Совершенствование анализа результатов секвенирования ДНК по Сэнгеру: компьютерная программа SeqBase

Танас А.С.[1], Симонова О.А.[1], Абрамычева Н.Ю.[2], Стрельников В.В.[1]

1 — ФГБНУ «Медико-генетический научный центр имени академика Н.П. Бочкова»
115522, г. Москва, ул. Москворечье, д. 1

2 — ФГБНУ «Научный центр неврологии»
125367, г. Москва, Волоколамское шоссе, д. 80

Введение. Программное обеспечение, предоставляемое производителями автоматических генетических анализаторов, в большинстве случаев позволяет провести адекватный анализ результатов секвенирования ДНК по Сэнгеру для матриц с составом нуклеотидов, близким к эквивалентному. Однако для рассмотрения результатов секвенирования матриц, отличающихся неэквивалентным нуклеотидным составом, требуется проводить анализ электрофореграмм с сохранением информации об интенсивности сигналов флуоресценции. В особенности это касается секвенирования ДНК, модифицированной бисульфитом натрия.

Цель: разработать и апробировать в практике научных исследований компьютерную программу для обеспечения адекватного анализа электрофореграмм секвенирования ДНК по Сэнгеру на основе бережного отношения к первичным данным и аккуратного определения базовых линий в спектральных каналах отдельных нуклеотидов.

Методы. Программа SeqBase написана на языке C#, программная платформа .NET Framework 4.0, и выполняется в среде исполнения CLR (Common Language Runtime) для операционных систем семейства Windows. Адрес установочного пакета программы SeqBase: http://www.epigenetic.ru/projects/seqbase.

Результаты. Разработана компьютерная программа, предназначенная для анализа первичных результатов секвенирования по Сэнгеру (хроматограмм капиллярного электрофореза), полученных на автоматических генетических анализаторах и представленных в файлах формата ABIF (*.ab1), обеспечивающая следующие возможности: 1) просмотр исходных электрофореграмм как в общем виде, так и раздельно по спектральным каналам; 2) кадрирование области анализа; 3) сглаживание сигналов; 4) ручная установка базовой линии по каждому из спектральных каналов; 5) сведение базовых линий по всем каналам; 6) ручная коррекция подвижности фрагментов ДНК в зависимости от типа флуоресцентной метки терминирующего нуклеотида. Апробация программы успешно проведена в рамках ряда исследований, результаты которых опубликованы в рецензируемых научных изданиях.

Заключение. Использование программы SeqBase целесообразно для анализа результатов секвенирования по Сэнгеру матриц ДНК с неэквивалентным нуклеотидным составом, в особенности, модифицированных бисульфитом натрия, во избежание получения ложных результатов и для уточнения количественных оценок.

Ключевые слова: секвенирование ДНК по Сэнгеру, бисульфитное секвенирование, анализ секвенограмм, компьютерная программа SeqBase.

Для цитирования: Танас А.С., Симонова О.А., Абрамычева Н.Ю., Стрельников В.В. Совершенствование анализа результатов секвенирования ДНК по Сэнгеру: компьютерная программа SeqBase. *Медицинская генетика* 2021; 20(10): 33-39.
DOI: 10.25557/2073-7998.2021.10.33-39

Автор для корреспонденции: *Стрельников Владимир Викторович*, **e-mail:** vstrel@list.ru
Финансирование. Работа выполнена в рамках государственного задания Минобрнауки России для ФГБНУ «МГНЦ».
Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Поступила: 25.09.2021.

Improving the Analysis of DNA Sanger Sequencing Results: SeqBase Computer Program

Tanas A.S.[1], Simonova O.A.[1], Abramycheva N. Yu.[2], Strelnikov V.V.[1]

1 — Research Centre for Medical Genetics
1 Moskvorechye st., Moscow, 115522, Russian Federation

2 — Research Center of Neurology
80 Volokolamskoye Shosse, Moscow, 125367, Russian Federation

Background. The software provided by the manufacturers of automatic genetic analyzers, in most cases, allows an adequate analysis of the results of Sanger DNA sequencing for templates with a nucleotide composition close to the equivalent. However, to consider the results of sequencing of templates with non-equivalent nucleotide composition, it is necessary to analyze electrophore-

grams with preservation of primary information on the intensity of fluorescence signals. This is especially important for the sequencing of DNA modified with sodium bisulfite.

Aim: to develop and validate in the practice of scientific research a computer program that ensures adequate analysis of electrophoregrams of Sanger DNA sequencing based on preservation of the primary data and on accurate determination of baselines in the spectral channels of individual nucleotides.

Methods. The SeqBase program is written in C#, the programming platform .NET Framework 4.0, and runs in the CLR (Common Language Runtime) for Windows operating systems. SeqBase installation package address is http://www.epigenetic.ru/projects/seqbase.

Results. A computer program has been developed designed to analyze the primary results of Sanger sequencing (chromatograms of capillary electrophoresis) obtained from automatic genetic analyzers and presented in files of the ABIF (*.ab1) format, which provides the following functions: 1) viewing the original electrophoregrams both in general form and separately by spectral channels; 2) cropping the area of analysis; 3) signal smoothing; 4) manual setting of the baseline for each of the spectral channels; 5) convergence of baselines on all channels; 6) manual correction of the mobility of DNA fragments depending on the type of fluorescent label of the terminating nucleotide. The program has been successfully tested in a number of studies, the results of which have been published in peer-reviewed scientific journals.

Conclusion. The use of the SeqBase program is advisable for the analysis of the results of Sanger sequencing of DNA templates with non-equivalent nucleotide composition, especially those modified with sodium bisulfite, to avoid false results and to clarify quantitative estimates.

Keywords: DNA Sanger sequencing, bisulfite sequencing, sequenogram analysis, SeqBase computer program.

For citation: Tanas A.S., Simonova O.A., Abramycheva N. Yu., Strelnikov V.V. Improving the Analysis of DNA Sanger Sequencing Results: SeqBase Computer Program. *Medicinskaya genetika [Medical Genetics]* 2021; 20(10): 33-39. (In Russ.)
DOI: 10.25557/2073-7998.2021.10.33-39

Corresponding author: *Vladimir V. Strelnikov*, **e-mail:** vstrel@list.ru
Funding. The research was carried out within the state assignment of Ministry of Science and Higher Education of the Russian Federation for RCMG.
Conflict of interest. The authors declare no conflict of interest.
Accepted: 25.09.2021.

Введение

Программное обеспечение, предоставляемое производителями автоматических генетических анализаторов, в большинстве случаев позволяет провести адекватный анализ результатов секвенирования ДНК по Сэнгеру для матриц с составом нуклеотидов, близким к эквивалентному. Однако для рассмотрения результатов секвенирования матриц, отличающихся неэквивалентным нуклеотидным составом, требуется проводить анализ электрофореграмм с сохранением информации об интенсивности сигналов флуоресценции. В особенности это касается секвенирования ДНК, модифицированной бисульфитом натрия.

При обработке геномной ДНК бисульфитом натрия происходит дезаминирование неметилированного цитозина с образованием урацила, при этом 5-метилцитозин дезаминированию не подвергается. Бисульфитиндуцированные различия последовательностей в образцах ДНК с разным содержанием 5-метилцитозина позволяют дифференцировать метилированную и неметилированную ДНК. Обработанная бисульфитом ДНК может быть напрямую использована для ПЦР-анализа, в котором урациловые и тиминовые остатки реплицируются как аденин и только 5-метилцитозиновые остатки реплицируются как гуанин [1] (**рис. 1**).

Секвенирование ПЦР-продукта, полученного с бисульфит-модифицированной ДНК [2], считается золотым стандартом анализа метилирования ДНК на уровне отдельных нуклеотидов. В то же время особенности последовательности ДНК после бисульфитной модификации предъявляют повышенные требования к технике собственно секвенирования и к инструментам анализа результирующих электрофореграмм.

В результате бисульфитной обработки происходит значительное снижение доли цитозина в последовательности ДНК. Предельный случай представляет полностью неметилированная ДНК, которая после бисульфитной обработки вовсе не содержит остатков цитозина (**рис. 2**).

Традиционное программное обеспечение для анализа результатов автоматического секвенирования предполагает усредненный нуклеотидный состав секвенируемых последовательностей ДНК. Отсутствие на первичных электрофореграммах бисульфитных сиквенсов сигналов, соответствующих цитозину, может приводить к компенсаторному усилению шумовых сигналов в соответствующем спектральном канале. Для обеспечения адекватного анализа электрофореграмм на основе бережного отношения к первичным

данным и аккуратного определения базовых линий в спектральных каналах отдельных нуклеотидов нами разработана компьютерная программа SeqBase.

Методы

Разработка программного обеспечения

Программа SeqBase написана на языке C#, программная платформа .NET Framework 4.0, и выполняется в среде исполнения CLR (Common Language Runtime) для операционных систем семейства Windows.

Загрузка и установка программы

Адрес установочного пакета SeqBase: http://www.epigenetic.ru/projects/seqbase. Установите программу, следуя инструкциям программы установки. Для запуска установленной программы SeqBase выберите исполняемый файл в меню «Программы», в папке RCMG.

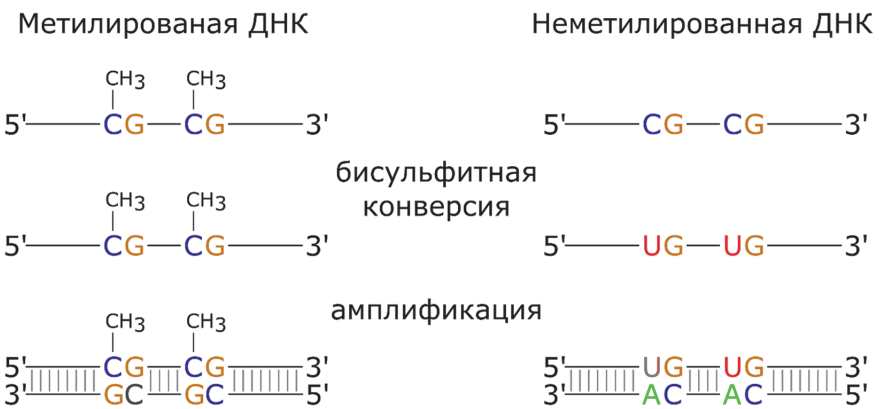

Рис. 1. Анализ метилирования ДНК методом ПЦР, опосредованный бисульфитной конверсией.

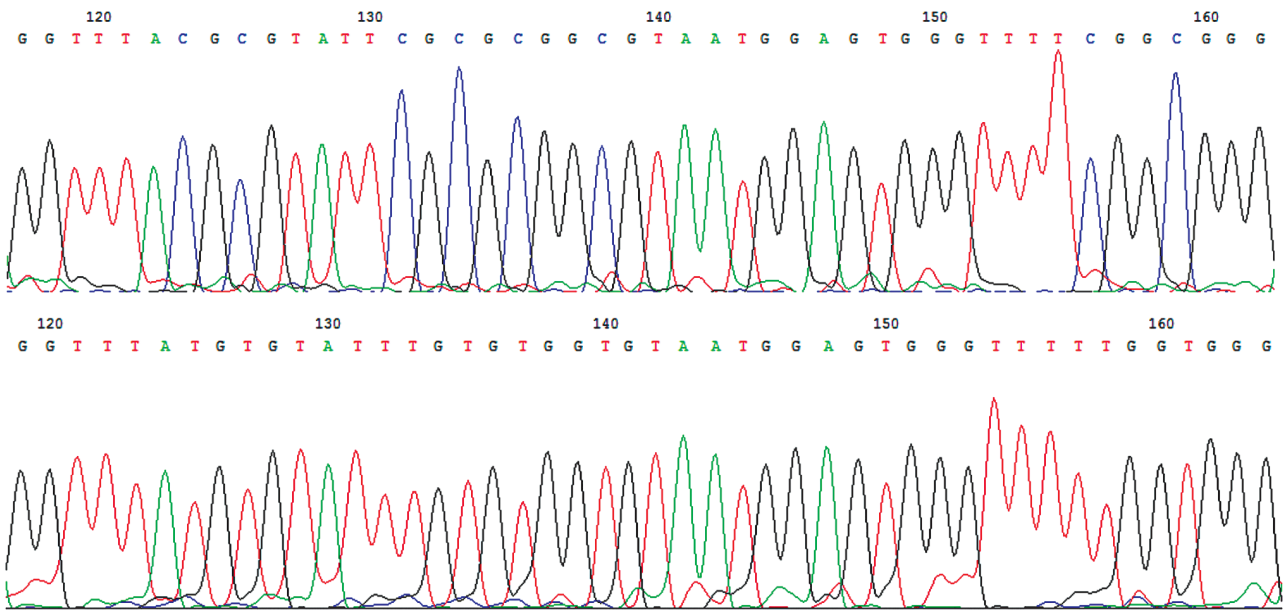

Рис. 2. Снижение доли цитозина в последовательности ДНК в результате бисульфитной обработки. Верхняя панель – электрофореграмма бисульфитного сиквенса полностью метилированной последовательности; наблюдается наличие сигналов цитозина в составе CpG-динуклеотидов. Нижняя панель – электрофореграмма бисульфитного сиквенса полностью неметилированной последовательности; наблюдается абсолютное отсутствие сигналов цитозина. Собственные результаты секвенирования; визуализация с использованием программы SeqBase.

Требования к пользовательской системе

32-разрядный (×86) или 64-разрядный (×64) процессор с тактовой частотой 1ГГЦ или выше, 250 Мб ОЗУ, 100 Мб свободного пространства на диске, ОС Windows XP или более поздней версии, .Net Framework 4.0.

Формат входных данных

Программа предназначена для анализа первичных результатов секвенирования по Сэнгеру (хроматограмм капиллярного электрофореза), полученных на автоматических генетических анализаторах и представленных в файлах формата ABIF (*.ab1).

Результаты исследования и их обсуждение

Типы ошибок анализа электрофореграмм секвенирования ДНК по Сэнгеру, обусловленных неэквивалентным нуклеотидным составом матриц

Традиционное программное обеспечение для анализа результатов автоматического секвенирования предполагает усредненный нуклеотидный состав секвенируемых последовательностей ДНК. Отсутствие в первичных электрофореграммах бисульфитных сиквенсов сигналов, соответствующих минорному или отсутствующему нуклеотиду последовательности, может приводить к компенсаторному усилению шумовых сигналов в соответствующем спектральном канале. Такие шумовые сигналы на последнем этапе анализа представляются избыточно усиленными сигналами минорного нуклеотида в составе ДНК, особенно при общем низком уровне сигналов и для полностью неметилированных последовательностей ДНК, лишенных цитозина в результате бисульфитной конверсии (**рис. 3**).

Результаты бисульфитного секвенирования полностью метилированных последовательностей ДНК, содержащих в своем составе достаточно большое количество цитозина (участки метилированных CpG-островков) разрешаются традиционным программным обеспечением более адекватно (**рис. 4**).

Ошибки анализа электрофореграмм продуктов реакции терминального мечения могут приводить к ложным выводам относительно наличия минорного нуклеотида в позициях последовательности ДНК. В частности, в случае бисульфитного секвенирования, – к ложным выводам относительно состояния метилирования CpG-динуклеотидов в исследуемых последовательностях (**рис. 5**, верхняя панель), в особенности, если интерпретация результатов осуществляется машинным методом.

Ошибки анализа электрофореграмм метилспецифического секвенирования бывают не только качественного порядка («ложное метилирование»), но и количественного. Стремление программного обеспечения компенсировать значительное снижение доли цитозина, повышая чувствительность по соответствующему спектральному каналу, приводит к невозможности адекватной количественной оценки метилирования конкретных CpG-динуклеотидов (**рис. 6**).

Компьютерная программа SeqBase, примеры результатов работы которой в сравнении с традицион-

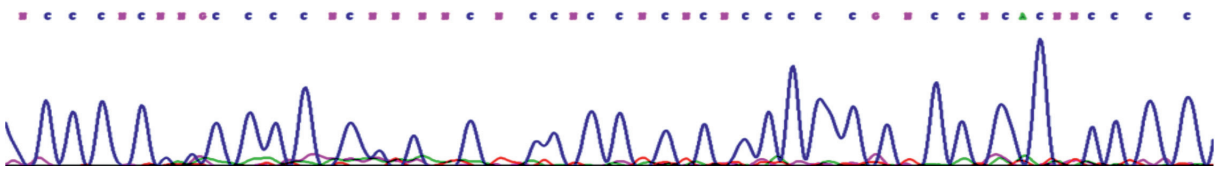

Рис. 3. Пример неадекватного компьютерного анализа результатов бисульфитного секвенирования полностью неметилированной последовательности ДНК вследствие компенсаторного усиления шумовых сигналов в спектральном канале, соответствующем цитозину. Собственные результаты секвенирования; визуализация с использованием традиционного программного обеспечения - Sequencing Analysis (ThermoFisher, США).

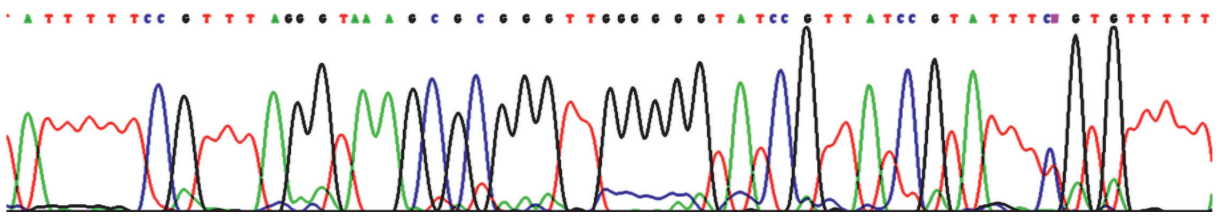

Рис. 4. Пример компьютерного анализа результатов бисульфитного секвенирования полностью метилированной последовательности ДНК, богатой CpG-динуклеотидами. Собственные результаты секвенирования; визуализация с использованием традиционного программного обеспечения.

ным программным обеспечением представлены на рис. 5 и 6, разработана для обеспечения адекватного анализа электрофореграмм на основе бережного отношения к первичным данным и аккуратного определения базовых линий в каналах отдельных нуклеотидов.

Основные функции компьютерной программы SeqBase: 1) просмотр исходных электрофореграмм как в общем виде, так и раздельно по спектральным каналам; 2) кадрирование области анализа; 3) сглаживание сигналов; 4) ручная установка базовой линии по каждому из спектральных каналов; 5) сведение базовых линий по всем каналам; 6) ручная коррекция подвижности фрагментов ДНК в зависимости от типа флуоресцентной метки терминирующего нуклеотида.

Программа SeqBase использовалась в ряде опубликованных научных работ для проведения адекватного анализа результатов бисульфитного секвенирования ДНК по Сэнгеру [3–5], уточнения наличия сигналов мозаичных аллелей с низкой долей представленности в образце биологического материала [6,7] и/или подготовки иллюстраций, наилучшим образом отражающих действительные результаты секвенирования по Сэнгеру [4, 5, 7].

В исследовании «Эпигенетика болезни Фридрейха: метилирование области экспансии (GAA)n-повторов

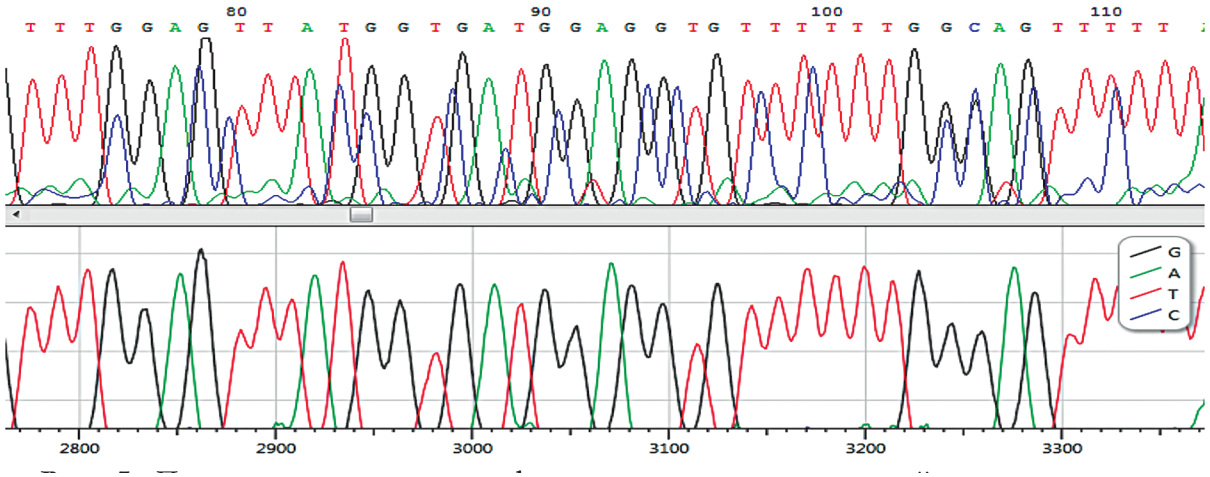

Рис. 5. Примеры анализа электрофореграммы неметилированной последовательности ДНК с использованием традиционного программного обеспечения (верхняя панель) и программы SeqBase (нижняя панель).

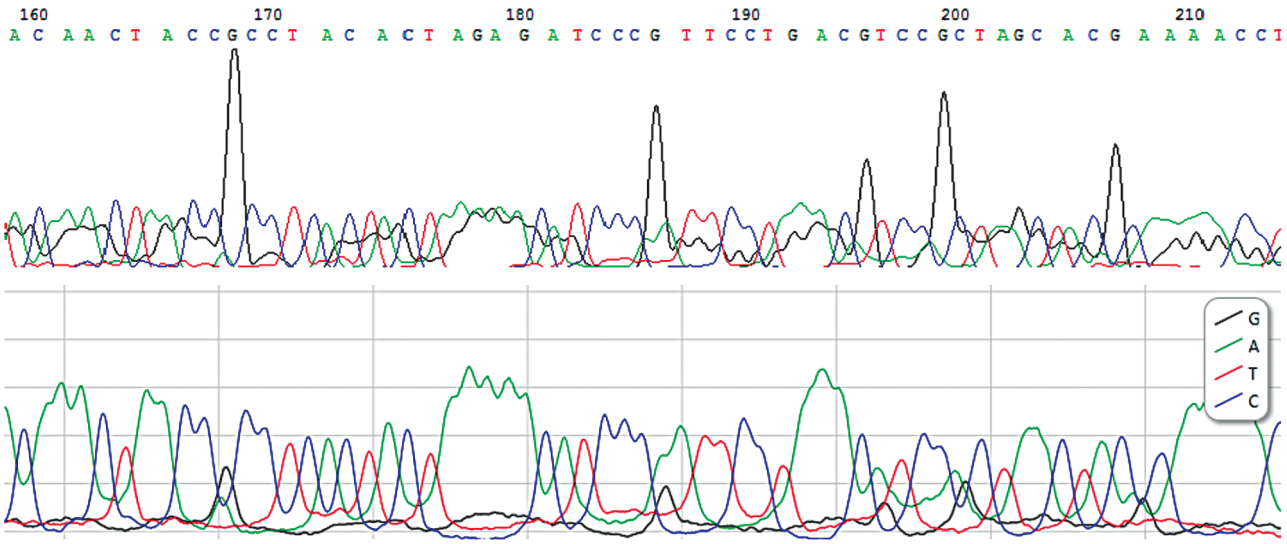

Рис. 6. Примеры анализа результатов бисульфитного секвенирования последовательности ДНК с использованием традиционного программного обеспечения (верхняя панель) и программы SeqBase (нижняя панель). На верхней панели видны чрезмерно усиленные сигналы гуанина, соответствующие цитозину в обратно комплементарной цепи.

гена *FXN*» изучалась степень метилирования CpG-сайтов в области GAA-повторов в интроне 1 гена *FXN*. Паттерн метилирования определяли методом прямого секвенирования соответствующих участков ДНК после бисульфитной обработки. Различия в количественной оценке представленности цитозина в составе бисульфитмодифицированной ДНК по результатам анализа секвенограмм, обработанных традиционным программным обеспечением и программой SeqBase, показаны на **рис. 7**. Ожидаемо, традиционное программное

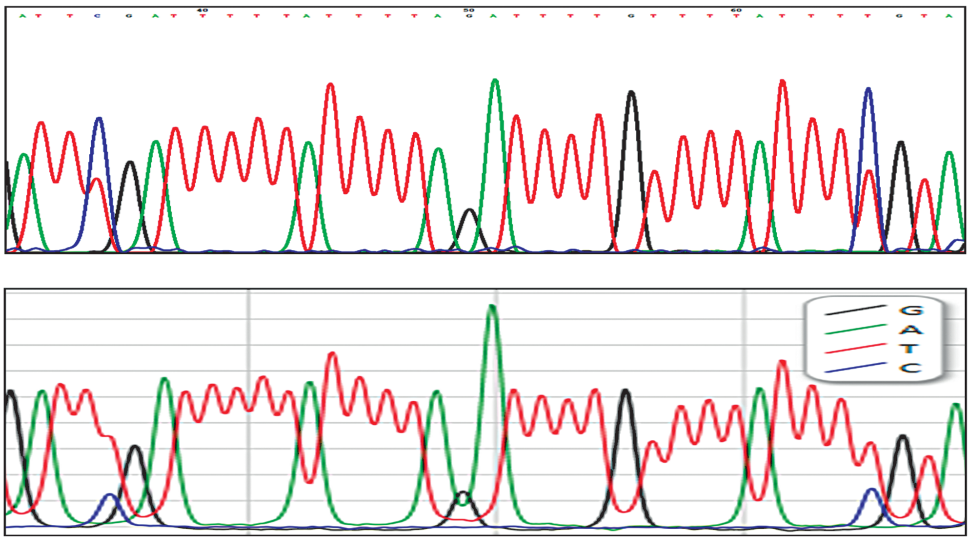

Рис. 7. Различия в количественной оценке представленности цитозина в составе бисульфитмодифицированной ДНК по результатам анализа секвенограмм, обработанных традиционным программным обеспечением (верхняя панель) и программой SeqBase (нижняя панель). Представлены результаты бисульфитного секвенирования области экспансии (GAA)n-повторов гена *FXN*, полученные при проведении исследования [3].

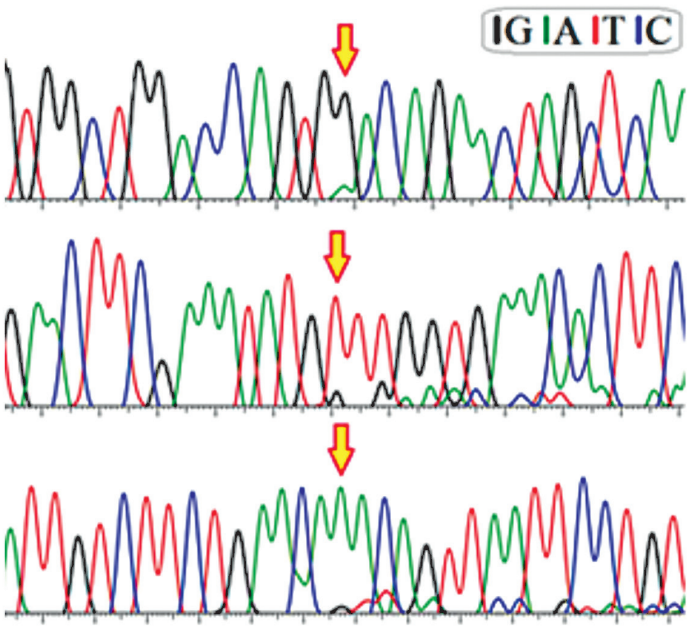

Рис. 8. Примеры иллюстраций наличия сигналов мозаичных аллелей с низкой долей представленности в образце биологического материала, выполненных с использованием компьютерной программы SeqBase. Верхняя панель – мозаичная точковая мутация. Средняя и нижняя панели – мозаичные тетрануклеотидные делеции [7].

обеспечение предлагает завышенные количественные оценки представленности цитозина в CpG-контексте бисульфитмодифицированной ДНК.

В исследованиях аномального деметилирования генов онконевральных белков аррестина и рекаверина в злокачественных новообразованиях почки «Autoantibody against arrestin-1 as a potential biomarker of renal cell carcinoma» [4] и «The cancer-retina antigen recoverin as a potential biomarker for renal tumors» [5] программа SeqBase была использована для картирования дифференциально метилированных участков изучаемых генов и для подготовки иллюстраций.

Программа SeqBase применялась для уточнения наличия сигналов мозаичных аллелей с низкой долей представленности в образце биологического материала в исследовании «Соматический мозаицизм при нейрофиброматозе первого типа» [6, 7], а также для подготовки иллюстраций по результатам исследования (**рис. 8**).

Заключение

Разработана компьютерная программа SeqBase, предназначенная для анализа первичных результатов секвенирования по Сэнгеру (хроматограмм капиллярного электрофореза), полученных на автоматических генетических анализаторах и представленных в файлах формата ABIF (*.ab1). Апробация программы успешно проведена в рамках ряда исследований, результаты которых опубликованы в рецензируемых научных изданиях. Использование программы SeqBase целесообразно для анализа результатов секвенирования по Сэнгеру матриц ДНК с неэквивалентным нуклеотидным составом, в особенности, модифицированных бисульфитом натрия, во избежание получения ложных результатов и для уточнения количественных оценок. Использование программы позволяет визуализировать сигналы мозаичных аллелей с низкой долей представленности в образце биологического материала, что важно при анализе мозаичных мутаций у пациентов с генетическими заболеваниями.

Литература

1. Hayatsu H. Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis — a personal account. *Proceedings of the Japan Academy. Series B* 2008; 84(8): 321-330. doi: 10.2183/pjab.84.321.
2. Frommer M., McDonald L.E., Millar D.S., Collis C.M., Watt F., Grigg G.W. et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences* 1992, 89(5): 1827-1831. doi: 10.1073/pnas.89.5.1827.
3. Abramycheva N.Y., Fedotova E.Y., Nuzhnyi E.P., Nikolaeva N.S., Klyushnikov S.A., Ershova M.V. et al. Epigenetics of Friedreich's Disease: Methylation of the (GAA) n-Repeats Region in FXN Gene. *Annals of the Russian academy of medical sciences* 2019; 74(2): 80-87.doi: 10.1038/s10038-019-0696-z.
4. Baldin A.V., Grishina A.N., Korolev D.O., Kuznetsova E.B., Golovastova M.O., Kalpinskiy A.S. et al. Autoantibody against arrestin-1 as a potential biomarker of renal cell carcinoma. *Biochimie* 2019; *157:* 26-37. doi: 10.1016/j.biochi.2018.10.019.
5. Golovastova M.O., Tsoy L.V., Bocharnikova A.V., Korolev D.O., Gancharova O.S., Alekseeva E.A. et al. The cancer-retina antigen recoverin as a potential biomarker for renal tumors. *Tumor Biolog.* 2016; 37(7): 9899-9907. doi: 10.1007/s13277-016-4885-5.
6. Karandasheva K., Pashchenko M., Tanas A. S., Strelnikov V. V., Kuznetsova E. Improving detection level of somatic mosaicism in neurofibromatosis type 1. *Annals of Oncology* 2019; 30: v23-v24. doi: 10.1093/annonc/mdz238.083.
7. Карандашева К. О., Пащенко М. С., Дёмина Н. А., Акимова И. А., Макиенко О. Н., Петухов, М. С., с соавт. Соматический мозаицизм при нейрофиброматозе первого типа. *Медицинская генетика* 2019; 18(5):28-36. doi: 10.25557/2073-7998.2019.05.28-36.

References

1. Hayatsu H. Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis — a personal account. *Proceedings of the Japan Academy. Series B* 2008; 84(8): 321-330. doi: 10.2183/pjab.84.321.
2. Frommer M., McDonald L.E., Millar D.S., Collis C.M., Watt F., Grigg G.W. et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences* 1992, 89(5): 1827-1831. doi: 10.1073/pnas.89.5.1827.
3. Abramycheva N.Y., Fedotova E.Y., Nuzhnyi E.P., Nikolaeva N.S., Klyushnikov S.A., Ershova M.V. et al. Epigenetics of Friedreich's Disease: Methylation of the (GAA) n-Repeats Region in FXN Gene. *Annals of the Russian academy of medical sciences* 2019; 74(2): 80-87.doi: 10.1038/s10038-019-0696-z.
4. Baldin A.V., Grishina A.N., Korolev D.O., Kuznetsova E.B., Golovastova M.O., Kalpinskiy A.S. et al. Autoantibody against arrestin-1 as a potential biomarker of renal cell carcinoma. *Biochimie* 2019; *157:* 26-37. doi: 10.1016/j.biochi.2018.10.019.
5. Golovastova M.O., Tsoy L.V., Bocharnikova A.V., Korolev D.O., Gancharova O.S., Alekseeva E.A. et al. The cancer-retina antigen recoverin as a potential biomarker for renal tumors. *Tumor Biolog.* 2016; 37(7): 9899-9907. doi: 10.1007/s13277-016-4885-5.
6. Karandasheva K., Pashchenko M., Tanas A. S., Strelnikov V. V., Kuznetsova E. Improving detection level of somatic mosaicism in neurofibromatosis type 1. *Annals of Oncology* 2019; 30: v23-v24. doi: 10.1093/annonc/mdz238.083.
7. Karandasheva K.O., Pashchenko M.S., Demina N.A., Akimova I.A., Makienko O.N., Petuhova M.S., et al. Somaticheskiy mozaitsizm pri neyrofibromatoze pervogo tipa [Somatic mosaicism in neurofibromatosis type 1]. *Meditsinskaya genetika [Medical Genetics]* 2019; 18(5): 28-36. (In Russ.). doi: 10.25557/2073-7998.2019.05.28-36.