DOI 10.23946/2500-0764-2017-2-2-77-82

СОВРЕМЕННЫЕ ТЕНДЕНЦИИ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ И ПРЕДСТАВЛЕНИЯ РЕЗУЛЬТАТОВ В КАНДИДАТНЫХ ГЕНЕТИКО-ЭПИДЕМИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

КУТИХИН А.Г.[1], ЮЖАЛИН А.Е.[2], ПОНАСЕНКО А.В.[1]

[1]ФГБНУ «Научно-исследовательский институт комплексных проблем сердечно-сосудистых заболеваний», г. Кемерово, Россия
[2]Оксфордский институт радиационной онкологии, Оксфордский университет, г. Оксфорд, Соединенное Королевство Великобритании и Северной Ирландии

LECTURE

HOW TO ANALYZE AND PRESENT GENETIC EPIDEMIOLOGY DATA IN CANDIDATE STUDIES

ANTON G. KUTIKHIN[1], ARSENIY E. YUZHALIN[2], ANASTASIA V. PONASENKO[1]

[1]Research Institute for Complex Issues of Cardiovascular Diseases (6, Sosnovy Boulevard, Kemerovo, 650002), Russian Federation
[2]Department of Oncology, Cancer Research UK and Medical Research Council Oxford Institute for Radiation Oncology, University of Oxford (Old Road Campus Research Building, Roosevelt Drive, Oxford OX3 7DQ), United Kingdom

Резюме

В статье излагаются основные положения текущих трендов относительно статистической обработки данных и представления результатов в кандидатных генетико-эпидемиологических исследованиях. Описывается методология генетической эпидемиологии и три основных этапа генетико-эпидемиологических исследований, выполняемых с помощью кандидатного подхода: 1) разработка дизайна исследования; 2) получение биоматериала, выделение ДНК и генотипирование; 3) статистический анализ и представление результатов. Кроме того, описываются основные аспекты таких исследований с позиции рецензента.

Ключевые слова: генетическая эпидемиология, генетико-эпидемиологическое исследование, кандидатные гены, генные полиморфизмы, статистический анализ, представление результатов.

Abstract

Here we describe recent trends in statistical analysis and data presentation in candidate genetic association studies. We first discuss methods of genetic epidemiology following talking about the three key steps in candidate genetic association studies: 1) study design; 2) isolation of biomaterial, DNA extraction and genotyping; 3) statistical analysis and data presentation. In addition, we consider the crucial issues of these studies from the reviewer's point of view.

Keywords: genetic epidemiology, genetic association studies, candidate genes, gene polymorphisms, statistical analysis, data presentation.

Методология генетической эпидемиологии

Стремительное развитие молекулярно-генетических методов исследования в последние

три десятилетия XX века и расшифровка генома человека в 2003 году привели к возможности изучения на больших выборках связи отдельных вариантов нуклеотидной последовательности внутри генов (так называемых генных полиморфизмов, в русскоязычной терминологии также принят термин «вариабельные сайты») с различными заболеваниями и обусловливающими их патологическими процессами. Работы подобного рода всегда являются междисциплинарными, так как для подбора необходимых биомаркеров требуются специфические знания в патологии и областях клинической медицины, для определения полиморфизмов применяются технологии генотипирования, а для разработки дизайна исследования и статистического анализа используются эпидемиологические методы. Поэтому раздел биомедицинской науки, изучающий обозначенную проблему и занимающийся подобными работами, наиболее правильно называть *генетической эпидемиологией*. Данное понятие является общепринятым как в англоязычной, так и в русскоязычной терминологии.

Генетико-эпидемиологические исследования выполняются при помощи двух противоположных подходов: 1) *кандидатного*, при котором тестируются только изначально определенные исследователем гены и полиморфизмы; 2) *полногеномного*, при котором поиск генетических ассоциаций с тем или иным состоянием осуществляется по всему спектру полиморфизмов в геноме. Важными преимуществами кандидатного подхода являются относительная доступность метода, возможность разработки оригинальных алгоритмов выбора генных полиморфизмов для анализа, а также возможность свободного варьирования стоимости исследования (от сотен тысяч рублей до неограниченных пределов). В то же время кандидатный подход характеризуется относительно слабой воспроизводимостью результатов выполненных при помощи него исследований и позволяет охватить лишь малое количество (до полусотни) генных полиморфизмов, которые могут быть ответственны за развитие той или иной патологии. Оба данных недостатка чрезвычайно сложно устранимы в реальной научно-исследовательской практике. Для улучшения воспроизводимости результатов требуется набор нескольких больших выборок (объем выборок зависит от предполагаемой мощности исследования, обусловливаемой пенетрантностью полиморфизмов-кандидатов). В свою очередь, охват необходимого количества маркеров для тестирования затрудняет колоссальное число полиморфизмов в геноме (более 165 миллионов согласно базе данных dbSNP). Поэтому в настоящее время ведущие мировые лаборатории применяют кандидатный подход лишь для верификации результатов исследований, выполненных посредством полногеномного подхода, который требует достаточно дорогостоящего оборудования и расходных материалов, и вследствие этого редко применяется в условиях ограниченных ресурсов. Поскольку данная проблема характерна для огромного количества научных коллективов, распространенность применения кандидатного подхода все еще остается высокой. Поэтому, учитывая собственный значительный опыт в таких исследованиях [1-5], в данной статье мы решили сфокусироваться именно на кандидатном подходе.

Основными этапами генетико-эпидемиологических исследований, выполняемых с помощью кандидатного подхода, являются: 1) разработка дизайна исследования, включая выбор генных полиморфизмов для анализа; 2) получение биоматериала, выделение ДНК и генотипирование; 3) статистический анализ и представление полученных результатов.

Этап 1. Разработка дизайна исследования

1) следует убедиться, что для достижения поставленной научной цели необходимо провести именно кандидатное генетико-эпидемиологическое исследование, а также определиться с критериями включения в исследование и критериями исключения из него;

2) необходимо определить планируемые к изучению звенья этиопатогенеза изучаемого заболевания, а затем - с ответственными за осуществление этих звеньев белками и кодирующими их генами;

3) по данным литературы и баз данных dbSNP, SNPinfo и SNPnexus следует выбрать полиморфизмы-кандидаты (вариабельные сайты кандидатных генов) по следующему алгоритму:

– локализация полиморфизмов в ответственных за планируемые к изучению звенья этиопатогенеза заболевания генах (генах-кандидатах);

– предполагаемая частота минорного аллеля по полиморфизмам в популяции не менее 5% по данным литературы, пилотных исследований, баз данных HapMap, 1000 Genomes или иных;

– предполагаемая по данным литературы или анализа *in silico* функциональная важность полиморфизмов (связь с функционально важными изменениями структуры или изменение экспрессии кодируемых генами-кандидатами РНК или белков);

– малое количество (не более трех), а в идеале отсутствие исследований о роли полиморфизмов в развитии изучаемой Вами патологии.

4) на основании частоты минорного аллеля в популяции и предполагаемой пенетрантности (способности проявляться в фенотипе и изменять риск) полиморфизмов-кандидатов следует рассчитать необходимую мощность исследования (величину уровня риска, которую можно будет определить в исследовании) и, соответственно, требуемый объем выборки. Это можно сделать в калькуляторе, находящемся в открытом доступе по ссылке: http://clincalc.com/stats/samplesize.aspx.

Этап 2. Получение биоматериала, выделение ДНК и генотипирование

1) Как правило, в качестве биоматериала в кандидатных генетико-эпидемиологических исследованиях используется периферическая венозная кровь, а в случае невозможности ее получения – буккальные соскобы. Выделение ДНК возможно проводить либо классическим фенол-хлороформным методом, оптимальным по соотношению цена/качество, но довольно трудоемким, либо используя коммерческие наборы, позволяющие значительно повысить скорость выделения. Используя на практике целый ряд методов выделения ДНК, авторы не считают необходимым рекомендовать какой-либо конкретный, поскольку они все обладают достаточно высокой эффективностью и в целом пригодны для выполнения кандидатных генетико-эпидемиологических исследований.

2) Хранение ДНК, как правило, осуществляется при температуре от слабоположительных температур (+4°C) до низких (-20 или -80°C). Повторные циклы заморозки и оттаивания губительно действуют на структуру ДНК, поэтому рекомендуется использовать раствор ДНК сразу же после разморозки. Если планируется проведение генотипирования по нескольким кандидатным генам в разные отрезки времени, целесообразно приготовить несколько аликвот раствора ДНК небольшого объема, достаточного для их использования в течение одного дня. Сроки и температурный режим хранения зависят от длительности сбора материала (частота встречаемости патологии в анализируемой популяции, планируемый объем выборки и другие сопутствующие факторы), планирования использования собранного материала в других научных проектах, условий финансирования научного исследования и возможности осуществления длительного ответственного хранения. При планировании завершения молекулярно-генетического тестирования в течении ближайших шести месяцев допускается хранение выделенной (после проверки чистоты выделения на наличие белковых и химических примесей) при температуре +4°C в пробирках с плотно подогнанными крышками. Предполагаемая длительная архивация требует условий хранения при низких температурах.

3) Генотипирование также может осуществляться посредством ряда методов, обладающих разной степенью производительности. Как правило, выбор технологии генотипирования зависит от имеющегося в наличии оборудования. В условиях дефицита ресурсов обычно используются технологии первого (аллель-специфичная полимеразная цепная реакция (ПЦР) или ПЦР-анализ полиморфизма длины рестрикционных фрагментов с детекцией результата путем электрофореза в агарозном геле) или второго (аллель-специфичная ПЦР с флюоресцентной детекцией результата в режиме реального времени) поколения. При доступности более производительных методов генотипирования (к примеру, на основе биочипов), естественно, стоит делать выбор в их пользу.

Этап 3. Статистический анализ и представление результатов

1) Для статистического анализа кандидатных генетико-эпидемиологических исследований авторы рекомендуют программу SNPStats, которая находится в открытом доступе по ссылке: http://bioinfo.iconcologia.net/SNPstats. Преимуществами данной программы является то, что она позволяет:

– сравнивать группы по пяти возможным моделям наследования (кодоминантной, доминантной, рецессивной, сверхдоминантной и лог-аддитивной). Каждая из этих моделей отражает различные варианты сравнения генотипов: отдельные сравнения гетерозиготного и вариантного гомозиготного с референтным гомозиготным (кодоминантная), объединенное сравнение

гетерозиготного и вариантного гомозиготного с референтным гомозиготным (доминантная), объединенное сравнение гетерозиготного и референтного гомозиготного с вариантным гомозиготным (рецессивная), объединенное сравнение двух гомозиготных генотипов с гетерозиготным (сверхдоминантная) и отдельные сравнения одного и двух вариантных аллелей с референтным (лог-аддитивная). Биологический смысл каждой из этих моделей наследования заключается в том, что при доминантной модели предполагается, что для изменения риска достаточно хотя бы одного вариантного аллеля, при рецессивной для этого требуются оба вариантных аллеля, при сверхдоминантной предполагается, что присутствие обоих аллелей изменяет риск в сравнении с двумя референтными или двумя вариантными, при кодоминантной – что каждый генотип может изменять риск независимо от остальных (неаддитивно), а при лог-аддитивной – что каждый вариантный аллель изменяет риск в аддитивной манере (т.е. что два вариантных аллеля увеличивают риск в два раза в сравнении с одним). Наиболее вероятная для каждого конкретного генного полиморфизма модель наследования имеет наименьшее значение информационного критерия Акаике, также вычисляемого данной программой;

– вносить поправки на воздействие сопутствующих факторов (confounders): безусловно необходимых (пол, возраст) и модифицируемых (клиникопатологических, этнографических, социально-бытовых, климатических, поведенческих и других), с большой вероятностью оказывающих влияние на патологический процесс вне зависимости от генома индивидуума;

– рассчитывать равновесие Харди-Вайнберга, отражающее частотное распределение аллелей в изучаемой популяции и необходимое для контроля качества генотипирования (значение менее 0,05 для контрольной группы асимптоматичных субъектов свидетельствует о неудовлетворительном качестве);

– рассчитывать модификаторы риска как для отдельных генных полиморфизмов, так и для их сочетаний (гаплотипов).

2) Поскольку кандидатные генетико-эпидемиологические исследования практически всегда характеризуются большим количеством сравнений, требуется оптимальный метод внесения поправки на множественные сравнения для расчета вероятности отвергнуть верную нулевую гипотезу p (p-значения). Авторы рекомендуют использовать для этой цели среднюю долю ложных отклонений гипотез (false discovery rate), которая рассчитывает скорректированные

Таблица 1. Пример представления результатов кандидатного генетико-эпидемиологического исследования. Статистически значимые различия между выборками выделены жирным шрифтом.

Table 1. Results of the genetic association study performed using candidate approach. Statistically significant differences are marked bold.

Модель наследования Model of inheritance	Генотип Genotype	Без заболевания Without disease	С заболеванием With disease	ОШ (95% ДИ) OR (95% CI)	p p value	ИКА AIC	РХВ HWE
IL1B rs1143634							
Кодоминантная Codominant	G/G	154 (51,3%)	82 (67,8%)	1,00	0,0029	472,5	0,89
	G/A	123 (41%)	28 (23,1%)	0,43 (0,26-0,72)			
	A/A	23 (7,7%)	11 (9,1%)	0,97 (0,43-2,18)			
Доминантная Dominant	G/G	154 (51,3%)	82 (67,8%)	1,00	0,0036	473,8	
	G/A-A/A	146 (48,7%)	39 (32,2%)	0,51 (0,32-0,81)			
Рецессивная Recessive	G/G-G/A	277 (92,3%)	110 (90,9%)	1,00	0,52	481,8	
	A/A	23 (7,7%)	11 (9,1%)	1,30 (0,59-2,88)			
Сверхдоминантная Overdominant	G/G-A/A	177 (59%)	93 (76,9%)	1,00	**0,0016**	470,6	
	G/A	123 (41%)	28 (23,1%)	**0,43 (0,26-0,71)**			
Лог-аддитивная Log-additive	---	---	---	0,70 (0,48-1,00)	0,046	478,2	
IL6 rs1554606							
Кодоминантная Codominant	G/G	92 (30,7%)	31 (25,4%)	1,00	0,31	483	0,99
	G/T	149 (49,7%)	62 (50,8%)	1,38 (0,82-2,33)			
	T/T	59 (19,7%)	29 (23,8%)	1,57 (0,84-2,95)			
Доминантная Dominant	G/G	92 (30,7%)	31 (25,4%)	1,00	0,15	481,2	
	G/T-T/T	208 (69,3%)	91 (74,6%)	1,44 (0,87-2,36)			
Рецессивная Recessive	G/G-G/T	241 (80,3%)	93 (76,2%)	1,00	0,36	482,5	
	T/T	59 (19,7%)	29 (23,8%)	1,28 (0,76-2,17)			
Сверхдоминантная Overdominant	G/G-T/T	151 (50,3%)	60 (49,2%)	1,00	0,58	483	
	G/T	149 (49,7%)	62 (50,8%)	1,13 (0,73-1,76)			
Лог-аддитивная Log-additive	---	---	---	1,26 (0,92-1,72)	0,14	481,2	

ОШ – отношение шансов, ДИ – доверительный интервал, ИКА – информационный критерий Акаике, РХВ – равновесие Харди-Вайнберга

OR – odds ratio, CI – confidence interval, AIC – Akaike information criterion, HWE – Hardy-Weinberg equilibrium

с учетом поправки на множественные сравнения q-значения из изначальных p-значений, полученных при анализе в SNPStats. Калькулятор для расчета q-значений находится в открытом доступе по ссылке: http://users.ox.ac.uk/~npike/fdr/ (файл с Excel-шаблоном FDR.xls в правом нижнем углу). Также можно использовать и другие калькуляторы (к примеру, http://www.sdmproject.com/utilities/?show=FDR).

3) Вследствие больших массивов данных результаты кандидатных генетико-эпидемиологических исследований, как правило, представляются в виде таблиц, как в клинических статьях. Пример оформления таблицы приведен в табл. 1. В табл. 1 p-значение представлено до его коррекции методом средней доли ложных отклонений гипотез. Скорректированное q-значение равно 0,0032, что также меньше 0,05 и, следовательно, является статистически значимым (**таблица 2**). Для расчета q-значений берется только p-значение по наиболее вероятной для каждого генного полиморфизма модели наследования (с наименьшим информационным критерием Акаике).

p-значение p value	q-значение q value
0,0016	0,0032
0,14	0,14

Таблица 2. Представление скорректированных p-значений после поправки на множественные сравнения при помощи средней доли ложных отклонений гипотез (q-значений) из таблицы 1.

Table 2. False discovery rate-corrected p values (q values) from Table 1.

4) Кроме того, в таблицах необходимо максимально подробно представлять характеристики выборки и описывать все проанализированные генные полиморфизмы. Если выборка может быть описана в произвольной форме, то для генных полиморфизмов авторы рекомендуют в обязательном порядке указывать следующие характеристики: референтный (универсальный) номер полиморфизма (rs number), нуклеотидную замену и функциональное последствие (аминокислотную замену или присутствие в некодирующих регионах), хромосомную позицию и нуклеотидную последовательность праймеров для ПЦР. Пример подобного описания приведен в **таблице 3.**

Полиморфизм Polymorphism	Нуклеотидная замена Nucleotide substitution	Хромосомная позиция Chromosomal position	Аминокислотная замена Amino acid substitution	5'-3' (F) и 3'-5' (R)-праймеры для полимеразной цепной реакции Forward 5'-3' and reverse 3'-5' polymerase chain reaction primers
Ген *IL1B* *IL1B* gene				
rs1143634	G>A	113590390	Phe105Phe	F: cataagcctcgttatcccatgtgtc R: aagaagataggttctgaaatgtgga
Ген *IL6* *IL6* gene				
rs1554606	T>G	22768707	Интронный intronic	F: ttagttcatcctgggaaaggtactc R: cagggccttttccctctctggctgc
rs1800796	G>C	22766246	5'-upstream	F: atggccaggcagttctacaacagcc R: ctcacagggagagccagaacacaga
rs2069827	G>T	22765456	5'-upstream	F: gcccaacagaggtcactgttttatc R: atcttgaagagatctcttcttagca

Таблица 3. Характеристики изученных генных полиморфизмов

Table 3. Features of the tested gene polymorphisms

Ключевые аспекты исследования: взгляд рецензента

Стоит отметить, что при оценке кандидатных генетико-эпидемиологических исследований, помимо общих для научных статей аспектов, рецензенты в основном обращают внимание на:

1) объем выборки и мощность исследования (реально ли с заявленным объемом выборки получить статистически и биологически значимые результаты);

2) выбранные звенья развития изучаемой патологии (будет ли их изучение иметь научную новизну и значимость);

3) алгоритм выбора генных полиморфизмов для анализа (почему были выбраны именно эти, а не какие-то иные генные полиморфизмы);

4) полноту описания характеристик набранной выборки (достаточно ли их для оценки всех сопутствующих факторов, которые могли бы повлиять на результаты исследования);

5) качество проведенного статистического анализа и представление результатов.

В заключение стоит отметить, что, несмотря на то, что кандидатные генетико-эпидемиологические исследования имеют значительные недостатки и зачастую малодоказательны сами по себе, они тем не менее могут быть успешно использованы для верификации результатов полногеномных генетико-эпидемиологических исследований и патофизиологических гипотез.

Литература / References:

1. Ponasenko AV, Khutornaya MV, Kutikhin AG, Rutkovskaya NV, Tsepokina AV, Kondyukova NV, Yuzhalin AE, Barbarash LS. A Genomics-Based Model for Prediction of Severe Bioprosthetic Mitral Valve Calcification. Int J Mol Sci. 2016;17 (9). pii: E1385.

2. Kutikhin AG, Ponasenko AV, Khutornaya MV, Yuzhalin AE, Zhidkova II, Salakhov RR, Golovkin AS, Barbarash OL, Barbarash LS. Association of TLR and TREM-1 gene polymorphisms with atherosclerosis severity in a Russian population. Meta Gene. 2016; 9: 76-89.

3. Golovkin AS, Ponasenko AV, Yuzhalin AE, Salakhov RR, Khutornaya MV, Kutikhin AG, Rutkovskaya NV, Savostyanova YY, Barbarash LS. An association between single nucleotide polymorphisms within TLR and TREM-1 genes and infective endocarditis. Cytokine. 2015; 71 (1):16-21.

4. Golovkin AS, Ponasenko AV, Khutornaya MV, Kutikhin AG, Salakhov RR, Yuzhalin AE, Zhidkova II, Barbarash OL, Barbarash LS. Association of TLR and TREM-1 gene polymorphisms with risk of coronary artery disease in a Russian population. Gene. 2014; 550 (1): 101-109.

5. Kutikhin AG, Yuzhalin AE, Volkov AN, Zhivotovskiy AS, Brusina EB. Correlation between genetic polymorphisms within IL-1B and TLR4 genes and cancer risk in a Russian population: a case-control study. Tumour Biol. 2014; 35 (5): 4821-4830.

Сведения об авторах

Кутихин Антон Геннадьевич, *кандидат медицинских наук, старший научный сотрудник лаборатории геномной медицины отдела экспериментальной и клинической кардиологии ФГБНУ «Научно-исследовательский институт комплексных проблем сердечно-сосудистых заболеваний», г. Кемерово, Россия.*
Вклад в статью: *написание статьи.*

Южалин Арсений Евгеньевич, *аспирант Оксфордского института радиационной онкологии, Оксфорд, Соединенное Королевство Великобритании и Северной Ирландии.*
Вклад в статью: *написание статьи.*

Понасенко Анастасия Валериевна, *кандидат медицинских наук, заведующая лабораторией геномной медицины отдела экспериментальной и клинической кардиологии ФГБНУ «Научно-исследовательский институт комплексных проблем сердечно-сосудистых заболеваний», г. Кемерово, Россия.*
Вклад в статью: *написание статьи.*

Корреспонденцию адресовать:
Кутихин Антон Геннадьевич
650002, г. Кемерово, ул. Сосновый бульвар, 6
Тел.: +79609077067
E-mail: antonkutikhin@gmail.com

Authors

Dr. Anton G. Kutikhin, *MD, PhD, Senior Researcher, Laboratory for Genomic Medicine, Division of Experimental and Clinical Cardiology, Research Institute for Complex Issues of Cardiovascular Diseases, Kemerovo, Russian Federation.*
Contribution: *wrote the article.*

Mr. Arseniy E. Yuzhalin, *MSc (Res), PhD Student, Department of Oncology, Cancer Research UK and Medical Research Council Oxford Institute for Radiation Oncology, University of Oxford, Oxford, United Kingdom.*
Contribution: *wrote the article.*

Dr. Anastasia V. Ponasenko, *MD, PhD, Head of the Laboratory for Genomic Medicine, Division of Experimental and Clinical Cardiology, Research Institute for Complex Issues of Cardiovascular Diseases, Kemerovo, Russian Federation.*
Contribution: *wrote the article.*

Acknowledgements: *There was no funding for this article.*

Corresponding author:
Dr. Anton G. Kutikhin,
Sosnovy Boulevard 6, Kemerovo, 650002, Russian Federation
E-mail: antonkutikhin@gmail.com

Статья поступила: 23.05.17 г.
Принята в печать: 27.05.17 г.