Государственное учреждение
«Республиканский научно-практический центр детской онкологии, гематологии и иммунологии»
Министерства здравоохранения Республики Беларусь

БИОИНФОРМАТИЧЕСКИЙ АЛГОРИТМ АВТОМАТИЗИРОВАННОГО ВЫЯВЛЕНИЯ И ИНТЕРПРЕТАЦИИ ГЕНЕТИЧЕСКИХ ВАРИАНТОВ, АССОЦИИРОВАННЫХ С ПЕРВИЧНЫМИ ИММУНОДЕФИЦИТАМИ



Докладчик: **Бобрик Павел Юрьевич**,
младший научный сотрудник лаборатории молекулярно-генетических исследований научного отдела РНПЦ ДОГИ.

bobrykpavel2001@gmail.com

Что такое первичный иммунодефицит (ПИД)?

Первичные иммунодефициты – широкая группа заболеваний иммунной системы, в основе которых лежат разнообразные генетические детерминанты. Нарушения в работе различных звеньев иммунитета приводят к ослаблению защитных свойств организма.

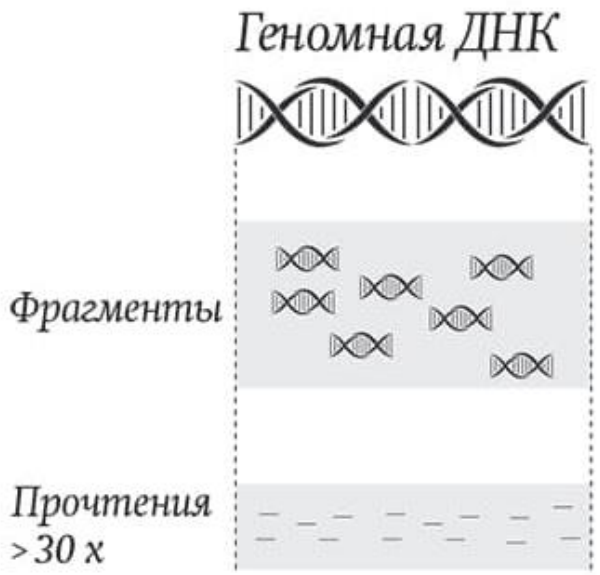
Генетическая основа первичных иммунодефицитов крайне разнообразна: фенотип заболевания зависит не только от нарушения в том или ином гене, но и от типа и локализации мутации в структуре гена, от влияния генов-модификаторов и соматических мутаций, а также от эпигенетических факторов.





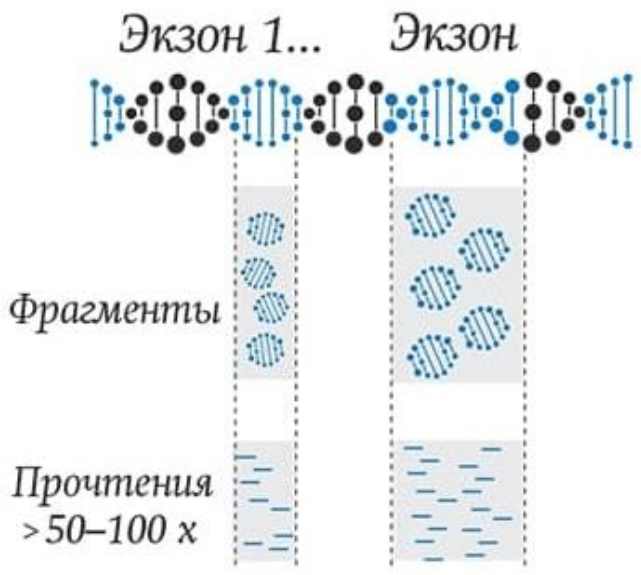
Высокопроизводительное секвенирование (NGS)

Полногеномное секвенирование



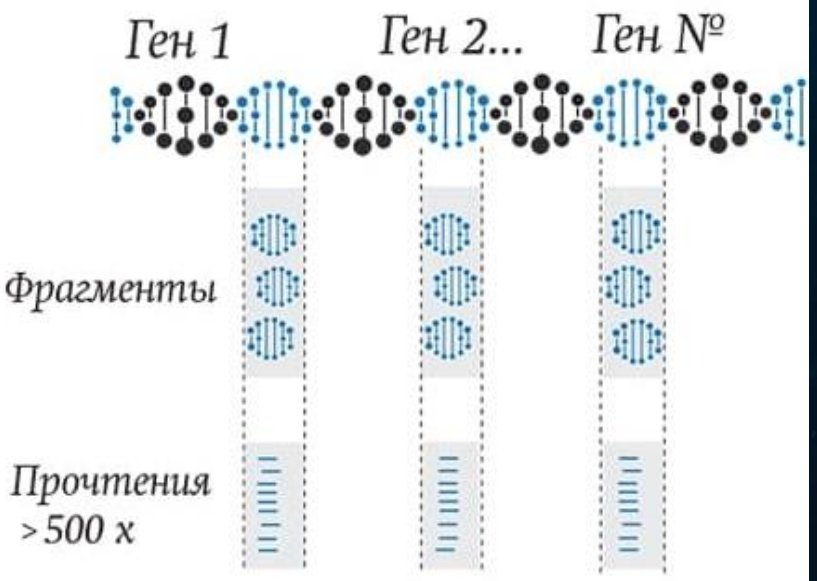
Цель: *все кодирующие и некодирующие последовательности ДНК*

Секвенирование экзома



Цель: *кодирующие последовательности ДНК (20–25 тыс. генов)*

Секвенирование панели генов



Цель: *~10–500 генов*

Этапы биоинформатической обработки данных высокопроизводительного секвенирования

Первый этап

0) .bcl → .fastq (для данных Illumina)
1) Оценка качества «сырых» данных (FastQC / MultiQC)
2) Тримминг прочтений (Trimmomatic)
3) Удаление адаптерных последовательностей (Trimmomatic, Cutadapt)

Второй этап

1) Выравнивание данных на референсный геном (BWA, minimap2)
.fastq → .bam
2) Дедупликация (Picard)
3) Оценка глубины покрытия (samtools)
4) Вызов вариантов (GATK)
.bam → .vcf

Третий этап

1) Аннотация и фильтрация генетических вариантов (Annovar, SnpEff, Ensembl VEP)
2) Приоритизация генетических вариантов с использованием различных баз данных и предикторов патогенности

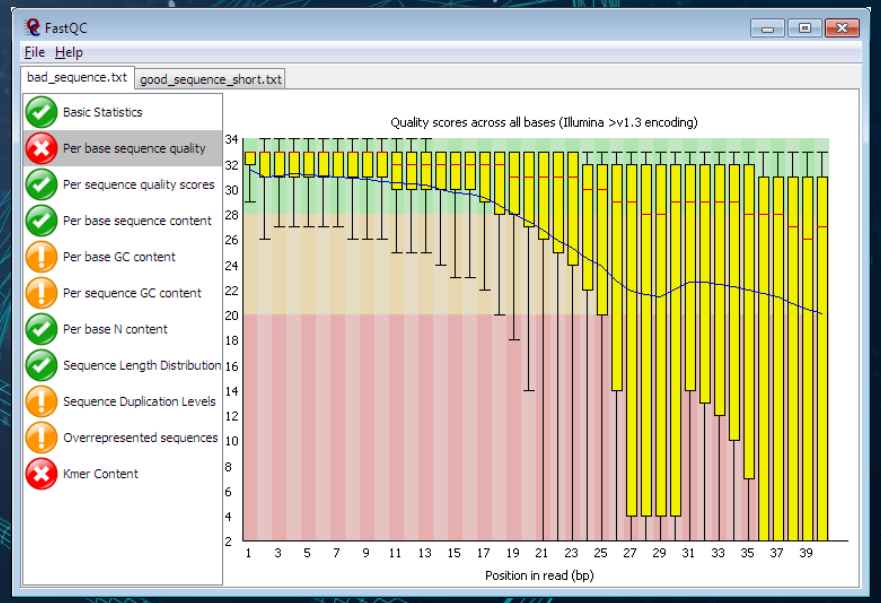
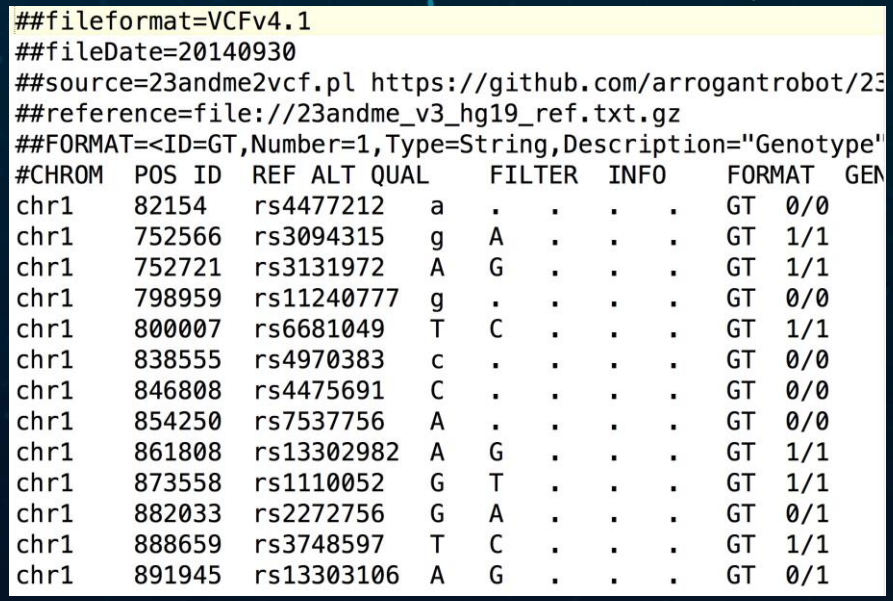


Рис. 1 – Интерфейс программы FastQC



```
##fileformat=VCFv4.1
##fileDate=20140930
##source=23andme2vcf.pl https://github.com/arrogantrobot/23
##reference=file://23andme_v3_hg19_ref.txt.gz
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype"
#CHROM  POS ID  REF ALT QUAL  FILTER  INFO  FORMAT  GEN
chr1  82154  rs4477212  a  .  .  .  .  GT  0/0
chr1  752566  rs3094315  g  A  .  .  .  GT  1/1
chr1  752721  rs3131972  A  G  .  .  .  GT  1/1
chr1  798959  rs11240777  g  .  .  .  .  GT  0/0
chr1  800007  rs6681049  T  C  .  .  .  GT  1/1
chr1  838555  rs4970383  c  .  .  .  .  GT  0/0
chr1  846808  rs4475691  C  .  .  .  .  GT  0/0
chr1  854250  rs7537756  A  .  .  .  .  GT  0/0
chr1  861808  rs13302982  A  G  .  .  .  GT  1/1
chr1  873558  rs1110052  G  T  .  .  .  GT  1/1
chr1  882033  rs2272756  G  A  .  .  .  GT  0/1
chr1  888659  rs3748597  T  C  .  .  .  GT  1/1
chr1  891945  rs13303106  A  G  .  .  .  GT  0/1
```

Рис. 2 – Структура VCF-файла



Рис. 3 – Характеристики замены

Как оценить клиническую значимость варианта?



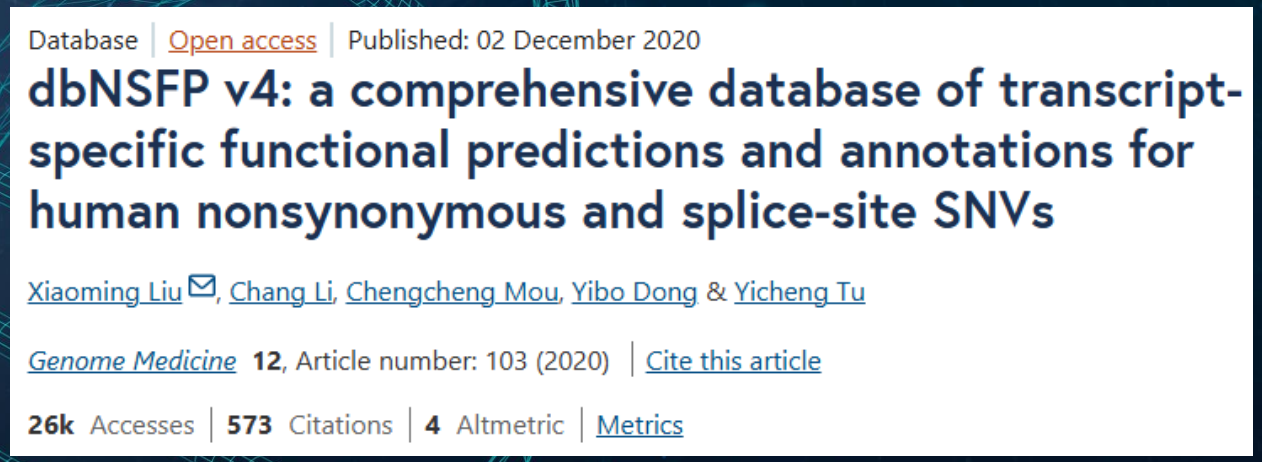
- Для оценки частоты встречаемости в популяциях можно использовать базы данных **gnomAD**, **All of Us**, **RUSeq** и т.п.
- **ClinVar** позволяет классифицировать вариант в одну из пяти категорий (Benign, Likely benign, Uncertain significance, Likely pathogenic, Pathogenic).
- Инструменты для предсказания патогенности *in silico* используют различные подходы, в том числе основанные на алгоритмах машинного обучения.
- Для поиска упоминаний варианта в литературе можно использовать **PubMed** (поиск по rsID), **VarSome/Franklin**, а также ресурс **Varaico** (поиск по координате). В идеальном случае в работе описано более одного пациента, проведён генетический анализ родственников пробанда, с помощью функционального анализа показан механизм развития заболевания, показана генотип-фенотипическая корреляция либо её отсутствие.

Что такое dbNSFP?

dbNSFP – это комплексный, универсальный ресурс для функциональной аннотации и прогнозирования патогенности всех возможных несинонимичных однонуклеотидных вариантов (nsSNVs) в геноме человека.

Текущая версия ресурса (v5.4, 01.08.2026) включает в себя более 86 млн. несинонимичных однонуклеотидных замен и более 2,5 млн. замен в сайтах сплайсинга (ssSNVs), полученных из всех известных белок-кодирующих генов человека.

Ресурс упоминается в различных научных публикациях, в том числе самими авторами dbNSFP при крупных обновлениях (v3.x→v4.x→...).


Database | Open access | Published: 02 December 2020

dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs

Xiaoming Liu, Chang Li, Chengcheng Mou, Yibo Dong & Yicheng Tu

Genome Medicine **12**, Article number: 103 (2020) | Cite this article

26k Accesses | **573** Citations | **4** Altmetric | Metrics

Рис. 4 – Публикация в журнале Genome Medicine, посвящённая выходу dbNSFP 4-ой версии

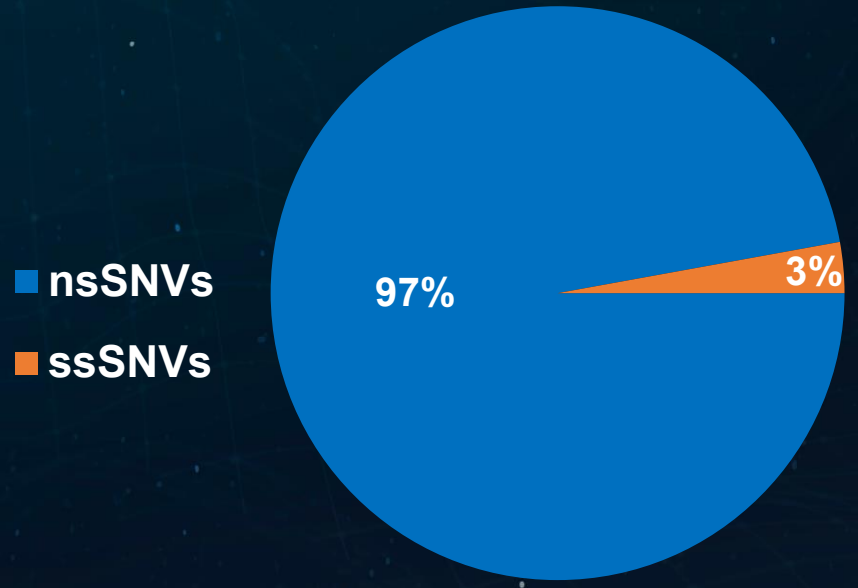


Рис. 5 – Распределение вариантов в базе данных dbNSFP

Почему недостаточно использовать только dbNSFP?

Несмотря на то, что большая часть клинически значимых замен присутствует в dbNSFP, существует ряд ограничений:

- Отсутствие аннотации для делеций / инсерций
- Отсутствие аннотации для интронных замен
- Отсутствие аннотации для синонимичных вариантов

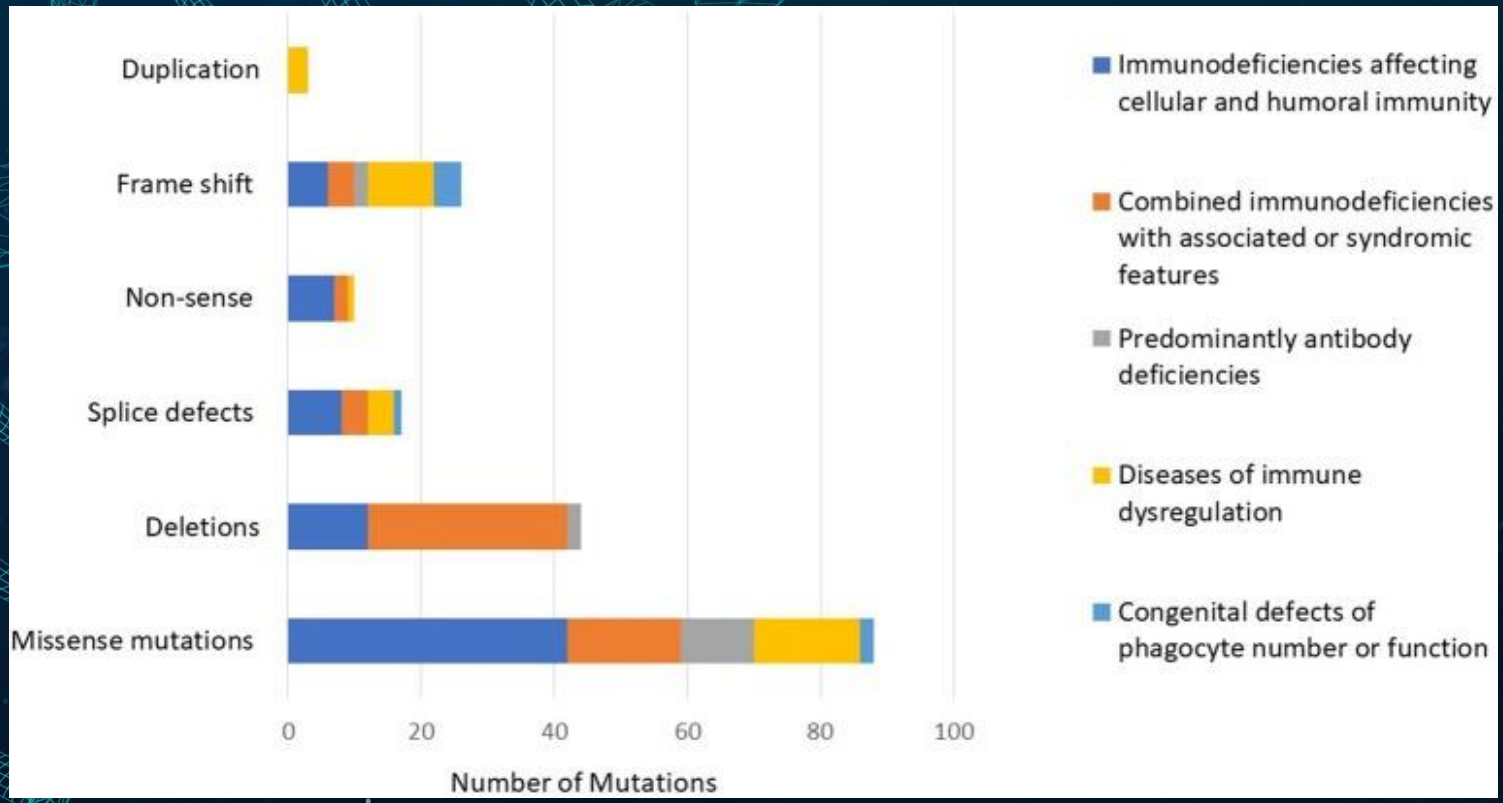
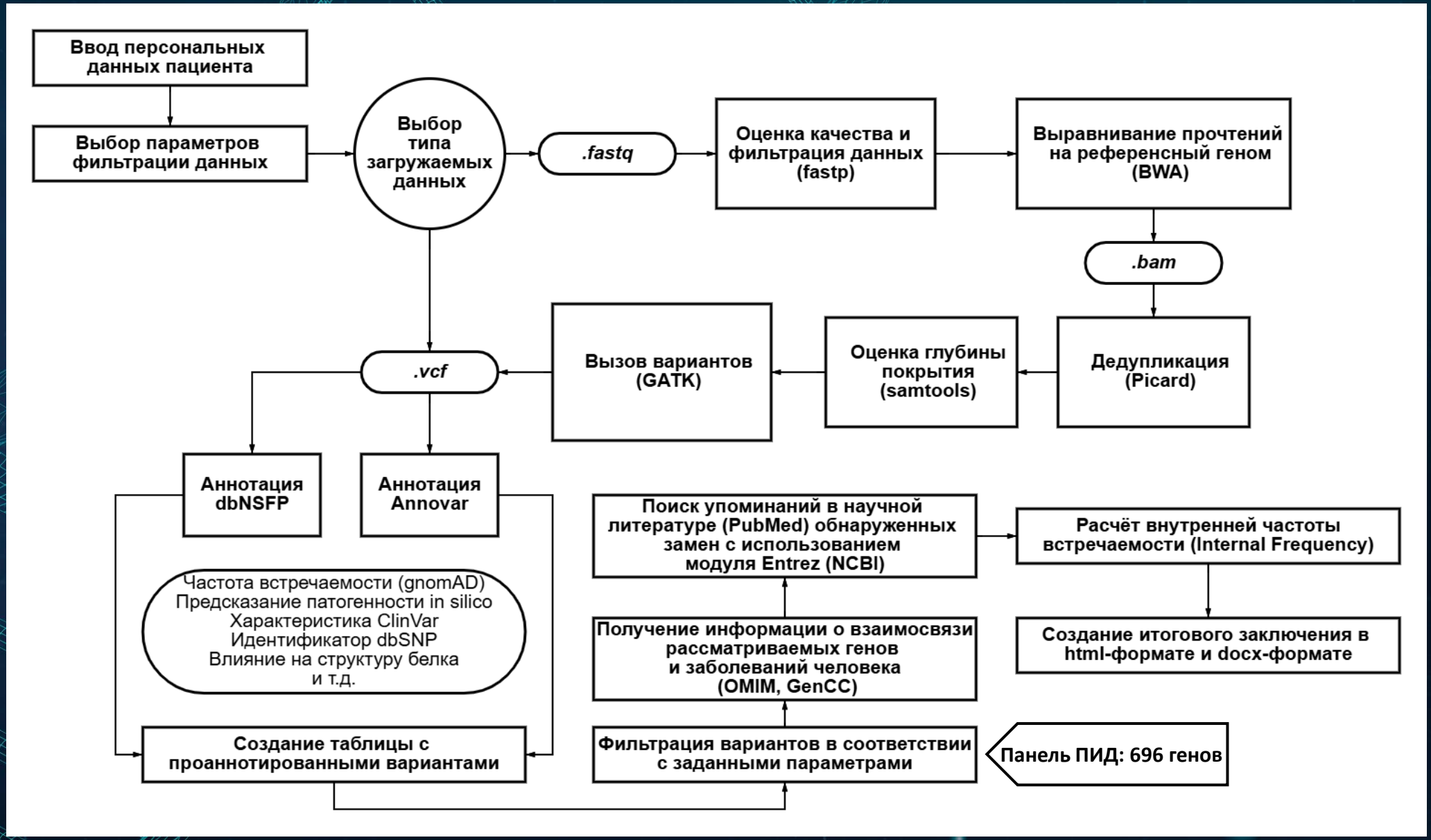
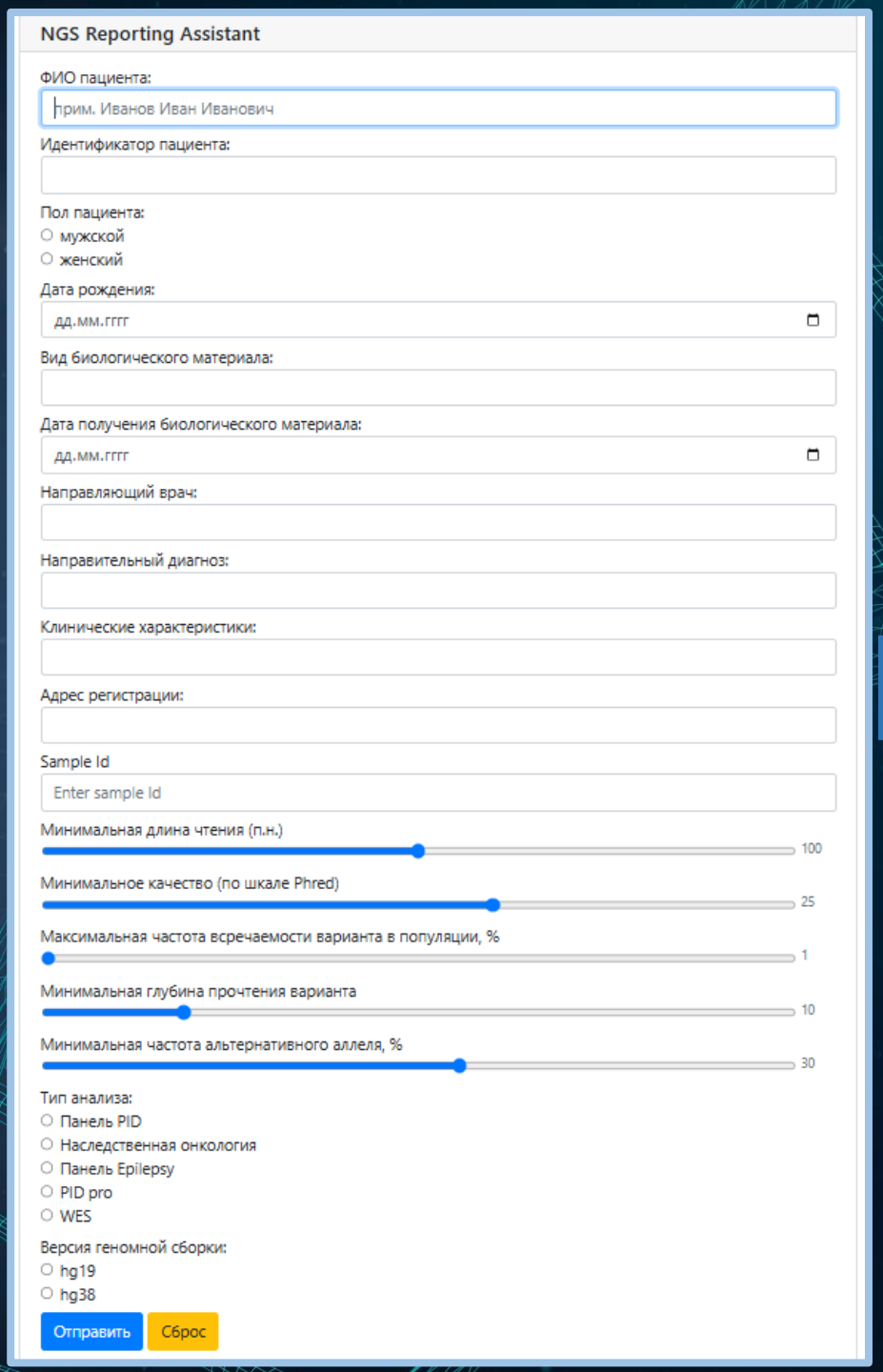


Рис. 6 – Типы генетических нарушений, характерных для пациентов с ПИД (Waleed Al-Herz et al., PMID: 30697212, 2019 г.)

Биоинформатический алгоритм автоматизированного выявления и интерпретации генетических вариантов, ассоциированных с первичными иммунодефицитами





VCF or FASTQ

Приложение 1

Включает в себя выявленные генетические варианты с патогенным или вероятно патогенным эффектом, а также варианты неопределённого клинического значения с частотой встречаемости в популяциях <1% (gnomAD), общей глубиной прочтения (Total Read Depth) >10 и частотой альтернативного аллеля (Variant Allele Frequency) >30%:

Таблица 1. Дополнительные данные:

Ген*	Транскрипт	CDC	Белок	Тип наследования	Зигот-ность	Глубина прочтения	ID dbSNP**	Классификация варианта***
CHEK2	NM_007194.4	c.470T>C	p.Ile157Thr	AD	het	64/161	rs17879961	Likely pathogenic
SLX4	NM_032444.4	c.2125G>A	p.Val709Met	AR	het	177/385	rs151026877	Uncertain significance

* названия генов даны по HUGO Gene Nomenclature Committee;
** при наличии, приведён идентификатор варианта в базе данных dbSNP;
*** классификация вариантов представлена на основании обработки данных ClinVar.

1) Согласно Online Mendelian Inheritance in Man (OMIM), мутации в гене *CHEK2* описаны у пациентов с такими заболеваниями, как Osteosarcoma [OMIM #259500], Prostate cancer [OMIM #176807], а также Tumor predisposition syndrome 4 (breast/prostate/colorectal) [OMIM #609265].

В гене *CHEK2* обнаружена замена нуклеотида в гетерозиготном состоянии c.470T>C, приводящая к замене аминокислоты p.Ile157Thr. Для данного варианта в базах данных аллельных вариантов человека (gnomAD Exomes) частота альтернативного аллеля достигает 0.00412.

Данная замена упоминается в следующих литературных источниках: Seung Hun Han et al., Integrative Analysis of Germline Rare Variants in Clear and Non-clear Cell Renal Cell Carcinoma, PMID: 38496821, 2024 г. || Joanna Huszno et al., Checkpoint Kinase 2 (CHEK2) Mutation in Renal Cell Carcinoma: A Single-Center Experience, PMID: 29682443, 2018 г. || Chuan Liu et al., The CHEK2 I157T variant and colorectal cancer susceptibility: a systematic review and meta-analysis, PMID: 22901170, 2012 г. || Outi Kilpivaara et al., CHEK2 variant I157T may be associated with increased breast cancer risk, PMID: 15239132, 2004 г.

Данные от предикторов патогенности in silico соотносятся следующим образом: количество благоприятных прогнозов – 8 (SIFT, SIFT4G, MutationAssessor, PROVEAN, MetaRNN, PrimateAI, DEOGEN2, ClinPred); количество неблагоприятных прогнозов – 9 (MutationTaster, VEST4, MetaSVM, MetaLR, REVEL, MVP, CADD, FATHMM-XF, PolyPhen-2).

По данным ClinVar вариант классифицирован как Pathogenic (13) | Likely pathogenic (16) | Uncertain significance (9).

Результаты

Для тестирования алгоритма использовалось 50 vcf-файлов (WES) (≈ 85 тыс. замен в файле).

Средняя продолжительность работы алгоритма составляла 17 мин. 43 сек. (±50 сек.), где примерно ¾ времени пришлось на аннотацию в dbNSFP, и ¼ – в ANNOVAR.

Доля вариантов в кодирующих регионах составляла в среднем 29,02% (±0,64%).

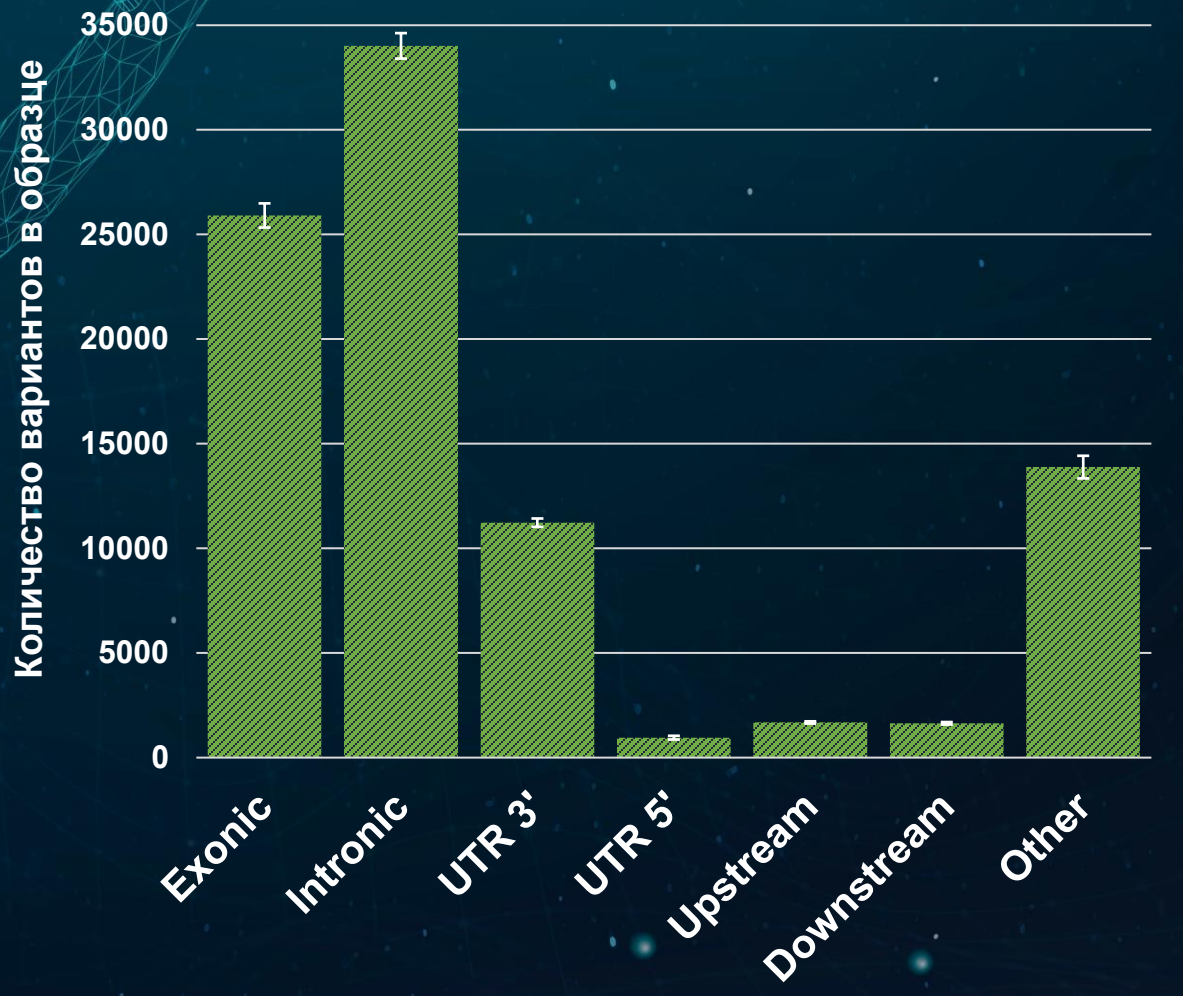


Рис. 7 – Среднее число выявленных вариантов в разных регионах

Результаты

Около 14% замен было аннотировано в dbNSFP, ещё примерно 76% – в Annovar. С помощью алгоритма удалось подробно проаннотировать все миссенс и нонсенс замены, а также часть интронных вариантов.

Преимущества данного решения:

- Подробная аннотация вариантов в кодирующих регионах;
- Высокая скорость обработки данных;
- Экономия ресурсов за счёт агрегации предикторов патогенности в одном месте при использовании dbNSFP;
- Графический интерфейс на основе PyWebIO, позволяющий пользователю, не имеющему опыта взаимодействия с командной строкой, анализировать данные путём загрузки vcf/fastq-файлов на сервер.

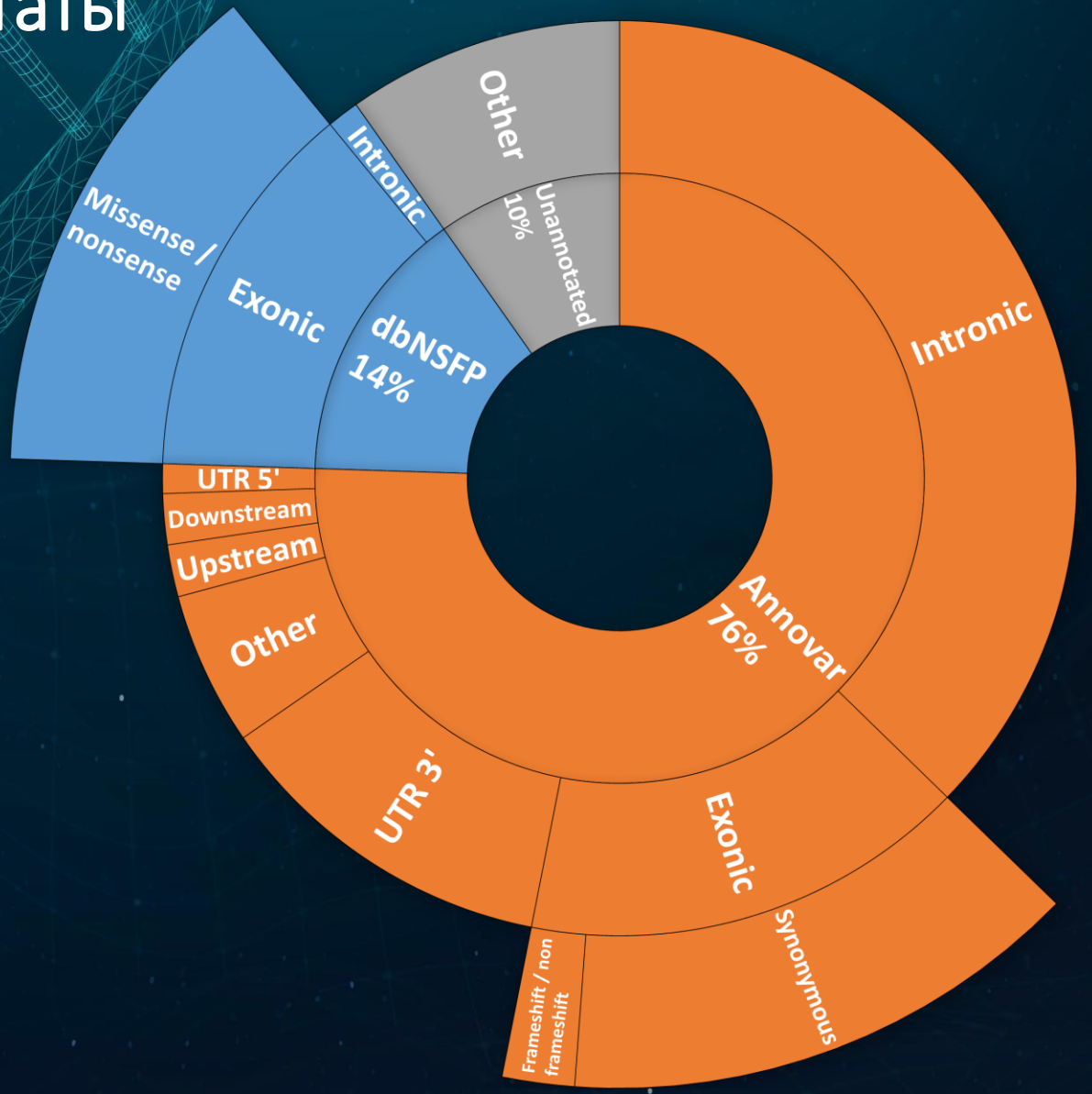


Рис. 8 – Доля проаннотированных вариантов модулям Annovar и dbNSFP

Целевая аудитория

Научно-исследовательские лаборатории медицинского профиля и учреждения здравоохранения.





СПАСИБО ЗА ВНИМАНИЕ

