# Проект Zenome: Белая книга
## *геномная платформа на блокчейне*

Николай Кулемин  Сергей Попов  Алексей Горбачёв

6 октября, 2017

## Аннотация

Автоматизация, масштабируемость и обмен данными в производственных технологиях — вот ключевые особенности протекающей в данный момент четвёртой промышленной революции. Также к ней относят искусственный интеллект, виртуальную реальность, интернет вещей и анализ больших данных. Геномика — ещё один типичный пример из числа таких технологий, для развития которой потребуется решение многих актуальных задач. Среди них проблема хранения и анализа больших данных — с возможностью свободного доступа к ним со стороны исследователей и сохранением конфиденциальности для носителей генетических данных.

На данный момент в геномной индустрии прослеживается ситуация неравноправности. Заключается она в том, что подавляющая часть персональных геномных данных расположена в дата-центрах геномных корпораций, государств, научных и медицинских учреждений, а также фармацевтических компаний. Существует также проблема законных ограничений, накладываемых на доступ к персональным геномным данным, при этом доноры данных не контролируют доступ к ним, а возможность свободного обмена данными практически отсутствует. Такая монополизация существенно сдерживает развитие геномики и, в частности, персонализированной медицины.

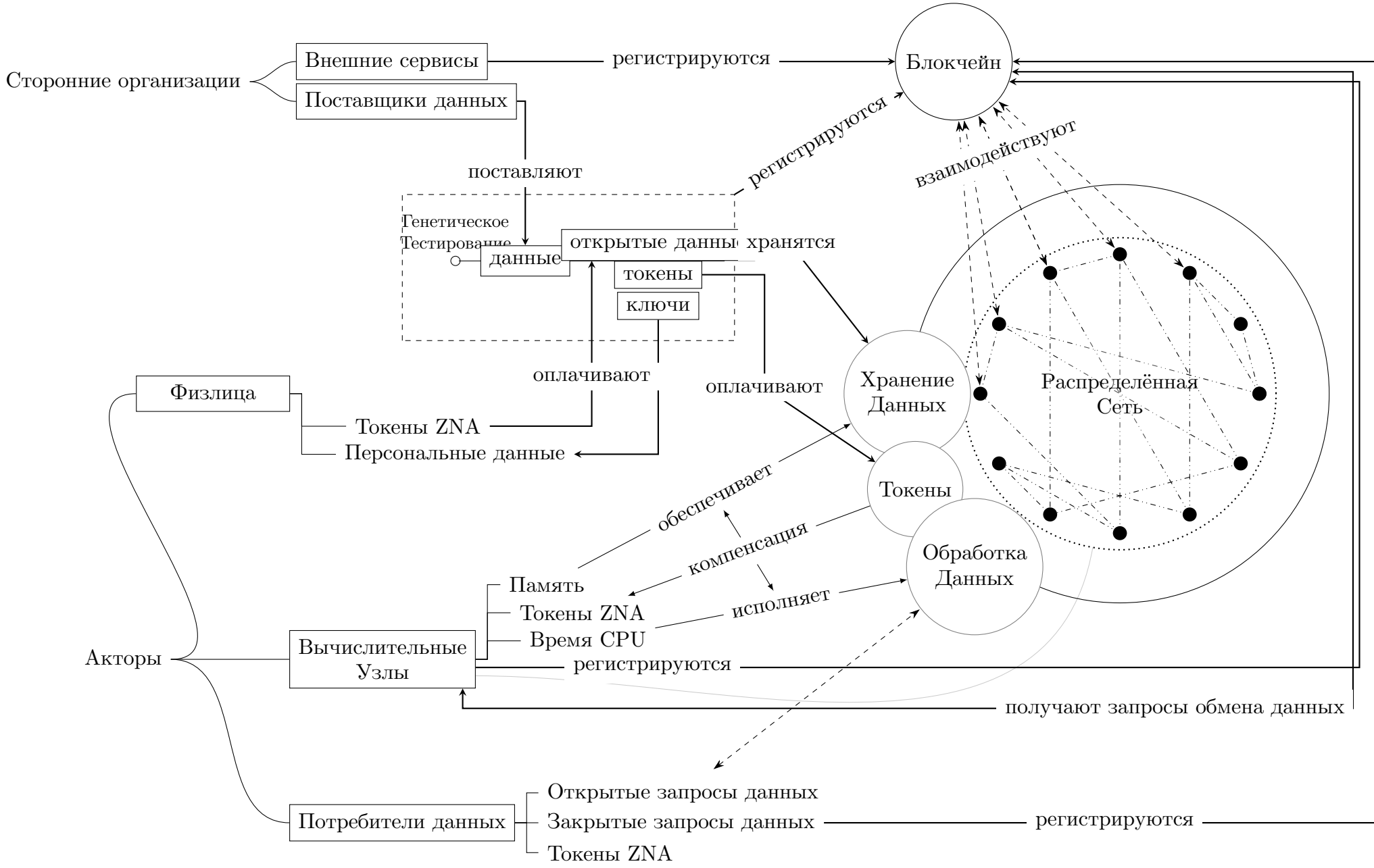
Появление криптовалют и лежащей в их основе технологии блокчейн существенно влияет на процессы трансформации многих экономических отраслей. Эта же технология может быть взята за основу для построения инфраструктуры для персональной геномики, где каждый человек — полноправный владелец своих генетических данных.

Проект Zenome как раз и ставит своей целью построение такой инфраструктуры. Наша платформа обеспечивает возможность управления персональными геномными данными, позволяя продавать доступ к различным частям своего генома, одновременно сохраняя конфиденциальность. Это должно поставить разработчиков лекарств в равные условия и подтолкнуть развитие научных и медицинских технологий.

Zenome — это, по сути, новое экономическое пространство, в основании которого лежат геномные данные и блокчейн. Реализация нашей концепции решает следующие задачи:

- создаёт инфраструктуру хранения больших геномных данных с использованием распределённых баз данных
- открывает, причём с сохранением конфиденциальности, доступ к данным миллионов человеческих геномов

- даёт каждому человеку возможность участвовать в научных и клинических исследованиях и зарабатывать на этом
- стимулирует совершенствование геномных наук в развивающихся странах и демонополизацию геномных данных в развитых


Сторонние организации
Внешние сервисы
Поставщики данных
регистрируются
Блокчейн
поставляют
регистрируются
взаимодействуют
Генетическое
Тестирование
данные
открытые данные
токены
ключи
хранятся
оплачивают
оплачивают
Физлица
Токены ZNA
Персональные данные
Хранение
Данных
Распределённая
Сеть
Токены
обеспечивает
компенсация
Обработка
Данных
Память
Токены ZNA
Время CPU
исполняет
Акторы
Вычислительные
Узлы
регистрируются
получают запросы обмена данных
Потребители данных
Открытые запросы данных
Закрытые запросы данных
Токены ZNA
регистрируются

# Оглавление

## Геномная Токеномика

## Глава 1 - Геномика



## Глава 2 - Концепция

# Вступление: Геномная Токеномика

*В следующих параграфах мы дадим общее описание платформы Zenome, приведём основания для выпуска собственных токенов, а также обсудим возможную выгоду для инвесторов.*

---

Как уже было сказано, подавляющая часть геномной информации хранится в базах данных, финансируемых государствами или крупными корпорациями. При этом в каждой такой базе данных содержится недостаточно данных для качественного скачка в геномике и персонализированной медицине. Но если посмотреть с другой стороны, каждая такая база содержит столько информации, что ее обработка уже не под силу одной отдельно взятой компании.

Получается, что обмен генетической информации — задача крайней важности. Перспективный генетический рынок должен гарантировать защиту от возможных злоупотреблений вообще и генетической дискриминации в частности. Особенно важно одновременно обеспечить прозрачность экономических взаимодействий, равный доступ для всех участников рынка и конфиденциальность генетических данных.

Глобальный обмен генетической информацией должен учитывать следующие проблемы:

- фрагментированность генетических данных
- ограничение доступа учёных, медиков и компаний к генетическим данным
- слабую доступность генетического тестирования, особенно в развивающихся странах
- недостаточность конфиденциальности добровольцев, выкладывающих свои генетические данные в открытый доступ

- ограниченность вычислительных ресурсов

Zenome ставит своей целью создание инфраструктуры персональной геномики, которая позволит участникам:

- загружать генетическую информацию и управлять ею
- безопасно хранить её
- зарабатывать на продаже доступа к ней или её части
- проходить генетическое тестирование в обмен на право использования полученной генетической информации
- получать персональные диетические рекомендации или программы упражнений, подобранные под индивидуальный генетический состав
- использовать другие генетические сервисы.

Главными потребителями генетической информации являются компании, заинтересованные в генетическом таргетинге, такие как Google, Facebook, Unilever и фармацевтические компании.

На платформе Zenome существуют разные типы информации — геномная, личная и финансовая — и все они неразрывно связаны. Особенности каждого типа определяют и метод хранения такой информации. Финансовые данные, в число которых входят и записи транзакций, хранятся на блокчейне. Анонимизированные геномные данные хранятся в распределённой сети. А личные данные хранятся только на собственных компьютерах пользователей. Разные подходы к обработке разных типов данных обеспечивают сохранение конфиденциальности и возможность масштабирования нашей системы.

Поскольку все транзакции, включая покупку и продажу данных, управляются смарт-контрактами, тем самым отражая децентрализованную природу нашей платформы, любая сделка в конечном итоге описывается только её балансом, хранимым на блокчейне. Использование для этой цели любых уже существующих токенов повлекло бы неоправданную зависимость нашей платформы от стоимости чужого токена или криптовалюты. По этой причине у нашей платформы должен быть собственный токен. Это, в частности, означает, что покупка генетических данных на платформе будет возможна только с использованием наших токенов, а не обычных денежных средств.

Zenome DNA (ZNA) и является таким токеном. Ценность ZNA таким образом привязывается к общему успеху платформы.

В этой белой книге мы подробно обсудим наиболее острые проблемы в области геномики и те их решения, которые предлагает платформа Zenome.

# Глава 1

# Геномика

## 1.1. Предыстория

*В этой секции мы дадим определения терминов "геном" и "геномика", обсудим историю первых попыток секвенирования генома и появления так называемых методов секвенирования нового поколения (СНП), опишем основных производителей реагентов и используемого для получения геномных данных оборудования, рассмотрим проблемы, связанные с накоплением геномных данных, и вопросы стоимости их анализа, проведём обзор актуальных баз геномных данных.*

---

### Геном

**Геном** — это полный набор генетических инструкций, находящийся в клетке [1].

Геном содержит биологическую информацию, необходимую для развития и функционирования организма. Человеческий геном образуется двойной спиралью линейных ДНК молекул и организован в виде 22 пар хромосом и 2 половых хромосом — X и Y. Вся информация, содержащаяся в геноме, закодирована четвертичным кодом из последовательности 4 нуклеотид, обозначаемых A, T, C и G. Выражение "прочесть геном" означает "определить последовательность нуклеотид методом секвенирования" [2].

Индивидуальная генетическая последовательность определяет широкое разнообразие черт организма, включая внешность, склонность к определённым заболеваниям, атлетические способности, метаболизм, пищевые предпочтения, совместимость половых партнёров (возможность зачать ребёнка) и многое другое.

## Международный проект "Геном Человека" (the Human Genome Project)

В 1990ом году с целью определить полную последовательность гаплоидного генома человека Национальным Институтом Здоровья США (the National Institute of Health, или NIH) был запущен международный проект "Геном Человека"[1]. Первым руководителем проекта стал один из первооткрывателей структуры ДНК лауреат Нобелевской премии Джеймс Уотсон.

Первый, черновой вариант последовательности генома человека был получен в середине 2000 года и опубликован в начале 2001 в журнале Nature. Стоимость этого международного проекта, финансируемого за счёт государственных средств, составила приблизительно 3 миллиарда долларов. В 1998 году в гонку секвенирования генома человека включилась частная компания Celera Genomics. Руководителем этого проекта, развивавшегося параллельно государственному, стал знаменитый учёный и предприниматель Крейг Вентер, которому удалось собрать 300 миллионов долларов частных инвестиций. Используя новой подход в секвенировании, называемый методом дробовика, и более эффективные вычислительные алгоритмы, Вентер получил секвенированную последовательность практически одновременно с международным консорциумом и опубликовал свои результаты в 2001 году[3] в журнале Science. "Полный" геном человека был опубликован в 2007 году, однако некоторые геномные участки, с трудом поддающиеся секвенированию, остаются непрочитанными до сих пор.

## Развитие геномного анализа

Масштабные инвестиции, огромное количество выдающихся участников из научной среды и конкуренция между частными и государственными организациями обеспечили значительный импульс развитию технологий геномного анализа. В результате появились современные технологии секвенирования, так называемые технологии СНП (секвенирование нового поколения)[2], а вместе с ними и новая ветвь науки под названием биоинформатика — молодая отрасль исследований, находящаяся на пересечении математики, информационных технологий и биологии, которая ставит своей задачей развитие техник и алгоритмов анализа эффективными вычислительными методами больших биологических данных.

Появление второго и третьего поколений технологий секвенирования (СНП) привёло к весомому падению стоимости геномного анализа. Ещё в 2009 году стоимость полного секвенирования генома составляла около 100 000 долларов,

[1] https://en.wikipedia.org/wiki/Human_Genome_Project

[2] https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-you-will-learn/what-next-generation-dna-

Таблица 1: Разработчики оборудования и поставщики реагентов, необходимых для генетического секвенирования. Рыночная капитализация взята с Yahoo Finance.

| **Компания** | **Продукт** | **Капита-лизация (долл.)** | **Страна** |
|---|---|---|---|
| Illumina | оборудование, реагенты, расходные материалы, программное обеспечение | 28.06 млрд. | США |
| Thermo Fisher Scientific | оборудование, реагенты, расходные материалы, программное обеспечение | 68.98 млрд. | США |
| Oxford Nanopore Technologies | оборудование, реагенты, расходные материалы | 534.41 млн. | Великобритания |
| Pacific BioScience | оборудование, реагенты расходные материалы | 436,93 млн. | США |
| Roche | оборудование, реагенты, расходные материалы, программное обеспечение | 213.44 млрд. | Швейцария |
| Agilent Technologies | оборудование, реагенты, расходные материалы, программное обеспечение | 19.32 млрд. | США |

а сегодня средняя цена за тот же анализ упала до уровня ниже 1000 долларов (см. картинку 1 и таблицу 3).


Рис. 1: Закон Мура и снижение стоимости геномного анализа. Резкое падение в 2008 году связано с изобретением технологий секвенирования нового поколения (СНП)[4].

https://en.wikipedia.org/wiki/Moore's_law

## Влияние на другие науки

Развитие геномики (науки, изучающей различные геномы) привело к трансформации многих научных областей — от биологии и антропологии до медицины и гуманитарных наук. Целый ряд ведущих коммерческих компаний, в число которых входят Google, Apple, IBM, Amazon и Alibaba, собираются использовать геномику для адаптирования своих продуктов и услуг под геномные профили пользователей. Такие изменения позволят этим компаниям очень точно выстраивать отношения между пользователями и предугадывать возможные нужды и действия своих клиентов [3].

## Генетические базы данных

Сокращение стоимости секвенирования привело к экспоненциальному росту объёмов доступных геномных данных. В зависимости от глубины секвенировнания, полный необработанный геном человека (”сырые” данные) может составлять от 50 гигабайт до 2 терабайт данных. Для хранения столь больших объёмов геномных данных были созданы специальные геномные базы, рассчитанные на конкретный тип данных: на необработанные данные (риды), полученные прямо от геномных секвенсеров, на последовательности генов и белков, на наборы кодирующих регионов генома, называемых экзомами, и даже на полные последовательности геномов. Некоторые из этих баз данных содержат клинически важную информацию о взаимосвязях, существующих между генетическими особенностями организма и болезнями. Большинство из них управляются централизованно и финансируются государствами и крупными корпорациями. Учёные со всего мира вовлечены в постоянный процесс добавления в эти базы новых данных, способствуя быстрому обновлению и синхронизации. В Таблице 2 приведены некоторые хорошо известные базы данных.

Большинство этих баз данных размещаются в развитых странах и централизованно управляются и контролируются государствами. Доступ к некоторым из них ограничен даже для научного сообщества или предоставляется только по коммерческой подписке. И хотя создатели геномных баз данных заверяют, что хранение осуществляется безопасным способом в анонимном виде, в действительности хранимые данные только псевдонимны, и известны случаи, когда люди идентифицировались по их геномной информации [5].

[3] https://www.smeal.psu.edu/fcfe/documents/innovations-in-medical-genomics-pdf

Таблица 2: Геномные базы данных

| GenBank | http://exac.broadinstitute.org |
|---|---|
| Владелец: | NCBI-NIH, USA |
| Описание: | База данных геномных последовательностей |
| На хранении: | более 199 341 377 различных генетических цепочек |
| **ExaC** | **www.ncbi.nlm.nih.gov/genbank** |
| Владелец: | Broad Institute of MIT and Harvard, USA, ODC Open Database License (ODbL) |
| Описание: | Консорциум Экзомной Агреграции |
| На хранении: | 60 706 последовательной экзомов человека |
| **UniprotKB** | **www.ebi.ac.uk/uniprot/** |
| Владелец: | EMBL-EBI, SIB, PIR, UK, Switzerland, USA |
| Описание: | Open Knowledge Base. Ручное курирование данных. Генетические и белковые последовательности. |
| На хранении: | более 555 100 вручную проверенных и аннотированных записей. |
| **ClinVar** | **https://www.ncbi.nlm.nih.gov/clinvar/** |
| Владелец: | NCBI-NIH, USA |
| Описание: | свободно доступный архив интерпретаций клинически значимых вариантов генома в контексте связанных с ними болезнями |
| На хранении: | более 158 000 интерпретаций более 125 000 генетических вариантов |
| **HGMD** | **http://www.hgmd.cf.ac.uk/ac/index.php** |
| Владелец: | QiaGen |
| Описание: | коммерческая база данных |
| На хранении: | 208 368 аннотированных записей мутаций человека |
| **SNPedia** | **https://www.snpedia.com/index.php/SNPedia** |
| Владелец: | открытая база данных |
| Описание: | викитека генетической информации человека |
| На хранении: | 107 037 ОНПов и связанных записей |
| **1000 Genomes Project** | **www.1000genomes.org** |
| Владелец: | EMBL-EBI, Wellcome Trust |
| Задача: | найти большую часть генетических вариантов с частотой встречаемости не менее 1% |
| На хранении: | более 2504 экземпляров генома и генетических последовательностей |
| **100000 Genomes Project** | **http://www.genomicsengland.co.uk/** |
| Владелец: | NHS, правительство Великобритании |
| Описание: | База данных, содержащая более 100 000 геномов |
| На хранении: | 32 642 последовательности полных геномов |

## 1.2. Обзор геномного рынка

*В этом разделе мы дадим краткий анализ рынка геномных технологий и приведём примеры наиболее популярных геномных продуктов и компаний, предоставляющих геномные услуги.*

## Динамика роста геномного рынка

Рынок геномных технологий стремительно растёт и открывает большие перспективы. На данный момент полный объём рынка оценивается в 25 миллиардов долларов с потенциалом практически десятикратного роста: от 5.9 миллиардов долларов в 2010 до 60 миллиардов долларов в 2020.


Рис. 2: Динамика роста геномного рынка за период 2010-2024.

## Виды геномной продукции

Как ключевой компонент 4ой промышленной революции геномика может быть использована в широком диапазоне приложений в практически любой отрасли экономики. Основные типы продукции современного геномного рынка таковы:

- **Пренатальный неинвазивный скрининг**

| | |
|---|---|
| **Описание:** | определение анеуплоидии плода[6] и пола ребёнка по ДНК плода, циркулирующей в материнской крови, уже на 6 недели беременности. |
| **Объём рынка:** | 4 млрд. долларов |
| **Лидера рынка:** | Illumina, Natera, Ariosa, Sequenom. |

- **Преимплантационный генетический скрининг**

| | |
|---|---|
| **Описание:** | генетическое профилирование эмбрионов, полученных при экстракорпоральным оплодотворении, методами ДНК-микрочипов или СНП-секвенированием перед имплантированием в матку |
| **Объём рынка:** | 336.4 млн. долларов |
| **Лидеры рынка:** | Illumina, Agilent Technologies |

- **Потребительское генетическое тестирование, SNP[4]-генотипирование**

[4]**Single nucleotide polymorphism (рус. *Однонуклеотидный полиморфизм*, или *ОНП*).**

| | |
|---|---|
| **Описание:** | SNP-генотипирование методом ДНК-микрочипов с целью получения рекомендаций по питанию, диетам, физическим упражнениям, определения статуса генетического переносчика. |
| **Объёмы рынка** | 2 млрд. долларов |
| **Лидеры рынка:** | AncestryDNA, 23andMe, DNAfit, deCode genetics. |

- **Остальная генетическая диагностика (включая онкогеномику)**

| | |
|---|---|
| **Описание:** | секвенирование генетических панелей, экзомов, полных геномов, жидких биопсий. |
| **Объёмы рынка:** | 16 млрд. долларов |
| **Лидеры рынка:** | Pathway Genomics, Human Longevity, Inc, Laboratory Corporation of America, Quest Diagnostics. |

# 1.3. Препятствия на пути геномики

*В этом разделе мы обсудим основные проблемы современной геномики, решение которых будет необходимо для реализации концепции Геномики 2.0 — повсеместного внедрения и использования персональных геномных технологий*

## Слабая доступность геномного анализа

Цены на различные виды геномного анализа приводятся в Таблице 3 и, как правило, начинаются от 1000 долларов. В сравнении со стоимостью чтения генома 5-10 лет назад, эта ценовая шкала чрезвычайно низка.[7].

Однако цена секвенирования и биоинформатической интерпретации всё ещё остаётся достаточно высокой для того, чтобы геномный анализ стал доступен всем слоям населения, особенно в развивающихся странах.

Получение больших объёмов геномной информации (последовательностей геномов вместе с фенотипическими характеристиками) жителей развивающихся стран является чрезвычайно важным с точки зрения широты обхвата всего генетического разнообразия населения планеты. В свою очередь это бы значительно простимулировало развитие рынка анализа больших геномных данных. Даже в большинстве развитых стран менее 2% населения когда-либо проходили генетическое исследование (микрочипное, экзомное секвенирование или секвениерование всего генома).

Таблица 3: Типы геномного анализа.

| **ДНК-микрочип** | |
|---|---|
| **Описание:** | анализ 1-5 миллионов заранее выбранных ОНП |
| **Ценовой диапазон:** | $100-500 |
| **Объём информации:** | 0.033% полного генома |
| **Поставщики:** | 23andMe, AncestryDNA, DNAfit. |
| **Генетические панели** | |
| **Ценовой диапазон:** | $100-2000 |
| **Объём информации:** | 0.001-1% полного генома |
| **Поставщики:** | Pathway Genomics, CeGaT. |
| **Секвенирование экзомов** | |
| **Описание:** | секвенирование кодирующих частей генома (экзомов) |
| **Ценовой диапазон:** | $250-3000 |
| **Объём информации:** | 2% полного генома |
| **Поставщики:** | BGI, CeGaT. |
| **Секвенирование полного генома** | |
| **Ценовой диапазон:** | $600-10000 |
| **Объём информации:** | 80-98% полного генома |
| **Поставщики:** | BGI, FullGenomes, Human Longevity. |

*Замечание: в действительности некоторая часть генома не секвенируется. Размер такой части зависит от глубины секвенирования и от нюансов приготовления образца. Поэтому термин "полное геномное секвенирование" на самом деле означает получение последовательности нуклеотид, покрывающей чуть более 80% генома.*

## Потеря конфиденциальности как цена участия в открытых биомедицинских исследованиях.

Осознанное согласие на сбор и обработку персональных данных — ключевой аспект любого биомедицинского исследования. Каждый проект, включающий исследование генома человека, начинается с получения подтверждения понимания участниками целей исследования и согласия с условиями. Форма такого согласия зависит от проекта и может включать разрешение на использование данных в будущих проектах, последствия чего могут быть непредсказуемы.

Например, в проекте "Персональный Геном" (the Personal Genome Project) Гарвардской Школы Медицины участники добровольно соглашались, что их данные и экземпляры их генетического материала могут быть использованы много раз и быть переданы другим лабораториям для собственных исследований. Участники явным образом информировались, что могут быть деанонимизированы и что их частные данные могут стать общественно доступными. Этот проект ставит своей целью обеспечить доступность геномных данных как можно бо́льшего числа людей и тем самым

простимулировать новые исследования и развитие геномной индустрии. Его руководители уверены, что если мы не обеспечим свободный обмен информацией и открытый доступ к геномным данным, то рискуем оказаться в ситуации тысяч изолированных друг от друга частных коллекций геномных данных (коллекций фармацевтических компаний, геномных корпораций и научных центров), каждая из которых, однако, не будет обладать необходимым минимумом данных для обеспечения прорывных исследований.

## Проведение международных научных и клинических исследований

Для проведения новых исследований в области геномики требуется проводить научные и клинические испытания на крупных выборках, полученных от разных групп населения. На данный момент сбор генетических экземпляров от людей с разным этническим происхождением сопряжён со сложностями организации специальных экспедиций и получения разрешения от местных регулирующих структур.

Сегодня только один стартап (`https://www.dnasimple.org/`) пытается решать эту задачу — за небольшую оплату и с обещанием соблюдения анонимности.

## Создание биобанков, хранящих материал от различных индивидов

Биобанки, храня и обрабатывая биологические материалы (экземпляры крови, ткани костного мозга), выполняют функцию посредников между донорами и исследователями. В целом, биобанки — это ключевой инструмент в развитии персонализированной медицины и разработки лекарств. Одной из главнейших функций биобанков является сбор донорского материала, включая кровь, ткани костного мозга и даже половые клетки, для последующего использования.

На сегодняшний момент биобанки активно разиваются во многих странах. Однако не исключена вероятность того, что наиболее ценные образцы биоматериала будут принадлежать самым обеспеченным с точки зрения финансов биобанкам, таким, например, как биобанки крупных фармацевтических компаний, что приведёт к неравноправному доступу разных категорий исследователей к биоматериалу. Таким образом существует необходимость создания биобанков в как можно б*о*льшем количестве стран и городов, вплоть до индивидуальных биорезервуаров человека.

## Обработка: скорость обработки генома ограничена сложностью биоинформатических алгоритмов. Адаптация аналитического программного обеспечения с целью подготовки к массовому использованию.

На данный момент скорость получения геномных данных с использованием технологий геномного анализа (секвенирования) высока и превосходит скорость обработки получаемых данных. Если изучить любое крупное научное исследование большого числа геномов, экспериментальные этапы, связанные с получением геномных данных, занимают не более 20% продолжительности всего исследования, а существенно б*о*льшая часть времени приходится на этапы обработки. Здесь под обработкой данных мы подразумеваем этапы исследования от получения необработанных, секвенированных данных до интерпретации результатов и поиска различных ассоциаций.

Ещё одна проблема с современным программным обеспечением связана с тем, что программы геномного анализа писались учёными для учёных и требуют соответствующей адаптации для их широкого применения в среде медиков и потребителей. Поскольку уже существуют геномные анализаторы размером с USB-флешку, использование персональных секвенсеров в ближайшем будущем может перестать быть научной фантастикой и стать такой же будничной реальностью, как использование персональных компьютеров сегодня.

## Интерпретация геномных данных: математические модели рисков развития заболеваний

Для ранжирования рисков развития заболеваний на основании генетических данных используются различные модели и алгоритмы. Первоначальные типы этих моделей основываются на типе рассматриваемого наследования: моногенное, полигенное или мультифакторное. Оценка рисков развития мультифакторных и сложных заболеваний требует учёта влияния и взаимодействия многих генов, равно как и факторов среды. Для более подробного ознакомления с существующими методами оценивания рисков развития сложных заболеваний смотрите [8, 9].

Для создания новой модели оценки таких рисков требуется участие большого числа учёных. А также исследование публикаций по тематике анализируемого заболевания с целями идентификации типа наследования, определения полиморфизмов и мутаций, способствующих развитию состояния, и выработки "геномной алгебры", то есть набора правил, по которым и проводится оценка таких рисков. Когда модель готова и подтверждена компьютерным симулиро-

ванием, должны быть проведены клинические испытания для оценки её применимости на практике. Такой подход на данный момент является наиболее точным, но и затратным как по времени, так и по общим усилиям.

## Интерпретация геномных данных: использование алгоритмов машинного обучения.

Использование алгоритмов машинного обучения для оценивания рисков мультифакторных заболеваний сегодня активно исследуется, но пока что по причинам малых размеров обучающих выборок алгоритмы машинного обучения проигрывают математическим моделям, разработанным биологами.

Однако уже сейчас машинное обучение успешно используется для предугадывания некоторых сложных характеристик тела человека. Примером такого использования может служить прогнозирование внешности человека в работах Крейга Вентера и его коллег. Суть их работы заключается в анализировании геномных данных и примерно 30000 базовых точек лица, собранных с нескольких тысяч волонтёров. На основании этих данных для алгоритмов машинного обучения были построены обучающие выборки — и тем самым выявлены зависимости между генетическими особенностями и индивидуальными чертами внешности. Благодаря этой работе, машины научились точно восстанавливать внешность человека на основании только его или её геномных данных [10, 11].

Результаты этого проекта позволяют составлять фото-робот преступника или изображение внешности ещё не родившегося ребёнка на самых ранних стадиях беременности. Получив образец крови от беременной женщины и выделив из него ДНК плода ребёнка, мы можем предсказать его внешность на момент наступления 18 лет.

Во время выполнения данного проекта Крейг Вентер пользовался услугами одного из лучших специалистов по машинному обучению из Гугла Фрэнка Оча, главного архитектора сервиса Google Translate [12].

Пока машинное обучение широко для исследования заболеваний не используется, так как требуется большое число правильно структурированных образцов для тренировки алгоритма. Создание полноценной базы данных генома человека с привязкой к результатам детализированных опросников о состоянии здоровья может подтолкнуть развитие тренинга алгоритмов машинного обучения в геномике, что приведёт к высокой точности предсказания рисков развития заболеваний. В то же время эти данные будут публичными и доступными для всех пользователей системы, исключая тем самым возможности их монополизации. Такая доступность чрезвычайно важна, так как концентрация больших объёмов данных в корпоративных базах данных будет вести к монополизации и в области применения машинного обучения к геномике.

## Безопасное хранение большого объёма данных

Безопасность персональных данных крайне важна: мы все стараемся предотвратить кражу данных наших банковских карт, данных страховых полисов, номеров банковских счетов и медицинской информации любого вида. Кража геномной информации многим людям может сегодня показаться маловажной. Однако подобное развитие событий может привести к серьёзным последствиям, которые невозможно предугадать. Например, к подбрасыванию на место преступления синтезированной цепочки генома человека с целью подлога.

Сегодняшние решения этой проблемы включают в себе зашифрованное хранение данных на центральном сервере, как это реализовано [13] у `https://www.pathway.com/`, `https://www.23andme.com/` и `http://www.humanlongevity.com/`. Такой тип закрытого хранения данных относительно безопасен, но не предоставляет возможности обмена данными и доступа к ним учёным со всего мира, что является критически важным условием для развития современной геномной науки.

Ещё одна проблема, связанная с хранением данных, относится к большому размеру самого генома и к экспоненциальному росту количества геномных данных, поскольку всё больше и больше людей проходят геномное секвенирование. В одном исследовании [14] предсказывалось, что к 2025 году полный объёма хранимых геномных данных (за размер одного полного генома бралось 100 гигабайт) достигнет 40 экзобайт в год, а хранение геномных данных называлось одним из главных потребителей рынка услуг хранения и обработки информации.

Таблица 4: 4 домена Больших Данных в 4025

| | |
|---|---|
| **Астрономия** | |
| **Скорость накопления** | 25 зеттабайт в год |
| **Объёмы хранения** | 1 1 экзобайт в год |
| **Анализ** | урезание данных in situ; обработка в реальном времени; массивные тома; |
| **Распространение** | выделенные линии от антенн до серверов (600 ТБ в секунду) |
| **Twitter** | |
| **Скорось накопления** | 0.5-15 миллиардов твитов в год |
| **Объёмы хранения** | 1-17 петабайт в год |
| **Анализ** | майнинг данных тематики и мнений; анализ метаданных; |
| **Распространение** | малые единицы |
| **YouTube** | |
| **Скорость накопления** | 500-900 миллионов часов видео в год |
| **Объёмы хранения** | 1-2 экзобайт в год |
| **Анализ** | ограниченные требования |
| **Распространение** | основной наполнитель интернет-канала современного пользователя (10 МБ в секунду) |
| **Геномика** | |
| **Скорость накопления** | 1 зеттабайт в год |
| **Объёмы хранения** | 2-40 экзобайт в год |
| **Анализ** | гетерогенные данные и анализ; отбор вариантов, ∼2 триллиона процессоро-часов; выравнивание генетических последовательностей, ∼10000 триллионов процессоро-часов. |
| **Распространение** | Много небольших (10 МБ в секунду) и ещё меньше массивых (10 ТБ в секунду) пересылок данных |

## Общая база данных с непрерывно обновляемыми опросниками.

В то же время недостаток публичных баз данных, выполненных с использованием концепции распределённого хранения и использования open-source программного обеспечения, может привести к полному доминированию рынка компаниями, обладающими мощным серверным оборудованием, вроде Google и Amazon[5]. Если корпорации и фармацевтические компании станут монополистами в области геномной информации, развитие медицины будет происходить долго и дорого, а в лечении заболеваний продолжат доминировать современные подходы — вместо подходов будущего, когда болезни предотвращаются ещё до их развития или на самых ранних стадиях симптоматики.

[5] `https://cloud.google.com/genomics/`

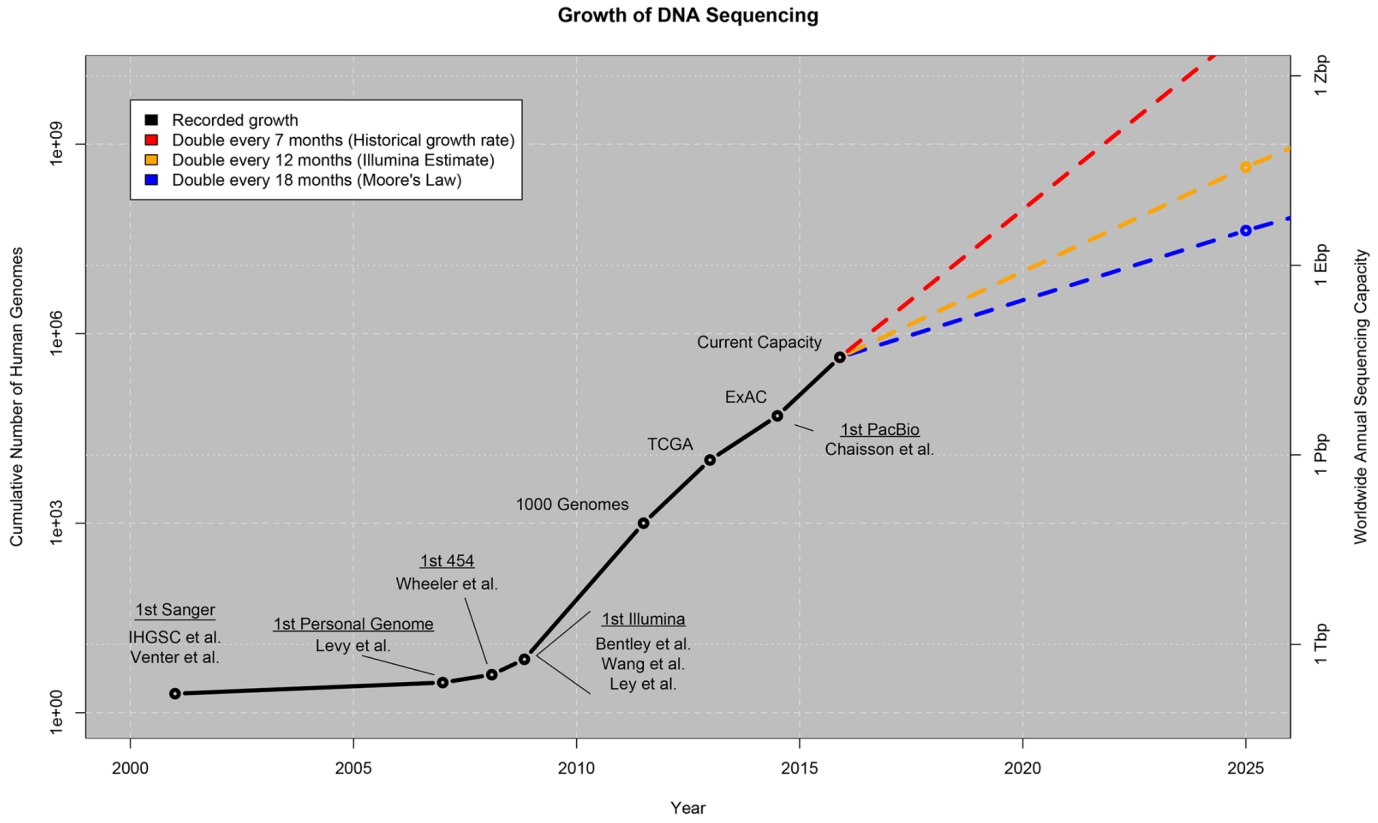



Рис. 3: рост секвенированных ДНК-данных. На этом графике показан рост секвенированных ДНК, выраженный в числе всех уже просеквенированных геномов человека (на левой оси) и в мировом годовом потенциале секвенирования (на правой оси: Тера-пары (Tbp), Пета-пары (Pbp), Экзо-пары (Ebp) и Зета-пары (Zbp)). Значения до 2015 года основаны на опубликованных исторических данных с пометками знаковых для секвенирования событий (от первого секвенирования Сенгером до первой публикации полного генома человека PacBio), а также трёх выдающихся проектов широкомасштабного секвенирования: Проект Тысячи Геномов, собравший к 2012 году сотни человеческих геномов; Атлас Раковой Геномики (или TCGA), собравший несколько тысяч парных соответствий генов опухоли и нормальной ткани; и Консорциум Агрегирования Экзомов (или ExAC), собравший больше 60000 человеческих экзомов. Многие собранные к сегодняшнему дню геномы по сути являются полными экзомами, а не полными геномами, но в будущем можно ожидать выправления этой ситуации в пользу полных геномов. Приведённые на картинке значения для года 2015 и позднее являются прогнозом в трёх возможных сценариях. Взято из [14].

### Ближайшее будущее геномики

В завершении этой главы опишем некоторые гипотетические, но технически реализуемые перспективы и риски, которые могут встать на пути у геномной индустрии и общества вообще:

- снижение стоимости геномного анализа и миниатюризация инструментов геномного секвенирования — вплоть до уровня подключаемых к мобильному телефону модулей
- взрывной рост геномных данных и их хранения и появление геномных хакеров и проблем конфиденциальности и защиты данных
- широкое распространение геномной и теле-медицины
- изменения в индустрии питания, связанные с возможностью подбора по геному персональных диет
- развитие терапии персонализированными лекарствами
- услуги знакомств на основе геномной совместимости
- персональная идентификация по геномной информации, включая возможность осуществлять платежи и получать услуги
- увеличение средней продолжительности жизни во всех странах, продление периода активного долгожительства и сильный сдвиг в сторону старения населения, поздних беременностей и уменьшения рождаемости
- развитие технологий редактирования геномов
- дизайн внешности будущих детей и другие возможности, которые сложно предсказать. К примеру, техническая возможность выбора здоровых детей и определения их будущей внешности как на стадии эмбриона, так и через получение ДНК плода от беременной женщины уже существует [10].

## 1.4. Этические вопросы персональной геномики

*В этом разделе мы рассмотрим проблемы этики, связанных с развитием и всесторонним внедрением геномных технологий, такие как вопросы конфиденциальности, публичных баз данных,*

*открытого доступа к ним для исследователей, возможных злоупотреблений и угроз личным свободам от распространения геномики*

---

Заметка

В нашей белой книге мы приводим только краткое описание основных проблем и перспектив развития геномной индустрии. Для более детального обсуждения вопросов конфиденциальности и защиты информации обратитесь к `https://www.smeal.psu.edu/fcfe/documents/innovations-in-medical-genomics-pdf`.

## Конфиденциальность

Для многих людей тема персональной геномной информации является деликатной. Однако многие до конца не понимают, что по их геномной информации можно определить их продолжительность жизнь, склонность к принятию эмоциональных решений, вероятность развития различных душевных расстройств и риск внезапной смерти, например, от сердечной аритмии.

Такая информация может оказаться неблагоприятной при устройстве на работу, принятии участия в выборах и при покупке медицинской страховки. Не исключается и возможность подбрасывания фрагментов ДНК, идентичных известному геному, на место совершения преступления для подлога и компрометирования конкретного человека. Возможны и отказы в медицинском лечении (или требования более высокой оплаты) или недопуск к желаемой работе.

Корпорации и государства могут также умышленно влиять на индивидуальные решения и покупки, используя знания о "слабостях" в геномной информации человека. Таким образом для обеспечения равных прав всех категорий людей совершенно необходима защита конфиденциальности геномных данных.

В то же время некоторые исследования показывают, что технически возможно определение людей по их анонимным геномам [5, 15].

И более того, некоторые компании (`http://www.humanlongevity.com/media/`) обладают алгоритмами, основанными на машинном обучении, которые могут точно реконструировать внешность человека только по его или её геному [16].

## Публичность

Такой подход подстегнёт развитие превентивной медицины, благодаря чему анализ больших данных позволит предугадывать развитие заболевания ещё до постановки диагноза, позволяя принимать меры, направленные на увеличение продолжительности жизни и улучшение её качества [17], а также идентификацию доноров всего мира. К сожалению, на сегодня не существует готовых решений, допускающих публичное использование геномной информации и одновременно охраняющих неприкосновенность частной жизни. Однако можно упомянуть некоторые стартапы, работающие в этой области и использующие технологию блокчейн.

Encrypgen — стартап, недавно проведший ICO, описал существующие проблемы и отношения частного и публичного с точки зрения блокчейна[18]. Однако их белая книга лишена описания технических деталей, которые собственно решают проблемы конфиденциальности и доступности.

Ещё один проект в этой области — стартап DNAbits [19], чей основатель Дрор Сэмьюэл Брама запатентовал общий подход к хранению и передаче данных через блокчейн [20]. Однако эта компания за последние 3 года так технически и не реализовала свою концепцию.

## Должно ли существовать право собственности на геномную информацию?

На данный момент не существует законного определения права на владение собственной генетической информацией. В некоторых развитых странах, включая США, Германию и Австрию, граждане не имеют права доступа и владения своей генетической информацией с целью её интерпретации[21]. Требуется обращение к некоторым посредникам в виде врача или медицинского центра с такими правами. Такой подход используется компаниями Pathway Genomics в США и CeGaT в Германии (`http://www.cegat.de/en/`).

Для прохождения генетического анализа необходимо направление врача, который может выступить и исполнителем генетического анализа, и только такой врач имеет право интерпретировать информацию, полученную в результате этого анализа.

В США существуют сервисы, работающие в области "развлекательной генетики", такие как 23andMe и Ancesty.com, которые продают генетические тесты сразу потребителю, минуя врача, но эти компании имеют право предоставлять только информацию об этническом происхождении или определённом наборе характеристик, относящихся к здоровью человека (например, спортивные характеристики), и не имеют права предоставлять

наиболее ценную с медицинской точки зрения информацию. Эти ограничения, наложенные регулирующими структурами, такими как, например, FDA, не мешают 23andMe продавать доступ к собранным ими генетическим данным фармацевтическим компаниям. Некоторые из таких сделок хорошо известны: сделка с Genentech (подразделение фармацевтического гиганта Roche) за 60 миллионов долларов [22] для исследования болезни Паркинсона и сделка с другим фармацевтическим гигантом, Pfizer, для исследования воспалительных заболеваний кишечника (болезни Крона) [23]. Некоторые доклады утверждают, что 23andMe также вела переговоры с Novartis об участии в исследовании болезни Альцгеймера [24].

Таким образом мы сегодня предоставляем крупным компаниям право управлять нашей геномной информацией, хранить её и зарабатывать на ней. Корпорации, прячась за маской благих намерений, монополизируют геномные данные, и мы не можем предсказать, как это монопольное положение скажется на ценах будущих лекарств и на совершении медицинских открытий.

## Право доступа к геномной информации

Существует ещё один этический вопрос, требующий внимания. Выше мы отмечали, что нарушения конфиденциальности и доступ к геномной информации могут нелегально производиться, например, работодателем. Человек может быть уволен или ему может быть отказано в повышении только на основании его генетического профиля. Для тех, чья работа связана с безопасностью людей и систем, например, для водителей автобусов или грузовиков, пилотов, операторов атомных станции и т.д., состояние здоровья критически важно, и их геномная информация может предотвратить аварию или даже катастрофу. В некоторых профессиях потенциальная опасность может угрожать самому работнику, а не случайным прохожим, например, шахтёру с заболеваниями лёгких. Здесь потребуется обсуждение с профессионалами и экспертами, а также со всем обществом, необходимости разработки законных правил, регулирующих использование геномной информации работодателями в таких случаях.

# Глава 2

# Концепция

## 2.1. Проект Zenome

### Философский взгляд

Информированность населения о геномной медицине всё ещё остаётся довольно низкой в развитых странах и практически нулевой в развивающихся. Это значит, что люди в целом слабо понимают и вероятные выгоды геномики, и те угрозы, которые могут от неё исходить. Во многих странах это привело к развитию чрезвычайно сложного комплекса административных ограничений, направленных на защиту генетической информации от возможных злоупотреблений — процедур, которые замедляют научный прогресс.

Платформа Zenome ставит своей целью повышение осведомленности геномной медициной, что позволит людям принимать осознанные решения в отношении своих данных. Чтобы это стало возможным, Zenome исходит из следующих фундаментальных принципов:

**Частное владение персональной генетической информацией** Каждый участник обладает всеми правами на свои генетические данные.

**Свобода выбора** Каждый участник свободно выбирает, как использовать свою генетическую информацию, в частности в каких научных/клинических испытаниях участвовать.

**Право совместного пользования** Каждый участник может предоставить доступ к своей генетической информации третьей стороне, делая ненужным копирование данных.

**Конфиденциальность** Шифрование частных данных делает невозможным доступ к персональной генетической информации без явного разрешения пользователя.

**Распределённое хранение данных** Благодаря масштабированию данных и резервному копированию, архитектура распределённых баз данных обеспечивает высокую доступность и устойчивость к ошибкам.

**Распределённая обработка данных** Данные обрабатываются на нескольких сетевых узлах одновременно. Стать таким узлом может каждый пользователь, предоставив дисковое пространство и вычислительные мощности своего компьютера.

**Масштабируемость** Архитектура платформы допускает высокую масштабируемость и гибкость системы.

## Экосистема платформы Zenome

На платформе Zenome пользователь[1] участвует во взаимодействиях разных типов. Эти взаимодействия происходят на разных уровнях системы, не мешая друг другу и используя разные шаблоны взаимодействий. Таким образом они должны быть представлены как отдельные элементы, исполняющие отличные роли.

На платформе Zenome существуют следующие роли:

**Узел вычислений/хранения** предоставляет дисковое пространство и вычислительные мощности за вознаграждение.

**Физлицо** загружает персональные генетические данные на платформу и возможно пользуется предоставляемыми на ней генетическими услугами.

**Аналитик** анализирует генетическую информацию на платформе. Это может быть учёный-специалист по обработке данных, научная организация и т.д.

**Поставщик услуг** предоставляет пользователям генетические услуги (возможно платные). Проще говоря, это любая организация, для которой обработка генетических данных — бизнес.

Короче

**Каждый пользователь занят в целом ряде взаимодействий разных типов, принимая на себя различные роли. Некоторые из них, например Поставщик услуг и Аналитик , требуют специальных знаний, а другие (Узел и Физлицо ) — нет.**

Каждая роль ещё будет обсуждаться в деталях.

[1] В широком смысле — это может быть и человек, и запущенная им программа.

## Обзор системной архитектуры

Платформа Zenome — это распределённое приложение, состоящее из 3 основных уровней.

**Сетевой уровень (уровень доступа к данным)** : обеспечивает уровень абстракции, инкапсулирующий взаимодействие с сетью, и предоставляет верхним уровням архитектуры интерфейс работы с распределённой средой.

| | Блокчейн | DHT Kademlia |
|---|---|---|
| Цена хранения данных | Высокая | Низкая |
| Неизменяемость данных | Есть | Нет |
| Производительность | Низкая | Высокая |
| Детерминированность результата | Есть | Нет |

Этот уровень состоит из двух распределённых систем совершенной разной природы:

**Распределённый реестр** (основанный на блокчейне) проверяемым и долговременным способом записывает транзакции между участниками. Для доступа к блокчейну программное обеспечение узла запускает встроенный Ethereum-клиент.

**Распределённая хэш-таблица** (основанная на протоколе Kademlia) комбинирует физические узлы в оверлейную сеть и делает возможным обмен сообщениями между узлами и распределённым хранилищем данных.

Роль

**Узел** (вычисления/хранения) работает на сетевом уровне.

**Средний уровень** содержит управление аккаунтами, целостный интерфейс к функциям безопасности подлежащих уровней и высокоуровневые API для программного обеспечения, исполняющегося на прикладном уровне.

Роль

Средним уровнем управляет **Поставщик услуг** . Так называемый API Платформы Внешних Сервисов позволяет третьим сторонам предоставлять генетические услуги прямо на нашей платформе.

**Прикладной уровень** Приложение Zenome отличается продвинутым пользовательским интерфейсом, который целостным образом переводит действия пользователя на средний уровень. Интерфейс может быть легко расширяем, чтобы генетические сервисы могли исполняться прямо в нашем приложении.

## Рынок генетических услуг

Рынок генетических услуг на данный момент развивается в следующих направлениях:

1. **исследование и внедрение технологий** на рынок
2. предоставление **геномно-диагностических услуг**
3. **государственная сертификация** генетических технологий
4. **разработка правовой базы**. В частности, мер для защиты генетической информации.

Заметка

Структура генетического рынка весьма сложна. Некоторые игроки, столкнувшись с быстро растущими потребительскими потребностями, пытаются найти своё место на рынке, развиваясь сразу в нескольких направлениях.

На этом рынке можно обнаружить участников следующих типов:

**Научные корпорации** работают над разработкой и внедрением новых технологий на рынок. В их число входят:

- Фармацевтические корпорации, биотехнологические и диагностические компании, такие как Pfizer и Myriad
- Компании, разрабатывающие и продающие все необходимые реагенты, такие как Life Technologies

**IT-биоинформационные компании** заняты разработкой методов вычислительной обработки данных. В этом секторе ключевой проблемой до сих пор остаётся большие объёмы и сложные типы получаемых данных.

**Научные и медицинские центры** играют ключевую роль в предоставлении и развитии генетических диагностических услуг.

**Коммерческие лаборатории** предоставляют быстрый, эффективный и, как правило, относительно дешевый доступ к генетическо-диагностическим услугам. Такие лаборатории обладают большими ресурсами, включая финансовые.

**Компании, оказывающие услуги конечным потребителям** повышают интерес среди населения к генетической диагностике. На данный момент этот сегмент очень мал, но в будущем он может вырасти в одну из ведущих частей рынка и быть принят в медицинскую практику.

Таблица 5: Сравнение с другими продуктами на рынке

| | Zenome | GeneCoin | Encrypgen | 23andMe | Pathway Genomics | Snpedia (Promethease) | Human longevity |
|---|---|---|---|---|---|---|---|
| Децентрализованность | ✓ | ✓ | ✓ | – | – | – | – |
| Пригодность для нечеловеческих организмов | ✓ | ✓ | ✓ | – | – | – | – |
| Клиент является владельцем своих данных | ✓ | ✓ | ✓ | – | – | ✓ | – |
| Возможность загрузки собственных данных | ✓ | ✓ | ✓ | – | – | ✓ | – |
| Открытость неличной информации | ✓ | ✓ | – | – | – | – | ✓ |
| Самостоятельно анализирует данные | ✓ | ✓ | – | ✓ | ✓ | ✓ | ✓ |
| Предоставляет отчёты клиентам | ✓ | – | – | ✓ | ✓ | ✓ | – |
| Использует AI и машинное обучение | ✓ | – | – | – | – | – | ✓ |
| Обмен без передачи массивных данных | ✓ | – | ✓ | – | – | – | – |
| Возможность заработка за счёт своих данных | ✓ | – | – | – | – | – | – |
| Открыта для учёных | ✓ | – | – | – | – | ✓ | – |
| Платформа для других инструментов | ✓ | – | ✓ | – | – | – | – |

## 2.2. Роли на платформе Zenome

### Точка зрения ресурсных узлов

Опишем систему с точки зрения вычислительного узла.

**Узел** — участник, предоставляющий платформе ресурсы своего компьютера (дисковое пространство и вычислительные мощности) с целями распределённого хранения и обработки генетических данных за вознаграждение токенами ZNA.

Чтобы стать вычислительным узлом, пользователь запускает на своём компьютере программу Zenome и активирует роль `Узел` в графическом интерфейсе пользователя. Программа должна исполняться непрерывно в фоновом режиме.

Заметка

Для превращения своего компьютера в вычислительный узел существует специальная версия приложения Zenome в виде командной строки.

Настройки предоставления системных ресурсов и управления задачами могут гибко меняться самим владельцем узла:

- максимальный объём дискового пространства, которое может быть использовано приложением
- режим использования вычислительных ресурсов:

  **Фиксированное использование** процессора/графического процессора — количество ядер и их максимальная загрузка (в процентах для каждого ядра)

  **Динамическое использование** — ресурсы предоставляются программе таким образом, чтобы это не мешало остальной работе пользователя.

- управление приоритетами вычислительных процессов (по идентификатору).
- временное выключение узла. Все данные, недоступные другим узлам, переносятся с выключаемого узла.
- постоянное отключение узла. Все данные, недоступные другим узлам, переносятся с отключаемого узла.

## Физлицо/Пользователь

**Физлицо** — человек, желающий предоставить свою генетическую информацию в платформу Zenome с целями заработка на продаже своих данных или получения генетических услуг на платформе.

Для работы с персональной генетической информацией пользователь устанавливает программу Zenome.

Используя графический интерфейс, пользователь может:

- Создать личный аккаунт.
- Управлять токенами: получать, переводить, использовать, платить за хранение данных, тратить на генетические услуги.
- Загружать генетическую информацию (формат данных определяется автоматически).
- Управлять персональными данными и заполнять опросники.
- Обрабатывать индивидуальные предложения.
- Использовать генетические услуги и настраивать уровень конфиденциальности для каждой услуги в отдельности.

При загрузке генетической информации может быть выбран один из следующих уровней конфиденциальности:

**Полная конфиденциальность** В этом случае все данные хранятся в зашифрованном виде и с пользователя взимается полная плата за их хранение.

**Стандартная конфиденциальность** Генетическая информация хранится в виде фрагментов, не допуская тем самым идентификацию по ней пользователя. Каждый фрагмент хранится в явном виде. Информация о связи фрагмента с пользователем закрыта. В этом случае дисковое хранение данных субсидируется системой.

**Полный (открытый) доступ** Все данные хранятся в открытом виде. Хранение бесплатно.

Заметка

Внимание! Хотя технических ограничений против повышения уровня конфиденциальности нет, это будет позволено сделать только в отношении свежих данных. Данные, однажды сделанные открытыми, уже никогда не

станут частными. Имейте это в виду при загрузке данных в первый раз.

### Потребитель данных

**Потребитель данных** — это учёный, коммерческая компания, научная организация или любой другой участник платформы, который заинтересован в анализе генетических данных с использованием возможностей платформы. У потребителя данных есть возможность отправлять пользователям запрос и объявлять вознаграждение за отклик на него.

Заметка

На подобные запросы накладываются некоторые ограничения. Это делается для того, чтобы предотвратить деанонимизацию пользователю и другие злоупотребления платформой. Строгость ограничений зависит от рейтинга потребителя данных в системе.

### Поставщик услуг

**Поставщик услуг** — организация, для которой использование генетической информации является частью бизнеса. То есть это те организации, которые предоставляют пользователям платформы генетические услуги.

Заметка

*Пользователь свободен в выборе, какие данные он или она желает предоставлять поставщикам услуг. Пользователь предупреждается, если запрашиваемые данные могут быть использованы для его или её идентификации.*

Заметка

Поставщик услуг, используя прямой запрос, не может увидеть более 70% списка мутаций в явном виде. Или получить информацию, сопоставляющую сырые данные (например, fastq-файл) с пользовательским опросником.

## 2.3. Геномные данные

### Типы геномных данных

На платформе Zenome используется 3 типа геномных данных, отличающихся по их ценности с точек зрения оплаты и содержащейся информации:

- Открытые данные, не являющиеся ценными для владельцев, но важные для учёных.
  **Пример:** геном разновидности бактерий Helicobacter pylori.
- Открытые данные, ценные как для владельца, так и для всего сообщества.
  **Пример:** геномы большинства участников сети.
- Данные с ограниченным использованием, которые просто хранятся в сети.
  *Для служебных проектов разных коммерческих и государственных организаций.*

## Предобработка генетических данных

Обработка СНП-геномных данных, как правило, состоит из двух следующих независимых шагов:

1. **Первичная обработка сырых данных.**
2. **Целенаправленный анализ генетических последовательностей** с целями выработки персональных рекомендаций или проведения исследований.

Заметка

**Референсный геном** — оцифрованный набор ДНК-данных, считающихся характерными для геномов определённого вида организмов.

Предобработка СНП-геномных данных состоит из следующих шагов:

1. Выравнивание СНП-ридов по референсному геному.
2. Поиск мутаций и других отличий от референсного генома и сохранение их в списке gVCF-формата.

Заметка

*Для организмов, отличных от человеческого, последовательность шагов та же самая. Меняется только референсный геном.*

Если ДНК-данные получены с использованием технологии микрочипного генотипирования (формат файлов сервиса 23andMe), они могут быть загружены на платформу точно так же, как и данные формата gVCF, поскольку форматы данных этих типов файлов совпадают.

## Проблема фальшивых данных

Если вместо правильного генома случайно или умышленно загружаются генетические данные другого (не человеческого) организма, система выявит подлог ещё на этапе предобработки сырых данных и уведомит об этом пользователя. Если пользователь загружает искажённые (фальшивые) генетические данные умышленно, это будет выявлено общеизвестными методами верификации до отгрузки данных в хранилище.

Заметка

Для мотивирования пользователей не загружать фальшивые данные используется экономическое стимулирование в виде требования оплаты услуг хранения данных вперёд за весь год.

## Проблема идентификации пользователя по генетическим данным

Открытый доступ к генетической информации поднимает проблему идентификации пользователя по их геномным или другим данным. Если пользователь решает не открывать полностью свою генетическую информацию, необходимо предпринять соответствующие меры на каждом этапе обработки и хранения данных. Для решения этой проблемы все взаимодействия в системе должны быть выстроены таким образом, чтобы на каждом этапе ни один из узлов не мог бы быть в состоянии определить владельца генетического материала или даже город, в котором он или она живёт.

Заметка

Генетические различия жителей одного города составляют примерно 0.01% ДНК-последовательностей.

На каждом этапе эта цель достигается разными способами:

1. На стадии предобработки — через разделение исходного файла на части так, что покрытие в среднем оказывается меньше уровня доверия (6 копий).
2. На стадии хранения — через фрагментирование данных по длине.

## Хранение геномных данных

Геномные данные расположены в распределённой сети, основанной на протоколе DHT Kademlia. Участники, предоставляющие сети свои ресурсы

(см. описание роли `Узел`) получают за это выплаты в виде токенов ZNA. Для получения платежа требуется подтверждение размещённости данных на компьютерах получателя. Процедура такой проверки основана на использовании блокчейна и, где это необходимо, использует шифрование.

Заметка

*В будущем также будет реализована интеграция с узлами Storj и FileCoin.*

Как уже было сказано, данные могут сырыми и обработанными.

| Тип данных | Сырые | Обработанные |
|---|---|---|
| Features | | |
| Формат | fastq / bam | (список мутаций) gvcf / vcf + bed / 23me(txt) |
| Размер, Гб | 50 | 2 |
| Назначение | Улучшение технологий секвенирования и обработки данных (для развития оборудования) | Проведение исследований и составление отчётов |
| Условия хранения | | |
| Количество копий | По крайней мере 3 в независимых узлах | По крайней мере 5 в независимых узлах |

Геномные данные хранятся в виде фрагментов такой длины, что не позволяет однозначно определить, геному какого человека мог бы принадлежать данный фрагмент.

Заметка

*Информация о том, геному какого пользователя соответствуют геномные фрагменты, является частной и может быть получена только с разрешения самого пользователя.*

## 2.4. Личные профили

Заполнение персональных опросников значительно увеличивает применимость геномных данных. Пользователи заполняют опросник, используя графический интерфейс.

Заметка

Каждый аналитик системы может создать собственный опросник и разместить его на платформе. Если опросник станет популярным, приложение предложит пользователям заполнить его.

## Характеристика опросника

Число опросников может быть велико, поэтому необходимо введение понятие характеристика опросника.

**Характеристика опросника** — это полное описание всех полей опросника и допустимых ответов.

Заметка

Говоря формально, характеристика опросника содержит ссылку на автора, описание и упорядоченный список записей, каждая из которых соответствует одному полу опросника.

Поля ответов могут быть нескольких типов:

**Числовые поля** — в качестве ответа принимается целое число.

**Множественный выбор** — в качестве ответа принимается номер правильного ответа.

Заметка

Ответы на вопросы этих типов могут быть прочитаны всеми, так как это не угрожает конфиденциальности.

**Строковое поле** — в качестве ответа принимается строка. Так как по введённому ответу такого типа возможна идентификация пользователя, это поле является закрытым.

**Частный блок** — позволяет сделать любое поле частным, вне зависимости от его типа.

## Запросы к системе

Статистические данные, построенные по геномной информации, если владелец не принял решение о её шифровке, а также доступные ответы

к опросникам, будут открытыми. Таким образом каждый может узнать, например, количество 25-летних пользователей системы с мутацией rs6026 (фактор коагуляции V).

Архитектура системы делает невозможной извлечение полной базы данных:

- При создании ассоциативного запроса заказчик не получает доступа к сырым данным.
- Начальная плата включает ограниченное количество запросов в день. Плата за дополнительные запросы растёт экспоненциально в течение дня.
- Если результат ассоциативного запроса содержит менее 100 пользователей, результат не будет предоставлен заказчику.

Если какой-либо пользователь зашифрует свои данные, тогда только он или она может решить, кому передавать свои данные или их фрагменты. Ни один аналитик не будет знать, какого типа данные зашифрованы.

## Безопасная передача персональных данных

Процесс передачи персональных данных между участниками системы должен обладать следующими свойствами:

1. Передаваемые данные во всей полноте должны быть доступны только покупателю и продавцу.
2. Перевод токенов должен осуществляться, только если передача данных была успешна.
3. Попытки продажи неверных данных должны определяться и блокироваться.
4. Попытки умышленных обвинений продавца в предоставлении неверных данных должны выявляться.
5. Передача данных не может доверяться третьей стороне.

Заметка

Для безопасной передачи данных используется блокчейн. Однако должно учитываться, что хранение (и передача) больших объёмов информации на блокчейне ресурсоёмко. Поэтому:

6. Через блокчейн разрешается передача только небольшого количества данных. Основная часть передаётся через обычный зашифрованный канал.

## 2.5. Система рейтинга

На платформе будет создан рейтинг для:

- каждого фрагмента генетической информации и блока персональных данных
- организаций
- поставщиков услуг
- поставщиков данных (лабораторий по секвенированию ДНК)

Заметка

В системе не будет индивидуального рейтинга пользователя, так как он может быть представлен в виде суммы рейтингов его или её генетических фрагментов, недоступных всему сообществу. Если генетические данные загружаются с аккаунта организации, их базовый рейтинг будет автоматически увеличен на величину рейтинга организации.

Факторы, влияющие на рейтинг генетических фрагментов:

**Подтверждение лабораторией качества данных** повышает рейтинг загружаемых данных в пропорции к рейтингу самой лаборатории, в которой генетический материал был получен. Подтверждение может осуществляться методом цифровой подписи лаборатории или через загрузку данных самой лабораторией по просьбе своего клиента.

**Проверка правдоподобности** позволяет проверять генетическую информацию с использованием статистических моделей частот полиморфизма и генетических связей. На данный момент этот модуль находится в разработке.

**Участие в исследованиях** повышает рейтинг участвующего в исследовании фрагмента. Если результаты такого исследования оказываются неправдоподобными, рейтинг фрагмента понижается.

## 2.6. Примеры использования

### Физлицо/Рядовой пользователь

Благодаря Zenome, для рядовых пользователей появляется возможность получить свою генетическую информацию и превратить её в источник дохода.

Комбинация генома и его взаимодействия с окружающей средой — важный источник информации. Наша платформа позволит пользователю безопасно управлять этим источником.

Наша платформа даёт возможность безопасно хранить и обмениваться генетической информацией, позволяя пользователю получать различные генетические услуги. Приведём несколько примеров:

- отчёты и рекомендации по питанию, рискам развития заболеваний, косметологии, диетам, фитнес-программам
- поиск родственников и определение родословной
- услуги знакомств
- индивидуальный подбор одежды, обуви, климатических настроек в доме, направлений для путешествий и мест проживания
- разные типы генетических отчётов для малых групп, например спортивных команд или рабочих коллективов
- и многое другое, что будет придумано компаниями, использующими нашу платформу — ведь практически каждый аспект человеческой жизни подвержен влиянию генетики.

Помимо множества сервисов, пользователь получает возможность заработать на своей генетической уникальности, предоставляя исследовательским компаниям данные заполненных опросников. Таким образом персональные данные вместе с генетической информацией человека становятся аналогом товара или природного ресурса.

## Здоровье

Современная система здравохранения и персонализированная медицина не может быть представлена без использования геномных технологий. Наша платформа позволит пациентам безопасно обмениваться генетической информацией, используемой в клинике, с медицинским персоналом:

- индивидуальные дозировки и непереносимость лекарств (например, подбор индивидуальной дозировки антикоагулянта варфарина на основе генетического профиля)
- собственные диапазоны допустимых биологических параметров (например, уровень маркера ПСА)

- генетическая предрасположенность к разным заболеваниям (например, оценка риска макулярной дегенерации с целью предупреждения развития заболевания)

- трансплантация и донорство органов. Пользователи могут безопасно обмениваться информацией касательно типа их HLA-антигенов для определения совместимости в случае трансплантации. Таким образом, возможно появление безопасной базы данных доноров и волонтёров для спасения жизни через донорство.

## Компании

Есть 2 типа компаний: первые предоставляют услуги на основе геномных данных, вторые заинтересованы в получении генетических данных для проведения исследований с ними.

Первый тип компаний уже был описан в разделе примеров применения пользователями. Второй тип может быть описан следующим образом: компания покупает геномные данные пользователя для проведения исследований и улучшения качества потребительских продуктов, генетического таргетинга своих продуктов и рекламы. Вот несколько примеров:

- Фармацевтическая компания планирует выпуск нового лекарства против мутировавшего белка раковых клеток. Используя нашу систему, компания может найти пользователей, которые пережили болезнь, заплатить им за генетическую информацию и получить частоты мутаций в гене, кодирующем белок, который и является целью препаратов компании.

- Потребительская компания собирается выйти на новый рынок и хочет узнать, как будет принят потребителем её продукт. Хорошо известно, что ароматизация вызывает отторжение у носителей особых генетических вариантов генов вкусовых рецепторов. Через нашу сеть эта компания посылает предложение носителям таких генов поучаствовать в исследовании, по результатам которого может быть выбран другой аромат или другой целевой рынок.

## Научное сообщество

Для научного сообщества наша системы открывает возможность хранить, обмениваться и проводить исследования с большим разнообразием геномных данных. Поскольку платформа не ограничена работой только с геномом человека, она может быть использована и для хранения и обработки геномных

данных важных для, например, сельского хозяйства (геномы растений, животных, микроорганизмов).

В целом, наличие нашей платформы обогащает научное сообщество доступом к данным генеральной популяции, даже и без данных индивидуальных опросников. А кроме того, с разрешения пользователей они также могут стать частью научного исследования.

Наша платформа также обеспечивает вычислительные мощности, доступ к которым позволит обрабатывать огромные объёмы генетических данных (аналогично адаптированной для работы с генетическими данными AWS).

# Глава 3

# Техническая часть

## 3.1. Распределённые объекты

### Концепция распределённой подсистемы

**(Распределённая) подсистема** — это набор из части системных (платформенных) элементов и процессов, которые в объектно-ориентированной логике могут быть представлены как отдельная сущность, с точки зрения внешнего наблюдателя демонстрирующая вполне определённое поведение.

Для детального описания характеристик этой подсистемы перечислим её аспекты:

1. **Структура**: формирующие подсистему элементы и процессы.
2. **Внешнее поведение**: взаимодействие подсистемы как единого целого с другими участниками. В частности:
   - **Интерфейсы**: набор допустимых запросов к подсистеме.
   - **Действия**: действия, выполняемые подсистемой в отношении других участников системы.
3. **Внутреннее состояние**: обобщённое внутреннее состояние подсистемы.

Подсистема может быть представлена как **квазиобъект**, с которым взаимодействуют другие участники.

Заметка

Префикс «квази-» означает, что в действительности взаимодействие происходит с элементами подсистемы, которые, в свою очередь, взаимодей-

ствуют друг с другом. В совокупности же это может рассматриваться как взаимодействие с неким цельным объектом.

Заметка

*Всюду далее мы не будем делать разницы между подсистемой и её представлением в виде квазиобъекта.*

Взаимодействия с подсистемой могут быть описаны как последовательность небольшого числа базовых операций, таких как:

- действия участников в отношении подсистемы.
- действия в отношении других участников.
- внутренние процессы, меняющие внутреннее состояние системы.

### Внутренние процессы

**Внутренний процесс** в широком смысле — это набор всех внутренних процессов каждого элемента подсистемы и взаимодействий между этими элементами.

**Внутренний процесс** в узком смысле — это процессы внутри подсистемы как единого целого, изменяющие её внутреннее состояние. Полное описание внутренних процессов должно содержать описание поведения подсистемы без конкретизации, как именно оно реализовано.

## 3.2. Основные распределённые подсистемы платформы

Платформа выстроена на следующих организационных уровнях:

- Основой системный уровень (критическая инфраструктура)
- Уровень хранения и обработки данных
- Высоко-уровневые взаимодействия

На каждом уровне располагаются следующие подсистемы:

1. Основной системный уровень

**Низко-уровневое внутреннее взаимодействие** Базовая подсистема обмена сообщениями между узлами сети. Также позволяет создавать распределённые хэш-таблицы.

**Авторизация** Инфраструктура аккаунтов и управления доступом к частной информации.

2. Уровень хранения и обработки данных

**Хранение** Уровень абстракции доступа к распределённой файловой системе.

**Обработка** Инфраструктура распределённых вычислений.

3. Высоко-уровневые взаимодействия

**Безопасные запросы** Подсистема, формирующая предложения покупки геномной информации, которые могут быть показаны только нужным пользователям.

**Операции с открытыми данными** Инструменты управления с открытыми (не частными) данными.

**Платформа внешних сервисов** API для подключения сторонних, централизованно управляемых сервисов к платформе. Организует передачу данных по обычным веб-протоколам.

## 3.3. Уровень низко-уровневых внутренних взаимодействий

### Распределённая P2P сеть

Базис платформы — P2P сеть на протоколе Kademlia из компьютеров пользователей, установивших приложение Zenome.

**Узел** — узел распределённой P2P сети, соответствующий компьютеру пользователя с установленным приложением Zenome.

Заметка

Таким образом между устройствами пользователей выстраивается оверлейная сеть. По сути, это виртуальная сеть, в которой каждый пользователь опознаётся по предоставляемому идентификатору "NodeId", а не по IP-адресу устройства. В такой сети каждый узел хранит список "ближайших" узлов, рассчитанный по формуле дальности на основе "NodeId". С обычным расстоянием эта дальность не имеет ничего общего.

Узлы хранят данные, используя распределённые хэш-таблицы.

### Детали реализации распределённой системы

В приложении используется изменённый протокол со следующими отличиями:

- Узлы могут обмениваться друг с другом произвольными сообщениями. Для пересылки сообщения узлу достаточно знать "NodeId" получателя.
- Внутри системы допускается существование нескольких хэш-таблиц. Хэш-таблица идентифицируется по строке-ключу.
- Разные таблицы могут подчиняться разными правилам хранения и удаления данных.
- Данные, передаваемые с одного узла на другой, шифруются (см. далее).

### Обмен сообщениями в распределённой среде

Участники пиринговой сети могут обмениваться следующими сообщениями:

| | |
|---|---|
| `PING` | Проверка подключённости узла к сети |
| `STORE(T,K,V)` | Сохранение значения V по ключу K в таблице T на узле-получателе сообщения |
| `FIND_NODE(N)` | Получение информации о ближайших узлах узла N |
| `FIND_VALUE(T,K)` | Извлечение значения V из таблицы T по ключу K на узле-получателе (если таблица в месте ключа пуста, узел возвращает список ближайших узлов) |
| `SEND(M,N,D?)` | Отправка сообщения M с данными D ключу N. Для локализации узла используется 'FIND_NODE'. |

Заметка

Этот уровень взаимодействия по своей сути является транспортным уровнем платформы.

## 3.4. Блокчейн

### Основная информация

Платформа использует блокчейн Ethereum, который представляет единую децентрализованную виртуальную машину (EVM). Нужная логика системы реализуется через смарт-контракты.

Цитата

> Контракт — это набор из кода (функций контракта) и данных (состояния контракта), расположенный по определенному адресу на блокчейне Ethereum. — Introduction to Smart Contracts (Solidity manual)

Таким образом смарт-контракты способны хранить данные. Для данных блокчейна система абстрактных типов, описанная выше, также применима (с некоторыми ограничениями).

## Подсистема работы с блокчейном

Заметка

> Внимание: конкретная реализация архитектуры зависит от используемой платформы. Далее описывается случай платформы PC.

Для обеспечения доступа к блокчейну программное обеспечение узлов полностью реализует Ethereum Node, обеспечивая доступ средствами 'JSON-RPC 2.0'.

Закрытые ключи хранятся в зашифрованном хранилище. От пользователя при первом запуске приложения требуется задать пароль.

Заметка

> Хотя технически и возможно использование существующего аккаунта, рекомендуется создание нового.*

Интерфейс предложения позволяет:

1. Создать новый аккаунт
2. Импортировать существующий аккаунт
3. Создать резервную копию частного хранилища

Заметка

> Внимание: резервные копии рекомендуется размещать на облачных сервисах — это позволит сохранять доступ к аккаунту даже в случае физической недоступности компьютера.

Заметка

*Приложение Zenote предоставляет доступ к командной строке узлов Ethereum. Эта функция предназначается в первую очередь для отладки (дебагинга). Использование командной строки без понимания её назначения не рекомендуется.*

## Внутренние токены

Экономические взаимодействия внутри системы обеспечиваются использованием внутренних токенов ZNA, которые могут покупаться и продаваться на онлайн-биржах. По своей сути токены ZNA — это Ethereum-токены.

## Концепция аккаунта

**Аккаунт** платформы позволяет пользователю взаимодействовать с системой, играя одновременно несколько ролей. Каждая роль аккаунта соответствует отдельному смарт-контракту на блокчейне.

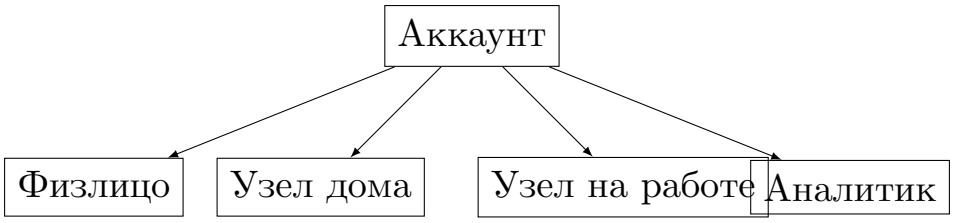



Рис. 4: Аккаунт пользователя, предоставившего свой геном платформе (роль "Физлицо"), который работает биоинформатиком и поддерживает 2 вычислительных узла - дома и на работе.

**Роль** — это характеристика отдельных типов взаимодействий пользователя системы, в которых он или она принимает участие. Каждый участник может играть несколько ролей одновременно.

Заметка

За распределение ролей отвечает отдельная подсистема.

# 3.5. Распределённая сеть хранения данных

## Принципы работы распределённой сети

Для хранения данных различных типов используется распределённая сеть, основанная на протоколе DHT Kademlia. Узлы этой сети представлены зарегистрированными в системе поставщиками ресурсов (вычислительных и ресурсов хранения), которые получают вознаграждение за хранение информации и исполнение вычислительных задач на своих компьютерах. Выплачивает подобное вознаграждение отгрузивший свои данные участник, но в целом ряде практически важных случаев эти траты полностью или частично субсидируются нашей системой. Например, в случае хранения информации, важной с точки зрения науки.

Единица хранения — это произвольный блок данных и уникальный ключ от него.

Заметка

Для предотвращения злоупотреблений размер блока данных ограничен.

**Продажа внешних сервисов.** В будущем будет реализована возможность интеграции с узлами Storj и FileCoin, а также с централизованно управляемыми хранилищами.

## Гарантии надёжности хранения данных

С учетом того, что узлы могут свободно подключаться и отключаться от системы, для обеспечения надёжности хранения данные независимо хранятся на нескольких узлах одновременно. Число узлов, на которых дублируются данные, определяется типом данных.

Таблица 6: Примерные параметры генетических данных.

| | Сырые | Обработанные |
|---|---|---|
| Формат | fastq/bam | gvcf/vcf + bed/23me(txt) |
| Размер | 50 Гб | 2 Гб |
| Узлы | $\geq 3$ | $\geq 5$ |

Каждый раз, когда число узлов изменяется, хранимые данные перераспределяются — для удовлетворения требования о минимальном количестве узлов.

В виду того, что каждый, кто хочет, может стать узлом в распределённой сети, невозможно полностью гарантировать, что узел хранит данные на основании только того, что так утверждает сам владелец узла. Для этого в соответствии с протоколом уровня безопасности производится периодическая проверка размещённости данных. Результаты такой проверки вносятся в блокчейн и могут быть поводом для вознаграждения узла или изменения его рейтинга.

### Гарантии конфиденциальности хранимых данных

Конфиденциальность хранения данных в описанной распределённой сети становится возможной благодаря применению асимметричной криптографии. Фактический метод шифрования определяется соответствующим протоколом безопасности.

Заметка

Заметим, что механизм субсидирования не может быть применим при хранении шифрованных данных.

### Некоторые особенности хранения геномной информации

Принадлежность некоторого геномного фрагмента конкретному индивиду может быть однозначно определена в случае, если фрагмент достаточно велик. Поэтому геномы в распределённой сети хранятся в виде довольно коротких фрагментов.

Фрагментация генома всегда осуществляется на основании референсного генома. Для каждого референсного генома (например, другого варианта референсного генома человека или генома другого вида) фрагментация выбирается единожды и каждому получившемуся фрагменту присваивается идентификатор, уникальный для референсного генома.

## 3.6. Личные данные пользователей

Запись — минимальная единица передаваемой информации. Таким образом невозможно передать (или продать) часть информации записи. На основе части данных записи возможно создание новой записи, которая с точки зрения потенциальных покупателей, однако, может оказаться ввиду её низкого рейтинга намного менее привлекательной.

Схема данных определяет, какая информация и в каком формате может содержаться записью. Использование схемы даёт гибкий инструмент унифика-

ции форматов обмениваемых между участниками системы данных. Идентификатор схемы данных может быть произвольной строкой, по которой пользователь сможет в сети найти описание схемы — например, гипер-ссылка или адрес смарт-контракта с описанием. Основная цель схем данных — объединение представлений о том, что является данными с точек зрения покупателей и продавцов.

## Характеристики опросников

Заполнение индивидуальных опросников значительно повышает ценность геномных данных. Существует большое число опросников, поэтому видится необходимым ввести концепцию характеристики опросника.

**Характеристика опросника** — это полное описание всех полей опросника со всеми возможными вариантами ответов в каждом поле.

Говоря формально, характеристика опросника содержит ссылку на автора, описание и упорядоченный список записей, каждая из которых соответствует одному полу опросника. **Числовое поле** В случае числового поля ответ пользователя должен укладываться в некоторый диапазон допустимых значений $[a, b]$. Ответом является беззнаковое целое, начало допустимых значений которого отсчитывается от левой границы диапазона.

Заметка

Ответы на вопросы такого типа могут быть прочитаны всеми, так как это не угрожает конфиденциальности.

**Множественный выбор** Ответом пользователя является беззнаковое целое, представляющее порядковый номер в списке допустимых ответов. Если ответ равен нулю, это означает, что пользователь предпочёл не отвечать на данный вопрос.

Заметка

Внимание: ответы на вопросы такого типа могут быть прочитаны всеми, так как это не угрожает конфиденциальности.

**Строковый ответ** Ответ пользователя хранится в виде строки

Заметка

Ответы на вопросы такого типа являются частной информацией.

**Заполненный опросник** — структура данных, содержащая ответы пользователя на вопросы опросников.

Заметка

*Если необходимо, заполненные опросники могут быть зашифрованы. Они хранятся на компьютере пользователя и могут быть загружены в зашифрованном виде на блокчейн.*

Запросы на предоставление пользователями своих личных данных отличаются следующими чертами:

- Не все пользователи соответствуют критериям конкретного исследования. Для проверки такого соответствия может потребоваться доступ к персональной информации.
- Персональная информация не должна передаваться за пределы пользовательского компьютера ни в явном (прямо), ни в неявном ("свидетельство канарейки") видах.
- В случае если пользователь удовлетворяет критериям, он или она получает предложение передать персональные данные за вознаграждение.

Вывод 1

Так как передача данных не допускается, проверка соответствия должна проводиться в изолированном окружении. В этом окружении исполняющийся код имеет полный доступ к частной информации, но не может взаимодействовать с другими частями системы.

Вывод 2

Результаты исполнения такой изолированной процедуры не могут быть посланы потенциальным покупателям из-за угрозы конфиденциальности.

Вывод 3

Представляется разумным выбрать ограниченный набор персональных данных, на основе которых можно было бы сделать необходимые выводы. При этом пользователь может видеть, какие данные для этого отбираются.

Вывод 4

Данные, подлежащие передаче, демонстрируются пользователю в явном виде на экране. Только данные, необходимые для проверки, могут быть переданы неявным образом, но и список с этими данными предоставляется пользователю для ознакомления.

Запрос персональных данных:

- **На первом шаге** запрашивается доступ к персональным данным пользователя.

  Проверяющей функции будут доступны только те данные, которые были указаны в запросе.

- На втором шаге внутри изолированной среды запускается код проверяющей функции, и результат её исполнения представляет из себя готовое предложение по обмену данными или отказ от него. В последнем случае данные не передаются.

  В первом случае пользователь уведомляется о сформированном предложении, после чего он или она может проверить сделанное предложение, список данных, использованных во время проверки, и данных, затребуемых для передачи.

  Если пользователь отказывается от предложения, передачи данных не происходит.

  Если пользователь принимает предложение, происходит передача только затребованных данных.

# Список литературы


[1] NHGRI. *Talking Glossary of Genetic Terms. Word «Genome».*
URL: https://www.genome.gov/glossary/index.cfm?id=90.

[2] Venter J.C., Smith H.O., Adams M.D. "The Sequence of the Human Genome". В: *Clinical Chemistry* 61.9 (2015), с. 1207—1208.
URL: http://clinchem.aaccjnls.org/content/61/9/1207.long.

[3] Adams M.D. Venter J.C. Smith H.O. "The Sequence of the Human Genome". В: *Science* 291.5507 (2001), с. 1304—1351. ISSN: 0036-8075.
DOI: 10.1126/science.1058040.
URL: http://science.sciencemag.org/content/291/5507/1304.

[4] Wetterstrand KA. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).*
URL: www.genome.gov/sequencingcostsdata.

[5] Melissa Gymrek и др. "Identifying Personal Genomes by Surname Inference". В: *Science* 339.6117 (2013), с. 321—324. ISSN: 0036-8075.
DOI: 10.1126/science.1229566.
URL: http://science.sciencemag.org/content/339/6117/321.

[6] H. Christina Fan и др. "Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood". В: *Proceedings of the National Academy of Sciences* 105.42 (2008), с. 16266—16271.
DOI: 10.1073/pnas.0808319105.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2562413/.

[7] Sboner, Andrea and Mu, Xinmeng Jasmine and Greenbaum, Dov and Auerbach, Raymond K. and Gerstein, Mark B. "The real cost of sequencing: higher than you think!" В: *Genome Biology* 12.8 (авг. 2011), с. 125. ISSN: 1474-760X.
DOI: 10.1186/gb-2011-12-8-125.
URL: https://doi.org/10.1186/gb-2011-12-8-125.

[8] Rachel R. J. Kalf и др. "Variations in predicted risks in personal genome testing for common complex diseases". В: *Genet Med* 16.1 (янв. 2014), с. 85—91. ISSN: 1098-3600.
DOI: 10.1038/gim.2013.80.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3883880/.

[9] Karen Norrgard. “Calculation of Complex Disease Risk”. B: *Nature* (2008).
URL: https://www.nature.com/scitable/topicpage/calculation-of-complex-disease-risk-756.

[10] Kathryn Nave. “How Craig Venter is fighting ageing with genome sequencing”. B: *WIRED UK* (2016).
URL: http://www.wired.co.uk/article/craig-venter-human-longevity-genome-diseases-ageing.

[11] Amalio Telenti и др. “Deep sequencing of 10,000 human genomes”. B: *Proceedings of the National Academy of Sciences* 113.42 (2016), с. 11901—11906.
DOI: 10.1073/pnas.1613365113.
eprint: http://www.pnas.org/content/113/42/11901.full.pdf.
URL: http://www.pnas.org/content/113/42/11901.abstract.

[12] Luke Timmerman. “Google Translate Star Leaves Venter’s Human Longevity For Illumina-Backed Grail”. B: *Forbes* (2016).
URL: https://www.forbes.com/sites/luketimmerman/2016/09/27/google-translate-star-leaves-venters-human-longevity-for-illumina-backed-grail.

[13] João Sá Sousa и др. “Efficient and secure outsourcing of genomic data storage”. B: *BMC Medical Genomics* 10.2 (июль 2017), с. 46. ISSN: 1755-8794.
DOI: 10.1186/s12920-017-0275-0.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5547444/.

[14] Zachary D. Stephens и др. “Big Data: Astronomical or Genomical?” B: *PLOS Biology* 13.7 (июль 2015), с. 1—11.
DOI: 10.1371/journal.pbio.1002195.
URL: https://doi.org/10.1371/journal.pbio.1002195.

[15] Erika Check Hayden. “Privacy protections: The genome hacker”. B: *Nature* (2013).
URL: http://www.nature.com/news/privacy-protections-the-genome-hacker-1.12940.

[16] Laure-Anne Pessina. “Reconstructing a face from DNA: an EPFL alumnus takes the stage at the 2016 TED Conference”. B: *School of Engineering (Federal Institute of Technology Lausanne)* (2016).
URL: http://sti.epfl.ch/page-129921-en.html.

[17] Melanie Swan. “Health 2050: The Realization of Personalized Medicine through Crowdsourcing, the Quantified Self, and the Participatory Biocitizen”. B: *Journal of Personalized Medicine* 2.3 (2012), с. 93—118. ISSN: 2075-4426.
DOI: 10.3390/jpm2030093.
URL: http://www.mdpi.com/2075-4426/2/3/93.

[18] Patrick Lin. “Blockchain: The Missing Link Between Genomics and Privacy?” B: *Forbes* (2017).
URL: https://www.forbes.com/sites/patricklin/2017/05/08/blockchain-the-missing-link-between-genomics-and-privacy.

[19] Justin Zimmerman. “DNA Block Chain Project Boosts Research, Preserves Patient Anonymity”. B: *CoinDesk* (2014).
URL: https://www.coindesk.com/israels-dna-bits-moves-beyond-currency-with-genes-blockchain/.

[20] D.S. Brama. “Method, System and Program Product for Transferring Genetic and Health Data”. US Patent App. 14/218,865. Июль 2015.
URL: https://www.google.com/patents/US20150205929.

[21] Melanie Swan. *Blockchain: Blueprint for a New Economy*.
URL: https://www.goodreads.com/book/show/24714901-blockchain.

[22] Matthew Herper. “Surprise! With $60 Million Genentech Deal, 23andMe Has A Business Plan”. B: *Forbes* (2015).
URL: https://www.forbes.com/sites/matthewherper/2015/01/06/surprise-with-60-million-genentech-deal-23andme-has-a-business-plan.

[23] “23andMe, Pfizer to Launch Inflammatory Bowel Disease Genetics Study”. B: *GenomeWeb* (2014).
URL: https://www.genomeweb.com/clinical-genomics/23andme-pfizer-launch-inflammatory-bowel-disease-genetics-study.

[24] “Genetic Wild West: 23andMe Raw Data Contains 75 Alzheimer’s Mutations”. B: *Alzforum* (2017).
URL: http://www.alzforum.org/news/community-news/genetic-wild-west-23andme-raw-data-contains-75-alzheimers-mutations.