Вавиловский журнал генетики и селекции. 2023;27(7):728-736
DOI 10.18699/VJGB-23-85

Перевод на английский язык https://vavilov.elpub.ru/jour

Human_SNP_TATAdb – база данных о SNP, статистически достоверно изменяющих сродство TATA-связывающего белка к промоторам генов человека: полногеномный анализ и варианты использования

С.В. Филонов[1, 2], Н.Л. Подколодный[1, 3] ✉, О.А. Подколодная[1], Н.Н. Твердохлеб[1], П.М. Пономаренко[1], Д.А. Рассказов[1], А.Г. Богомолов[1], М.П. Пономаренко[1]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, Новосибирск, Россия
[3] Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, Новосибирск, Россия
✉ pnl@bionet.nsc.ru

Аннотация. Ранее было показано, что уровень экспрессии генов человека положительно коррелирует с аффинностью TBP к промоторам этих генов. В свою очередь, однонуклеотидные полиморфизмы (SNP) в промоторах генов человека могут влиять на аффинность белка TBP к ДНК и, как следствие, на экспрессию генов. В ИЦиГ СО РАН разработан метод предсказания аффинности TBP к промоторам генов на основе трехшагового механизма связывания, включающего скольжение TBP по ДНК, остановку TBP в месте связывания, фиксацию комплекса TBP–промотор за счет изгиба спирали ДНК. Метод показал высокую корреляцию теоретических предсказаний с измеренными значениями при многократной экспериментальной проверке независимыми группами исследователей. На основе этой модели в ИЦиГ СО РАН ранее были разработаны веб-сервисы SNP_TATA_Z-tester и SNP_TATA_Comparator, позволяющие вычислять статистическую оценку вызванного SNP изменения аффинности связывания TBP с промотором гена человека и прогнозировать изменение экспрессии, которые могут быть связаны с генетической предрасположенностью к заболеваниям или фенотипическими особенностями организма. В настоящей работе проведена интеграция в единой базе данных информации об однонуклеотидных полиморфизмах в промоторах генов человека, полученной путем автоматической экстракции из различных гетерогенных источников данных, а также результатов оценки аффинности TBP к промотору с использованием трехшаговой модели связывания и оценки их влияния на экспрессию генов для промоторов дикого типа и промоторов с однонуклеотидным полиморфизмом. Показана возможность использования базы данных Human_SNP_TATAdb для аннотации и выявления кандидатных SNP-маркеров заболеваний. Представлены результаты полногеномного анализа данных, включая особенности распределения генов по количеству транскриптов, распределение SNP, влияющих на аффинность TBP к ДНК по позициям внутри промоторов, а также закономерности, связывающие между собой аффинность TBP к промотору, специфичность сайта связывания TBP с промотором и другие характеристики промоторов. Результаты полногеномного анализа показали, что аффинность TBP к промотору и специфичность его сайта связывания статистически связаны с другими характеристиками промоторов, важными для функциональной классификации промоторов и исследования особенностей дифференциальной экспрессии генов.

Ключевые слова: TATA-бокс; аффинность; TBP; однонуклеотидный полиморфизм; база данных; полногеномный анализ.

Для цитирования: Филонов С.В., Подколодный Н.Л., Подколодная О.А., Твердохлеб Н.Н., Пономаренко П.М., Рассказов Д.А., Богомолов А.Г., Пономаренко М.П. Human_SNP_TATAdb – база данных о SNP, статистически достоверно изменяющих сродство TATA-связывающего белка к промоторам генов человека: полногеномный анализ и варианты использования. *Вавиловский журнал генетики и селекции*. 2023;27(7):728-736. DOI 10.18699/VJGB-23-85

Human_SNP_TATAdb: a database of SNPs that statistically significantly change the affinity of the TATA-binding protein to human gene promoters: genome-wide analysis and use cases

S.V. Filonov[1, 2], N.L. Podkolodnyy[1, 3] ✉, O.A. Podkolodnaya[1], N.N. Tverdokhleb[1], P.M. Ponomarenko[1], D.A. Rasskazov[1], A.G. Bogomolov[1], M.P. Ponomarenko[1]

[1] Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Institute of Computational Mathematics and Mathematical Geophysics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
✉ pnl@bionet.nsc.ru

Abstract. It was previously shown that the expression levels of human genes positively correlate with TBP affinity for the promoters of these genes. In turn, single nucleotide polymorphisms (SNPs) in human gene promoters can affect TBP affinity for DNA and, as a consequence, gene expression. The Institute of Cytology and Genetics SB RAS (ICG) has developed a method for predicting TBP affinity for gene promoters based on a three-step binding mechanism: (1) TBP slides along DNA, (2) TBP stops at the binding site, and (3) the TBP-promoter complex is fixed due to DNA helix bending. The method showed a high correlation of theoretical predictions with measured values during repeated experimental testing by independent groups of researchers. This model served as a base for other ICG web services, SNP_TATA_Z-tester and SNP_TATA_Comparator, which make a statistical assessment of the SNP-induced change in the affinity of TBP binding to the human gene promoter and help predict changes in expression that may be associated with a genetic predisposition to diseases or phenotypic features of the organism. In this work, we integrated into a single database information about SNPs in human gene promoters obtained by automatic extraction from various heterogeneous data sources, as well as the estimates of TBP affinity for the promoter obtained using the three-step binding model and predicting their effect on gene expression for wild-type promoters and promoters with SNPs. We have shown that Human_SNP_TATAdb can be used for annotation and identification of candidate SNP markers of diseases. The results of a genome-wide data analysis are presented, including the distribution of genes with respect to the number of transcripts, the distribution of SNPs affecting TBP-DNA affinity with respect to positions within promoters, as well as patterns linking TBP affinity for the promoter, the specificity of the TBP binding site for the promoter and other characteristics of promoters. The results of the genome-wide analysis showed that the affinity of TBP for the promoter and the specificity of its binding site are statistically related to other characteristics of promoters important for the functional classification of promoters and the study of the features of differential gene expression.
Key words: TATA box; affinity; TBP; single nucleotide polymorphism; database; genome-wide analysis.

For citation: Filonov S.V., Podkolodnyy N.L., Podkolodnaya O.A., Tverdokhleb N.N., Ponomarenko P.M., Rasskazov D.A., Bogomolov A.G., Ponomarenko M.P. Human_SNP_TATAdb: a database of SNPs that statistically significantly change the affinity of the TATA-binding protein to human gene promoters: genome-wide analysis and use cases. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding*. 2023;27(7):728-736. DOI 10.18699/VJGB-23-85

Введение

Разработка методов предсказания влияния мутаций на уровень экспрессии генов различных организмов имеет важное значение для решения задач в области биотехнологии, селекции растений, медицины и так далее. Мутации в геноме человека могут быть ассоциированы со множеством физиологических особенностей и заболеваний, и знание о наличии и причине их безусловно необходимо для активно развивающегося подхода персонализированной медицины. Самым распространенным типом мутаций в геноме человека являются однонуклеотидные полиморфизмы (Single Nucleotide Polymorphism, SNP) – отличия последовательности ДНК размером в один нуклеотид. Однонуклеотидные полиморфизмы могут локализоваться в различных функциональных районах генома, от чего зависит характер их проявления. Наиболее изучены мутации в кодирующих районах гена, они непосредственно влияют на структуру транскрибируемой мРНК и синтезируемого белка. Однако полногеномные ассоциативные исследования (GWAS) показали, что большинство однонуклеотидных полиморфизмов, которые в значительной степени связаны с предрасположенностью к заболеванию, лежит в некодирующих областях (Hindorff et al., 2009; French, Edwards, 2020; Chandra et al., 2021), а более 90 % из них расположены в регуляторных элементах (Maurano et al., 2012). Одним из наиболее изученных регуляторных районов на данный момент является район TATA-бокса в промоторе, от последовательности которого зависит сродство к нему белка TBP (TATA Binding Protein), – ключевого фактора инициации транскрипции. Мутации в этом районе могут влиять на связывание белка TBP с промотором и, как следствие, на экспрессию гена (Савинкова и др., 2007).

В ИЦиГ СО РАН разработан метод предсказания аффинности TBP к промоторам генов на основе трехшагового механизма связывания (Пономаренко и др., 2008). Метод показал высокую корреляцию теоретических предсказаний с измеренными значениями аффинности при многократной экспериментальной проверке независимыми группами исследователей (Delgadillo et al., 2009; Savinkova et al., 2013; Oshchepkov et al., 2022). На основе этой модели в ИЦиГ СО РАН разработан веб-сервис SNP_TATA_Z-tester (Рассказов и др., 2013), позволяющий вычислять статистическую оценку вызванного SNP изменения аффинности связывания TBP с промотором гена человека и прогнозировать изменение экспрессии. С помощью этого веб-сервиса мы ранее выявили кандидатные SNP-маркеры аутоиммунных заболеваний (Ponomarenko et al., 2016а), поведенческих расстройств (Chadaeva et al., 2016), хронопатологий (Ponomarenko et al., 2016б) и других заболеваний.

В настоящей работе проведена интеграция в единой базе данных информации об однонуклеотидных полиморфизмах в промоторах генов человека, полученной путем автоматической экстракции из различных гетерогенных источников данных, а также результатов оценки аффинности TBP к промотору и специфичности сайта связывания TBP с использованием трехшаговой модели связывания и оценки их влияния на экспрессию генов для промоторов из референсного генома и промоторов с однонуклеотидным полиморфизмом.

Ключевым вариантом использования базы данных Human_SNP_TATAdb является аннотация промоторов и генов с целью поиска кандидатных SNP-маркеров заболеваний. Учитывая, что к настоящему времени уже вы-

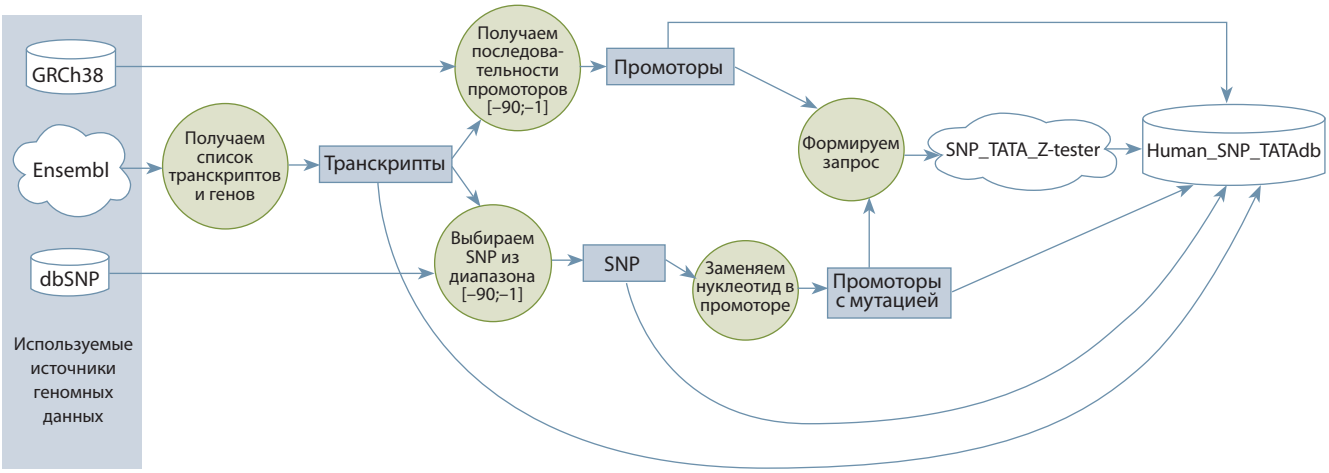

Рис. 1. Схема потока данных для инициализации базы данных Human_SNP_TATAdb.

полнено много исследований, в которых проводилась такого рода аннотация, мы привели в качестве примера один из вариантов.

Представлены результаты полногеномного анализа данных, включая особенности распределения генов по количеству транскриптов, распределение SNP, влияющих на аффинность TBP к ДНК по позициям внутри промоторов, а также закономерности, связывающие между собой аффинность TBP к промотору, специфичность сайта связывания TBP с промотором и другие характеристики промоторов, важные для функциональной классификации промоторов и исследования особенностей дифференциальной экспрессии генов.

Материалы и методы

Ниже представлены этапы работы по интеграции данных и создания базы данных (рис. 1). Данные о генах и их атрибутах, стартах транскрипции и транскриптах получены с веб-сервиса Ensembl (Birney et al., 2004). Для доступа к сервисам и базам данных использована библиотека Bioconductor языка R со следующими пакетами:

1. biomaRt[1] – пакет, который обеспечивает интерфейс для коллекции баз данных Ensembl, позволяя извлекать большие объемы данных унифицированным способом и использовать при анализе данных в Bioconductor.
2. BSgenome.Hsapiens.NCBI.GRCh38[2] – пакет, обеспечивающий доступ к последовательностям генома *Homo sapiens* (Human), предоставленным NCBI (GRCh38.p13).
3. SNPlocs.Hsapiens.dbSNP155.GRCh38[3] – пакет для доступа к dbSNP 155, включающий информацию о 949021448 SNP в хромосомах 1–22, X, Y и MT.

Для выявления старта транскрипции необходимо использовать транскрипты с качественной аннотацией, которая включает эту информацию и для которых доказана их биологическая релевантность. При описании транскриптов в Ensembl для определения наиболее качественно аннотированных ставят специальные метки. Мы включили в базу данных только те транскрипты, качество аннотации которых соответствует метке GENCODE Basic[4]. В соответствии со спецификацией Ensembl GENCODE Basic содержит по крайней мере один транскрипт для каждого гена в генетическом наборе GENCODE независимо от биотипа, т.е. каждый ген представлен в базовом наборе GENCODE. Для генов, кодирующих белок, в базовый набор GENCODE включены только полноразмерные транскрипты, кодирующие белок.

Для заданных координат старта транскрипции определяются координаты и нуклеотидные последовательности соответствующего им промотора ([–90; –1] от старта транскрипции). Данные о SNP получены с использованием базы данных dbSNP[5] (Sherry et al., 2001). Для каждого промотора выделены SNP, локализованные в пределах [–90; –1] от старта транскрипции. Минорные варианты последовательности промотора созданы автоматически путем внесения в основные варианты последовательностей соответствующих замен нуклеотидов из базы данных dbSNP (вып. 155). Для выявления TATA-содержащих промоторов использована весовая матрица Бухера (Bucher, 1990).

Аффинность TBP к ДНК рассчитывали с применением трехшаговой модели связывания, разработанной ранее в ИЦиГ СО РАН (Ponomarenko et al., 2008) и реализованной нами многопоточной высокопроизводительной версии программы SNP_TATA_Z-tester. Эта программа также позволяет оценить статистическую значимость изменения аффинности белка TBP к промотору при точечных заменах нуклеотидов (SNP) в промоторе с использованием z-критерия.

Аффинность, или сродство, TBP описывается константой ассоциации комплекса TBP/ДНК. Однако в настоящее время вместо константы ассоциации обычно используют обратную меру – константу диссоциации K_d. В этом случае аффинность TBP к ДНК, измеренная в наномолях на литр (нМ/л), будет равна $A = 10^9/K_d$. Чем меньше K_d, тем

[1] https://bioconductor.org/packages/release/bioc/html/biomaRt.html
[2] https://bioconductor.org/packages/release/data/annotation/html/BSgenome.Hsapiens.NCBI.GRCh38.html
[3] https://bioconductor.org/packages/release/data/annotation/html/SNPlocs.Hsapiens.dbSNP155.GRCh38.html
[4] https://www.ensembl.org/info/genome/genebuild/transcript_quality_tags.html
[5] https://www.ncbi.nlm.nih.gov/snp/

выше сродство TBP к промотору и сильнее взаимодействие TBP с промотором.

Второй вариант, представленный в базе данных, – логарифмическая форма аффинности $\alpha = 9*\ln(10) - \ln(K_d)$, которая удобна для сравнения показателей аффинности TBP к промотору, так как имеет близкое к нормальному распределение. При увеличении α возрастают сродство TBP к промотору и сила их взаимодействия.

Расчеты аффинности проведены для референсных последовательностей ДНК всех промоторов и минорных вариантов последовательностей этих промоторов с одним однонуклеотидным полиморфизмом. Для каждой минорной последовательности оценивали отклонение аффинности TBP к промотору от аффинности, полученной для последовательности ДНК промотора из референсного генома. При этом определяли уровень статистической значимости этих изменений.

Ранее показано, что аффинность TBP к промотору статистически достоверно коррелирует с уровнем экспрессии соответствующего транскрипта (Mogno et al., 2010). Поэтому при статистически достоверном увеличении или уменьшении аффинности TBP в базе данных указывается оценка соответствующего изменения уровня экспрессии транскрипта. На основе оценок аффинности белка TBP к промотору введены дополнительные характеристики, например специфичность сайта связывания белка TBP с промотором, который можно использовать для классификации промотора и биологической аннотации групп промоторов или генов.

Специфичность сайта связывания TBP с промотором гена соответствует максимальной нормированной аффинности TBP к промотору гена относительно средней аффинности TBP по каждой позиции скользящего окна (Ponomarenko et al., 2015), не включая 10 позиций ближайших к старту транскрипции (всего 55 значений). Специфичность Z рассчитывали следующим образом:

$$Z = \frac{\alpha_{\max} - \overline{\alpha}}{\sigma_\alpha}, \ \sigma_\alpha = \sqrt{\frac{1}{54}\sum_1^{55}(\alpha_i - \overline{\alpha})},$$

где α_i – оценка аффинности TBP к промотору в позиции i, $\overline{\alpha}$ – среднее значение α_i, σ_α– несмещенная оценка среднеквадратичного отклонения α_i, Z – специфичность сайта связывания белка TBP с промотором.

Еще один важный показатель, описывающий вызванное SNP изменение аффинности TBP к промотору, – натуральный логарифм отношения K_d для референсных (*wt*) и минорных (*mt*) аллелей рассматриваемого SNP:

$$k_{snp} = \ln(K_{d,\,wt}/K_{d,\,mt}).$$

Положительные или отрицательные значения k_{snp} указывают на то, что экспрессия гена для минорного аллеля соответственно выше или ниже, чем для случая референсного варианта. Этот показатель использовался для выявления кандидатных SNP-маркеров, которые могут быть связаны с генетической предрасположенностью к заболеваниям; в частности, сделаны предсказания, которые согласуются с клиническими данными о недостаточной экспрессии этого гена у пациентов с вариабельным иммунодефицитом, инсультом и преэклампсией (Ponomarenko et al., 2017).

Результаты и обсуждение

База данных

В ИЦиГ СО РАН разработана база данных Human_SNP_TATAdb (рис. 2). Базу данных заполняли в соответствии со сценарием интеграции данных и инициализации базы данных (см. рис. 1). База данных реализована на основе СУБД MySQL[6] версии 8.0 и включает 6 основных таблиц (chromosomes, genes, transcripts, snps, promoters, promoters_has_snps), 10 вспомогательных таблиц и словарей (см. рис. 2). Работа с базой данных осуществляется через SQL-запросы.

Таблица chromosomes включает идентификатор хромосомы, длину, количество нуклеотидов и вид организма.

Таблица genes содержит информацию об идентификаторах гена в разных базах данных, в том числе в Ensembl, символьное имя гена, ссылку на хромосому, цепь, биотип гена.

Таблица transcripts включает информацию об идентификаторах транскрипта, координаты транскрипта в геноме, биотип транскрипта и ссылку на промотор и ген.

Таблица snps включает следующую информацию: идентификаторы SNP, позиции SNP в геноме, ссылка на хромосому и аллель. За один SNP здесь и далее принимается однозначный вариант изменения генома. Полиморфизмы, имеющие один rs идентификатор, но допускающие несколько вариантов замены нуклеотида, считаются по количеству таких вариантов.

Необходимо отметить, что одна и та же нуклеотидная замена может попадать в разные промоторы гена и поразному изменять уровень аффинности белка TBP к этим промоторам, и поэтому в базе данных заданы две таблицы для описания промоторов promoters и promoters_has_snps с отношением 1:N (на один промотор может оказывать влияние несколько SNP), а таблицы snps и promoters_has_snps также связаны отношением 1:N (один SNP может входить в несколько промоторов).

В таблицу promoters включена следующая информация: идентификатор промотора, последовательность ДНК, соответствующая району [–90; –1] от старта транскрипции, координаты старта и конца промотора в геноме, аффинность белка TBP к промотору с ошибкой, ссылка на ген.

Таблица promoters_has_snps содержит информацию об идентификаторе промотора, ссылку на SNP, координаты SNP в промоторе и относительно старта транскрипции, последовательность промотора дикого типа и промотора с SNP, аффинность TBP к промотору с ошибкой, характер изменения экспрессии гена при мутации в промоторе, уровень значимости статистического теста.

Таблица source_snp_dbs включает информацию, которая необходима для автоматизированного обновления базы данных Human_SNP_TATAdb: об источниках данных, версии баз данных, ссылки на базы данных.

Типы отношений между таблицами задают ограничения, которые соответствуют природе данных и поэтому важны для сохранения целостности базы данных, а также обеспечивают дополнительный контроль данных и уменьшают возможность ошибок. В частности, у каждого гена

[6] https://www.mysql.com/

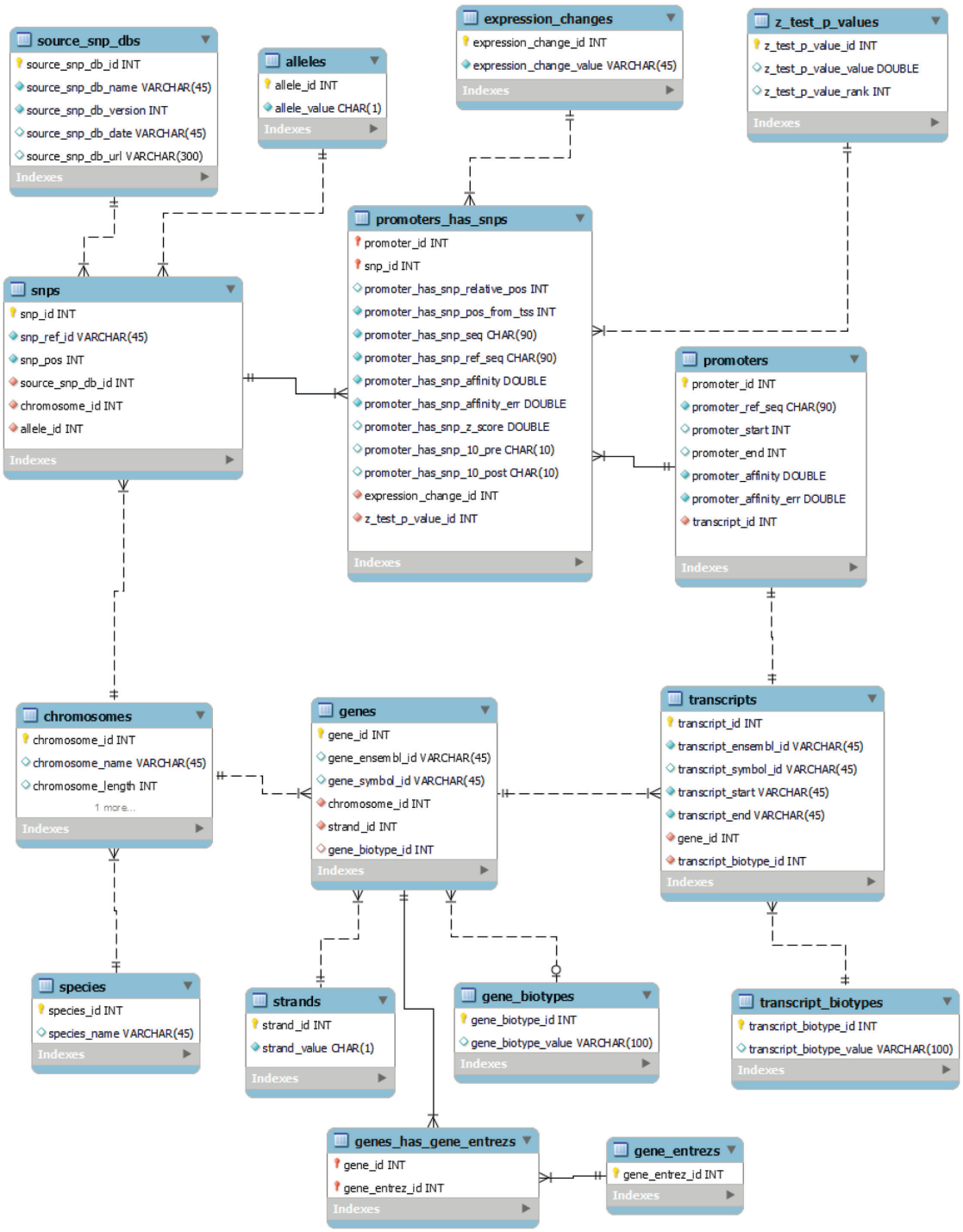

Рис. 2. Схема базы данных Human_SNP_TATAdb.

может быть один или несколько промоторов, каждый промотор может регулировать экспрессию одного или несколько транскриптов.

В итоге база данных содержит информацию о:

- 62603 генах, из которых 19314 кодируют белки;
- 117414 транскриптах, из которых 63141 кодирует белки;
- 5305816 вариантов SNP в промоторах генов в интервале [–90; –1] от старта транскрипции, из них 3199285 в промоторах белок кодирующих генов;
- для 445875 вариантов SNP в промоторе белок кодирующего гена предсказано, что они статистически значимо (p-value < 0.05) изменяют уровень аффинности TBP к этому промотору.

Варианты использования базы данных Human_SNP_TATAdb

Представленные в базе данных аффинность белка TBP к промотору, специфичность сайта связывания TBP с промотором и оценки изменения этих характеристик при

однонуклеотидном полиморфизме важны для поиска маркеров генетической предрасположенности заболеваний, выявления и функциональной интерпретации классов промоторов, схожих по механизму регуляции ранней стадии инициации транскрипции и так далее.

База данных Human_SNP_TATAdb также может быть использована для аннотации генов или группы генов в терминах аффинности TBP к промотору или специфичности сайта связывания TBP с промотором. Чтобы определить характеристику гена, связанную со спецификой связывания TBP с промоторами гена с целью проведения GO-анализа, можно использовать средние значения аффинности TBP к промоторам гена или аффинности TBP к промотору, соответствующему единственному для гена транскрипту, который определен экспертами Ensembl в качестве канонического и задается в базе данных меткой Ensembl Canonical[7], который в целом наиболее консервативен, наиболее экспрессируем, имеет самую длинную кодирующую последовательность и представлен в других ключевых ресурсах, таких как NCBI и UniProt. Мы помечаем соответствующий ему промотор как канонический и используем такие характеристики, как аффинность TBP к каноническому промотору и специфичность сайта связывания TBP с каноническим промотором, для аннотации гена или группы генов.

Корреляционный анализ показал, что между аффинностью TBP к каноническому промотору гена и средней аффинностью промоторов гена наблюдается сильная линейная зависимость (R = 0.88, d.f. = 19308), поэтому оба варианта дают сходные результаты. Однако использование аффинности TBP к каноническому промотору гена, по-видимому, биологически более обосновано. Безусловно, ключевым вариантом применения базы данных Human_SNP_TATAdb является аннотация генов и поиск кандидатных SNP-маркеров предрасположенности к заболеваниям.

Учитывая, что к настоящему времени уже выполнено много исследований, в которых проводилась такого рода аннотация, мы приведем в качестве примера использование базы данных Human_SNP_TATAdb для аннотации и выявления кандидатных SNP-маркеров атерогенеза, атеросклероза и атеропротекции работу (Bogomolov et al., 2023).

Предварительно были отобраны 1068 генов человека, связанных с этими заболеваниями. Информация о SNP в промоторах этих генов человека, результатах оценки аффинности TBP к промоторам и оценки их влияния на экспрессию генов для промоторов дикого типа и промоторов с однонуклеотидным полиморфизмом получена из базы данных Human_SNP_TATAdb. Эта информация была дополнена аннотацией отобранных генов, подготовленной экспертами, и сформировано представление базы данных, ориентированное на анализ генов, связанных с атерогенезом, атеросклерозом и атеропротекцией, внешний доступ к которой осуществляется через веб-интерфейс[8].

Анализ *in silico* всех 5112 SNP в их промоторах выявил 330 кандидатов в маркеры SNP, статистически значимо изменяющих аффинность TATA-связывающего белка (TBP) к этим промоторам. Далее сравнили соответствующие частоты SNP, которые увеличивают и уменьшают сродство TBP к промоторам одних и тех же генов. Сравнение было сделано для анализа того, находятся ли эти гены под действием естественного отбора или нейтрального дрейфа. Мы обнаружили, что естественный отбор действует против недостаточной экспрессии хаб-генов атерогенеза, атеросклероза и атерозащиты и благодаря усиленной атеропротекции способствует улучшению здоровья человека (Bogomolov et al., 2023).

Примеры использования базы данных Human_SNP_TATAdb для полногеномного анализа

Разработанная база данных позволяет проводить анализ полногеномной статистики и распределения указанных показателей в различных группах промоторов, например TATA-содержащих. Для полногеномного анализа мы использовали белок кодирующие гены и транскрипты, отобранные по значениям полей ‘gene_biotype’ и ‘transcript_biotype’ равными ‘protein_coding’.

Альтернативные промоторы и аффинность TBP/ДНК

Следует отметить, что один ген может иметь несколько транскриптов, инициация транскрипции которых происходит с использованием разных промоторов, для которых оценивается аффинность белка TBP. Как видно из рис. 3, наибольшее число белок кодирующих генов (29.77 % генов) имеет единственный транскрипт и, как следствие, один промотор. Пять процентов белок кодирующих генов имеют не менее 9 белок кодирующих транскриптов. Анализ распределения генов по числу транскриптов показал, что среднее число транскриптов на ген – 3.27, а медиана – 2 транскрипта на ген. Максимальное число (87) белок кодирующих транскриптов наблюдается у гена *Mapk10* (*mitogen-activated protein kinase 10*).

Наш анализ показал, что распределение средней аффинности TBP к каноническим промоторам в группах генов, разбитых по числу транскриптов, близко к равномерному. Таким образом, нет необходимости нивелировать эффекты, обусловленных разным числом транскриптов у гена при проведении анализа данных с использованием аффинности TBP.

Распределение SNP, изменяющих экспрессию генов по позициям промотора

Распределение SNP, статистически значимо изменяющих экспрессию генов по позициям от старта транскрипции, имеет ярко выраженное отклонение от равномерного (рис. 4). В районе [–35; –20], соответствующем обычному расположению TATA-бокса, число таких SNP заметно выше, чем в других районах промотора.

Число SNP, уменьшающих экспрессию генов в районе [–35; –20], соответствующие расположению TATA-бокса, более чем в полтора раза выше, чем в других районах промотора. Это может быть связано с тем, что SNP в этом районе, как правило, разрушают TATA-бокс.

Число SNP, увеличивающих экспрессию генов, выше на флангах наиболее частых локализаций TATA-бокса. Пики локализованы в –24 и –32 позициях от старта транскрипции. Следует отметить, что распределение всех SNP по позициям промоторов белок кодирующих генов равномерно.

[7] https://www.ensembl.org/info/genome/genebuild/canonical.html
[8] http://www.sysbio.ru/Human_SNP_TATAdb

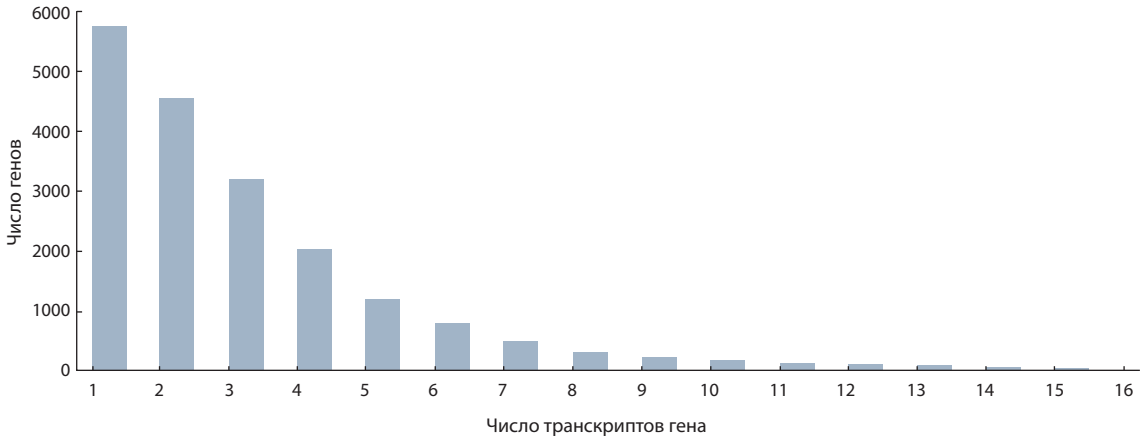

Рис. 3. Распределение белок кодирующих генов по числу транскриптов.

Рис. 4. Распределение числа SNP, увеличивающих (excees) и уменьшающих (deficiency) аффинность TBP к ДНК промоторов белок кодирующих генов в зависимости от позиции SNP относительно старта транскрипции.

Это говорит о том, что увеличение на флангах TATA-бокса числа SNP, увеличивающих экспрессию генов, может иметь функциональное значение.

Аффинность TBP к TATA-содержащим и TATA-не содержащим промоторам белок кодирующих генов

Анализ зависимости показателей аффинности TBP/ДНК, измеренной в логарифмической шкале ($\alpha = 9*\ln(10) - \ln(K_d)$) для TATA-содержащих и TATA-не содержащих промоторов белок кодирующих генов (рис. 5), показал, что в группе TATA-содержащих промоторов наблюдается более высокая аффинность TBP/ДНК, что соответствует более сильному сходству TBP к промотору.

Функциональные SNP, влияющие на аффинность TBP к ДНК промоторов и специфичность сайта связывания белка TBP

Проведен анализ зависимости доли SNP, статистически значимо влияющих на аффинность TBP к ДНК промоторов белок кодирующих генов, от специфичности сайта связывания белка TBP (рис. 6). Показано, что SNP в промоторах с низкой специфичностью сайта связывания TBP с промотором, как правило, приводят к увеличению экспрессии генов, а в промоторах с высокой специфичностью доля SNP, понижающих экспрессию, повышена.

Анализ таблицы сопряженности показал, что низкие значения специфичности сайта связывания TBP с промотором (Spec < 2.5) чаще наблюдаются на промоторах без TATA-бокса (TATA–) ($\chi^2 = 10385$, p-value < 1.0e–228).

Заключение

В настоящей работе описана база данных Human_SNP_TATAdb, которая включает информацию о SNP в промоторах генов человека, полученную путем автоматической экстракции из различных гетерогенных источников данных, результатах оценки аффинности TBP к промотору с использованием трехшаговой модели связывания и оценки их влияния на экспрессию генов для промоторов дикого типа и промоторов с однонуклеотидным полиморфизмом.

Представленные в базе данных аффинность белка TBP к промотору, специфичность сайта связывания TBP с про-

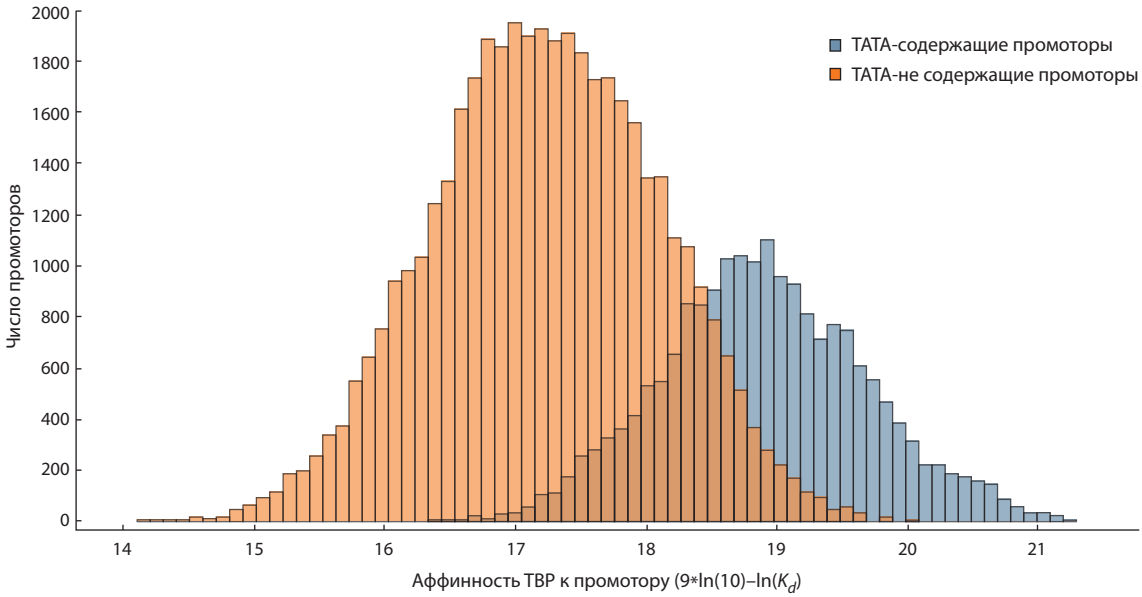

Рис. 5. Распределение промоторов белок кодирующих генов по аффинности TBP в группах TATA-содержащих промоторов и промоторах без TATA-бокса. Оценка аффинности TBP к промотору по оси *x* задается в логарифмической шкале.

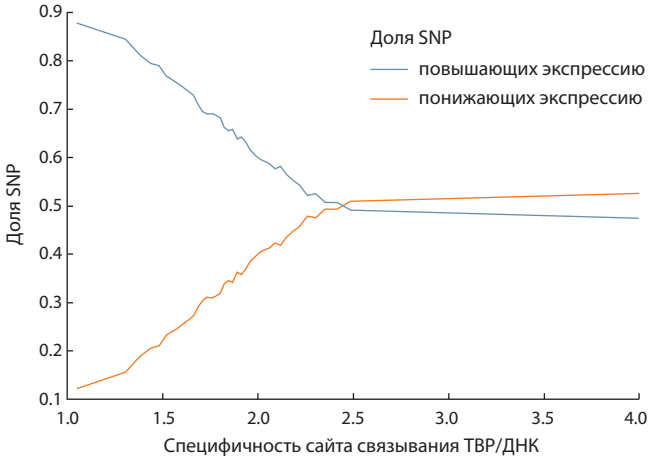

Рис. 6. Доля SNP в промоторах, повышающих и понижающих экспрессию белок кодирующих генов в зависимости от специфичности сайта связывания TBP к ДНК промотору.

Таблица сопряженности специфичности сайта связывания TBP с промотором и наличия TATA-бокса в промоторе

Специфичность	TATA–	TATA+	Всего
Spec < 2.5	**29114**	10379	39493
Spec ≥ 2.5	14538	9109	23647
Всего	43652	19488	63140

мотором и оценки изменения этих характеристик при SNP важны для поиска кандидатных маркеров генетической предрасположенности заболеваний, выявления и функциональной интерпретации классов промоторов, схожих по механизму регуляции ранней стадии инициации транскрипции, и так далее. База данных Human_SNP_TATAdb также может быть использована для аннотации генов или групп генов в терминах аффинности TBP к промотору или специфичности сайта связывания TBP с промотором.

Результаты полногеномного анализа показали, что аффинность TBP к промотору и специфичность его сайта связывания статистически связаны с другими характеристиками промоторов, важными для функциональной классификации промоторов и исследования особенностей дифференциальной экспрессии генов. Использование базы данных Human_SNP_TATAdb для аннотации генов и выявление кандидатных SNP-маркеров атерогенеза, атеросклероза и атеропротекции – один из примеров, в результате которого становятся доступны новые знания о влиянии различных одиночных полиморфизмов на предрасположенность к тем или иным заболеваниям.

Список литературы / References

Рассказов Д.А., Гунбин К.В., Пономаренко П.М., Вишневский О.В., Пономаренко М.П., Афонников Д.А. SNP_TATA_COMPARATOR: web-сервис применения уравнения равновесия TBP/TATA-комплекса в сравнительной оценке SNPS промоторов генов, связанных с болезнями человека. *Вавиловский журнал генетики и селекции.* 2013;17(4/1):599-606
[Rasskazov D.A., Gunbin K.V., Ponomarenko P.M., Vishnevsky O.V., Ponomarenko M.P., Afonnikov D.A. SNP_TATA_COMPARATOR: web service for comparison of SNPS within gene promoters associated with human diseases using the equilibrium equation of the TBP/TATA complex. *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding.* 2013;17(4/1):599-606 (in Russian)]

Савинкова Л.К., Драчкова И.А., Пономаренко М.П., Лысова М.В., Аршинова Т.В., Колчанов Н.А. Взаимодействие рекомбинантного TATA-связывающего белка с TATA-боксами промоторов генов млекопитающих. *Экологическая генетика.* 2007;5(2):44-49. DOI 10.17816/ecogen5244-49

[Savinkova L.K., Drachkova I.A., Ponomarenko M.P., Lysova M.V., Arshinova T.V., Kolchanov N.A. Interaction of recombinant TATA-binding protein with mammals gene promoter TATA boxes. *Ekologicheskaya genetika = Ecological genetics.* 2007;5(2):44-49. DOI 10.17816/ecogen5244-49 (in Russian)]

Birney E., Andrews T.D., Bevan P., Caccamo M., Chen Y., Clarke L., Coates G., ..., Cox A., Hubbard T., Clamp M. An overview of Ensembl. *Genome Res*. 2004;14(5):925-928. DOI 10.1101/gr.1860604

Bogomolov A., Filonov S., Chadaeva I., Rasskazov D., Khandaev B., Zolotareva K., Kazachek A., … Kolchanov N., Tverdokhleb N., Ponomarenko M. Candidate SNP markers significantly altering the affinity of TATA-binding protein for the promoters of human hub genes for atherogenesis, atherosclerosis and atheroprotection. *Int. J. Mol. Sci.* 2023;24(10):9010. DOI 10.3390/ijms24109010

Bucher P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol*. 1990;212(4):563-578. DOI 10.1016/0022-2836(90)90223-9

Chadaeva I.V., Ponomarenko M.P., Rasskazov D.A., Sharypova E.B., Kashina E.V., Matveeva M.Yu., Arshinova T.V., Ponomarenko P.M., Arkova O.V., Bondar N.P., Savinkova L.K., Kolchanov N.A. Candidate SNP markers of aggressiveness-related complications and comorbidities of genetic diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *BMC Genomics.* 2016;17(Suppl. 14):995. DOI 10.1186/s12864-016-3353-3

Chandra V., Bhattacharyya S., Schmiedel B.J., Madrigal A., Gonzalez-Colin C., Fotsing S., Crinklaw A., Seumois G., Mohammadi P., Kronenberg M., Peters B., Ay F., Vijayanand P. Promoter interacting expression quantitative trait loci are enriched for functional genetic variants. *Nat. Genet*. 2021;53(1):110-119. DOI 10.1038/s41588-020-00745-3

Delgadillo R.F., Whittington J.E., Parkhurst L.K., Parkhurst L.J. The TATA-binding protein core domain in solution variably bends TATA sequences via a three-step binding mechanism. *Biochemistry.* 2009; 48(8):1801-1809. DOI 10.1021/bi8018724

French J.D., Edwards S.L. The role of noncoding variants in heritable disease. *Trends Genet.* 2020;36(11):880-891. DOI 10.1016/j.tig.2020.07.004

Hindorff L.A., Sethupathy P., Junkins H.A., Manolio T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA*. 2009; 106(23):9362-9367. DOI 10.1073/pnas.0903103106

Maurano M.T., Humbert R., Rynes E., Thurman R.E., Haugen E., Wang H., Reynolds A.P., … Sunyaev S.R., Kaul R., Stamatoyannopoulos J.A. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337(6099):1190-1195. DOI 10.1126/science.1222794

Mogno I., Vallania F., Mitra R.D., Cohen B.A. TATA is a modular component of synthetic promoters. *Genome Res*. 2010;20(10):1391-1397. DOI 10.1101/gr.106732.110

Oshchepkov D., Chadaeva I., Kozhemyakina R., Zolotareva K., Khandaev B., Sharypova E., Ponomarenko P., Bogomolov A., Klimova N.V., Shikhevich S., Redina O., Kolosova N.G., Nazarenko M., Kolchanov N.A., Markel A., Ponomarenko M. Stress reactivity, susceptibility to hypertension, and differential expression of genes in hypertensive compared to normotensive patients. *Int. J. Mol. Sci*. 2022;23(5):2835. DOI 10.3390/ijms23052835

Ponomarenko P.M., Savinkova L.K., Drachkova I.A., Lysova M.V., Arshinova T.V., Ponomarenko M.P., Kolchanov N.A. A step-by-step model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism. *Dokl. Biochem. Biophys*. 2008;419:88-92. DOI 10.1134/S1607672908020117

Ponomarenko M., Rasskazov D., Arkova O., Ponomarenko P., Suslov V., Savinkova L., Kolchanov N. How to use SNP_TATA_Comparator to find a significant change in gene expression caused by the regulatory SNP of this gene's promoter via a change in affinity of the TATA-binding protein for this promoter. *Biomed Res. Int*. 2015;2015:359835. DOI 10.1155/2015/359835

Ponomarenko M.P., Arkova O., Rasskazov D., Ponomarenko P., Savinkova L., Kolchanov N. Candidate SNP markers of genderbiased autoimmune complications of monogenic diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *Front. Immunol*. 2016a;7:130. DOI 10.3389/fimmu.2016.00130

Ponomarenko P., Rasskazov D., Suslov V., Sharypova E., Savinkova L., Podkolodnaya O., Podkolodny N.L., Tverdokhleb N.N., Chadaeva I., Ponomarenko M., Kolchanov N. Candidate SNP markers of chronopathologies are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *Biomed Res. Int.* 2016б;2016:8642703. DOI 10.1155/2016/8642703

Ponomarenko M., Rasskazov D., Chadaeva I., Sharypova E., Ponomarenko P., Arkova O., Kashina E., Ivanisenko N., Zhechev D., Savinkova L., Kolchanov N. SNP_TATA_Comparator: genomewide landmarks for preventive personalized medicine. *Front. Biosci.* (*Schol. Ed.*). 2017;9(2):276-306. DOI 10.2741/s488

Savinkova L., Drachkova I., Arshinova T., Ponomarenko P., Ponomarenko M., Kolchanov N. An experimental verification of the predicted effects of promoter TATA-box polymorphisms associated with human diseases on interactions between the TATA boxes and TATA-binding protein. *PLoS One*. 2013;8(2).e54626. DOI 10.1371/journal.pone.0054626

Sherry S.T., Ward M.H., Kholodov M., Baker J., Phan L., Smigielski E.M., Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29(1):308-311. DOI 10.1093/nar/29.1.308

ORCID ID

N.L. Podkolodnyy orcid.org/0000-0001-9132-7997
O.A. Podkolodnaya orcid.org/0000-0003-3247-0114
P.M. Ponomarenko orcid.org/0000-0003-2715-9612
D.A. Rasskazov orcid.org/0000-0003-4795-0954
A.G. Bogomolov orcid.org/0000-0003-4359-6089
M.P. Ponomarenko orcid.org/0000-0003-1663-318X

Благодарности. Работа выполнена при поддержке бюджетных проектов FWNR-2022-0020, № 0251-2022-0005 и Федеральной научно-технической программы развития генетических технологий России.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию 22.08.2023. После доработки 15.09.2023. Принята к публикации 19.09.2023.