

Биобанк России — онлайн-платформа для хранения и анализа результатов GWAS-исследований полигенных заболеваний, а также разработка фенотипических предсказательных моделей на базе когорты Биобанка России

Усольцев Д. А. (Университет ИТМО), О.П. Ротарь (НМИЦ им. Алмазова), А.А. Костарева (НМИЦ им. Алмазова), А.О. Конради (НМИЦ им. Алмазова), Е.В.Шляхто (НМИЦ им. Алмазова), Артемов Н.Н. (Институт геномной медицины, Детская больница Нейшенвайд, США)

Научный руководитель: к. т. н., доцент Сергушичев А. А. (Университет Вашингтона в Сент-Луисе, США)

Большинство заболеваний и признаков человека являются полигенными, то есть их развитие зависит от множества генов, а также факторов окружающей среды. Одним из наиболее надежных методов для поиска связи между геном и заболеванием является полногеномное ассоциативное исследование (genome-wide association study, GWAS). В данной работе представлена онлайн-платформа, содержащая результаты GWAS для выборки из российской популяции, включающей 4145 человек и 464 признака. Фенотипические данные платформы использованы для создания простых в использовании предсказательных моделей рисков для распространенных заболеваний.

Ключевые слова: Биобанк России, GWAS, риски заболеваний, предсказательные модели.

Введение. Полногеномное ассоциативное исследование (GWAS) представляет собой статистический тест, оценивающий связь между исследуемым признаком и каждым ДНК-вариантом. Результатом GWAS является определение размера эффекта генотипа на исследуемый признак и р-значения для каждого ДНК-варианта [1]. Популяционный биобанк является ценным ресурсом, содержащим данные о фенотипах, генетической информации и результатах GWAS [2]. Необходимость создания популяционно-специфичных биобанков обусловлена трудностью переноса генетических моделей рисков заболеваний между различными популяциями [3]. Российская популяция, включающая 150 этнических групп — от европейских до азиатских, делает популяционные исследования в России важным инструментом для изучения широкого спектра генетических вариантов как европейского, так и азиатского происхождения, при сравнительно небольшом объеме когорты и ограниченных ресурсах. Также, одним из важных применений биобанка России является возможность разработки предсказательных моделей для оценки рисков возникновения заболеваний.

В данной работе представлена онлайн-платформа результатов GWAS-исследований полигенных заболеваний в российской популяции (4145 человек и 464 признака), а также онлайн-платформа для предсказания рисков развития метаболически нездорового ожирения.

Онлайн платформа Биобанка России. Результаты GWAS для 464 признаков были агрегированы в единую платформу, доступную по адресу <https://biobankrus.almazovcentre.ru/>. Платформа включает страницу фенотипа, на которой представлена графическая визуализация результатов GWAS в виде Manhattan plot [1], показывающего статистическую значимость для каждого ДНК-варианта (рис. 1, а).

Пользователи могут скачать данные GWAS в виде таблицы, а также отфильтровать ДНК-варианты в зависимости от их влияния на белки, кодируемые соответствующими генами.

Дополнительно на странице фенотипа представлены показатели качества GWAS в виде графика квантилей р-значений, а также генетические корреляции данного фенотипа с другими фенотипами. Все ДНК-варианты, для которых р-значение меньше 1×10^{-6} , отображаются на Manhattan plot как интерактивные элементы. При наведении на них появляется дополнительная информация, например частота варианта в российской популяции. При клике на конкретный вариант пользователю предоставляется возможность приблизить генетические координаты до масштаба отдельных генов (рис. 1, b).

Кроме того, для каждого ДНК-варианта доступны все известные генетические ассоциации, полученные из таких баз данных, как GWAS-каталог и dbSNP

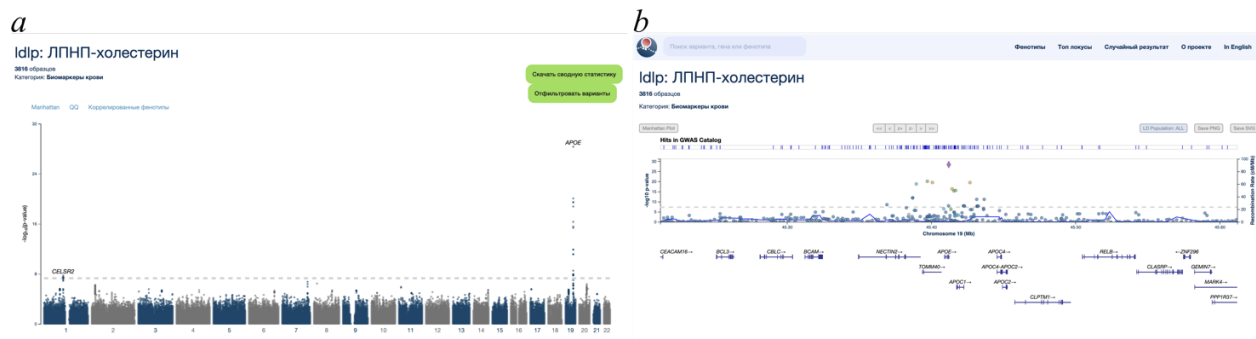


Рис. 1. Биобанк России, а – страница фенотипа (ЛПНП-холестерин), b – приближение генетических координат до масштабов отдельных генов.

Биобанк России является ресурсом, содержащим информацию о 7,5 миллионах ДНК-вариантов, встречающихся в российской популяции [4]. На странице каждого ДНК-варианта предоставлена информация об ассоциациях данного варианта со всеми фенотипами, включенными в базу данных (рис. 2, a) [5]. Это позволяет пользователю исследовать связь между конкретными генетическими вариациями и различными признаками здоровья. Кроме того, ассоциации для всех вариантов, генетические позиции которых пересекаются с известными генами, представлены на отдельной странице гена (рис. 2, b). Это дает возможность пользователям исследовать, как изменения в определенных генах могут влиять на фенотипы, а также как генетические взаимодействия могут способствовать развитию различных заболеваний.

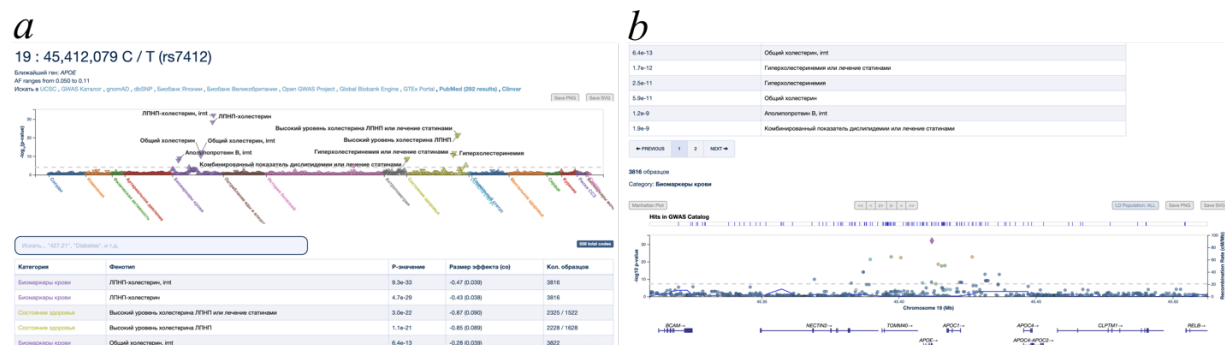


Рис. 2. Биобанк России, а – страница ДНК-варианта (rs7412), b – страница гена (APOE)

Онлайн платформа для предсказания рисков развития метаболически нездорового ожирения. Проблема ожирения становится все более актуальной в России и продолжает набирать масштабы. Одним из наиболее опасных типов ожирения является метаболически нездоровое ожирение — состояние, при котором ожирение сопровождается нарушениями метаболического профиля, что значительно повышает риски сердечно-сосудистых заболеваний, диабета второго типа и преждевременной смерти [6].

Фенотипическая информация Биобанка России представляет собой ценный ресурс для оценки рисков заболеваний в российской популяции. В рамках исследования была разработана онлайн-платформа для оценки рисков перехода от метаболически здорового ожирения к метаболически нездоровому. Для этого использованы данные 6-летнего наблюдения за 58 пациентами, изначально имеющими метаболически здоровое ожирение. Платформа позволяет предсказать вероятность развития метаболически нездорового ожирения на основе различных факторов с помощью такой модели машинного обучения как случайный лес. ROC AUC модели составлял 0,86 [0,76 – 0,97] (рис. 3).

MH sustainability predictor MHO Terms of use FAQ Tutorial Contacts Citation

Single sample input

Age 30	Have you taken any blood sugar medication for the last 2 weeks?*** ☐ yes ☒ no ☐ I don't know	Systolic blood pressure, mmHg* 127	Diastolic blood pressure, mmHg** 80
Gender* ☒ male ☐ female	Have you taken cholesterol-lowering drugs in the last 2 weeks?*** ☐ yes ☒ no ☐ I don't know	Insulin, pmol/L* 82.95	Height, cm*** 164.5
Have your doctor ever told you that you have a high blood pressure?* ☐ yes ☒ no ☐ I don't know	How often do you consume a cottage cheese?* ☒ I do not consume ☐ 1-2 times a month ☐ 1-2 times a week ☐ Daily ☐ I don't know	Total cholesterol, mmol/L* 5.585	Weight, kg*** 91.1
Have you taken any blood pressure medications for the last 2 weeks?*** ☐ yes ☒ no ☐ I don't know		Low density lipoproteins, mmol/L* 3.645	Waist circumference, cm** 101.25
		High density lipoproteins, mmol/L** 1.385	Glucose, mmol/L* 5.115
		Uric acid, µmol/L* 315.5	Triglycerides, mmol/L** 1.055

* Highly important features.
 ** Needed to determine metabolic health status.
 *** Needed for BMI calculation to determine obesity status.

Apply

Bulk upload for multisample input

Рис. 3. Онлайн платформа для оценки риск развития метаболически нездорового ожирения

Выводы. Биобанк России представляет собой крупнейшее GWAS-исследование в российской популяции, результаты которого стали широкодоступными научному сообществу, в том числе благодаря онлайн-платформе, представленной в данной работе. Несмотря на относительно небольшой размер когорты, представленные на платформе результаты GWAS могут быть успешно использованы в метаанализах крупных мировых биобанков, а информация о популяционной частоте ДНК-вариантов может служить референсом для дальнейших генетических исследований в России. Онлайн-платформа для предсказания рисков развития метаболически нездорового ожирения может стать полезным инструментом для приоритизации пациентов, находящихся в зоне риска, что способствует раннему вмешательству и профилактике заболеваний.

Литература:

1. Uffelmann E. et al. Genome-wide association studies // Nat Rev Methods Primers. 2021. V. 1. P. 59.
2. Bycroft C. et al. The UK Biobank resource with deep phenotyping and genomic data // Nature. 2018. V. 562. P. 203–209.
3. Martin A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities // Nat. Genet. 2019. V. 51. P. 584–591.
4. Бойцов С.А. et al. Исследование ЭССЕ-РФ (Эпидемиология сердечно-сосудистых заболеваний и их факторов риска в регионах Российской Федерации). Десять лет спустя // Кардиоваскулярная терапия и профилактика. 2021. V. 20. N 5. P. 3007.
5. Gagliano Taliun, S. A. et al. Exploring and visualizing large-scale genetic associations by using PheWeb // Nat. Genet. 2020. V. 52. P. 550–552.
6. Iacobini C. et al. Metabolically healthy versus metabolically unhealthy obesity // Metabolism. 2019. V. 92. P. 51–60.