

Е.Ю. Брагина, С.В. Буйкин

ВЛИЯНИЕ ЧИСЛЕННОСТИ КОНТРОЛЬНОЙ ВЫБОРКИ НА ЗНАЧИМОСТЬ АССОЦИАЦИЙ ГЕНЕТИЧЕСКИХ МАРКЕРОВ С РАЗВИТИЕМ МУЛЬТИФАКТОРИАЛЬНЫХ ЗАБОЛЕВАНИЙ

УДК 575.174.015.3:599.9

В настоящем исследовании оценено влияние увеличения численности контрольной выборки на значимость ранее полученных ассоциаций полиморфных вариантов генов *IL12B* (rs3212227), *IL4RA* (rs1805010) и *GSTM1* (del) с развитием вторичного туберкулеза, рассеянного склероза, atopической бронхиальной астмы и сахарного диабета 1-го типа у русских жителей г. Томска. Установлено, что для увеличения доказательности исследования генетической предрасположенности к различным мультифакториальным заболеваниям возможно увеличение численности контрольной выборки за счет популяционного контроля.

Ключевые слова: мультифакториальные заболевания, дизайн исследования случай-контроль, полиморфные варианты, ассоциации.

For association study results for SNP *IL12B* (rs3212227), *IL4RA* (rs1805010) and *GSTM1* (del) with secondary tuberculosis, atrophy bronchial asthma, 1 type diabetes, we estimated effect of sample size on association significance. We have shown that it is possible to increase significance level of association with some common diseases by increasing number of controls.

Keywords: common diseases, case-control design, SNP, association.

Введение

Дизайн исследования случай-контроль для определения роли генетических вариантов в развитии мультифакториальных заболеваний (МФЗ) является общепринятым и предполагает высокую мощность полученных ассоциаций [18]. Для исследований случай-контроль используется сравнение распределения вариантов аллелей и генотипов между группами больных и здоровых индивидов. Данный подход наряду с другими алгоритмами исследований генетических ассоциаций (когортное, семейное) имеет свои недостатки и преимущества, и в большей степени различные подходы дополняют друг друга. Распространенной проблемой дизайна случай-контроль является слабая воспроизводимость результатов, полученных в различных исследованиях. Лишь незначительная часть полученных ассоциаций находит свое подтверждение [8]. Одной из возможных причин таких результатов могут быть немногочисленные группы для исследования. Если частота «редкого» аллеля превышает 5%, а численность исследуемых групп невелика, то могут быть пропущены многие ассоциации МФЗ с генетическими вариантами [15]. Другой причиной может быть неоднородность этнического состава исследуемых групп больных и здоровых индивидов, а также принадлежность их к определенной популяции [5]. Сбор материала с учетом вышеперечисленных критериев несколько затрудняет работу исследователей, но тем не менее в эру популяризации полногеномных ис-

следований, требующих большой численности исследуемых групп, дизайн случай-контроль остается наиболее эксплуатируемым.

Цель исследования. Анализ воспроизводимости ранее полученных ассоциаций для ряда заболеваний с увеличением численности контрольной группы.

Материал и методы исследования

В работе были использованы ранее полученные и опубликованные данные о распределении генотипов и аллелей в контрольных выборках [1, 2, 4, 6]. Исследуемые были объединены в 3 группы: I - популяционная выборка г. Томска (n=96): 38 мужчин и 52 женщин; II - контрольная выборка для исследования генетических причин atopического дерматита; русские, проживающие на территории г.Томска (n=150): 70 мужчин и 80 женщин; III - контрольная выборка для исследования генетических причин бронхолегочных заболеваний, включая хронический обструктивный бронхит, бронхиальную астму, туберкулез; русские, проживающие на территории г. Томска (n=140) - 60 мужчин и 80 женщин.

Для анализа были выбраны несинимичные полиморфизмы генов *IL12B* (rs3212227), *IL4RA* (rs1805010) и протяженная делеция в гене *GSTM1* (порядка 10 т.п.о.), приводящая к отсутствию соответствующего белкового продукта. В качестве модели заболеваний для сравнения с контрольными группами использовали группу больных вторичным туберкулезом, рассеянным склерозом, atopической бронхиальной астмой и сахарным диабетом 1-го типа, индивиды во всех исследуемых группах - русские жители г. Томска. Возраст больных и здоровых индивидов в данном контексте исследования не

принимался во внимание. Распределение генотипов для рассматриваемых генетических вариантов проверяли на соответствие равновесию Харди-Вайнберга с помощью точного теста Фишера [3]. Оценку значимости различий между группами проводили с помощью критерия χ^2 и двустороннего точного теста Фишера и значимыми считали отличия при $p \leq 0,05$. Расчеты проводили с использованием программ «STATISTICA 6.0» и «Microsoft Excel».

Результаты и обсуждение

Во всех исследованных выборках для гена *IL12B* (rs3212227) наблюдаемое распределение генотипов соответствовало ожидаемому при равновесии Харди-Вайнберга (табл.1). Наблюдаемая гетерозиготность составила 32-33% и во всех группах отмечалась однонаправленная тенденция к незначительному избытку гетерозигот. При сравнении частот аллелей и генотипов между контрольными выборками значимых различий не выявлено ($p=0,587-0,967$), вследствие чего они были объединены и численность суммарной контрольной выборки составила 374 индивида. В суммарной контрольной выборке сохранялось равновесие Харди-Вайнберга, наблюдаемая гетерозиготность составила 32,9% и соответствовала предыдущим группам.

Далее проводилась проверка воспроизводимости ассоциации с заболеванием, с использованием суммарного контроля. Ранее для локуса *IL12B* (rs3212227) была получена ассоциация с вторичным туберкулезом [4]. Первоначально численность контрольной выборки составила 129, а больных вторичным туберкулезом - 237 индивидов, и при сравнении распределения генотипов и аллелей меж-

ду этими группами были найдены различия с уровнем значимости $p=0,0340$ и $p=0,0410$ для частот генотипов и аллелей соответственно. С использованием суммарной контрольной выборки ($n=374$) ассоциация сохранилась, и уровень значимости различий между группой больных и суммарной контрольной выборкой значительно вырос ($0,0179$ и $0,0150$ для генотипов и аллелей соответственно).

Подобные результаты были получены в исследовании наследственной предрасположенности рассеянного склероза [1]: ген *IL12B* (rs3212227) ассоциирован с развитием заболевания ($0,00006$ и $0,0003$ для генотипов и аллелей соответственно). Численность группы больных составила 90, а контрольной – 129 индивидов. При использовании суммарной контрольной выборки ($n=374$) ассоциация сохраняется, а уровень значимости различий возрастает ($0,00004$ и $0,00003$ для генотипов и аллелей соответственно).

Несколько иная картина наблюдалась для гена *GSTM1* (del). Распространенность делеции в контрольных выборках варьирует от 55,2% до 64,6%, однако в распределении генотипов группы не различаются между собой ($p=0,241-1,000$). В суммарной контрольной выборке частота делеции составила 60,5%. Ранее для данного варианта была получена ассоциация с развитием атопической бронхиальной астмы, частота делеции преобладала у больных индивидов (71,0%) по сравнению со здоровыми (55,2%), уровень значимости различий составил 0,0108 [2]. При использовании суммарной контрольной выборки ($n=347$) ассоциация сохраняется, однако уровень значимости достигнутых различий снизился и составил 0,0429. Возможно, данный факт является смещением, за счет значительного увеличения контрольной выборки (374 по сравнению с 134) и гораздо меньшей численностью группой больных ($n=131$).

Увеличение численности групп больных и контроля оправдано, особенно в случае исследования SNPs, с частотой «редкого» аллеля менее 5%. По данным проектов HarMap, Encode для 40-60% SNPs в геноме человека частота «редкого» аллеля составляет менее 5%, в этом случае, чем ниже риск и частота «редкого» аллеля, тем больший размер выборки необходим для исследования (табл. 2).

В случае если частота «редкого» аллеля превышает 5% при небольшой численности исследуемых групп отно-

сительный риск развития заболевания будет невысоким, порядка 1,2-1,6 для большинства из найденных ассоциаций [15].

Основным недостатком исследований случай-контроль является необходимость стратификации исследуемого

населения по ряду параметров [9, 13, 14]. Исследования с использованием семейного материала позволяют избежать этой проблемы, более того, их несомненным плюсом является гомогенность всех членов семьи в отношении воздействия факторов окружающей среды. Однако сбор хорошо охарактеризованного семейного материала достаточно сложен и требует длительного времени, поэтому при изучении генетических ассоциаций с МФЗ гораздо реже используется именно этот дизайн исследования. В 2006 г. были опубликованы результаты сравнения ассоциаций, полученных с использованием дизайна случай-контроль и семейного исследования [11]. Работа представляла мета-анализ данных по генам-канди-

датам широко распространенных заболеваний человека, включая астму, сахарный диабет, туберкулез и другие. Анализируя порядка 93 исследования, авторам удалось показать, что, несмотря на возможные отклонения и необходимость стратификация населения, ассоциации, полученные путем сравнения случай-контроль в основном подтверждают результаты исследований семейного материала [10, 21]. В большинстве случаев оба дизайна являются взаимодополняющими, и совмещение их результатов повышает мощность исследования [7, 12, 16, 17, 22].

Для оценки влияния численности контрольной группы на значимость ассоциаций с заболеванием для двух дизайнов (случай-контроль и семейный анализ) были использованы результаты ранее проведенного исследования сахарного диабета 1 типа [6]. В этой работе была получена ассоциация

Таблица 1

Уровни гетерозиготности
в контрольных выборках

Ген/SNP	Группа	Гетерозиготность	χ^2
<i>IL12B</i> rs3212227	I	ho=0,3368±0,0485 he=0,3071±0,0352	0,8880
	II	ho=0,3200±0,0381 he=0,3036±0,0281	0,4352
	III	ho=0,3333±0,0415 he=0,2880±0,0307	3,1973
	Суммарная контрольная выборка n=374	ho=0,3289±0,0243 he=0,2992±0,0179	3,6744
<i>IL4RA</i> rs1805010	II	ho=0,4932±0,0414 he=0,4886±0,0090	0,0124
	III	ho=0,5246±0,0452 he=0,4837±0,0116	0,8700
	Суммарная контрольная выборка n=276	ho=0,5075±0,0305 he=0,4865±0,0071	0,4965

Примечание. n- численность индивидов в группе; ho и he – наблюдаемая и ожидаемая гетерозиготность $\pm S.D.$

Таблица 2

Оптимальная численность групп
для выявления вариантов рисковой значимости
аллеля [составлена по: [19]]

Относительный риск	Частота «редкого» аллеля	Количество исследуемых образцов		
		Кол-во «случаев», n	Кол-во «контроля», n	Суммарно, n
1,5	5%	931	931	1.862
2,0	5%	290	290	580
1,5	2,5%	1.710	1.710	3.420
2,0	2,5%	525	525	1.050
1,5	1%	4.060	4.060	8.120
2,0	1%	1.225	1.225	2.450

Примечание. n – численность индивидов в группе.

гена *IL4RA* (rs1805010) с развитием сахарного диабета 1 типа ($p=0,0183$) с помощью Transmission Disequilibrium Test, показывающего насколько чаще передается «патологический» аллель больному ребенку от его родителей. В то же время, используя сравнение между группой больных и контрольной выборкой, значимых различий автором не обнаружено ($p=0,3880$ и $p=0,792$ при сравнении генотипов и аллелей соответственно). Причиной отсутствия ассоциации при сравнении случай-контроль, могла быть недостаточная численность контрольной выборки ($n=122$), которая почти в два раза по численности меньше, чем группа больных индивидов ($n=207$). Для увеличения численности контрольной выборки мы использовали контрольные группы, рассмотренные ранее (табл.1) В результате анализа, по локусу *IL4RA* (rs1805010) значимых различий между контрольными группа-

ми не обнаружено ($p=0,7951-0,9225$). Наблюдаемое распределение генотипов соответствовало ожидаемому при равновесии Харди-Вайнберга во всех контрольных выборках, которые были объединены для дальнейшего анализа. В суммарной контрольной выборке ($n=268$) сохранялось равновесие Харди-Вайнберга, наблюдаемая гетерозиготность составила порядка 50,8% и соответствовала предыдущим трем группам. При сравнении распределения генотипов и аллелей между суммарной контрольной выборкой и группой больных сахарным диабетом 1 типа не было найдено значимых отличий (0,3772 и 0,5420 для генотипов и аллелей соответственно). Таким образом, увеличение контрольной выборки не помогло выявить ассоциаций с заболеванием. В данном случае, первичное предположение о недостаточности контроля по отношению к случаю не подтвердилось.

Настоящее исследование основывалось на подходе увеличения численности контрольной выборки, который использовался консорциумом Wellcome Trust Case Control Consortium [20]. На основании данных полногеномного анализа был проведен поиск ассоциаций с МФЗ в британской популяции, включая коронарный атеросклероз, артериальная гипертензия, биполярный психоз, сахарный диабет 1 типа, сахарный диабет 2 типа, болезнь Крона, ревматоидный артрит. Для исследования были выбраны две контрольные выборки: 1500 индивидов – британская когорта 1958 года рождения (58BC) и 1500 отобранных из образцов донорской крови. По результатам генотипирования 500,568 однонуклеотидных полиморфизмов авторами отмечается, что было мало статистически значимых различий между двумя контрольными выборками, несмотря на то, что эти группы отличались по возрасту и месту проживания. Контрольные группы были объединены для дальнейшего поиска ассоциаций с заболеваниями уже с использованием контрольной группы численностью 3000 индивидов. Так же в данной работе был использован метод наращивания контрольной группы

за счет включения в ее состав индивидов с болезнями, отличающимися по патогенезу от изучаемого заболевания. Такой подход не только подтвердил полученные ассоциации, но и позволил выявить новые локусы с высоким уровнем статистической значимости ($p<5\times 10^{-7}$). Авторы указывают, что основной недостаток при использовании популяционного контроля для МФЗ, это наличие в контроле индивидов с исследуемым заболеванием. Однако эффект его на мощность статистических критериев довольно умеренный, за исключением случаев, когда это смещение существенно. Например, если в 5% контроля было бы это заболевание, то это привело бы к потере в мощности, эквивалентной уменьшению размеров выборки на 10%.

Заключение

В каждом конкретном случае исследователь делает выбор между уровнем детализации в описании фенотипа и стоимостью исследования. Использование подходов, снижающих стоимость исследования при незначительном влиянии на мощность статистических критериев, имеет несомненные преимущества, однако необходимо накопление практического опыта их применения в изучении генетических основ МФЗ.

Полученные предварительные результаты позволяют заключить, что для широко распространенных заболеваний для увеличения доказательности исследования оправдан подход увеличения численности контрольной выборки за счет популяционного контроля.

Исследование выполнено при финансовой поддержке РГНФ в рамках научно-исследовательского проекта РГНФ № 08-06-00514а.

Литература

1. Бабенко С.А. Роль аллельных вариантов генов иммунного ответа в развитии рассеянного склероза: автореф. дис...канд.мед.наук / С.А. Бабенко. – Томск: НИП, 2008. – 23 с.
2. Вейр Б. Анализ генетических данных. Пер. с англ. / Б. Вейр. – М.: Мир, 1995. – 400 с.
3. Колоколова О.В. Аллельные варианты генов-кандидатов подверженности туберкулезу у русского населения Западной Сибири: автореф.

дис...канд.мед.наук / О.В. Колоколова. – Томск: НИП, 2005. – 18 с.

4. Кучер А.Н. Исследование ассоциаций аллельных вариантов генов с мультифакториальными заболеваниями: проблема формирования контрольных групп / А.Н. Кучер // Генетик человека и патология: сб. науч. тр. – Томск: Изд-во «Печатная мануфактура», 2007. – Вып. 8. — С. 328-331.

5. Тарасенко Н.В. Патогенетика сахарного диабета 1 типа и его осложнений: гены синтаз оксида азота и цитокинов: автореф. дис...канд.мед.наук / Н.В. Тарасенко – Томск: НИП, 2008. – 23 с.

6. A comparison of case-control and family-based association methods: The example of sickle-cell and malaria / H. Ackerman [et al.] // Ann. Hum. Genet. – 2005. – V. 69. – P. 559–565.

7. A comprehensive review of genetic association studies / J.N. Hirschhorn [et al.] // Genet. Med. – 2002. – V. 4. – P. 45–61.

8. Cardon L.R. Population stratification and spurious allelic association / L.R. Cardon, L.J. Palmer // Lancet. – 2003. – V. 361. – P. 598–604.

9. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. S. Wacholder [et al.] // Cancer Epidemiol. Biomarkers Prev. – 2002. – V. 11. – P. 513–520.

10. Family-based versus unrelated case-control designs for genetic associations / E. Evangelou [et al.] // Plos Genetics. – 2006. – V. 2 (8). – P. 1147–1155.

11. Genetic association analysis using data from triads and unrelated subjects / M.P. Epstein [et al.] // Am. J. Hum. Genet. – 2005. – V. 76. – P. 592–608.

12. Genome-wide association studies: Theoretical and practical concerns / W.Y. Wang [et al.] // Nat. Rev. Genet. – 2005. – V. 6. – P.109–118.

13. Hattersley A.T. What makes a good genetic association study? / A.T. Hattersley, M.I. McCarthy // Lancet. – 2005. – V. 366. – P. 1315–1323.

14. Ioannidis J.P. Genetic associations: False or true? / J.P. Ioannidis // Trends. Mol. Med. – 2003. – V. 9. – P. 135–138.

15. Kazeem G.R. Integrating case-control and TDT studies / G.R. Kazeem, M. Farrall // Ann. Hum. Genet. – 2005. – V. 69. – P. 329–335.

16. Mitchell L.E. Relationship between case-control studies and the transmission/disequilibrium test / L.E. Mitchell // Genet. Epidemiol. – 2000. – V. 19. – P. 193–201.

17. Phases of biomarker development for early detection of cancer / M. S. [et al.] Pepe // J. Nat. Cancer Inst. – 2001. – V. 93. – P. 1054–1061.

18. Shifting paradigm of association studies: value of rare SNPs / I.P. Gorlov [et al.] // Am.J. Hum. Genet. – 2008. – V. 82. – P. 100–112.

19. The Wellcome trust Case Control Consortium. Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls // Nature. – 2007. – V. 447. – P. 661–678.

20. Thomas D.C. Point population stratification: A problem for case-control studies of candidate-gene associations? / D.C. Thomas, J.S. Witte // Cancer Epidemiol. Biomarkers Prev. – 2002. – V. 11. – P. 505–512.

21. Weinberg C.R. A hybrid design for studying genetic influences on risk of diseases with onset early in life / C.R. Weinberg, D.M. Umbach // Am. J. Hum. Genet. – 2005. – V. 77. – P. 627–636.