализация — в этих случаях используются окна длиной до 1000 нуклеотидов. Затем для каждого образца в каждом окне вычисляется количество последовательностей, картированных в конкретное окно, и из этих величин формируется матрица, которая и является основой для последующей нормализации и дифференциального анализа (рис. 1,б, в). Следует отметить, что многие методы были разработаны для исследования эпигенома клеточных культур на небольшом количестве образцов и программные пакеты не всегда адаптированы для исследований больших выборок, как этого требует сравнительный анализ, например, при исследовании болезней человека.

Нормализация на глубину секвенирования

Один из очевидных факторов, влияющих на возможность корректного сравнения между образцами, является количество прочтений, полученных после глубокого секвенирования, поскольку эта величина слабо зависит от исходного материала, но определяется, во-первых, стоимостью секвенирования и, во-вторых, сложностью эквимолярного смешивания библиотек и тем, как работают приборы для секвенирования. Поэтому один из самых простых методов нормализации данных ChIP-seq (и многих других данных глубокого секвенирования) является нормализация измерения количества прочтений, картированных в конкретный локус (например, ген или пик обогащения), к общему количеству прочтений, полученных для образца [13–15]. В этом случае элементы исходной матрицы делятся на сумму величин в каждой строке матрицы (т.е. на суммарное количество прочтений конкретного образца, рис. 1,б).

TMM-нормализация

Другим подходом, который получил первоначальное развитие в анализе экспрессии генов, является нормализация образцов, опирающаяся на идею того, что большая часть локусов не должна иметь значительных изменений в одних образцах по сравнению с другими. Исходя из этого, при анализе нескольких образцов сначала выбирается случайно один из образцов, называемый референс-

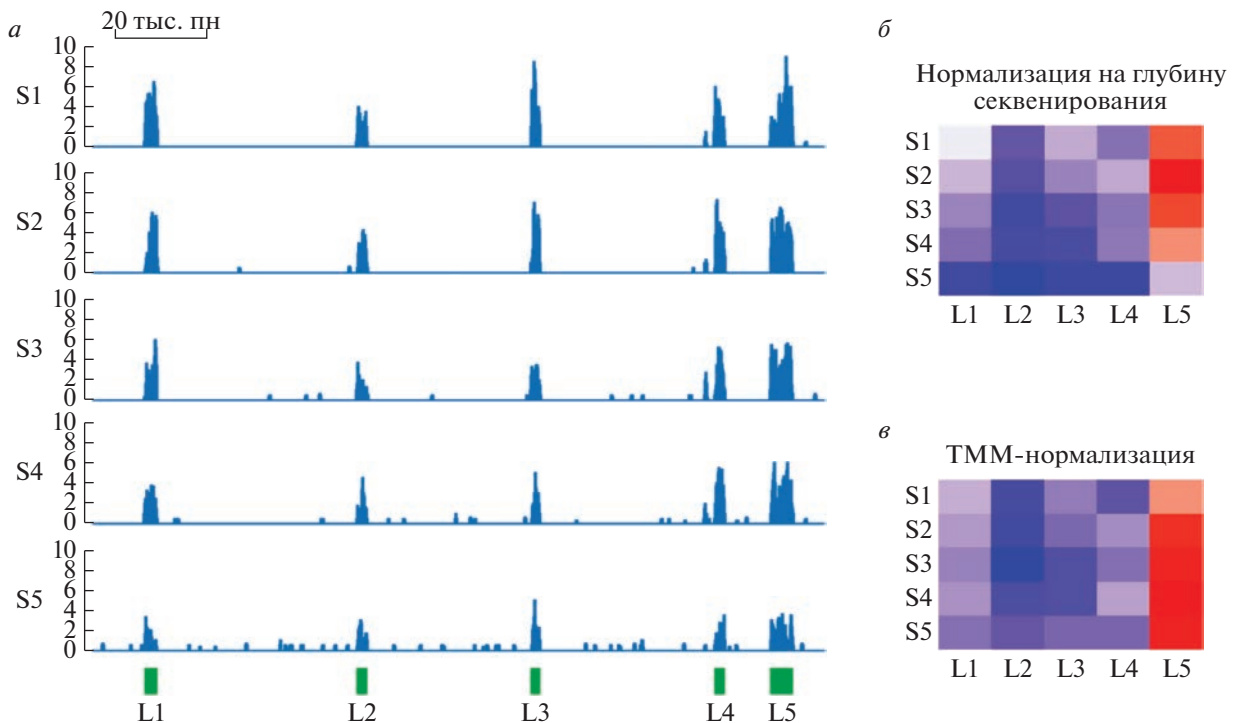


Рис. 1. Демонстрация двух методов нормализации. На панели *a* приведены пять симулированных профилей ChIP-seq. При этом в симуляции использован один и тот же сигнал, но с разным соотношением сигнала и шума (отражая различную степень эффективности антител в реальном эксперименте), поэтому ожидается, что все пять профилей должны быть похожи друг на друга после нормализации. Зелеными прямоугольниками выделены симулированные локусы обогащения. На панелях *б* и *в* изображены результаты двух методов нормализации этих профилей: строки соответствуют образцам панели *a*, а столбцы — локусам обогащения, цвет характеризует оценку сигнала. На панели *б* изображены результаты нормализации на суммарную глубину секвенирования, а на панели *в* — результаты TMM-нормализации. Этот пример демонстрирует преимущество TMM-нормализации над нормализацией на глубину секвенирования в данной ситуации, поскольку на панели *б* наблюдается зависимость нормализованных значений от соотношения сигнал/шум в профиле. Следует отметить, что обе нормализации не предполагают нормализации значений между локусами обогащения, а предполагают только нормализацию между различными профилями.

ным, и остальные нормализуются относительно него. Для каждого вычисляется обрезанное среднее М-величин (trimmed mean of M-values, TMM), которое равно взвешенному среднему М-величин (логарифм разности плотности прочтений между двумя образцами), исключая локусы, в которых М-величина и А-величина (половина логарифма суммы плотности прочтений в двух образцах) принимают экстремальные значения (определенные по умолчанию как 30- и 10%-ные квантили соответственно). Этот подход успешно применяется как для транскриптомных данных, для которых был изначально разработан, так и для других геномных данных, включая ChIP-seq [13, 16]. Однако, как отмечают авторы программного пакета *csaw* [13] для сравнительного анализа ChIP-seq данных, в некоторых случаях исходная посылка метода TMM не выполняется (например, когда сравнивают образцы, в которых происходит значительная перестройка эпигенетического статуса генома) и предлагают в таких случаях использовать только локусы, в которых наблюдается высо-

кое связывание целевого белка с ДНК. Однако и метод нормализации на глубину секвенирования, и TMM-нормализация не моделируют напрямую основного фактора, специфического для ChIP-seq: вероятности неравной эффективности антитела в исследуемых образцах, что приводит к отличающимся соотношениям сигнал-шум и может существенно влиять на анализ.

CisGenome, SPP, CCAT и NCIS

Как отмечают авторы программного пакета *csaw* [13], метод нормализации TMM, очень успешно применяемый для анализа данных RNA-seq, не всегда подходит для анализа ChIP-seq. Существует целое семейство методов, разработанных для нормализации данных ChIP-seq, основанных на исключении части локусов из анализа, что позволяет более эффективно использовать метод нормализации на глубину секвенирования или TMM-нормализацию. Каждый из этих методов исключает из общего числа некоторые локусы, основываясь на

определенном критерии. После этого коэффициент нормализации вычисляется одним из вышеописанных способов, используя только картированные в оставшиеся участки генома последовательности.

В методе, предложенном авторами программного пакета CisGenome [17], предлагают использовать только участки генома, в которых суммарно обнаружено малое количество прочтений, тем самым нормализуя уровень шума между образцами. Авторы метода SPP используют распределение Пуассона, чтобы оценить количество прочтений, картированных в локус, и используют для нормализации только те, в которых у всех исследуемых образцов P -значение меньше 0.05 [18].

В программном пакете CCAT [19] для определения локусов для оценки коэффициента нормализации используется итерационный процесс, в котором на первом этапе используется нормализация с помощью глубины секвенирования. Затем на каждом шаге для анализа используются только регионы генома, в которых различается количество последовательностей в образцах, картированных только на прямую цепь референсного генома, а оценка коэффициента нормализации обновляется с помощью последовательностей, картированных только на комплементарную цепь референсной последовательности. Этот процесс повторяется пока коэффициент нормализации не перестает изменяться.

Во всех этих подходах используются заранее заданные критерии размера локусов, количества последовательностей, ограничение P -значения или другие. Авторы метода NCIS [20] развивают эти подходы, но предлагают алгоритм, адаптивный к исходным данным, для выбора размера локусов и ограничения на суммарное количество прочтений в образцах. Они отмечают, что выбор ограничения на количество прочтений позволяет маневрировать между смещением оценки коэффициента нормализации и его дисперсией. Анализируя данные ChIP-seq, они строят график данных разницы логарифмов количества прочтений в двух образцах относительно суммарного количества последовательностей в локусе (своеобразный аналог МА-графика) и находят на нем точку, разделяющую локусы с обогащением от локусов, содержащих только шум. Аналогичным образом получают оценку необходимого размера регионов генома. На основании этих двух величин вычисляется коэффициент нормализации с помощью метода, аналогичного методу CisGenome.

MANorm & MANorm2

Метод MANorm [21] применяется для нормализации пары образцов относительно друг друга исходя из того, что если в двух различных образцах много общих локусов связывания с целевым

белком то в этих регионах связывание должно определяться общими механизмами. Исходя из этого, авторы развивают идею нормализации, которую успешно применяют для нормализации данных экспрессии генов с помощью графика Бланта—Альтмана, который еще называют МА-графиком. В методе MANorm в этом графике на оси абсцисс нанесена логарифмированная средняя плотность прочтений в двух образцах (A), а на оси ординат — логарифмированная разница в плотности прочтений между двумя образцами (M), но для построения используются только те участки генома, в которых у каждого из образцов наблюдается обогащение. На основании такого графика определяется устойчивая линейная модель, связывающая среднюю плотность последовательностей с разницей в этой плотности. Исходя из данной линейной модели, производится трансформация пространства значений (M, A) таким образом, чтобы ось абсцисс совпадала с полученной линейной моделью. Рассматривая теперь значения M и A после трансформации во всех пиках, а не только общих, можно получить нормализованные значения M и A, т.е. нормализованные значения разницы между двумя образцами. Таким образом, метод MANorm позволяет провести сравнение двух образцов с учетом нормализации, но не позволяет получить нормализованные профили ChIP-seq.

Метод MANorm2 [22] является развитием метода MANorm, позволяя применить похожий подход не для сравнения двух образцов, а для сравнения двух групп образцов. В частности, авторы предлагают учесть наличие потенциальной структуры (гетерогенности образцов) и для этого на одном из первых этапов отдельно нормализуют группы. Внутри группы образцов авторы проводят нормализацию путем линейной трансформации M- и A-величин таким образом, чтобы среднее M-величин было нулевым и ковариация M- и A-величин тоже была нулевой.

ChIPnorm

Авторы подхода ChIPnorm [23] предлагают похожий на MANorm метод, но используют два последовательных шага. На первом этапе определяются участки генома, в которых наблюдается значимое обогащение последовательностей в двух анализируемых образцах по сравнению с контрольной библиотекой и со случайным распределением. На втором этапе проводится нормализация двух образцов на основании регионов, которые имели значимое обогащение в обоих образцах на первом этапе (т.е. исходя из того же принципа, что в методе MANorm, но сами регионы выбираются по другому принципу). В предлагаемом авторами программном пакете также реализованы некоторые другие методы нормализации: нормализация к единице и квантильная нормализация, которые

похожи на нормализацию на глубину секвенирования или ТММ-нормализацию и были ранее разработаны для анализа данных экспрессии генов с помощью микроматриц.

CHIPIN

Авторы программного пакета для языка программирования R под названием CHIPIN (ChIP-seq Inter-sample Normalization) [24] предлагают подход для нормализации данных ChIP-seq, который можно использовать в тех случаях, когда не используются калибровочный контроль, но зато доступны данные по экспрессии генов. Авторы предлагают нормализовать сигнал ChIP-seq между различными образцами исходя из предположения, что для генов, экспрессия которых постоянна, регуляторные участки также должны иметь одинаковую активность.

Исходя из этого, авторы предлагают проводить нормализацию в два этапа: на первом определяются локусы генов, имеющих неизменную экспрессию, а на втором — линейная или квантильная нормализация определяется так, чтобы разница в локусах, пересекающихся с локусами “неизменных” генов, была минимальной.

Оценивая качество нормализации на основании анализа средних значений нормализованных профилей покрытия в локусах начала транскрипции, авторы, помимо нормализации предлагают использовать этот метод для оценки качества работы (специфичности) антител, используемых для ChIP-seq. На примере использования двух различных антител (против H3K27ac и против H3K27me3) в клеточных линиях авторы демонстрируют, что при использовании первого антитела средний профиль плотности прочтений в окрестности сайтов начала транскрипции не зависит от степени экспрессии гена, а для второго антитела профили генов с высокой и средней степенью экспрессии похожи друг на друга и на профили, полученные для первого, но сильно отличаются от профилей для генов с низкой экспрессией. Это подтверждает ранее опубликованные результаты о том, что второе антитело связывается не только с метилированными лизинами, но и с ацетилированными, т.е. не является специфичным.

Идейно близкий подход также реализован в программном пакете THOR [25], однако в нем в качестве “постоянных” генов используются так называемые гены домашнего хозяйства (т.е. гены необходимы в клетках для поддержания базовой жизнеспособности).

Калибровочный контроль

Природа данных глубокого секвенирования предполагает исследование относительных вели-

чин: абсолютные значения крайне трудно измерить. Однако в ряде случаев происходят глобальные изменения в структуре хроматина (например, в случае мутации в гистоне), для исследования которого сравнительного анализа недостаточно. Для того чтобы решить эту проблему, авторы метода ChIP-Rx [9] разработали подход, позволяющий учитывать количество клеток, использованных в исследовании. Вдохновляясь аналогичными подходами в исследованиях транскриптомики [26, 27], они предлагают добавлять в каждый исследуемый образец внешний калибровочный контроль в отношении, постоянном между образцами. При этом в качестве дополнения используется материал от другого биологического вида. Например, реализуя эту стратегию на иммортализованной человеческой клеточной линии Jukart, авторы добавляют в образцы клетки плодовой мушки в пропорции две клетки мухи к одной клетке человека. Гистоны являются одними из наиболее эволюционно консервативных белков и поэтому антитела, вероятно, будут одинаково эффективно работать и для млекопитающих, и для плодовой мушки.

Выполнив иммунопреципитацию и последующее секвенирование, можно затем в каждом образце картировать последовательности и на геном человека, и на геном плодовой мушки. Затем определить отношение количества прочтений на два генома и использовать эту величину для нормализации.

В другой похожем подходе [28] предлагается использовать комбинацию экзогенного хроматина плодовой мушки и специфического для него антитела, тем самым ослабляя требование универсальности используемого антитела. Более того, авторы предполагают, что такой подход легче в использовании, потому что позволяет заранее подготовить большое количество экзогенного материала до проведения иммунопреципитации в целевых образцах.

ChIPseqSpikeInFree

Несмотря на то что использование калибровочного контроля показало свою эффективность как в анализе данных экспрессии генов, так и для ChIP-seq, лишь малая доля исследований использует этот подход. Это связано с экспериментальными трудностями (например, требуется оптимизация количества добавляемого хроматина в зависимости от целевой модификации гистонов или транскрипционных факторов). Тем не менее, как утверждают авторы метода ChIPseqSpikeInFree [29], в таких случаях все же возможно провести адекватную нормализацию и провести сравнение образцов, исходя только из данных секвенирования ChIP-seq.

Для этого авторы разбивают весь референсный геном на участки длиной 1 тпн и определяют распределение плотности прочтений в этих участках. Они обнаружили, что во многих ChIP-seq экспериментах с помощью анализа плотности все участки генома можно разбить на три класса: 1) те, в которых нет обогащения целевым белком, 2) те, в которых низкое обогащение, и 3) те, в которых очень высокое обогащение. Далее авторы заметили, что многие локусы с высоким обогащением инвариантны относительно глобальных изменений структуры хроматина и их обогащение обратно пропорционально глобальным уровням целевого белка. На основании этого они предлагают нормализовать образцы исходя из того, что соотношение между необогащенными локусами и высокообогащенными локусами генома не зависит от исследуемого образца.

Используя этот подход на примере нескольких экспериментов, в которых происходит глобальное изменение уровней химических модификаций гистонов, авторы показали, что результаты соответствуют ожиданиям. Например, они подтвердили глобальное уменьшение уровней H3K79me2 и сохранение уровней H3K4me3 в клетках MV4-11, обработанных фенилбутазоном, по сравнению с необработанными, а также увеличение уровней H3K27me3 в клеточных линиях с нокаутом гистона H3.3^{K27M} и в нескольких других экспериментах [9, 30].

Нелинейная нормализация

В дополнение к предыдущим методам, которые в большей степени производят нормализацию линейным преобразованием данных, были также предложены методы, предоставляющие большую гибкость. Например, в методе S3norm [31] производится не линейная, а логарифмическая нормализация таким образом, чтобы не только локусы обогащения имели в среднем одинаковый сигнал, но локусы, в которых нет обогащения (отрицательный контроль), так же имели в среднем одинаковые значения сигнала (отличающиеся от значения для локусов обогащения). Таким образом, напрямую производится нормализация соотношения сигнала и шума. В пакете csaw [13] предложен метод нелинейной нормализации, который производит трансформацию данных (методом локальной полиномиальной регрессии) таким образом, чтобы соотношение сигнала между образцами не зависело от уровня среднего значения сигнала.

Выбор конкретного метода нормализации

Такое большое количество методов для одного из первых этапов анализа данных ChIP-seq порождает очевидный вопрос: какой из них необхо-

димо использовать? Авторы каждого из методов аргументируют в пользу того, который разработали они сами, поэтому в первую очередь необходимы независимые исследования применимости этих методов, какие были уже опубликованы для RNA-seq.

Авторы одного из таких исследований для ChIP-seq [32] предлагают выбирать подходящий метод из нескольких, основываясь на анализе адекватности сделанной ими оценки коэффициента нормализации с помощью так называемого диагностического графика. Для построения такого графика последовательность референсного генома разбивается на множество непересекающихся участков (окон), которые затем кластеризуются таким образом, чтобы в них последовательно содержалось примерно одинаковое количество картированных последовательностей. Авторы показали, что отношение количества картированных последовательностей в каждом кластере имеет форму нормального распределения, в котором среднее равно логарифму коэффициента нормализации. Отображая на графике эти распределения и сравнивая их с оценками, полученными одним из вышеописанных методов, можно получить представление о достоверности нормализации. В качестве примера авторы сравнивают методы нормализации CisGenome, CCAT и NCIS и в приведенных примерах метод CCAT дает заниженную оценку коэффициента нормализации.

Следует учесть, что предложенный метод диагностических графиков можно использовать только для методов нормализации, которые выполняют линейную трансформацию данных, и он не учитывает возможность глобального изменения структуры хроматина. Поэтому при выборе способа нормализации необходимо учитывать специфический характер исследования и маловероятно, что существует единый способ нормализации, подходящий под все возможные эксперименты.

НОРМАЛИЗАЦИЯ ДРУГИХ ФАКТОРОВ

Вместе с техническими эффектами, свойственными именно ChIP-seq, исследованиям хроматина свойственны те же проблемы, что и другим молекулярным методам, в частности неравномерность выборки и/или сильное влияние факторов, которые не представляют интереса для исследования. Например, при исследовании клеток мозга человека методом ATAC-seq [33] было установлено, что более 466 локусов генома меняют степень открытости хроматина в зависимости от продолжительности временного интервала, прошедшего от момента смерти до получения образцов тканей мозга (postmortem interval, PMI). В исследованиях нейropsychических заболеваний некоторыми такими факторами являются курение человека, которое значительно влияет на эпигенетический профиль [34],

в том числе клеток мозга, и использование медикаментов, которые тоже по-разному могут влиять на состояние хроматина. Более общей трудностью при исследованиях клеток мозга является гетерогенность клеточного состава клеток мозга, который содержит, по разным оценкам, более 10 различных типов клеток (например, нейроны различных типов, олигодендроциты, астроциты, микроглия и пр.), в результате чего полученный молекулярный профиль представляет собой некое усреднение сигнала по разным типам клеток (за исключением применения методов секвенирования отдельных единичных клеток или ядер). Эти и другие факторы требуют более тщательного подхода к анализу, нежели классический метод сравнения двух групп индивидов, который в этом случае будет иметь меньшую статистическую силу.

Для многих молекулярных методов успешно используется подход, в котором наблюдаемый сигнал раскладывается на несколько компонент с помощью линейной модели. Этот подход позволяет учитывать в модели не только исследуемый фенотип, но и другие характеристики исследуемых образцов. Например, в эту модель включают возраст, пол, статус курения и пр., что позволяет при проведении статистического теста на ассоциацию с фенотипом исключить их влияние. Этот подход используется в программных пакетах *limma* [35], *edgeR* [16], *DEseq* [36] и других. Другим вариантом является линейная регрессия, с помощью которой можно получить сигнал, из которого исключено влияние факторов, не представляющих интереса для исследования [37, 38]. Считается, что второй подход по сравнению с первым уменьшает статистическую силу анализа, но зато позволяет использовать другие виды анализа (например, PCA). Таким образом возможно учесть различные биологические (и технические тоже) факторы при анализе конкретного фенотипа.

В частности, в качестве таких факторов можно использовать клеточный состав исследуемых образцов (т.е. процент содержания различных типов клеток), который можно оценить по уже полученному профилю ChIP-seq. Одним из наиболее надежных способов является декомпозиция на основании ранее полученных профилей отдельных типов клеток (в похожих экспериментальных условиях, например для того же транскрипционного фактора или модификации гистонов), которые можно получить из ранее опубликованных данных. Например, некоторые исследования клеток используют метод сортировки ядер нейронов, чтобы отделить нейроны мозга от других типов клеток мозга: получены нейронально-специфичные профили для некоторых модификаций хро-

матина (в частности H3K4me3), а также соответствующие профили ненейрональных клеток [39, 40]. В одном из исследований получены профили модификаций хроматина для четырех различных типов клеток [41]. Наконец, в рамках проекта ENCODE [42] получены различные данные ChIP-seq для клеточных линий, производных от астроцитов, олигодендроцитов, нейронов и прочих. В этом случае возможно использование одного из многих методов декомпозиции [43, 44]. Тем не менее следует учитывать, что эти методы в большей степени разработаны для анализа экспрессии генов и ДНК метилирования, но аналогичные подходы можно применять и для анализа ChIP-seq [45], хотя этой теме посвящено малое количество исследований.

Наконец, в некоторых случаях оправдано определение неизвестных заранее факторов, не связанных с исследуемыми фенотипами. В частности, довольно успешно и часто применяются методы анализа суррогатных переменных (SVA) [46] или RUVseq [47]. Однако следует учитывать, что они основаны на линейных моделях и могут при применении оказывать нежелательное влияние на анализ исследуемых фенотипов. В частности, эти методы могут определять пропорции различных типов клеток как один из таких нежелательных факторов.

НОРМАЛИЗАЦИЯ ChIP-seq В ИССЛЕДОВАНИЯХ ТКАНЕЙ МОЗГА В НОРМЕ И ПАТОЛОГИИ

Одним из центральных проектов последнего десятилетия в исследованиях эпигенетики является проект ENCODE [42], в основном направленный на исследование эпигенетического статуса. В какой-то мере его продолжением является проект psychENCODE [48], направленный на изучение регуляции экспрессии генов при психиатрических заболеваниях и, в частности, его участники применяют ChIP-seq для исследования молекулярных факторов патологий. Например, в исследовании органоидов мозга человека (т.е. полученных путем репрограммирования фибробластов) [49] для сравнительного анализа ChIP-seq профилей с маркером H3K27ac была применена TMM-нормализация. В исследовании профилей ацетилирования гистонов сортированных нейронов лобной доли [50] также был применен метод TMM-нормализации и поправка на многие другие факторы (возраст, пол, клеточный состав образца и пр.), как и в продолжении этой работы для сравнения пациентов с шизофренией и контрольной выборки [51] (но уже не только для метки H3K27ac, но и для H3K4me3). Однако в анало-

гичном исследовании ацетилирования гистонов пациентов с аутизмом была применена нормализация на глубину секвенирования (но также вместе с поправкой на возраст, пол, пропорцию нейрональных клеток и пр.).

В исследованиях нейродегенеративных заболеваний, таких как болезнь Альцгеймера, также применяется метод ChIP-seq для исследования эпигенетического состояния хроматина. Например, для исследования профилей H3K4me3 и H3K27me3 клеток энторинальной коры головного мозга больных и здоровых индивидов [52] была применена нормализация на общее количество прочтений (т.е. на глубину секвенирования). В более широком исследовании 699 индивидов с различной степенью деменции [53] был также применен метод нормализации на глубину секвенирования. В исследованиях факторов развития глиобластомы с помощью ChIP-seq против метки активных генов H3K27ac [54] была использована нормализация методом, близким к ТММ-нормализации из пакета DESeq2 [55]. В другом исследовании глиом [56] была использована нормализация на глубину секвенирования.

Таким образом, в исследованиях модификаций гистонов методом ChIP-seq в клетках мозга человека наиболее часто используются сортировка нейронов отдельно от других типов клеток, а также нормализация на глубину секвенирования или ТММ-нормализация. Несмотря на то что оба эти метода нормализации не были изначально разработаны специально для данных ChIP-seq, их применение оправдано в тех случаях, когда уровень шума примерно одинаков в используемых данных (что необходимо контролировать [11]), либо когда другие факторы (пол, возраст и пр.) вносят большой вклад и нормализация предназначена в первую очередь для учета этих различий.

ЗАКЛЮЧЕНИЕ

Широкий спектр методов нормализации данных ChIP-seq позволяет выбрать наиболее подходящий под условия конкретного эксперимента. Несмотря на то что многое можно сделать во время анализа уже полученных данных, изначальное планирование эксперимента может иметь решающее значение для статистической силы и достоверности последующего анализа. Например, уже упомянутое ранее добавление калибровочного контроля для анализа позволяет провести нормализацию данных, но имеет также и недостатки. Одним из подходов к частичному решению проблемы гетерогенности клеток мозга может быть использование сортировки клеток/ядер до прове-

дения ChIP-seq для получения профиля, что часто применяется для исследования нейропсихиатрических и нейродегенеративных заболеваний. Наконец, использование одноклеточных технологий (single cell/nuclei) для исследования состояния хроматина на уровне отдельных клеток (scATAC-seq и scChIP-seq) является одним из наиболее перспективных методов в этой области.

Работа выполнена при финансовой поддержке РНФ (номер гранта 19-75-30039).

Настоящая статья не содержит каких-либо исследований с использованием в качестве объекта животных.

Настоящая статья не содержит каких-либо исследований с участием в качестве объекта людей.

Авторы заявляют, что у них нет конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. *Fyodorov D.V., Zhou B.-R., Skoultchi A.I., Bai Y.* Emerging roles of linker histones in regulating chromatin structure and function // *Nat. Rev. Mol. Cell. Biol.* 2018. V. 19. № 3. P. 192–206. <https://doi.org/10.1038/nrm.2017.94>
2. *Park P.J.* ChIP-seq: advantages and challenges of a maturing technology // *Nat. Rev. Genet.* 2009. V. 10. № 10. P. 669–680. <https://doi.org/10.1038/nrg2641>
3. *Furey T.S.* ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions // *Nat. Rev. Genet.* 2012. V. 13. № 12. P. 840–852. <https://doi.org/10.1038/nrg3306>
4. *Altman N.* Batches and blocks, sample pools and subsamples in the design and analysis of gene expression studies // *Batch Effects and Noise in Microarray Experiments*. UK, Chichester: John Wiley & Sons, Ltd, 2009. P. 33–50. <https://doi.org/10.1002/9780470685983.ch4>
5. *Goh W.W.B., Wang W., Wong L.* Why batch effects matter in omics data, and how to avoid them // *Trends Biotechnol.* 2017. V. 35. № 6. P. 498–507. <https://doi.org/10.1016/j.tibtech.2017.02.012>
6. *Jung Y.L., Luquette L.J., Ho J.W.K. et al.* Impact of sequencing depth in ChIP-seq experiments // *Nucl. Acids Res.* 2014. V. 42. № 9. <https://doi.org/10.1093/nar/gku178>
7. *Sundaram A.Y.M., Hughes T., Biondi S. et al.* A comparative study of ChIP-seq sequencing library preparation methods // *BMC Genomics.* 2016. V. 17. № 1. P. 816. <https://doi.org/10.1186/s12864-016-3135-y>
8. *Teng M., Du D., Chen D., Irizarry R.A.* Characterizing batch effects and binding site-specific variability in ChIP-seq data // *NAR Genomics and Bioinformatics.*

2021. V. 3. № 4.
<https://doi.org/10.1093/nargab/lqab098>
9. Orlando D.A., Chen M.W., Brown V.E. et al. Quantitative ChIP-seq normalization reveals global modulation of the epigenome // *Cell Reports*. 2014. V. 9. № 3. P. 1163–1170.
<https://doi.org/10.1016/j.celrep.2014.10.018>
10. Gu B., Lee M.G. Histone H3 lysine 4 methyltransferases and demethylases in self-renewal and differentiation of stem cells // *Cell & Bioscience*. 2013. V. 3. № 1.
<https://doi.org/10.1186/2045-3701-3-39>
11. Nakato R., Sakata T. Methods for ChIP-seq analysis: A practical workflow and advanced applications // *Methods*. 2021. V. 187. P. 44–53.
<https://doi.org/10.1016/j.jymeth.2020.03.005>
12. Price E.M., Robinson W.P. Adjusting for batch effects in DNA methylation microarray data, a lesson learned // *Front. Genet.* 2018. V. 9.
<https://doi.org/10.3389/fgene.2018.00083>
13. Lun A.T.L., Smyth G.K. csaw: A Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows // *Nucl. Acids Res.* 2016. V. 44. № 5.
<https://doi.org/10.1093/nar/gkv1191>
14. Diaz A., Park K., Lim D.A., Song J.S. Normalization, bias correction, and peak calling for ChIP-seq // *Stat. Appl. Genet. Mol. Biol.* 2012. V. 11. № 3.
<https://doi.org/10.1515/1544-6115.1750>
15. Stark R., Brown G. DiffBind: Differential Binding Analysis of ChIP-seq Peak Data. Bioconductor version: Release (3.16), 2022.
<https://doi.org/10.18129/B9.bioc.DiffBind>
16. Robinson M.D., McCarthy D.J., Smyth G.K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data // *Bioinformatics*. 2010. V. 26. № 1. P. 139–140.
<https://doi.org/10.1093/bioinformatics/btp616>
17. Ji H., Jiang H., Ma W., Wong W.H. Using CisGenome to analyze ChIP-chip and ChIP-seq data // *Curr. Protoc. Bioinformatics*. 2011.
<https://doi.org/10.1002/0471250953.bi0213s33>
18. Kharchenko P.V., Tolstorukov M.Y., Park P.J. Design and analysis of ChIP-seq experiments for DNA-binding proteins // *Nat. Biotechnol.* 2008. V. 26. № 12. P. 1351–1359.
<https://doi.org/10.1038/nbt.1508>
19. Xu H., Handoko L., Wei X. et al. A signal-noise model for significance analysis of ChIP-seq with negative control // *Bioinformatics*. 2010. V. 26. № 9. P. 1199–1204.
<https://doi.org/10.1093/bioinformatics/btq128>
20. Liang K., Keleş S. Normalization of ChIP-seq data with control // *BMC Bioinformatics*. 2012. V. 13. № 1.
<https://doi.org/10.1186/1471-2105-13-199>
21. Shao Z., Zhang Y., Yuan G.-C. et al. MAnorm: A robust model for quantitative comparison of ChIP-Seq data sets // *Genome Biol.* 2012. V. 13. № 3.
<https://doi.org/10.1186/gb-2012-13-3-r16>
22. Tu S., Li M., Chen H. et al. MAnorm2 for quantitatively comparing groups of ChIP-seq samples // *Genome Res.* 2021. V. 31. № 1. P. 131–145.
<https://doi.org/10.1101/gr.262675.120>
23. Nair N.U., Sahu A.D., Bucher P., Moret B.M.E. ChIP-norm: A statistical method for normalizing and identifying differential regions in histone modification ChIP-seq libraries // *PLoS One*. 2012. V. 7. № 8.
<https://doi.org/10.1371/journal.pone.0039573>
24. Polit L., Kerdivel G., Gregoricchio S. et al. CHIPIN: ChIP-seq inter-sample normalization based on signal invariance across transcriptionally constant genes // *BMC Bioinformatics*. 2021. V. 22. № 1. P. 407.
<https://doi.org/10.1186/s12859-021-04320-3>
25. Allhoff M., Seré K., Pires J. et al. Differential peak calling of ChIP-seq signals with replicates with THOR // *Nucl. Acids Res.* 2016. V. 44. № 20.
<https://doi.org/10.1093/nar/gkw680>
26. Lovén J., Orlando D.A., Sigova A.A. et al. Revisiting global gene expression analysis // *Cell*. 2012. V. 151. № 3. P. 476–482.
<https://doi.org/10.1016/j.cell.2012.10.012>
27. Kanno J., Aisaki K., Igarashi K. et al. “Per cell” normalization method for mRNA measurement by quantitative PCR and microarrays // *BMC Genomics*. 2006. V. 7. № 1.
<https://doi.org/10.1186/1471-2164-7-64>
28. Egan B., Yuan C.-C., Craske M.L. et al. An alternative approach to ChIP-Seq normalization enables detection of genome-wide changes in histone H3 lysine 27 trimethylation upon EZH2 inhibition // *PLoS One*. 2016. V. 11. № 11.
<https://doi.org/10.1371/journal.pone.0166438>
29. Jin H., Kasper L.H., Larson J.D. et al. ChIPseqSpikeIn-Free: a ChIP-seq normalization approach to reveal global changes in histone modifications without spike-in // *Bioinformatics*. 2020. V. 36. № 4. P. 1270–1272.
<https://doi.org/10.1093/bioinformatics/btz720>
30. Pathania M., De Jay N., Maestro N. et al. H3.3K27M cooperates with Trp53 loss and PDGFRA gain in mouse embryonic neural progenitor cells to induce invasive high-grade gliomas // *Cancer Cell*. 2017. V. 32. № 5. P. 684–700. e9.
<https://doi.org/10.1016/j.ccell.2017.09.014>
31. Xiang G., Keller C.A., Giardine B. et al. S3norm: Simultaneous normalization of sequencing depth and signal-to-noise ratio in epigenomic data // *Nucl. Acids Res.* 2020. V. 48. № 8. P. e43.
<https://doi.org/10.1093/nar/gkaa105>
32. Angelini C., Heller R., Volkinshtein R., Yekutieli D. Is this the right normalization? A diagnostic tool for ChIP-seq normalization // *BMC Bioinformatics*. 2015. V. 16. № 1. P. 150.
<https://doi.org/10.1186/s12859-015-0579-z>
33. Bryois J., Garrett M.E., Song L. et al. Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia // *Nat. Commun.* 2018. V. 9.

- № 1. P. 3121.
<https://doi.org/10.1038/s41467-018-05379-y>
34. Tsai P.-C., Glastonbury C.A., Eliot M.N. et al. Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health // *Clin. Epigenetics*. 2018. V. 10. P. 126.
<https://doi.org/10.1186/s13148-018-0558-0>
 35. Ritchie M.E., Phipson B., Wu D. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies // *Nucl. Acids Res.* 2015. V. 43. № 7. P. e47.
<https://doi.org/10.1093/nar/gkv007>
 36. Love M.I., Huber W., Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 // *Genome Biol.* 2014. V. 15. № 12. P. 550.
<https://doi.org/10.1186/s13059-014-0550-8>
 37. Zhang Y., Parmigiani G., Johnson W.E. ComBat-seq: Batch effect adjustment for RNA-seq count data // *NAR Genomics and Bioinformatics*. 2020. V. 2. № 3.
<https://doi.org/10.1093/nargab/lqaa078>
 38. Johnson W.E., Li C., Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods // *Biostatistics*. 2007. V. 8. № 1. P. 118–127.
<https://doi.org/10.1093/biostatistics/kxj037>
 39. Shulha H.P., Cheung I., Guo Y. et al. Coordinated cell type-specific epigenetic remodeling in prefrontal cortex begins before birth and continues into early adulthood // *PLoS Genetics*. 2013. V. 9. № 4.
<https://doi.org/10.1371/journal.pgen.1003433>
 40. Gusev F.E., Reshetov D.A., Mitchell A.C. et al. Epigenetic-genetic chromatin footprinting identifies novel and subject-specific genes active in prefrontal cortex neurons // *The FASEB J.* 2019. V. 33. № 7. P. 8161–8173.
<https://doi.org/10.1096/fj.201802646R>
 41. Nott A., Holtman I.R., Coufal N.G. et al. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association // *Science*. 2019. V. 366. № 6469. P. 1134–1139.
<https://doi.org/10.1126/science.aay0793>
 42. Dunham I., Kundaje A., Aldred S.F. et al. An integrated encyclopedia of DNA elements in the human genome // *Nature*. 2012. V. 489. № 7414. P. 57–74.
<https://doi.org/10.1038/nature11247>
 43. Ouyang Z., Bourgeois-Tchir N., Lyashenko E. et al. Characterizing the composition of iPSC derived cells from bulk transcriptomics data with CellMap // *Sci. Rep.* 2022. V. 12. № 1. P. 17394.
<https://doi.org/10.1038/s41598-022-22115-1>
 44. Jew B., Alvarez M., Rahmani E. et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information // *Nat. Commun.* 2020. V. 11. № 1. P. 1971.
<https://doi.org/10.1038/s41467-020-15816-6>
 45. Li H., Sharma A., Luo K. et al. DeconPeaker, A deconvolution model to identify cell types based on chromatin accessibility in ATAC-seq data of mixture samples // *Frontiers in Genet.* 2020. V. 11.
 46. Leek J.T. svaseq: removing batch effects and other unwanted noise from sequencing data // *Nucl. Acids Res.* 2014. V. 42. № 21. P. e161.
<https://doi.org/10.1093/nar/gku864>
 47. Risso D., Ngai J., Speed T.P., Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples // *Nat. Biotechnol.* 2014. V. 32. № 9. P. 896–902.
<https://doi.org/10.1038/nbt.2931>
 48. Akbarian S., Liu C., Knowles J.A. et al. The psychENCODE project // *Nat. Neurosci.* 2015. V. 18. № 12. P. 1707–1712.
<https://doi.org/10.1038/nn.4156>
 49. Amiri A., Coppola G., Scuderi S. et al. Transcriptome and epigenome landscape of human cortical development modeled in organoids // *Science*. 2018. V. 362. № 6420.
<https://doi.org/10.1126/science.aat6720>
 50. Girdhar K., Hoffman G.E., Jiang Y. et al. Cell-specific histone modification maps in the human frontal lobe link schizophrenia risk to the neuronal epigenome // *Nat. Neurosci.* 2018. V. 21. № 8. P. 1126–1136.
<https://doi.org/10.1038/s41593-018-0187-0>
 51. Girdhar K., Hoffman G.E., Bendl J. et al. Chromatin domain alterations linked to 3D genome organization in a large cohort of schizophrenia and bipolar disorder brains // *Nat. Neurosci.* 2022. V. 25. № 4. P. 474–483.
<https://doi.org/10.1038/s41593-022-01032-6>
 52. Persico G., Casciaro F., Amatori S. et al. Histone H3 Lysine 4 and 27 Trimethylation Landscape of Human Alzheimer's Disease // *Cells. Multidisciplinary Digital Publ. Institute*. 2022. V. 11. № 4.
<https://doi.org/10.3390/cells11040734>
 53. Klein H.-U., McCabe C., Gjoneska E. et al. Epigenome-wide study uncovers large-scale changes in histone acetylation driven by tau pathology in the aging and Alzheimer human brain // *Nat. Neurosci.* 2019. V. 22. № 1. P. 37–46.
<https://doi.org/10.1038/s41593-018-0291-1>
 54. Mack S.C., Singh I., Wang X. et al. Chromatin landscapes reveal developmentally encoded transcriptional states that define human glioblastoma // *J. Exp. Med.* 2019. V. 216. № 5. P. 1071–1090.
<https://doi.org/10.1084/jem.20190196>
 55. Anders S., Huber W. Differential expression analysis for sequence count data // *Genome Biol.* 2010. V. 11. № 10.
<https://doi.org/10.1186/gb-2010-11-10-r106>
 56. Stępiak K., Machnicka M.A., Mieczkowski J. et al. Mapping chromatin accessibility and active regulatory elements reveals pathological mechanisms in human gliomas // *Nat. Commun.* 2021. V. 12. № 1. P. 3621.
<https://doi.org/10.1038/s41467-021-23922-2>

Methods for ChIP-seq Normalization and Their Application for Analysis of Regulatory Elements in Brain Cells

F. E. Gusev^{a, b, *}, T. V. Andreeva^{a, b, **}, and E. I. Rogayev^{a, b, ***}

^a*Vavilov Institute of General Genetics Russian Academy of Sciences, Moscow, 119991 Russia*

^b*Center for Genetics and Life Science, "Sirius" University of Science and Technology, Krasnodar krai, p. Sirius, 354340 Russia*

^{*}*e-mail: gusev@vigg.ru*

^{**}*e-mail: an_tati@vigg.ru*

^{***}*e-mail: evivrog@gmail.com*

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) has become one of the major tools to elucidate gene expression programs. Similar to other molecular profiling methods, ChIP-seq is sensitive to several technical biases which affect downstream results, especially in cases when material quality is difficult to control, for example, frozen post-mortem human tissue. However methods for bioinformatics analysis improve every year and allow to mitigate these effects after sequencing by adjusting for both technical ChIP-seq biases and more general biological biases like post-mortem interval or cell heterogeneity of the sample. Here we review a wide selection of ChIP-seq normalization methods with a focus on application in specific experimental settings, in particular when brain tissue is investigated.

Keywords: genomics, epigenomics, chromatin, normalization, immunoprecipitation, brain.