



УДК 57.012+57.016.4

# СТРУКТУРНЫЙ АНАЛИЗ БЕЛКА LZTFL1 МЕТОДОМ ГЛАВНЫХ КОМПОНЕНТ (PCA-seq)

© 2024 г. И. И. Хегай\*, #, С. Юй\*\*, В. М. Ефимов\*, \*\*

\* ФИЦ Институт цитологии и генетики СО РАН, Россия, 630090 Новосибирск, просп. Акад. Лаврентьева, 10

\*\* Новосибирский государственный университет, Россия, 630090 Новосибирск, ул. Пирогова, 2

Поступила в редакцию 05.02.2024 г.

После доработки 16.02.2024 г.

Принята к публикации 17.02.2024 г.

Однонуклеотидная мутация rs17713054G>A в промоторной области гена *LZTFL1* (leucine zipper transcription factor like 1) относится к факторам тяжелого течения коронавирусной инфекции COVID-19. Компьютерно-статистический анализ аминокислотной последовательности белка LZTFL1 методом главных компонент (PCA-seq) выявил наличие высокой корреляции между первой главной компонентой транскрибируемой аминокислотной последовательности и одиннадцатью аминокислотными индексами базы данных AAindex, характеризующими физико-химические и биохимические свойства белка. Индексы BEGF750102, CHOP780209, PALJ810110, GEIM800107, QIAN880121, LEVM780102 и PRAM900103 связаны с параметрами β-складчатой укладки. Белок LZTFL1 входит в состав белковых комплексов синдрома Барде–Бидля (BBS), регулирующих внутриклеточный транспорт в мерцательном эпителии легких. Предполагается, что присутствие β-складчатых элементов в структуре белка LZTFL1 играет важную роль в эндоцитозе, опосредованном рецептором ACE2, стимулируя интенсивность рециркуляции ангиотензинпревращающего фермента ACE2 и ускоряя доставку адгезированных вирионов коронавируса SARS-CoV-2 внутрь клетки в процессе инициации тяжелого острого респираторного синдрома COVID-19.

**Ключевые слова:** белок LZTFL1, мерцательный эпителий, рецептор ACE2, аминокислотный индекс, BBS-белки, метод главных компонент

**DOI:** 10.31857/S0132342324060075, **EDN:** NfJQUO

## ВВЕДЕНИЕ

Белок LZTFL1 представляет собой продукт повсеместно экспрессируемого гена *LZTFL1* и локализуется преимущественно в цитоплазме на поверхности внутриклеточных органелл. Белок входит в состав ресничек мерцательного или реснитчатого эпителия и вовлечен в регуляцию связывания других белков на плазматической мембране для последующей их доставки внутрь клетки [1, 2]. Мерцательный эпителий образован монослоем реснитчатых клеток и выстилает все воздухоносные пути, от носовой полости до трахей и бронхов. На апикальном конце данные клетки

имеют множественные отростки цитоплазмы в виде ресничек с повышенным содержанием микротрубочек. Реснички имеют особый белковый и липидный состав, обеспечивающий селекцию внешних сигнальных молекул и внутриклеточный трафик [3, 4]. В 2021 г. было установлено, что ген *LZTFL1* может быть фактором тяжелого течения COVID-19. Методом полногеномного поиска ассоциаций (GWAS) была показана корреляция мутаций в хромосомном локусе 3p21.31 с риском дыхательной недостаточности. В составе локуса идентифицирован структурный ген *LZTFL1* и мутантный аллель rs17713054G>A, содержащий

Сокращения: LZTFL1 – транскрипционный фактор 1, подобный лейциновой застежке-молнии (leucine zipper transcription factor like 1); COVID-19 – коронавирусное заболевание (COroNaVirus Disease); PCA-seq – метод главных компонент (principal component analysis); AAindex – аминокислотный индекс (amino acid index); BBS – синдром Барде–Бидля (Bardet–Biedl Syndrome); ACE2 – ангиотензинпревращающий фермент 2 (angiotensin-converting enzyme 2); SARS-CoV-2 – коронавирус 2, связанный с тяжелым острым респираторным синдромом (severe acute respiratory syndrome-related coronavirus 2); RBD – домен, связывающийся с рецептором (receptor-binding domain); Arl6 – ГТФаза 6, подобная фактору рибозилирования АДФ (ADP Ribosylation Factor Like GTPase 6); CEBPB – транскрипционный фактор бета, связывающийся с энхансерным мотивом CCAAT (CCAAT/enhancer-binding protein beta).

# Автор для связи: (эл. адрес: khegay@bionet.nsc.ru).

однонуклеотидную замену в регуляторной области гена [5].

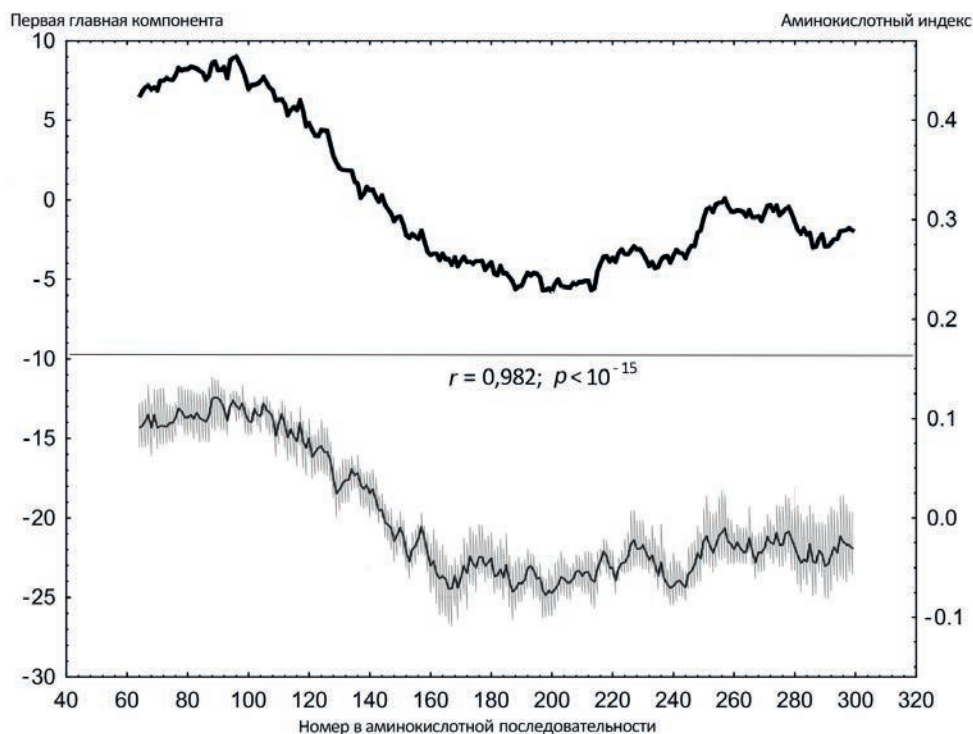
На фоне масштабных молекулярно-биологических исследований геногеографии распределения полиморфных вариантов локуса 3p21.31 и их связи с этиологией COVID-19 [6–8] актуальной задачей становится анализ функции кодируемого белка LZTFL1 в патогенезе коронавируса SARS-CoV-2. Вследствие того, что структура белка остается слабо изученной, целью данной работе было провести компьютерно-статистический анализ аминокислотной последовательности белка LZTFL1 методом главных компонент (PCA-seq).

## РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

На основании теоремы Такенса первичная последовательность белка LZTFL1 была преобразована в траекторную матрицу, по которой были вычислены главные компоненты аминокислотной последовательности относительно набора фрагментов длиной 64 а.о., непрерывно считываемых слева направо со сдвигом рамки считывания в одну аминокислоту в каждой следующей серии. На рис. 1 в верхней части представлена траектория первой главной компоненты аминокислотной последовательности LZTFL1 относительно номера фрагмента белка. Индивидуальные свой-

ства аминокислот учитывались в форме аминокислотных индексов. Аминокислотные индексы представляют наборы из 20 числовых значений, рассчитанных на основании кластерного анализа аминокислот и коэффициентов корреляции с отдельными физическими параметрами. В нижней части рис. 1 сплошной линией показана арифметическая средняя 11 аминокислотных индексов, наиболее коррелирующих по модулю с первой главной компонентой. Диапазон соответствующих минимальных и максимальных значений отображен серым фоном. Очевидно, что верхняя и нижняя кривые имеют практически идентичную форму, коэффициент корреляции между первой главной компонентой и средней 11 аминокислотных индексов равен 0.982, что соответствует  $pB\text{-value} < 10^{-15}$ .

В результате корреляционного анализа количественных параметров физико-химических свойств и первой главной компоненты LZTFL1 выявлено 11 аминокислотных индексов с коэффициентом корреляции  $>0.900$ , их список представлен в табл. 1. Аминокислотные индексы DESM900101 и RACS820105 связаны с внутриклеточной локализацией белка. Индексы PONP800106 и CHAM830103 ориентированы, соответственно, на гидрофобные свойства и размеры боковых



**Рис. 1.** Динамика позиционной изменчивости первой главной компоненты аминокислотной последовательности белка LZTFL1 (вверху) и нормированных аминокислотных индексов (внизу). Серым фоном обозначен разброс данных.

**Таблица 1.** Аминокислотные индексы с очень высоким коэффициентом корреляции с первой главной компонентой LZTFL1

№	Аминокислотный индекс	Коэффициент корреляции	Краткое описание
1	BEGF750102	0.964	Конформационные параметры $\beta$ -структуры
2	CHOP780209	0.958	Нормированная частота C-концевого $\beta$ -листа
3	DESM900101	0.941	Параметры трансмембранного домена
4	PALJ810110	0.940	Нормированная частота $\beta$ -листов в составе $\beta$ -структур
5	RACS820105	0.937	Средняя встречаемость во внешнем слое эндоплазматического ретикулума
6	GEIM800107	0.931	Индекс $\beta$ -цепей для $\alpha/\beta$ -элементов вторичной структуры
7	QIAN880121	0.916	Вес $\beta$ -листа в первом позиционном окне
8	PONP800106	0.909	Гидрофобность в области точки перегиба
9	CHAM830103	0.908	Количество атомов в боковой цепи
10	LEVM780102	0.908	Нормированная взвешенная частота $\beta$ -листов
11	PRAM900103	0.908	Относительная частота в составе $\beta$ -листа

цепей. Остальные аминокислотные индексы относятся к кластеру организации  $\beta$ -складчатых элементов в структуре белка. Максимальный коэффициент корреляции установлен для индекса BEGF750102. Данный аминокислотный индекс обязательно учитывается при статистическом анализе и построении конформации  $\beta$ -цепи, а следующий за ним в таблице индекс CHOP780209 принимается в расчет при оценке частоты присутствия антипараллельных  $\beta$ -структур на C-конце белка.

Таким образом, большинство аминокислотных индексов с наиболее высокими коэффициентами корреляции с первой главной компонентой аминокислотной последовательности LZTFL1 преимущественно связаны с параметрами  $\beta$ -складчатой укладки.

Начальное событие коронавирусной инфекции SARS-CoV-2 – взаимодействие S-белков шпиковых пепломеров вирусной оболочки с мембранными белками заражаемой клетки. Функцию основных рецепторов SARS-CoV-2 выполняют молекулы ангиотензинпревращающего фермента ACE2 (КФ 3.4.15.1), локализованные на плазматической мембране клетки хозяина. Для агрегации S-белка и ACE2 характерна высокая аффинность связывания с чувствительностью на уровне наномолярных концентраций [9]. Во взаимодействии участвуют RBD (рецептор-связывающий домен) S-белка и N-концевой пептидазный домен ACE2. Особенность RBD – наличие антипараллельного  $\beta$ -листа из семи  $\beta$ -цепей в структуре рецептор-связывающего мотива, непосредственно контактирующего с активным центром пептидазного домена ACE2 [10, 11]. Кар-

боксипептидаза ACE2 состоит из внеклеточного акцепторного участка, трансмембранного домена и цитоплазматического хвоста. Каталитический центр локализован во внеклеточной области и одновременно является рецептором для коронавируса SARS-CoV-2 [12]. Рецепция вируса усиливает ферментативную активность N-концевого пептидазного домена в молекуле ACE2 [13]. ACE2 экспрессируется в эпителиальных тканях большинства органов, но преимущественно локализуется в легочных превоцитах II типа, энтероцитах тонкого кишечника и эндотелиальных клетках артерий и вен [14, 15]. ACE2, встроенный в апикальную мембрану альвеолярных бронхиол и альвеол, служит котранспортером для SARS-CoV-2 в клетки легких. Проникновение вируса в цитоплазму клеток-мишеней происходит за счет рециркуляции ACE2 вследствие регулируемого эндоцитоза [16]. Данный процесс – ключевой для вирусной инфекции.

Методом полногеномного поиска ассоциаций GWAS (genome-wide association studies) было установлено, что тяжесть патогенеза COVID-19 коррелирует с полиморфизмом в хромосомном локусе 3p21.31 [5]. На данном участке хромосомы локализован ген *LZTFL1*, кодирующий цитоплазматический белок – регулятор внутриклеточного транспорта. В настоящее время установлено, что белок LZTFL1 входит в семейство BBS-белков и, согласно данной номенклатуре, обозначается как BBS17 [17]. Белки семейства BBS образуют сложные гетерополимерные комплексы из различных типов BBS-белков, одни из которых взаимодействуют с мембранными гликопротеинами, другие – с внутриклеточным

цитоскелетом. Октамерные комплексы BBS-белков функционируют как адаптеры регулируемого эндоцитоза в клетках мерцательного эпителия. Белки BBS1 и BBS3 связываются с мембранной GTPазой Arl6, а субъединицы BBS4, BBS5, BBS8 и BBS9 – с фосфоинозидами. Аффинность регулируется стерическими эффектами BBS2 и BBS7. Механизм действия более поздних по времени открытия белков BBS изучен недостаточно подробно. Мутации в комплексах BBS влияют на регулируемый эндоцитоз и являются причиной полиорганных дефектов вследствие нарушения структуры эпителиальных тканей [18, 19]. Белок LZTFL1 относится к слабоизученным элементам семейства BBS. Косвенно это отражено в его порядковом номере BBS17 в соответствующей номенклатуре белков.

Функционально активные комплексы BBS представляют собой гетерооктамеры с центральной коровой частью, собранной из консервативных субъединиц BBS1, BBS4, BBS5, BBS8 и BBS9. Наиболее важную роль в распознавании котранспортируемых белков играет BBS1, содержащий на *N*-конце домен специфического взаимодействия с GTPазой Arl6 [20]. *N*-Концевой домен белка BBS1 имеет структуру  $\beta$ -пропеллера. Данный вариант  $\beta$ -укладки характеризуется наличием четного количества высокосимметричных  $\beta$ -листов, расположенных в форме лопастей вокруг центрального канала. Отдельная лопасть состоит из нескольких антипараллельных  $\beta$ -цепей [21].  $\beta$ -Пропеллеры также идентифицированы в белках BBS7 и BBS9. Домены с  $\beta$ -пропеллерной организацией участвуют в фиксации внешних сигнальных и структурных белков, а также способны взаимодействовать между собой при образовании внутри- и межмолекулярных связей. В частности, регуляторная функция BBS7 реализуется через стерические эффекты контактирующих  $\beta$ -пропеллеров BBS1 и BBS7 [17]. Функция LZTFL1 в составе комплексов BBS не исследована. Выполненный нами компьютерно-статистический анализ выявил высокую вероятность наличия множественных  $\beta$ -складчатых элементов в структуре белка LZTFL1. Потенциально это может быть связано с участием белка в регуляции свойств BBS-комплексов и иметь отношение к трафику с поверхности внутрь клетки.

Полиморфный вариант rs17713054G>A, детектированный в локусе 3p21.31, усиливает в 2 раза тяжесть проявления COVID-19. Однонуклеотидная замена G>A, зафиксированная в промоторной области структурного гена *LZTFL1*,

расположенного в данном локусе, формирует дополнительный энхансерный мотив ССААТ для связывания транскрипционного фактора СЕВРВ и увеличивает скорость транскрипции гена *LZTFL1* [5]. Можно предположить, что повышение уровня белка LZTFL1 модулирует регуляцию адаптерных свойств BBS-комплексов и стимулирует интенсивность рециркуляции ACE2, в конечном итоге ускоряя доставку адгезированных вирионов SARS-CoV-2 внутрь клетки.

Таким образом, прогнозируемая  $\beta$ -складчатая структура белка LZTFL1 может стать ключом для понимания механизма эндоцитоза, опосредованного рецептором ACE2, в патогенезе COVID-19.

### ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Аминокислотная последовательность белка (299 а.о.), кодируемого геном *LZTFL1*, в формате FASTA была взята из GenBank NCBI (CAB95836.1). Аминокислотную последовательность белка LZTFL1 исследовали компьютерно-статистическим методом главных компонент (PCA-Seq) [22]. На основании теоремы Такенса [23] исходную последовательность длины *N* нарезали на фрагменты скользящим окном ширины *L*. Номер фрагмента – позиция последней аминокислоты фрагмента в исходной последовательности (первый фрагмент получает номер *L*, последний – *N*, всего  $N - L + 1$  фрагментов). Каждый фрагмент преобразовывали в 20-мерный вектор частот аминокислот. Между всеми фрагментами вычисляли матрицу квадратов евклидовых расстояний, которые однозначно определяют взаимное расположение фрагментов как точек в евклидовом пространстве. Методом главных координат (РСо) Гауэра [24] вычисляли координатные оси этого пространства. Проекции всех точек на эти оси образуют новые признаки, не коррелирующие между собой. Признаки упорядочиваются по убыванию дисперсии и являются главными компонентами.

Оптимальная ширина скользящего окна для анализируемой последовательности LZTFL1 составила 64 а.о., что в итоге сгенерировало 236 аминокислотных фрагментов. В качестве расстояния между фрагментами было выбрано хордовое расстояние Кавалли–Сфорца [25], по определению являющееся евклидовым. При  $n = 236$  и числе коэффициентов корреляции 553 все коэффициенты при  $r > 0.406$  достоверны с  $pB\text{-value} < 10^{-6}$ . Одиннадцать коэффициентов корреляции аминокислотных индексов с первой главной компонентой LZTFL1 превысили уровень



$r > 0.900$ , что соответствует  $pB$ -value  $< 10^{-15}$  (минимальное ненулевое  $p$ -value, допускаемое пакетом Statistica).

Физико-химические и биохимические свойства аминокислот учитывали после преобразования в числовые ряды в форме аминокислотных AA-индексов. Из актуальной базы данных AAindex были взяты 553 индекса 20 протеиногенных аминокислот [26]. Все AA-индексы предварительно стандартизировали (центрировали и нормировали). Для каждого фрагмента каждая аминокислота была заменена на соответствующее значение AA-индекса и было вычислено их среднее по фрагменту. Таким способом для множества фрагментов получили 553 физико-химических признака. Между ними и всеми 19 главными компонентами вычислили коэффициенты корреляции Пирсона с оценкой достоверности  $pB$ -value, т.е.  $p$ -value с поправкой на множественные сравнения по критерию Бонферрони [27].

Статистическую обработку полученных данных выполняли с использованием пакетов программ PAST 4 [28], Statistica 12 (StatSoft®), Jacobi 4 [29] и MS Excel 2016 (Microsoft®).

### ЗАКЛЮЧЕНИЕ

Компьютерно-статистический анализ с использованием метода главных компонент (PCA-seq) выявил в аминокислотной последовательности белка LZTFL1 главные компоненты, коррелирующие с физико-химическими параметрами структурной организации белков. Из рассмотренных 553 физико-химических признаков установлена группа из одиннадцати аминокислотных индексов с очень высокой корреляцией, семь из которых указывают на особенности  $\beta$ -цепей. В частности, максимальный коэффициент корреляции установлен для индекса BEGF750102, учитываемого при статистическом расчете конформации  $\beta$ -цепи, а следующий за ним по модулю индекс CHOP780209 оценивает частоту присутствия антипараллельных  $\beta$ -структур на C-конце белка. Основной вывод из полученных результатов – аминокислотные индексы с наиболее высокими коэффициентами корреляции с первой главной компонентой аминокислотной последовательности LZTFL1 преимущественно связаны с параметрами  $\beta$ -складчатой укладки. Прогнозируемое с высокой вероятностью присутствие  $\beta$ -складчатых элементов в структуре белка LZTFL1 имеет практическое значение для построения модели регуляции эндоцитоза вирионов SARS-CoV-2, опосредованного

рецептором ACE2, в патогенезе COVID-19, что в перспективе может быть использовано в процессе разработки новых средств терапии.

### ФОНДОВАЯ ПОДДЕРЖКА

Работа выполнена при поддержке бюджетного проекта Института цитологии и генетики СО РАН FWNR-2022-0021, FWNR-2022-0019.

### СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Статья не содержит описания исследований, выполненных с участием людей или использованием животных в качестве объектов.

### КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

### ВКЛАД АВТОРОВ

Автор СЮ выполнил числовое преобразование данных и компьютерные вычисления. Авторы ИИХ и ВМЕ в равной степени внесли вклад в статистический анализ и подготовку текста статьи.

### ДОСТУПНОСТЬ ДАННЫХ

Данные, подтверждающие выводы настоящего исследования, можно получить у корреспондирующего автора по обоснованному запросу.

### СПИСОК ЛИТЕРАТУРЫ

1. Seo S., Zhang Q., Bugge K., Breslow D.K., Searby C.C., Nachury M.V., Sheffield V.C. // *PLoS Genet.* 2011. V. 7. P. e1002358.  
<https://doi.org/10.1371/journal.pgen.1002358>
2. Huang Q., Li W., Zhou Q., Awasthi P., Cazin C., Yap Y., Mladenovic-Lucas L., Hu B., Jeyasuria P., Zhang L., Granneman J.G., Hess R.A., Ray P.F., Kherraf Z.-E., Natarajan V., Zhang Z. // *Dev. Biol.* 2021. V. 477. P. 164–176.  
<https://doi.org/10.1016/j.ydbio.2021.05.006>
3. Fliegau M., Benzing T., Omran H. // *Nat. Rev. Mol. Cell Biol.* 2007. V. 8. P. 880–893.  
<https://doi.org/10.1038/nrm2278>
4. Lyu Q., Li Q., Zhou J., Zhao H. // *J. Cell Biol.* 2024. V. 223. P. e202307150.  
<https://doi.org/10.1083/jcb.202307150>
5. Downes D.J., Cross A.R., Hua P., Roberts N., Schwesinger R., Cutler A.J., Munis A.M., Brown J., Mielczarek O., de Andrea C.E., Melero I., COMBAT Consortium, Gill D.R., Hyde S.C., Knight J.C., Todd J.A., Sansom S.N., Issa F., Davies J.O.J., Hughes J.R. // *Nat. Genet.* 2021. V. 53. P. 1606–1615.  
<https://doi.org/10.1038/s41588-021-00955-3>

6. *Anderson R.M., Heesterbeek H., Klinkenberg D., Déirdre Hollingsworth T.D.* // *Lancet*. 2020. V. 395. P. 931–934.  
[https://doi.org/10.1016/S0140-6736\(20\)30567-5](https://doi.org/10.1016/S0140-6736(20)30567-5)
7. *Tang X., Wu C., Li X., Song Y., Yao X., Wu X., Duan Y., Zhang H., Wang Y., Qian Z., Cui J., Lu J.* // *Natl. Sci. Rev.* 2020. V. 7. P. 1012–1023.  
<https://doi.org/10.1093/nsr/nwaa036>
8. *Hu B., Guo H., Zhou P., Shi Z.-L.* // *Nat. Rev. Microbiol.* 2021. V. 19. P. 141–154.  
<https://doi.org/10.1038/s41579-020-00459-7>
9. *Lu J., Sun P.D.* // *J. Biol. Chem.* 2020. V. 295. P. 18579–18588.  
<https://doi.org/10.1074/jbc.RA120.015303>
10. *Lan J., Ge J., Yu J., Shan S., Zhou H., Fan S., Zhang Q., Shi X., Wang Q., Zhang L., Wang X.* // *Nature*. 2020. V. 581. P. 215–220.  
<https://doi.org/10.1038/s41586-020-2180-5>
11. *Hajizadeh F., Khanizadeh S., Khodadadi H., Mokha-ryeri Y., Ajorloo M., Malekshahi A., Heydaria E.* // *Microb. Pathog.* 2022. V. 168. P. 105595.  
<https://doi.org/10.1016/j.micpath.2022.105595>
12. *Wysocki J., Schulze A., Batlle D.* // *Biomolecules*. 2019. V. 9. P. 886.  
<https://doi.org/10.3390/biom9120886>
13. *Lu J., Sun P.D.* // *J. Biol. Chem.* 2020. V. 295. P. 18579–18588.  
<https://doi.org/10.1074/jbc.RA120.015303>
14. *Guy J.L., Lambert D.W., Warner F.J., Hooper N.M., Turner A.J.* // *Biochim. Biophys. Acta*. 2005. V. 1751. P. 2–8.  
<https://doi.org/10.1016/j.bbapap.2004.10.010>
15. *Iwasaki M., Saito J., Zhao H., Sakamoto A., Hirota K., Ma D.* // *Inflammation*. 2021. V. 44. P. 13–34.  
<https://doi.org/10.1007/s10753-020-01337-3>
16. *Ren Y., Lv L., Li P., Zhang L.* // *J. Infect.* 2022. V. 85. P. e21–e23.  
<https://doi.org/10.1016/j.jinf.2022.04.019>
17. *Klink B.U., Gatsogiannis C., Hofnagel O., Wittinghofer A., Raunser S.* // *eLife*. 2020. V. 9. P. e53910.  
<https://doi.org/10.7554/eLife.53910>
18. *Muller J., Stoetzel C., Vincent M.C., Leitch C.C., Laurier V., Danse J.M., Hellé S., Marion V., Bennouna-Greene V., Vicaire S., Megarbane A., Kaplan J., Drouin-Garraud V., Hamdani M., Sigaudy S., Francannet C., Roume J., Bitoun P., Goldenberg A., Philip N., Odent S., Green J., Cossée M., Davis E.E., Katsanis N., Bonneau D., Verloes A., Poch O., Mandel J.L., Dollfus H.* // *Hum. Genet.* 2010. V. 127. P. 583–593.  
<https://doi.org/10.1007/s00439-010-0804-9>
19. *Liu P., Lehtreck K.F.* // *Proc. Natl. Acad. Sci. USA*. 2018. V. 115. P. E934–E943.  
<https://doi.org/10.1073/pnas.1713226115>
20. *Jin H., White S.R., Shida T., Schulz S., Aguiar M., Gygi S.P., Bazan J.F., Nachury M.V.* // *Cell*. 2010. V. 141. P. 1208–1219.  
<https://doi.org/10.1016/j.cell.2010.05.015>
21. *Pereira J., Lupas A.N.* // *Front. Mol. Biosci.* 2022. V. 9. P. 895496.  
<https://doi.org/10.3389/fmolb.2022.895496>
22. *Ефимов В.М., Ефимов К.В., Ковалева В.Ю.* // *Вавиловский журнал генетики и селекции*. 2019. Т. 23. С. 1032–1036.  
<https://doi.org/10.18699/VJ19.584>
23. *Takens F.* // *Dynamical Systems and Turbulence, Lecture Notes in Mathematics*. 1981. V. 898. P. 366–381.  
<https://doi.org/10.1007/BFb0091924>
24. *Gower J.C.* // *Biometrika*. 1966. V. 53. P. 325–338.  
<https://doi.org/10.1093/biomet/53.3-4.325>
25. *Cavalli-Sforza L.L., Menozzi P., Piazza A.* // *J. Asian Studies*. 1995. V. 54. P. 2173–2219.  
<https://doi.org/10.2307/2058750>
26. *Kawashima S., Pokarowski P., Pokarowska M., Kolinski A., Katayama T., Kanehisa M.* // *Nucleic Acids Res.* 2008. V. 36. P. D202–D205.  
<https://doi.org/10.1093/nar/gkm998>
27. *Benjamini Y., Hochberg Y.* // *J. R. Statist. Soc. B*. 1995. V. 57. P. 289–300.  
<https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
28. *Hammer Ø., Harper D.A., Ryan P.D.* // *Palaeontologia Electronica*. 2001. V. 4. P. 1–9.  
[https://palaeo-electronica.org/2001\\_1/past/issue1\\_01.htm](https://palaeo-electronica.org/2001_1/past/issue1_01.htm)
29. *Polunin D., Shtaliger I., Efimov V.* // *bioRxiv*. 2019. P. 803684.  
<https://doi.org/10.1101/803684>

# Structural Analysis of LZTFL1 Protein by the Principal Component Analysis Method (PCA-seq)

I. I. Khegay<sup>\*,#</sup>, X. Yu<sup>\*\*</sup>, and V. M. Efremov<sup>\*,\*\*</sup>

<sup>#</sup> Phone: +7 (383) 363-49-63; e-mail: khegay@bionet.nsc.ru

<sup>\*</sup> Federal Research Center Institute of Cytology and Genetics, Siberian Branch, RAS,  
prosp. Akad. Lavrentieva 10, Novosibirsk, 630090 Russia

<sup>\*\*</sup> Novosibirsk State University, ul. Pirogova 2, Novosibirsk, 630090 Russia

The single-nucleotide mutation rs17713054G>A in the promoter region of LZTFL1 (leucine zipper transcription factor like 1) gene is a factor in the severe course of coronavirus infection COVID-19. Computer statistical analysis of the gene by principal component analysis (PCA-seq) revealed the presence of a high correlation between the first principal component of the translated amino acid sequence and eleven amino acid indices of the AAindex database, characterizing the physicochemical and biochemical properties of the protein. The indices BEGF750102, CHOP780209, PALJ810110, GEIM800107, QIAN880121, LEVM780102, PRAM900103 are associated with  $\beta$ -folding parameters. The LZTFL1 protein is part of the Bardet-Biedl Syndrome (BBS) protein complexes that regulate intracellular transport in the ciliated epithelium of the lungs. It is assumed that the presence of  $\beta$ -sheet elements in the structure of the LZTFL1 protein plays an important role in ACE2 receptor-mediated endocytosis, stimulating the rate of angiotensin-converting enzyme 2 recycling and accelerating the delivery of adherent coronavirus SARS-CoV-2 virions into the cell during the initiation of severe acute respiratory syndrome COVID-19.

**Keywords:** LZTFL1 protein, ciliary epithelium, ACE2 receptor, amino acid indices, BBS proteins, principal component analysis