МЕМОРАНДУМ О ВЗАИМОПОНИМАНИИ

ПРИЛОЖЕНИЕ III: ПРАВИЛА ДОСТУПА К ДАННЫМИ И РЕСУРСАМ IBOL

[Рассмотрено советом директоров iBOL (Исполнительным комитетом) 22 августа 2011]

ВВЕДЕНИЕ

Международный проект «Штрихкодирование живых форм» формирует референтную библиотеку коротких, стандартизированных участков последовательностей (ДНК-штрихкод) с целью молекулярной идентификации известных и облегчения выявления новых видов. На конференции в Форт Лаудердэйл[1], посвященной доступу к данным, полученным в ходе крупных научно-исследовательских проектов, был введен термин «проект общих ресурсов» ("community resource project"), определенный как «исследовательский проект, спланированный и осуществленный с целью создания информации, веществ или других материалов, непосредственное применение которых направлено на использование широким научным сообществом». Участники международного проекта «Штрихкодирование живых форм» (iBOL) в ходе проекта создают такой «общий ресурс» и, согласно с рекомендациям, принятыми на конференции в Форт-Лаудердэйл, проект iBOL направлен на быстрое введение данных в общий доступ. Такая позиция характерна для большинства крупных международных проектов в области генетики (например, международный проект «HapMap») и отражает политику введения данных в общий доступ, принятую в таких организациях как национальный институт исследования генома человека (National Human Genome Research Institute), Геном Канада (Genome Canada), Фонд Гордона и Бетти Мур (The Gordon and Betty Moore Foundation) в США, а также Велком Траст (The Welcome Trust) в Великобритании.

Политика открытости и доступности ресурсов (как отмечено выше) стремится ускорить получение такого результата, который принесет пользу человечеству путем создания возможности быстрого доступа к основной информации, создаваемой iBOL, – первичным последовательностям ДНК, ассоциированным с таксономическими определениями видов. Это соответствует рекомендациям по уровню открытости данных крупных проектов в области геномики (Field et al., 2008)[2]. Публикация более детального анализа последовательностей ДНК и их аннотаций (например, включающих множественное выравнивание, а также экспертную видовую идентификацию по отношению к каждому из сиквенсов) остается не менее важным аспектом в обеспечения публичности данных и ожидается что члены iBOL опубликуют данные результаты в разумные сроки. Участие в iBOL предполагает признание и следование следующим правилам:

[1] http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf

[2] Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, dePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kravitz S, Kyrpides N, Leebens-Mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, Methe B, Mizrachi I, Moxon R, Nelson K, Parkhill J, Proctor L, White O, Sanson S-A, Spiers A, Stevens R, Swift, P, Taylor C, Tateno Y, Tett A, Turner S, Ussery D, Vaughan B, Ward N, Whetzel T, San Gil I, Wilson G, Wipat A. The minimum information about a genome sequence (MIGS) specification. 2008. Nature Biotechnology 26, 541 - 547)

ТИП ДАННЫХ, СОЗДАВАЕМЫХ ПРОЕКТОМ IBOL:

В ходе реализации проекта iBOL появляются различные типы данных, некоторые из которых близки либо идентичны данным, получаемым в ходе геномных или молекулярно-биологических проектов, в то время как другие типы данных уникальны для iBOL. Основная цель проекта iBOL состоит в создании библиотеки референтных «штрихкодовых» последовательностей ДНК, полученных для 500 тысяч видов и представленных 5 миллионами особей. Каждая добавленная последовательность ассоциирована с записями, отражающими происхождение индивидуума в коллекции (такие как дата, место и обстоятельства сбора образца) и включает реально существующий ваучерный экземпляр, депонированный в обозначенной коллекции (музее). Ваучерный образец, цифровая фотография этого образца и точные географическое положение точки сбора должны быть привязаны к каждой депонированной последовательности ДНК, ДНК-штрихкоду, за исключением обстоятельств, когда указание точных координат нахождения может поставить под угрозу существование вида или популяции. Ключевым элементом аннотации библиотеки ДНК-штрихкодов является экспертное определение видовой принадлежности для каждого образца, причем видовое название должно соответствовать списку валидных таксономических названий.

Общепризнанно, что такой полный набор данных требует согласованных шагов и часто многоэтапной и сложной таксономической идентификации. Во многих случаях это требует экспертного заключения нескольких специалистов-систематиков и поэтому требуется намного больше времени, чем обычная публикация геномных данных. В то же время, отсрочка в публикации данных о последовательностях ДНК по причине длительной верификации таксономического определения каждой особи несовместима с политикой общедоступности данных, принятой в iBOL. Для того чтобы оценить прогресс в выполнении проекта и укрепить взаимное сотрудничество, члены iBOL должны быстро публиковать первичные данные последовательностей (файлы электрофореграмм) и предварительную (например, высокого уровня) таксономическую информацию, согласно процедуре, описанной ниже. Это упростит участие более широкого круга исследователей в плане признания необходимости дальнейшего уточнения таксономических аннотаций и курирования базы данных последовательностей после их предварительной публикации. Первоочередной задачей iBOL является создание базы данных ДНК- штрихкодов, то есть, последовательностей ДНК, которая была бы доступна и соответствовала нуждам, как широкой общественности, так и научному сообществу. Для успеха iBOL необходимо чтобы полноценная таксономической аннотация и процесс валидации таксономического определения проводился до того уровня, который представляется возможным в момент депонирования, и полный набор данных публиковался по возможности быстро. Ниже дается описание типов данных, участвующих в процессе создания библиотеки.

1. Данные об образце

Как только образец собран и представлен для анализа, ему назначается уникальный BOLD–идентификационный номер, а также идентификационный номер образца. Также

присутствуют биологические и географические данные, связанные с данным образцом или экземпляром. Ниже представлена минимально необходимая информация об образце:

- Страна/океан, где образец был собран;
- дата сбора.

Другие данные также могут быть представлены вместе с образцом или экземпляром, но они не строго обязательны при первичной публикации данных. Они могут включать:

- Цифровой образ образца (экземпляра). Хотя это не обязательно, но крайне желательно, чтобы фотография была представлена на каком-нибудь из этапов процесса обработки материала.
- GPS-координаты точки сбора (хотя это требование не является обязательным, но крайне желательно, чтобы GPS-координаты были приложены[3], за исключением определенных обстоятельств).

2. Название таксона/идентификация

Предварительная таксономическая принадлежность обычно приводится при представлении образца для секвенирования и необходима при первичной публикации данных. Она может быть на уровне семейства или иметь описательный характер, например, природная проба (environmental sample), и не предполагает конечного видового определения. Реальное определение рода и вида является целью, но это может быть невозможно в течение короткого времени, так что публикация данных не должна зависеть от аннотации последовательности с окончательной таксономической идентификацией.

3. Генетические данные

- Указание участка гена;
- Последовательность ПЦР-праймеров и условия ПЦР;
- Файлы хроматограмм, используемые в:
- сборке контига (конечной суммарной последовательности, ДНК-штрихкода)
- идентификационный номер таксона (BIN – BOLD Barcode Index Number)

КОНТРОЛЬ КАЧЕСТВА / УРОВЕНЬ ДОСТОВЕРНОСТИ ГЕНЕТИЧЕСКИХ ДАННЫХ

Перед публикацией данные должны пройти процесс оценки уровня достоверности / контроля качества (Quality Assessment/Quality Control, QA/QC). Предварительные данные, описанные выше, должны соответствовать следующим QA/QC критериям:

[3] http://www.nature.com/nature/journal/v453/n7191/pdf/453002a.pdf

i. Длина конечной последовательности должна включать больше 75% принятой длины маркера штрихкода[4] (например, 500 пар нуклеотидов для COI) с ожидаемой двукратной дешифровкой.
ii. Качество последовательности должно быть разумным (т.е. меньше 1% ошибочных оснований в окончательном обрезанном контиге, с ожидаемым средним значением по Фреду, Phred Score > 20).
iii. Последовательность не должна совпадать с обычными загрязнениями (например, с человеческой ДНК).
iv. Последовательность должна соответствовать предполагаемому таксономическому положению высокого уровня (например, ДНК-штрихкод должен кластеризоваться с родственными таксонами).

Части i – iii могут быть автоматизированы. Часть iv является важной и может требовать участия ручной обработки.

БАЗА ДАННЫХ ШТРИХКОДИРОВАНИЯ ЖИВЫХ СУЩЕСТВ (BOLD)

Вне зависимости от того, был ли образец обработан и ДНК-штрихкод получен в Центре ДНК-штрихкодирования Института Биоразнообразия Онтарио (BIO), либо в другом центре штрихкодирования – участнике iBOL, научное сообщество iBOL использует BOLD и/или международные зеркальные серверы BOLD как первичный инструмент для составления и публикации данных ДНК-штрихкодирования.

ВРЕМЕННЫЕ РАМКИ ПУБЛИКАЦИИ ДАННЫХ В ОБЩЕСТВЕННЫЕ БАЗЫ ДАННЫХ (GenBank)

Данные, представленные в BOLD проектами, связанными с iBOL, будут перенесены в GenBank до момента опубликования авторами статей по полученным результатам секвенирования. Публикация данных происходит в два этапа и осуществляется поквартально.

Этап I включает обнародование всех полученных последовательностей и таксономическую информацию высокого уровня. Эта ранняя публикация предназначена для предоставления информации, которая может быть полезна для других исследователей, а также для мониторинга прогресса роста числа ДНК-штрихкодов для каждой из рабочих групп iBOL. Это будет происходить в автоматическом режиме, включая данные, которые могут быть опубликованы после компьютеризированной проверки качества и присвоения последовательностям BIN-индексов. В частности, следующие данные будут опубликованы на первом этапе в течение одной недели после получения последовательностей:

- Место сбора: вся доступная информация;

[4] CBOL/INSDC approved BARCODE marker (e.g. 5' COI for animals)

- Дата сбора;
- Таксономическая информация: определение до уровня отряда и BIN;
- Информация о последовательностях: автоматически собранная последовательность, хроматограммы, использованные праймеры, название центра, в котором получена последовательность.
- Идентификаторы базы данных: идентификационный номер BOLD и идентификаторы экземпляра (ваучерный номер, коллекция, коллекционный номер).

Этап II состоит в публикации дополнительной информации, которая требует ручной обработки и детального таксономического определения. Этап II обычно происходит, когда рукопись предоставляется для публикации. Однако исследователи могут предпочесть безотлагательную публикацию всех элементов данных, даже когда ранняя версия включает существенные ошибки таксономического определения. Эти ошибки могут быть исправлены в ходе последующего уточнения. Следующие данные поступают из BOLD в GenBank либо сразу, либо на этапе отправки рукописи для публикации:

- Информация о месте сбора: GPS-координаты, высота/глубина, область/страна, точное место сбора и фамилии участвовавших в сборе образца.
- Таксономическая информация: определение до уровня вида (если требуется, то до подвида), фамилии лиц, участвовавших в идентификации экземпляра.
- Информация о последовательности: собранные вручную и выверенные последовательности ДНК-штрихкодов.

ТИПЫ РЕСУРСОВ

Также как несколько типов данных создается в процессе проекта iBOL, также и несколько типов ресурсов будут переводиться в общий доступ консорциумом iBOL.

Это включает:

1. Биоматериалы. В эту категорию попадают экземпляры или образцы тканей. Члены консорциума iBOL следуют правилам, принятым Конвенцией о Биоразнообразии. Любая передача этих ресурсов между членами iBOL проводится с учетом всех ограничений по отношению к передаче материалов и руководствуется соглашением о передаче материалов (Materials Transfer Agreement, MTA) или похожими соглашениями, которые должны быть подписаны до того, как происходит передача материала. Соглашение должно включать описание условий доступа и использования пересылаемых материалов другими исследователями, а также их хранения и курирования. Некоторые страны или институты могут предпочесть ограничение доступа и использования этих биоматериалов, и эти условия должны быть четко описаны в MTA или другом подобном соглашении.

2. База данных ДНК-штрихкодов (BOLD). Записи ДНК-штрихкодов в BOLD являются ресурсом сообщества и будут становиться общедоступными. Данные общедоступной базы данных будут бесплатны для пользователей. Единственным условием использованием базы данных BOLD – это обязательная ссылка на тех исследователей консорциума iBOL,

которые участвовали в получении использованных последовательностей. iBOL настоятельно рекомендует использование публичных данных BOLD для развития приложений, которые приведут к развитию технологий, улучшения здоровья народонаселения, мониторинга состояния природной среды и развития любых других технологий.

3. **Анализ информации**. Средства анализа информации, используемые в BOLD и других аспектах проекта iBOL, рассматриваются как общественный ресурс внутри iBOL. Для примера, система управления информацией лаборатории, используемая в Центре штрихкодирования Института Биоразнообразия Онтарио (BIO) является частью BOLD и может быть доступна как всем членам iBOL, так и более широкому научному сообществу.

АКРЕДИТАЦИЯ ОПУБЛИКОВАННЫХ ДАННЫХ

Члены iBOL признают важность для всех исследователей, как из академических, так и из государственных или промышленных учреждений, для аспирантов, кандидатов наук (докторантов), пост-докторантов или преподавателей, что данные, которые они получили и сделали доступными для широкого научного сообщества, при использовании имели бы соответствующую ссылку на их авторство. Это важно по нескольким причинам, включая тот факт, что публикации и цитирование широко используется при оценке качества и успешности как проектов, так и исследователей. С другой стороны, для исследователя или любого другого, кто желает использовать эти опубликованные данные, важно знать, кто получил эти данные. Поскольку это дает в дальнейшем возможность партнерства или сотрудничества, либо получение дополнительной информации или других данных, которые будут полезны исследователям, получившим последовательности. Таким образом, важно, что данные были опубликованы так, как описано выше и содержали информацию о том, кто их опубликовал и чтобы имелся соответствующий текст, позволяющий привести цитирование и ссылку на авторство. Пользователи опубликованных в базах данных последовательностей должны при публикации давать ссылку на источник данных.

Существуют другие механизмы, когда исследователи могут получить соответствующее признание при первичной публикации данных. Два таких пути настоятельно рекомендуются членами iBOL: описание проекта и публикация об обнародовании данных в печати. Это служит для многих целей, таких как (а) приведение информации о научных коллективах, чтобы эти исследования могли быть цитируемы, (б) информирование команды iBOL и более широкой научной общественности о возможностях добавления и обновления данных, которые могут быть использованы для уточнения и исправления первичных данных, а также (с) дать возможность обмена информацией, что позволит возникнуть новому партнерству и привлечению дополнительного финансирования.

Публикация описания проекта - перспективное описание больших проектов, таких как проекты, выполняемые в рамках iBOL. Они могут содержать некоторые предварительные данные, не представлять данных проекта вовсе, но содержат описание проекта, его цели,

участников, используемые методические основы, временные рамки и предполагаемые результаты, а также механизм и время публикации полученных данных. Существуют примеры других проектов, направленных на создание общедоступных ресурсов, например:

- The International HapMap Project.http://www.hapmap.org/. Nature. 2003, Dec 18; 426(6968):789-96.
- The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004, Oct 22; 306(5696):636-40.

Публикация об обнародовании данных, представленная в рецензируемом журнале, описывает первичные массивы данных проекта. При этом в публикации указывается, что эти данные первичны и в дальнейшем будут уточнены и обработаны на более поздних стадиях проекта. Например: Hubert at al., 2008[5]. Хотя в этой статье данные и таксономические определения относительно полно представлены, дополнительное уточнение в процессе продолжающихся исследований этой группы предоставит значительно более ценную информацию для научного сообщества и для общедоступного ресурса данных.

ИНТЕЛЛЕКТУАЛЬНАЯ СОБСТВЕННОСТЬ

Члены iBOL рассматривают все данные ДНК-штрихкодов в BOLD как ресурс сообщества, который становится общедоступным согласно терминологии и условиям, очерченным в настоящих правилах. Эти данные не попадают под категорию Интеллектуальной Собственности.

ЧАСТНАЯ ЖИЗНЬ/ВОПРОСЫ ЭТИКИ

1. Обычно аспекты частной жизни редко возникает при ДНК-штрихкодировании. Однако, некоторые ограничения могут налагаться в случае, когда данные происходят из образцов, используемых в рабочей группе по патогенам человека. Любой исследователь, работающий в рамках iBOL и участвующий в сборе образцов на человеке, руководствуется этическими нормами, принятыми в соответствующих странах или организациях. Вся соответствующая документация должна быть представлена до начала каждого такого подпроекта. В согласии с соответствующим законодательством о частной жизни и другими правилам, любые данные, ассоциированные с таким проектом, должны быть достаточно обезличенным, так чтобы никакая персональная идентификация не была бы возможна.
2. Во время сборов экземпляров для библиотеки ДНК-штрихкодов iBOL, некоторые из образцов будут собранны в экологически уязвимых местах обитания. Некоторые страны могут проявлять озабоченность в опубликовании информации, которая может быть использована для нахождения этих экологически уязвимых местообитаний. Например, доступ к GPS-координатам мест нахождения видов орхидей в тропических лесах могут

[5] Hubert N, Hanner R, Holm E, Mandrak NE, Taylor E, Burridge M, Watkinson D, Dumont P, Curry A, Bentzen P, Zhang J, April J, Bernatchez L. 2008. Identifying Canadian Freshwater Fishes through DNA Barcodes. PLoS One 3(6): e2490

увеличить риск для этих видов. Если возникает подобная озабоченность, iBOL-исследователи обязуются держать эти данные в конфиденциальности или использовать другие способы обезличивания данных. Публикация GPS-координат не является обязательной в iBOL.

3. Некоторые участники консорциума iBOL являются исследователями государственных организаций, в чьи обязанности входит мониторинг инвазивных и вредных видов на территории их страны или района. Такие отделы и организации имеют прописанные протоколы по публикации данных об инвазивных или других вредных видах. Возможны опасения, что ранняя публикация по ДНК-штрихкодированию, касающаяся инвазивных видов может вступать в противоречие с государственным законодательством. Исследователь несет ответственность за идентификацию инвазивного вида согласно протоколу контроля качества, описанного выше. Если предварительные данные указывают на инвазивный и вредный вид, то такие данные могут не публиковаться до тех пор, пока необходимые требования, предъявляемые институтом или государством, не будут выполнены. Совет директоров iBOL будет проинформирован при возникновении подобной ситуации, которая ограничивает публикацию данных.

ОБЩЕЕ РУКОВОДСТВО

Все члены iBOL придерживаются изложенных правил. Любые запросы о продлении временных рамок, трактовке формулировок и другие вопросы направляются в совет директоров iBOL.