

**Математическая модель построения генетических шкал риска развития
неинфекционных заболеваний**

Сопленкова Анна Глебовна

Студент (специалист)

Московский государственный университет имени М.В.Ломоносова,
Механико-математический факультет, Кафедра теории вероятностей, Москва, Россия
E-mail: anna.mail10@mail.ru

Основной задачей современной генетики является предсказание количественного фенотипа человека, который может характеризоваться совокупностью признаков, включающих в себя, напр., наличие неинфекционного заболевания, уровень какого-либо белка в плазме крови и т.п., по его генотипу, который определяется последовательностью ДНК [2]. Отличие последовательности ДНК от референсной размерностью в один нуклеотид называется однонуклотидным полиморфизмом (ОНП) [6]. Таким образом, ОНП может трактоваться как дискретная переменная. За исключением отдельных хорошо изученных случаев [7], значение, которое принимает ОНП, не оказывает сильного влияния на организм человека. Чаще за фенотипический признак отвечает некоторое количество ОНП. Иногда количество ОНП, связанных с фенотипическим признаком, доходит до нескольких тысяч. Обнаружить индивидуальные эффекты таких ОНП сложно: для этого требуются достаточно большие размеры выборок и современные компьютеры с мощными процессорами [1]. Отметим, что представленное исследование было бы невозможно еще десятилетие назад. Цель работы — получить ассоциации изучаемых фенотипических признаков с массивом данных, представляющий собой последовательности ОНП в выделенных участках 13 генов (экзонов [1]) у 1680 человек из Ивановской области. В качестве фенотипических признаков рассматривались различные липидные показатели [8] в плазме крови, а также атеросклеротическое поражение периферических сосудов. С этой целью для каждого ОНП был вначале проверен на выполнение закон Харди-Вайнберга [2]. Затем ОНП, для которых закон Харди-Вайнберга отвергался, были исключены из исследования, поскольку принято считать, что в подавляющем большинстве случаев это означает ошибку генотипирования [2]. В исследованиях, направленных на поиск ассоциаций между фенотипом и ОНП применяются различные статистические методы, наиболее распространенными среди которых являются линейная регрессия [4] и дисперсионный анализ [5]. В работе к имеющимся данным применены оба этих метода и проведено сравнение полученных результатов. Линейная регрессия обнаружила наибольшее число значимо связанных с изучаемым признаком ОНП: в случаях, когда эффект ОНП носит аддитивный характер, линейная регрессия оказывается наиболее подходящим методом, поскольку она нацелена на поиск именно аддитивных эффектов. Однако с помощью дисперсионного анализа выявлены несколько ОНП, которые не были обнаружены линейной регрессией. При более детальном изучении этих вариантов выявлено, что их эффект не является аддитивным: в случае одной мутантной аллели среднее значение признака уменьшалось, а в случае двух мутантных аллелей — резко увеличивалось. После сравнения классических для генетических исследований статистических методов стояла задача построить математическую модель предсказания фенотипа, которая опирается не только на наличие или отсутствие определенных ОНП, но также и на взаимное положение этих ОНП друг относительно друга. Более того, изучена связь между фенотипическими признаками и последовательным накоплением мутаций. Для этого каждому человеку из выборки сопоставлялась некоторая траектория накопления мутаций. В качестве меры накопления бралась площадь между этой траекторией

и некоторой “нулевой” траекторией, соответствующей референсному геному. Найдена зависимость между фенотипическими признаками и количеством накопленных мутаций в изучаемой выборке. В то же время обнаружить влияние взаимного расположения мутаций на фенотипические признаки не удалось. Данные для исследования предоставлены лабораторией молекулярной генетики ФГБУ «Национальный медицинский исследовательский центр терапии и профилактической медицины» МЗ РФ.

Источники и литература

- 1) Graffelman, J., Jain, D., and Weir, B. (2017). A genome-wide study of Hardy–Weinberg equilibrium with next generation sequence data. *Hum. Genet.* 136, 727–741. doi: 10.1007/s00439-017-1786-7
- 2) Falconer D.S., *Introduction to Quantitative Genetic*
- 3) Kimmel M., Axelrod D.E., *Branching Processes in Biology*
- 4) Buzkova, P. (2013). Linear regression in genetic association studies. *PLOS One*, 8(2):e56976.
- 5) Zhang, X., Zou, F., Wang, W. (2008). FastANOVA: an Efficient Algorithm for Genome-Wide Association Study. *KDD : proceedings. International Conference on Knowledge Discovery and Data Mining*, 821–829.
- 6) Den Dunnen J. T. Recommendations for the description of sequence variants (англ.) // *Human Genome Variation Society : journal.* — 2008.
- 7) Weedon MN, Jackson L, Harrison JW, Ruth KS, Tyrrell J, Hattersley AT, Wright CF (2021). Use of SNP chips to detect rare pathogenic variants: retrospective, population based diagnostic evaluation. *BMJ* 2021;372:n214
- 8) Metelskaya Victoria A., Shalnova Svetlana A., Yarovaya Elena B., Kutsenko Vladimir A., Boytsov Sergey A., Shlyakhto Eugeny V., Drapkina Oxana M. (2022) Lipoprotein Profile in Populations from Regions of the Russian Federation: ESSE-RF Study. *International Journal of Environmental Research and Public Health*, MDPI (Basel, Switzerland), том 19, № 2, с. 931 DOI