ՀԱՅԱՍՏԱՆԻ ՀԱՆՐԱՊԵՏՈՒԹՅԱՆ
ԳԻՏՈՒԹՅՈՒՆՆԵՐԻ ԱԶԳԱՅԻՆ ԱԿԱԴԵՄԻԱ
ՄՈԼԵԿՈՒԼԱՅԻՆ ԿԵՆՍԱԲԱՆՈՒԹՅԱՆ ԻՆՍՏԻՏՈՒՏ

Արսեն Արտաշեսի Առաքելյան

ԿԵՆՍԱԻՆՖՈՐՄԱՏԻԿԱԿԱՆ ՄՈՏԵՑՈՒՄՆԵՐԻ ՄՇԱԿՈՒՄ ՄԱՐԴՈՒ ՔՐՈՆԻԿ ՀԻՎԱՆԴՈՒԹՅՈՒՆՆԵՐԻ ՉԱՐԳԱՑՄԱՆ ՄՈԼԵԿՈՒԼԱՅԻՆ ՄԵԽԱՆԻԶՄՆԵՐԻ ՈՒՍՈՒՄՆԱՍԻՐՈՒԹՅԱՆ ՀԱՄԱՐ

Գ.00.03 – «Մոլեկուլային և բջջային կենսաբանություն» մասնագիտությամբ կենսաբանական գիտությունների դոկտորի գիտական աստիճանի հայցման ատենախոսության

ՍԵՂՄԱԳԻՐ

ԵՐԵՎԱՆ – 2019

НАЦИОНАЛЬНАЯ АКАДЕМИЯ НАУК РЕСПУБЛИКИ АРМЕНИЯ
ИНСТИТУТ МОЛЕКУЛЯРНОЙ БИОЛОГИИ

Аракелян Арсен Арташесович

РАЗРАБОТКА БИОИНФОРМАТИЧЕСКИХ ПОДХОДОВ ДЛЯ ИЗУЧЕНИЯ МОЛЕКУЛЯРНЫХ МЕХАНИЗМОВ РАЗВИТИЯ ХРОНИЧЕСКИХ ЗАБОЛЕВАНИЙ ЧЕЛОВЕКА

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
доктора биологических наук по специальности
03.00.03 – “Молекулярная и клеточная биология”

ЕРЕВАН – 2019

Ատենախոսության թեման հաստատվել է ՀՀ ԳԱԱ Մոլեկուլային կենսաբանության ինստիտուտի գիտական խորհրդում:

Գիտական խորհրդատուներ՝	ՀՀ ԳԱԱ թղթ. անդ., ֆ.-մ.գ.դ., պրոֆ. Լ.Հ. Ասլանյան
	ՀՀ ԳԱԱ թղթ. անդ., կ.գ.դ., պրոֆ. Ա.Ս. Բոյաջյան
Պաշտոնական ընդդիմախոսներ՝	ՀՀ ԳԱԱ թղթ. անդ., կ.գ.դ., պրոֆ. Է.Ս. Գևորգյան ֆ.-մ.գ.դ., պրոֆ. Վ.Բ. Առաքելյան կ.գ.դ., պրոֆ. Պ.Ֆ. Շտադլեր
Առաջատար կազմակերպություն՝	Հարավային դաշնային համալսարանի Դ.Ի. Իվանովսու անվան կենսաբանության և կենսատեխնոլոգիայի ակադեմիա

Ատենախոսության պաշտպանությունը տեղի կունենա 2019թ. դեկտեմբերի 12-ին, ժամը 14^{00}-ին ՀՀ ԳԱԱ Մոլեկուլային կենսաբանության ինստիտուտում, 042 մասնագիտական խորհրդում (ՀՀ, 0014, ք. Երևան, Հասրաթյան 7):

Ատենախոսությանը կարելի է ծանոթանալ ՀՀ ԳԱԱ Մոլեկուլային կենսաբանության ինստիտուտի գրադարանում և http://molbiol.sci.am կայքում:

Սեղմագիրն առաքվել է 2019թ. նոյեմբերի 1-ին:

042 մասնագիտական խորհրդի
գիտ. քարտուղար, կենս. գիտ. թեկնածու Գ.Մ. Մկրտչյան

Тема диссертации утверждена на заседании ученого совета Института молекулярной биологии НАН РА.

Научные консультанты:	член-корр. НАН РА, д.ф.-м. н., проф. Асланян Л.А.
	член-корр. НАН РА, д.б.н., проф. Бояджян А.С.
Официальные оппоненты:	член-корр. НАН РА, д.б.н., проф. Геворгян Э.С. д.ф.-м. н., проф. Аракелян В.Б. д.б.н., проф. Штадлер П. Ф.
Ведущая организация:	Академия биологии и биотехнологии им. Д.И. Ивановского Южного федерального университета

Защита диссертации состоится 12 декабря 2019г. в 14^{00} часов на заседании специализированного совета 042, в Институте молекулярной биологии НАН РА (РА, 0014, г. Ереван, ул. Асратяна 7).

С диссертацией можно ознакомиться в библиотеке Института молекулярной биологии НАН РА и на сайте http://molbiol.sci.am.

Автореферат разослан 1 ноября 2019 г.

Ученый секретарь специализированного совета 042, кандидат биол. наук Мкртчян Г.М.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы

Хронические неинфекционные заболевания (НИЗ) человека являются наиболее распространенными нозологиями, приводящими к высокой смертности среди населения и грозят стать новыми эпидемиями XXI века (Dexter et al., 2010; WHO, 2017). Благодаря значительным научным достижениям, полученным в конце XX столетия, было выявлено, что этиопатогенез НИЗ человека обусловлен генетическим компонентом (Katsanis, 2016), а многочисленные междисциплинарные исследования показали, что ключевым фактором для идентификации молекулярных мишеней, разработки новых, эффективных диагностических, профилактических и терапевтических подходов для борьбы с хроническими заболеваниями является понимание молекулярных механизмов их развития (Mias & Snyder, 2013; Gandal et al., 2016).

Большинство современных поисковых исследований молекулярных механизмов развития заболеваний основано на анализе глобальной экспрессии генов (транскриптома) ввиду относительной простоты методов измерений уровней РНК и их корреляции с уровнями белка. Показано, что исследования на уровне транскриптома важны для интерпретации функциональных элементов генома и выявления молекулярных механизмов, лежащих как в основе нормальной деятельности клеток и тканей, так и для понимания процессов развития и патомеханизмов заболеваний.

При анализе транскриптома для получения достоверных результатов и их корректной интерпретации требуется применение адекватных методов классической и многомерной статистики и машинного обучения, а также наличие репрезентативной выборки. Между тем высокая стоимость экспериментов и ограничения, связанные с доступностью исследуемых образцов, приводят к тому, что формируемая выборка не всегда соответствует указанному требованию. Кроме того, в используемых современных методах анализа транскриптома игнорируются функциональные связи между белками, кодируемыми соответствующими РНК и объединенными в биологические пути, что не может не отражаться на корректной интерпретации полученных результатов. В связи с этим была поставлена задача поиска новых подходов анализа глобальной экспрессии генов.

Цель и задачи исследования

Цель работы: разработать биоинформатические алгоритмы и программные пакеты для исследования глобальной экспрессии генов и активности биологических путей с целью изучения молекулярных механизмов развития хронических неинфекционных заболеваний человека.

Задачи работы:

- разработать биоинформатические подходы для анализа экспрессии генов и идентификации потенциальных биомаркеров
- разработать биоинформатические методы для моделирования активности биологических путей

На основе сконструированных биоинформатических инструментов

- оценить влияние топологии биологических путей на результаты функционального анализа глобальной экспрессии генов
- проанализировать топологическую резистентность биологических путей к мутациям
- исследовать общие и специфические особенности активации биологических путей при онкологических и хронических воспалительных заболеваниях легких

- провести сравнительный анализ профилей активации биологических путей при аутоиммунных и аутовоспалительных заболеваниях
- исследовать влияние единичных мутаций на профиль экспрессии генов и активации биологических путей при моногенных аутовоспалительных заболеваниях
- оценить дифференциальную экспрессию генов при посттравматическом стрессовом расстройстве.

В качестве экспериментальных данных, проанализированных с помощью разработанных автором компьютерных методов, использовались результаты исследований по полно- и широкогеномной экспрессии генов, публично доступных в базе данных Gene Expression Omnibus (GEO) (Barrett et al., 2011). Автор диссертации выражает коллегам благодарность за предоставление этой информации.

Научная новизна, теоретическое значение и научно-практическая ценность работы

В ходе работы разработан комплекс новых биоинформатических методов для исследования молекулярных механизмов патогенеза заболеваний с использованием данных о глобальной экспрессии генов и топологии биологических путей. В частности, нами был предложен алгоритм растущих опорных множеств, основанный на методах добычи данных и выделения признаков, для идентификации дифференциально экспрессируемых генов и их групп, которые могут служить биомаркерами развития патологических состояний. Кроме того, с помощью этого метода возможно реконструировать и выявлять новые биологические пути.

Разработаны программы для редактирования и визуализации биологических путей KEGG Pathways. В них впервые была внедрена возможность полуавтоматической коррекции взаимодействий между компонентами биологических путей, а также модификации по тканевой специфичности и белок-белковым взаимодействиям, что существенно повысило точность получаемых результатов биоинформатического анализа и их интерпретации.

С помощью этих программ создана коллекция сигнальных, метаболических и регуляторных путей, использованная в дальнейших исследованиях.

Разработан алгоритм и ряд программных пакетов, позволяющих осуществить непосредственный переход от по-генного подхода анализа дифференциальной экспрессии генов к методам системной биологии, в частности, на уровне биологических путей. Уникальной особенностью данного биоинформатического инструмента является возможность мониторинга изменений активности биологических путей как в целом, так и в их отдельных ответвлениях, что позволяет точнее предсказать биологические процессы, которые будут затронуты в зависимости от изменений профиля активации конкретных путей.

Предложенные вычислительные методы успешно использованы для изучения молекулярных механизмов патогенеза целого спектра неинфекционных заболеваний человека. В частности, получены новые сведения о специфичности и сходстве профилей активации биологических путей при раковой и фиброзной трансформации легких, предложена молекулярная классификация интерстициальных легочных заболеваний. Показана роль дисбаланса иммунного и воспалительного ответа в патогенезе аутоиммунных и аутовоспалительных заболеваний. Выявлена зависимость резистентности биологических путей к мутациям и идентифицированы ключевые гены, мутации в которых могут существенно влиять на нормальные функции клеток. Более того, спектр применения разработанных нами методов гораздо шире. Они могут использоваться при анализе любых типов динамических данных (протеом, метаболом, метилом).

Вышеизложенное теоретическое значение и научно-практическая ценность диссертации подтверждается активным цитированием ключевых статей (Arsen Arakelyan - Google Scholar Citations: https://scholar.google.com/citations?user=z8SFEuAAAAAJ&hl=en), грантовой поддержкой работ (гранты ANSEF NS-molbio-2319,NS molbio-3808, NS molbio-3808; ГКН МинОбр РА: 15T-1F150, 16GE-025 и 18RF-112; грант Федерального министерства образования и научных исследований Германии FFE-034).

Публикации по теме работы и апробация

По теме диссертации опубликованы 24 статьи в отечественных и международных рецензируемых журналах и сборниках, 2 обзорные главы в зарубежных изданиях и 8 тезисов конференций.

Основные положения диссертационной работы доложены на VII, VIII, IX международных конференциях "Computer Science and Information Technologies" (2009, 2011, 2013, Ереван), международных конгрессах Европейского респираторного общества (Барселона 2013, Мюнхен 2014, Амстердам 2015), Международной научно-практической конференции "Перспективы развития гематологии и трансфузиологии" (Ереван, 2008), IV съезде Российского общества биохимиков и молекулярных биологов (Новосибирск, 2008), школе молодых ученых «Биоинформатика и системная биология» (Новосибирск, 2010), II Международной научно-практической конференции «Постгеномные методы анализа в биологии, лабораторной и клинической медицине: геномика, протеомика, биоинформатика» (Новосибирск, 2011), VIII международной конференции «Биоинформатика Регуляции и Структуры Геномов\Системная Биология - БГРС\СБ-2012» (Новосибирск, 2012), конференции молодых ученых «Новые аспекты молекулярной биотехнологии и биохимии» (Ереван, 2013), Международном симпозиуме по нейробиологии и биологической психиатрии «STRESS AND BEHAVIOR» (Ереван 2013), 23-й Международной годичной конференции «Annual International Conference on Intelligent Systems for Molecular Biology и 14-й Европейской конференции по вычислительной биологии 2015» (Дублин 2015).

Структура и объем работы

Диссертация состоит из введения, списка сокращений и основных обозначений, четырех основных глав: «Обзор литературы», «Разработка логико-комбинаторных подходов анализа экспрессии генов», «Разработка логико-комбинаторных подходов анализа экспрессии генов» и «Исследование молекулярных механизмов патогенеза комплексных заболеваний человека», заключения и выводов. Объем диссертации составляет 231 машинописную страницу, включая 63 рисунка и 16 таблиц. Список литературы содержит 428 источников.

РАЗРАБОТКА ЛОГИКО-КОМБИНАТОРНЫХ ПОДХОДОВ АНАЛИЗА ЭКСПРЕССИИ ГЕНОВ

Основные типы задач обработки данных в анализе транскриптома

В зависимости от поставленной задачи о нахождении функциональных зависимостей между генотипом и фенотипом можно выделить три основных типа сравнительного анализа экспрессии (анализа дифференциальной экспрессии) генов.

"Определение эффективных групп генов для сравнения классов" используется для анализа дифференциальной экспрессии генов в группах с заранее определенными свойствами. Эта стратегия позволяет исследовать механизмы развития заболеваний, а также идентифицировать гены (группы генов), которые могут служить диагностическими и прогностическими биомаркерами, мишенями для дизайна лекарств, и т.д. Примером подобного исследования является, в частности, идентификация дифференциальной

экспрессии генов при системной красной волчанке (Crow & Wohlgemuth, 2003). "Определение эффективных групп генов для сравнения классов", рассмотрено, среди прочих, в статье Tarca et al. (2006). Данный тип исследования нацелен на определение тех генов и биологических процессов, которые либо ответственны за фенотипические различия между классами, либо сами являются его последствиями. Исследуется вся доступная выборка объектов, но не ставится цель предсказания класса новых объектов (Tarca et al., 2006). В теории распознавания образов этой задаче соответствует так называемый этап отбора признаков (feature selection). Основная гипотеза, лежащая в основе этого типа анализа транскриптома, может быть сформулирована следующим образом: ***насколько и какие флуктуации в уровнях экспрессии генов связаны с фенотипическими различиями исследуемых классов, а не вызваны случайными факторами.***

При такой постановке задачи традиционными подходами для ее решения являются методы математической статистики, в частности, методы проверки гипотез (Arakelyan et al., 2013). Сегодня наиболее широко используемым инструментом является по-генный анализ экспрессии, который оценивает статистическую значимость изменений для каждого гена в отдельности между двумя или более классами. Результатом такого анализа является список дифференциально экспрессируемых генов, чьи абсолютные значения превышают заданный исследователем порог. Как отмечено выше, при наличии достаточно большой выборки с помощью методов математической статистики можно на вероятностном уровне сделать выводы о свойствах исследуемых объектов и выдвинуть некоторые предположения. Методы проверки гипотез, корреляций, регрессий и генерализованных линейных моделей для этих задач являются основными компонентами статистического подхода (Arakelyan et al., 2013). Однако допущения о нормальности распределения, независимости исследуемых переменных и, зачастую, малый размер выборки делают применение этих походов неадекватными и уменьшают статистическую мощность исследования.

Другим методом анализа экспрессии генов является логико-комбинаторное машинное обучение и распознавание образов, которое было разработано начиная с 70-х годов прошлого века в школе Ю.И. Журавлева (Журавлев, 1998). Эти подходы разрабатывались как альтернатива методам классической статистики и, в большинстве своем, не нуждаются в соблюдении множества статистических условий и предпосылок. Методы машинного обучения основаны на выявлении закономерностей по эмпирическим данным (обучение по прецедентам) и эвристике, являясь более гибкими в применении. Основной отличительной чертой методов машинного обучения является необходимость в обучающей выборке, которая должна быть высокого качества и достаточного размера для описания всех закономерностей исследуемого феномена. Базовые подходы, используемые в распознавании образов, включают модель распознавания PAC (probably approximately correct, вероятностно-приблизительно-корректное обучение), бустинг, опорных векторов (support vector machine) и другие, а также метрические алгоритмы, такие как многопараметрические алгоритмы голосования, логического разделения, нейронных сетей и т.д. Однако алгоритмы анализа экспрессии генов, основанные на машинном обучении, плохо справляются с «проклятием» размерности, когда размер выборки (исследуемых образцов, объектов) намного меньше количества признаков (генов). Таким образом, вопрос разработки подходов для анализа данных экспрессии генов по классам (что является известной и хорошо исследованной задачей), до сих пор содержит ряд нерешеных вопросов, таких как адаптация к малым размерностям данных обучения, снижения размерности модели, анализ и поиск значимых и незашумленных групп признаков и т.п.

Формально задачу анализа фенотипов по экспрессиям групп генов можно определить следующим образом. Задано множество обучения/таблица классификации M, состоящая из d объектов столбцов по t признакам-строкам, принадлежаших к исследуемым классам (ИК). Может быть задана и таблица $\mathbf{K}$, состоящая из c аналогичных объектов, составляюших контрольную выборку, представленных в виде c столбцов таблицы K, и t признаков (в строках таблицы). В случае экспериментов по анализу транскриптома таблицы обучения и контроля представляют собой матрицы экспрессии генов. Целью анализа является нахождение признаков и их групп, способных с определенной точностью классифицировать объекты в M и K, что вписывается в тип задач классификации образов и распознавания изображений (Журавлев и Никифоров, 1971).

В введенных терминах любая строка/признак матрицы экспрессии M может служить классификатором, с определенной точностью классифицирущей объекты множества M, например, при помощи простой линейной гиперплоскости. С другой стороны, очевидно, что точность классификации разных строк матрицы будет отличаться. Кроме того, строки могут комбинироваться, формируя пары, триплеты и т.д., которые также способны классифицировать объекты матрицы M. Все возможные комбинации значений наборов из t признаков можно представить в виде решетки, начинающейся с пустого множества, затем на 1-м уровне – 1-элементные наборы, на 2-м – 2-элементные наборы и т.д. На уровне k представлены все k-элементные наборы, которые связаны со всеми своими k $(k-1)$-элементными подмно-жествами (рис. 1), а также $t-k$ $(k+1)$-элементными над-множествами. Таким образом, полное множество всех непустых комбинаций признаков составляет $2^t - 1 = \sum_{k=1}^{t} \frac{t!}{k!(t-k!)}$, где k – количество признаков. Очевидно, что при достаточно большом значении t простой перебор всех возможных комбинаций k признаков является проблемой неприемлемой вычислительной сложности.

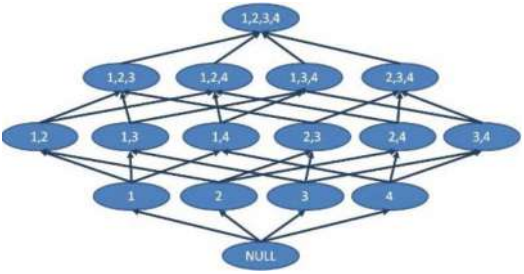

Рисунок 1. Возможные комбинации классификаторов из 4-х признаков (генов).

Для преодоления данной проблемы мы обратимся к модели алгоритмов распознавания образов, предложенной Ю. Журавлевым (Журавлев и Никифоров, 1971) и ее адаптированной версией к использованию для задач добычи данных (дата майнинг) (Aslanyan & Sahakyan, 2009). В основе этих моделей лежит механизм уменьшения вычислительной сложности алгоритмов за счет определения системы опорных множеств. Система опорных множеств представляет собой минимальный набор подмножеств признаков, по которым распознаюший алгоритм максимально/достаточно правильно классифицирует объекты. Рассмотренные здесь опорные множества отличаются от таковых, введенных в симплекс методе решения задачи линейного программирования, а также от тех, которые используются в алгоритмах опорных векторов машинного обучения. В первом случае опорное множество соответствует вершинам симплекса/многогранника допустимых решений задачи линейного программирования (Aslanyan & Sahakyan, 2017), во втором – указывает на граничные вершины/объекты разделяющей гиперплоскости с максимальным отступом (margin) (Sain & Vapnik, 2006).

При исследовании биологической проблемы определение области поиска опорных множеств может быть либо основанной на априорных знаниях задачи, либо на автоматизированном анализе и оценке данных экспериментов. Априорные знания могут сушественно сузить область поиска опорных множеств, например, в нашей задаче –

ограничить его генами, вовлеченными в один или несколько биологических путей или процессов. Однако при этом существенно ограничивается возможность обнаружения новых ассоциаций и получения новых знаний об интересующем биологическом феномене. С другой стороны, перебор всей системы опорных множеств дает большую свободу и возможность нахождения новой информации, но по причине обычных ограничений вычислительных ресурсов такой подход требует применения особо эффективных и обоснованных эвристических подходов.

Функциональный анализ экспрессии генов

Экспрессия гена определяется количеством синтезированных копий молекул РНК с матрицы гена в клетке в определенных условиях и в определенный момент времени. На самом деле экспрессия гена не является константной величиной; это динамический процесс, в который вносит свой вклад целый ряд факторов, влияющих как на синтез, так и на деградацию РНК. Кроме того, экспрессия генов в основном измеряется не в единичной клетке, а в их популяциях, что, в свою очередь, вносит дополнительный элемент случайности (Kuznetsov et al., 2002). Исходя из вышеотмеченного, можно считать, что истинный уровень экспрессии гена является случайной переменной, выбранной из специфического для данной клеточной популяции распределения значений рассматриваемого параметра. Наконец, технические операции, необходимые для экспериментального измерения экспрессии генов, а также особенности детекции вносят дополнительный шум в конечный результат. Таким образом, измеренное значение экспрессии гена представляет собой стохастическую переменную, полученную из суммарного распределения объединенных профилей анализируемой характеристики и распределения шума. Все вышеперечисленное является причиной того, что идентификация дифференциально экспрессируемых генов, ассоциированных с исследуемым состоянием, на фоне «случайных» флуктуаций остального транскриптома становится нетривиальной задачей.

Предположим, что в нормальном состоянии экспрессия какого-либо гена является случайной переменной G_{norm} с распределением, равным суммарному распределению истинного значения экспрессии в клетке, B, и шуму измерительного прибора/процедуры измерения, N, так что $G_{norm} = B + N$. В случае сдвига (displacement) от нормального состояния (например, развитие болезни, прием лекарственных препаратов и т.д.) к распределению гена добавляется аналогичная переменная D (истинное распределение сдвига экспрессии гена в "ненормальном" состоянии), если данный ген связан со сдвигом, или 0, если ген не связан с ним: $G_{disp} = B + D + N$. Опираясь на опубликованные ранее исследования по анализу глобальной экспрессии генов, можно предположить, что число дифференциально экспрессируемых генов (т.е. связанных с исследуемым состоянием) составляет только малую долю генома (Lai, 2006). В этом случае в экспериментах по глобальному анализу транскриптома профили экспрессии дифференциально экспрессируемых генов нелегко отличить от большого числа независимо экспрессируемых генов, аналогичные характеристики которых могут быть независимо распределены и совпадать с профилем дифференциальной экспрессии.

В тоже время, уровни экспрессии независимых и связанных с исследуемым состоянием генов могут совпадать из-за перекрывания распределений значений экспрессии в нормальном состоянии, $G_{norm} = B + N$ и при сдвиге от нормы, $G_{disp} = B + D + N$. Это может привести как к ложноположительным (гены, не связанные с исследуемым состоянием, но их экспрессия превышает заданный порог), так и к ложноотрицательным (гены, которые ассоциированы с исследуемым состоянием, но их экспрессия ниже порогового значения) результатам.

Большинство существующих алгоритмов анализа экспрессии генов не справляется с вышеизложенной проблемой, выбирая дифференциально экспрессируемые гены на основе пороговых значений. Это, в свою очередь, зачастую приводит к частичной или полной несовместимости результатов, полученных в разных экспериментах (Chen et al., 2007).

Вернемся к матрице классификации M, определенной выше. По этой матрице рассмотрим не только отдельные гены и их экспрессию, но и комбинации генов с экспрессиями, т.е. профили экспрессий. Возможное число различных профилей экспрессии генов по таблице ограничено числом $2^{(d_1+d_2)}$, где d_1 и d_2 – мощности множеств объектов классов (в случае двух классов). В обычном высокопроизводительном эксперименте (микрочипы, широкогеномный поиск ассоциаций, полногеномное секвенирование) $2^{(d_1+d_2)} \ll t$, где t – число рассматриваемых генов. Таким образом, при отсутствии специфических ограничений значительное количество профилей экспрессии генов будет повторяться в разы, сопоставимые с t. В этом контексте, т.е. для данного профиля экспрессии ИК-ассоциированного гена, можно также ожидать наличие аналогичных профилей у большого числа ИК-независимых генов. Предполагается, что такая ситуация может возникнуть из-за очень грубой оцифровки, а также в случае, когда природа экспрессии генов не позволяет четко различать дифференциально экспрессируемые и инвариантные гены. Кроме того, многие алгоритмы анализа экспрессии генов учитывают не весь профиль, а некоторые из его интегральных характеристик, таких как среднее, дисперсия и частотное распределение, что увеличивает вероятность совпадений профилей экспрессии инвариантных и дифференциально экспрессируемых генов.

Наконец, значительное количество экспериментальных и теоретических расчетов показало, что общее количество транскриптов мРНК в отдельной клетке варьирует в пределах от 519688 до 851087 молекул мРНК при рассмотрении их синтеза с 8000 до 20000 генов. При этом уровень экспрессии для отдельно взятого гена может составлять 0.1-20000 копий мРНК (Kuznetsov et al., 2002). Однако, как уже было отмечено, эмпирические относительные частоты распределения экспрессии генов характеризуются сильным сдвигом влево, а значения уровней рассматриваемого параметра в основном находятся в диапазоне от нескольких единиц до нескольких сотен. При этом многие гены экспрессируются на одинаковых уровнях.

Таким образом, исходя из предпосылок, что: **1)** экспрессия генов исследуется в случайном наборе клеток, **2)** оценки экспрессии содержат стохастическую составляющую, сравнимую с величиной экспрессии большинства генов, **3)** для основных уровней экспрессии количество экспрессируемых генов сравнимо с общим числом генов, **4)** ИК-ассоциированные гены действуют независимо и одновременно, а эффективная разница экспрессии D становится смешанной и случайной, и **5)** применяя неравенство Чебышева к этой модели, можно сделать вывод, что основным профилям экспрессии ИК-ассоциированных генов, как и в случае отдельно взятых генов, может соответствовать большое количество аналогичных профилей, не связанных с ИК.

Обобщая вышеизложенное, необходимо отметить, что для решения проблемы корректной идентификации дифференциально экспрессируемых генов механистические алгоритмы, основанные на математических и статистических особенностях анализируемых данных, несостоятельны; они должны быть дополнены алгоритмами, основанными на использовании накопленных знаний о биологической природе исследуемого состояния и о связи генов и их экспрессии с ним. Только при таком подходе возможно осуществить переход от простого списка дифференциально экспрессируемых генов к системному пониманию фундаментальных молекулярных механизмов, лежащих в основе развития того или иного биологического процесса (функциональный анализ экспрессии генов). В настоящее время это достигается путем поиска и комбинации знаний

в литературных источниках и ряде публичных баз данных по данной проблематике. Сегодня такими источниками являются Gene Ontology – аннотация биологических процессов, клеточной локализации и молекулярной функции (Ashburner et al., 2000), KEGG и Reactome – биологические пути (Ogata et al., 1999; Kanehisa et al., 2010), Pfam – белковые домены (Finn et al., 2009), TRANSFAC – факторы транскрипции (Matys et al., 2006), BIND – белок-белковые взаимодействия (Bader & Hogue, 2000).

Алгоритм растущих опорных множеств

Опорное множество – это определенный набор признаков, который предоставляет классификатор необходимой точности, тогда как подмножества и надмножества этого множества ***менее*** и ***не более*** точны, соответственно.

При построении модели и разработке нашего алгоритма мы исходили из нескольких предпосылок: 1) если задана матрица экспрессии M, состоящая из d_1 объектов, принадлежащих к одному классу (ИК_1), и d_2 объектов, принадлежащих к другому классу (ИК_2), t экспрессий генов для каждого объекта, и наблюдается биологический процесс, который отвечает за фенотипические различия в классах, то, соответственно, возможно поставить и решить задачу алгоритмической классификации этих классов; 2) при наличии подобного биологического процесса предполагается, что признаки (гены), которые ассоциированы с ним, присутствуют в таблице M и они способны обеспечивать классификацию объектов в классах; 3) набор признаков имеет более высокую вероятность правильной классификации объектов, если известно, что его подмножества также подразделяют эти классы; 4) чем большее количество наборов признаков ассоциировано с одним и тем же биологическим процессом, тем проще и точнее можно идентифицировать сам процесс. Общая схема разработанного алгоритма представлена на рис. 2.

Даны: M - матрица экспрессий генов (строки ($g_i, i = 1,2,\dots,t$) - гены), столбцы ($o_j^1, j = 1,2,\dots d_1$) и ($o_j^2, j = 1,2,\dots d_2$) - образцы, *supp* - пороговое значение выбора опорных множеств.

Шаг 1

Цикл по строкам матрицы M, определение эмпирической ошибки классификации по множеству обучения, $E(h,\varpi)$, для каждого гена $\varpi = g_i, i = 1,2,\dots,t$.

Выбор генов с $E(h,\varpi) \geq supp$ в множество F_1 ($F_1 \equiv$ множество-аналог частых наборов)

Выбор генов с $E(h,\varpi) < supp$ в множество S_1 ($S_1 \equiv$ множество 1-опорных наборов)

Шаг 2

Для $(k = 1; k < n; k++)$

Пока $(F_{k-1} \neq \emptyset)$

Генерация k-«частых наборов» из F_{k-1}. При этом генерируются все k наборы, пересекающиеся с элементами F_{k-1}, проверяется все ли k-1-подмножества содержатся в F_{k-1}, и далее – что эти кандидатные подмножества имеют $E(h,\varpi) \geq supp$. Этим путем строится множество F_k.

Кандидаты с $E(h,\varpi) < supp$ включаются в множество S_k.

Шаг 3

Для всех генов в составе опорных множеств провести функциональный анализ по базам данных KEGG Pathways, Reactome и GeneOntology.

Рисунок 2. Псевдокод алгоритма растущих опорных множеств.

На первом шаге алгоритма определяются классификаторы, состоящие из одного ($k = 1$) признака (гена). Для этого необходимо пройтись по всем признакам в матрице M и подсчитать для них точность классификации. Классификаторы ранжируются по точности классификации и в дальнейшем рассматриваются только те из них, чья точность выше, чем установленный порог для кандидатов (например, >60%). Соответственно, классификаторы, прошедшие порог, называются "кандидатами". Для $k \geq 2$ классификаторы, подлежащие рассмотрению, формируются, если только его все $k-1$ подклассификаторы являются "кандидатами", а точность классификации данного классификатора больше, чем точность каждого из его $k-1$ подмножеств классификаторов. Состоящие из k признаков классификаторы, чья точность классификации превышает пороговое значение и которые не представляют собой часть другого $k+1$ классификатора, являются опорными множествами. Это, как уже отмечалось выше, просто насыщенные наборы генов, которые обеспечивают требуемую точность классификации. В экспериментальной части работы, на основе анализа литературы, в качестве порога точности классификации были выбраны следующие значения: 100% – если размер выборки не превышает 15, 90% – если ее объем равен 15-30, и 80% – если выборка содержит 30 и более объектов (Arakelyan et al., 2013). Ниже рассматриваются отдельные элементы алгоритма.

Классификаторы

В первую очередь необходимо определить элементарные классификаторы. Они представляют собой функции типа «признак-класс объекта», характеризующиеся определенной точностью классификации. Классификаторы первого уровня (1-классификатор) определяются одной строкой (значениями экспрессии одного гена). Классификаторы второго уровня (2-классификатор) представляют из себя комбинации двух строк (комбинации экспрессии двух генов). Классификатор n –го уровня (n – классификатор), соответственно, представляет собой комбинацию из n строк (комбинаций экспрессии n генов), при этом $n \leq (d_1 + d_2)$. Гены, связанные с классификаторами, называются опорными множествами.

В данном случае для классификации объектов использовался метод опорных векторов (МОВ, SVM). Данный подход особенно эффективен в ситуациях, когда малое количество классифицируемых объектов имеет высокую размерность признаков. МОВ был реализован методом линейного программирования, который позволяет найти оптимальную разделяющую гиперплоскость в случае линейной неразделимости классов. Задача нахождения опорных векторов формулируется следующим образом: для N точек $\{x_i, y_i\}$, где $i = 1,2. \ldots, N, x \in \Re^p$ и классов $y = \{-1,1\}$, опорные вектора должны удовлетворять следующим условиям (Sain & Vapnik, 2006):

$$\begin{cases} W(x_i) = w^T x_i + b \geq +1, if\ y_i = +1 \\ W(x_l) = w^T x_i + b \leq -1, if\ y_i = -1 \end{cases},$$

где $w^T x_i + b = 0$ – разделяющая гиперплоскость. При минимизации $w^T w$, граница разделения между классами возрастает. Знаковое расстояние $dist_i$ точки x_i до разделяющей гиперплоскости составляет $dist_i = W(x_i)\|w\|$.

Оценка ошибки/точности классификации

В случае линейной неразделимости классификатор может неправильно назначать членство новых объектов, что приводит к ошибочной классификации. Истинная ошибка классификации – это доля ошибок классификатора при его тестировании на истинном распределении объектов (Nolan, 1997). Однако, поскольку истинное распределение ошибки классификации, как правило, неизвестно, необходимо выработать правильную оценку истинной ошибки классификации. Недавно Braga-Neto и Dougherty (Braga-Neto &

Dougherty, 2004) предложили новый метод оценки ошибки классификации, называемый «усиленная оценка ошибки» (bolstered error estimate). Основная идея данного подхода заключается в обучении классификатора на всех объектах выборки и последующей генерации тестовой выборки исходя из класс-зависимой дисперсии. В простых случаях усиление осуществляется путем построения p-мерных сфер вокруг объектов выборки (p - число признаков), в пределах которых генерируется тестовая выборка (рис. 3).

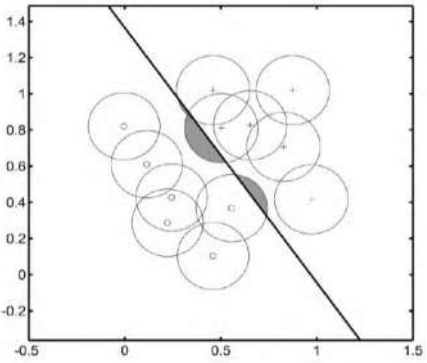

Рисунок 3. Принцип метода «усиленная оценка ошибки».
Точки (o, +), являющиеся центрами окружностей, представляют собой объекты исследуемой выборки. Окружности вокруг точек обозначают пространства, в пределах которых генерируется тестовая выборка. Серые сферические сегменты показывают вклад каждого объекта исследуемой выборки в формирование ошибки классификации. Общая ошибка вычисляется в виде средних значений индивидуальных ошибок.

Усиленная оценка ошибок сочетает в себе высокую вычислительную скорость метода подстановки ("resubstitution") и точность контроля по отдельным объектам ("leave-one-out").

В алгоритме опорных множеств мы использовали аналитическое решение, включающее расчет объема сферического сегмента, рассекаемого линейной гиперплоскостью. Радиус сферического ядра (spherical bolstering kernel) рассчитывался в соответствии с ранее предложенным методом и равен класс-зависимому стандартному отклонению, умноженному на поправочный коэффициент (значение обратного кумулятивного распределения χ^2со степенями свободы, равными p, в точке 0.5) (Braga-Neto & Dougherty, 2004).

Для расчета объема -мерного сферического сегмента была использована формула (Li, 2011):

$V_n^{cap}(n) = \int_{\theta-\varphi}^{\theta-0} V_{n-1}(r\sin\theta)d\cos\theta = \frac{1}{2}V_n(r)I_{sin^2\theta}\left(\frac{n+1}{2},\frac{1}{2}\right)$,

где $I_{sin^2\theta}\left(\frac{n+1}{2},\frac{1}{2}\right)$ - неполная бета функция, а $0 \le \varphi \le \pi/2$ - полярный угол (коширота).

Результаты симуляций показали, что предложенный вариант оценки ошибки классификации характеризуется точностью, сравнимой с методами перекрестной проверки, и скоростью подстановки (Arakelyan, et al. 2014).

Валидация алгоритма растущих опорных множеств

Для валидации разработанного алгоритма мы использовали набор данных по измерению экспрессий генов при остром лимфобластном лейкозе (ОЛЛ) и остром миелобластном лейкозе (ОМЛ) (Golub et al., 1999). Набор содержит данные по экспрессии 6871 гена в образцах костного мозга 27 и 11 больных ОЛЛ и ОМЛ, соответственно. Для поиска опорных множеств были установлены следующие параметры: порог точности классификации $e \le 0.15$, порог выбора кандидатных множеств $0.15 < e \le 0.25$. Классификация проводилась с помощью метода опорных векторов, а оценка точности классификации – методом «усиленной оценки ошибки». Значимость частоты встречаемости генов в опорных множествах, содержащих два и более генов, была оценена

на основе распределения рассматриваемого показателя при генерации случайных множеств того же размера.

Функциональный анализ генов в опорных множествах был проведен с помощью программных пакетов DAVID Bioinformatics Resources 6.8 (Huang et al., 2008), STRING 10.5 (von Mering et al., 2004) и Webgestalt (Zhang et al., 2005) с использованием ресурсов Gene Ontology и KEGG.

Проведенный поиск выявил 87 1-классификаторов (1 ген), 21,123 2-классификатора (2 гена), 108,397 3-классификаторов (3 гена) и 662 4-классификатора (4 гена).

Таблица 1.

Ассоциация генов, идентифицированных методом растущих опорных множеств, с лейкемиями

Описание функционального набора	Кол-во генов	P	FDR
Leukemia, Myeloid, Acute	30	0	0
Leukemia, T-Cell	22	6.01E-12	2.52E-09
Leukemia, Lymphoid	25	2.36E-11	6.85E-09
Lymphatic Diseases	28	4.79E-11	1.25E-08
Leukemia, Myelocytic, Acute	14	1.71E-07	0.000102
Acute Promyelocytic Leukemia	8	2.21E-07	0.000102

В первую очередь мы провели сравнение списка генов, связанных с 1-классификаторами (87 генов), со списком генов (53 гена) из публикации Golub и соавт. (Golub et al., 1999). Из 53 генов 33 были в списке 1-классификаторов. Точность остальных 20 генов классификации была ниже установленного порога.

В своей работе Golub и соавт. не провели функциональный анализ; вместо него был проведен выборочный анализ генов на основании ранее опубликованных данных. Нами, в свою очередь, был проведен анализ обогащенности, по категориям DisGenNet, Glad4U и OMIM, который подтвердил ассоциацию этих генов с развитием лейкемий. Проведенный нами анализ биологических процессов по категориям Gene Ontology (GO) показал, что дифференциально экспрессируемые гены, представленные в работе Golub и соавт. (1999), связаны с процессами активации и пролиферации лейкоцитов (в частности, лимфоцитов) и регуляции иммунного ответа. Однако попытка получить более подробные представления о патогенезе этих заболеваний на уровне репрезентативности биологических путей не выявила сколь-либо значимых ассоциаций. Таким образом, использованный авторами метод хорошо подходит для анализа и идентификации биомаркеров, но при этом не обладает достаточной мощностью для функционального анализа патогенеза заболеваний.

Таблица 2

Биологические пути, ассоциированные с патогенезом лейкемий

Название	Кол-во генов	P	FDR
Hematopoietic cell lineage	12	1E-05	0.003
Transcriptional misregulation in cancer	15	0.0001	0.009
Innate Immune System	73	2E-13	0.000
Cytokine Signaling in Immune system	38	6E-06	0.001
Antigen activates B Cell Receptor	6	0.0002	0.017
Platelet calcium homeostasis	5	0.0006	0.031
Apoptosis	12	0.0008	0.036
B Cell Receptor Signaling Pathway	12	0.0001	0.022
TSLP Signaling Pathway	7	0.0007	0.044
Hematopoietic Stem Cell Differentiation	7	0.0008	0.044

С другой стороны, предложенный нами метод опорных множеств позволяет обойти указанные недостатки. Для этого на начальном этапе был проведен анализ репрезентативности генов, связанных с классификаторами (опорными множествами). Были выбраны те гены, частота встречаемости которых значимо отличалась от таковой при генерации случайных множеств того же размера. Окончательный список содержал 320 генов. Функциональный анализ подтвердил ассоциацию идентифицированных генов с развитием лейкемий (таблица 1). Кроме того, список генов, полученных с помощью нашего алгоритма опорных множеств, позволил получить информацию о биологических процессах, лежащих в основе патогенеза этих заболеваний.

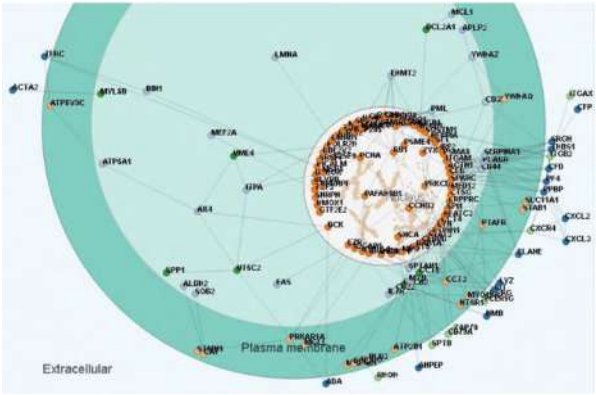

Рисунок 4. Пространственная ориентация белков в идентифицированной белок-белковой сети.

Анализ репрезентативности биологических путей в базах данных KEGG Pathways, Reactome, Wikipathways и Panter позволил выявить биологические пути, нарушения в которых связаны с этиопатогенезом лейкемий (таблица 2).

Наконец, используя информацию о белок-белковых взаимодействиях ресурса String-db (von Mering et al., 2004) и список генов, ассоциированных с опорными множествами, была реконструирована сеть белок-белковых взаимодействий. Пространственная ориентация белков в клетке показала, что полученная сеть объединяет белки-лиганды, рецепторы, внутриклеточные мессенджеры, нуклеарные белки и транскрипционные факторы, что позволяет классифицировать данную сеть как биологический путь (рис. 4).

РАЗРАБОТКА БИОИНФОРМАТИЧЕСКИХ МЕТОДОВ ДЛЯ МОДЕЛИРОВАНИЯ АКТИВНОСТИ БИОЛОГИЧЕСКИХ ПУТЕЙ

Биологические пути как функциональная единица жизнедеятельности клетки

Согласно определению, предложенному Национальным институтом исследования генома человека США (National Human Genome Research Institute), биологический путь – это "серия взаимодействий между молекулами в клетке, приводящих к синтезу определенного продукта или к изменению функционального состояния клетки". Биологическому пути свойственно направленное распространение информации (сигнальный поток) от принимающих узлов к эффекторным через промежуточные узлы. С функциональной точки зрения биологические пути подразделяются на три типа: метаболические, регуляторные и сигнальные. Биологические пути представляют собой одну из фундаментальных основ жизнедеятельности клетки и организма, и, поэтому, исследования, направленные на изучение их функционального состояния для выявления механизмов развития заболеваний, определения молекулярных мишеней для разработки лекарств, входят в разряд наиболее актуальных задач современной биоинформатики (Liu & Chance, 2013).

С появлением массово-параллельных методов транскриптомики, протеомики и метаболомики стала возможной глобальная оценка состояния биологических путей, т.е. переход от анализа на генном уровне к системной биологии, что открывает принципиально новые перспективы для понимания биологических процессов, протекающих в клетке в норме и патологии. Во-первых, группирование десятков тысяч генов, белков и/или других биомолекул по нескольким сотням биологических путей существенно снижает размерность данных и уменьшает сложность исследуемой биологической системы. Во-вторых, идентификация различий в активации путей гораздо лучше может объяснить фенотипические несовпадения в исследуемых состояниях, чем список дифференциально экспрессируемых генов (Glazko & Emmert-Streib, 2009). Эти соображения дали мощный толчок разработке множества математических алгоритмов и компьютерных программ "анализа биологических путей".

Проблему анализа биологических путей можно подразделить на две независимые задачи: 1) создание коллекций топологий биологических путей, которые будут максимально оптимизированы для последующего компьютерного анализа, и 2) разработка алгоритмов, способных моделировать динамические процессы в биологических путях.

Визуализация и редактирование биологических путей баз данных KEGG pathways

База данных KEGG Pathway представляет собой набор карт сигнальных, регуляторных и метаболических путей. Каждая карта создана вручную и представляет собой компиляцию текущих знаний о молекулярных взаимодействиях и метаболических реакциях (Kanehisa et al., 2010). В картах KEGG Pathway содержится дополнительная информация о биологических процессах, регуляция которых ассоциируется с конкретным путем. Для каждого биологического пути, представленного в базе данных, помимо графической карты, имеется также файл KGML (язык KEGG Markup), который предназначен для автоматизации компьютерного анализа и моделирования метаболических и сигнальных путей.

Биологический путь в KGML файле представлен в виде списка узлов (продуктов экспрессии генов (белков и нуклеиновых кислот), метаболитов и других биологических путей), а также перечня физических и химических взаимодействий между узлами. Однако формат KGML имеет ряд ограничений, которые значительно уменьшают возможность использования KGML файлов в автоматизированном анализе без предварительной обработки. Анализ несоответствий пяти случайно выбранных карт KEGG (Tight junction, Ubiquitin mediated proteolysis, Toll-like receptor signaling, Autoimmune thyroid disease, и Homologous recombination) показал, что сопутствующие файлы KGML, по сравнению с графической картой, содержат в среднем 13, 26, 3 и 10 несоответствий, касающихся текстовых меток, отсутствия или неправильного направления взаимодействий и аннотации групповых узлов, соответственно. Подобные несоответствия могут оказывать значительное искажающее воздействие на сигнальные потоки в путях, что, в свою очередь, может привести к получению неточных результатов и к их неправильной интерпретации. Таким образом, перед использованием KGML файлов в автоматическом анализе необходима предварительная обработка информации, содержащейся в файле KGML.

Однако даже с учетом этих недостатков база данных KEGG Pathway по-прежнему является ценным и наиболее используемым ресурсом в системной биологии и геномике. В связи с этим интенсивно разрабатываются синтаксические анализаторы, способные преобразовывать информацию KGML файлов в различные типы объектов, удобные для последующего анализа.

Разработка программного пакета KEGGParser для среды MATLAB и SciLab

Для решения вышеперечисленных проблем, связанных с автоматизацией анализа и визуализации, мы создали программу KEGGParser (полуавтоматический синтаксический анализатор / редактор путей) для среды MATLAB. KEGGParser использует функции языка MATLAB и его дополнительных пакетов Bioinformatics toolbox версии 3.x и Image processing toolbox 2.x и не требует установки сторонних приложений. Кроме того, мы также портировали KEGGParser в Scilab 5.4 – бесплатную альтернативу MATLAB с открытым исходным кодом. KEGGParser для MATLAB и Scilab находятся в свободном доступе (http://www.mathworks.com/matlabcentral/fileexchange/37561-keggparser--parsing-and-editing-kegg-pathway-maps-in-matlab).

Рабочий процесс KEGGParser включает в себя пять основных этапов: 1) загрузка копии KGML файла, описывающего биологический путь; 2) трансляция данных KGML формата в объект графа; 3) редактирование узлов и ребер графа; 4) визуализация; 5) сохранение пути в виде графа.

KEGGParser предоставляет три способа загрузки карт в формате KGML. После загрузки файла с картой биологического пути KGML подвергается предварительному синтаксическому анализу, в результате чего создается объект графа. На этом этапе различные подтипы взаимодействий, определенные в KGML, обобщаются в три основных типа, в зависимости от их эффекта: активация, ингибирование и связывание. Последний этап используется, когда эффект явно не упоминается в описании взаимодействия. Кроме того, на этом этапе KEGGParser имеет встроенный функционал для автоматической коррекции несоответствий в KGML файле. Программа может автоматически исправлять несоответствия, связанные с белок-метаболит-белковыми взаимодействиями, «групповыми» узлами, ориентацией взаимодействия типа «binding» (рис. 5).

После предварительной автоматической обработки следующие операции могут выполняться вручную: 1) добавление текстовых меток; 2) изменение направления ребра взаимодействия; 3) добавление или удаление ребер и узлов.

Визуализация путей, созданных KEGGParser, максимально приближена к статическим изображениям в базе данных KEGG pathway. При этом сохраняются размеры узлов и их относительные позиции. Поскольку биологический путь представлен в виде графа, можно использовать широкий спектр алгоритмов теории графов, поддерживаемых MATLAB.

Необходимость предварительного редактирования карт биологических путей демонстрирует нижеследующий пример по редактированию биологического пути «RIG-I-like receptor signaling pathway». Рецепторы и внутреклеточные сигнальные каскады для RIG-I-like генов являются ключевыми элементами в распознавании вирусных патогенов и инициации ответа врожденной иммунной системы против них (Kaneda, 2013). Нарушение регуляции данного пути связано со многими

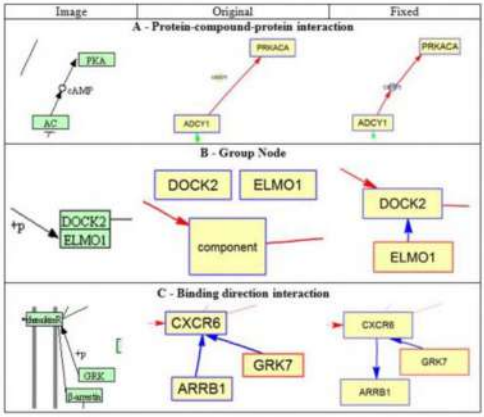

Рисунок 5. Примеры автоматической коррекции несоответствий между изображением карты биологического пути и его описания в KGML формате с помощью программы KEGGParser.

аутоиммунными заболеваниями, такими как системная красная волчанка и синдром Айкарди-Гутьера (Kato & Fujita, 2015). Чтобы проанализировать основные параметры топологии данного биологического пути, мы использовали KEGGParser для синтаксического анализа и редактирования соответствующего файла KGML. Карта пути, а также неотредактированные и отредактированные графы представлены на рис. 6-8, соответственно. Как можно заметить из рис. 7 и 8, анализируемый граф, созданный на основе синтаксического разбора соответствующего KGML файла, существенно отличается от графического изображения карты и в таком виде не пригоден для анализа топологических характеристик.

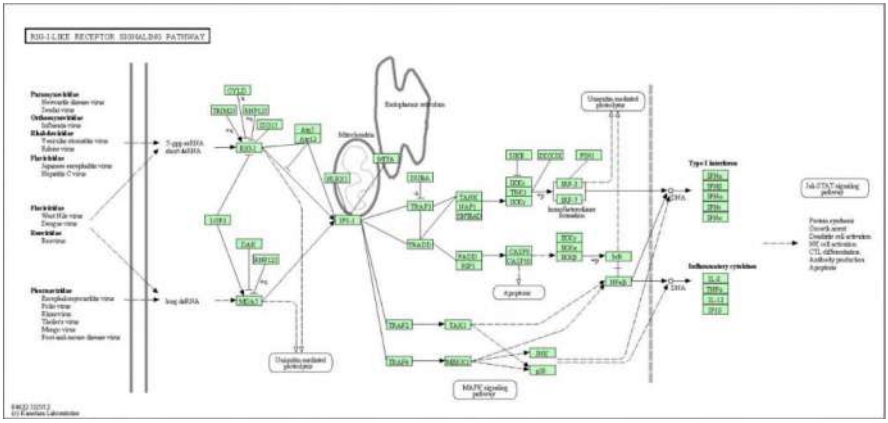

Рисунок 6. Графическое изображение биологического пути «RIG-I-like receptor signaling pathway» из базы данных KEGG Pathways.

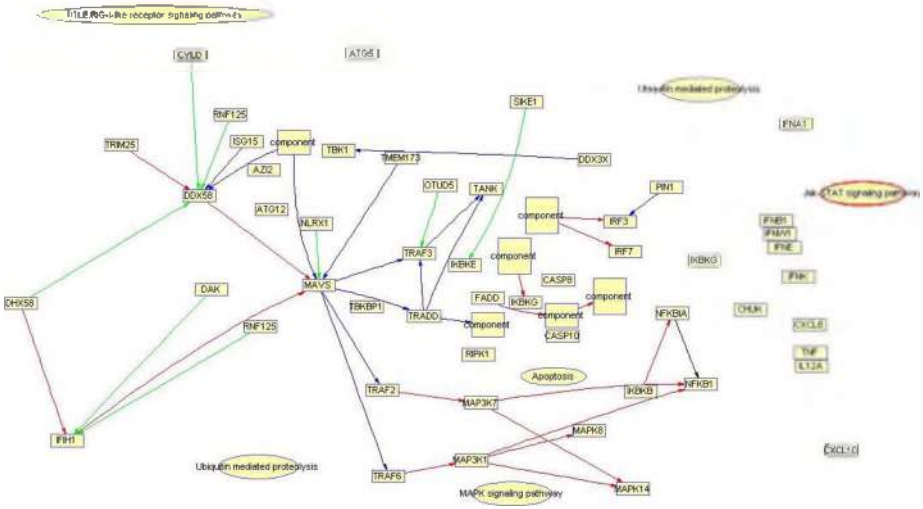

Рисунок 7. Объект графа, полученный в KEGGParser при синтаксической трансляции файла KGML для «RIG-I-like receptor signaling pathway».

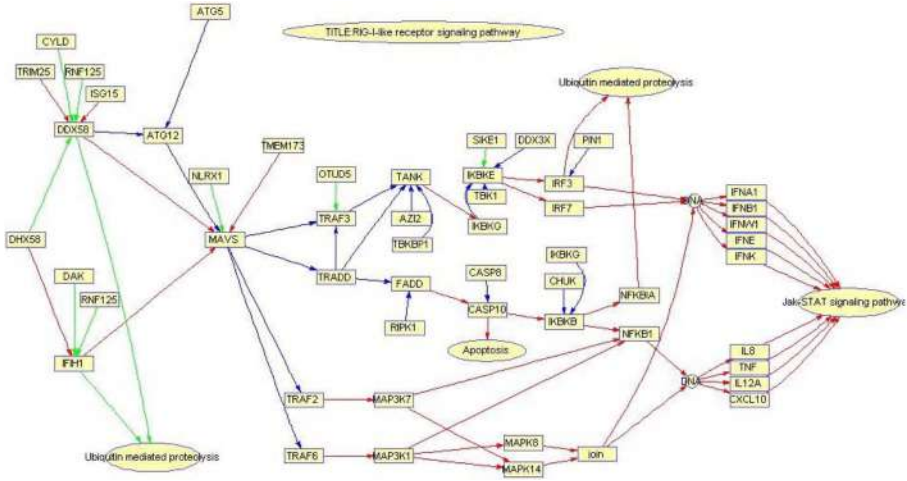

Рисунок 8. Отредактированный в KEGGParser граф биологического пути «RIG-I-like receptor signaling pathway».

Это связано с тем, что файл KGML содержит информацию только о белок-белковых взаимодействиях, а сведения о других типах взаимодействий, присутствующих в изображениях карты, теряются во время трансляции файла. В среде KEGGParser мы вручную отредактировали полученный граф, восстановили всю необходимую информацию (рис. 9) и провели анализ, используя функции теории графов, реализованные в наборе пакета Bioinformatics для MATLAB. Анализ гистограмм распределения валентности узлов на неотредактированном графе показал значительный перекос к узлам с нулевой валентностью по сравнению с отредактированным графом (рис. 9А, Б). Затем мы провели оценку сильно связанных компонентов в графах. Результаты показали, что в оригинальном графе имеются пять многоузловых сильно связанных компонентов (в среднем содержащих четыре узла), а в отредактированном графе идентифицируются шесть компонентов, содержащих в среднем семь узлов (рис. 9В, Г). Эти компоненты соответствуют узлам, которые образуют разные кластеры или ветви в биологическом пути, связанные с различными функциональными процессами, которые ассоциированы с активацией этого пути. Наконец, распределение длин всех пар кратчайших путей в оригинальном и отредактированном графах показало значительные отличия; при этом средняя длина была больше после редактирования, что согласуется с графическим изображением карты пути (рис. 9Д, Е). Таким образом, использование KEGGParser приводит к лучшему компьютерному представлению биологических путей в MATLAB и способствует адекватному анализу топологий биологических путей.

Разработка програмного пакета CyKEGGParser для среды Cytoscape

CyKEGGParser представляет собой расширенную версию KEGGParser, доступную в виде приложения для Cytoscape (Nersisyan et al., 2014). Программный код для CyKEGGParser был написан на Java.

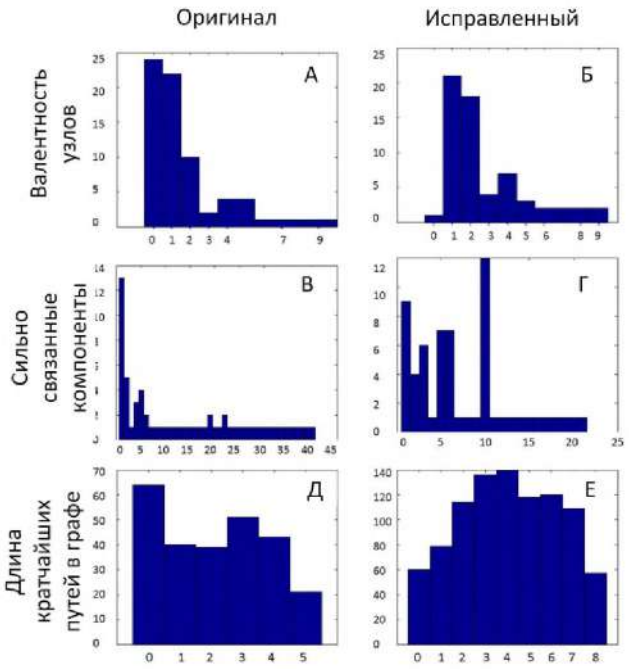

Рисунок 9. Анализ характеристик топологии оригинального и отредактированного графа биологического пути «RIG-I-like receptor signaling pathway».

Программа поддерживает все функции, реализованные в KEGGParser, кроме того, она содержит новые функции, которые существенно расширили возможность редактирования и визуализации биологических путей и их использования в автоматическом анализе (рис. 10). CyKEGGParser доступен в качестве приложения для Cytoscape 3 (http://apps.cytoscape.org/apps/cykeggparser). По состоянию на 01.06.2017 CyKEGGParser был загружен 14226 раз.

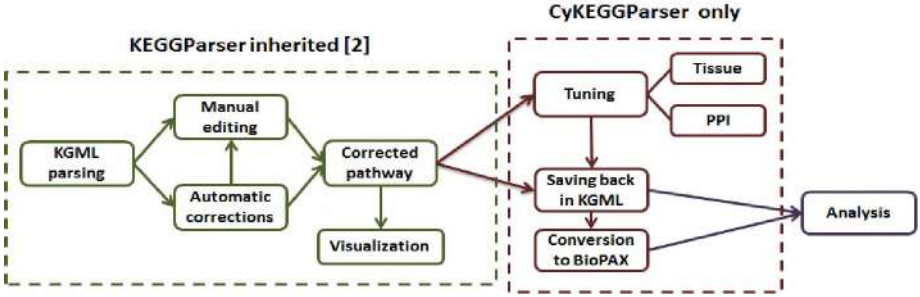

Рисунок 10. Дизайн программы CyKEGGParser. Зеленым отмечен функционал, унаследованный от KEGGParser, красным - новые возможности, реализованные в CyKEGGParser.

Основным преимуществом CyKEGGParser, по сравнению с KEGGParser, является возможность проведения специального типа редактирования по тканевой специфичности и белок-белковым взаимодействиям.

Для редактирования по тканевой специфичности CyKEGGParser использует наборы данных по экспрессии генов для нормальных тканей и раковых клеток BioGPS (http://biogps.org/), доступных в базе данных GeneCards (www.genecards.org). Кроме того, пользователь может создать собственный набор данных и проанализировать его. Для перестройки пути на основе белок-белковых взаимодействий (ББВ) CyKEGGParser использует базу данных String (http://string-db.org/).

Таким образом, в рамках данного раздела исследований были разработаны приложения KEGGParser и CyKEGGParser, которые позволяют загружать, корректировать, визуализировать и настраивать биологические сети, представленные в базе данных KEGG Pathways. Хотя подобные программы были разработаны и другими научными группами, полуавтоматическая коррекция и разные возможности по более гибкой настройке являются уникальными особенностями KEGGParser и CyKEGGParser. С помощью этих программ нами были подготовлены библиотеки сигнальных, метаболических и регуляторных путей KEGG Pathways, которые затем использовались для биоинформатического анализа изменений активностей сигнальных и метаболических каскадов при различных патологических состояниях.

Разработка алгоритма оценки сигнальных потоков Pathway Signal Flow

Для оценки изменения активности биологических путей нами был разработан алгоритм оценки сигнальных потоков в сетях (Pathway Signal Flow) (Nersisyan, et al., 2015; Arakelyan et al., 2016). Исходный алгоритм доступен в виде скриптовых пакетов для R и MATLAB.

Сигнальный поток в биологическом пути (pathway signal flow, PSF), или пертурбация, – это поток, генерируемый при распространении сигнала, начиная с входных узлов, проходя через промежуточные узлы в ветвях и накапливаясь на выходных узлах. Таким образом, PSF может быть индикатором активности пути. Более того, PSF дает возможность оценивать функциональный статус отдельных ответвлений пути, что невозможно осуществить с помощью описанных выше топологических методов.

Алгоритм оценивает распространение сигнала от входных узлов к выходным узлам биологического пути в зависимости от значений дифференциальной экспрессии генов (FC) в узлах пути и типов взаимодействий между ними (топологии биологического пути).

Значение FC для генов присваиваются соответствующим узлам; если узел содержит несколько генов, значения FC усредняются.

Взаимодействия между узлами биологического пути обобщаются в два основных типа в зависимости от их действия: активация/связывание и ингибирование. Поток сигналов между двумя соединенными узлами сети рассчитывается как:

$\boldsymbol{FC}_{(\text{узел 1})} * \boldsymbol{FC}_{(\text{узел 2})}$, если узел 1 активирует узел 2 и $\frac{1}{\boldsymbol{FC}_{(\text{узел 1})}} * \boldsymbol{FC}_{(\text{узел 2})}$, если узел 1 ингибирует узел 2.

После этого входной сигнал равный 1 задается на входные узлы исследуемого пути, а оценка состояния исследуемого пути проводится по значению сигнального потока на выходных узлах. Алгоритм PSF откалиброван таким образом, что экспрессия гена FC = 1 во всех узлах (нормальная экспрессия гена) дает значения PSF = 1. Значения PSF < 1 свидетельствуют о понижении, тогда как значения PSF > 1 – о повышении активности биологического пути.

Оценка значимости изменений сигнального потока в биологическом пути проводится методом пермутации значений экспрессии в узлах и построения распределения случайных значений изменений сигнального потока. В качестве источника топологий биологических

путей используется библиотека, созданная на основе базы данных KEGG Pathway с помощью программ KEGGParser и CyKEGGParser (Arakelyan & Nersisyan, 2013; Nersisyan et al., 2014).

Разработка приложение Pathway Signal Flow Calculator для Cytoscape

Нами было разработано приложение Pathway Signal Flow Calculator (PSFC) для Cytoscape с целью анализа распространения сигнального потока в биологическом пути на основе данных дифференциальной экспрессии генов и топологии биологических сетей (Nersisyan et al., 2014). В PSFC встроены различные варианты распространения сигнала, которые используются в известных алгоритмах анализа потока сигналов, а также предоставлена возможность создания новых правил. Таким образом, программа позволяет экспериментировать с результатами, полученными различными (существующими и настраиваемыми) подходами в рамках единого приложения, и оценивать соответствие различных моделей взаимодействия к реальным условиям.

Программа PSFC реализована в Java и доступна как приложение для Cytoscape 3 (http://apps.cytoscape.org/apps/psfc). Алгоритм PSFC состоит из нескольких шагов:

1. Сортировка графа. Эта операция необходима для правильного расположения узлов графа, что достигается присвоением топологических уровней узлам в биологическом пути для распространения сигнала от узлов более низкого к более высоким уровням. Биологические сети часто содержат петли обратной связи, которые создают циклы в графах. На первом этапе PSFC выполняет поиск в глубину и удаляет обратные ребра в графе, затем проводит топологическую сортировку на полученном ациклическом графе, после чего удаленные ребра восстанавливаются.

2. Расчет распространения сигнального потока. В биологических путях сигнал распространяется через взаимодействия между парами узлов. В результате этого определенный уровень сигнала (значение PSF) накапливается в каждом из узлов пути. На рис. 11 показано распространение сигнала в зависимости от применения различных правил распространения сигнала. В качестве основного правила PSFC использует алгоритм PSF, описанный выше.

3. Правила для взаимодействия между узлами. В контексте анализа распространения сигнала каждому ребру в графе присваивается функция передачи сигнала, которая зависит от типа взаимодействия. PSFC позволяет пользователю определять типы взаимодействий для каждого ребра и присваивать конкретные математические функции любой сложности. Функция передачи сигнала должна иметь вид f(сигнал исходного узла, экспрессия конечного узла). На рис. 11 представлен результат присвоения различных функций.

4. Правила для множественных входящих и исходящих сигналов. Как правило, интенсивность взаимодействий между молекулами во многом зависит от их концентрации и состояния активации. Однако если узел в биологическом пути имеет несколько взаимодействующих партнеров, они могут конкурировать друг с другом, а "емкость" взаимодействия узла может быть «разделена» между этими партнерами.

Таким образом, существует возможность пропорционального разделения сигнала между несколькими ребрами графа, начиная с одного исходного узла или заканчивая на конечном узле. Сигналы на нескольких ребрах, заканчивающихся на одном узле, могут трактоваться одним из следующих трех способов: они могут вычисляться отдельно на каждом ребре, а затем либо суммироваться, либо умножаться, или итеративно обрабатываться путем обновления сигнала на конечном узле. Очередность, в которой ребра взаимодействий обрабатываются в последнем случае, может быть скорректирована с помощью манипуляций с порядком узлов (рис. 11).

5. Правила обработки петель обратной связи в биологических путях. Петли отрицательной и положительной обратной связи в биологических путях имеют первостепенное значение для тонкой регуляции их функциональной активности. PSFC предоставляет несколько вариантов обработки циклов: 1) Игнорировать петли обратной связи: в этом случае циклические обратные ребра игнорируются во время вычислений PSF; 2) Предварительное вычисление сигналов в циклах: в этом режиме алгоритм сначала находит ребра обратной связи, вычисляет их сигналы и обновляет их значения на конечном узле. Затем алгоритм работает на всем графе в режиме «игнорировать петли обратной связи»; 3) Итерация до схождения: алгоритм работает в течение нескольких итераций, пока не будет достигнуто схождение значений потока сигнала. Конвергенция достигается, если процент изменения сигнала между двумя итерациями меньше заданного порога сходимости во всех узлах. Если конвергенция не достигается, алгоритм останавливается после определенного количества итераций.

6. Расчет значимости изменений распространения сигнала в биологических путях. Значение значимости потоков сигналов на каждом узле вычисляется с использованием метода бутстрэппинга. Пользователь может выбирать между бутстрэпом образцов или генов. В первом случае значения экспрессии узлов выбираются случайно – методом перестановки. В случае бутстрэппинга по генам, показатель экспрессии каждого узла выбирается случайным образом, например, из матрицы экспрессии.

С помощью программы PSFC было оценено изменение сигнального потока в пути MAPK на основе опубликованных экспериментальных данных (Nelander et al., 2008; Feiglin et al., 2012). Согласно результатам вычислений, во всех случаях понижения экспрессии генов в сигнальном пути наблюдалось существенное повышение сигнала на узлах «арест клеток в G1-фазе» и «апоптоз», что полностью соответствует результатам, полученным Feiglin и соавт. (Feiglin et al., 2012).

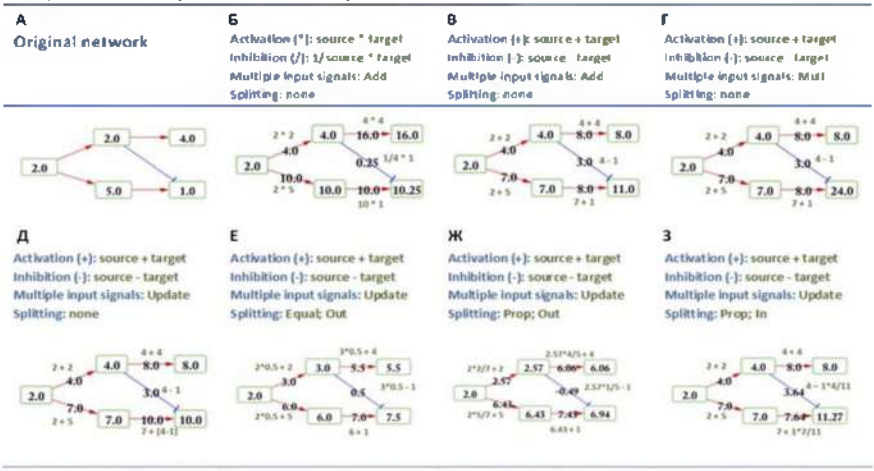

Рисунок 11. Вычисление сигнала PSF в зависимости от правил распространения потока. Синие и красные линии указывают на активацию и ингибирование. Правила объединения множественных входящих и исходящих сигналов вычисляются по сложению (Add), умножению (Mult) или обновлению значений сигнала на узлах (Update). Разделение сигнала с множественных узлов устанавливается как «none» (без разделения) , либо «equal» (равноценное разделение) или «prop» (пропорциональное), и выполняется либо на нескольких исходящих (Out), либо на нескольких входящих краях (In).

ИССЛЕДОВАНИЕ МОЛЕКУЛЯРНЫХ МЕХАНИЗМОВ ПАТОГЕНЕЗА КОМПЛЕКСНЫХ ЗАБОЛЕВАНИЙ ЧЕЛОВЕКА

Исследование молекулярных механизмов развития и течения заболеваний является необходимым условием для разработки эффективных методов и средств прогностики, диагностики и лечения. При этом ключевым фактором, определяющим патогенез заболевания, являются динамические изменения в уровне биомолекул и активности объединяющих их биологических путей и каскадов. Появление высокопроизводительных методов "-омики", таких как транскриптомика, протеомика и т.д., позволили накопить огромные массивы данных по уровням "-омов" при различных заболеваниях, что создает уникальную возможность для исследования их патомеханизмов.

В данном разделе представлены результаты наших исследований молекулярных механизмов широкого спектра моно- и полигенных заболеваний с помощью разработанных биоинформатических алгоритмов и программных пакетов, использованных для анализа доступных наборов данных по глобальной экспрессии генов.

Оценка влияния топологии биологических путей на результаты функционального анализа глобальной экспрессии генов

В настоящее время анализ обогащенности функциональными группами (gene set enrichment analysis, GSEA) является основным методом функциональной аннотации в экспериментах по экспрессии генов (Subramanian et al., 2005; Hung, 2013). При исследовании экспрессии генов или других геномных данных GSEA использует информацию о функциональных группах генов, связанных с определенным биологическим процессом или функцией. Однако GSEA не учитывает информацию о взаимодействиях между генами/белками в пределах функционального набора, что зачастую приводит к искаженной интерпретации результатов исследований. Поэтому методы функциональной аннотации, принимающие во внимание эти особенности, должны быть более аккуратными, особенно в случаях, когда функциональными группами являются биологические пути.

В данном исследовании мы провели сравнительный анализ эффективности алгоритмов PSF и GSEA для функциональной аннотации биологических путей по результатам исследования глобальной экспрессии генов.

Использованные наборы данных и алгоритмы

Были использованы наборы данных из хранилища Gene Expression Omnibus (Services, 2007; Barrett et al., 2011; NCBI Resource Coordinators, 2016), содержащие профили экспрессии генов в различных тканях при псориазе (GSE13355), рассеянном склерозе (GSE13732) и кардиоэмболическом ишемическом инсульте (GSE58294).

Для данного исследования были выбраны 24 биологических пути из базы данных KEGG (B cell receptor signaling pathway, Hedgehog signaling pathway, Calcium signaling pathway, HIF-1 signaling pathway, Chemokine signaling pathway, Jak-STAT signaling pathway, ErbB signaling pathway, MAPK signaling pathway, Fc epsilon RI signaling pathway, mTOR signaling pathway, Fc gamma R-mediated phagocytosis, Natural killer cell mediated cytotoxicity, FoxO signaling pathway, Notch signaling pathway, NOD-like receptor signaling pathway, PI3K-Akt signaling pathway, TNF signaling pathway, Rap1 signaling pathway, Ras signaling pathway, TGF-beta signaling pathway, RIG-I-like receptor signaling pathway, Toll-like receptor signaling pathway, T cell receptor signaling pathway, VEGF signaling pathway), которые, согласно литературным данным, вовлечены в развитие указанных заболеваний. GSEA был проведен с использованием классического алгоритма, реализованного в Java

приложении GSEA (Subramanian et al., 2005), и значений параметров, установленных по умолчанию.

Оценка активности биологических путей была осуществлена с применением алгоритма PSF. Значения экспрессии генов предварительно были трансформированы в кратные изменения (fold change, FC) по отношению к средним значениям экспрессии генов в контрольной группе (для каждого набора данных по отдельности). В качестве источника топологии биологических путей использована библиотека, созданная нами на основе базы данных KEGG Pathway с помощью программ KEGGParser и CyKEGGParser. Значимость изменения активности сигнального потока в биологическом пути рассчитывалась с применением процедуры бутстрапа (200 циклов).

Анализ ассоциации изменений активности биологических путей при НИЗ

Анализ активности биологических путей при псориазе. Псориаз является воспалительным гиперпролиферативным заболеванием кожи и суставов, с выраженной генетической компонентой (Deng et al., 2016). Набор данных GSE13355 содержал log2-трансформированные профили экспрессии генов (чип - Affymetrix HG-133 2 Plus), полученных из образцов биопсии кожного покрова 58 пациентов и 64 здоровых лиц (Nair et al., 2009). Мы ожидали, что методы функционального анализа позволят идентифицировать изменения в биологических путях, связанных с регуляцией иммунного ответа. Согласно полученным результатам, алгоритм PSF выявил значимые нарушения активности всех 24 путей, при этом GSEA смог идентифицировать только 5 из них (рис. 12).

Анализ активности биологических путей при рассеянном склерозе. Рассеянный склероз является органоспецифическим аутоиммунным заболеванием, вызванным воспалительной демиелинизацией нервных волокон центральной нервной системы (Corvol et al., 2008). Методами PSF и GSEA, была проведена оценка дерегуляции биологических путей с использованием набора данных GSE13732, который содержал профили экспрессии генов в нативных CD4 + Т-клетках больных рассеянным склерозом и здоровых лиц (платформа Affymetrix HG 133 2 Plus) (Corvol et al., 2008). Анализ GSEA у пациентов выявил изменения в сигнальном пути TGF-β по сравнению с контролем. Напротив, PSF идентифицировал значимые изменения в 20 путях (рис. 13), в том числе в сигнальных путях В- и Т-клеток, которые играют ключевую роль в патогенезе этого заболевания (Holley et al., 2014; Blauth et al., 2015).

Анализ активности биологических путей при ишемическом инсульте. Ишемический инсульт составляет около 80% всех случаев инсульта мозга и примерно 70% всех острых цереброваскулярных заболеваний. При этом, почти четверть случаев инсульта регистрируется у людей трудоспособного возраста (Di Napoli et al., 2006), и только 1/3 всех пациентов с инсультом достигают полной социальной и профессиональной реинтеграции. Острый локальный и системный воспалительный ответ, наряду с активацией апоптоза, является важным фактором патогенеза инсульта и предиктором течения заболевания (Di Napoli et al., 2006). В данном случае были проанализированы профили экспрессии генов в цельной крови 69 пациентов с ишемическим инсультом и 23 здоровых лиц (Stamova et al., 2014). И вновь PSF позволил выявить изменения активности в 22 сигнальных путях, регулирующих иммунный и воспалительный ответ, в то время как GSEA обнаружил значимые изменения только в четырех путях (рис. 14).

Оценка роли топологии при идентификации нарушений активности биологических путей. Основным недостатком GSEA является его неспособность правильно оценить влияние взаимодействий между белками или генами в биологическом пути. Статистика GSEA основана на ранжировании экспрессии генов, в то время как функциональный эффект изменения экспрессии просто игнорируется.

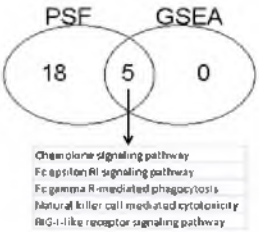

Рисунок 12. Сравнение результатов анализа PSF и GSEA при псориазе.

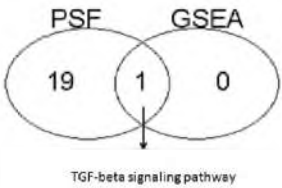

Рисунок 13. Сравнение результатов анализа PSF и GSEA при рассеянном склерозе.

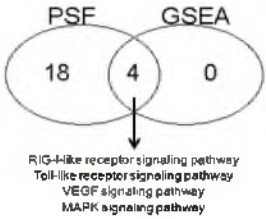

Рисунок 14. Сравнение результатов анализа PSF и GSEA при ишемическом инсульте.

В качестве примера мы проанализировали дерегуляцию сигнального пути TGF при рассеянном склерозе. Метод GSEA не идентифицировал значимого различия активности этого пути у пациентов по сравнению с контролем (FDR = 0.17, рис. 15A). Такой результат получился из-за почти равного количества генов, чья экспрессия была либо снижена, либо повышена по сравнению с нормой (рис. 15B). Дальнейшая проверка показала, что большинство генов с пониженной экспрессией являются ингибиторами эффекторных генов, проводящих сигнал от рецепторов к транскрипционным факторам и ДНК.

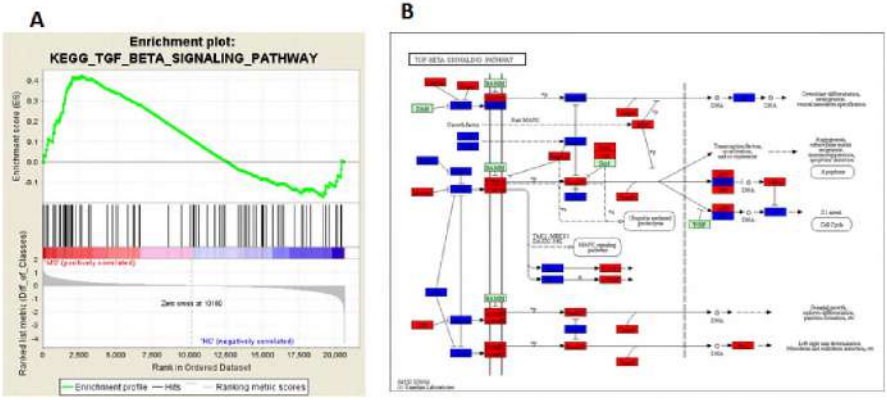

Рисунок 15. Результаты анализа GSEA по активации TGF-beta signaling pathway.

Таким образом, снижение их экспрессии должно привести к увеличению интенсивности сигнала, поступающего на выходные узлы и, соответственно, общей активации этого пути. С другой стороны, в алгоритме PSF учтена топология пути, что позволяет точно оценивать эффекты изменения уровней экспрессии на результирующую активность пути. Кроме того, с применением PSF возможно оценить дерегуляцию в каждой отдельной ветви в разветвленных путях. Так, результаты анализа PSF показали, что активность всех семи ответвлений сигнального пути TGF значимо меняется при инсульте по сравнению с нормой (*ответвление ROCK1: PSF = 1.36, p = 0.045; ответвление RPS6KB1: PSF = 1.47, p = 0.0065; ответвление ID1: PSF = 1.51, p > 10E-6; ответвление SMAD4: PSF = 1.59, p > 10E-6; ответвление SMAD4: PSF = 1.90, p > 10E-6; ответвление Cell cycle: PSF = 0.60, p > 10E-6; ответвление Apoptosis: PSF = 1.73, p > 10E-6*).

Таким образом, данное исследование продемонстрировало, что хотя GSEA нужно применять для работы с наборами генов (Subramanian et al., 2005), при функциональном анализе биологических путей следует использовать методы и алгоритмы (например, PSF), учитывающие особенности топологии и характер взаимодействий между генами или белками.

Анализ топологической резистентности биологических путей к мутациям
Состояние проблемы и цель исследования

Изучение влияния экспрессии генов на изменение активности биологических путей в контексте этиопатогенеза заболеваний, дизайна лекарств и нормальной физиологии организма является одним из наиболее перспективных направлений современной молекулярной биологии, чему во многом способствовало развитие технологий высокопроизводительных транскриптомных (микрочипы, секвенирование РНК) и протеомных (масс-спектрометрия) исследований, а также соответствующих биоинформатических подходов (Arakelyan et al., 2013; 2016). Между тем, влияние мутаций на взаимодействие между белками, или другими биомолекулами, т.е. изменение топологии биологических путей, изучено гораздо хуже, хотя их значимость была продемонстрирована в многочисленных экспериментальных исследованиях.

В данном исследовании была поставлена задача оценить устойчивость сигнальных путей к изменениям топологии, вызванными мутациями.

Использованные наборы данных и алгоритмы

Топологии биологических путей в виде ориентированных графов были получены с помощью пакета "graphite" для языка программирования R, который содержит карты путей из четырех различных баз данных, включая KEGG Pathways (Sales et al., 2012). В пакете "graphite" реализованы специальные правила для генерации белок-белок (или другой лиганд) взаимодействий, а также для максимально корректного восстановления сигнальных потоков в биологическом пути.

Для данного исследования были выбраны 11 сигнальных путей из базы данных KEGG Pathway («B cell receptor signaling pathway»(id: hsa04662), «Chemokine signaling pathway» (id: hsa04062), «ErbB signaling pathway» (id: hsa04012), «NOD-like receptor signaling pathway» (id: hsa04621), «Notch signaling pathway» (id: hsa04330), «T cell receptor signaling pathway» (id: hsa04660), «TGF-beta signaling pathway» (id: hsa04350), «TNF signaling pathway» (id: hsa04668), «Toll-like receptor signaling pathway» (id: hsa04620), «VEGF signaling pathway» (id: hsa04370), «Wnt signaling pathway» (id: hsa04310)). Все эти пути характеризуются комплексной топологией и содержат несколько, часто пересекающихся, ответвлений, которые могут передавать сигнал от входных к выходным узлам, даже если передача сигнала в каком-либо ответвлении нарушена.

Кроме того, для проверки зависимости степени нарушения передачи сигнала от меры сложности топологии пути были также созданы «синтетические» биологические пути на основе случайных графов с возрастающей связностью узлов (graph node connectivity), начиная с линейной топологии. Моделирование мутаций в биологических путях проводилось поэтапно. Для каждой пары соединенных узлов в графе: а) ребро графа было удалено для моделирования эффекта мутации, приводящей к полному нарушению взаимодействия между белками; б) функциональный атрибут взаимодействия был изменен на противоположный (активация была изменена на ингибирование, ингибирование – на активацию) для моделирования мутаций, приводящих к изменению эффекта взаимодействия. Таким образом, для каждого из выбранных путей были созданы топологии, учитывающие все возможные эффекты мутаций.

Анализ возмущений сигнальных потоков, вызванных мутациями, был проведен с использованием алгоритма PSF (Nersisyan et al., 2014; 2015). В данном исследовании значения экспрессии для всех генов были приняты равными единице. После этого были рассчитаны значения сигнальных потоков на выходных узлах для «референсных» (без изменения топологии) и «мутированных» топологий для каждого пути. Степень изменения уровня передачи сигнала, связанного с мутацией, затем была рассчитана относительно референсных значений сигнальных потоков.

Оценка устойчивости сигнальных путей к изменениям топологии, вызванным мутациями

Нами проведена *in silico* оценка изменений активности биологических путей, вызванных связанных с мутациями нарушениями в топологии. Мы проанализировали 12 сигнальных путей, которые, как было показано ранее, играют важную роль в физиологических процессах как в норме, так и в патологии (Newton & Dixit, 2012; Ito et al., 2014; Buch, 2014). Полученные результаты показали, что во всех случаях изменение взаимодействия для любого гена вызывает отклонения от исходного состояния активности, за исключением четырех генов в сигнальных путях B- и T-клеток. Распределение изменений активностей в пути было смещено вправо, с более высокой частотой для низких значений возмущений (рис. 16).

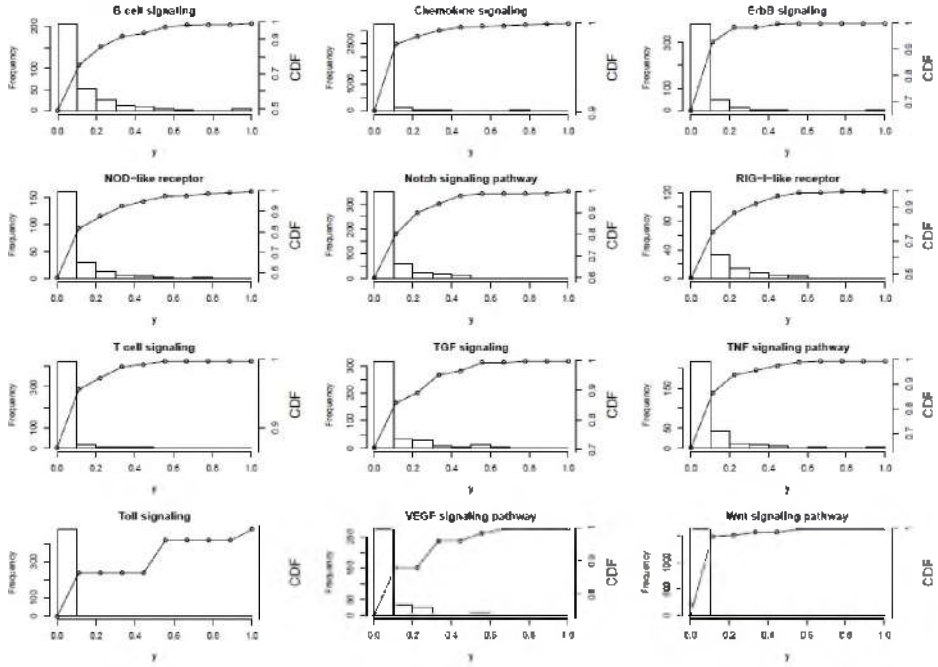

Рисунок 16. Распределение изменений активностей в биологических путях в зависимости от изменения топологии.

Это указывает на то, что в большинстве случаев поток сигналов пути может проявлять определенной степени устойчивость к изменениям топологии (или мутациям). Это достигается благодаря наличию большого количества ответвлений и возможных «обходов», которые обеспечивают непрерывную передачу сигнала от входных к выходным узлам через альтернативные маршруты.

Для подтверждения этих наблюдений мы промоделировали влияние мутаций на возмущения активности пути, используя «синтетические» биологические пути, полученные на основе случайных графов. Как показали результаты, линейная топология биологических путей наиболее уязвима для мутаций, а разветвленность повышает их устойчивость (рис. 17). Затем мы попытались определить, существует ли набор генов, мутации в которых могут в значительной степени повлиять на изменение сигнальных потоков в путях. Из каждого распределения значений возмущений пути мы выбрали пары генов, «мутации» в которых вызывали значимые изменения активности пути.

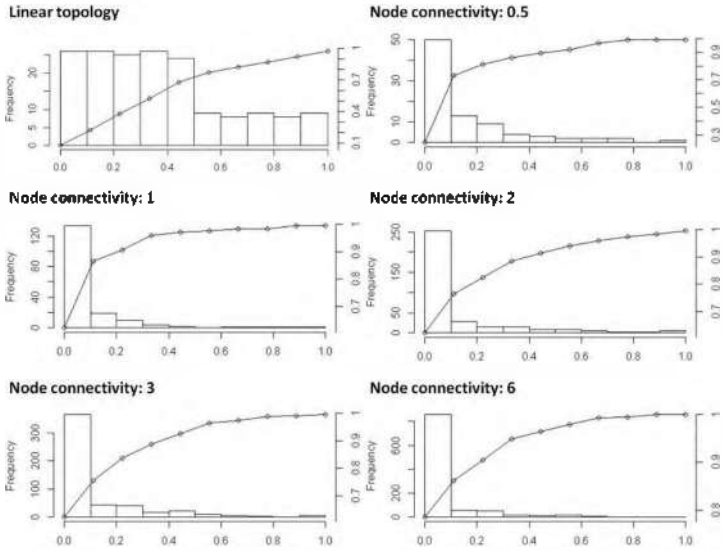

Рисунок 17. Распределение изменений сигнальных потоков в синтетических графах, начиная с линейной конфигурации графа с пошаговым увеличением связности узлов.

Следует отметить, что эти гены являются компонентами практически всех исследуемых путей. Функциональная аннотация выбранных генов показала, что они являются частью сигнальных каскадов MAPK и PI3K и играют ключевую роль в передаче сигналов от рецепторных к эффекторным узлам в пути (табл. 3). Более того, анализ ассоциации генов и заболеваний показал, что мутации в этих генах связаны с большим спектром моногенных и комплексных заболеваний человека. Наконец, нами было показано, что среднее число известных патогенных аллелей для этих генов выше (12 патогенных аллелей на ген), чем для генов, которые лишь умеренно влияли на изменения активности биологических путей (восемь патогенных аллелей на ген).

Полученные результаты показывают, что биологические пути содержат «узкие места», или гены-концентраторы, мутации в которых могут сильно влиять на их активность.

Однако следует также отметить, что активность пути зависит не только от топологии, но и от экспрессии генов. Фактически, мутации, которые могут резко изменить топологию пути, были описаны главным образом при моногенных и онкологических заболеваниях (Ito et al., 2014; Vajjhala et al., 2014). В случае комплексных полигенных заболеваний такие события редки, в то время как изменение экспрессии генов (Emilsson et al., 2008; Cookson et al., 2009) и степени сродства в белок-белковых взаимодействиях (например, изменение аффинности связывания вследствие генетических полиморфизмов) играют большую роль (Zhao et al., 2014). Эту ситуацию можно смоделировать путем корректировки весов ребер в графе, представляющем топологию пути, однако недостаточные знания о стохиометрических параметрах белков в сигнальных путях являются основным ограничивающим фактором развития исследований в данном направлении.

Таблица 3.

Функциональная аннотация генов, мутации которых оказывают наибольший эффект на изменение активности биологических путей.

Гены	Биологический путь из базы данных KEGG	Заболевание
PIK3R5, SOS2, IKBKG, HRAS, CHUK, RAF1, RELA, KRAS, PIK3R3, MAP2K2, MAP2K1, IKBKB, MAP3K7, AKT1, AKT2, PIK3CA, pik3cb, AKT3, PIK3CD, NRAS, GRB2, MAPK3, NFKB1, NFKBIB, PIK3CG, NFKBIA, PIK3R1, PIK3R2, SOS1, MAPK1	PI3K-Akt signaling pathway (p = 2.10E-29), MAPK signaling pathway (p = 5.02E-19)	Cancer (з=1e-08), Metabolic (P=6e-08), Infection (P=8e-07), Pharmacogenomic (P=4e-06), Neurological (P=0.03), Unknown Ethilogy (P=0.03)

В целом полученные нами результаты показывают, что сигнальные пути в определенной степени устойчивы к мутациям, которые влияют на белок-белковые взаимодействия из-за комплексных разветвленных топологий. Однако пути могут содержать узлы-концентраторы и узкие места, где мутации могут вызывать большие возмущения и существенно влиять на общую активность данного пути.

Исследование общих характеристик и специфических особенностей активации биологических путей при легочных заболеваниях

Состояние проблемы и цель исследования

Развитие и клиническое течение легочных заболеваний являются комплексными процессами, в которые вовлечены как генетические причины, так и факторы окружающей среды (Pouladi et al., 2015). При этом фенотипическое проявление болезни не всегда отражает лежащие в ее основе молекулярные механизмы. Дисфункция одного гена может способствовать развитию нескольких заболеваний, вызывать различные клинические проявления, и, наоборот, схожие клинические симптомы болезни могут быть обусловлены дисфункцией в разных генах (Lewis et al., 2008; Pennings et al., 2008).

Таким образом, исследование патогенеза легочных заболеваний на системном уровне позволит получить более детальное представление о молекулярных механизмах развития и течения этих патологий, охарактеризовать сходство и различия в патогенезе. В данном исследовании мы применили системный подход, направленный на изучение патобиологии широкого спектра злокачественных и хронических заболеваний легких на уровне дерегуляции активности биологических путей.

Использованные наборы данных и алгоритмы

Было использовано шесть наборов данных, доступных в базе данных Gene Expression Omnibus (GEO) (Edgar et al., 2002; Barrett et al., 2011), содержащей сигнатуры глобальной экспрессии генов при заболеваниях легких. Данные представлены в виде матриц экспрессии и содержат дополнительную информацию об исследуемых образцах, методах калибровки и нормализации. Для этого исследования были выбраны образцы пациентов, не получавших медикаментозное лечение. Всего в исследование было включено 948 образцов, отобранных у пациентов (острая интерстициальная пневмония (AIP, N=1); криптогенная пневмония (COP, N=3); хроническое обструктивное заболевание легких (COPD, N=220); десквамативная интерстициальная пневмония (DIP, N=4); идиопатический легочный фиброз (LF, N=10); гиперчувствительный пневмонит (HP, N=30); другие интерстициальные заболевания легких (ILD_OTHER, N=9); интерстициальные заболевания легких неизвестного происхождения (ILD_UNK, N=14); неспецифическая интерстициальная пневмония (NSIP, N=14); дыхательный бронхиолит - интерстициальное заболевание легких (RB-ILD, N=12); легочный саркоидоз (SARC, N=6); идиопатический легочный фиброз (UIP_IPF, N=157); аденокарцинома легких (ADC, N=85); базалоидный рак легких (BAS, N=39); карциноидный рак легких (CARCI, N=24); крупноячеистая карцинома легких (LCC, N=3); крупноклеточный нейроэндокринный рак легких (LCNE, N=56); рак легких, неидентифицированной природы (LCO, N=4); мелкоклеточная карцинома легких (SCC, N=21); плоскоклеточная карцинома легких (SQC, N=61)) и здоровых лиц (HC, N=170).

Оценка изменения активности биологических путей была выполнена с использованием разработанного нами алгоритма PSF (Arakelyan et al., 2013; Binder et al., 2014; Nersisyan et al., 2015). В качестве источника топологии биологических путей была использована библиотека, созданная нами на основе базы данных KEGG Pathway (Kanehisa et al., 2016) с помощью программ KEGGParser и CyKEGGParser (Arakelyan & Nersisyan, 2013; Nersisyan, et al., 2014). Анализ профилей активации биологических путей был проведен с помощью метода самоорганизующихся карт (self-organizing maps, SOM), доступного в пакете oposSOM для среды R (Löffler-Wirth et al., 2015).

Ландшафт изменений PSF для каждого заболевания описывается значениями экспрессии мета-PSF («индивидуальные» портреты заболеваний). Кластеры мета-PSF расположены в соответствии с базовой сеткой SOM и визуализируются с помощью соответствующего цветового градиента. Индивидуальные портреты взаимно сопоставимы. Алгоритм SOM располагает сходные профили мета-PSF на соседних фрагментах карты, тогда как отличающиеся расположены более удаленно. Мета-PSF, расположенные в одном и том регионе сетки SOM, образуют "пятна" или модули. Повышенная или пониженная экспрессия модулей в каждом «индивидуальном» портрете определяется как кластеры мета-PSF, значения которых выше (ниже) заданного порога (90% от максимума или минимума значений мета-PSF) (Wirth et al., 2011). Модули с «индивидуальных» портретов заболеваний затем переносятся на одну основную карту для визуализации глобальной картины изменений активности биологических путей.

Анализ значимости для дифференциальных значений PSF был проведен с использованием модификации t-теста Стьюдента и оценки доли ложных срабатываний (FDR) для множественных проверок (Wirth et al., 2011; 2012). Значения $p < 0.05$, и $FDR < 0.2$ были приняты как статистически значимые.

Тестирование сходства заболеваний легких проводилось на основе «индивидуальных» портретов с использованием анализа независимых компонент (independent component analysis, ICA) и иерархической кластеризации, доступных в пакете oposSOM (Wirth et al., 2011, 2012), а также методом поиска подмножеств в графе с методом случайного блуждания (random walktrap), реализованного в пакете igraph R (Csardi & Nepusz, 2006).

Функциональное состояние биологических путей при легочных заболеваниях

Профили PSF 138 метаболических и сигнальных путей KEGG были вычислены в 948 образцах легочной ткани (21 заболевание и здоровая легочная ткань) и затем трансформированы в серию двумерных изображений, называемых «портретами», которые визуализируют активность выходных узлов биологических путей для каждого заболевания (рис. 18).

На портретах прослеживаются модули, выделенные красными и синими цветами, обозначающие повышение и понижение активности на выходных узлах биологических сетей, соответственно.

Визуальное сравнение портретов заболеваний показало, что профили активации биологических путей в здоровых легких и при хронических заболеваниях имеют четко выраженные отличия по сравнению с онкологическими заболеваниями аналогичной локализации (рис. 18). Большинство онкологических состояний характеризуется наличием на портрете активного модуля (красное пятно), расположенного в правом нижнем углу, в сочетании с неактивным модулем (синее пятно) в левом верхнем углу. Хронические заболевания легких, в свою очередь, показывают зеркальное распределение активных и неактивных модулей. Кроме того, для этой группы нозологий характерна большая вариабельность распределения модулей по сравнению с патологиями онкологической этиологии.

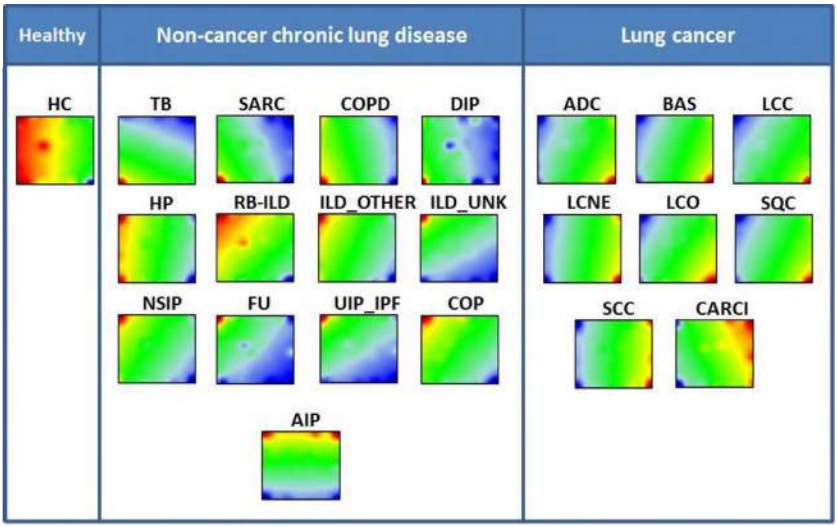

Рисунок 18. Индивидуальные портреты изменений активностей биологических путей (значений PSF) при легочных заболеваниях. Цветовая гамма от темно-синего (понижение) до темно красного (повышение) показывает степень изменения активностей биологических путей

Можно заметить, что портреты TB, SARC, COPD, DIP, в отличие от NSIP, FU, UIP / IPF, COP, ILD_OTHER, ILD_UNK, RB_ILD, имеют совпадающее распределение модулей. Распределение модулей на портрете HP является переходным между первой и второй группой хронических заболеваний, в то время как на портретах AIP, FU, UIP / IPF и

ILD_UNK этот параметр можно рассматривать как промежуточную форму между хроническими и онкологическими заболеваниями.

С целью получения глобальной картины дерегуляции активности биологических путей, связанных с конкретным заболеванием, модули всех отдельных портретов заболеваний были интегрированы в единую карту глобальной дерегуляции (рис. 19A). На данной карте модули представляют собой кластеры коррелированных и согласованно дерегулированных профилей активности выходных узлов биологических путей при одном или нескольких заболеваниях. Модули A, B, C и D в углах глобальной карты являются признаками, позволяющими разграничить хронические и злокачественные заболевания легких. Остальные модули обеспечивают более тонкую структуру состояний активации биологических путей при различных заболеваниях. В целом было выявлено, что с модулями связан 51 биологический путь, характеризующийся изменением активности, по меньшей мере, одного выходного узла/ответвления.

На следующем этапе была проведена полная функциональная аннотация биологических процессов, связанных с дерегуляцией активности выходных узлов биологических путей, расположенных в основных модулях глобальной карты. Поскольку выходные узлы являются продуктами генов, непосредственно связанных с каким-либо функциональным процессом, с использованием программы WebGestalt был проведен анализ обогащения категорий GO для каждого модуля (Zhang et al., 2005) (рис. 19B). Результаты показали, что модули A и C (характерные для хронических заболеваний), в основном связаны с иммунной/воспалительной реакцией, пролиферацией и ингибицией апоптоза, тогда как модули B и D (характерные для онкологических заболеваний легких) ассоциированы с регуляцией клеточного цикла, апоптозом и обменом углеводов.

Для оценки общности и различий между патомеханизмами легочных заболеваний данные об ассоциации биологических путей с функциональными модулями и заболеваниями были трансформированы в объект графа, где вершины отображают заболевания, а ребра – наличие общих изменений активности биологических путей между парами заболеваний. Затем был проведен поиск подмножеств (кластеров) в графе с использованием метода случайного блуждания по вершинам графа. В общей сложности были идентифицированы четыре кластера, содержащие три и более заболеваний, которые характеризуются однонаправленным изменением активности биологических путей (рис. 20).

Согласно полученным результатам, кластер 1 объединяет хроническое обструктивное заболевание легких (COPD), легочный саркоидоз (SARC) и туберкулез (TB). Многочисленные экспериментальные исследования, в том числе и наши собственные, показали сходные характеристики вовлечения иммунных/воспалительных процессов, таких как активация Toll-like рецепторов, фагоцитоз и сигнализация хемокинов, в патомеханизмы этих заболеваний (Arakelyan et al., 2009; Haspel & Choi, 2011; Kriegova et al., 2011; An et al., 2012; Maertzdorf et al., 2012; Pabst et al., 2013; Pugazhendhi et al., 2013). Более того, сигнатуры дифференциальной экспрессии генов, а также ассоциация генетических полиморфизмов, связанных с этими заболеваниями, характеризуются значительным сходством (Arakelyan et al., 2009; Haspel & Choi, 2011). Кроме того известно, что особенности патофизиологии этих заболеваний также очень похожи (Maertzdorf et al., 2012). Таким образом, результаты проведенного биоинформатического анализа полностью соответствуют имеющимся данным, подтверждающим наличие общей картины дерегуляции иммунной системы при хроническом обструктивном заболевании легких, легочном саркоидозе и туберкулезе.

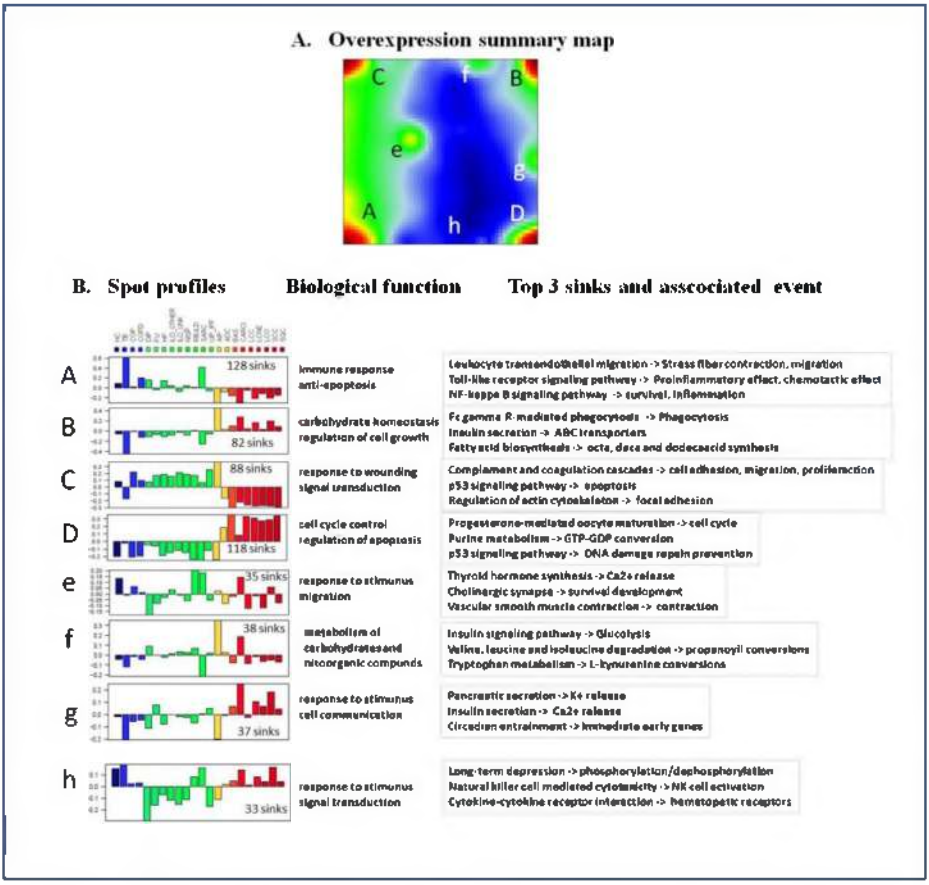

Рисунок 19. Глобальная картина дерегуляции активностей биологических путей при легочных заболеваниях.
(A) На карте отмечены основные (A, B, C, D) и промежуточные (e,f,g,h) модули дерегуляции биологических путей.
(B) Функциональная аннотация модулей по категориям GO.

Второй кластер объединяет дыхательный бронхиолит (RB-ILD), гиперчувствительный пневмонит (HP) и криптогенную пневмонию (COP) (Leslie, 2009; Meyer et al., 2012). Патомеханизмы этих заболеваний изучены недостаточно, хотя и есть немногочисленные публикации, описывающие кластеризацию RB-ILD, HP и COP на основе глобальной экспрессии генов (Cho et al., 2011; Lee & Yang, 2013).

Наши результаты показали, что связь между этими заболеваниями может быть реализована посредством общей дерегуляции путей, ассоциированных с клеточным циклом (сигнальные пути Hippo и p53), трансдукцией внеклеточных сигналов (цитокины, гормоны щитовидной железы) и метаболизмом пуриновых и пиримидиновых нуклеиновых кислот.

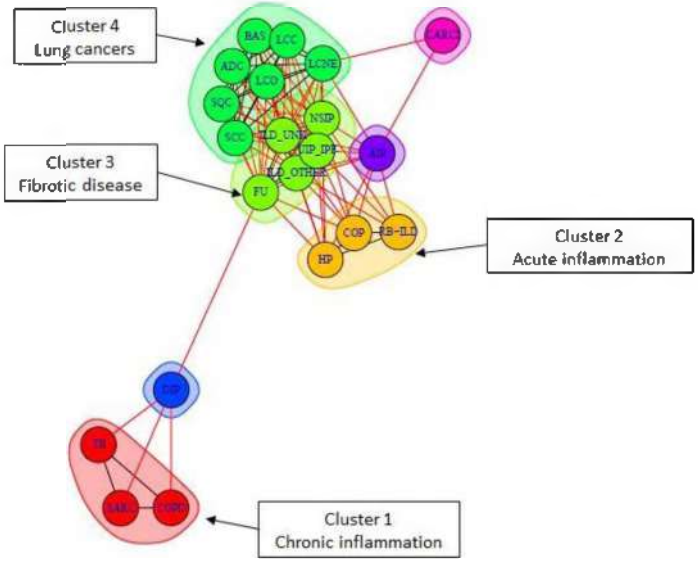

Рисунок 20. Идентификация кластеров заболеваний, характеризующихся наиболее схожей картиной изменений активностей биологических путей.

Наконец, онкологические заболевания легких (кластер 4) характеризовались снижением активности биологических путей, связанных с иммунным ответом, а также повышением активности клеточного цикла, пролиферацией и метаболизмом, что соответствует ранее опубликованным данным (Han et al., 2014; Domagala-Kulawik, 2015). Кроме того, из всех типов рака легких, включенных в исследование, профили активации биологических путей карциноидной опухоли легкого (CARCI) заметно отличались от других онкологических состояний (Anbazhagan et al., 1999), что может указывать на ее нейроэндокринное происхождение (Rekhtman, 2010).

Используя разработанный нами метод оценки активности биологических путей PSF в паре с алгоритмом SOM, в данном исследовании была проведена оценка сходства профилей дерегуляции биологических путей при различных легочных патологиях путем исследования данных о глобальной экспрессии генов и топологий сигнальных и метаболических путей. Наши результаты выявили значительные различия в дерегуляции путей, связанных с онкологическими и хроническими заболеваниями легких. В то время как рак легких характеризуется активацией биологических путей, сопряженных с пролиферацией клеток и метаболизмом, хронические заболевания характеризуется изменениями в путях, связанных с иммунным ответом и фиброзным ремоделированием легочной ткани. Кроме того, при интерстициальных заболеваниях легких мы наблюдали значительную гетерогенность профилей активации биологических путей. Полученные данные позволили идентифицировать три различных кластера, показав, что развитие характерных патологических процессов, таких как фиброз, может быть инициировано дерегуляцией в разных сигнальных путях. Эти результаты указывают на необходимость разработки новой классификации интерстициальных заболеваний с учетом молекулярных механизмов патогенеза, а не клинических проявлений. Наконец, мы также выявили существенное сходство между онкологическими новообразованиями и интерстициальными заболеваниями, характеризующимися развитием фиброза, что

свидетельствует о наличии общих молекулярных механизмов, участвующих в их патогенезе.

Таким образом, полученные данные в значительной степени расширяют наше понимание молекулярных механизмов развития онкологических и хронических заболеваний легких. Более того, результаты нашей работы формируют новое представление о молекулярных механизмах ряда интерстициальных легочных заболеваний, которые были изучены в меньшей степени по сравнению с интерстициальным легочным фиброзом, хронической обструктивной болезнью легких и легочным саркоидозом.

Сравнительный анализ профилей активации биологических путей при аутовоспалительных и аутоиммунных заболеваниях

Состояние проблемы и цель исследования

Эпидемиологические исследования указывают на значительный рост распространенности заболеваний, связанных с дисфункцией иммунной системы, в том числе аутоиммунных и аллергических состояний (Bach, 2002; Eaton et al., 2010; Mackay et al., 2010).

Фенотипическая гетерогенность аутоиммунных и аутовоспалительных заболеваний не обязательно отражает фундаментальные генетические или механистические различия между этими группами. Действительно, хотя некоторые варианты и полиморфизмы генов специфичны для конкретного заболевания (WTCC, 2007; Costenbader et al., 2012), другие ассоциированы с предрасположенностью к развитию множественных расстройств, что указывает на то, что общие мутации могут влиять на общие гены или биологические пути, вовлеченные в патогенез нескольких заболеваний (Melanitou et al., 2003; Castiblanco et al., 2013). Более того, одинаковый ответ на лекарственные препараты (например, глюкокортикоиды) (Baughman & Lower, 2014; Ciccarelli et al., 2014; Zhao et al., 2014) указывает на наличие общих молекулярных мишеней при различных заболеваниях. Однако глобальная картина молекулярных механизмов, лежащих в основе сходства и особенностей хронических воспалительных заболеваний, недостаточна изучена.

Целью данного исследования являлось комплексное изучение изменений активности биологических путей при хроническом воспалении, связанном с аутоиммунными и аутовоспалительными заболеваниями. Была проведена оценка сходства и различий в изменении активности биологических путей в комбинированном наборе данных, содержащих профили глобальной экспрессии генов в двенадцати аутоиммунных и аутовоспалительных заболеваниях с применением биоинформационного алгоритма, комбинирующего методы самоорганизующихся карт и оценки сигнальных потоков.

Использованные наборы данных и алгоритмы

В исследовании были использованы наборы данных глобальной экспрессии, содержащиеся в базе данных GEO (Edgar et al., 2002). Поиск был проведен по ключевым словам “autoinflammation”, “autoimmunity” и ограничен только клиническими образцами. Целью отбора являлось получение максимально гомогенной выборки, которая соответствовала разработанному дизайну проведенных экспериментов, типу исследованной ткани и платформы, использованной для измерения экспрессии генов. С учетом этих требований были выбраны наборы данных, полученные из образцов мононуклеарных клеток периферической крови, платформа Affymetrix, содержащих образцы пациентов и здоровых лиц (исследования типа «случай-контроль»).

Конечная выборка состояла из восьми наборов данных микрочипов, содержащих профили экспрессии генов в образцах мононуклеарных клеток периферической крови пациентов с аутоиммунными (диабет 1-го типа (T1D/СД1), N=12; рассеянный склероз

(MS/РС), N=12; системная красная волчанка (SLE/СКВ), N=61; синдром Шегрена (SS/СС), N=11) и аутовоспалительными заболеваниями (криопирин-ассоциированные периодические синдромы (CAPS/КАПС), N=23; синдром гипериммуно-глобулинемии d (мутации гена mvk) (HIDS/СГГД), N=8; синдром стерильного пиогенного артрита в сочетании с гангренозной пиодермией (PAPA/ССПАГП), N=6; синдром, ассоциированный с рецептором фактора некроза опухоли (TRAPS/САРФНО), N=29; болезнь бехчета (BD/ББ), N=15; болезнь крона (CD/БК), N=59; язвенный колит (UC/ЯК), N=26; ювенильный идиопатический артрит (ЛА/ЮИА), N=22), а также здоровых лиц (N=254). Для расчетов значений экспрессии генов были использованы исходные файлы измерений интенсивности сигналов на микрочипах в формате Affymetrix CEL. Для набора данных GSE3365 исходные файлы были недоступны, поэтому в анализе использовались заранее рассчитанные значения экспрессии генов, доступные в GEO. Преобразование интенсивности сигнала зондов, нормализация RMA и аннотация зондов для микрочипов серии Affymetrix Human Genome были выполнены с использованием пакета «affy» для R (Gautier et al., 2004), а для микрочипа Affymetrix Human Exon 1.0 ST Array – с помощью пакета «oligo» для R (Carvalho & Irizarry, 2010). Средние значения экспрессии генов в контрольных образцах использовались для расчета кратности изменения экспрессии в натуральной шкале (fold change, FC).

Оценка изменения активности биологических путей была выполнена с использованием разработанного нами алгоритма PSF-SOM (Arakelyan et al., 2017), основанного на совмещении алгоритмов PSF (Arakelyan et al., 2013; Binder et al., 2014; Nersisyan et al., 2015) и метода SOM (Löffler-Wirth et al., 2015). Расчет значений PSF во всех образцах был проведен для 1825 ответвлений в 168 биологических путях. Полученная матрица профилей активации биологических путей являлась входными данными для последующего анализа методом SOM. В данном случае использовалась карта нейронной сетки размером 35x35; обучение нейронной сети выполнялось со стандартными параметрами. Идентификация функциональных модулей на карте SOM выполнялась по заданным пороговым значениям (90% от максимума или минимума значений мета-PSF) (Wirth et al., 2011), кластеризация – методом поиска подмножеств в графе с использованием алгоритма случайного блуждания (random walktrap), реализованного в пакете igraph R (Csardi & Nepusz, 2006), а функциональный анализ – с использованием биоинформатических программ DAVID и WebGestalt (Huang et al., 2008; 2009; Wang et al., 2013).

Функциональное состояние биологических путей при воспалительных заболеваниях

Профили активации биологических путей при хроническом воспалении

Метод PSF-SOM позволил получить «усредненный» портрет профилей активации биологических путей для каждого заболевания. Цветовая гамма портрета – от темно-синего (низкая) до темно-красного (высокая) – отражает степень изменения активности кластера биологических путей, связанных с определенным участком на «портрете» (рис. 21).

Визуальное сравнение портретов заболеваний выявило значительное сходство: почти все нозологии, за исключением синдрома гипериммуно-глобулинемии D (MVK), характеризовались наличием активного модуля в левом нижнем углу. Кроме того, на молекулярном портрете рассеянного склероза (MS) присутствует дополнительный активный модуль в верхнем правом углу, тогда как синдром гипериммуно-глобулинемии D (MVK), криопирин-ассоциированные периодические синдромы (CAPS), рассеянный склероз (MS) и, в меньшей степени, синдром стерильного пиогенного артрита в сочетании с гангренозной пиодермией (PAPA) характеризовались наличием дополнительного модуля вблизи нижнего правого угла.

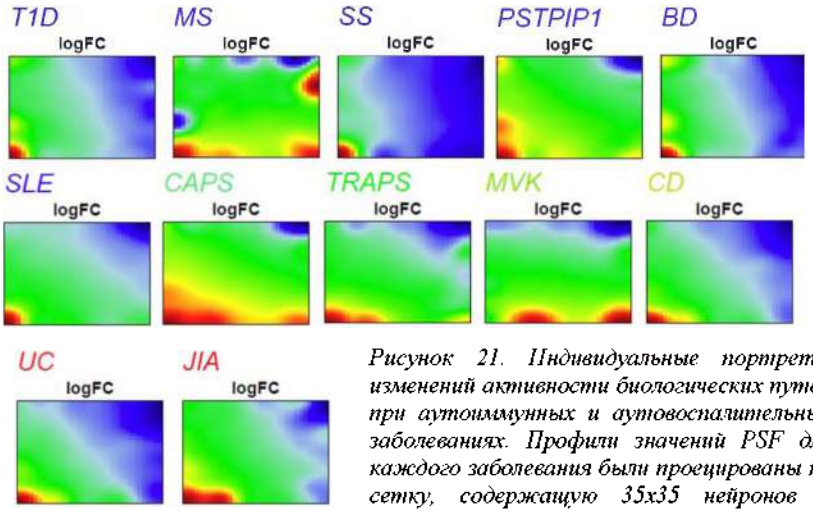

Рисунок 21. Индивидуальные портреты изменений активности биологических путей при аутоиммунных и аутовоспалительных заболеваниях. Профили значений PSF для каждого заболевания были проецированы на сетку, содержащую 35x35 нейронов и визуализированы в виде двумерных карт.

Цветовая гамма портрета от темно-синего (низкая) до темно-красного (высокая) отражает степень изменения активности кластера биологических путей, связанных с определенным участком на «портрете».

Наконец, следует отметить, что три моногенных аутовоспалительных заболевания (CAPS, TRAPS и HIDS), а также JIA и MS и, в меньшей степени, UC и CD содержали дополнительный модуль вблизи нижнего левого угла.

Оценка схожести исследуемых заболеваний

Для оценки степени сходства профилей активации биологических путей при аутоиммунных и аутовоспалительных заболеваниях был построен граф, где вершины отображают заболевания, а ребра – наличие общих функциональных модулей биологических путей в парах заболеваний. Затем был проведен поиск подмножеств (кластеров) в графе с использованием метода случайного блуждания по вершинам графа. В общей сложности были идентифицированы два кластера, характеризующиеся однонаправленным изменением активности биологических путей (рис. 22).

Кластер 1 объединяет 4 аутоиммунных (T1D, BD, SS и MS) и одно аутовоспалительное заболевание (PAPA). Этот кластер характеризуется активными функциональными модулями D и F. Аутоиммунные заболевания в кластере 1 в значительной степени имеют общие симптомы и клинические проявления. Так, в опубликованных работах по клиническим случаям сообщается, что первичные проявления синдрома Шергена (SS) могут имитировать таковые при рассеянном склерозе (MS) (Jung et al., 2000; de Seze et al., 2001; Solomon et al., 2013). Более того, синдром Шергена часто сопровождает диабет как у людей (Binder et al., 1989), так и в мышиной модели сахарного диабета 1-го типа (T1D), не отягощенного ожирением (Non-Obese Diabetic, NOD) (Brayer et al., 2000). Кроме того, исследования у мышей показали, что гены устойчивости к сахарному диабету 1-го типа в специфических хромосомных локусах (*Idd3* и *Idd5*) защищают организм от воспаления и от дисфункции слюнных желез (Brayer et al., 2000). Рассеянный склероз имеет сходство с

неврологическими проявлениями болезни Бехчета (Ashjazadeh et al., 2003), которая является коморбидным для диабета 1-го типа (Zapatero & Colin, 2008; Song et al., 2014).

Примечательно, что синдром стерильного пиогенного артрита в сочетании с гангренозной пиодермией (PAPA), являясь моногенной аутовоспалительной патологией, характеризующейся активацией воспалительных процессов, не содержал активированного модуля С и оказался в одном кластере с аутоиммунными заболеваниями. Заметные различия между этим синдромом и другими аутовоспалительными заболеваниями также были описаны в литературе. Так, пациенты с синдромом стерильного пиогенного артрита в сочетании с гангренозной пиодермией менее чувствительны к противовоспалительному лечению, направленному на угнетение IL-1 и TNF сигнальных каскадов (Demidowich et al., 2012). Кроме того, было показано, что в мышиной модели PAPA ген *PSTPIP1* не является регулятором активации воспаления, что указывает на альтернативные эффекты мутаций *PSTPIP1* при PAPA (Wang et al., 2013). Кроме того, модуль С является кластером биологических путей, ассоциированных с актин-опосредованной миграцией клеток, в то время как мутации в *PSTPIP1* приводят к нарушению образования подосом (Cortesio et al., 2010). В целом, полученные нами результаты указывают на определенные особенности молекулярных механизмов синдрома стерильного пиогенного артрита в сочетании с гангренозной пиодермией по сравнению с другими аутовоспалительными заболеваниями.

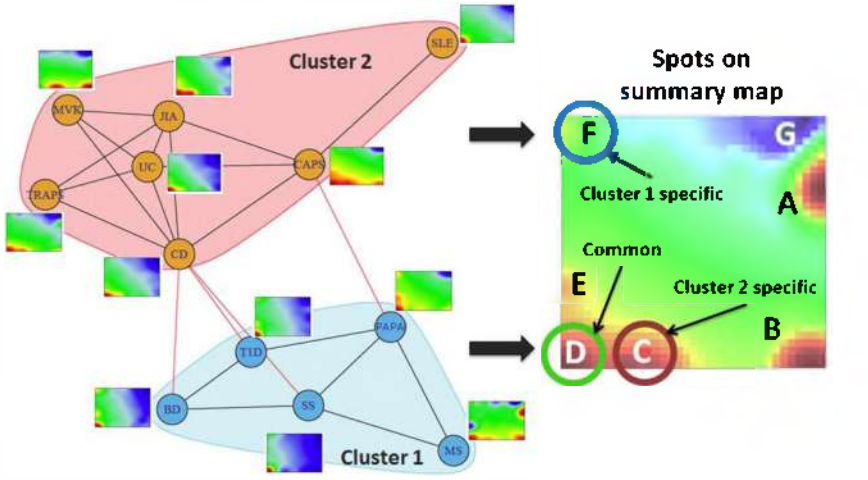

Рисунок 22. Идентификация кластеров заболеваний, характеризующихся наиболее похожей картиной изменений активностей биологических путей.

Исходя из неожиданной кластеризации PAPA с аутоиммунными нарушениями и наблюдаемой разницы между аутовоспалительными заболеваниями, мы изучили изменения PSF сигнального пути «NOD-like receptor signaling pathway» при всех заболеваниях, поскольку именно в этом каскаде описаны процессы, связанные с активацией инфламасом (Shaw et al., 2010).

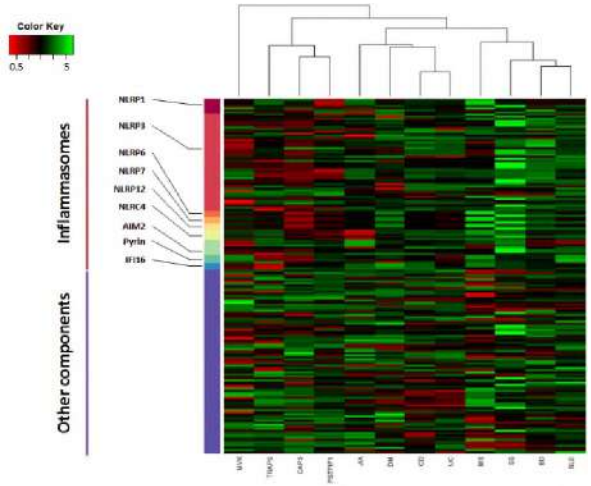

Рисунок 23. Тепловая карта (heatmap) значений PSF в узлах биологического пути «NOD-like receptor signaling pathway».

Результаты анализа показали ожидаемое группирование всех аутовоспалительных заболеваний (HIDS, PAPA, TRAPS и CAPS) в один кластер, а аутоиммунных – в другой (рис. 23). Это, в свою очередь, указывает на то, что специфические особенности, наблюдаемые для синдрома PAPA, не связаны со сборкой и активацией инфламасом, а скорее – с дерегуляцией активности других биологических путей.

Кластер 2 объединяет моногенные (HIDS, CAPS и TRAPS) и полигенные аутовоспалительные (JIA, CD, UC) заболевания (McGonagle & McDermott, 2006; Ciccarelli et al., 2014) и характеризуется высокой активностью модулей C и D (рис. 22).

Системная красная волчанка (SLE), будучи аутоиммунным заболеванием, также является членом данного кластера. Системный характер волчанки, при которой иммунная система атакует здоровые ткани в нескольких частях тела, может объяснить наличие общих модулей с другими аутовоспалительными синдромами в кластере (Lisnevskaia et al., 2014). Недавно опубликованные данные свидетельствуют о том, что системная красная волчанка имеет сходные молекулярные сигнатуры с моногенными аутовоспалительными синдромами, опосредованными интерфероном 1-го типа (Rodero & Crow, 2016). Более того, показано, что аутовоспалительные процессы играют важную роль в патогенезе этого заболевания (Kahlenberg & Kaplan, 2014). Недавние исследования, проведенные Shin и соавт. (Shin et al., 2013) и Zhang и соавт. (Zhang et al., 2016), показали, что аутоантитела к двухцепочечной ДНК в моноцитах могут индуцировать активацию NLRP3 инфламасомы, приводя к увеличенной продукции IL-1β. Последнее, в свою очередь, индуцирует Th17 клеточный ответ при системной красной волчанке. В другом исследовании Kahlenberg и соавт. (Kahlenberg et al., 2014) было установлено, что каспаза-1, центральный фермент, участвующий в активации воспаления, имеет важное значение для развития SLE в модели индуцируемой волчанки у мышей.

Полученные нами результаты свидетельствуют о том, что независимо от механизмов инициации заболевания, профили активности нисходящих биологических путей имеют значительное сходство, которое может быть обусловлено хроническим воспалением. Последнее, в свою очередь, является существенным патологическим фактором, связанным со всеми исследуемыми заболеваниями. Таким образом, формальная демонстрация общих профилей дерегуляции путей, связанных с воспалением, между этими заболеваниями

свидетельствует о том, что наблюдаемые нарушения могут создать условия для развития осложнений и сопутствующих синдромов, как это было показано выше.

Функциональный анализ активных модулей биологических путей

Для функциональной аннотации путей, ассоциированных с активными модулями, был проведен анализ ORA с использованием программ WebGestalt и REVIGO (Supek et al., 2011; Wang et al., 2013). Полученные данные показывают, что во всех модулях наблюдалась повышенная репрезентативность категорий GO, связанных с различными аспектами иммунного ответа (рис. 24).

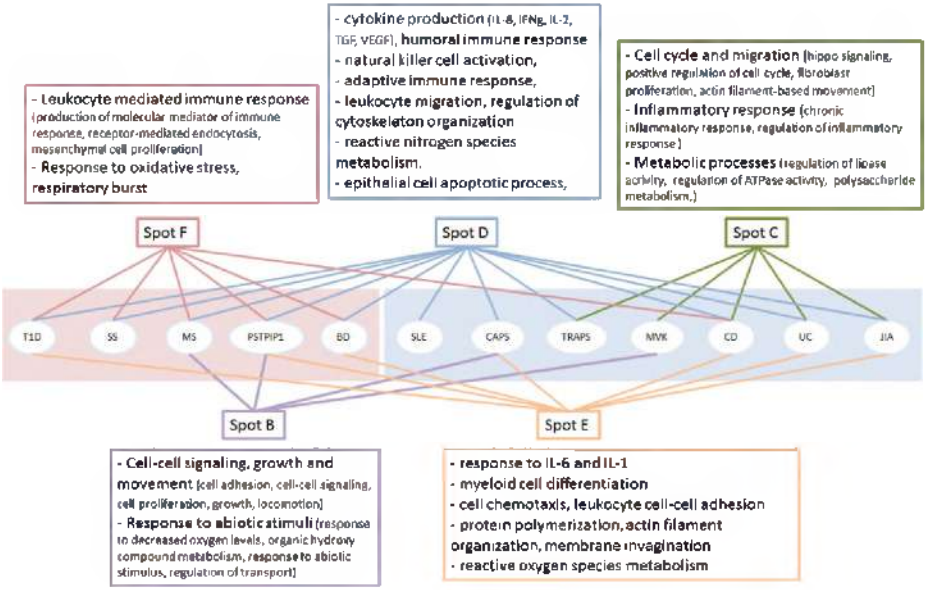

Рисунок 24. Функциональная аннотация процессов, связанных с биологическими путями в активных модулях на сводной карте профилей активации биологических путей. Кластеры заболеваний показаны розовым и синим цветом.

Полученные нами данные подтверждают избыточную активацию основных патологических событий, приводящих к развитию хронического воспаления, о чем свидетельствует наличие модуля D при всех заболеваниях (за исключением синдрома гипериммуно-глобулинемии D). Между тем, наблюдается определенная степень специфичности, обусловленная наличием других модулей, каждый из которых связан с тем или иным кластером заболеваний. Модуль F специфичен для кластера 1 (аутоиммунные заболевания) и ассоциирован с категориями GO, связанными с иммунным ответом, тогда как модуль C – для кластера 2 (аутовоспалительные заболевания) ассоциируется с воспалительным ответом. Два других модуля (B и E), по-видимому, играют вспомогательную роль, являясь либо амплификаторами иммунных/воспалительных реакций (как, например, в случае криопирин-ассоциированных периодических синдромах или рассеянного склероза), либо обеспечивают альтернативные механизмы активации иммунного ответа (как в случае синдрома гипериммуно-глобулинемии D).

В общей сложности 131 из 168 исследованных сигнальных и метаболических путей характеризовался нарушением активности, по меньшей мере, в одном ответвлении при одном заболевании. Таким образом, нельзя утверждать, что молекулярные механизмы исследуемых заболеваний связаны с нарушением активности в специфических биологических путях. Более того, нами было отмечено, что разные ответвления одного биологического пути связаны с разными функциональными модулями (в среднем три модуля на путь), что указывает на необходимость учета степени разветвленности пути и разнообразия функциональных событий, связанных с одним путем. Интересно, что дерегулированные ответвления указанных путей сконцентрированы в модулях C, D, E и F. Все эти пути напрямую связаны с хроническим воспалением, что, как и ожидалось, указывает на их центральную роль в этиологии изучаемых заболеваний (рис. 25).

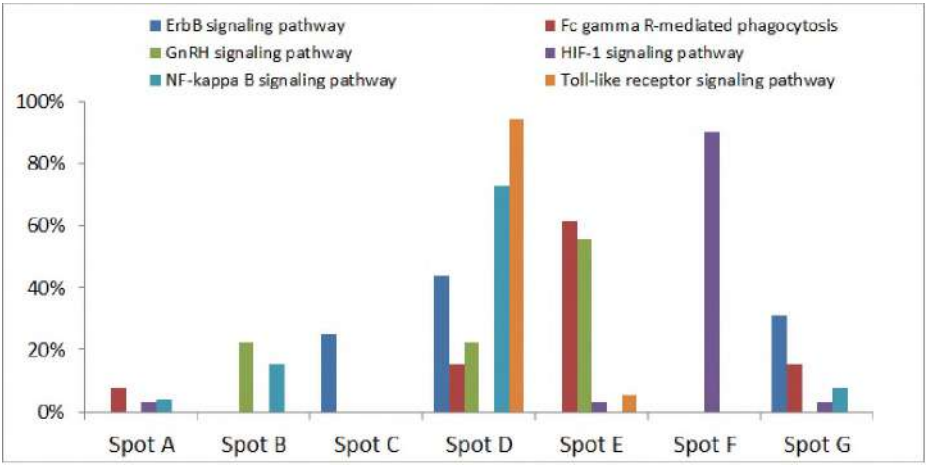

Рисунок 25. Распределение ответвлений и выходных узлов биологических путей по функциональным модулям.

Полученные результаты показывают, что и аутоиммунные, и аутовоспалительные заболевания характеризуются развитием воспалительного ответа, направленного на собственные ткани и органы, а также нарушениями гомеостаза канонических цитокиновых каскадов. При этом сравнительный анализ изменений профилей активации биологических путей позволяет оценивать наблюдаемые явления с точки зрения системной биологии. Результаты предыдущих исследований широкогеномных вариаций, экспрессии генов и воздействия лекарственных препаратов указывали на общность патомеханизмов хронических воспалительных заболеваний (Ciccarelli et al., 2014; Lee et al., 2014; Chimenti et al., 2015). Однако проведенные исследования, как правило, ограничивались выяснением количества сходных заболеваний, а также носили локальный характер – без попыток получить общую картину задействованных механизмов.

В данном случае мы использовали подходы системной биологии, которые позволили выявить универсальность нарушений дисфункции биологических путей при аутоиммунных и аутовоспалительных расстройствах, а также раскрыть новые аспекты сходства в патомеханизмах этих заболеваний. Результаты данного исследования четко указывают на комплексный характер механизмов, вовлеченных в патогенез аутоиммунитета и аутовоспаления. С одной стороны, картина довольно сложная и обусловлена влиянием большого количества факторов. С другой, кластеризация,

основанная на активности биологических путей, показала, что болезни, характеризующиеся сильно различающимися клиническими проявлениями, в основном имеют сходные патомеханизмы. В некоторых случаях связь заболеваний в выделенных кластерах подтверждалась результатами ранее опубликованных работ, а в других случаях наше исследование привело к описанию новых фактов. Так, в нашем исследовании мы выделили особенности патомеханизма синдромов стерильного пиогенного артрита в сочетании с гангренозной пиодермией (Wise et al., 2002b) и гипериммуно-глобулинемии D (Mulders-Manders & Simon, 2015). Предыдущие исследования были в основном сконцентрированы на роли NLRP3 и пириновых инфламасом в активации каспазы-1 и повышенной секреции IL-1β при обоих синдромах (Wilson & Cassel, 2010; Xu et al., 2014; Park et al., 2016). Интересно, что синдром гипериммуно-глобулинемии D, в отличие от других аутовоспалительных заболеваний, опосредован мутациями в гене *MVK*, кодирующем метаболизирующий фермент мевалонат-киназу (Drenth et al., 1999; Waterham et al., 1999; Park et al., 2016), а не в генах инфламасом. Мевалонатный путь является важным регулятором клеточных процессов через метаболизм стерольных (холестерин) и нестерольных изопреноидов (долихол, гем-А, убихинон). Нестерольные изопреноиды, будучи ключевыми мессенджерами клеточного роста и дифференциации, являются потенциальными молекулярными мишенями при онкологических, аутоиммунных и нейродегенеративных заболеваниях (Buhaescu & Izzedine, 2007). В контексте этих заболеваний мутации в гене *MVK* приводят к инициации внутриклеточных каскадов, запускающих активацию пириновой инфламасомы и секреции IL-1β (Wilson & Cassel, 2010; Favier & Schulert, 2016). Это может частично объяснить отличие активных модулей, ассоциированных с указанными патологиями, от остальных аутовоспалительных синдромов.

В целом полученные нами результаты показывают, что в патомеханизме аутоиммунных и аутовоспалительных заболеваний задействованы общие функциональные модули биологических путей, относящихся к врожденному и приобретенному иммунитету. При этом характер патологии во многом зависит от баланса между нарушениями в этих двух системах (Aziz et al., 2009). Идентификация общих путей способствует более полному пониманию патомеханизмов аутоиммунных и аутовоспалительных заболеваний и может способствовать разработке терапевтических методов, направленных на модуляцию активности биологических путей.

Анализ профилей активации биологических путей при аутовоспалительных заболеваниях

Состояние проблемы и цель исследования

Аутовоспалительные заболевания характеризуются аномальной активацией врожденного иммунного ответа экзогенными или эндогенными стимулами без участия аутоантител и/или аутореактивных Т-клеток. Семейные аутовоспалительные заболевания в большинстве своем являются моногенными, т. е. развитие болезни вызывается мутациями в одном конкретном гене (Touitou & Koné-Paut, 2008). Врожденные и de novo мутации в генах могут вызывать приобретение функции (gain-of-function) через механизмы, связанные с физиологической ролью белка в сигнальной трансдукции, или путем активации других процессов, таких как ответ на мисфолдинг белков (protein misfolding), стресс эндоплазматического ретикулума (ЭР-стресс), образование свободных радикалов или активация воспалительного ответа (Park et al., 2012). Тем не менее, несмотря на очевидные факты, в настоящее время все еще непонятно, как одиночная мутация приводит к глобальным изменениям клеточной физиологии при моногенных аутовоспалительных заболеваниях.

Целью данного исследования являлась оценка изменений глобальной экспрессии генов и общей активности сигнальных путей при моногенных аутовоспалительных заболеваниях. В данном исследовании мы попытались выяснить общие изменения указанных факторов в зависимости от различных мутаций при аутовоспалительных заболеваниях.

Использованные наборы данных и алгоритмы

В данном исследовании был использован набор данных GSE43553, депонированный в хранилище Gene Expression Omnibus (Barrett et al., 2011) и содержащий профили экспрессии генов мононуклеарных клеток периферической крови больных с CAPS, TRAPS, гипериммуноглобулинемией D (HIDS) и стерильным пиогенным артритом в сочетании с гангренозной пиодермией (PAPA), а также здоровых лиц в качестве контроля. Информация о мутациях была доступна в соответствующем описании образца для каждого пациента.

Для анализа экспрессии генов были использованы исходные файлы измерений интенсивности сигналов на микрочипах в формате Affymetrix CEL. Преобразование интенсивности сигнала зондов, нормализация RMA и аннотация зондов на микрочипах были выполнены с использованием пакета «affy» для R (Gautier et al., 2004). Средние значения экспрессии генов в контрольных образцах использовались для расчета кратности изменения экспрессии в логарифмической шкале (log fold change, logFC).

Анализ экспрессии генов выполнялся с использованием алгоритма самоорганизующихся карт (SOM) пакета «oposSOM» (Wirth et al., 2011; Löffler-Wirth et al., 2015). Этот алгоритм переводит высокоразмерную N x M (N – число генов, M – количество образцов) матрицу экспрессий генов (logFC) в матрицу метаданных K x M (K – число метагенов). Каждый метаген служит в качестве репрезентативного прототипа кластера реальных генов с аналогичными профилями экспрессии. Профиль экспрессии метагенов, в свою очередь, соответствует усредненному профилю экспрессии реальных генов. Использовалась двумерная сетка SOM прямоугольной топологии размерностью 40 × 40 нейронов и стандартные параметры обучения нейронной сети.

Оценка значимости изменений экспрессии генов проводилась модифицированным T-тестом Стьюдента, учитывающим стандартную ошибку значений экспрессии каждого гена в реплицированных измерениях. Пакет «fdrtool» R был далее использован для расчета доли ложных срабатываний (FDR) (Strimmer, 2008). Функциональную аннотацию генов в дерегулированных модулях проводили с использованием биоинформатических программ DAVID и WebGestalt (Huang et al., 2008; 2009; Wang et al., 2013).

Анализ активности биологических путей проводился с использованием алгоритма PSF, описанного выше. В качестве источника топологий биологических путей была использована библиотека, созданная нами на основе базы данных KEGG Pathway (Kanehisa et al., 2016) с помощью программ KEGGParser и CyKEGGParser.

Несмотря на то, что мутации в генах *MVK*, *NLRP3*, *PSTPIP1* и *TNFRSF1A*, вызывающих развитие аутовоспалительного фенотипа, достаточно хорошо описаны, их влияние на изменение белок-белковых взаимодействий и топологию биологических путей все еще остается невыясненным. По этой причине из данного исследования были исключены биологические пути, включающие вышеуказанные гены (Terpenoid backbone biosynthesis для *MVK*, NOD-like receptor signaling pathway для *NLRP3* и *PSTPIP1*, MAPK signaling pathway, cytokine-cytokine receptor interaction, NF-kappa B signaling pathway, Sphingolipid signaling pathway, mTOR signaling pathway, TNF signaling pathway, Adipocytokine signaling pathway, Apoptosis, и Osteoclast differentiation для *TNFRSF1A*).

Дифференциальный анализ активности биологических путей был выполнен с помощью алгоритма самоорганизующихся карт (SOM) с использованием стандартных

параметров обучения нейронной сети размерностью 30 x 30, по методу PSF-SOM (Arakelyan et al., 2017). Идентификация функциональных модулей на карте SOM проводилась по заданным пороговым значениям (90% от максимума или минимума значений мета-PSF) (Wirth et al., 2011), кластеризация – методом поиска подмножеств в графе с использованием алгоритма случайного блуждания (random walktrap), реализованного в пакете igraph R (Csardi & Nepusz, 2006), а функциональный анализ – с использованием биоинформатических программ DAVID и WebGestalt (Huang et al., 2008; 2009; Wang et al., 2013).

Влияние мутаций на транскриптомный ландшафт и активность биологических путей при аутовоспалении

Анализ глобальной экспрессии генов показал значительную вариабельность транскриптомного ландшафта у пациентов с различными мутациями (рис. 26), а также в распределении функциональных кластеров дифференциально экспрессируемых генов, называемых модулями в методе SOM.

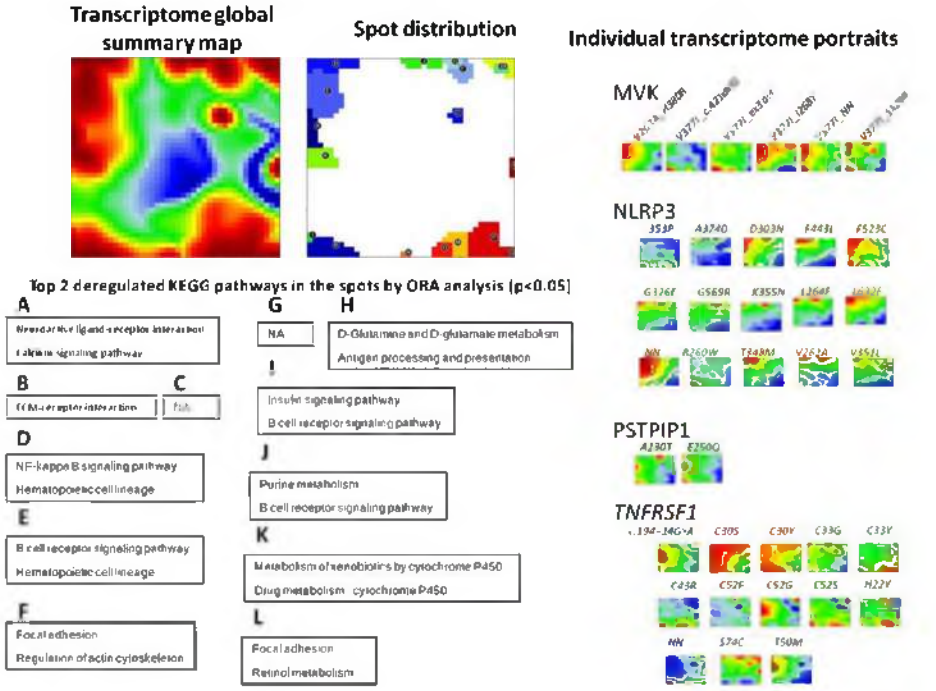

Рисунок 26. Молекулярные портреты экспрессии генов при аутоиммунных заболеваниях. Слева представлены индивидуальные портреты для каждой мутации.

Цветовая гамма от темно-синего (понижение) до темно красного (повышение) показывает уровень экспрессии по сравнению с контролем. Справа представлена сводная карта глобальных изменений транскриптома и функциональная аннотация активных модулей в категориях KEGG pathways, полученных с помощью анализа ORA.

Результаты показали, что профили экспрессии генов у пациентов с мутациями *MVK* V377I / S329R, *TNFRSF1A* C30S, *TNFRSF1A* C30Y, *TNFRSF1A* H22Y, *TNFRSF1A* NN и *TNFRSF1A* T50M аналогичны экспрессии генов в здоровых контрольных образцах. Транскриптомный ландшафт остальных пациентов содержал от одного до четырех дифференциально экспрессируемых функциональных модулей. Функциональная аннотация модулей экспрессии с использованием программ DAVID и WebGestalt не смогла выявить нарушения в биологических путях, характерных для аутовоспалительных синдромов (рис. 26). Кроме того, кластерный анализ, основанный на корреляции модулей, не выявил единой картины экспрессии генов, что позволяет разграничить аутовоспалительные синдромы друг от друга (рис. 27). Между тем анализ активности биологических путей продемонстрировал более ожидаемые результаты (рис. 28). Глобальная карта дерегуляции активностей биологических путей содержала шесть модулей. Все мутации гена *MVK* характеризовались наличием модуля A, в то время как большинство мутаций *NLRP3* (9 из 15 образцов) – наличием модулей A и C. Две мутации в гене *PSTPIP1* были связаны с модулями A/F и C/D. Наконец, мутации *TNFRSF1A* показали наибольшую вариабельность степени ассоциации с функциональными модулями. Из 13 образцов у пяти присутствовал модуль A, у четырех – модуль D, два образца характеризовались модулем F; а в двух образцах (*MVK* V377I / S329R, *TNFRSF1A* H22Y) связи с функциональными модулями обнаружено не было.

Все функциональные модули характеризовались нарушением активности ключевых путей врожденного иммунитета и воспаления, вовлеченных в развитие и прогрессию аутовоспаления (рис. 28). Полученные данные указывают также на изменения профиля активации PI3K и Ras-сигнальных путей, которые играют важную роль в регуляции иммунного ответа, генерации активных форм кислорода и инициации эндоплазматического стресса (Kastner et al., 2010; Xie et al., 2014) и могут способствовать как инициации, так и амплификации аутовоспалительного процесса.

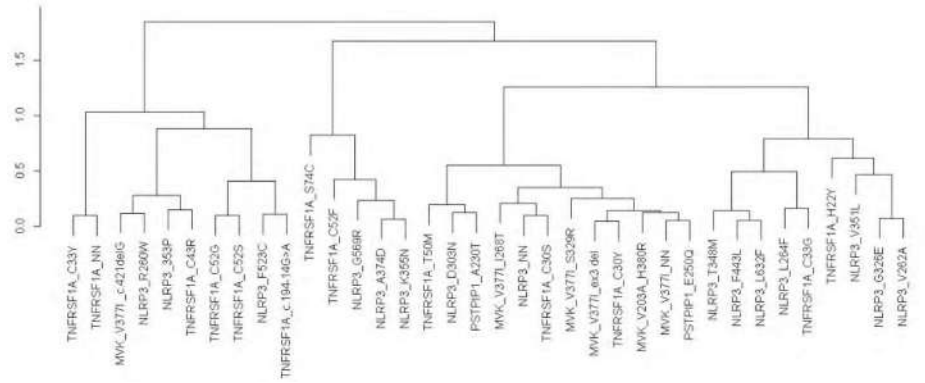

Рисунок 27. Кластерный анализ аутовоспалительных заболеваний по данным экспрессии генов.
В качестве метрики расстояния был использован коэффициент Пирсона. Построение дендрограммы было проведено методом полной связи (complete linkage).

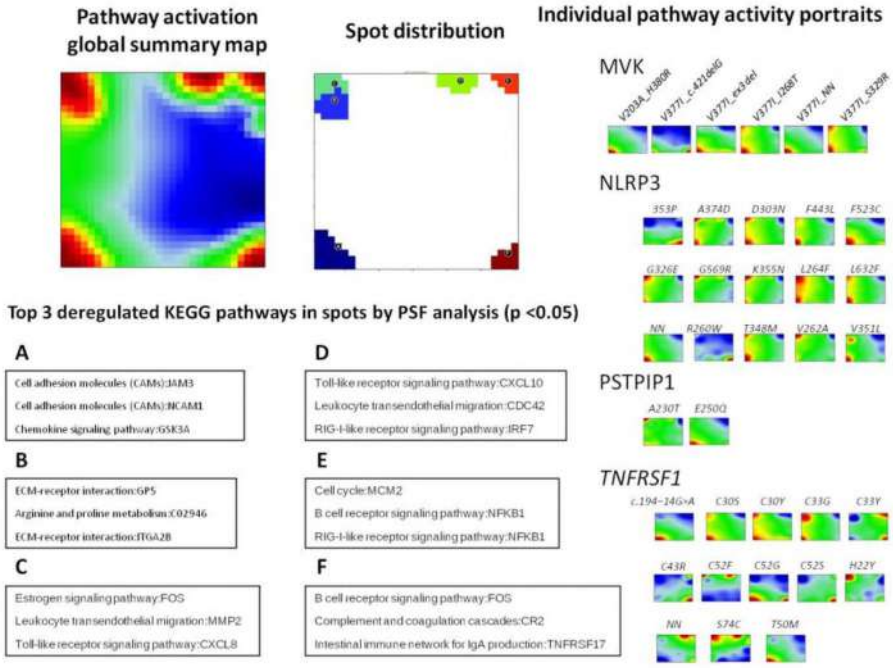

Рисунок 28. Портреты изменений активности биологических путей при аутоиммунных заболеваниях.
Слева представлены индивидуальные портреты для каждой мутации. Цветовая гамма от темно-синего (понижение) до темно красного (повышение) показывает уровень изменения активности биологического пути по сравнению с контролем.
Справа представлена сводная карта глобальных изменений профилей активации и их функциональная аннотация.

Значительные изменения активности были также отмечены в сигнальных путях Toll-like, Fc-gamma и B клеточных рецепторов, а также системы комплемента. Все эти пути являются хорошо известными триггерами аутовоспалительных сигналов (Masters et al., 2009; de Jesus et al., 2015). Кроме того, были отмечены значительные изменения активности метаболических путей, что могло привести к образованию вторичных метаболитов, таких как сукцинат и цитрат в цикле трикарбоновых кислот или арахидоновая кислота в метаболизме фосфолипидов (Fitzpatrick & Young, 2013). Наконец, мы отметили дерегуляцию в путях, которые обычно связаны с развитием опухолей, таких как Wnt-, Hedgehog- и Hippo, что соответствует известным представлениям о связи хронического воспаления с развитием опухолей (Del Prete et al., 2011). Кроме того, кластерный анализ, основанный на корреляции модулей активации биологических путей, показал более четкое разделение аутовоспалительных синдромов (рис. 29).

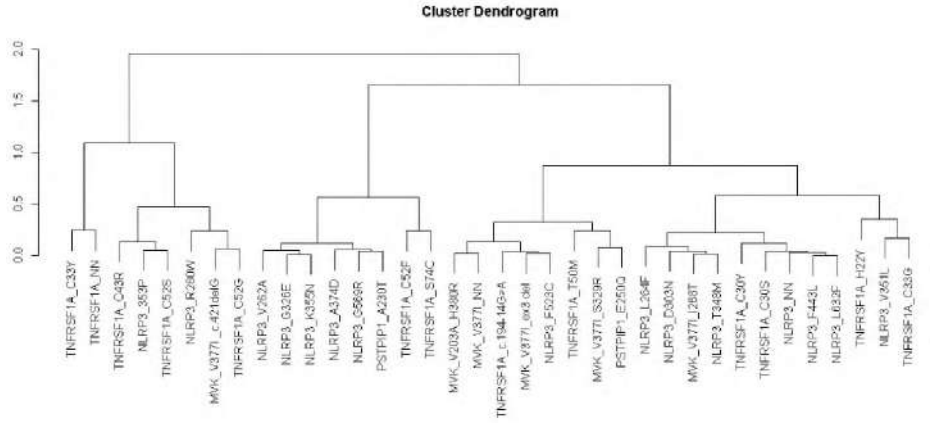

Рисунок 29. Кластерный анализ аутовоспалительных заболеваний по значениям активностей биологических путей.
В качестве метрики расстояния был использован коэффициент Пирсона. Построение дендрограммы было проведено методом полной связи (complete linkage).

Наше исследование показывает, что независимо от типа мутации, инициирующей развитие аутовоспалительного процесса, профили активации биологических путей в основном сходны при исследованных заболеваниях и затрагивают процессы, связанные с хроническим воспалением, изменением метаболизма и клеточной пролиферацией.

Анализ дифференциальной экспрессии генов при посттравматическом стрессовом расстройстве

Состояние проблемы и цель исследования

Молекулярные механизмы ПТСР мало изучены, однако полученные данные свидетельствуют о комплексной природе этой болезни, обусловленной взаимодействием генетических и средовых факторов, а также о том, что дисфункция нейро-эндокринной и иммунной системы могут быть в значительной степени ответственны за его развитие и течение (Koenen et al., 2008; Boyajyan et al., 2015). Большинство исследований патомеханизмов ПТСР направлены на определение генетических локусов, ассоциированных с предрасположенностью к развитию заболевания, однако наличие данных по анализу транскриптома сегодня дает возможность для более полного понимания молекулярных механизмов, лежащих в основе этиопатогенеза ПТСР.

В данном исследовании была поставлена задача провести анализ изменений транскриптома на ранних и поздних этапах развития ПТСР с помощью разработанного нами метода анализа дифференциальной экспрессии генов.

Использованные наборы данных и алгоритмы

В исследовании был использован набор данных, содержащий уровни экспрессии генов в мононуклеарных клетках крови у переживших травму 33 лиц, которые были распределены по группам ПТСР (n = 17) и контроля (n = 16) на основе диагностических критериев DSM IV через один и четыре месяца (Segman et al., 2005).

Измерение уровня глобальной экспрессии генов проводилось дважды: сразу после поступления в отделение неотложной помощи и через четыре месяца. Набор данных

(GDS1020) доступен в хранилище Gene Expression Omnibus (Services, 2007; Barrett et al., 2011; NCBI Resource Coordinators, 2016).

Для анализа экспрессии генов был использован разработанный нами алгоритм растущих опорных множеств. В качестве исходных данных были использованы лог-трансформированные значения экспрессии генов. Анализ транскриптома был проведен в двух подгруппах:

- сравнение транскриптома в группах ПТСР и контроля на момент поступления в отделение неотложной помощи (ER),
- сравнение транскриптома в группах ПТСР и контроля через четыре месяца (M4),

Порогом для включения генов в опорные множества была выбрана точность классификации в 90% случаев, а включения в кандидатные множества – 80%. Анализ белок-белковых взаимодействий и функциональная аннотация генов в опорных множествах была выполнена с помощью программы StringApp для Cytoscape (http://apps.cytoscape.org/apps/stringapp) и Webgestalt (B. Zhang et al., 2005) и InnateDB (Breuer et al., 2013).

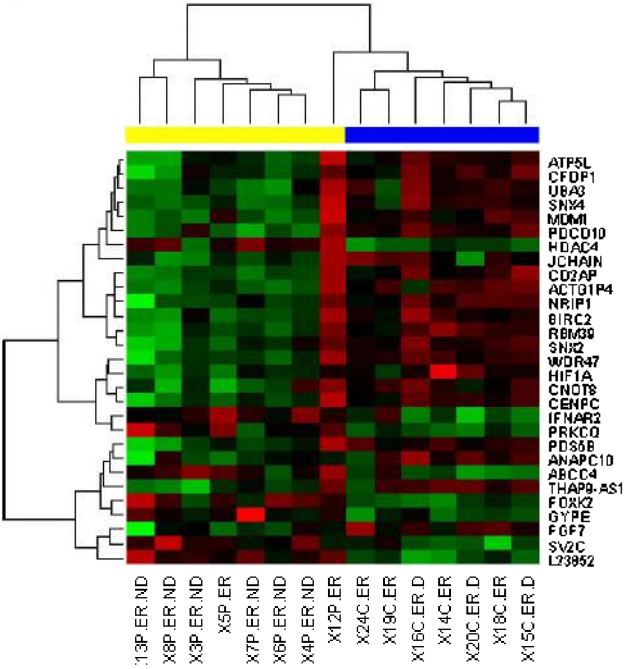

Рисунок 30. Кластерный анализ дифференциальной экспрессии генов опорных множеств в периферической крови на момент травмы.

Дифференциальная экспрессия генов при ПТСР

Анализ дифференциальной экспрессии генов в группах ПТСР и контроля на момент поступления в отделение неотложной помощи с использованием разработанного нами алгоритма позволил идентифицировать 53 опорных множества генов, позволивших с 90%-й точностью предсказать развитие ПТСР сразу после травмы (рис. 30). Три опорных множества состояли из одного гена, 50 – из пары генов. В целом опорные множества

содержали 29 генов (таблица 4). Функциональная аннотация генов показала, что непосредственно после травмы развитие ПТСР характеризуется повышением уровня экспрессии генов, ассоциированных с рецепторами факторов роста и трансферрина, что согласуется с ранее опубликованными результатами (Heinzelmann et al., 2014; M. Zhao et al., 2016).

Одновременно в группе ПТСР наблюдалось понижение уровня экспрессии генов, связанных с модификаций гистонов (*HDAC4*), клеточным ответом на интерфероны (*IFNAR2*) и на лекарственные препараты (*ABCC4, PRKCQ, HDAC4*). Примечательно также, что, согласно имеющимся данным, уровень экспрессии и метилирования, а также функциональные полиморфизмы гена *HDAC4* ассоциированы с развитием ПТСР у женщин (Zhao et al., 2016).

Таблица 4.

Гены опорных множеств классификации ПТСР.

Символ	Описание	t.stat	t.pval
L23852	Cluster Incl L23852:Homo sapiens (clone Z146) retinal mRNA	-5.09	0.000211
WDR47	WD repeat domain 47	2.62	0.030195
HDAC4	histone deacetylase 4	-4.20	0.001741
ACTG1P4	actin gamma 1 pseudogene 4	2.70	0.020037
ABCC4	ATP binding cassette subfamily C member 4	-3.59	0.003304
MDM1	Mdm1 nuclear protein	1.73	0.117897
GYPE	glycophorin E (MNS blood group)	-2.98	0.012893
IFNAR2	interferon alpha and beta receptor subunit 2	-3.23	0.007935
PRKCQ	protein kinase C theta	-3.51	0.006066
CENPC	centromere protein C	2.08	0.069033
SV2C	synaptic vesicle glycoprotein 2C	-3.28	0.005978
FOXK2	forkhead box K2	-4.04	0.002361
FGF7	fibroblast growth factor 7	2.83	0.014866
CD2AP	CD2 associated protein	2.50	0.02766
PDCD10	programmed cell death 10	2.75	0.021655
BIRC2	baculoviral IAP repeat containing 2	2.27	0.044572
NRIP1	nuclear receptor interacting protein 1	2.84	0.017403
PDS5B	PDS5 cohesin associated factor B	2.23	0.055347
RBM39	RNA binding motif protein 39	2.99	0.011698
UBA3	ubiquitin like modifier activating enzyme 3	2.01	0.068687
HIF1A	hypoxia inducible factor 1 alpha subunit	2.71	0.022162
JCHAIN	joining chain of multimeric IgA and IgM	1.44	0.1733

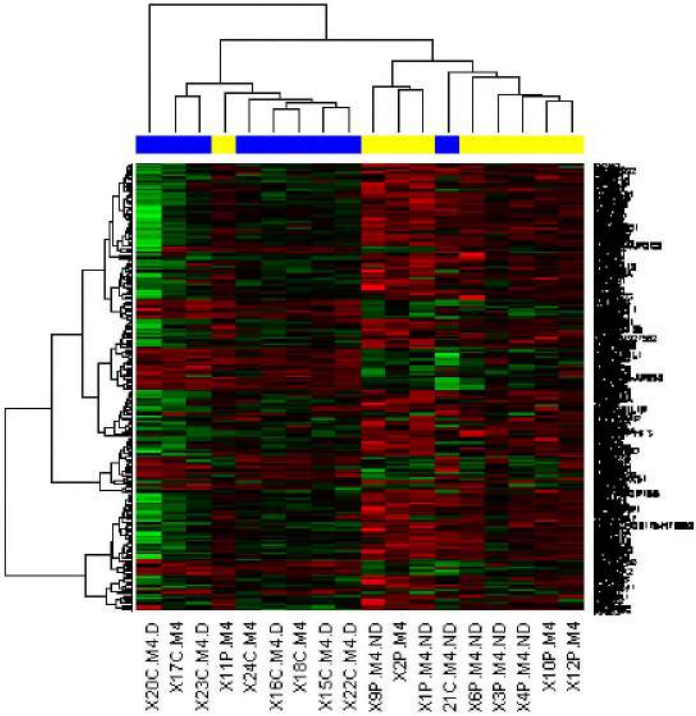

Рисунок 31. Кластерный анализ дифференциальной экспрессии генов опорных множеств в периферической крови спустя 4 месяца после травмы.

Однако мы не смогли обнаружить достаточно большие группы генов, которые объединяются в функциональные категории генов и могут свидетельствовать о конкретных молекулярных механизмах развития болезни. Это может быть связано с тем, что анализ транскриптома был проведен на клетках периферической крови, несмотря на их очевидную второстепенную роль в патогенезе ПТСР. В то же время показано, что изменения глобальной экспрессии генов в клетках периферической крови наблюдаются как при остром эмоциональном стрессе (Aloe et al., 1994), так и при различных психических заболеваниях (Rollins et al., 2010; Munkholm et al., 2015; Xu et al., 2016). Тем не менее, идентифицированные группы генов могут служить информативными маркерами развития ПТСР, что дает основу для разработки методов ранней диагностики.

Вместе с тем, гораздо более серьезные и видимые изменения транскриптома в периферических клетках наблюдаются на более поздних этапах, когда сформировался клинический фенотип заболевания. Анализ транскриптома спустя четыре месяца позволил идентифицировать уже 5918 опорных множеств генов, позволяющих с 90%-й точностью классифицировать группы ПТСР и контроля (рис. 31). Пять множеств состояли из одного гена, 1610 – из двух и 4303 – из трех генов.

Всего опорные множества были составлены из 406 генов. Их функциональная аннотация показала, что они вовлечены в целый спектр биологических процессов – сигнальных, регуляторных и метаболических путей (таблица 5).

Таблица 5.

Идентифицированные функциональные категории, связанные с дифференциальной экспрессией генов в клетках периферической крови на позднем этапе развития ПТСР

Биологический путь	Кол-во генов	P	FDR
Regulation of bad phosphorylation	4	0.0002	0.0462
Integrin cell surface interactions	6	0.0008	0.0563
Apoptosis	6	0.0010	0.0657
Beta2 integrin cell surface interactions	4	0.0005	0.0777
Deregulation of cdk5 in alzheimers disease	2	0.0025	0.0784
ALK1 signaling events	3	0.0022	0.0793
Androgen Receptor	7	0.0059	0.0866
C-MYB transcription factor network	5	0.0036	0.0880
Terminal pathway of complement	2	0.0035	0.0896
Cell cycle	6	0.0064	0.0899
GPCR signaling	10	0.0053	0.0931
Alternative complement pathway	2	0.0072	0.0964
Classical complement pathway	2	0.0142	0.0986
Postsynaptic nicotinic acetylcholine receptors	2	0.0142	0.0986
Syndecan-1-mediated signaling events	2	0.0142	0.0986
Caspase-mediated cleavage of cytoskeletal proteins	2	0.0104	0.0997
Lectin induced complement pathway	2	0.0104	0.0997
RNA Polymerase II Pre-transcription Events	4	0.0084	0.1001
Jak-STAT signaling pathway	6	0.0180	0.1003
Signal transduction through il1r	3	0.0108	0.1007
Chromatin remodeling by hswi/snf atp-dependent complexes	2	0.0184	0.1007
IL6-mediated signaling events	3	0.0188	0.1008
Ion transport by P-type ATPases	3	0.0188	0.1008
B cell survival pathway	2	0.0122	0.1014
AP-1 transcription factor network	4	0.0117	0.1024
Complement and coagulation cascades	4	0.0129	0.1029
GPCR Adenosine A2A receptor signaling pathway	3	0.0116	0.1065
Il-2 receptor beta chain in t cell activation	3	0.0266	0.1268
Tgf beta signaling pathway	2	0.0281	0.1271
IL3-mediated signaling events	2	0.0308	0.1337

ВЫВОДЫ

1. Разработан комплекс алгоритмов и программных пакетов для анализа и интерпретации данных о глобальной экспрессии генов, который был использован при изучении молекулярных механизмов патогенеза широкого спектра хронических заболеваний человека.
2. Для идентификации дифференциально экспрессируемых генов и их групп разработан алгоритм «растущих опорных множеств», основанный на методах «добычи данных» и

выделения признаков, что позволяет распознать биомаркеры развития патологических состояний, а также реконструировать и выявлять новые биологические пути.

3. Предложены программные пакеты для редактирования и визуализации биологических путей KEGG Pathways, в которые впервые внедрена возможность полуавтоматической коррекции взаимодействий между компонентами биологических путей, модификации по тканевой специфичности и белок-белковым взаимодействиям, что существенно повысило точность результатов биоинформатического анализа и их интерпретацию.
4. Разработаны алгоритм и ряд программных пакетов, позволяющих осуществить непосредственный переход от погенного метода анализа дифференциальной экспрессии генов к анализу на уровне сигнальных путей, что значительно улучшает точность определения биологических процессов, вовлеченных в развитие патологических состояний.
5. Сигнальные и метаболические пути устойчивы к мутациям, влияющим на белок-белковые взаимодействия благодаря множественным дублирующим ответвлениям и разветвленным топологиям. В то же время они могут содержать узлы-концентраторы и «узкие» места, мутации в которых могут существенно влиять на общую активность данного пути.
6. Рак легких характеризуется активацией биологических путей, связанных с пролиферацией клеток и ускорением метаболизма, а хронические заболевания легких сопровождаются изменениями, связанными с иммунным ответом и фиброзным ремоделированием легочной ткани.
7. В основе патогенеза аутоиммунных и аутовоспалительных заболеваний лежит смещение баланса между дисфункцией иммунного и воспалительного ответов; патогенез ряда аутоиммунных заболеваний характеризуется вовлечением биологических путей, связанных с аутовоспалением.
8. Генетическая гетерогенность и различия в экспрессии генов при моногенных аутовоспалительных синдромах практически не отражаются на профиле активации биологических путей, связанных с иммунным и воспалительным ответами, что может объяснить сходную клиническую картину при разных мутациях, затрагивающих один и тот же ген.
9. Привлечение методов биоинформатики и вычислительной биологии к анализу больших объемов экспериментальных данных, основанных на использовании микрочипов и секвенирования следующего поколения, позволяет всесторонне исследовать молекулярные механизмы функционирования живых систем на всех уровнях их организации в норме и патологии.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Loeffler-Wirth H, Kreuz M, Hopp L, **Arakelyan A**, Haake A, Cogliatti SB, Feller AC, Hansmann ML, Lenze D, Möller P, Müller-Hermelink HK, Fortenbacher E, Willscher E, Ott G, Rosenwald A, Pott C, Schwaenen C, Trautmann H, Wessendorf S, Stein H, Szczepanowski M, Trümper L, Hummel M, Klapper W, Siebert R, Loeffler M, Binder H; German Cancer Aid consortium Molecular Mechanisms for Malignant Lymphoma. A modular transcriptome map of mature B cell lymphomas. // Genome Med. 2019; 11(1):27. doi:10.1186/s13073-019-0637-7.

2. Hopp L, Loeffler-Wirth H, Nersisyan L, **Arakelyan A**, Binder H. Footprints of sepsis framed within community acquired pneumonia in the blood transcriptome. // Front Immunol. 2018; 9:1620. doi:10.3389/fimmu.2018.01620.
3. Çakır MV, Loffler-Wirth H, **Arakelyan A.**, Binder H. Dysregulated signal propagation in a MYC-associated Boolean gene network in B-cell lymphoma. // Biol Eng Med. 2017; 2:2. doi:10.15761/BEM.1000115.
4. Binder H, Hopp L, Schweiger MR, Hoffmann S, Jühling F, Kerick M, Timmermann B, Siebert S, Grimm C, Nersisyan L, **Arakelyan A**, Herberg M, Buske P, Loeffler-Wirth H, Rosolowski M, Engel C, Przybilla J, Peifer M, Friedrichs N, Moeslein G, Odenthal M, Hussong M, Peters S, Holzapfel S, Nattermann J, Hueneburg R, Schmiegel W, Royer-Pokora B, Aretz S, Kloth M, Kloor M, Buettner R, Galle J, Loeffler M. Genomic and transcriptomic heterogeneity of colorectal tumours arising in Lynch syndrome. // J Pathol. 2017; 243(2):242-254. doi: 10.1002/path.4948.
5. **Arakelyan A**, Nersisyan L, Poghosyan D, Khondkaryan L, Hakobyan A, Löffler-Wirth H, Melanitou E, Binder H. Autoimmunity and autoinflammation: A systems view on signaling pathway dysregulation profiles. // PLoS One. 2017; 12(11):e0187572. doi: 10.1371/journal.pone.0187572.
6. **Arakelyan A**. Analysis of biological pathway activation profiles in monogenic autoinflammatory disorders. // Electronic Journal of Natural Sciences. 2017; 1(28):30-34.
7. **Arakelyan A**. Topology dependent pathway tolerance to mutations. // Electronic Journal of Natural Sciences. 2017; 1(28):25-29.
8. **Arakelyan A**. Functional gene sets in post-traumatic stress disorder. // Proceedings of the Yerevan State University. 2016; 1:43-48
9. **Arakelyan A**. Analysis of somatic mutation enrichment in gene expression landscapes for cancer cell lines. // Biolog. Journal of Armenia. 2016; 3(68):54-58.
10. **Arakelyan A**. On importance of topology on functional annotation on biological pathways in gene expression experiments. // IJITK. 2016; 10(1):82-90.
11. **Arakelyan A**, Nersisyan L, Petrek M, Löffler-Wirth H, Binder H. Cartography of pathway signal perturbations identifies distinct molecular pathomechanisms in malignant and chronic lung diseases. // Front. Genet. 2016; 7:79. doi:10.3389/fgene.2016.00079.
12. **Arakelyan A**, Nersisyan L, Hakobyan A. Application of MATLAB in -omics and systems biology. Applications from engineering with MATLAB concepts. Jan Valdman (Ed.), InTech, Croatia. 2016. p.171-187. doi: 10.5772/62847. ISBN: 978-953-51-2459-7.
13. Hopp L, Nersisyan L, Löffler-Wirth H, **Arakelyan A**, Binder H. Epigenetic heterogeneity of B-cell lymphoma: Chromatin modifiers. // Genes. 2015; 6(4):1076-1112. doi:10.3390/genes6041076.
14. Nersisyan L, Johnson G, Riel-Mehan M, Pico A, **Arakelyan A**. PSFC: a pathway signal flow calculator app for cytoscape. // F1000Res. 2015; 4:480. doi:10.12688/f1000research. 6706.1.
15. Nersisyan L, Hakobyan A, **Arakelyan A**. Telomere-associated gene network in lung adenocarcinoma. // Eur Respir J. 2015; 46(suppl 59). doi: 10.1183/13993003.congress-2015.OA3493.
16. Arakelyan A, Nerisyan L, Gevorgyan A, **Boyajyan A**. Geometric approach for Gaussian-kernel bolstered error estimation for linear classification in computational biology. // IJITA. 2014; 21(2):170-181.
17. Nersisyan L, Löffler-Wirth H, **Arakelyan A**, Binder H. Gene set- and pathway-centered knowledge discovery assigns transcriptional activation patterns in brain, blood, and colon cancer: a bioinformatics perspective. // IJKDB. 2014; 4(2):46-69. doi:10.4018/ IJKDB.2014070104.

18. Binder H, Wirth H, **Arakelyan A**, Lembcke K, Tiys ES, Ivanisenko VA, Kolchanov NA, Kononikhin A, Popov I, Nikolaev EN, Pastushkova L, Larina IM. Time-course human urine proteomics in space-flight simulation experiments. // BMC Genomics. 2014; 15(Suppl 120:S2. doi:10.1186/1471-2164-15-S12-S2.
19. Nersisyan L, Samsonyan R, **Arakelyan A**. CyKEGGParser: tailoring KEGG pathways to fit into systems biology analysis workflows. // F1000Res. 2014; 3:145. doi:10.12688/f1000research.4410.2.
20. Nersisyan L, Wirth H, Gevorgyan A, Binder H, **Arakelyan A**. Methylation associated pathway activity deregulation in lung adenocarcinoma. // Eur Respir J. 2014; 44(Suppl. 58):403.
21. **Arakelyan A**, Nersisyan L, Wirth H, Binder H. Mining common pathway deregulation profiles in lung diseases. // Eur Respir J. 2014; 44(Suppl. 58):2021.
22. **Arakelyan A**, Nersisyan L. KEGGParser: parsing and editing KEGG pathway maps in Matlab. // Bioinformatics. 2013; 29(4):518-519. doi:10.1093/bioinformatics/bts730.
23. **Arakelyan A**, Aslanyan L, Boyajyan A. Gene expression data analytics. Proceedings of the IX International Conference "Computer Science and Information Technologies", Yerevan, September 23-27, 2013; p.174-176.
24. **Arakelyan A**, Aslanyan L, Boyajyan A. On knowledge-based gene expression data analysis. // IEEE Xplore. 2013; 1-6. doi:10.1109/CSITechnol.2013.6710349.
25. **Arakelyan A**, Aslanyan L, Boyajyan A. High-throughput gene expression analysis concepts and applications. In: Genomics II - Bacteria, Viruses and Metabolic Pathways. iConcept Press Ltd., USA 2013. p.71-95. ISBN:978-1-922227-04-1.
26. **Arakelyan A.** Molecular pathway perturbations in pulmonary sarcoidosis. // Eur Respir J. 2013; 42(Suppl. 57):540.
27. **Arakelyan A.**, Wirth H., Cakir V., Binder H. Molecular pathway perturbations in B-cell lymphoma subtypes. // Biolog. Journal of Armenia. 2013; 65(suppl 1):35-36.
28. **Arakelyan A**. Pathway signal flow analysis for high-throughput gene expression data. Abstracts of the VIII International Conference "Bioinformatics of genome regulation and structure/system biology". Novosibirsk, Russia. 2012, p.41.
29. **Arakelyan A**. Growing support sets for pathway specific analysis of sarcoidosis. Abstracts of the II International Conference "Postgenomic methods of analysis in biology and clinical medicine: genomics, proteomics and bioinformatics". Novosibirsk, Russia. 2011, p.113.
30. **Arakelyan A**, Boyajyan A, Sahakyan H, Aslanyan L, Ivanova K, Mitov I. Growing support set systems in analysis of high-throughput gene expression data. // New trends in classification and data mining. ITHEA. 2010:47-53.
31. **Arakelyan A**, Boyajyan A, Aslanyan L. Algorithmic analysis of functional pathways affected in post-traumatic stress disorder. Abstracts of the YSS "Bioinformatics and systems biology", Novosibirsk, Russia. 2010, p.16.
32. **Arakelyan A**, Boyajyan A, Aslanyan L, Muradyan D, Sahakyan H Algorithmic analysis of functional pathways affected by typical and atypical antipsychotics Proceedings of the Conference "Computer Science and Information Technologies", Yerevan, 2009, p361-363.
33. **Arakelyan A**, Boyajyan A, Aslanyan L, Muradyan D, Chavushyan A, Hovsepyan T, Nersisyan L. Functional gene sets in posttraumatic stress disorder: analysis of disease related gene expression Proceedings of the International Conference "Biotechnology and Health-3", Yerevan. 2009, p57-61.
34. Boyajyan A, Hovhannisyan L, Mkrtchyan G, Sukiasian S, Ayvazyan V, **Arakelyan A**, Arakelova E, Khoyetsyan A, Manukyan L, Tsakanova G, Avetisyan G Immune system abnormalities in posttraumatic stress disorder Proceedings of the International Conference "Biotechnology and Health-2", Yerevan. 2008, p41-46.

ԱՐՍԵՆ ԱՐՏԱՇԵՍԻ ԱՌԱՔԵԼՅԱՆ

ԿԵՆՍԱԻՆՖՈՐՄԱՏԻԿԱԿԱՆ ՄՈՏԵՑՈՒՄՆԵՐԻ ՄՇԱԿՈՒՄ ՄԱՐԴՈՒ ՔՐՈՆԻԿ ՀԻՎԱՆԴՈՒԹՅՈՒՆՆԵՐԻ ԶԱՐԳԱՑՄԱՆ ՄՈԼԵԿՈՒԼԱՅԻՆ ՄԵԽԱՆԻԶՄՆԵՐԻ ՈՒՍՈՒՄՆԱՍԻՐՈՒԹՅԱՆ ՀԱՄԱՐ

Ամփոփում

Առանքային բառեր: մոլեկուլային մեխանիզմներ, գենի էքսպրեսիա, կենսաբանական ուղիների կենսաինֆորմատիկա, գենոմիկա, քրոնիկ հիվանդություններ

Քրոնիկ ոչ վարակիչ հիվանդությունները հանդիսանում են մարդկության մահացության հիմնական պատճառն ամբողջ աշխարհում՝ հանգեցնելով տնտեսական և ժողովրդագրական ծանր կորուստների: Այդ պատճառով այս հիվանդությունների կանխարգելման, ախտորոշման և բուժման նոր մոտեցումների մշակումը հանդիսանում է կենսաբժշկության և առողջապահության գերակա խնդիրներից մեկը:

Հիվանդությունների զարգացման մոլեկուլային մեխանիզմների ուսումնասիրության ոլորտում ՌՆԹ քանակական վերլուծությունն առավել արդյունավետ և տարածված մեթոդն է՝ հաշվի առնելով դրա համեմատաբար պարզությունը, ինչպես նաև բջջում ՌՆԹ-ի և սպիտակուցային մակարդակների հարաբերական կոռելացիան:

Գեների գլոբալ էքսպրեսիայի չափման եղանակների մշակմանը զուգահեռ՝ ստեղծվել են որակապես նոր հնարավորություններ կենդանի օրգանիզմում բջիջների, բջջային պոպուլյացիաների, հյուսվածքների և այլ մակարդակներում ֆիզիոլոգիական և պաթոֆիզիոլոգիական գործընթացների համապարփակ ուսումնասիրության համար: Այս նոր մոտեցումները հատկապես արժեքավոր են պոլիգեն, բազմագործոն հիվանդությունների հետազոտությունների համար: Այսօր գեների գլոբալ էքսպրեսիայի արդեն իսկ կուտակված հսկայածավալ տվյալները առաջ են քաշում նոր խնդիր՝ մշակել կենսաինֆորմատիկական վերլուծության մեթոդներ, որոնք բարձրարդյունավետ տրանսկրիպտոմային հետազոտությունների արդյունքները համարժեքորեն կփոխարկեն հիվանդությունների զարգացման մոլեկուլային մեխանիզմների վերաբերյալ գիտելիքների:

Այսպիսով, ներկայացվող աշխատանքի նպատակն է մշակել գեների գլոբալ էքսպրեսիայի և կենսաբանական ուղիների ակտիվության վերլուծության կենսաինֆորմատիկական ալգորիթմներ և ծրագրային փաթեթներ՝ մարդու քրոնիկ ոչ վարակիչ հիվանդությունների զարգացման մոլեկուլային մեխանիզմների ուսումնասիրության համար:

Աշխատանքում իրականացվել է տրանսկրիպտոմի հետազոտության հիմնական ձևերի տեսական վերլուծություն: Կատարվել է գեների էքսպրեսիայի մոդելի տեսական վերլուծություն և ցույց է տրվել, որ էքսպրեսիայի հավանական և բազմագործոն բնույթը ֆիզիոլոգիական և ախտաբանական վիճակներում, ինչպես նաև ընտրանքի փոքր ծավալը հանդիսանում են մի կողմից դասական վիճակագրական և մեքենայական ուսուցման մեթոդների օգտագործումը

սահմանափակող, իսկ մյուս կողմից նոր ալգորիթմական մոտեցումների մշակում պահանջող գործոններ:

Մեքենայական ուսուցման և պատկերների ճանաչողության մեթոդների և օրինաչափությունների հիման վրա մշակվել է «աճող օժանդակ հավաքածուների» ալգորիթմ դիֆերենցիալ գեների վերլուծության համար:

Մշակվել են մի շարք ծրագրային փաթեթներ և ալգորիթմներ, որոնք հնարավորություն են ընձեռում անցում կատարել եզակի գենի էքսպրեսիայի վերլուծությունից դեպի կենսաբանական ուղիների մակարդակում հետազոտություններ: Մասնավորապես՝ մշակվել են KEGGParser և CyKEGGParser ծրագրային փաթեթները, որոնք թույլ են տալիս ստեղծել կենսաբանական ուղիների գրադարաններ KEGG Pathway տվյալների շտեմարանում պահվող տեղեկատվության հիման վրա: Այս ծրագրերի միջոցով ստեղծվել են ազդանշանային, նյութափոխանակության և կարգավորող ուղիների հավաքածու, որն օգտագործվել է մեր հետագա ուսումնասիրություններում:

Կենսաբանական ուղիներում ազդանշանային հոսքերի մոդելավորման համար մշակվել է «Pathway Signal Flow» ալգորիթմը, որը կենսաբանական ուղիների տոպոլոգիայի և գեների էքսպրեսիայի տվյալների հիման վրա թույլ է տալիս գնահատել կենսաբանական ուղու ակտիվությունն օրգանիզմի ֆիզիոլոգիական և տարբեր ախտաբանական վիճակներում:

Օգտագործելով մշակված ալգորիթմները և ծրագրային փաթեթները՝ ուսումնասիրվել են մի շարք հիվանդությունների զարգացման մոլեկուլային մեխանիզմները: Այսպես, ցույց է տրվել կենսաբանական ուղիների կայունության կախվածությունը մուտացիաների նկատմամբ, ինչպես նաև բացահայտվել են այն հիմնական գեները, որոնց մուտացիաները կարող են էապես ազդել բջիջների բնականոն գործունեության վրա:

Կենսաբանական ուղիների ակտիվության վերլուծությունը թույլ է տվել նույնականացնել թոքային հիվանդությունների նոր մոլեկուլային ենթատիպեր, որոնք բնութագրվում են սուր և քրոնիկ բորբոքային պատասխանի, հյուսվածքի ֆիբրոզ ձևափոխման, բջջային պրոլիֆերացիայի և ապոպտոզի խանգարման տարբեր աստիճաններով: Բացի այդ, ցույց է տրվել, որ թոքերի հյուսվածքի ֆիբրոզ ձևափոխման հետ կապված կենսաբանական ուղիները զգալի դեր են խաղում թոքերի քաղցկեղի պաթոգենեզում:

Պոլի- և մոնոգեն աուտոիմունային և աուտոբորբոքային հիվանդությունների զարգացման պաթոմեխանիզմների ուսումնասիրությունը ցույց է տվել, որ այս հիվանդությունները բնութագրվում են իմունային և բորբոքային պատասխանի ուղիների ներգրավվածության տարբեր աստիճաններով: Բացի այդ, ցույց է տրվել, որ մի շարք աուտոիմունային հիվանդությունների պաթոգենեզը բնութագրվում է աուտոբորբոքային պատասխանի հետ ասոցացված կենսաբանական ուղիների ներգրավմամբ:

Պարզվել է, որ մոնոգեն աուտոբորբոքային հիվանդություններում ազդանշանային ուղիների ակտիվության խանգարումների դինամիկան ավելի ճիշտ է արտացոլում այդ ախտաբանությունների հիմքում ընկած պաթոգենետիկական օրինաչափությունները, քան՝ առանձին գեների էքսպրեսիայի փոփոխությունները:

Բացահայտվել է, որ հետտրավմատիկ սթրեսային խանգարման սկզբնական փուլում ծայրամասային արյան բջիջներում գեների էքսպրեսիան հանդիսանում է հիվանդության զարգացման կենսամարկեր, իսկ հետագա փուլերում՝ կապված է հիմնական նյարդահոգեբուժական չափորոշիչների հետ:

Ստացված արդյունքները զգալիորեն հարստացնում են մի շարք քրոնիկ ոչ վարակիչ հիվանդությունների պաթոգենետիկական մեխանիզմների վերաբերյալ պատկերացումները: Մշակված ալգորիթմները և ծրագրային փաթեթները, որոնք արդեն իսկ հաջողությամբ կիրառվում են կենսաբանության ոլորտի հետազոտություններում, թույլ են տալիս իրականացնել կենսաբանական ուղիների խանգարումների մակարդակում համակարգային վերլուծություն, որն առավել արդյունավետ մոտեցում է հիվանդությունների պաթոմեխանիզմների ուսումնասիրության համար, քան գեների դիֆերենցիալ էքսպրեսիայի վերլուծությունը:

ARSEN ARTASHES ARAKELYAN

DEVELOPMENT OF BIOINFORMATIC APPROACHES FOR STUDYING THE MOLECULAR MECHANISMS OF THE CHRONIC HUMAN DISEASES

Summary

Key words: molecular mechanisms, gene expression, biological pathways bioinformatics, genomics, chronic diseases

Chronic noncommunicable diseases (NCDs) are the leading cause of mortality and are responsible for significant demographic and economic losses worldwide. Therefore, the development of prognostic, diagnostic and therapeutic approaches for prevention and control of NCDs is considered a priority challenge for biomedicine and healthcare. In this regard, the understanding of regulatory mechanisms that underlie disease development and progression is an important driver for progress in these fields.

Quantitative analysis of RNA levels (gene expression or transcriptomics) has become a method of choice for evaluation of disease development and progression mechanisms due to simplicity of the measurements and considerable correlation with protein levels. With the advent of high-throughput transcriptomics methods new opportunities have emerged for comprehensive understanding of physiological and pathophysiological processes that occur in a living organism at the level of cells, cell populations, tissues, etc. These approaches have become especially useful in studies of the molecular mechanisms of polygenic complex diseases. To date, huge amounts of transcriptome measurement data have already accumulated. This, in turn, raises the need for developing new bioinformatics analysis approaches that adequately translate the results of high-throughput experiments into knowledge about molecular mechanisms of disease development.

This study was aimed at developing bioinformatics algorithms and software packages for analysis of global gene expression and biological pathway activity to disentangle the molecular mechanisms of development of NCDs.

In the framework of the study, formal analysis of current designs of transcriptome analysis experiments as well as theoretical analysis of the probabilistic model of gene expression was performed. It has been shown that the stochastic and multifactorial nature of gene expression in normal and pathological conditions, as well as high dimensionality of the transcriptome data

coupled with small sample sizes, limit the use of standard statistical methods and machine learning algorithms thus justifying the need for new data analysis approaches.

Here, several algorithms and software packages addressing this challenge have been developed: the growing support set algorithm for robust identification of differentially expressed genes; the software packages KEGGParser and CyKEGGParser for parsing, visualization, and analysis of KEGG pathways that were utilized for generation of a collection of signaling, metabolic, and regulatory pathways used in downstream analysis; and finally, the "Pathway Signal Flow" (PSF) algorithm for modeling the propagation of signal flows in biological pathways based on pathway topology and gene expression data, used for assessment of pathway activity deregulations in different conditions, for simulation studies on network dynamics, etc.

The bioinformatics algorithms and software packages developed herein were used to study the molecular mechanisms of development of several complex human diseases.

It was shown that due to complex branched topologies signaling pathways are robust to mutations that alter protein-protein interactions. On the other hand, pathways may contain hubs and bottleneck genes, whereby mutations may cause large perturbations and significantly affect the overall activity of a given pathway and normal function of a cell.

The systems-levels analysis of lung disease datasets identified three novel molecular subtypes of interstitial lung diseases characterized by different levels of involvement of pathways associated with immune/inflammatory response and fibrotic tissue remodeling. Furthermore, our analysis showed that lung cancers were characterized by pathways implicated in cell proliferation, metabolism, while non-malignant lung diseases were characterized by deregulations in pathways involved in inflammation and fibrosis.

The studies of molecular mechanisms of autoimmune and autoinflammatory diseases revealed that clinically divergent disease groups are characterized by different levels of involvement of immune and inflammatory response-related pathways. Further, it was shown that autoinflammatory processes are implicated in pathophysiology of several autoimmune diseases.

Analysis of transcriptome and pathway deregulation profiles in monogenic autoinflammatory syndromes demonstrated that regardless of the initiating mutation event, the downstream pathway activity deregulations are mainly shared, while gene expression profiles showed considerable heterogeneity.

Our study on post-traumatic stress disorder (PTSD) showed that gene expression signatures in peripheral blood cells are informative prognostic markers and, in the late stages of the disease, correlate with the main neuropsychiatric parameters.

Overall, the results of these studies significantly enrich the current understanding of the pathogenesis of chronic and oncological lung diseases, PTSD, autoimmune and autoinflammatory diseases. Software packages and algorithms developed in the framework of this study allow for performing systems biology level analyses and offer a more efficient way to study disease molecular mechanisms compared to more common single-gene level analyses.