Базы данных коллекций биологического материала: организация сопроводительной информации

***Буйкин С.В.*[1], *Брагина Е.Ю.*[1], *Конева Л.А.*[1], *Пузырёв В.П.*[1, 2]**

Databases of biological collection: organization of associated information

Buikin S.V., Bragina Ye.Yu., Koneva L.A., Puzyryov V.P.

[1] *НИИ медицинской генетики СО РАМН, г. Томск*
[2] *Сибирский государственный медицинский университет, г. Томск*

Данный обзор посвящен проблеме организации и хранения сопроводительной информации для биологических коллекций. Рассмотрены основные принципы разработки структуры баз данных для этих целей. Проведен обзор баз данных биологических коллекций и биобанков в России и мире. Представлена структура базы данных биобанка НИИ медицинской генетики СО РАМН (г. Томск).

Ключевые слова: биобанк, база данных, многофакторные заболевания.

The review discusses problem of organization and storage of associated information for biological collection. Basic principles of database structure development are presented. Databases for biological collections and biobanks in Russia and other countries are reviewed. Structure of database of Biobank of Institute of Medical Genetics SB RAMS is described.

Key words: biobank, database, complex diseases.

УДК 57.082.5

Введение

Современные исследования в области медицины, в частности изучение генетических основ подверженности сложным заболеваниям, требуют комплексного анализа большого массива клинической информации и молекулярно-генетических данных, характеризующих индивидуальные особенности организма [1, 15, 24]. Коллекции биологического материала, или биобанки, играют центральную роль в объединении этих двух потоков информации и консолидируют большой объем биологических образцов и сопроводительной информации [2]. Биобанки сегодня — новое направление, которое развивается как самостоятельная область исследования со многими специфическими компонентами, требующая специализированного персонала [50].

Согласно публикациям за последние годы, в развитии биобанков можно выделить стремление к их интеграции, что позволяет проводить исследования с участием многих научно-исследовательских центров из разных стран [9, 30]. Появление новых биотехнологий и развитие современных концепций изучения заболеваний ужесточили требования к высококачественным, хорошо аннотированным биологическим образцам для биомедицинских исследований [29]. В настоящее время описаны и утверждены стандарты в отношении забора и хранения биологических образцов [49]. В то же время требования к спектру, полноте и описанию сопроводительной информации и организации структуры системы управления данными в биобанке сильно варьируют и не имеют общих стандартов или рекомендаций. В данном сообщении представлены сведения о структуре различных баз данных, касающиеся коллекций биологического материала, предназначенных для научных исследований, об опыте работы по их созданию и дальнейшему использованию, а также описана база данных для биобанка НИИ медицинской генетики СО РАМН (г. Томск).

Базы данных биобанков

Базой данных (БД) называют логически согласованное собрание взаимосвязанных данных с конкретным значением, предназначенное для определенной цели и являющееся единым хранилищем информации. БД состоит из записей — самостоятельных, внутренне связанных пакетов информации, представляя собой информационную модель предметной области. Базы данных необходимы для сбора и хранения данных, обеспечения удобных для пользователя функций доступа и поиска, а также стандартизации представления данных и их организации в знания. Обращение к БД и управление данными осуществляется с помощью системы управления базами данных (СУБД). Главные цели создания БД: а) уменьшение избыточности данных; б) достижение независимости данных [4].

Сегодня биологические данные и коллекции биологического материала собирают во всех уголках мира. Базы данных — необходимый элемент организации структуры и эффективного функционирования биобанков.

БД функционирующих биобанков. В организации биобанков выделяют два больших формата с несколькими подтипами — это популяционные биобанки и болезньориентированные биобанки [9]. Первый формат включает следующие подтипы: лонгитудинальные популяционные биобанки, биобанки популяционных изолятов и близнецовые регистры. Болезньориентированные биобанки включают коллекции биологического материала из клинических исследований «случай — контроль» и банки биологических тканей. Кроме функциональной организации можно разделить биобанки по размеру — от небольших специфических собраний биологического материала для изучения редких болезней [32, 43] до огромных популяционных биобанков NBSBCCC [36]. Цель создания биобанков также подразумевает различный формат их организации: при медицинских центрах (архивы патологии для медицинской диагностики) [37], при исследовательских институтах (прежде всего для научных целей) [21]. Несмотря на такие различия, основная функция биобанков — это сбор и хранение биологических образцов и сопроводительной информации к ним.

Хотя вопросы сбора и хранения биологических образцов поднимались и ранее [35, 44], первыми об организации биобанков ДНК и тканей официально заявили в Великобритании и скандинавских странах [10]. В дальнейшем отмечался рост количества и разнообразие специализаций создаваемых биобанков [16], формировались и разрабатывались БД для их сопровождения [41]. В литературе описание БД биобанков встречается значительно реже, чем публикации о структуре и организации самих биобанков. При поисковом запросе «база данных биобанка» в БД научных публикаций PubMed (http://www.ncbi.nlm.nih.gov/pubmed) из 35 найденных ссылок только 2 публикации были посвящены организации БД биобанка, в остальных статьях внимание концентрировалось на общих вопросах организации и функционирования биобанков или представлялись исследования, проводимые с использованием биобанков. При аналогичном запросе в поисковой системе Google было найдено более 95 тыс. ссылок. Из числа первых 50 ссылок только 27 описывали биоколлекции (биобанки) при различных медицинских и исследовательских центрах. В большинстве случаев упоминалось только наличие БД, а структура биобанка сводилась к перечислению патологий и общему количеству образцов. Одним из объяснений этого может быть то, что биобанки часто создаются под конкретный исследовательский проект и имеют определенную структуру и специализацию, в результате БД биобанка становится узкоспециализированным ресурсом и является интеллектуальной собственностью организации, инициирующей создание биобанка. Такая ситуация характеризует низкую информационную освещенность структуры БД конкретных биобанков и ставит задачу разработки общих рекомендаций и требований по организации БД и системы управления данными биобанков.

Одним из удачных примеров системы управления БД для коллекций биологического материала в биомедицинских исследованиях является разработанная в National Human Genome Research Institute (США) система GeneLink [17], которая позволяет хранить и управлять огромными массивами информации, необходимыми для генетического картирования комплексных признаков человека. GeneLink представляет собой доступную через Интернет, защищенную паролем базу данных, созданную на основе платформы Sybase. GeneLink является эффективным инструментом, позволяющим легко объединять генотипические данные с родословными и обширными фенотипическими данными. Эта система является платформенно независимой и специально разработана для облегчения круп-

номасштабных (многоцентровых) исследований генетического сцепления или анализа ассоциаций.

Изначально система GeneLink была создана при изучении генетической подверженности рака простаты. Для реализации данного проекта были привлечены исследователи из США, Финляндии и Швеции [18]. В ходе его выполнения было изучено 496 семей, включающих 5 247 индивидов, прогенотипировано 2 374 образца ДНК по более чем 400 микросателлитным маркерам. Таким образом, при выполнении проекта были получены данные примерно для 1 млн генотипов. Учитывая значительное число генотипов, требующее их анализа, перед исследователями встала задача разработки системы управления базой данных, которая может обрабатывать такие массивы данных.

Структура базы данных GeneLink представляет собой систему взаимосвязанных таблиц. Данные хранятся в 11 основных таблицах. Таблица «Семья» содержит клиническую информацию относительно каждой семьи. Таблица «Родословная» включает запись на человека в пределах семьи и биологические связи (ID отца, ID матери) для каждого человека. Определяемые признаки и их значения представлены в таблицах с соответствующими названиями. Таблица «Маркеры» хранит информацию относительно всех маркеров, изучаемых в данном проекте, и условий генотипирования. Таблица «Праймеры» обеспечивает дополнительные данные для каждого маркера, например последовательность праймеров. Таблица карт включает генетические карты локализации маркера в геноме, относительный порядок маркеров на хромосоме и расстояние между соседними маркерами. В таблице «Генотип» отражены итоговые данные результатов генотипирования, информация о лаборатории, в которой они получены. Таблица «Класс подверженности» включает состояние относительно классов предрасположенности, которые определяются, используя любую комбинацию признаков, например возраст, пол, статус болезни. Наконец, таблица «ДНК» представляет сведения обо всех доступных образцах ДНК в лаборатории (включает специальные данные относительно даты забора образца, концентрации ДНК, ее количества и места хранения).

Таблицы в БД могут быть заполнены либо импортом множества записей из текстового файла, либо последовательным заполнением отдельных записей через веб-интерфейс. Каждая из 11 таблиц GeneLink имеет встроенные механизмы проверки качества записей, такие как регистрация всех операций редактирования записей в таблице «История операций», контроль пригодности формата записей для проводимых операций, также предусмотрена проверка пола индивидов, наличия родственных связей и многих других признаков в различных связанных таблицах БД. Безопасность данных в GeneLink обеспечивается предоставлением разных уровней доступа пользователям БД [18]. В настоящее время проводится модернизация БД GeneLink с целью ее использования для хранения и администрирования информации, полученной в результате семейных исследований, а также исследований по типу «случай — контроль» [17].

В качестве примера процесса создания базы данных биобанка специализированного биомедицинского исследования можно привести инициированное в 2003 г. фармакогенетическое изучение народов Африки по маркерам, участвующим в метаболизме лекарственных средств. По окончании исследования создан биобанк крови и ДНК, а также фармакогенетическая база данных, содержащая информацию о генетическом разнообразии по полиморфным вариантам наиболее значимых генов лекарственного метаболизма [45]. Объем коллекции биобанка представлен 1 488 образцами ДНК и охватывает девять этнических групп Африки. Каталогизация образцов и характеристика генотипов по функционально значимым полиморфным вариантам генов ферментов метаболизма ксенобиотиков для каждого индивидуума представлена в формате приложения Microsoft Access. Особое внимание уделяется доступу к информации базы данных, который строго лимитирован и разрешен исключительно для авторизированных лиц, вовлеченных в исследовательский проект [33].

Существуют биобанки, где представлены образцы биологического материала из многих стран с редко встречающимися синдромами, например биобанк синдрома Ретта в Италии (www.biobank.unisit.it), в Австралии (http://www.ibahc.org) [27, 39]. Начиная с 1998 г. отдел медицинской генетики университетского госпиталя г. Сиены (Италия) приступил к сбору биологического материала лимфобластоидных клеточных линий пациентов с указанным заболеванием. В течение последующих лет число образцов, включая кровь, ДНК и клеточные линии пациентов значительно возросло и было организовано в биобанк. Быстрое увеличение

образцов биобанка способствовало созданию on-line базы данных (www.biobank.unisit.it), которая доступна с 2004 г. [40]. Она поддерживается и обновляется каждые 3 мес на сервере университета Сиены. На общей домашней странице есть ссылки на независимые БД: Х-сцепленная умственная отсталость (XLMR); синдром Ретта; ретинобластома и др. На февраль 2010 г. в биобанке хранилось более 1 200 образцов ДНК (340 пациентов с синдромом Ретта и 860 родственников). БД содержит подробную описательную информацию фенотипа каждого пациента согласно международным диагностическим критериям [22], а также данные молекулярно-генетического анализа о мутациях в генах *MECP2* и *CDKL5*. БД организована в соответствии с руководством Итальянского общества генетики человека для биобанков, обеспечивает анонимность и конфиденциальность пациентов согласно международным критериям [19]. Кроме того, информированное согласие охватывает все аспекты хранения образцов и использования анкетных данных [19]. БД синдрома Ретта входит в консорциум InterRett, собирающий информацию о БД из 28 стран мира (всего 2 089 случаев).

В научно-исследовательских работах, базирующихся на крупномасштабных исследованиях биобанков, основная проблема состоит в интеграции генотипических и фенотипических данных, полученных из разных источников и хранящихся в учреждениях разных стран. Эта проблема относится прежде всего к международным проектам создания глобальных биобанковских сетей EuroBioBank (www.eurobiobank.org), GenomEUtwin (www.genomeutwin.org), HERACLES (www.redheracles.net) и др. [12, 31, 38].

EuroBioBank объединяет 16 биобанков из восьми европейских стран, основное направление деятельности — это сбор и изучение редких заболеваний. В настоящее время сеть содержит сведения о более 440 тыс. биологических образцов, информация о которых организована в каталог. Каталог разделяет образцы по видам биопроб (ДНК, клетки, ткань), а также по клиническому диагнозу и месту сбора и хранения. При содействии EuroBioBank ежегодно инициируются десятки исследовательских проектов.

Близнецовый регистр GenomEUtwin создан для изучения сложных фенотипов и объединяет семь близнецовых когорт (с 1870 по 1990 гг. рождения). На данный момент создана БД, содержащая фенотипическую и генотипическую информацию индивидов семи европейских стран и Австралии, которая администрируется через сеть TwinNET. Пользователь имеет доступ к стандартному интерфейсу с необходимыми элементами данных для объединенных исследований [34].

БД биологических коллекций в России. В русскоязычной литературе сведения об организации систем управления информацией биобанков мало представлены. В то же время можно предполагать, что коллекции биологического материала и сопроводительная информация к ним имеются в большинстве исследовательских центров медико-биологического профиля. Так, на базе Медико-генетической лаборатории республиканской больницы № 1 создан банк ДНК наследственных патологий и популяций народов Республики Саха (Якутия). В настоящее время в биобанке насчитывается более 8 тыс. образцов биологического материала (больные с наследственными моногенными заболеваниями и их родственники, больные мультифакториальными заболеваниями и популяционный материал малочисленных народов Республики Саха). Разработана компьютерная программа INFOGEN, которая представляет собой пользовательский интерфейс, предназначенный для ведения учета пациентов, биологический материал которых был внесен в биобанк. Основными возможностями БД является поиск данных по критериям (шифр семьи или ДНК, фамилия, имя, отчество члена семьи, дата рождения, пол, национальность, диагноз, место рождения, место проживания), а также вывод результатов поиска в выходной документ Microsoft Exel и Microsoft Word и распечатка результатов поиска [5].

В Южно-Уральском институте биофизики создан биобанк ДНК облученных людей для оценки риска отдаленных последствий облучения. В настоящее время банк ДНК содержит генетический материал 1 тыс. человек бывших и нынешних работников атомного предприятия «Маяк». Создана электронная БД банка ДНК, включающая индивидуальные медико-демографические, профессиональные и дозовые характеристики, а также качественные и количественные показатели ДНК [6].

Можно отметить, что в настоящее время отмечается бурный рост направления биобанков в целом и тенденция к укрупнению и объединению разрозненных биобанков в консорциумы и сети. Учитывая

ситуацию в данной области, можно констатировать необходимость создания полноценных биологических банков в РФ, отвечающих международным стандартам. В качестве примера в данной области можно назвать биобанк НИИ медицинской генетики СО РАМН (г. Томск), который функционирует с 2008 г. и объединяет коллекции биологических образцов, собираемых с момента основания института (с 1982 г.). Основной целью биобанка НИИМГ СО РАМН является сбор, хранение и использование биологического материала для исследований в области медицинской генетики. В биобанке хранится несколько тысяч образцов ДНК индивидов с различными патологиями, включая сердечно-сосудистые, иммунозависимые, моногенные и другие заболевания, а также ДНК здоровых индивидов популяций Восточной, Западной Сибири и Средней Азии. Для оптимизации сохранности образцов и сопроводительной информации, а также для повышения эффективности научных исследований была разработана БД биобанка НИИМГ СО РАМН [3].

БД биобанка НИИМГ СО РАМН

На момент начала проектирования БД биобанка НИИМГ СО РАМН отсутствовали разработанные рекомендации и требования к информационным ресурсам подобного рода. Отдельные исследования, посвященные разработке общих принципов создания структур управления информацией в биобанках, стали доступны только в последние несколько лет [13, 46]. Тем не менее сравнение архитектуры и организация связей БД биобанка НИИМГ СО РАМН с описанием доступных БД других биобанков и анализ представленных рекомендаций позволяет сделать заключение о том, что в БД биобанка НИИМГ СО РАМН были учтены все особенности создания подобных информационных ресурсов.

В процессе создания БД разработчики столкнулись с рядом затруднений технического (выбор модели и системы управления БД, системы защиты информации БД) и организационного (анализ имеющегося биологического материала, определение объемов и типов сопроводительной информации) плана. Отдельно стояли вопросы выбора первичной структуры БД, формализации всех типов сопроводительной информации (разноплановые по своему содержанию данные, отличающиеся для разных патологий и хранящиеся в виде таблиц Excel, текстовых и графических файлов); разработка унифицированных форм и таблиц БД, максимально охватывающих разнообразие представленной информации, и т.д.

Сопроводительная информация биологических образцов биобанка НИИМГ СО РАМН, учитывая специфичность информации хранящихся образцов, содержит следующие данные:

1. Демографические данные об индивиде (Ф.И.О., пол, национальность, национальность родителей, место рождения, место жительства, семейный статус и т.д.).

2. Семейные данные (сведения о родителях, сведения о потомках).

3. Фенотипическое описание индивида (здоровый (больной); клинико-анамнестические данные; результаты лабораторных и инструментальных исследований).

4. Информация относительно этических вопросов, связанных с возможностью использования биологических образцов в научных исследованиях (информированные согласия индивидов на исследование с использованием их биологического материала).

5. Методические особенности работы с биологическими образцами (метод выделения образца ДНК, концентрация, качество, объем образца и т.д.).

6. Описание результатов генетического исследования (результаты генотипирования, секвенирования и т.д.).

БД биобанка условно делится на четыре блока (рис. 1): а) «Основная информация (индивид)» — включает в себя информацию об объекте исследования, его анамнезе, заболеваниях, результатах лабораторных исследований, родственных связях; б) «Образцы биологического материала» — представляет информацию о месте, дате и методе выделения ДНК, месте хранения, расходовании образца, название выборки, употреблявшееся до создания общей БД института; в) «Результаты молекулярно-генетического анализа (генотип)» — представляет справочную таблицу названий и описаний генов, по которым проводились исследования, таблицу полиморфных вариантов, генотипы, результаты секвенирования и анализ мтДНК; г) «Пользователи» — содержит информацию о пользователях БД, проводимых запросах, историях операций, перечень разрешенных IP-адресов.

Для обеспечения безопасности данных на уровне информационной системы были реализованы автори-

зированный доступ под паролем к БД, ограничение доступа пользователей с указанных IP-адресов, защита от автоматического подбора пароля к БД с блокировкой атакующего IP, предоставление доступа к персональным данным, а также доступа к внесению и редактированию данных ограниченному кругу пользователей. Хранение истории авторизации и запросов пользователей, а также истории на внесение изменений в БД.

В результате в биобанке НИИМГ СО РАМН разработан электронный ресурс, отвечающий требованиям специалистов, работающих в различных областях медицинской генетики, и соответствующий нормам федеральных законов и актов об обеспечении безопасности персональных данных.

Рис. 1. Структура организации БД биобанка НИИМГ СО РАМН

Проблемы разработки БД биобанка

Информацию, которая хранится в БД биобанка, обычно можно подразделить на несколько типов: клинические данные о пациенте (доноре), демографические сведения, характеристика образцов биологического материала и административная информация [50]. При формировании БД биобанка клиническая информация обычно вручную извлекается из историй болезни пациентов, анкетных данных, и каждый раз при инициации нового исследовательского проекта колоссальные интеллектуальные и материальные ресурсы затрачиваются на создание специфической архитектуры БД [20]. В результате вследствие отсутствия единых стандартов в организации клинической или другой сопутствующей информации в архитектуре БД биобанков возникают трудности по выбору и объединению информации, что ограничивает возможности сотрудничества и препятствует полномасштабному использованию ресурсов биобанков [47, 51].

Попыткой решения данной проблемы с целью разработки единого формата организации данных в БД биобанков является использование стандартизированных понятий — архетипов на основе электронных историй болезни. Исследователи из биобанка Ирландского консорциума изучения рака простаты (Irish Prostate Cancer Research Consortium (PCRC)) поставили задачу выяснить, каким образом можно модифицировать электронные истории болезни для целей научных исследований [46]. Разработка архетипов — это деятельность в пределах проекта openEHR, который представляет собой открытый стандарт управления, хранения и обмена электронными историями болезни (electronic health record (EHR)). Функциональные требования openEHR основаны на более чем 20-летнем опыте европейских и австралийских научных исследований в области электронных историй болезни [28]. Интегрированная среда openEHR согласуется со стандартами менеджмента качества и используется в правительственных программах в области здравоохранения в Великобритании, Австралии, Дании, Нидерландах, Швеции, Сингапуре, США, Словакии, Чили и Бразилии [48].

OpenEHR применяет двухуровневый подход к моделированию клинической информации, который означает, что правила представления информации в записи openEHR отражены в архетипах (которые могут усовершенствоваться и быть общедоступными), а части, из которых эти модели строятся, неизменны и полностью описываются референсной моделью [26]. В результате разработанное программное обеспечение может быть построено на стабильной референсной модели, а изменяющимися и развивающимися клиническими концепциями можно управлять в среде знаний — репозитории архетипов. Архетипы несут в себе правила, которые проверяют качество данных, и они могут быть использованы при вводе информации для обеспечения их корректности. Преимуществом такого подхода является то, что эволюция клинических кон-

цепций не вызывает необходимости изменения программного обеспечения на фундаментальном уровне.

В исследовании PCRC авторы разрабатывали электронный биомедицинский научный отчет (electronic biomedical research record (eBMRR)), используя методологию архетипа openEHR; eBMRR должен объединять клиническую и исследовательскую информацию и может быть адаптирован для определенного биобанка или научного исследования. Данные в биобанке PCRC администрируются с помощью Biobank Information Management System (BIMS). Исследователи из биобанка PCRC сначала проанализировали структуру, понятия и содержание базы данных BIMS. Было важно определить точное значение и контекст каждой области так, чтобы можно было установить соответствие между отдельными архетипами и областью базы данных. Выявлено, что концептуальная схема базы данных BIMS PCRC содержала 18 иерархически связанных таблиц, при этом поля в некоторых таблицах повторялись, создавая избыточность в хранении информации. Из этих 18 таблиц 14 содержали информацию, связанную главным образом с образцами биологического материала, а оставшиеся 4 таблицы содержали клиническую информацию. Авторы отметили, что результаты более сложных исследований, например omic экспериментов, не сохранялись в базе данных PCRC, а только на компьютерах отдельных исследователей и могли быть потеряны для будущих научных разработок. Процесс работы с базой данных предложен авторами в виде концептуальной схемы (рис. 2).

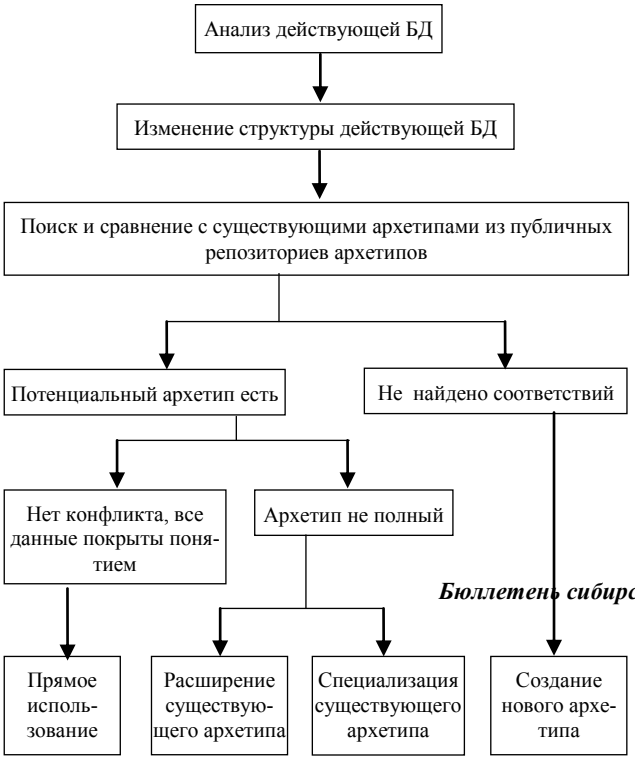

Рис. 2. Блок-схема моделирования концепций в БД BIMS PCRC с архетипами

В процессе анализа структуры БД BIMS PCRC авторы получили набор из 47 архетипов, который охватывал все виды данных биобанка. Из них 29 архетипов повторно использовались без изменения, 6 были изменены и (или) расширены, 1 был специализирован и 11 были созданы заново. Эти архетипы были систематизированы в 8 шаблонов, обязательных для этого биобанка.

Данная работа продемонстрировала множество проблем, которые явились следствием незрелости архетипного подхода: недостаточная разработанность инструментов поддержки моделирования архетипов; нехватка принципов управления и практических правил для моделирования архетипов; трудности в определении высококачественных архетипов; проблемы перекрывающихся архетипов. Кроме того, поиск и идентификация подходящих архетипов оказался очень трудоемким процессом в связи с множеством семантических конфликтов в процессе соотнесения структурных единиц из БД биобанка PCRC с существующими архетипами. Эти конфликты включали различия: в степени детализации документации, в используемой терминологии и словарях и в структуре регистрируемой информации.

Авторы делают вывод, что выявление подходящих архетипов в репозиториях и установление точного соответствия между областями в БД биобанка и элементами существующих архетипов, разработанных для клиницистов, выступает сложной и трудоемкой задачей, требующей совместного анализа и подхода со стороны медицинского и научного сообществ [46]. Несмотря на это, очевидны перспективы повсеместного использования универсальных архетипов в контексте биомедицинского исследования. Многие архетипы, первоначально развитые для электронных историй болезней, могут повторно использоваться, чтобы моделировать клиническую (фенотипическую) информацию и информацию о выборке в контексте биобанка [42].

Очень важными для создания и функционирования БД биобанков являются этические, юридические и социальные проблемы, поднятые инициаторами создания консорциумов и сетей биобанков, которые используют большие БД биобанков различных стран. Доступ к таким БД, где содержатся результаты широкомасштабного генетического скрининга большого количества образцов, становится все более и более важным в биомедицинских исследованиях [29]. Несмотря на процессы глобализации и интенсивное развитие единых стандартов, в том числе и для биобанков, перспективы универсального свода норм и правил по созданию и администрированию биобанков и их БД пока значительно отстают от роста количества биобанков и биологических коллекций [11].

Другим не менее важным вопросом создания и функционирования БД биобанка является защита персональных данных пациентов или доноров биобанка. Целью информационной безопасности БД является сохранность информации системы, защита и гарантия полноты и целостности данных, минимизация потерь в случае модификации или утраты данных [14]. Безопасность БД требует учета всех событий, в ходе которых информация создается, модифицируется, к ней обеспечивается доступ или она распространяется [7, 8]. Необходимость хранения персональной информации в БД биоколлекционных ресурсов требует усиленных мер технической и программной безопасности, а также введения ограничений доступа к хранимой информации для разных групп пользователей. При этом четкой и ясной методики комплексного решения задачи защиты баз данных, которую можно было бы применять во всех случаях, не существует, в каждой конкретной ситуации приходится находить индивидуальный подход.

Заключение

В отличие от исследований простых менделевских заболеваний изучение широкораспространенных болезней и комплексных признаков является намного более сложной задачей. В такие исследования вовлечены большие выборки индивидов (для которых получена информация о генотипах по огромному числу маркеров) зачастую из многих исследовательских центров, расположенных в разных странах. Создание биологических коллекций и связанных с ними данных необходимо для генетических и клинических исследований, фундаментального обоснования разработки новых подходов к диагностике и лечению заболеваний с наследственной предрасположенностью [23, 25]. Очень важным условием собранной коллекции является ее долгосрочное, правильное хранение и систематизация имеющихся данных [38], что невозможно без отлаженной, оптимально сконструированной и безопасной БД биобанка.

Развитие подходов к организации и оптимизации коллекций биологического материала привело к возникновению новых направлений исследования, касающихся вопросов унификации и интеграции клинической и молекулярно-генетической информации. Несмотря на малочисленность публикаций, в которых непосредственно охарактеризована организация БД биобанков, можно отметить, что существуют попытки выработки общих принципов и понятий в данной интенсивно развивающейся области исследований. Одним из примеров можно назвать архетипический подход, ориентированный на выявление стандартных архетипов как информационных доменов в системах хранения и управления биомедицинскими знаниями.

Отсутствие полноценных биобанков в РФ отчасти выступает результатом дефицита информационных ресурсов, ориентированных на выработку единых концепций организации БД коллекций биологического материала. В то же время конкурентоспособность российских биомедицинских исследований и разработок на международном уровне обязательно требует использования больших массивов качественного биологического материала и информационных ресурсов, позволяющих оперировать такими объемами данных. Таким образом, создание полноценной централизованной БД является одним из ключевых требований оптимальной и беспрерывной работы биобанка, что способствует результативности медико-генетических исследований и определяет их качество и эффективность.

Работа выполнена в рамках реализации ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009—2013 гг. (гос. контракт № П722).

Литература

1. *Брагина Е.Ю., Буйкин С.В.* Влияние численности контрольной выборки на значимость ассоциаций генетических маркеров с развитием мультифакториальных заболеваний // Якут. мед. журн. 2009. № 2 (26). С. 142—144.
2. *Брагина Е.Ю., Буйкин С.В., Пузырев В.П.* Биологические банки: проблемы и перспективы их использования в исследованиях генетических аспектов комплексных забо-

леваний человека // Мед. генетика. 2009. № 3. С. 20—27.
3. *Буйкин С.В., Брагина Е.Ю., Конева Л.А.* Разработка структуры базы данных для биобанков // Якут. мед. журн. 2011 № 1. С. 70—73.
4. *Игнасимуту С.* Основы биоинформатики. М.; Ижевск: НИЦ «Регулярная и хаотическая динамика»; Институт компьютерных исследований, 2007. 320 с.
5. *Кычкина О.И., Кононова С.К., Фёдорова С.А. и др.* О деятельности банка ДНК наследственных патологий и популяций народов Республики Саха (Якутия) // Тез. докл. III Междунар. науч.-практ. конф. 2006. 73 с.
6. *Русинова Г.Г., Адамова Г.В., Осовец С.В. и др.* Банк ДНК людей, подвергшихся профессиональному облучению. Цели и перспективы // Генетика. 2001. Т. 37, № 9. С. 1307—1310.
7. *Сабанов А.* Безопасность баз данных // Connect. 2006. № 4. С. 25—37.
8. *Соколов А.* Средства защиты персональных данных: проблемы оценки соответствия // Connect. 2008. № 12. С. 25—37.
9. *Asslaber M., Zatloukal K.* Biobanks: transnational, European and global networks // Brief. Funct. Genomics Proteomic. 2007. V. 6 (3). P. 193—201.
10. *Austin M.A., Harding S., McElroy C.* Genebanks: a comparison of eight proposed international genetic databases // Community Genet. 2003. V. 6 (1). P. 37—45.
11. *Beale T.* Archetypes: Constraint-based Domain Models for Future-proof Information Systems // http://www.openehr.org. 2008.
12. *Bingham S., Riboli E.* Diet and cancer — the European prospective investigation into cancer and nutrition // Nat. Rev. Cancer. 2004. V. 4. P. 206—215.
13. *Dangl A.* The IT-infrastructure of a biobank for an academic medical center // Stud. Health. Technol. Inform. 2010. V. 160. P. 1334—1338.
14. *Early W.* Database management systems for process safety // J. of Hazardous Materials. 2006. V. 130. P. 53—57.
15. *Feero W.G., Guttmacher A.E.* Genome wide Association Studies and assessment of the risk of disease // N. Engl. J. Med. 2010. V. 363. P. 166 —176.
16. *Galloux J.C.* An empirical survey on biobanking of human genetic material and data in six EU countries // Eur. J. Hum. Genet. 2003. V. 11. P. 475—488.
17. *GeneLink:* A data management system designed to facilitate genetic studies of complex traits. URL: http://research.nhgri.nih.gov/genelink.
18. *Gillanders E.M., Masiello A., Gildea D. et al.* GeneLink: a database to facilitate genetic studies of complex traits // BMC Genomics. 2004. V. 5, № 81. P. 1—14.
19. *Godard B., Schmidtke J., Cassiman J.J., Aymé S.* Data storage and DNA banking for biomedical research: informed consent, confidentiality, quality issues, ownership, return of benefits. A professional perspective // Eur. J. Hum. Genet. 2003. V. 11. P. 88—122.
20. *Grimson J.* Delivering the electronic healthcare record for the 21 century // Int. J. Med. Inform. 2001. V. 64. P. 111—127.
21. *Gurwitz D., Fortier I., Lunshof J.E., Knoppers B.M.* Children and Population Biobanks // Science. 2009. V. 325. № 5942. P. 818—819.
22. *Hagberg B., Hanefeld F., Percy A., Skjeldal O.* An update on clinically applicable diagnostic criteria in Rett syndrome // Eur. J. Paediatr. Neurol. 2002. V. 6. P. 293—297.
23. *Hansson M.G.* Combining efficiency and concerns about integrity when using human biobanks // Stud. Hist. Philos. Biol. Biomed. Sci. 2006. V. 37 (3). P. 520—532.
24. *Illig T., Gieger C., Zhai G. et al.* A genome-wide perspective of genetic variation in human metabolism // Nat. Genet. 2010. V. 42, № 2. P. 137—141.
25. *Kaiser J.* Biobanks: Population Databases Boom, From Iceland to the U.S. Science. 2002. V. 298. P. 1158—1161.
26. *Kristianson K.J., Ljunggren H., Gustafsson L.L.* Data extraction from a semi-structured electronic medical record system for outpatients: A model to facilitate the access and use of data for quality control and research // Health Informatics Journal. 2009. V. 15, № 4. P. 305—319.
27. *Laurvick C.L., de Klerk N., Bower C.* Rett syndrome in Australia: a review of the epidemiology // J. Pediatr. 2006. V. 48. P. 347—352.
28. *Leslie H.* International developments in openEHR archetypes and templates // Health information management journal. 2008. V. 37, № 1. P. 38—39.
29. *Little J., Higgins J.P., Ioannidis J.P. et al.* Strengthening the reporting of genetic association studies (STREGA): an extension of the STROBE Statement // Hum. Genet. 2009. V. 125. P. 131—151.
30. *Litton J.E.* Biobank informatics: connecting genotypes and phenotypes. Methods Mol. Biol. 2011. V. 675. P. 343—3461.
31. *Marrugat J., Lopez-Lopez J.R., Heras M. et al.* The HERACLES cardiovascular Network // Rev. Esp. Cardiol. 2008. V. 61, № 1. P. 66—75.
32. *Martin N., Krol P., Smith S. et al.* A national registry for juvenile dermatomyositis and other paediatric idiopathic inflammatory myopathies: 10 years' experience; the Juvenile Dermatomyositis National (UK and Ireland) Cohort Biomarker Study and Repository for Idiopathic Inflammatory Myopathies // Rheumatology. 2011. V. 50, № 1. P. 137—145.
33. *Matimba A., Oluka M.N., Ebeshi B.U. et al.* Establishment of a biobank and pharmacogenetics database of African populations // Eur. J. of Human Genetics. 2008. V. 16. P. 780—785.
34. *Muilu J., Peltonen L., Litton J.-E.* The federated database — a basis for biobanking-baesd post-genome studies, integrating phenome and genome date from 600000 twin pairs in Europe // Eur. J.l of Human Genetics. 2007. V. 15. P. 718—723.
35. *Narod S., Rosenblatt D., Lamothe E.* The banking of DNA for the prevention of genetic disease // Clin. Invest. Med. 1991. V. 14. P. 359—362.
36. *Pukkala E., Andersen A., Berglund G. et al.* Nordic biological specimen banks as basis for studies of and control more than 2 million sample donors, 25 million years and 100 000 prospective cancers // Acta Oncologica. 2007. V. 46. P. 286—307.
37. *Riegman P.H., Morente M.M., Betsou F. et al.* Biobanking for better healthcare // Mol. Oncol. 2008. V. 2. P. 213—222.
38. *Riegman P.H., Dinjens W.N., Oomen M.H. et al.* TuBaFrost 1: uniting local frozen tumour banks into a European network: an overview // Eur. J. Cancer. 2006. V. 42. P. 2678—2683.
39. *Robertson L., Hall S.E., Jacoby P.* The association between behavior and genotype in Rett syndrome using the Australian

Rett Syndrome Database // Am. J. Med. Genet. 2006. V. 5, № 41. P. 177—183.
40. *Sampieri K., Meloni I., Scala E. et al.* Italian Rett database and biobank // Hum. Mutat. 2007. V. 4. P. 329—335.
41. *Sampogna C.* Creation and governance of human genetic research databases Organization for Economic Co-operation and Development, OECD Publishing, 2006. 159 p.
42. *Savova G.K., Masanz J.J., Ogren P.V. et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications // J. Am. Med. Inform. Assoc. 2010. V. 17, № 5. P. 507—513
43. *Schena F.P. Cerullo G., Torres D.D. et al.* The IgA nephropathy Biobank. An important starting point for the genetic dissection of a complex trait // BMC Nephrol. 2005. V. 6. P. 14.
44. *Sherman J.K.* Synopsis of the use of frozen human semen since 1964: state of the art of human semen banking // Fertil. Steril. 1973. V. 24. P. 397—412.
45. *Sirugo G., Schim L., Sam O. et al.* A national DNA bank in the Gambia, West Africa, and genomic research in developing countries // Nat. Genet. 2004. V. 36. P. 785—786.
46. *Späth M.B., Grimson J.* Applying the archetype approach to the database of a biobank information management system // Int. J. Med. Inform. 2011. V. 80, № 3. P. 205—26.
47. *Tassé A.M., Budin-Ljøsne I., Knoppers B.M., Harris J.R.* Retrospective access to data: the ENGAGE consent experience // Eur. J. Hum. Genet. 2010. V. 18, № 7. P. 741—745.
48. *The* openEHR Foundation // http://www.openehr.org.
49. *Troyer D.* Biorepository standards and protocols for collecting, processing, and storing human tissues // Methods Mol. Biol. 2008. V. 441. P. 193—220.
50. *Watson P.H., Wilson-McManus J.E., Barnes R.O. et al.* Evolutionary concepts in biobanking—the BC BioLibrary // J. Trans. Med. 2009. V. 7. P. 95.
51. *Yuille M., Dixon K., Platt A. et al.* The UK DNA banking network: a «fair access» biobank // Cell Tissue Bank. 2010. V. 11. P. 241—251.

Поступила в редакцию 10.05.2011 г.
Утверждена к печати 22.12.2011 г.

Сведения об авторах

С.В. Буйкин — канд. мед. наук, науч. сотрудник лаборатории популяционной генетики НИИ медицинской генетики СО РАМН (г. Томск).

Е.Ю. Брагина — канд. биол. наук, науч. сотрудник лаборатории популяционной генетики НИИ медицинской генетики СО РАМН (г. Томск).

Л.А. Конева — канд. биол. наук, науч. сотрудник лаборатории популяционной генетики НИИ медицинской генетики СО РАМН (г. Томск).

В.П. Пузырёв — д-р мед. наук, профессор, академик РАМН, директор НИИ медицинской генетики СО РАМН (г. Томск), зав. кафедрой медицинской генетики СибГМУ (г. Томск).

Для корреспонденции

Буйкин Степан Вячеславович, тел. 8 (3822) 51-72-72; e-mail: stepan_buikin@mail.ru

Порядок рецензирования статей в журнале «Бюллетень сибирской медицины»

Все поступающие в редакцию рукописи после регистрации проходят этап обязательного двойного конфиденциального рецензирования членами редакционного совета либо внешними рецензентами. Рецензенты не имеют права копировать статью и обсуждать ее с другими лицами (без разрешения главного редактора).

При получении положительных рецензий работа считается принятой к рассмотрению редакционной коллегией журнала, которая окончательно решает вопрос о публикации материала в «Бюллетене сибирской медицины».

Редакция журнала извещает основного автора о результатах прохождения рецензирования и сроках публикации.

Редакция не принимает рукописи научно-практического характера, опубликованные ранее в других изданиях.

Все полученные редакцией журнала «Бюллетень сибирской медицины» рукописи будут рассмотрены без задержек и при получении положительных рецензий и решения редакционной коллегии опубликованы в течение одного года.

С правилами оформления работ можно ознакомиться в Интернете на сайте СибГМУ: http://ssmu.tomsk.ru.

Статьи и информация для журнала принимаются в редакционно-издательском отделе СибГМУ.