УДК 61:575; 658.011.56

ПРОГРАММНЫЙ КОМПЛЕКС SNP-MED ДЛЯ АНАЛИЗА ВЛИЯНИЯ ОДНОНУКЛЕОТИДНЫХ ПОЛИМОРФИЗМОВ НА ФУНКЦИЮ ГЕНОВ, СВЯЗАННЫХ С РАЗВИТИЕМ СОЦИАЛЬНО ЗНАЧИМЫХ ЗАБОЛЕВАНИЙ

© 2013 г. **Н.Л. Подколодный[1], Д.А. Афонников[1], Ю.Ю. Васькин[2], Л.О. Брызгалов[1], В.А. Иванисенко[1], П.С. Деменков[1], М.П. Пономаренко[1], Д.А. Рассказов[1], К.В. Гунбин[1], И.В. Процук[2], И.Ю. Шутов[2], П.Н. Леонтьев[2], М.Ю. Фурсов[2], Н.П. Бондарь[1], Е.В. Антонцева[1], Т.И. Меркулова[1], Н.А. Колчанов[1]**

[1] Федеральное государственное бюджетное учреждение науки Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия, e-mail: pnl@bionet.nsc.ru;
[2] ООО Новосибирский центр информационных технологий «УНИПРО», Новосибирск, Россия

Поступила в редакцию 15 августа 2013 г. Принята к публикации 5 сентября 2013 г.

В данной работе описана модульная компьютерная информационная система SNP-MED, предназначенная для оценки влияния однонуклеотидных полиморфизмов (ОНП) на функцию генов, связанных с развитием социально значимых заболеваний, включающая программные компоненты «Геномика», «Протеомика», «Генные сети» и базу данных «Информационный ресурс» (БДИР).

Ключевые слова: биоинформатика, однонуклеотидные полиморфизмы, персонализированная медицина.

ВВЕДЕНИЕ

Для современной постгеномной биологии характерно бурное развитие высокопроизводительных экспериментальных методик, позволяющих в одном эксперименте проводить измерения параметров целого генома, транскриптома, протеома. Разработанные ДНК-чиповые и транскриптомные технологии позволяют изучать динамику экспрессии десятков тысяч генов одновременно. Новое поколение методов высокоразрешающей масс-спектрометрии позволяет наблюдать за динамикой изменения концентраций РНК, белков, изучать потоки низкомолекулярных соединений в применении к фундаментальным медицинским проблемам. Новые технологии секвенирования геномов организмов (так называемые технологии секвенирования нового поколения, СНП) обеспечивают недорогой и эффективный способ ресеквенирования геномной ДНК человека, стоимость которого постоянно уменьшается и в ближайшее время может составить менее 1000 USD. Эти достижения позволяют перейти в медицине к новой парадигме, так называемой «персонализированной медицине», современной концепции здравоохранения, сформировавшейся в постгеномную эпоху, которая предполагает при проведении лечения учет индивидуальных геномных особенностей каждого пациента (генетическую предрасположенность к разным заболеваниям, воздействиям различных лекарств и т. п.). В основе персонализированной медицины лежат информация о персональных геномах и современные технологии оценки рисков заболеваний с учетом геномных полиморфизмов.

Выяснение молекулярных механизмов генетической предрасположенности к различным заболеваниям, таким как сердечно-сосудистые, психоневрологические и онкологические, является одной из основных проблем современной медицинской генетики, молекулярной физиологии и патологии. В рамках этой проблемы в настоящее время во всем мире проводятся широкомасштабные исследования, посвященные изучению связи вариаций геномной последовательности ДНК с различными патологиями.

В настоящее время интенсивно развиваются методы биоинформатики, направленные на оценку влияния полиморфизмов на различных уровнях описания молекулярно-генетических систем: генома, транскриптома, протеома и генных сетей.

Замены одного нуклеотида (однонуклеотидные полиморфизмы, ОНП) – наиболее распространенный и интенсивно изучаемый тип полиморфизма последовательностей ДНК. Данные по ОНП человека в настоящее время получаются в достаточно большом количестве, в частности в результате проектов по полногеномным исследованиям ассоциаций (Genome-Wide Association Studies, GWAS) (Torkamani *et al*., 2008). В связи с этим становится возможным детальное изучение влияния полиморфизмов на возникновение социально значимых заболеваний. Оно основывается преимущественно на двух методиках: а) анализе ассоциаций полиморфизмов с заболеваниями; б) системной биологии.

Первый подход основан на статистическом выявлении взаимосвязей между геномными вариациями и риском заболеваний (Psychiatric GWAS Consortium …, 2009). Например, исследования ассоциаций позволили установить взаимосвязь ряда полиморфизмов с предрасположенностью к раку кожи (Gerstenblith *et al*., 2010). Большое количество результатов таких исследований доступно в виде баз данных (Johnson, O'Donnell, 2009). Для реализации такого подхода необходима кропотливая работа по сбору данных и дальнейшей их верификации методами доказательной медицины. Данный подход имеет ряд недостатков. При статистическом анализе ОНП невозможно разделить функционально значимые полиморфизмы и полиморфизмы, сцепленные с признаком, что требует дополнительных исследований для применения результатов, полученных на одной популяции, перед использованием в клинике. Вторым недостатком является сложность изучения редких полиморфизмов в связи с трудностями создания достаточно большой выборки для статистической обработки.

Альтернативным подходом может служить метод предсказания эффекта единичных замен нуклеотидов *in silico*, основанный на информации о структурно-функциональной аннотации генома человека и особенностей его функционирования в рамках модели генных и метаболических сетей (Weston, 2004). Этот подход основан на том, что ОНП в генах могут влиять на человеческий фенотип на разных уровнях экспрессии гена: полиморфизмы в некодирующих регуляторных областях могут вносить повреждения в последовательности сайтов связывания транскрипционных факторов, сплайсинга, нарушая их функционирование на уровне транскрипции или трансляции. Полиморфизмы в кодирующих участках генов могут становиться причиной аминокислотных замен и приводить к изменениям функциональных или структурных свойств кодируемого белка. В совокупности такие повреждения на уровне отдельных генов могут влиять на функционирование генных сетей (Системная компьютерная биология ..., 2008) и приводить к фенотипическим нарушениям на уровне организма. Современные методы биоинформатики позволяют выявлять повреждающие эффекты мутаций как в некодирующих регуляторных участках генов (Савинкова и др., 2009), так и на уровне белков (Иванисенко и др., 2011).

Подход *in silico*, разумеется, не обладает такой степенью доказательности, как широкомасштабные исследования генетических ассоциаций, тем не менее, в последнее время он интенсивно развивается, так как в дополнение к результатам GWAS позволяет оценивать влияние мутаций на возникновение патологий, на основе современных знаний об их молекулярных механизмах (Na *et al*., 2013).

Необходимо отметить, что в настоящее время разработано большое количество информационных ресурсов (баз данных и компьютерных программ), которые нацелены на решение отдельных задач по оценке влияния ОНП на определенные функции генома (Mooney *et al*., 2010). Однако разнородность этих ресурсов, огромный объем информации, необходимой для всесторонней оценки риска возникновения патологий, ее сложность не позволяют экспертам-медикам или биоинформатикам систематизировать и анализировать ее вручную. Эффективное решение этой задачи возможно только при использовании компьютерной системы, которая позволяет в автоматическом режиме на основе данных персонального генома проводить первичную оценку рисков возникновения, прежде всего, социально значимых заболеваний. В основе работы такой системы должен лежать принцип интеграции разнородных данных, полученных в результате работы большого числа компьютерных программ при обработке большого числа баз данных. Результаты работы подобной системы, полученные как с учетом известных ассоциаций полиморфизмов с заболеваниями, так и предсказанные на основе биоинформационного анализа предрасположенности к тому или иному заболеванию, могут использоваться медиком-экспертом для учета индивидуальных особенностей пациента.

В настоящей работе описана модульная компьютерная информационная система (МКИС) для оценки влияния полиморфизмов на возникновение социально значимых заболеваний, включающая базы данных, алгоритмы и математические модели влияния полиморфизмов на функцию генов и генных сетей, особенно при таких социально значимых заболеваниях, как рак и заболевания, связанные с нарушением метаболизма.

МАТЕМАТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ АНАЛИЗА ВЛИЯНИЯ ОНП НА ФУНКЦИЮ ГЕНОВ, СВЯЗАННЫХ С ПОЯВЛЕНИЕМ СОЦИАЛЬНО ЗНАЧИМЫХ ЗАБОЛЕВАНИЙ

Полиморфизм кодирующих районов генов. Достаточно хорошо обоснованной в настоящее время является модель влияния ОНП на структуру и функцию белков – принципиальных компонент генных сетей. Аминокислотные замены в белках могут существенно влиять на функционирование генных сетей и даже приводить к качественному изменению динамики их функционирования. Так, например, мутации в некоторых транскрипционных факторах способны изменять спектр целевых генов, регулируемых данными регуляторными белками (Farnebo *et al*., 2010).

В общем виде мутации по механизму их действия на белок можно разделить на два больших класса: 1) мутации, нарушающие функцию белка, но сохраняющие его пространственную укладку, и 2) мутации, нарушающие структуру (Sanchez-Ruiz *et al*., 2010).

Нарушение функции белка может быть вызвано мутациями, непосредственно расположенными в функциональных центрах белка (каталитических сайтах, сайтах связывания, сайтах посттрансляционных модификаций и т. д.), а также и удаленными от функциональных центров мутациями, действие которых может приводить к изменениям пространственной структуры активных центров белков.

Нарушение структуры белков может проявляться на уровне формирования белковой укладки, т. е. образования белковых форм, принципиально отличающихся по третичной структуре от нативного белка либо вообще не имеющих компактной структуры, либо характеризующихся снижением термодинамической стабильности белков. В последнем случае мутантные белки могут обладать правильной укладкой, однако время существования белка в этой нестабильной форме оказывается ниже по сравнению с нативным белком, и такие белки с более высокой скоростью подвержены протеолитической деградации.

Масштабные исследования полиморфизмов при помощи компьютерных моделей показали, что 90 % единичных мутаций, связанных с заболеваниями, так или иначе понижают стабильность белковой глобулы. В то же время около 70 % наблюдаемых полиморфизмов (из их общего пула) являются, по данным компьютерного моделирования, нейтральными. Примерно 30 % мутаций могут обуславливать заболевания полигенного характера. Использование новых методов компьютерного анализа и представления данных позволило создать информационные ресурсы, посвященные анализу и аннотации ОНП в белках, и привязать информацию об ОНП к структурной и функциональной аннотации белка (Ramensky *et al*., 2002; Cavallo *et al*., 2005; Karchin *et al*., 2005). Это позволяет предсказывать роль ОНП в возникновении заболеваний человека и тем самым планировать ассоциативные исследования по мутациям в геноме человека (Yue *et al*., 2006).

Полиморфизм некодирующих районов генов. Гораздо меньше информации имеется о регуляторных ОНП (рОНП), способных влиять на уровень экспрессии генов-кандидатов. Имеющиеся в литературе примеры рОНП показывают, что такие однонуклеотидные замены часто приводят или к разрушению сайтов связывания различных транскрипционных факторов (ТФ), или к образованию новых сайтов, или же изменяют сродство ТФ к их сайтам. Эти события могут не только менять уровень транскрипции генов, но также радикально влиять на характер их экспрессии вплоть до изменений ее тканеспецифичности и способности реагировать на внешние сигналы. С развитием высокопроизводительных экспериментальных подходов к выявлению участков связывания транскрипционных факторов в масштабе генома (ChIP-seq) и накоплением больших массивов этой информации в результате работы международного постгеномного проекта ENCODE (http://genome.ucsc.edu/) появилась возможность масштабной идентификации регуляторных районов в последовательностях генов на основании данных о скоплении мест связывания различных ТФ в определенных геномных локусах. Это, в свою очередь, открывает возможность отбора ОНП, попадающих в эти локусы и потенциально способных влиять на связывание ТФ с данным районом ДНК. Исследования показали высокую предсказательную способность такого подхода к выявлению рОНП. Так, более 70 % нуклеотидных замен, попадающих в такие районы, способны менять связывание белков с ДНК. Таким образом, становится возможным предсказывать участие некодирующих ОНП в развитии различных патологий.

Влияние полиморфизма на функцию генных сетей. Системный подход к реконструкции механизмов, обеспечивающих проявление полиморфизмов через патологические режимы функционирования генных сетей, приобретает все большее значение и демонстрирует свою эффективность. В качестве примера укажем на работу А. Торкамани с соавт. (Torkamani *et al*., 2008), которые провели анализ молекулярных путей в генных сетях, связанных с заболеваниями человека (биполярное аффективное расстройство, заболевание коронарной артерии, болезнь Крона, гипертензия, ревматоидный артрит, диабеты 1 и 2 типов). Анализировались гены, участвующие в генных сетях, а также влияние на их функцию полиморфизмов, данные по которым были получены в результате полногеномных ассоциативных исследований. Результаты анализа показали, что механизмы возникновения патогенных состояний являются общими для многих заболеваний и обусловлены множеством как общих, так специфических факторов риска. В частности, одни и те же сигнальные пути в генных сетях могут отвечать за возникновение сразу нескольких заболеваний. К таким важным путям были отнесены пути передачи сигналов, вовлекающие рецепторы, аденилатциклазы, протеинкиназы, системы передачи кальциевых сигналов.

В качестве еще одного примера укажем, что к настоящему времени уже известны многие десятки мутаций в генных сетях, контролирующих пищевое поведение, ассоциированные с одним и тем же признаком – избыточной массой тела. При этом примечательно, что механизмы действия этих мутаций основаны на нарушениях регуляторных процессов функционирования указанных выше генных сетей.

Складывающийся к настоящему времени биоинформационный подход к выявлению молекулярно-генетических механизмов мультифакториальных заболеваний включает следующие этапы (Moore *et al*., 2010): реконструкция генных сетей, нарушения которых ассоциированы с теми или иными заболеваниями; идентификация генов и белков, участвующих в функционировании этих генных сетей; функциональная оценка влияния мутаций (полиморфизмов) на уровень экспрессии генов либо структуры/функции/активности белков; оценка характера функциональных нарушений на уровне локальных и интегральных генных сетей организма.

РЕАЛИЗАЦИЯ ПРОГРАММНОГО КОМПЛЕКСА SNP-MED

Архитектура программного комплекса SNP-MED. МКИС SNP-MED включает три функциональных модуля, реализованных в виде программных компонент (ПК) (рис. 1), и базу данных «Информационный ресурс», которая содержит всю необходимую информацию для работы этих компонент.

1. Программная компонента «Геномика» для оценки эффекта влияния ОНП на функционирование регуляторных районов генов. В основе работы этого модуля лежит широкомасштабный поиск совпадений полиморфизмов пациента с известными данными о полиморфизмах и их взаимосвязи с социально значимыми заболеваниями, представленными в общедоступных базах данных, а также предсказание влияния ОНП в регуляторных районах на функцию генов.

2. Программная компонента «Протеомика» для оценки эффекта влияния ОНП в участках генов, кодирующих белки. В основе работы этого модуля лежит широкомасштабный анализ влияния полиморфизмов пациента на нарушение структуры и функций белков, кодируемых генами человека.

3. Программная компонента «Генные сети» для оценки интегрального эффекта влияния ОНП на генные сети. В основе работы этого модуля лежит оценка влияния генетических рисков, найденных в результате работы программных компонент «Геномика» и «Протеомика» на уровне генов и регуляторных взаимодействий, на структуру и функцию генных сетей.

Входные данные в виде последовательности персонального генома или списка ОНП подаются на вход системы пользователем. Эти данные вначале поступают на обработку модулем «Геномика». В результате отбираются ОНП, аннотация которых уже содержится в базах данных, а также ОНП в регуляторных районах генов.

Затем производится обработка ОНП, локализованных в кодирующих участках генов (их влияние на термостабильность и активные центры). На последнем этапе анализа выявленные повреждающие мутации проецируются на структуру генных сетей, которые загружаются пользователем в систему в одном из стандартных форматов. В результате работы ПК пользователь получает аннотацию ОНП, их классификацию по степени повреждающего эффекта, оценку их влияния на функционирование генов, белков и структуру генных сетей.

Для создания МКИС SNP-MED была использована биоинформационная платформа UGENE, которая представляет собой один из широко используемых программных пакетов

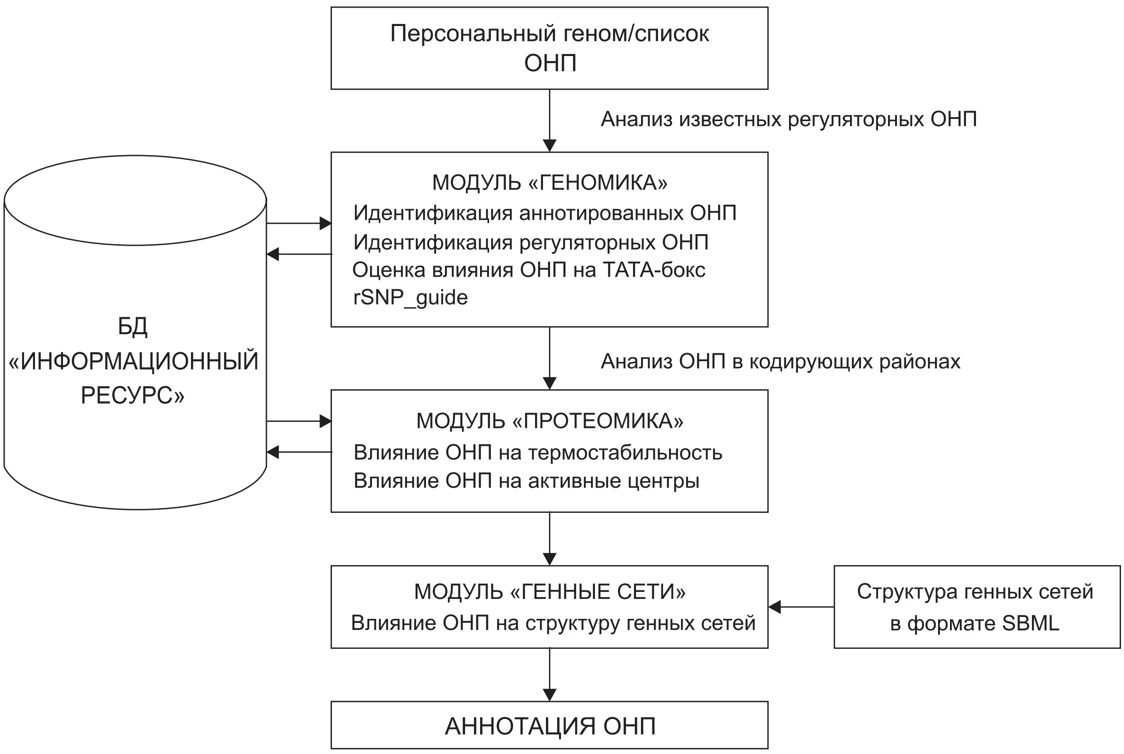

Рис. 1. Основные модули МКИС SNP-MED и их взаимосвязь.

для выполнения различных процедур анализа биологических данных (Okonechnikov, 2012). Высокая популярность UGENE обусловлена, прежде всего, широким спектром его возможностей, кроссплатформенностью, а также его свободным использованием, так как UGENE распространяется на условиях лицензии GNU General Public License, v 2.0.

Для аннотирования однонуклеотидных полиморфизмов используется конструктор вычислительных схем, входящий в состав UGENE, так называемый Workflow Designer (далее WD). Этот модуль позволяет создавать многоступенчатые конвейеры (вычислительные схемы) для обработки биологической информации. Каждая ступень такого конвейера – отдельный алгоритм. В процессе работы схемы они обмениваются между собой сообщениями, содержащими входные данные для одних алгоритмов и/или результаты работы других.

Внутренняя архитектура WD позволяет объединить компоненты МКИС в рамках общего интерфейса вычислительных схем за счет добавления новых вычислительных элементов, соответствующих отдельным алгоритмам. Такой подход позволит в дальнейшем использовать эти компоненты в любых вычислительных схемах WD.

Рассмотрим подробнее функционирование отдельных модулей МКИС SNP-MED.

Программная компонента «Геномика» включает следующие программные модули.

1. Программный модуль для поиска известных ОНП, ассоциированных с социально значимыми заболеваниями, представленных в общедоступных базах данных. К числу этих баз данных относятся такие, как Diseases, dbSNP (Sherry *et al*., 2001), Exome variant server, 1000 Genomes, HapMap. В работе этого модуля также используются последовательность ДНК генома человека и ее функциональная разметка: локализация генов, регуляторных районов, участков сегментных дупликаций и эволюционно-консервативных районов (http://genome.ucsc.edu/). Данный модуль на основе информации о локализации ОНП извлекает доступную аннотацию из упомянутых источников и подает ее на выход модуля.

2. Программный модуль для оценки вероятности нахождения ОНП в регуляторных районах генов. Этот модуль использует информацию по аннотации сайтов связывания десятков транскрипционных факторов с участками генома, полученного в ходе выполнения проекта ENCODE (Rosenbloom *et al*., 2013). Наличие полиморфизмов в этих участках потенциально способно влиять на связывание ТФ с данным районом ДНК. Так, по экспериментальным данным, более 70 % нуклеотидных замен, попадающих в такие районы, способны влиять на связывание белков с ДНК.

3. Программный модуль для оценки влияния ОНП на функционирование районов ТАТА-боксов. В основе анализа – модель взаимодействия белка TBP и фрагмента ДНК, содержащего ТАТА-бокс, описывающая связывание за четыре последовательных шага, отражающих критически значимые этапы функционирования ТАТА-бокса (Пономаренко и др., 2008). Эта модель с высокой точностью позволяет оценить значение аффинности, что было подтверждено экспериментальными исследованиями.

4. Программный модуль для оценки влияния ОНП на функционирование регуляторных районов на основе технологии rSNP_Guide. Данная технология позволяет определить тип транскрипционного фактора, связывание с которым наиболее сильно нарушается в результате мутации в регуляторном районе ДНК.

Программная компонента «Протеомика» включает:

1. Программный модуль для предсказания влияния ОНП на термодинамическую стабильность белков.

2. Программный модуль для идентификации ОНП в функциональных сайтах белков.

База данных «Информационный ресурс» (БДИР) включает информацию, необходимую для функционирования МКИС при анализе влияния ОНП на функцию генов, связанных с появлением социально значимых заболеваний. Содержит аннотацию человеческого генома и известных ОНП, которые использует модуль «Геномика», а также информацию, полученную в результате аннотации ОНП свободно распространяемыми программами:

1) результаты анализа белок-кодирующих последовательностей с использованием алгоритма SIFT (Ng, Henikoff, 2003), оценивающего влияние мутации на функцию белка;

2) результаты анализа белок-кодирующих последовательностей с использованием алгоритма PolyPhen (Adzhubei *et al*., 2010), позволяющего выявить повреждающие мутации в аминокислотных последовательностях;

3) результаты анализа белок-кодирующих последовательностей с использованием алгоритмов PhyloP, LRT и GERP (Pollard *et al*., 2010), оценивающих степень повреждающего эффекта влияния ОНП на основе анализа эволюционной консервативности последовательностей геномов.

Наличие БДИР не требует компьютерных расчетов для каждой ОНП, запрошенной пользователем; программа осуществляет поиск уже подготовленной информации, что обеспечивает быструю обработку данных.

При проектировании программного комплекса большая часть необходимых для аннотирования ОНП алгоритмов была разработана в виде отдельных Web-приложений. В локальном варианте остался доступен лишь анализ влияния полиморфизмов на функционирование TATA-боксов. Соответственно, при разработке функциональности новых вычислительных элементов WD было принято решение для удаленных алгоритмов использовать запросы на соответствующие сервисы, а локальные вызывать непосредственно.

В целом все алгоритмы с программной точки зрения имеют общий интерфейс – требуется передать некоторые атрибуты последовательности, содержащей нуклеотидную замену, а также идентификатор ОНП. На рис. 2 приведена общая схема взаимодействия вычислительных элементов, используемая в данном программном комплексе. Что касается упомянутой процедуры анализа, то в аналогичной диаграмме этапы формирования запроса и получения ответа отсутствуют, а вызов алгоритма производится вычислительным элементом.

Одним из атрибутов каждого вычислительного элемента (за исключением алгоритма «*SNP ChIP*») является локальный путь до базы данных БДИР, содержащей аннотацию генома человека. Обращения к ней позволяют получить параметры последовательности, содержащей полиморфизм, необходимые для вызова каждого из алгоритмов. Затем, используя эти данные, вычислительный элемент вызывает исполнение соответствующего сценария в среде Python, который, в свою очередь, производит обра-

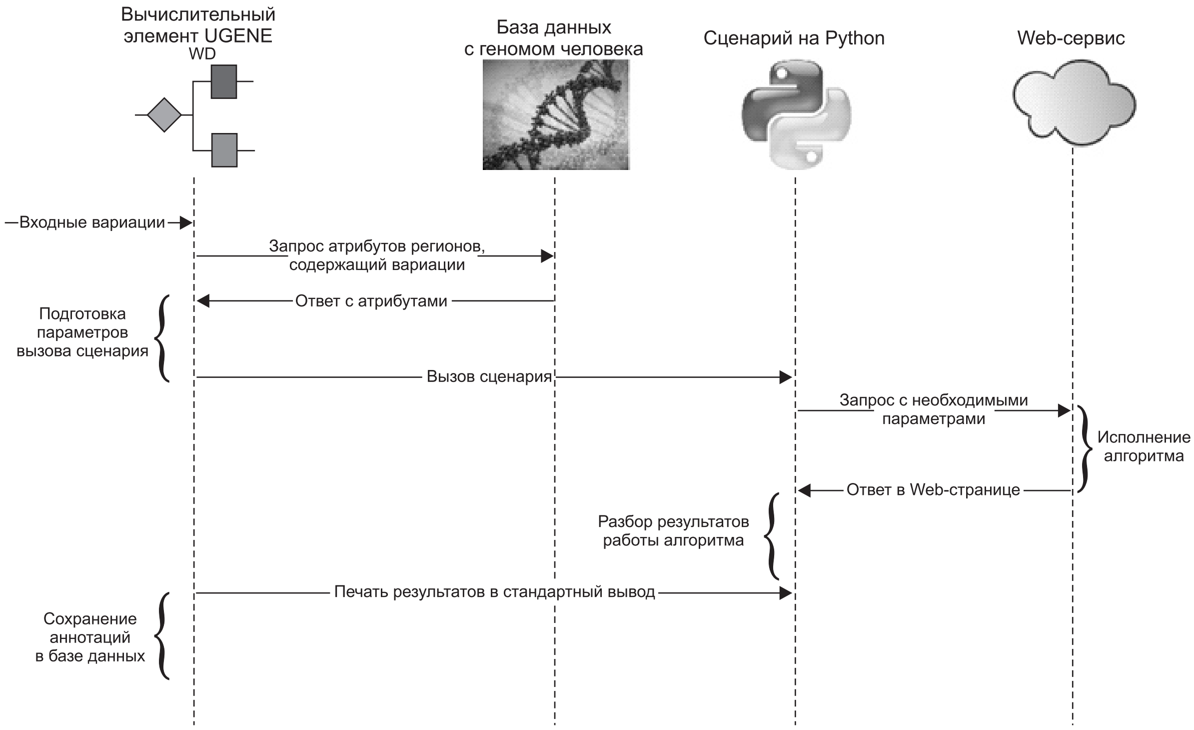

Рис. 2. Диаграмма взаимодействия вычислительного элемента МКИС SNP_MED и БДИР в процессе получения аннотации ОНП на основе использования технологии Workflow Designer UNIPRO.

щение к удаленному сервису, реализующему необходимый алгоритм. После завершения работы сервиса сценарий получает результаты обработки, аннотации ОНП в HTML-формате, которые впоследствии подвергаются разбору. Затем аннотации полиморфизма попадают в стандартный вывод сценария, откуда их считывание производит вычислительный элемент.

На последнем этапе работы того или иного алгоритма полученная информация заносится в общую для всех вычислительных элементов базу данных. Таким образом, последовательно проходя через различные ступени конвейера, аннотация каждой ОНП постепенно дополняется новыми атрибутами в этом хранилище. На основе его содержимого заключительный элемент вычислительной схемы («*Write SNP Report*» на рис. 3) формирует два текстовых отчета – о влиянии полиморфизмов на генную и регуляторную области соответственно (рис. 4, 5).

Данные выдаются в формате SNP-report. Это текстовый файл, в котором аннотация полиморфизма занимает одну или несколько строк. В них располагаются данные в виде столбцов, разделенных знаками табуляции. Число столбцов в каждом из отчетов зависит напрямую от числа задействованных в конвейере алгоритмов. К примеру, для его полной версии (рис. 3) каждой нуклеотидной замене в отчете будут представлены идентификаторы области генома, в котором она обнаружена, а также следующие параметры: идентификатор и местоположение поврежденного гена; идентификатор и кодон поврежденного белка; перечень связанных с поврежденной областью социально значимых заболеваний; оценка повреждающего эффекта с помощью алгоритма SIFT; оценка влияния полиморфизма на термодинамическую стабильность первичной и третичной структур белка; идентификация вариации в функциональных сайтах белка.

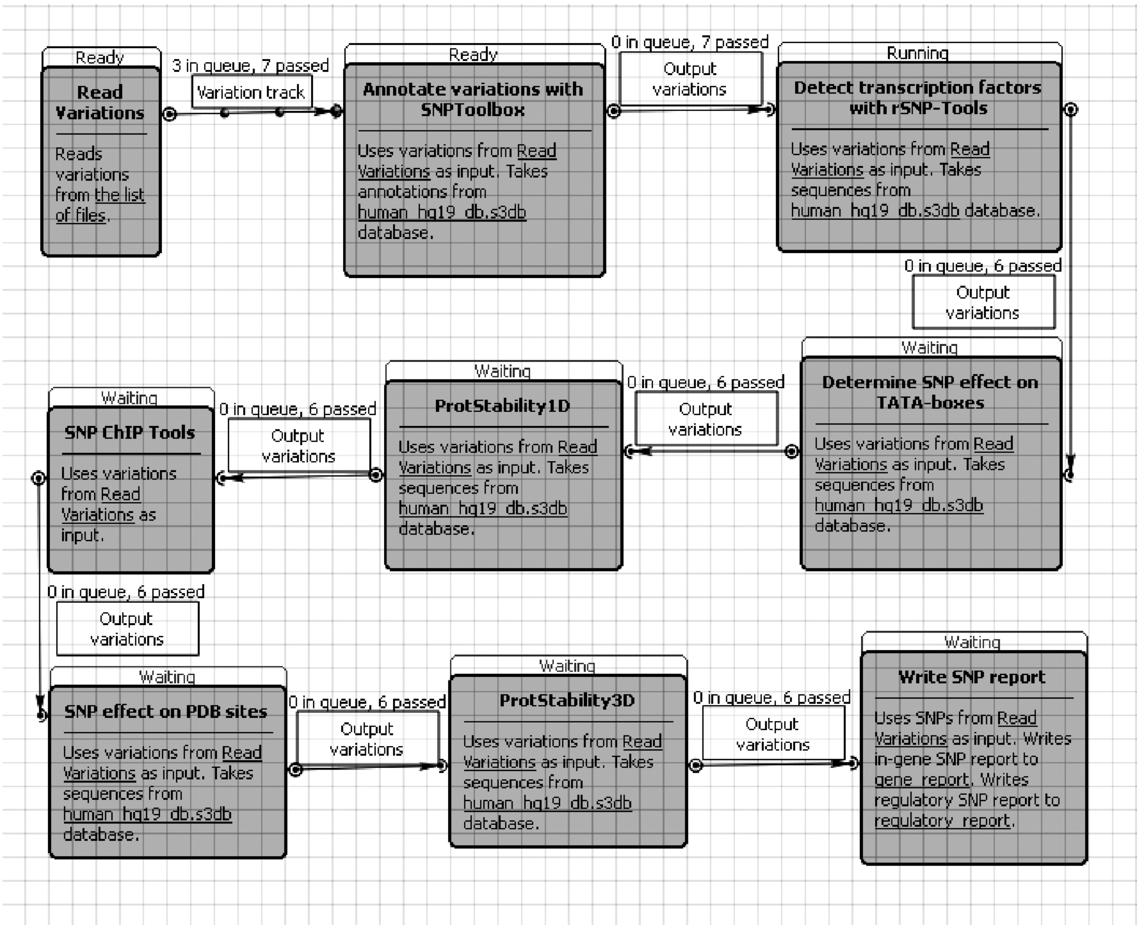

Рис. 3 Детальная последовательность обработки данных в процессе функционирования конвейера МКИС SNP_MED, реализованного в UGENE Workflow Designer в процессе вычислений.

1	#Chr	Position	Allele	dbSNP	Gene	Clinical_significance	Location	Protein	Codon	Substitution
2	chr11	94800448	A/G	-	B2R6B8	CDS.	Exon: 94800056..94800900	Q9BRL6	AAC->GAC	N20D
3	chr13	113634072	G/A	-	AB002360	Factor VII CDS.	Intron: 113623029..113669076	-	-	-
4	chr13	113634072	G/A	-	CCDS45070	Factor VII CDS.	Intron: 113623029..113669076	-	-	-
5	chr13	113634072	G/A	-	KIAA0362	Factor VII CDS.	Intron: 113623773..113669076	-	-	-
6	chr13	113634072	G/A	-	CCDS9527	Factor VII CDS.	Exon: 113633982..113634072.	E9PDN8	GAT->AAT	D31N
7							Donor splice-site.			
8	chr14	22356190	T/A	-	AJ004871	CDS.	Intron: 22192546..22447340	-	-	-
9	chr14	22356190	T/A	-	FR159098	CDS.	Exon: 22205173..23016490	-	-	-
10	chr14	22356190	T/A	-	TCRA	Cysticercosis(2.2),	Intron: 22294244..22994626	-	-	-
11						Taeniasis(2.2),				
12						Leukemia(1.4),				
13						Disease(1.2),				
14						Toxocariasis(1.2),				
15						Echinococcosis(1.1),				
16						Cancer(1.0),				
17						Lymphoma(0.8),				
18						Myoma(0.8),				
19						Pneumothorax(0.7),				
20						Arthritis(0.5)				
21	chr14	22356190	T/A	-	FR004500	CDS.	Exon: 22321073..22981890	-	-	-
22	chr14	22356190	T/A	-	M97714	CDS.	Intron: 22337544..22733720	-	-	-
23	chr14	22356190	T/A	-	AV2S1A1	CDS.	Exon: 22356037..22356190.	-	TGG->AGG	W18R
24							Donor splice-site.			
25	chr3	126386191	G/C	-	C3orf46	CDS.	Exon: 126385949..126386191.	Q8IVU5	AAG->AAC	K133N
26							Donor splice-site.			
27	chr4	265207	C/A	-	B4DXR9	CDS.	Exon: 264464..266419	B4DXR9	GGC->GTC	G448V
28	chr4	265207	C/A	-	B4DXR9	CDS.	Exon: 264464..266419	B4DXR9	GGC->GTC	G480V
29	chr4	152571673	C/G	-	FAM160A1CDS.	CDS.	Exon: 152570611..152571744	Q05DH4	CCA->CGA	P827R
30	chr5	33997584	G/C	-	AMACR	Cancer(2.0),	Intron: 33989608..33998745	-	-	-
31						Adenocarcinoma(2.0),				
32						Carcinoma(2.0),				
33						Disease(1.8),				
34						Adenoma(1.0),				
35						Prostatitis(1.0),				
36						Angiomyolipoma(0.8),				
37						Adrenoleukodystrophy(0.6)				
38	chr5	33997584	G/C	-	A5YM47	prostate cancer;	Intron: 33989608..33998745	-	-	-
39						effects in AMACR are				
40						the cause of				
41						alpha-methylacyl-				
42						CoAracemase deficiency				
43						(AMACRD) effects in				
44						AMACR are the cause of				
45						congenital bile				
46						acidsynthesis defect				
47						type 4 (CBAS4) CDS.				

Рис. 4. Пример отчета об известных ОНП, ассоциированных с заболеваниями, в формате SNP-report.

Верхняя строка содержит названия колонок отчета. Ниже представлены аннотации полиморфизмов, полученные при работе МКИС SNP-MED.

Пример такого отчета приведен на рис. 4.

Что касается отчета о влиянии ОНП на регуляторную область, то он включает следующую информацию о затронутом ОНП регионе: идентификатор промотора, соответствующего данной области; оценка повреждающего эффекта регуляторной области; перечень связанных с поврежденной областью социально значимых заболеваний; список поврежденных сайтов связывания транскрипционных факторов; оценка влияния вариации на функциональную активность ТАТА-боксов.

Пример аннотационного отчета о регуляторных ОНП приведен на рис. 5.

В рамках МКИС SNP-MED разработаны следующие сценарии биоинформационного анализа влияния ОНП на функции генов, связанных с появлением социально значимых заболеваний.

1. Сценарий биоинформационного анализа данных на геномном уровне для идентификации ОНП, ассоциированных с социально значимыми заболеваниями.

2. Сценарий оценки влияния ОНП на регуляторные районы генов. Пользователю выдается список выявленных ОНП, находящихся в регуляторных районах генов и известных в качестве ассоциированных с социально значимыми заболеваниями, с оценкой их функциональной значимости.

3. Сценарий влияния ОНП на функциональную активность ТАТА-боксов в регуляторных областях генов. Пользователю выдается список генов, ТАТА-боксы которых подвержены значимому изменению уровня активности в результате найденных ОНП, с описанием эффекта ОНП.

4. Сценарий оценки влияния ОНП на функциональную активность сайтов связывания транскрипционных факторов в регуляторных областях генов. Пользователю выдается список генов, у которых сайты связывания транскрип-

1	#Chr	Position	Allele	dbSNP	Promoter_of_Gene	Clinical_significance	From_transcroption_start	rSNPTools_factors	ChIPTools
2	chr17	39993435	T/C	-	C9JKC4	-911	-	-	-
3	chr17	39993435	T/C	-	5NT3L_HUMAN	-911	-	-	-
4	chr17	39993435	T/C	-	KLH10_HUMAN	-608	-	-	-
5	chr17	39993435	T/C	-	AK302141	-642	-	-	-
6	chr17	39993435	T/C	-	AK301797	-642	-	-	-
7	chr17	39993435	T/C	-	5NT3L_HUMAN	-946	-	-	-
8	chr17	39993435	T/C	-	C9JKC4	-911	-	-	-
9	chr17	39993435	T/C	-	5NT3L_HUMAN	-911	-	-	-
10	chr17	39993435	T/C	-	KLH10_HUMAN	-608	-	-	-
11	chr17	39993435	T/C	-	AK302141	-642	-	-	-
12	chr17	39993435	T/C	-	AK301797	-642	-	-	-
13	chr17	39993435	T/C	-	5NT3L_HUMAN	-946	-	-	-
14	chr17	39993435	T/C	-	C9JKC4	-911	-	-	-
15	chr17	39993435	T/C	-	5NT3L_HUMAN	-911	-	-	-
16	chr17	39993435	T/C	-	KLH10_HUMAN	-608	-	-	-
17	chr17	39993435	T/C	-	AK302141	-642	-	-	-
18	chr17	39993435	T/C	-	AK301797	-642	-	-	-
19	chr17	39993435	T/C	-	5NT3L_HUMAN	-946	-	-	-
20	chr17	39993435	T/C	-	C9JKC4	-911	-	-	-
21	chr17	39993435	T/C	-	5NT3L_HUMAN	-911	-	-	-
22	chr6	32635046	A/G	rs76356512	HLA-DQB1	-579	-	-	-
23						Disease(2.0),			
24						Lymphopenia(1.3),			
25						Scleroderma(1.1),			
26						Leukopenia(1.1),			
27						Podoconiosis(1.0),			
28						Sarcoidosis(1.0),			
29						Narcolepsy(0.8),			
30						Dermatitis(0.7),			
31						Schizophrenia(0.7),			
32						Elephantiasis(0.7),			
33						Thyroiditis(0.7),			
34						Glomerulonephritis(0.7),			
35						Hyperthyroidism(0.6),			
36						Arthritis(0.6),			
37						Nephritis(0.6),			
38						Cancer(0.6),			
39						Agammaglobulinemia(0.6)			
40	chr6	32635046	A/G	rs76356512	A2AAZ0	-579	-	-	-
41						Leprosy; diabetes,			
42						type 1; pancreatitis,			
43						autoimmune;			
44						pancreatitis, chronic			
45						calcifying;			
46						periodontitis;			
47						infertility, tubal			

Рис. 5. Пример отчета о влиянии полиморфизмов на регуляторную область в формате SNP-report.

Верхняя строка содержит названия колонок отчета. Ниже представлены аннотации полиморфизмов, полученные при работе МКИС SNP-MED.

ционных факторов подвержены изменению функциональной активности в результате найденных ОНП, с описанием эффекта ОНП.

5. Типовой сценарий биоинформационного анализа данных на протеомном уровне, включая предсказание влияния ОНП на термодинамическую стабильность белков, идентификацию ОНП в функциональных сайтах белков.

6. Типовой сценарий биоинформационного анализа влияния ОНП на генные сети.

ЗАКЛЮЧЕНИЕ

Разработана модульная компьютерная информационная система SNP-MED для анализа влияния ОНП на функцию генов, связанных с появлением социально значимых заболеваний, в состав которой входят:

1. Программная компонента «Геномика», включающая программные модули или интерфейсы к сервисам поиска известных ОНП, ассоциированных с социально значимыми заболеваниями, оценки вероятности нахождения ОНП в регуляторных районах генов, оценки влияния ОНП на функционирование районов ТАТА-боксов и регуляторных районов.

2. Программная компонента «Протеомика», включающая программные модули или интерфейсы к сервисам идентификации известных ОНП в кодирующих участках генов, ассоциированных с социально значимыми заболеваниями, предсказания влияния ОНП на термодинамическую стабильность белков и идентификации ОНП в функциональных сайтах белков.

3. Программная компонента «Генные сети», позволяющая оценивать эффект влияния ОНП на генные сети.

4. База данных «Информационный ресурс», включающая информацию, необходимую для функционирования МКИС SNP-MED при анализе влияния ОНП на функцию генов, связанных с появлением социально значимых заболеваний.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке Министерства образования и науки Российской Федерации (Госконтракт № 14.512.11.0094).

ЛИТЕРАТУРА

Иванисенко В.А., Деменков П.С., Иванисенко Т.В., Колчанов Н.А. Protein structure discovery: пакет программ для решения задач компьютерной протеомики // Биоорган. химия. 2011. Т. 37. № 1. С. 22–35.

Пономаренко П.М., Савинкова Л.К., Драчкова И.А. и др. Пошаговая модель связывания TBP/TATA-бокс позволяет предсказать наследственное заболевание человека по точечному полиморфизму // Докл. АН. 2008. Т. 419. С. 828–832.

Савинкова Л.К., Пономаренко М.П., Пономаренко П.М. и др. Полиморфизмы TATA-боксов промоторов генов человека и ассоциированные с ними наследственные патологии // Биохимия. 2009. Т. 4. № 4. С. 149–163.

Системная компьютерная биология // Под ред. Н.А. Колчанов, С.С. Гончаров, В.А. Лихошва, В.А. Иванисенко. Новосибирск: СО РАН, 2008.

Adzhubei I.A., Schmidt S., Peshkin L. *et al.* A method and server for predicting damaging missense mutations // Nature Meth. 2010. V. 7. No. 4. P. 248–249.

Cavallo A., Martin A.C. Mapping SNPs to protein sequence and structure data // Bioinformatics. 2005. V. 21. P. 1443–1450.

Farnebo M., Bykov V.J., Wiman K.G. The p53 tumor suppressor: a master regulator of diverse cellular processes and therapeutic target in cancer // Biochem. Biophys Res. Commun. 2010. P. 85–89.

Gerstenblith M.R., Shi J., Landi M.T. Genome-wide association studies of pigmentation and skin cancer: a review and meta-analysis // Pigment Cell Melanoma Res. 2010. V. 23. No. 5. P. 587–606.

Johnson A.D., O'Donnell C.J. An open access database of genome-wide association results // BMC Med. Genet. 2009. V. 10. No. 1. P. 6.

Karchin R., Diekhans M., Kelly L. *et al.* LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources // Bioinformatics. 2005. V. 21. P. 2814–2820.

Mooney S.D., Krishnan V.G., Evani U.S. Bioinformatic tools for identifying disease gene and SNP candidates // In Genetic Variation. 2010. P. 307–319.

Moore J.H., Asselbergs F.W., Williams S.M. Bioinformatics challenges for genome-wide association studies // Bioinformatics. 2010. V. 26. P. 445–455.

Na Y.J., Cho Y., Kim J.H. AnsNGS: An annotation system to sequence variations of next generation sequencing data for disease-related phenotypes // Healthcare Inform. Res. 2013. V. 19. No. 1. P. 50–55.

Ng P.C., Henikoff S. SIFT: Predicting amino acid changes that affect protein function // Nucl. Acids Res. 2003. V. 31. P. 3812–3814.

Okonechnikov K., Golosova O., Fursov M. *et al*. Unipro UGENE: a unified bioinformatics toolkit // Bioinformatics. 2012. V. 28. P. 1166–1167.

Pollard K.S., Hubisz M.J., Rosenbloom K.R., Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies // Genome Res. 2010. V. 20. No. 1. P. 110–121.

Psychiatric GWAS Consortium Steering Committee. A Framework for Interpreting Genome-Wide Association Studies of Psychiatric Disorders // Mol. Psychiatry. 2009. V. 14. No. 1. P. 10.

Ramensky V., Bork P., Sunyaev S. Human non-synonymous SNPs: server and survey // Nucl. Acids Res. 2002. V. 30. P. 3894–3900.

Rosenbloom K.R., Sloan C.A., Malladi V.S. *et al.* ENCODE Data in the UCSC Genome Browser: year 5 update // Nucl. Acids Res. 2013. P. D56–D63.

Sanchez-Ruiz J.M. Protein kinetic stability // Biophys. Chem. 2010. V. 148. P. 1–15.

Sherry S.T., Ward M.H., Kholodov M. *et al.* dbSNP: the NCBI database of genetic variation // Nucl. Acids Res. 2001. No. 29. P. 308–311.

Torkamani A., Topol E.J., Schork N.J. Pathway analysis of seven common diseases assessed by genome-wide association // Genomics. 2008. No. 92. P. 265–272.

Weston A.D., L H. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine // J. Proteome Res. 2004. V. 3. No. 2. P. 179–196.

Yue P., Melamud E., Moult J. SNPs3D: Candidate gene and SNP selection for association studies // BMC Bioinformatics. 2006. No. 7. P. 166.

THE SNP-MED SYSTEM FOR ANALYSIS OF THE EFFECT OF SINGLE-NUCLEOTIDE POLYMORPHISMS ON THE FUNCTION OF GENES ASSOCIATED WITH SOCIALLY SIGNIFICANT DISEASES

N.L. Podkolodnyy[1], D.A. Afonnikov[1], Yu.Yu. Vaskin[2], L.O. Bryzgalov[1], V.A. Ivanisenko[1], P.S. Demenkov[1], M.P. Ponomarenko[1], D.A. Rasskazov[1], K.V. Gunbin[1], I.V. Protsyuk[2], I.Yu. Shutov[2], P.N. Leontyev[2], M.Yu. Fursov[2], N.P. Bondar[1], E.V. Antontseva[1], T.I. Merkulova[1], N.A. Kolchanov[1]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: pnl@bionet.nsc.ru;
[2] Novosibirsk Center of Information Technologies «UNIPRO», Novosibirsk, Russia

Summary

This paper describes the SNP-MED modular computer-based information system for estimation of the influence of single nucleotide polymorphisms (SNPs) on the function of genes associated with the risk of socially significant diseases. The system includes software components Genomics, Proteomics, Gene networks and the Information resource database (BDIR).

Key words: bioinformatics, SNP, personalized medicine.