XXVII Симпозиум «Биоинформатика и компьютерное конструирование лекарств»

Идентификация генов, функционально значимых для прогрессии вич-инфекции, на основе интеллектуального анализа текстов

НАДЕЖДА ЮРЬЕВНА БИЗЮКОВА

Федеральное государственное бюджетное научное учреждение «Научно-исследовательский институт биомедицинской химии имени В. Н. Ореховича»

Год	2017	2018	2019
Кумулятивное число ВИЧ-позитивных лиц на территории РФ	1 257 886	1 364 571	1 458 364
Кумулятивное число умерших ВИЧ-позитивных лиц на территории РФ	248 438	280 833	314 462
Кумулятивное число умерших с диагнозом СПИД среди граждан РФ	67 182	79 102	90 210
Количество новых зарегистрированных случаев ВИЧ-инфекции	106 072	103 506	97 176

В мире (2019 год):

Общее число инфицированных: 38 млн

Не знали о наличии ВИЧ: 7,1 млн

Принимали ВААРТ: 25,4 млн

Число смертей за год: более 940 тыс

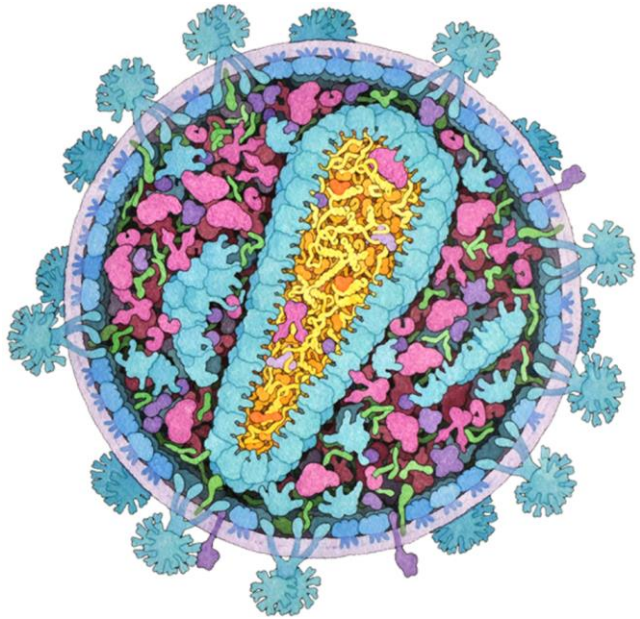
Источник: ВОЗ https://www.who.int

https://spid.center/ru/

https://www.cdc.gov/hiv/basics/whatishiv.html

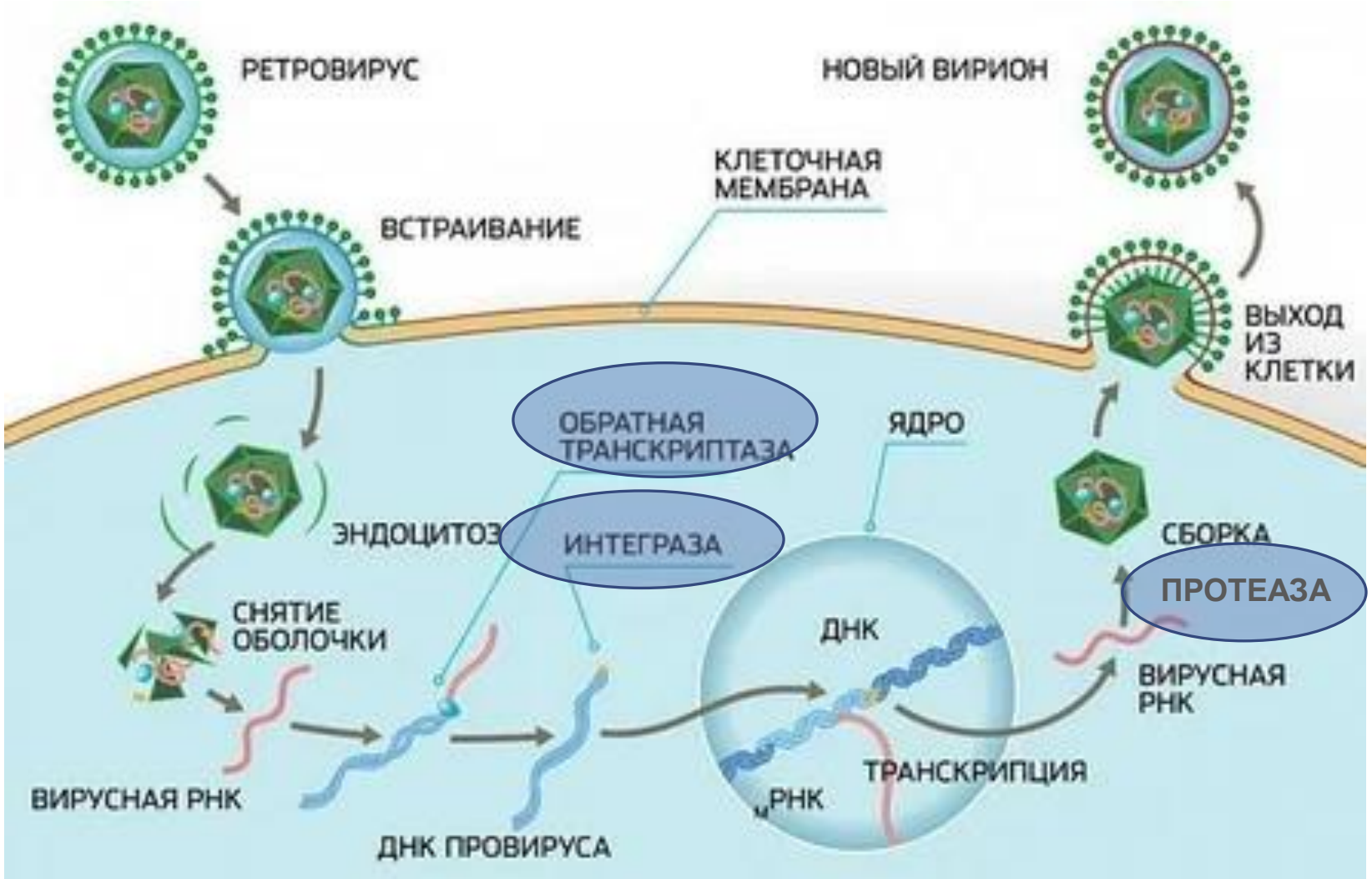
https://www.iasociety.org/

https://hivinfo.nih.gov/



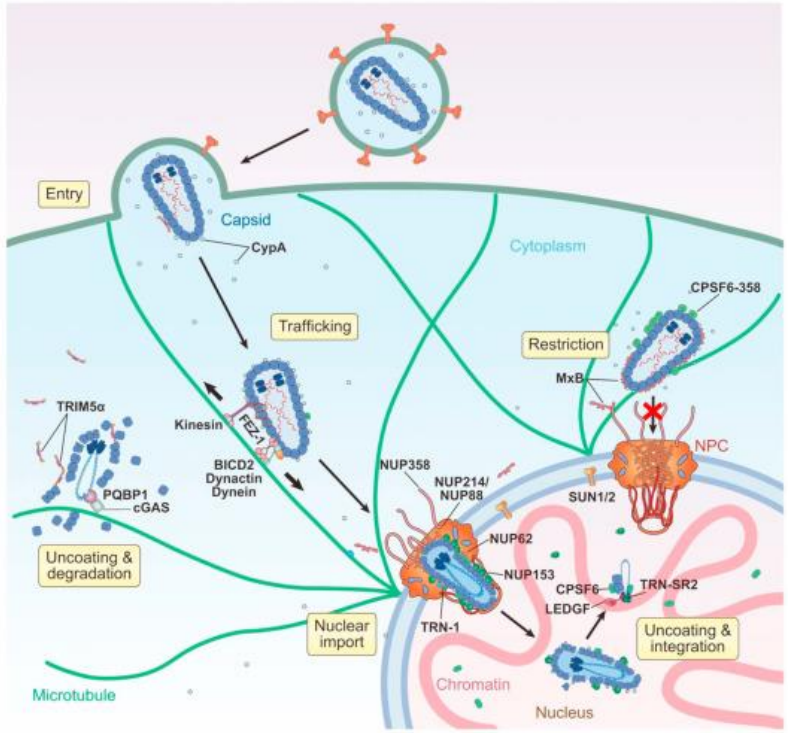
http://pdb101.rcsb.org/

Современные принципы терапии ВИЧ/СПИД – ингибирование основных структурных белков ВИЧ - ВААРТ

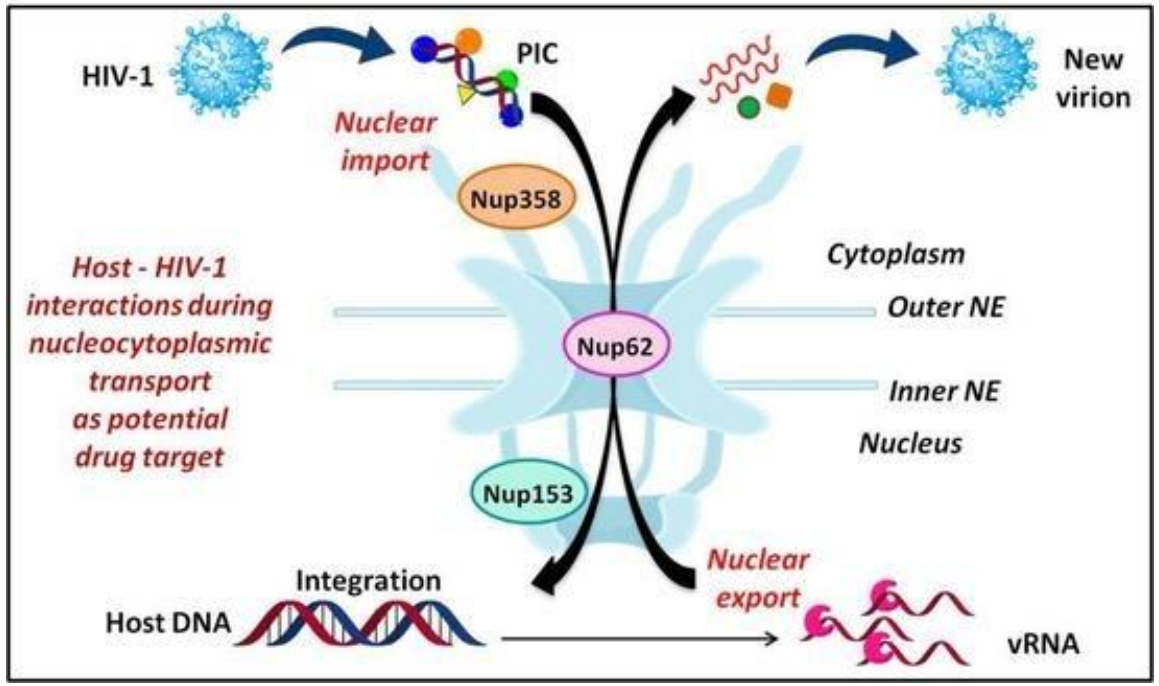


https://spid.center/ru/articles/2414

Поиск перспективных мишеней для терапии ВИЧ/СПИД среди белков человека, взаимодействующих с ВИЧ



Joshua Temple and co-authors, Current Research in Structural Biology, 2020



Joshua Temple and co-authors, Cells, 2020

Biziukova, N., Tarasova, O., Ivanov, S., and Poroikov, V. (2020). Automated Extraction of Information From Texts of Scientific Publications: Insights Into HIV Treatment Strategies. Front Genet 11, 618862. doi:10.3389/fgene.2020.618862.

Krallinger, M., and Valencia, A. (2005). Text-mining and information-retrieval services for molecular biology. Genome Biol 6, 224. doi:10.1186/gb-2005-6-7-224.

Цель

Извлечение наименований белков и генов, потенциально ассоциированных с молекулярными механизмами патогенеза ВИЧ-инфекции, из текстов научных публикаций

Алгоритм распознавания наименований белков и генов в текстах

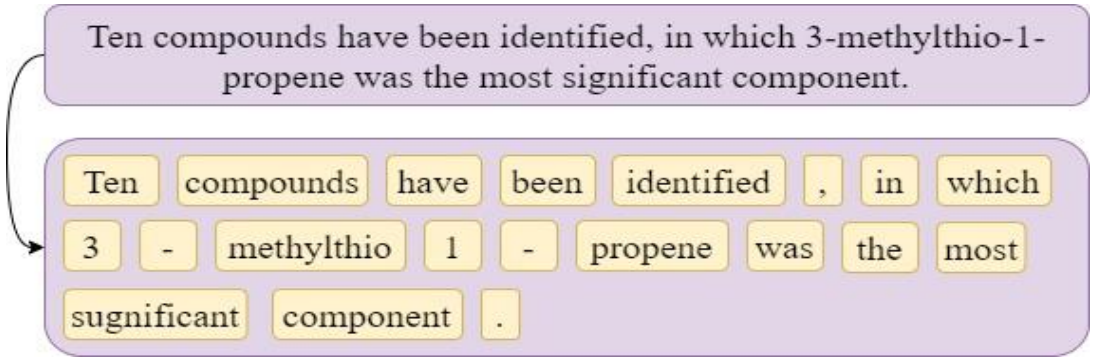
Аннотированный корпус ChemProt

```
23538162  T10  CHEMICAL  947  957  ICI 82,780
23538162  T11  CHEMICAL  1002 1005 Rg1
23538162  T12  CHEMICAL  72   87   ginsenoside Rg1
23538162  T13  GENE-Y    1330 1337 Aβ25-35
23538162  T14  GENE-Y    1391 1393 GR
```

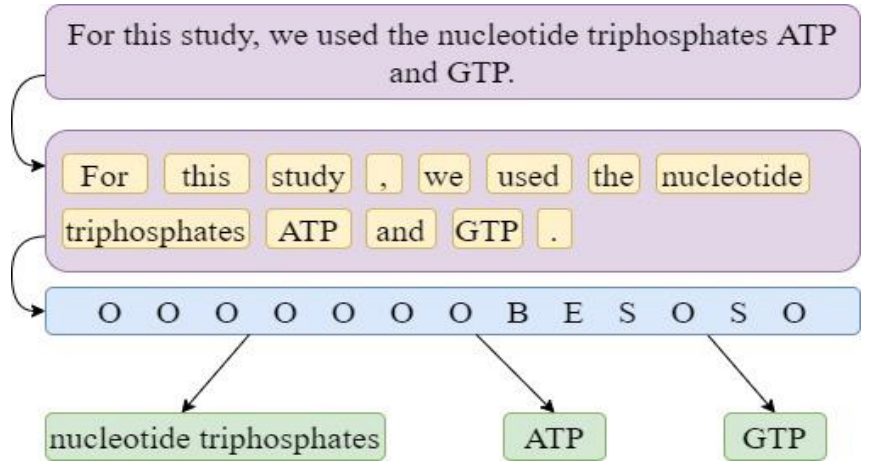
Тип	Число наименований, принадлежащих классам	Доля наименований, принадлежащих классам, %
CHEMICAL	32514	51%
GENE-Y	20544	32%
GENE-N	10378	16%
ВСЕГО	**63436**	**100%**

Препроцессинг

- Токенизация



- Соотнесение с метками принадлежности к наименованию



https://biocreative.bioinformatics.udel.edu/news/corpora/chemprot-corpus-biocreative-vi/

Признак	Тип	Значение
word	string	Токен
lower	string	Токен в нижнем регистре
isUpper	Boolean	Записан ли токен в верхнем регистре
isTitle	Boolean	Является ли первый символ токена заглавным
isDigit	Boolean	Является ли токен числом
hasDigits	Boolean	Содержит ли токен цифры
isNonSpecific	Boolean	Является ли токен неспецифическим термином
isStopWord	Boolean	Является ли токен стоп-словом
hasSymbols	Boolean	Содержит ли токен знаки
word[n-3:n]	string	Последние три символа токена
word[n-2:n]	string	Последние два символа токена
firstChar	string	Первый символ токена
length	integer	Число символов в токене
posTag	string	Часть речи

Неспецифические термины

compound
Inhibitor
activator
drug
chemical
molecule
molecules
derivative
…

Стоп-слова

all
also
by
down
for
get
in
many
the
…

{'word': 'pyrene', 'word.lower()': 'pyrene', 'word.isupper()': 0, 'word.istitle()': 0,
'word.isdigit()': 0, 'word.havedigits()': 0, 'word.isNonSpecific': 0,
'word.isStopWord': 0, 'word.isSymbol': 0, 'word[-3:]': 'ene', 'word[-2:]': 'ne',
'word.FirstSymbol': 'p', 'word.CharNumber': 6, 'postag': 'NN',
'-1:word': ')', '-1:word.lower()': ')', '-1:word.isupper()': 0, '-1:word.istitle()': 0,
'-1:word.isdigit()': 0, '-1:word.havedigits()': 0, '-1:word.isNonSpecific': 0,
'-1:word.isStopWord': 0, '-1:word.isSymbol': 1, '-1:word[-3:]': ')',
'-1:word[-2:]': ')', '-1:word.FirstSymbol': ')', '-1:word.CharNumber': 1,
'-1:postag': ')',
'+1:word': 'but', '+1:word.lower()': 'but', '+1:word.isupper()': 0, '+1:word.istitle()': 0,
'+1:word.isdigit()': 0, '+1:word.havedigits()': 0, '+1:word.isNonSpecific': 0,
'+1:word.isStopWord': 1, '+1:word.isSymbol': 0, '+1:word[-3:]': 'but',
'+1:word[-2:]': 'ut', '+1:word.FirstSymbol': 'b', '+1:word.CharNumber': 3,
'+1:postag': 'CC'}

$$[..., S, O, O, O, O, B, I, I, E, O, O, O, O, ...]$$

Алгоритм: условные случайные поля (Conditional Random Fields, CRF)

Реализация: Python 3.7

Оценка точности распознавания

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 * precision * recall}{precision + recall}$$

Качество распознавание наименований белков и генов (пятикратная кросс-валидация)

	Precision	Recall	F1-score	N меток в ChemProt	Доля меток в ChemProt, %
S	0,86	0,83	0,84	35793	5%
O	0,97	0,98	0,98	581830	84%
B	0,83	0,78	0,80	20396	3%
I	0,84	0,81	0,83	37577	5%
E	0,83	0,78	0,81	20396	3%
Avg	**0,87**	**0,84**	**0,85**	-	-

S – наименование, состоящее из одного токена

B – начальный токен составного наименования

I – промежуточные токены составного наименования

E – конечный токен составного наименования

O – токен, не относящийся к наименованию

Точность извлечения белков на тестовой выборке (100 публикаций):

- Precision: 0,84
- Recall: 0,79
- F1-score: 0,81

Сравнение с ранее реализованными алгоритмами

PMID	Метод	F1-score
32046638	BiLSTM-CRF	0,84
30564940	BiLSTM-CRF	0,79
32000677	BiLSTM-CRF	0,81
30239666	CNN-BiLSTM-CRF	0,8

Элитные контроллеры

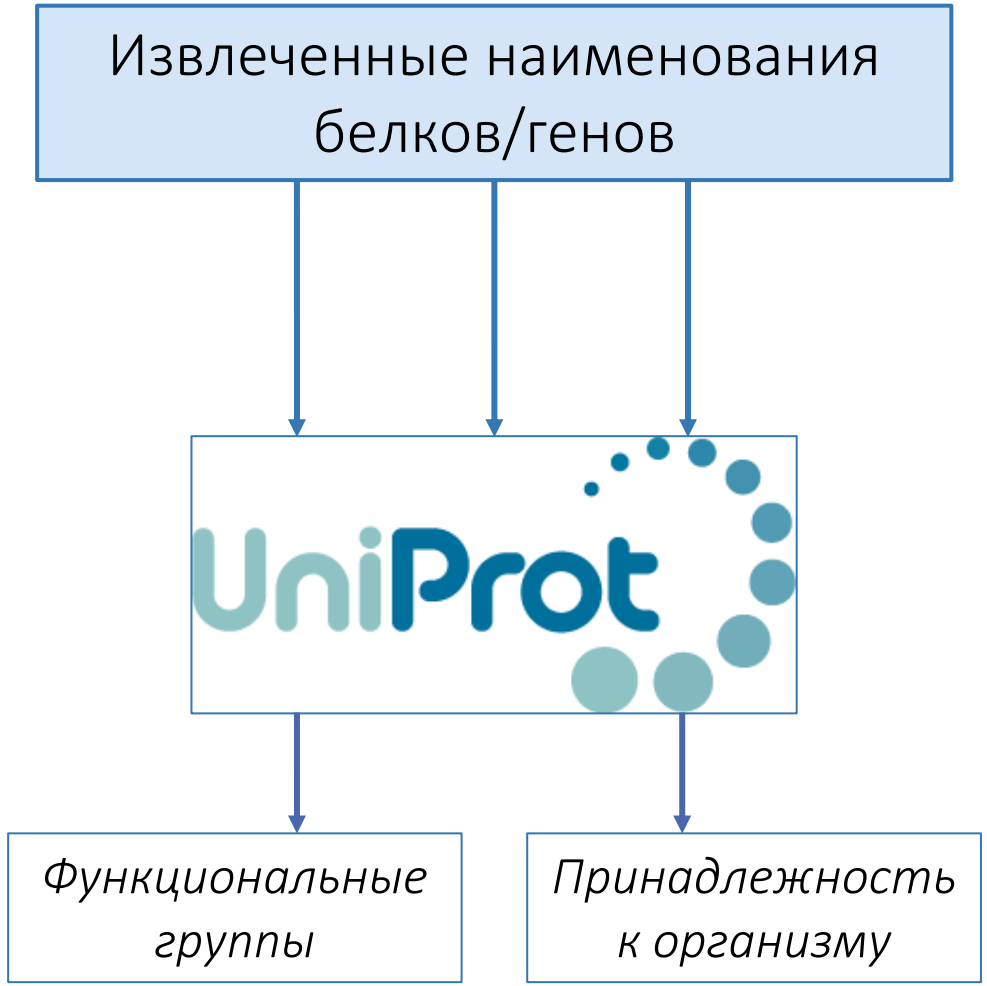
ВИЧ-положительные пациенты

“(HIV[Title/Abstract] OR "human immunodeficiency virus"[Title/Abstract] OR "HIV"[Mesh] OR AIDS[Title/Abstract] OR "acquired immunodeficiency syndrome"[Title/Abstract] OR "Acquired Immunodeficiency Syndrome"[Mesh]) AND ("elite control"[Title/Abstract] OR "Elite suppress*"[Title/Abstract])*

“HIV positive AND HIV/AIDS”

Алгоритм распознавания наименований белков и генов в текстах

Элитные контроллеры – ВИЧ-положительные пациенты, не принимающие компоненты ВААРТ, у которых в течение года сохранялась копийность РНК не более, чем 50 копий/мл



	Число уникальных наименований белков/генов, извлеченных из текстов
Элитные контроллеры (Группа 1)	478
ВИЧ-позитивные пациенты (Группа 2)	1443
Пересечение между Группами 1 и 2	75
Белки, специфичные только для Группы 1	403
Белки, специфичные только для Группы 2	1368

Biziukova, N., and co-authors. Frontiers in Genetics (2020) 11, 618862.

	Список белков
Иммунный ответ	Human leukocyte elastase; nkg2a; nkg2d receptor; nkp30; nkp44; antileukoproteinase (ALP), apolipoprotein A-I, c4b-binding protein alpha chain (C4bp), complement C4-B (Basic complement C4), interferon-gamma; Calcitonin gene-related peptide 1 (Alpha-type CGRP); Antileukoproteinase (ALP); C4b-binding protein alpha chain (C4bp); Complement C4-B (Basic complement C4); HLA class I histocompatibility antigen (HLA-B27K protein) (MHC class I antigen) (Major histocompatibility complex); HLA-B alpha chain (B*5703GB) (MHC class I antigen); Interferon beta (IFN-beta) (Fibroblast interferon), Apolipoprotein A-I (Apo-AI); Interferon beta (IFN-beta) (Fibroblast interferon)
Аутофагия	alpha-1A adrenergic receptor; putative peripheral benzodiazepine receptor-related protein; forkhead box protein O3; interferon gamma; microtubule-associated proteins 1A/1B; microtubule-associated protein 1S (MAP-1S); microtubule-associated protein tau (Neurofibrillary tangle protein); Platelet-activating factor acetylhydrolase IB subunit beta; Microtubule-associated proteins 1A; Nicotinamide phosphoribosyltransferase (NAmPRTase);
Воспалительный ответ	adenosine deaminase (EC 3.5.4.4) (Adenosine aminohydrolase); angiotensin-converting enzyme 2 ; antithrombin-III (ATIII) (Serpin C1); calcitonin; 5'-nucleotidase (5'-NT); Integrin beta-2; human leukocyte antigen b57 interferon-gamma; tumor necrosis factor(nf)- alpha, interleukin (il)-2, C-reactive protein; Interferon gamma (IFN-gamma) (Immune interferon); Platelet factor 4 (PF-4); prostaglandin F2-alpha receptor (PGF receptor) (PGF2-alpha receptor) (Prostanoid FP receptor); prothrombin (EC 3.4.21.5) (Coagulation factor II)
Негативная регуляция Т-клеточно опосредованной цитотоксичности	carcinoembryonic antigen-related cell adhesion molecule 1; leukocyte immunoglobulin-like receptor subfamily B member 1 (CD85 antigen-like family member J); leukocyte immunoglobulin-like receptor subfamily B member 1 (LIR-1); HLA class I histocompatibility antigen, alpha chain G (HLA G antigen)
Негативная регуляция активации CD8+ Т-клеток	leukocyte immunoglobulin-like receptor subfamily B member 1 (LIR-1) (cd85j receptor)
Регуляция экспрессии генов	Progonadoliberin-1 (Progonadoliberin I), lhrh, Prostaglandin F2-alpha receptor (PGF receptor); Angiotensin-converting enzyme (ACE); Myc proto-oncogene protein (Proto-oncogene c-Myc); Calcitonin receptor (CT-R); Leukocyte immunoglobulin-like receptor subfamily B member 1 (LIR-1); Core histone macro-H2A.2, C-reactive protein; Estrogen receptor (ER); Pro-epidermal growth factor (EGF); Fibronectin (FN); Interferon gamma (IFN-gamma)

Для некоторых из найденных белков, по данным литературы, существуют ассоциации с патогенезом ВИЧ-инфекции и скоростью патогенеза ВИЧ

1. **NKp44 (UniProt: O95944)**
   Fausther-Bovendo, H., Sol-Foulon, N., Candotti, D., Agut, H., Schwartz, O., Debré, P., et al. (2009). HIV escape from natural killer cytotoxicity: nef inhibits NKp44L expression on CD4+ T cells. AIDS 23, 1077–1087. doi:10.1097/QAD.0b013e32832cb26b.
2. **CD85j (UniProt: Q8NHL6), S100-A9 (UniProt: P06702)**
   Arnold, V., Cummings, J.-S., Moreno-Nieves, U. Y., Didier, C., Gilbert, A., Barré-Sinoussi, F., et al. (2013). S100A9 protein is a novel ligand for the CD85j receptor and its interaction is implicated in the control of HIV-1 replication by NK cells. Retrovirology 10, 122. doi:10.1186/1742-4690-10-122.

Biziukova, N., Tarasova, O., Ivanov, S., and Poroikov, V. (2020). Automated Extraction of Information From Texts of Scientific Publications: Insights Into HIV Treatment Strategies. Front Genet 11, 618862. doi:10.3389/fgene.2020.618862.

Спасибо за внимание!

Работа выполнена при поддержке гранта РНФ 19-75-10097 *«Компьютерный анализ взаимодействия с организмом человека вируса иммунодефицита с учетом применяемой терапии ВИЧ/СПИД»*