БИОФИЗИКА, 2020, том 65, № 6, с. 1045–1050

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 577.3

ЭФФЕКТИВНОСТЬ ОПРЕДЕЛЕНИЯ 5-МЕТИЛЦИТОЗИНА В ДНК КЛЕТОК *Escherichia coli*, НЕСУЩИХ ГЕНЫ БАКТЕРИАЛЬНЫХ ДНК-МЕТИЛТРАНСФЕРАЗ, С ПОМОЩЬЮ УСТАНОВКИ OXFORD NANOPORE

© 2020 г. В.В. Ильинский*, **, Е.М. Козлова*, *, С.Х. Дегтярев****, Н.К. Янковский*, *****, В.Ю. Макеев*, ***, ******, *********

**Институт общей генетики им. Н.И. Вавилова РАН, 119991, Москва, ул. Губкина, 3*

***ООО «Генотек», 105120, Москва, Наставнический пер., 17/1*

****Московский Физико-Технический Институт. 141701, Долгопрудный Московской области, Институтский пер., 9*

*****НПО «СибЭнзим», 630060, Новосибирск, ул. Академика Тимакова, 2/12*

******Московский государственный университет имени М.В. Ломоносова, Москва, 119991, Воробьевы горы, 1*

*******Институт молекулярной биологии РАН им. В.А. Энгельгардта, 119991, Москва, ул. Вавилова, 32*

********НИЦ «Курчатовский институт» – ГосНИИгенетика, 117545, Москва, 1-й Дорожный проезд, 1*

E-mail: vsevolod.makeev@vigg.ru

Поступила в редакцию 28.08.2020 г.
После доработки 28.08.2020 г.
Принята к публикации 24.09.2020 г.

С помощью системы MinION (Oxford Nanopore Technologies Ltd, Великобритания) проведено прямое секвенирование геномной ДНК двух рекомбинантных штаммов *Escherichia coli*. В одном случае в клетках присутствовала плазмида с геном ДНК-метилтрансферазы M.HpaII, метилирующей второй цитозин в сайте CCGG, во втором случае штамм *E. coli* содержал ДНК-метилтрансферазу M.HspAI, модифицирующую центральный цитозин в последовательности GCGC. В обоих случаях при метилировании образуется 5-метилцитозин. Показано, что в случае высокого покрытия при секвенировании ДНК, наличие 5-метилцитозина в ДНК определяется с высокой точностью. В частности, в ДНК *E. coli* с клонированным геном ДНК-метилтрансферазы M.HspAI при прочтении с покрытием 1300× 98.9% сайтов GCGC определяются как метилированные по первому цитозину. В то же время лишь 0.09% остальных тетрануклеотидов, имеющих в середине динуклеотид CpG, дают ложноположительный результат и определяются как метилированные по центральному цитозину. В присутствии гена ДНК-метилтрансферазы M.HpaII среди позиций, покрытых более чем 700 ридами, 91.3% всех сайтов CCGG определяются как метилированные, при этом только 0.13% других тетрануклеотидов с центральным CG-динуклеотидом определяются как содержащие 5-метилцитозин во втором положении. Делается вывод, что при используемой методике для надежного определения 5-метилцитозина и исключения ложноположительных результатов покрытие должно быть не менее 700–1000×.

Ключевые слова: нанопоры, 5-метилцитозин, метилирование, E. coli, DeepSignal.

DOI: 10.31857/S0006302920060010

Нанопорное секвенирование — это уникальная перспективная технология определения последовательности нуклеиновых кислот, разработанная компанией Oxford Nanopore Technologies Ltd (Великобритания). Основным элементом такого секвенатора является мембрана с отверстием порядка 10^{-9} м в диаметре. Мембрана помещается в электролитический раствор. Постоянное электрическое поле прикладывается в перпендикулярном к поверхности направлении. При этом возникает ионный ток, величина которого постоянно фиксируется цифровым преобразователем. Кроме того, через это же отверстие подается одноцепочечная ДНК, скорость подачи которой контролируется ферментативной системой, включающей хеликазу [1]. Использование хеликазы явилось ключевым изобретением при создании технологии, поскольку позволило контролировать скорость прохождения ДНК через пору. Скорость диффундирования свободной ДНК по градиенту электрического поля составляет приблизительно 10 мкс на нуклеотид и слишком велика, чтобы позволить разрешить последовательность нуклеотидов ДНК. В результате была сконструирована молекулярная машина, основанная на использовании хеликазы, фермента разделяю-

щего двунитевую ДНК на две однонитевых ДНК. В установке MinION хеликаза не только разделяет двунитевую ДНК на две однонитевых ДНК, но и разделяет нити, направляя одну нить в пору, а другую – возвращая в исходный раствор. Выбранная MinION хеликаза позволяет подавать одноцепочечную ДНК со скоростью порядка 500 нуклеотидов в секунду, что позволяет наиболее эффективно разрешать последовательность нуклеотидов [1]. Система позволяет получать очень длинные риды (одиночные прочтения, от английского «read»), до 100000 оснований, однако доля таких длинных ридов невелика, а большинство ридов не превышают по своей длине нескольких тысяч нуклеотидов.

Сила ионного тока определяется пропускной способностью отверстия, которая, в свою очередь, зависит от последовательности нуклеотидов исследуемой ДНК. Эти изменения достаточно велики, чтобы по ним можно было восстановить последовательность нуклеотидов. Таким образом, секвенаторы MinION, разработанные компанией Oxford Nanopore Technologies, осуществляют прямой анализ одноцепочечных молекул ДНК, не подвергая их амплификации и не используя химический синтез.

Последовательность импульсов, получаемая на выходе прибора, зависит не только от чередования классических оснований ДНК (аденина, гуанина, тимина и цитозина) [1], но и от наличия модификаций ДНК, таких как 5-метилцитозин [2, 3]. Возможность прямого определения 5-метилцитозина является важнейшим преимуществом нанопорных технологий и принципиально недоступна для систем, использующих технологию секвенирования на основе применения ДНК-полимеразы (sequencing by synthesis), реализованных, например, семейством установок HiSeq, созданных компанией Illumina (США).

Знание профиля метилирования ДНК чрезвычайно важно как для медицинских приложений, так и для фундаментальных исследований в области молекулярной биологии, поскольку эти модификации играют важнейшую роль в развитии, регуляции и поддержании жизнедеятельности как бактериальной, так и эукариотической клетки. В частности, они являются ключевыми элементами геномного импринтинга, инактивации Х-хромосомы, репрессии ретроэлементов и процесса старения. Метилирование промоторных участков, как правило, подавляет транскрипцию, и этот механизм реализуется, в частности, при формировании различных типов клеток позвоночных. Многие заболевания имеют характерный профиль метилирования ДНК. Например, локальное 5mC-гиперметилирование и общее гипометилирование генома в целом характерно для опухолевых клеток [4], а паттерны метилирования CpG островков позволяют различить нормальные и опухолевые ткани [5].

Фирмой Oxford Nanopore Technologies и сторонними разработчиками создан ряд декодеров, позволяющих переводить последовательность электронных импульсов в последовательность оснований ДНК. Согласно обзору [1], наиболее точным декодером является программа Guppy (https://community.nanoporetech.com). За последние несколько лет был разработан ряд методов, позволяющих с достаточной уверенностью определять наличие 5-метилцитозина в последовательности ДНК. Для этой цели используются статистические алгоритмы [3] и, в последнее время, нейросетевые технологии, которые реализованы в программах DeepMod, DeepSignal [6, 7]. Таким образом, в распоряжении исследователей впервые появился метод прямого определения модификаций одиночных молекул ДНК, не подвергавшихся амплификации. К сожалению, несмотря на большую работу, проведенную Oxford Nanopore Technologies Ltd, точность определения как классических оснований, так и их модификаций невелика, и при одиночном прочтении не превышает 89% [1]. Такая ошибка в определении оснований до определенной степени может быть исправлена путем глубокого секвенирования и статистической обработки полученной большой выборки перекрывающихся ридов.

В настоящем исследовании было проведено прямое секвенирование двух геномов рекомбинантных штаммов *Escherichia coli* с клонированными генами бактериальных ДНК-метилтрансфераз, осуществляющими высокоспецифичное метилирование сайтов GCGC и CCGG по центральному цитозину. Геномы бактерий можно секвенировать очень глубоко, и именно на геноме бактерий была впервые продемонстрирована возможность точного определения последовательности с помощью секвенатора MinION [8].

МАТЕРИАЛЫ И МЕТОДЫ

В работе использовали штамм *E. coli* ER2267, любезно предоставленный компанией NEB (США). Ген ДНК-метилтрансферазы M.HspAI из *Haemophilus* sp. A1 был клонирован в вектор pUC19, полученной плазмидой трансформировали клетки *E. coli* ER2267 (http://science.sibenzyme.com/soft-data/database/nucleotide-sequence-of-plasmid-dna-phspai2). Ген ДНК-метилтрансферазы M.HpaII из *Haemophilus parainfluenzae* был клонирован в вектор PMTL22 и клетки *E. coli* ER2267 трансформировали этой ДНК (http://www.sibenzyme.com/info624.php). Геномную ДНК из полученных рекомбинантных штаммов выделяли, как описано ранее [9].

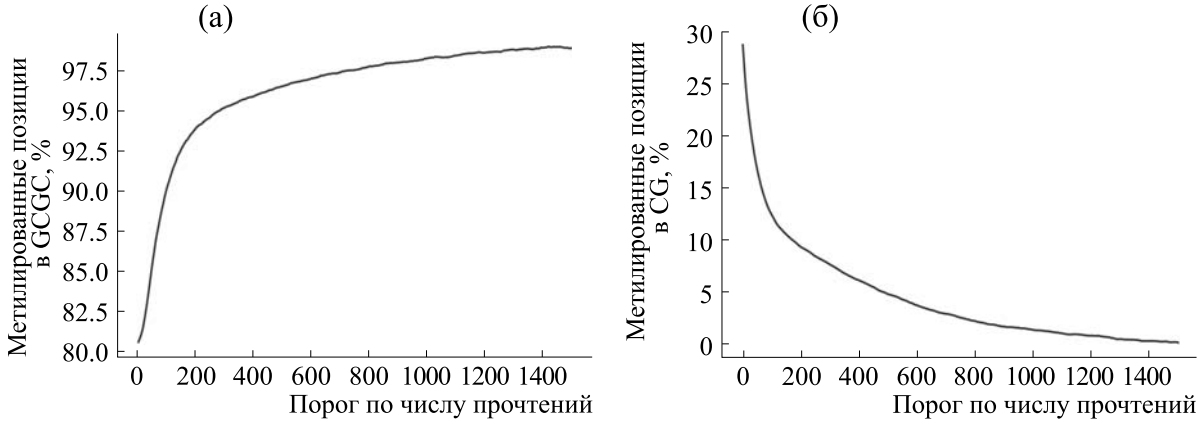

Рис. 1. Количество контекстов, определенных как метилированные, в зависимости от минимального покрытия ридами MinION: (а) – контексты GCGC, (б) – контексты NCGN, отличные от GCGC. Риды фильтрованы по индексу уверенности DeepSignal.

Оба препарата геномной ДНК были отсеквенированы с помощью MinION (R9.4.1) с применением протокола SQK-LSK108 и получением исходного среднего покрытия 2193× и 1832× для геномов клеток *E. coli*, несущих гены метилтрансфераз M.HpaII и M.HspAI соответственно. Риды в формате fast5 были обработаны программой Guppy. После чернового определения оснований результаты были согласованы с последовательностью генома *E. coli* штамма NRRL B-1109 с использованием программы tombo resquiggle (Oxford Nanopore Technologies). Этот алгоритм создает новое соответствие сигнала и нуклеотидов, используя последовательности из референсного генома. Полученные файлы с координатами ридов в референсном геноме и сигналом секвенирования использовали в качестве источника входных данных программы DeepSignal [7] для определения 5-метилцитозинов. Программа DeepSignal присваивает каждой CpG-паре индекс надежности метилирования цитозинов, представляющий собой число от 0 до 1 (0 – надежно неметилированный цитозин; 1 – цитозин надежно метилирован). Была проведена фильтрация ридов с сохранением ридов, содержащих CpG-пары с индексом менее 0.1 (неметилированные цитозины) или более 0.9 (метилированные цитозины). Среднее покрытие геномов после такой фильтрации составило 175 (M.HspAI), и 146 (M.HpaII). При этом дисперсия значений покрытия оказалась очень велика, и в геноме наблюдались участки с покрытием до 1500 ридов для отдельных CpG-пар. Цитозин, для которого после фильтрации было показано метилирование более чем в 50% покрывающих его ридов, считался метилированным. Статистический анализ полученных результатов проводили с использованием интерпретируемого языка Python.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Секвенирование генома *E. coli* с клонированным геном ДНК-метилтрансферазы M.HspAI (контекст метилирования GCGC). Распределение ридов MinION вдоль генома *E. coli* достаточно неоднородно. Поскольку работа в первую очередь направлена на анализ метилирования, было вычислено покрытие выявленных 5-метилцитозинов ридами MinION, которое варьировало от значений меньше 100× до значений, превышающих 1500×. Как видно из рис. 1а, доля определяемых сайтов G(5mC)GC существенно зависит от величины покрытия ридами MinION.

При покрытии 1500× определяется 98.9% всех сайтов G(5mC)GC (не определяется 31 сайт из 2789). Доля неопределяемых сайтов G(5mC)GC на уровне приблизительно 1% сохраняется и при дальнейшем росте покрытия. При максимальных покрытиях (больше 2000×) определяются как метилированные все сайты GCGC, но сайтов с таким покрытием очень мало (16). Приблизительно 2/3 цитозинов во втором положении искомого сайта имеют покрытие меньше, чем 100 ридов, однако и среди этих цитозинов более 80% определяются как метилированные.

В то же время для контекстов NCGN, отличных от GCGC, метилирование нехарактерно (рис. 1б). Низкое качество ридов приводит к тому, что такие контексты в некотором количестве определяются как метилированные. Число это не очень велико – на уровне покрытия 1500× имеется 59 из 3448 контекстов NCGN, отличных от GCGC и определяемых как метилированные (рис. 1б).

Индекс уверенности DeepSignal играет исключительно важную роль при определении метилирования конкретных CG-динуклеотидов. Если не фильтровать риды по этому индексу, многие кон-

Таблица 1. Процент различных контекстов NCGN, отличных от GCGC, метилированных в геноме *E. coli*, несущем рекомбинантную метилтрансферазу M.HspAI

	Контекст	Доля метилированных
0	ACGC	0.17
1	GCGT	0.16
2	CCGC	0.13
3	GCGA	0.12
4	GCGG	0.10
5	CCGG	0.07
6	CCGA	0.09
7	CCGT	0.05
8	ACGG	0.04
9	ACGT	0.03
10	ACGA	0.03
11	TCGC	0.01
12	TCGG	0
13	TCGA	0
14	TCGT	0

тексты, являющиеся мишенями ДНК-метилтрансферазы, при большом покрытии определяются как неметилированные, а с другой стороны, около 6% контекстов NCGN с высоким покрытием определяются как метилированные (данные не показаны). Подавляющее большинство этих позиций определится «правильно» (метилированными являются только сайты узнавания ДНК-метилтрансферазы) после фильтрации ридов по индексу уверенности DeepSignal.

Отдельный интерес представляет собой анализ случаев, в которых метилирование определяется вне контекстов, распознаваемых метилтрансферазой. В табл. 1 даны частоты разных контекстов NCGN, отличных от GCGC, в которых ошибочно определялось метилирование. Поскольку с ростом покрытия доля контекстов NCGN, распознаваемых как метилированные, падает практически до нуля, мы предполагаем, что имеем дело с ошибками определения сайтов метилирования, а не с аберрантным метилированием метилтрансферазой.

Из табл. 1 видно, что комплементарные тетрануклеотиды ACGC и GCGT имеют высокие и близкие по величине частоты ошибочно определяемого метилирования. Однако в случае других комплементарных тетрануклеотидов это правило не всегда соблюдается. Например, тетрануклеотид TCGC почти никогда не определяется как метилированный, в то же время комплементарный ему контекст GCGA, напротив, часто выявляется как метилированный.

Данные табл. 1 показывают также, что тетрануклеотиды TCGN практически никогда не определяются как метилированные, а тетрануклеотиды ACGN показывают незначительный уровень ошибочного метилирования по сравнению с остальными сайтами. Причины такой зависимости ошибки от контекста могут быть связаны как с формированием сигнала MinION, так и с особенностяими работы программы DeepSignal, определяющей метилирование цитозина.

Секвенирование генома *E. coli* с клонированным геном ДНК-метилтрансферазы M.HpaII (контекст метилирования CCGG). Покрытие генома штамма *E. coli* с ДНК-метилтрансферазой M.HpaII ридами MinION с высоким индексом уверенности DeepSignal оказалось более низким, чем у штамма, несущего ДНК-метилтрансферазу M.HspAI. Более низким оказался и уровень метилирования контекстов CCGG, являющихся мишенями M.HpaII (рис. 2а). Показательно, что даже среди сайтов, характеризующихся очень высоким покрытием ридами MinION, существует около 10% неметилированных контекстов, причем для малой доли сайтов с очень высоким покрытием процент контекстов, определенных как неметилированные, не падает, а даже растет. Так, из 47 сайтов CCGG с покрытием не менее 1700× 12 сайтов не метилированы, что составляет практически четвертую часть. Тем не менее при покрытии от 200× и более приблизительно 10% сайтов CCGG определяются как неметилированные, т. е. около 90% таких контекстов имеют последовательность C(5mC)GG (приблизительно 1400 сайтов в геноме). Для контестов CCGG с покрытием меньше 100× как метилированные определяются приблизительно 85% сайтов. Данные рестрикционного анализа генома не показывают гидролиза геномной ДНК рестриктазой HpaII, расщепляющей неметилированные сайты CCGG (данные не показаны), причем наличие даже менее 3% неметилированных сайтов давало бы видимую картину расщепления. Таким образом, 10% неметилированных контекстов CCGG с большей вероятностью объясняются ненадежной обработкой сигнала с помощью программы DeepSignal, а не реальным недометилированием геномной ДНК.

Если рассматривать позиции в контекстах NCGN, отличных от CCGG (рис. 2б), то в геноме *E. coli*, несущей ген метилтрансферазы M.HpaII, такие контексты определяются как практически неметилированные уже при покрытии в 200×. При покрытии 700× определяются как метилиро-

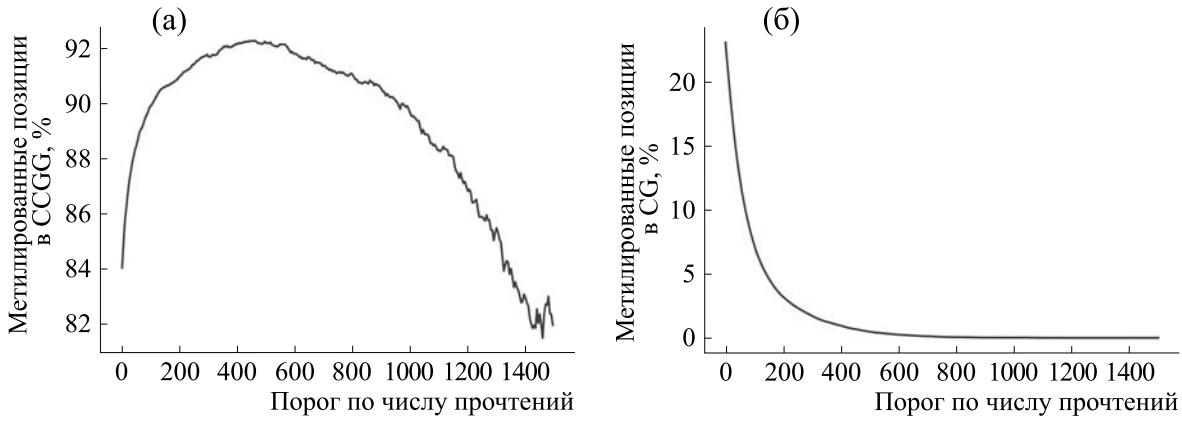

Рис. 2. Количество контекстов, определенных как метилированные, в зависимости от минимального покрытия ридами MinION: (а) – контексты CCGG, (б) – контексты NCGN, отличные от CCGG. Риды фильтрованы по индексу уверенности DeepSignal.

ванные только 0.13% контекстов NCGN, отличных от CCGG (53 контекста из 41002).

Как видно из табл. 2, контекст CCGT и комплементарный ему ACGG наиболее часто ошибочно определяются как метилированные. В случае сайта GCGC (табл. 1) максимальные значения ошибочного определения метилирования наблюдались для комплементарных тетрануклеотидов GCGT и ACGC. Таким образом, при определении метилированных сайтов GCGC (сайт узнавания M.HspAI) и CCGG (сайт узнавания M.HpaII) максимальное значение ошибочно установленного метилирования наблюдается для этих же сайтов, но с заменой последнего нуклеотида на T (или первого основания на A в случае комплемента).

Таблица 2. Процент различных контекстов NCGN, отличных от CCGG, метилированных в геноме *E. coli*, несущем рекомбинантную метилазу M.HpaII

	Контекст	Доля метилированных
0	CCGT	0.27
1	ACGG	0.21
2	GCGG	0.12
3	CCGA	0.18
4	CCGC	0.05
5	GCGA	0.05
6	GCGT	0.04
7	GCGC	0.02
8	ACGC	0.02
9	ACGA	0.03
10	ACGT	0.01
11	TCGG	0.01
12	TCGC	0
13	TCGA	0
14	TCGT	0

В случае M.HpaII, так же как и в случае с M.HspAI, тетрануклеотиды TCGN практически не определяются как метилированные, а тетрануклеотиды ACGN показывают незначительный уровень ошибочного метилирования по сравнению с остальными сайтами.

ЗАКЛЮЧЕНИЕ

В настоящей работе мы показали, что в случае высокого уровня покрытия при секвенировании ДНК система MinION может определять наличие 5-метилцитозина в ДНК, практически исключая ложноположительные результаты. При этом определение 5-метилцитозина в различных контекстах ДНК отличается. Для сайтов G(5mC)GC доля определяемых метилированных контекстов растет с ростом покрытия практически до 100%, однако имеется некоторое количество других тетрануклеотидов NCGN, которые ошибочно определяются прибором как метилированные. В случае сайта C(5mC)GG процент ошибочного определения 5-метилцитозина в тетрануклеотидах NCGN очень мал.

Для сайтов C(5mC)GG доля определяемых метилированных контекстов растет с ростом покрытия приблизительно до 90%, причем, по-видимому, часть контекстов не определяется как метилированные из-за неудачной работы программы DeepSignal. Почему DeepSignal показывает такую разную эффективность на двух контекстах, еще предстоит узнать.

Таким образом, полученные результаты секвенирования на приборе MinION двух геномных ДНК, имеющих 5-метилцитозин в различных контекстах, позволяют говорить о возможности использования прибора для достоверного определения 5-метилцитозина в геномной ДНК.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации (грант № RFMEFI60419X0218).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая работа не содержит описания исследований с использованием людей и животных в качестве объектов.

СПИСОК ЛИТЕРАТУРЫ

1. R. R. Wick, L. M. Judd, and K. E. Holt, Genome Biol. **20**, 129 (2019).
2. M. Stoiber, J. Quick, R. Egan, et al., Biorxiv 094672 (2017). DOI: 10.1101/094672
3. J. T. Simpson, R. E. Workman, P. C. Zuzarte, et al., Nat. Methods **14**, 407 (2017).
4. M. Ehrlich, Epigenomics **1** (2), 239 (2009).
5. G. P. Pfeifer, Int. J. Mol. Sci. **19** (4), 1166 (2018).
6. Q. Liu, L. Fang, G. Yu, D. Wang, et al., Nat. Commun. **10** (1), 2449 (2019).
7. P. Ni, N. Huang, Z. Zhang, et al., Bioinformatics **35** (22), 4586 (2019).
8. N. J. Loman, J. Quick, and J. T. Simpson, Nat. Methods, **12**, 733 (2015).
9. C. L. Smith, S. R. Klso, and C. R. Cantor, *Genome analysis: A Practical Approach*, Ed. by K. Davis (URL Press, Oxford, UK, 1987).

Efficiency of Identification of 5-Methylcytosine in *Escherichia coli* DNA Cells that Carry Genes of Bacterial DNA-Methyltransferases Using an Oxford Nanopore Device

V.V. Ilinsky*, **, E.M. Kozlova*, *, S. Kh. Degtyarev****, N.K. Yankovsky*, *****, and V.J. Makeev*, ***, ******, *********

**Vavilov Institute of General Genetics, Russian Academy of Sciences, ul. Gubkina 3, Moscow, 119991 Russia*
***Genotek, Nastavnicheskiy per. 17/1, Moscow, 105120 Russia*
****Moscow Institute of Physics and Technology, Institutskiy per. 9, Dolgoprudny, Moscow Region, 141701 Russia*
***** Scientific Production Association "SibEnzyme", ul. Akad. Timakova 2/12, Novosibirsk, 630060 Russia*
******Lomonosov Moscow State University, Vorobyovy Gory 1, Moscow, 119991 Russia*
*******Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, ul. Vavilova 32, Moscow, 119991 Russia*
********National Research Center «Kurchatov Institute» – GOSNIIGENETIKA, 1-i Dorozhnyi proezd 1, Moscow, 117545 Russia*

The MinION system (Oxford Nanopore Technologies Limited) was used for direct sequencing of genomic DNA of two recombinant *E. coli* strains. In one case, the cells contained a plasmid with the M.HpaII gene of DNA methyltransferase, which methylates the second cytosine in CCGG site; in the second case, the *E. coli* strain contained M.HspAI DNA methyltransferase, which modifies the central cytosine in the GCGC sequence. In both cases, DNA methyltransferases methylate cytosine to 5-methylcytosine. It has been shown that when DNA is sequenced at high coverage, the presence of 5-methylcytosine in DNA can be detected with high accuracy. In particular, in E. coli DNA containing the cloned gene of DNA methyltransferase M.HspAI, at 1300× coverage, 98.9% of the GCGC sites are identified as methylated at the first cytosine. At the same time, only 0.09% of the remaining tetranucleotides, which have the CpG dinucleotide in the middle, give a false positive result being identified as methylated at the central cytosine. In the presence of the gene of DNA methyltransferase M.HpaII, among the positions covered by more than 700 reads, 91.3% of all CCGG sites are identified as methylated, while only 0.13% of other tetranucleotides with a central CG-dinucleotide are identified as sites containing 5-methylcytosine in the second position. Therefore, when this method is used, at least 700–1000× coverage is needed for accurate measurements of 5-methylcytosine and elimination of false-positive results.

Keywords: nanopores, 5-methylcytosine, methylation, E. coli, DeepSignal