



Л е к ц и я № 5

Компьютерная геномика Секвенирование

Разработал профессор П. М. Васильев
Кафедра фармакологии и биоинформатики

Для студентов, обучающихся по направлению 06.03.01 «Биология»
профили Биохимия, Генетика
при изучении дисциплины «Биоинформатика»

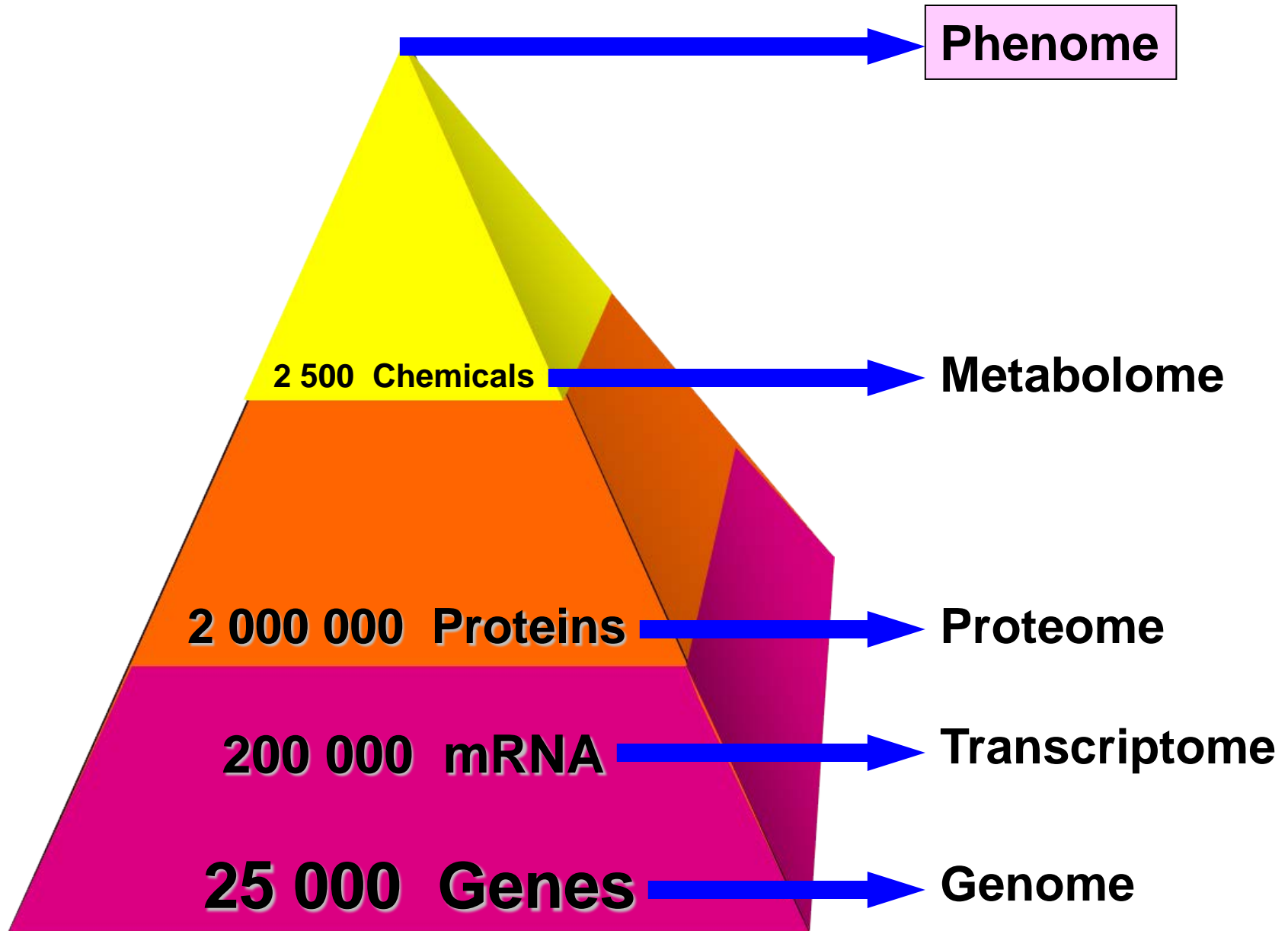
П л а н л е к ц и и

- **Геном, транскриптом, протеом, метаболом: механизмы формирования.**
- **Геномика, транскриптомика, протеомика, метаболомика.**
- **Методы секвенирования.**
- **Полимеразная цепная реакция (ПЦР).**
- **Генотипирование.**
- **Структура генома.**
- **Типы мутаций.**

Компьютерная геномика

- Предсказание генов в последовательностях
- Предварительная аннотация по сходству белковых последовательностей
- Сравнительный анализ геномов
- Исследование регуляции работы генов
- Поиск «пропущенных» генов
- Исследование транспортеров
- Полногеномный анализ

“Om's” иерархия живых систем



Определения

- Геном — совокупность всей наследственной информации: не только гены, но и пространство между ними (вся молекула ДНК)
- Экзом — совокупность всех генов, функционально-значимая часть генома; ~1.3% от его общего объема
- Транскриптом — полный набор транскриптов (молекул РНК) в данный период времени при данных условиях

Определения

- Протеом — совокупность экспрессированных белков в данный период времени при данных условиях
- Метаболом — комплекс всех низкомолекулярных метаболитов с массой < 1500 Da
- Феном — сумма фенотипических черт организма

Определения

- Кодон — тройка нуклеотидов в ДНК или РНК, обычно кодирующая одну аминокислоту
- Старт-кодон — кодон мРНК, сигнализирующий о начале синтеза белка
- Стоп-кодон — кодон мРНК, сигнализирующий об окончании синтеза белка

Определения

- Ген — участок хромосомной ДНК, кодирующий белок или функциональную РНК
- Экзон — белок-кодирующая часть гена
- Инtron — не кодирующая часть гена
- Транскрипт — молекула РНК, образуется в результате транскрипции
- Секвенирование — метод определения нуклеотидной последовательности молекулы ДНК или РНК

Размеры

- Геном — ~3.2 млрд. пар нуклеотидов, из которых ~8% не расшифровано
- Экзом — ~25 тыс. генов, ~1.3% генома
- Транскриптом — ~200 тыс. мРНК
- Протеом — >2 млн. белков
- Метаболом — ~2500 метаболитов и ~3500 пищевых компонентов

Human Genome Project (HGP)



National Human Genome
Research Institute

Begin your search here

1990 — 2003

About Genomics

Research Funding

Research at NHGRI

Health

Careers & Training

News & Events

About NHGRI

Home / About Genomics / The Human Genome Project

The Human Genome Project



The Human Genome Project (HGP) was one of the great feats of exploration in history. Rather than an outward exploration of the planet or the cosmos, the HGP was an inward voyage of discovery led by an international team of researchers looking to sequence and map all of the genes -- together known as the genome -- of members of our species, *Homo sapiens*. Beginning on October 1, 1990 and completed in April 2003, the HGP gave us the ability, for the first time, to read nature's complete genetic blueprint for building a human being.

<https://www.genome.gov/human-genome-project>

1000 Genomes Project

IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data

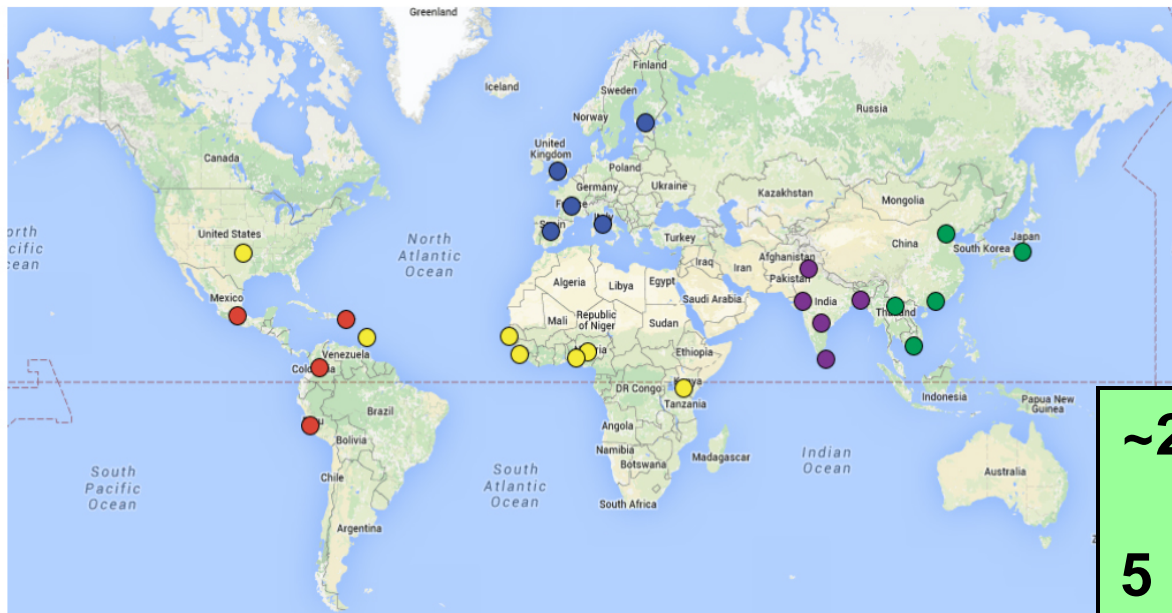
2008 — 2012

[Home](#) [About](#) [Data](#) [Portal](#) [Analysis](#) [Contact](#) [Browser](#) [FAQ](#)

Search 1000genomes



IGSR and the 1000 Genomes Project



Links

- [Announcements](#)
- [IGSR Sample Collection Principles](#)
- [1000 Genomes Project Publications](#)
- [File formats](#)
- [Software tools](#)
- [Download data](#)
- [Twitter](#)

**~250 тыс. геномов
26 популяций,
5 суперпопуляций**

<http://www.internationalgenome.org/>

Референсный геном

- Абстрактный геном, собранный по результатам секвенирования геномов большого числа людей и максимально проаннотированный.
- Представляет из себя искусственную конструкцию, реально не существующую.
- Выполняет роль «стандарта» при анализе геномов всех людей: сборке генома по результатам секвенирования, анализе мутаций и т.д.

Референсный геном человека

NCBI Resources How To Sign in to NCBI

Assembly Assembly Search Help

Advanced Browse by organism

Full Report Send to: Download Assembly

GRCh38.p13

Description: Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13)
Organism name: [Homo sapiens \(human\)](#)
BioProject: [PRJNA31257](#)
Submitter: Genome Reference Consortium
Date: 2019/02/28
Assembly type: haploid-with-alt-loci
Release type: patch
Assembly level: Chromosome
Genome representation: full
RefSeq category: reference genome
GenBank assembly accession: GCA_000001405.28 (latest)
RefSeq assembly accession: GCF_000001405.39 (latest)
RefSeq assembly and GenBank assembly identical: no ([hide details](#))

- Only in GenBank: 1 unplaced scaffold (in primary assembly-unit)
- Data displayed for RefSeq version

IDs: 2334371 [UID] 8687898 [GenBank] 8765528 [RefSeq]

History ([Show revision history](#))

Comment

The DNA sequence is composed of genomic sequence, primarily finished clones that were sequenced as part of the Human Genome Project. PCR products and WGS shotgun sequence have been added where necessary to fill gaps or correct errors. All ... [more](#)

Total ungapped length	2,948,583,725
Gaps between scaffolds	349
Number of scaffolds	472
Scaffold N50	67,794,873
Scaffold L50	16
Number of contigs	998

Access the data

- RefSeq Annotation Report
- BLAST the assembly
- Full sequence report
- Statistics report
- Regions report
- FTP directory for RefSeq assembly
- FTP directory for GenBank assembly

Related Information

- BioProject
- Genome
- Nucleotide INSDC

PubMed articles for this assembly

Finishing the finished human chromosome 22 sequence.
The DNA sequence and biological annotation of human chromosome 1.

See [Genome](#) Information for [Homo sapiens](#)

There are 244 assemblies for this organism
[See more](#)

Всего доступно 244 различных сборки

https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39

Сборка GRCh38.p13 — февраль 2019

Genome Reference Consortium Human Build 38 patch release 13

CHES

CHES

Comprehensive Human Expressed Sequences

2017 —

CHES 2.2 is a comprehensive set of human genes based on nearly 10,000 RNA sequencing experiments produced by the GTEx project. It includes a total of 20,352 protein-coding genes and 18,887 lncRNA genes. Adding antisense and other RNA genes, transcripts, 266,331 re

CHES contains virtual and 2,671 lncRNAs based on a paper (see reference b

Список доминантных генов
Включает данные из
RefSeq (NCBI) и GENCODE (EMBL)

protein-coding genes
2018 Genome Biology

CHES 2.2 data

42,611 генов и 323,258 транскриптов:
20,352 генов, кодирующих белки
18,887 генов, кодирующих lncRNA
266,331 транскрипт изоформ генов, кодирующих белки

Browser.

CHES gene list

This file is a table showing all 42,611 genes in CHES release 2.2, in a tab-delimited text file with one gene per line. For each gene it provides features such as gene ID, type, gene name, source of the annotation, location(s), GFF

[chess2.2.genes](#)

<http://ccb.jhu.edu/chess/>

Human Proteome Project (HPP)



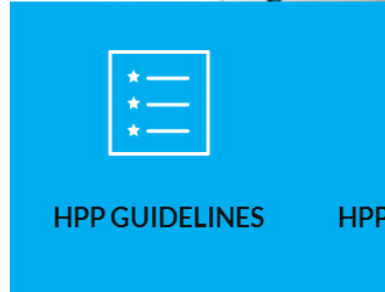
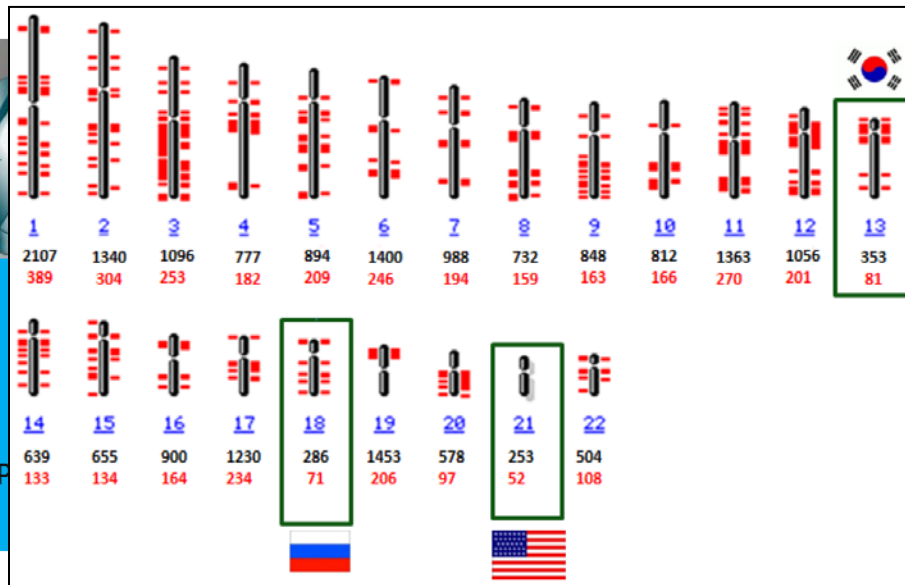
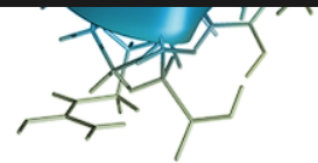
translating
the code of life



2010 —

Home > The Human Proteome Project (HPP)

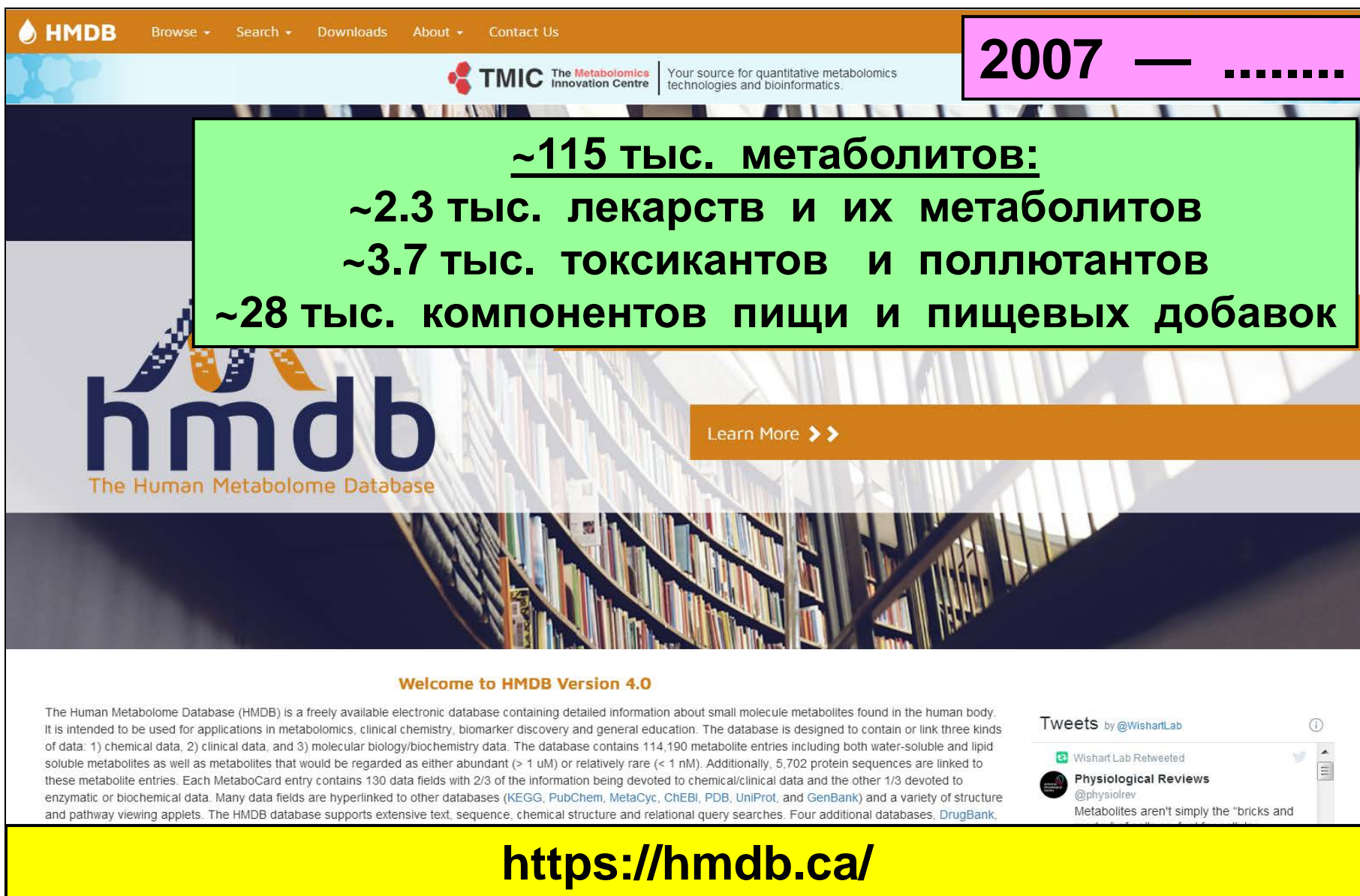
Human Proteome Project



- About the HPP
- Leadership
- C-HPP
- B/D-HPP
- HPP Resource Pillars ▶
- News
- Resources
- HPP Q & A

<https://hupo.org/human-proteome-project>

Human Metabolome Database (HMDB)



2007 —

~115 тыс. метаболитов:
~2.3 тыс. лекарств и их метаболитов
~3.7 тыс. токсикантов и поллютантов
~28 тыс. компонентов пищи и пищевых добавок

hmdb
The Human Metabolome Database

Learn More >>

Welcome to HMDB Version 4.0

The Human Metabolome Database (HMDB) is a freely available electronic database containing detailed information about small molecule metabolites found in the human body. It is intended to be used for applications in metabolomics, clinical chemistry, biomarker discovery and general education. The database is designed to contain or link three kinds of data: 1) chemical data, 2) clinical data, and 3) molecular biology/biochemistry data. The database contains 114,190 metabolite entries including both water-soluble and lipid soluble metabolites as well as metabolites that would be regarded as either abundant (> 1 uM) or relatively rare (< 1 nM). Additionally, 5,702 protein sequences are linked to these metabolite entries. Each MetaboCard entry contains 130 data fields with 2/3 of the information being devoted to chemical/clinical data and the other 1/3 devoted to enzymatic or biochemical data. Many data fields are hyperlinked to other databases (KEGG, PubChem, MetaCyc, ChEBI, PDB, UniProt, and GenBank) and a variety of structure and pathway viewing applets. The HMDB database supports extensive text, sequence, chemical structure and relational query searches. Four additional databases, DrugBank,

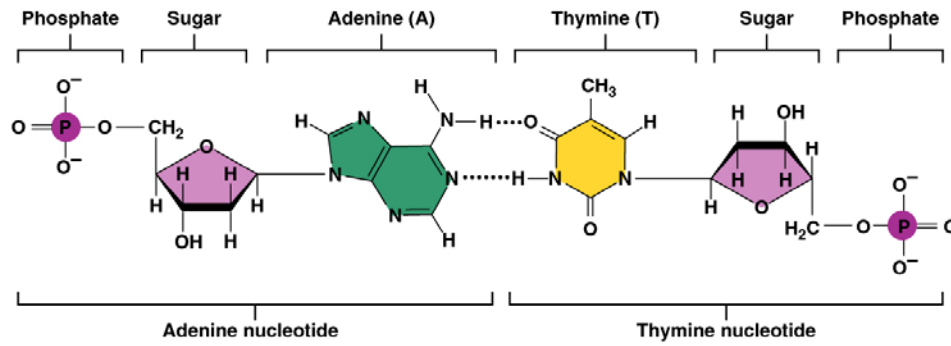
Tweets by @WishartLab

Wishart Lab Retweeted

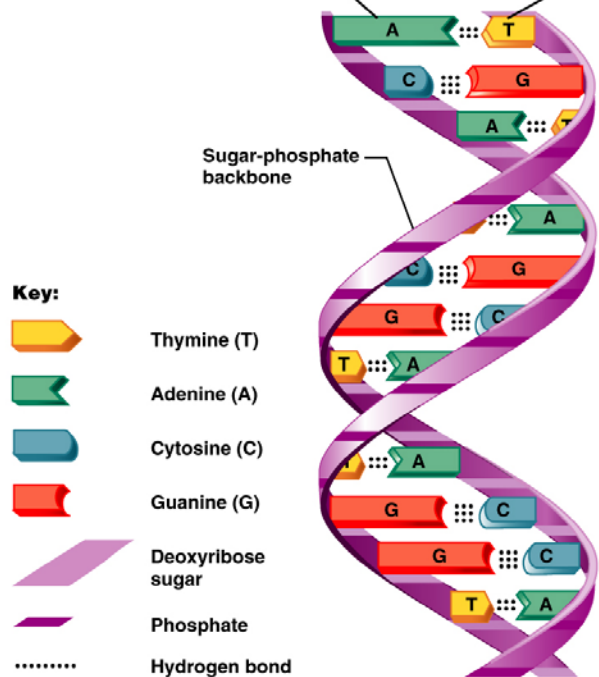
Physiological Reviews
@physiolrev
Metabolites aren't simply the "bricks and

<https://hmdb.ca/>








Строение ДНК



(a)

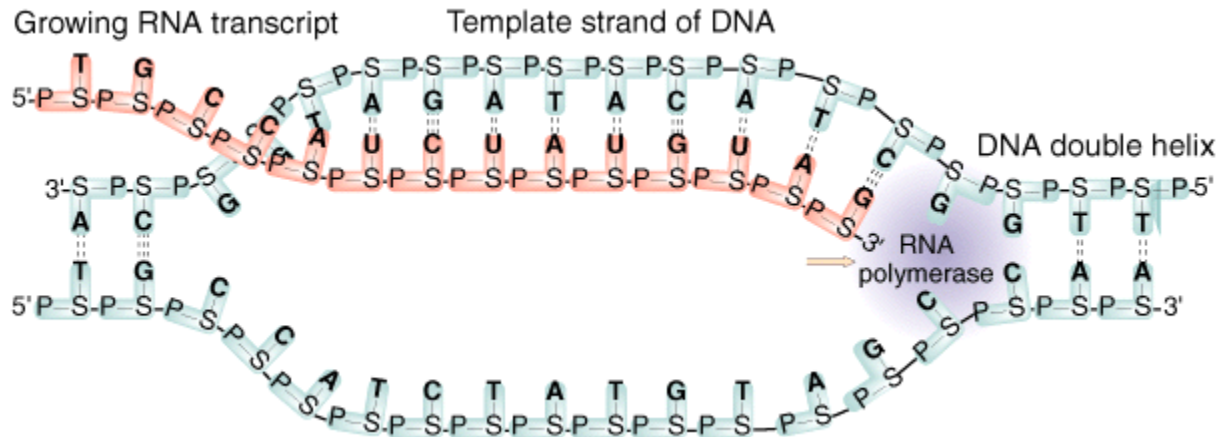


Key:

-  Thymine (T)
-  Adenine (A)
-  Cytosine (C)
-  Guanine (G)
-  Deoxyribose sugar
-  Phosphate
-  Hydrogen bond

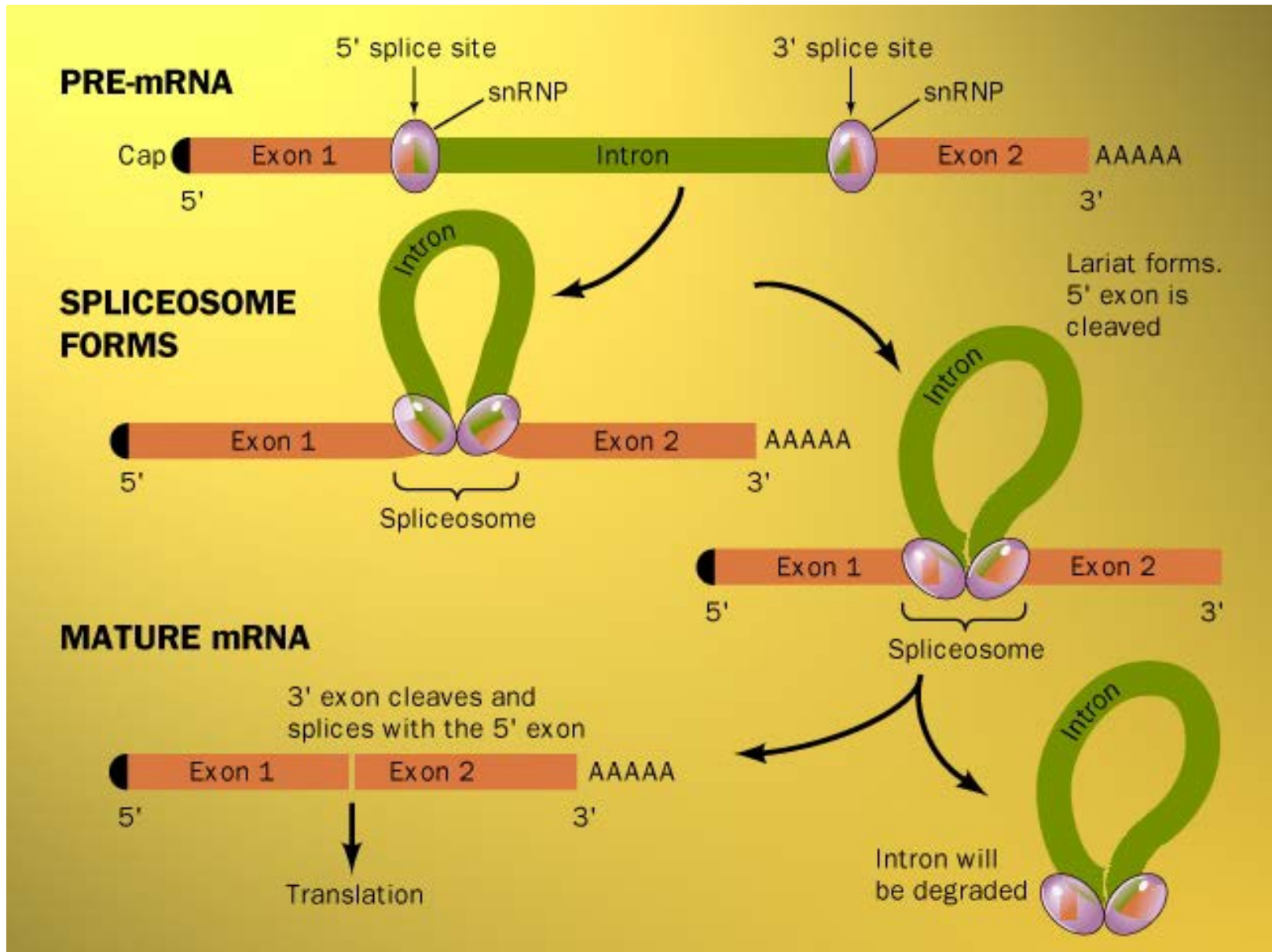
(b)

Транскрипция

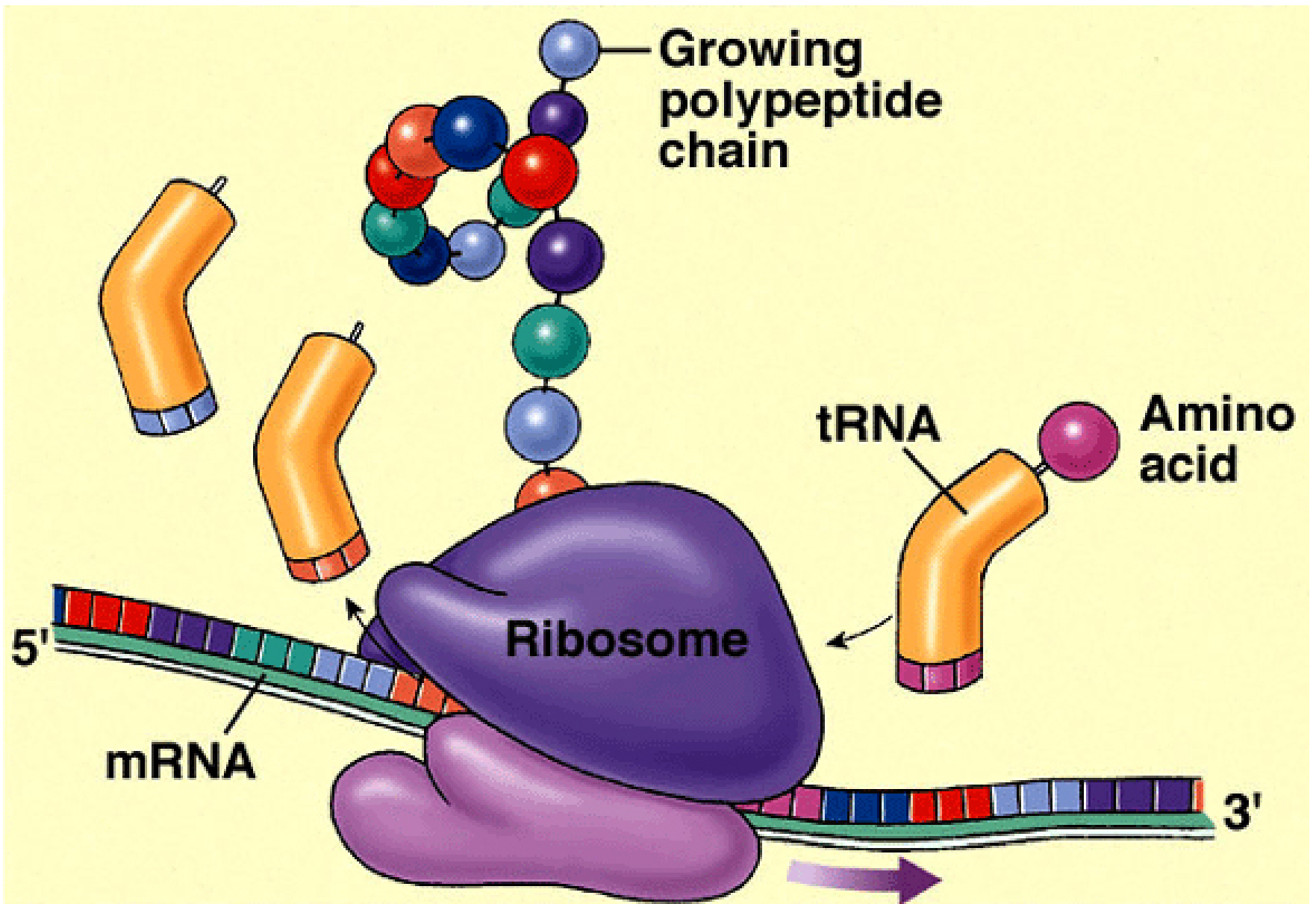


- РНК-продукт не остается комплементарно связанным с ДНК-матрицей
- По окончании синтеза двойная спираль ДНК восстанавливается
- Одноцепочечная РНК высвобождается
- синтезированные РНК – копии ограниченных участков ДНК

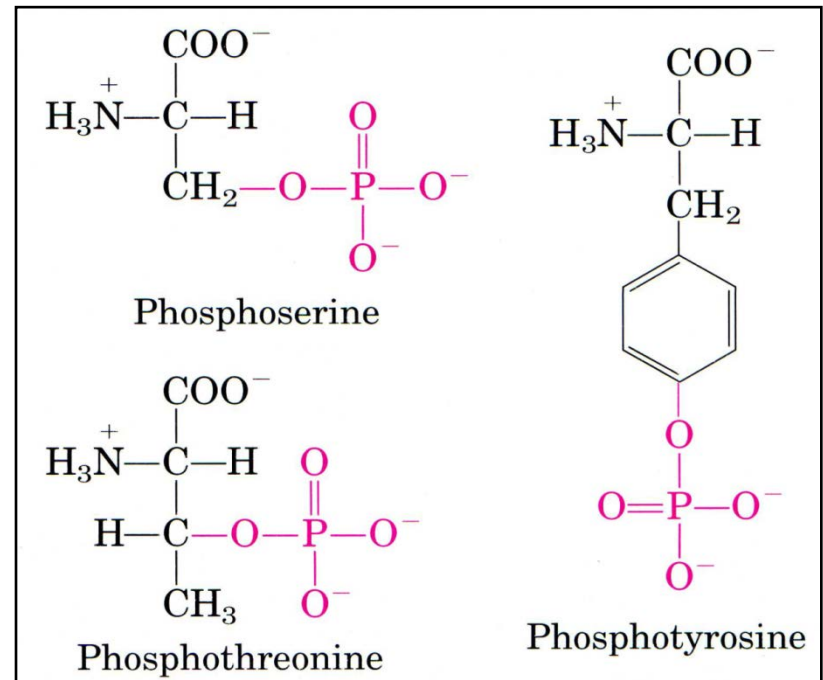
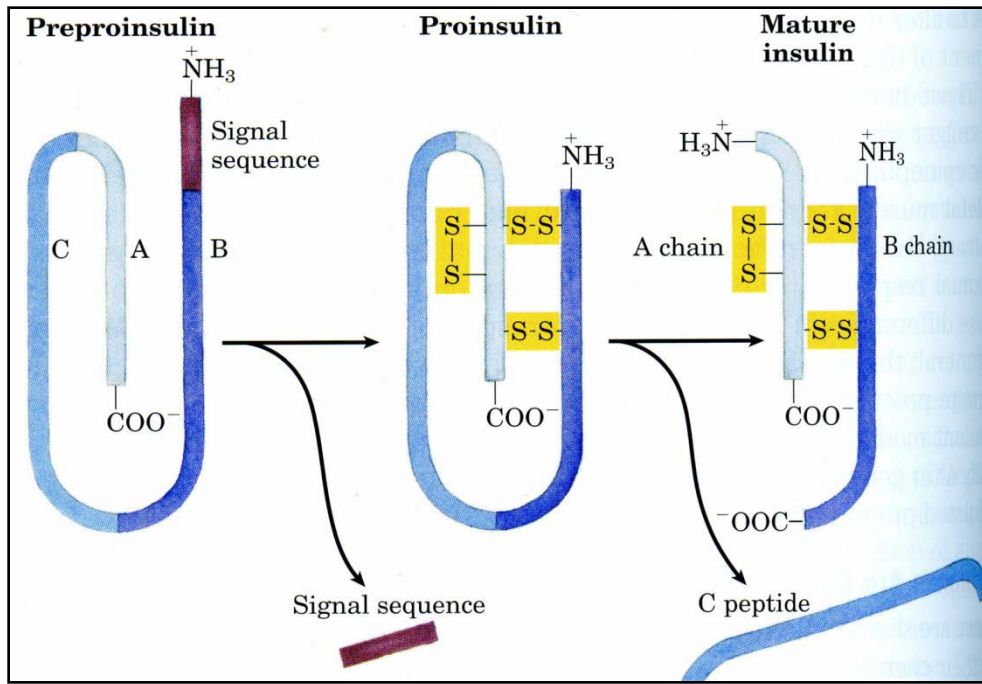
Сплайсинг



Трансляция



Процессинг



Фолдинг

Ограниченный протеолиз

Добавление сульфидных связей

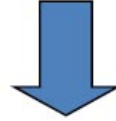
Ковалентные модификации:

- фосфорилирование
- ацетилирование
- метилирование
- гликозилирование

Иерархия Omic's технологий

Genomics (DNA- 25,000 genes)

Геномика



Transcriptomics (RNA - 100,000 mRNA's)

Транскриптомика

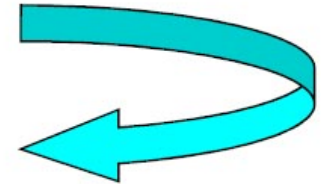
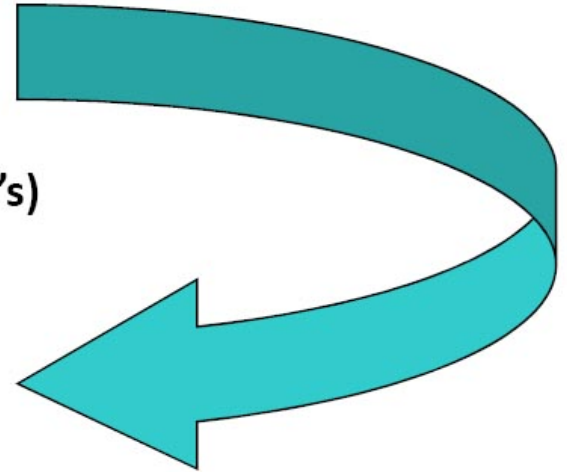
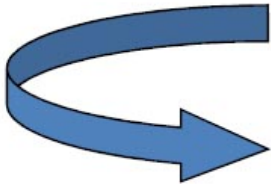
микроРНК (2000)

Proteomics (Proteins - 1,000,000 proteins)

Протеомика

Metabolomics (molecules)

Метаболомика



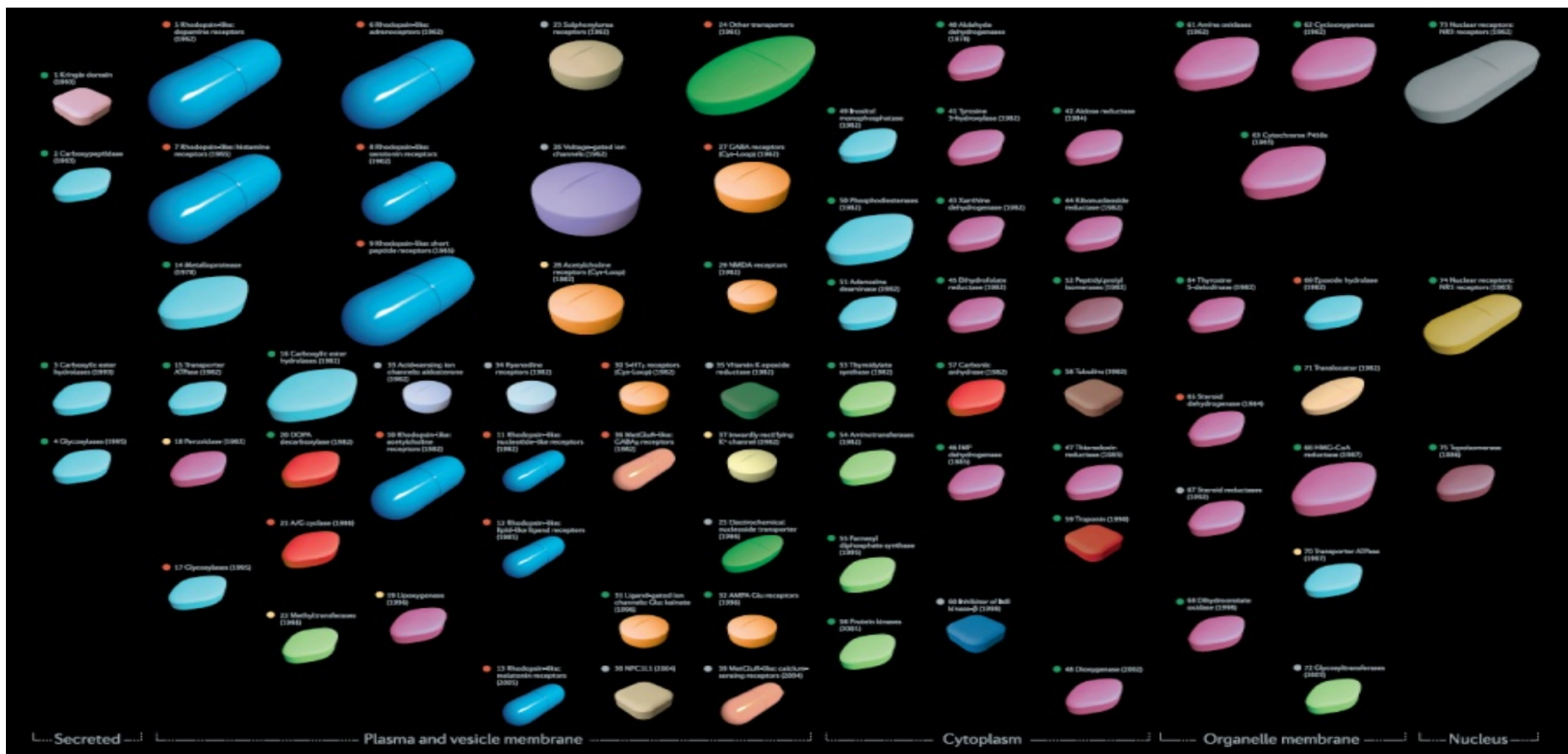
Omic's экспериментальные методы

- **Геномика**
 - **Высокоскоростное секвенирование ДНК (HT-DNA sequencing)**
- **Обнаружение мутаций**
 - **Метод однонуклеотидного полиморфизма (SNP)**
- **Транскриптомика**
 - **Измерение генной транскрипции (Gene/Transcript measurement)**
 - **Серийный анализ экспрессии генов (SAGE)**
 - **Генные чипы**
 - **Микрочипы (Microarrays)**
- **Протеомика**
 - **Масс-спектрометрия (MS)**
 - **2D-гель электрофорез (2D-PAGE)**
 - **Белковые чипы**
 - **Двугибридный анализ (Yeast-2-hybrid)**
 - **Рентгеноструктурный анализ (X-ray)**
 - **Ядерный магнитный резонанс (NMR)**
- **Метаболомика**
 - **Ядерный магнитный резонанс**
 - **Рентгеноструктурный анализ**
 - **Капиллярный электрофорез**

Протеомные проекты

- Проект «Протеом человека».
Россия исследует белки 18 хромосомы.
- Проект «Протеом плазмы крови».
Идентифицировано более 13 тыс. белков.
- Протеом плазмы крови включает 10 %
всех белков человека.
- Найдено более 1000 новых белков-
мишеней для создания лекарств.

Биомишени для создания лекарств



	GPCR		Ионные каналы
	Ферменты		Ядерные рецепторы
	Транспортеры		Прочие

Размер значка

Маленький – 1 ÷ 10 лекарств

Средний – 11 ÷ 50 лекарств

Большой – > 50 лекарств

Достоверность 3D-модели

- Высокая
- Умеренная
- Низкая
- Нулевая

Секвенирование

- Метод определения нуклеотидной последовательности молекулы ДНК
- ДНК многократно копируются и нарезаются на короткие отрезки
- Позволяет обнаружить мутации и полиморфизмы любого масштаба: от крупных хромосомных перестроек до однонуклеотидных замен

Задачи секвенирования

- Расшифровка неизвестных ДНК (секвенирование "de novo")
- Выявление индивидуальных отличий конкретного образца от референсной последовательности (ресеквенирование)
- Анализ генетических полиморфизмов, включая однонуклеотидные (SNP-типирование)
- Анализ эпигенетических модификаций ДНК (например, профиля метилирования)
- Анализ профиля экспрессии секвенированием кДНК, полученной из тотальной мРНК (диагностика вирусных и раковых заболеваний)

Стоимость секвенирования

Cost per Genome

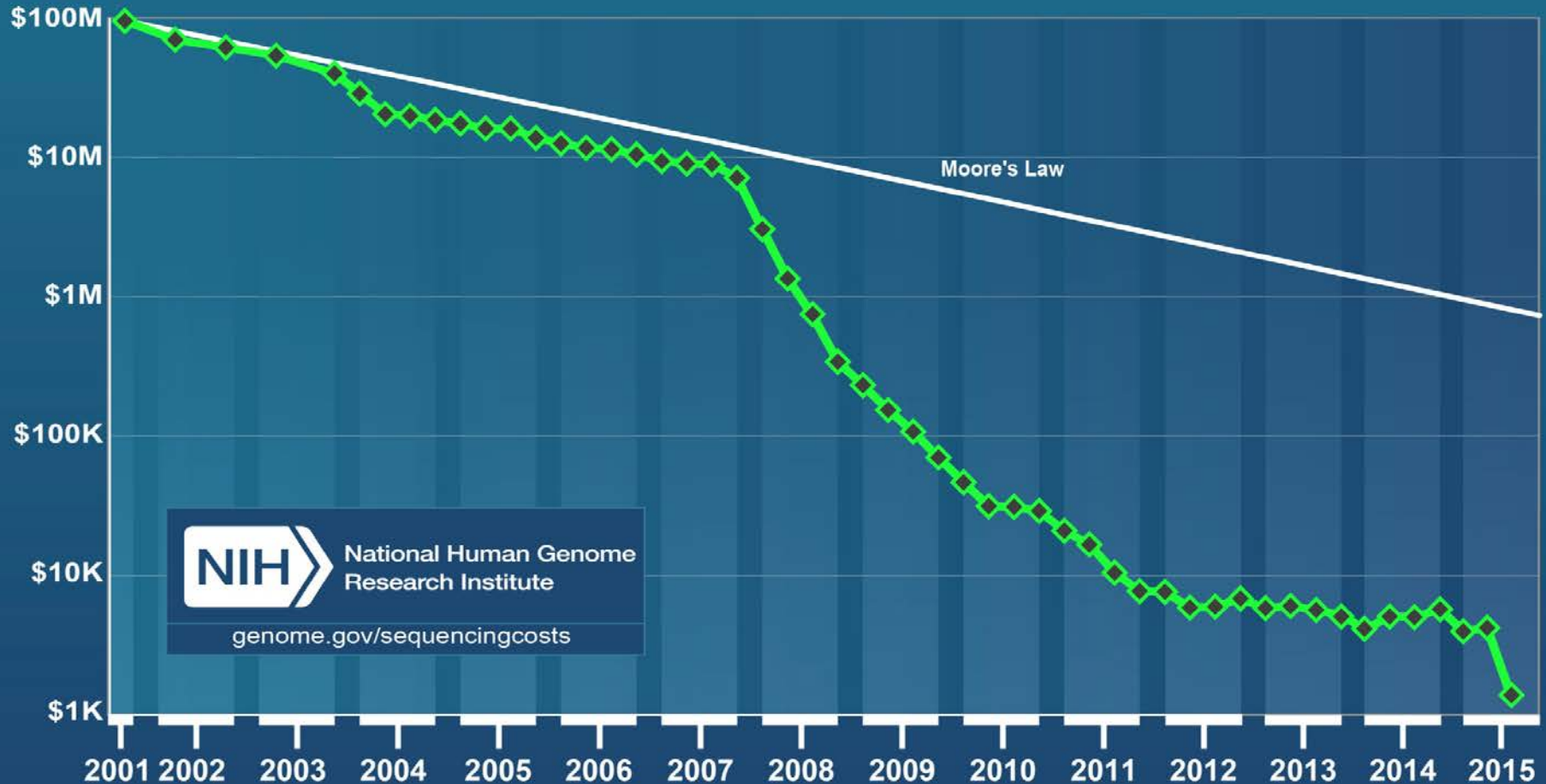


Схема секвенирования

Human Genome Sequencing

Generating a Reference Genome Sequence
(e.g., Human Genome Project)



Genomic DNA

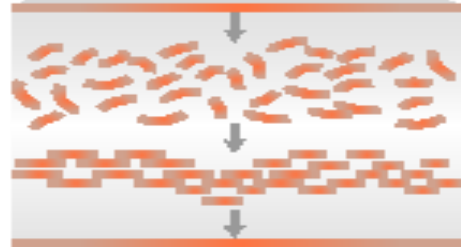
Break genome into large fragments and insert into clones



Order clones



Break individual clones into small pieces



Generate thousands of sequence reads and assemble sequence of clone

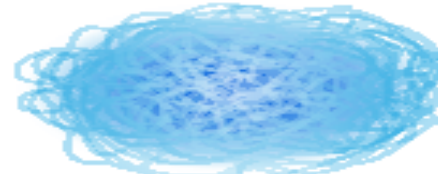


Assemble sequences of overlapping clones to establish reference sequence



Reference Sequence

Generating a Person's Genome Sequence
(e.g., Circa ~2016)



Genomic DNA

Break genome into small pieces



Generate millions of sequence reads

... TATGCGATGCGTATTCGTA ...

Align sequence reads to established reference sequence



Reference Sequence

Deduce starting sequence and identify differences from reference sequence



Амплификационные технологии

Полимеразная цепная реакция

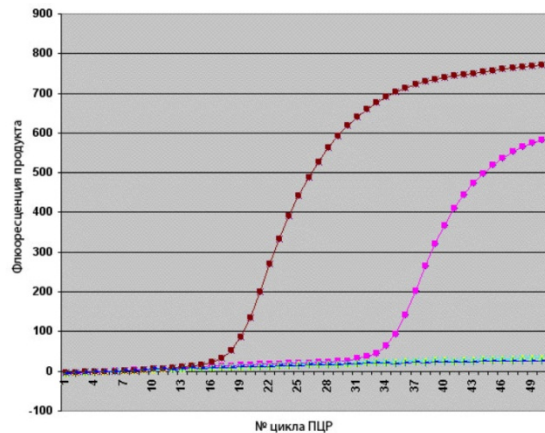
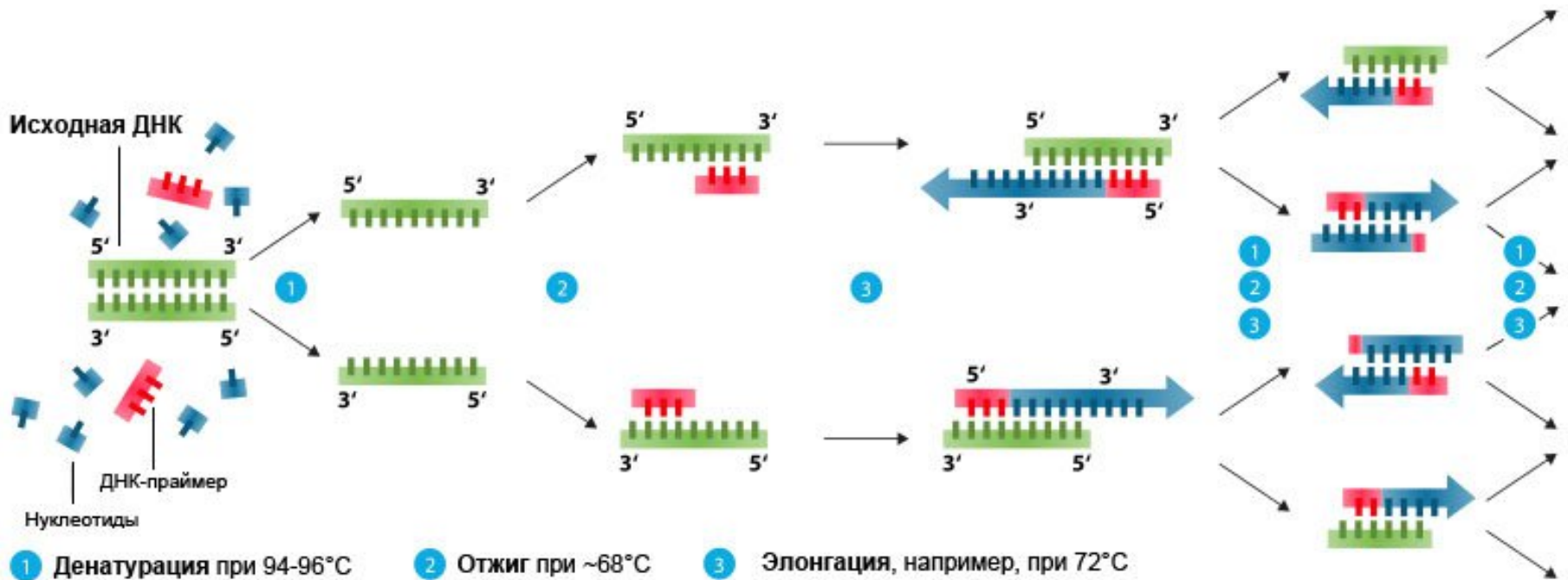
Кэри Муллис 1983 г.

Нобелевская премия 1993 г.

Компоненты ПЦР

- ДНК-матрица
- Праймеры – олигонуклеотиды длиной 18-30 пн, комплементарные противоположным концам разных цепей ДНК-матрицы
- дНТФ – дезоксирибонуклеозидтрифосфаты (dATP, dGTP, dCTP, dTTP)
- ДНК-полимераза (Taq, Tth)

Схема ПЦР



Виды ПЦР

- Количественная ПЦР (ПЦР в реальном времени, real-time PCR)
- ПЦР in situ – ПЦР непосредственно в клетках, результаты видны под микроскопом
- ПЦР с обратной транскрипцией – используется при секвенировании РНК
- Мультипраймерная ПЦР – одновременная амплификация двух и более ДНК

Технологии секвенирования

- 1-е поколение



Sanger sequencing

- 2-е поколение

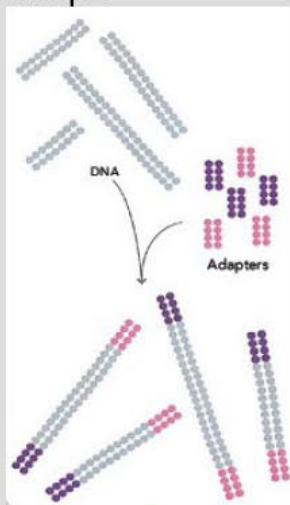


- 3-е поколение

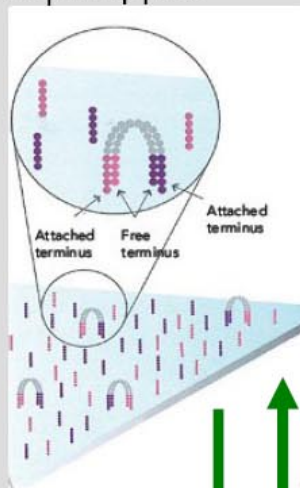


Секвенирование Illumina - принцип метода

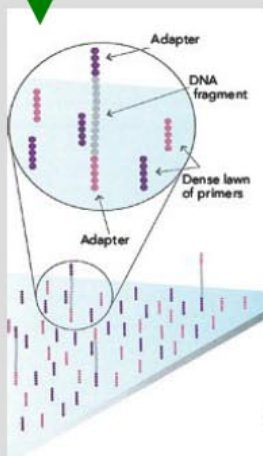
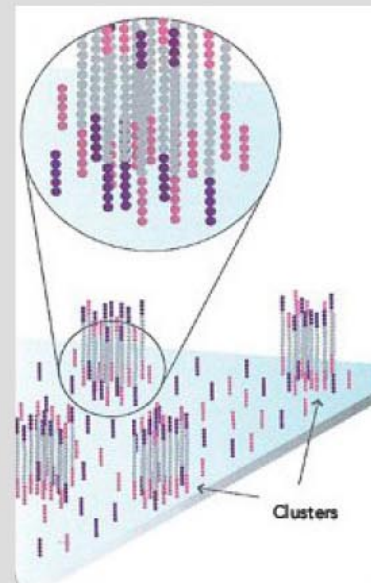
1. ДНК фрагментируют и лигируют к фрагментам адаптеры



3. Через ячейку пропускают реагенты для достраивания второй цепи ДНК



Стадии 3-4 повторяются 30-35 раз

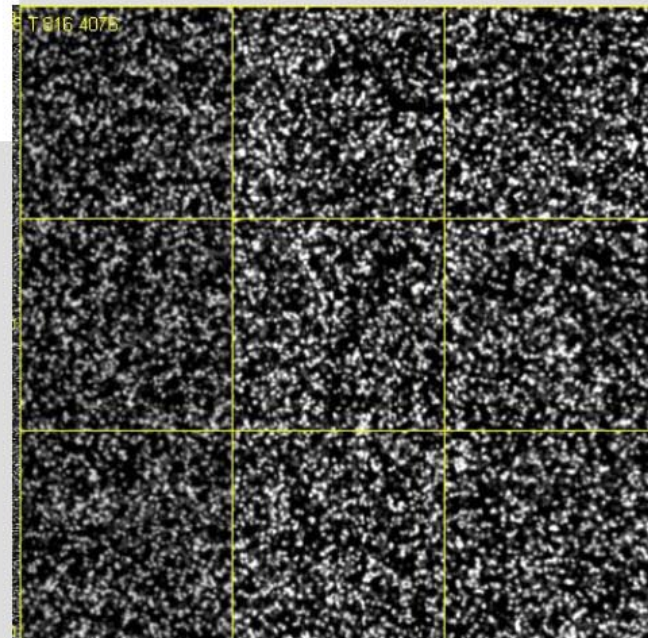
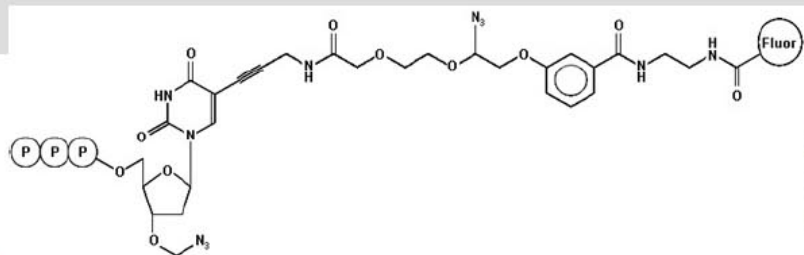
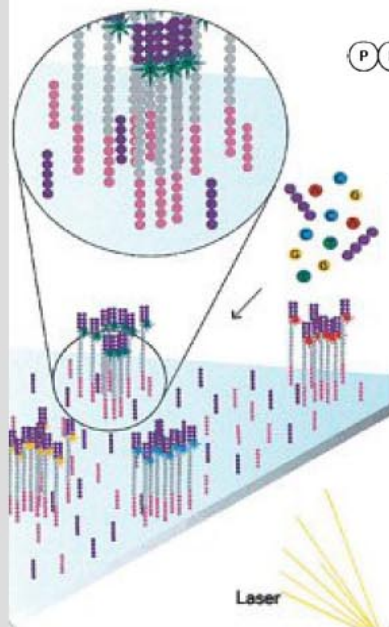


2. ДНК пропускают через каналы ячейки, покрытые праймерами, комплементарными концам адаптеров

4. Двухцепочечные фрагменты денатурируют

6. Каждый фрагмент оказывается окружен группой идентичных молекул («кластеры»).

Секвенирование Illumina - принцип метода



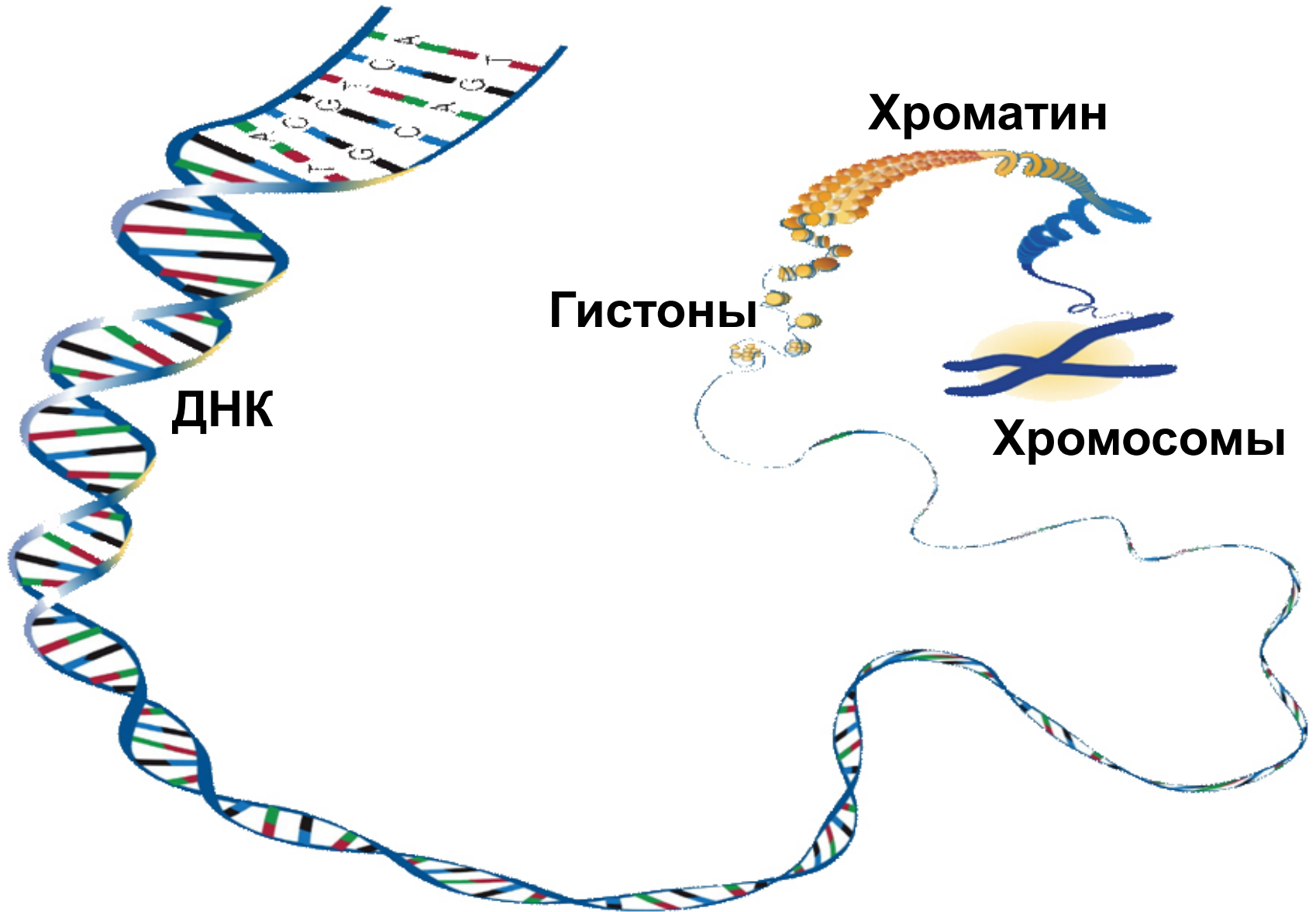
7. Через ячейку пропускают реагенты (флуоресцентно меченые терминированные dNTP и полимеразу)

10. Повторение 7-9
нужное число раз (50-
300). Число циклов
соответствует длине
чтения.

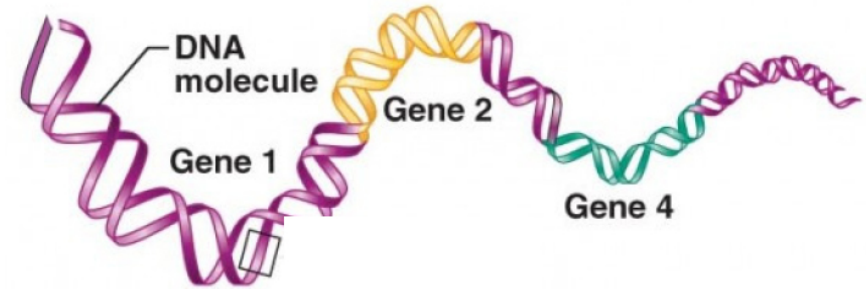
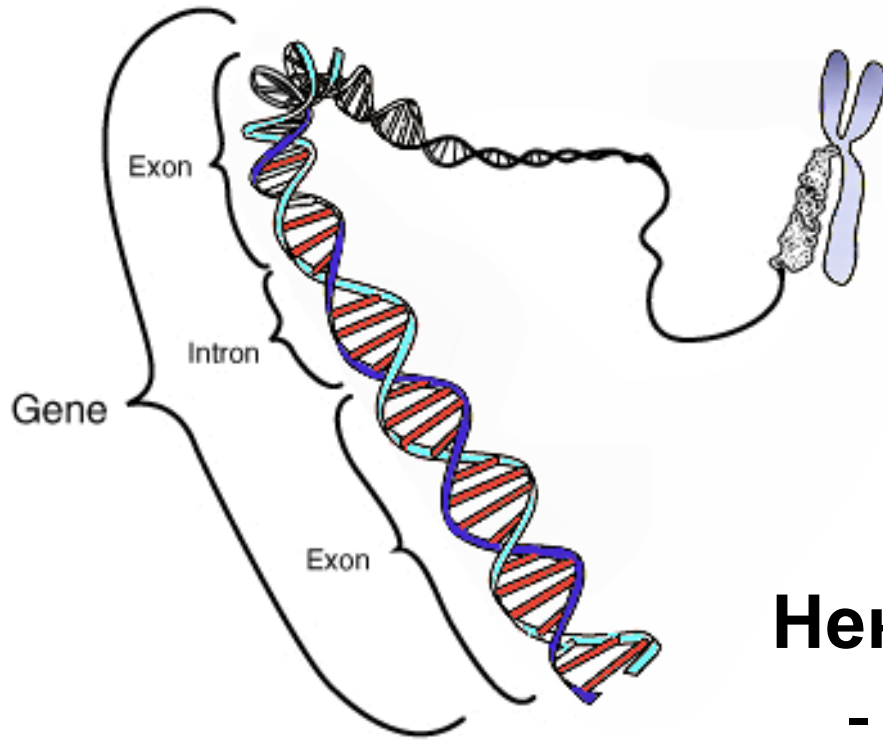
8. На ячейку светят лазером
и проводят съемку.

9. Через ячейку пропускают
реагенты, отщепляющие
флуорофор и терминатор

Структура генома



Структура генома



Некодирующая часть:

- области РНК
- регуляторные регионы
- ретро-транспозоны
- всездо-гены
- и др.

Факторы заболеваний

Deciphering Complex Disease

A systems biology approach using the SOLiD™ System

Detect and Sequence Insertions and Deletions

Discover and Quantitate Non-Coding RNA

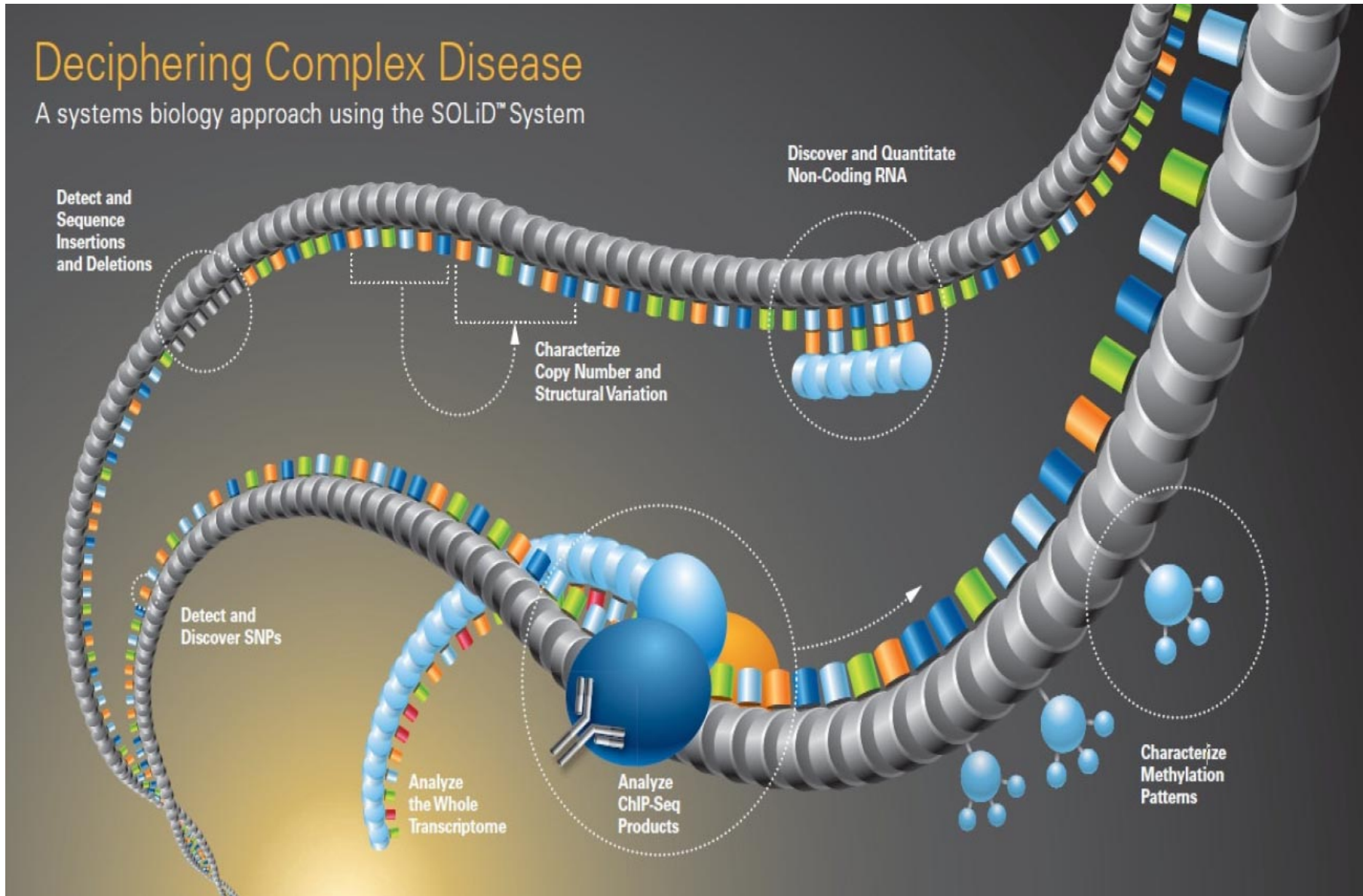
Characterize Copy Number and Structural Variation

Detect and Discover SNPs

Analyze the Whole Transcriptome

Analyze ChIP-Seq Products

Characterize Methylation Patterns



Мутации

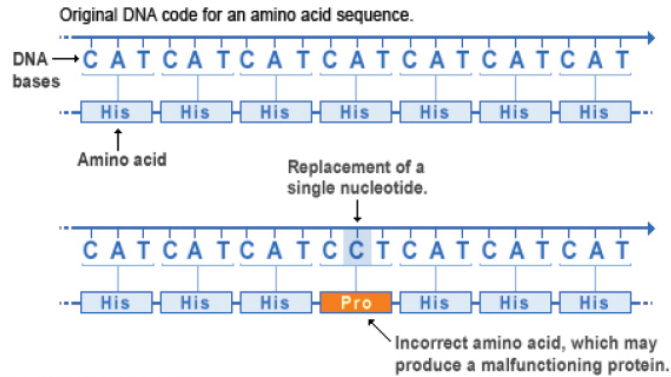
Наследуемые изменения ДНК, различаются:

- по размеру (точечные, хромосомные, геномные)
- по влиянию (нейтральные, негативные, позитивные)
- по способу возникновения (случайные, индуцированные)
- по возможности наследования (соматические, половые)
- по отклонению от нормы (прямые, обратные, супрессорные)

Мутации

MISSENSE MUTATION

Замены

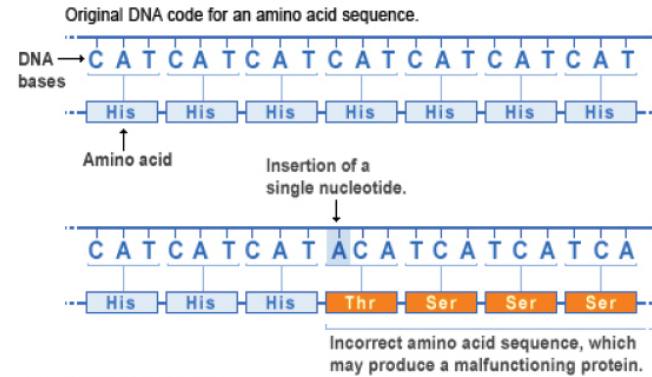


U.S. National Library of Medicine

This type of mutation is a change in one DNA base pair that results in the substitution of one amino acid for another in the protein made by a gene.

INSERTION

Вставки

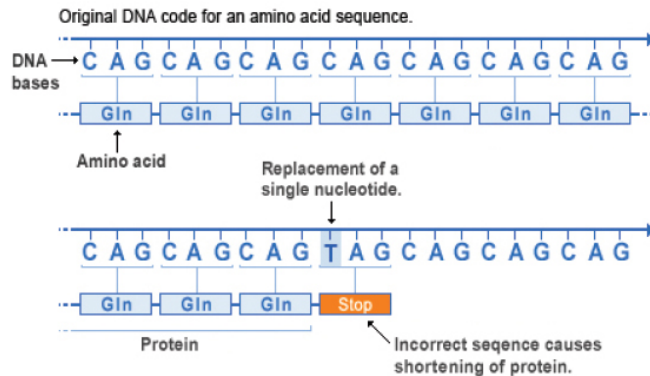


U.S. National Library of Medicine

An insertion changes the number of DNA bases in a gene by adding a piece of DNA. As a result, the protein made by the gene may not function properly.

NONSENSE MUTATION

Нонсенсы

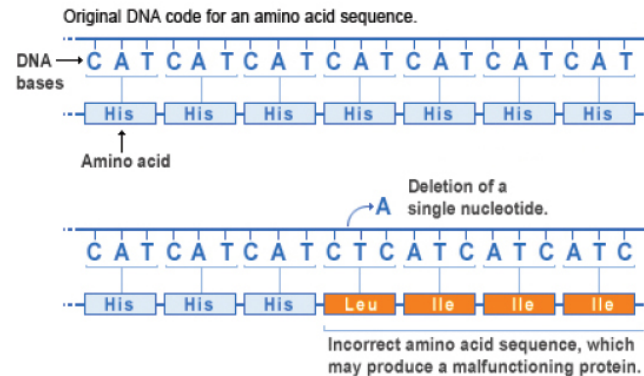


U.S. National Library of Medicine

A nonsense mutation is also a change in one DNA base pair. Instead of substituting one amino acid for another, however, the altered DNA sequence prematurely signals the cell to stop building a protein. This type of mutation results in a shortened protein that may function improperly or not at all.

DELETION

Делеции



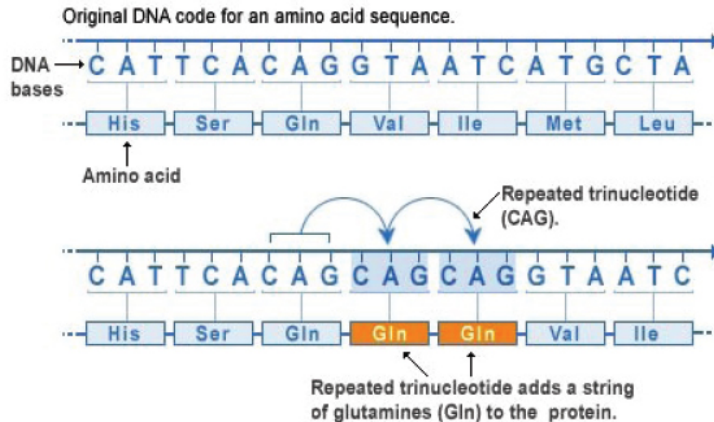
U.S. National Library of Medicine

A deletion changes the number of DNA bases by removing a piece of DNA. Small deletions may remove one or a few base pairs within a gene, while larger deletions can remove an entire gene or several neighboring genes. The deleted DNA may alter the function of the resulting protein(s).

Мутации

REPEAT EXPANSION

Повторы



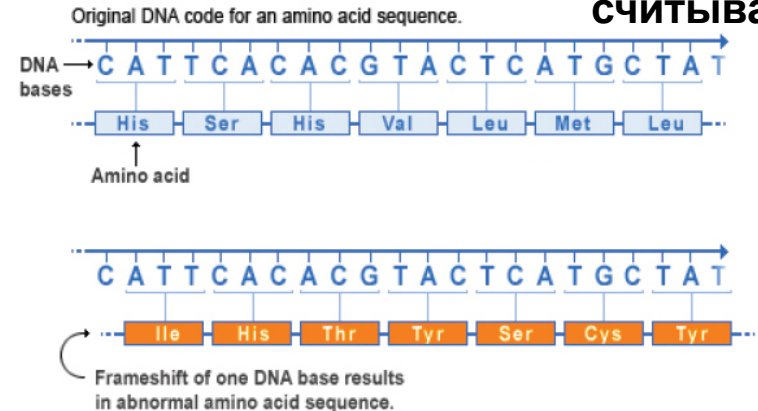
U.S. National Library of Medicine

Nucleotide repeats are short DNA sequences that are repeated a number of times in a row. For example, a trinucleotide repeat is made up of 3-base-pair sequences, and a tetranucleotide repeat is made up of 4-base-pair sequences. A repeat expansion is a mutation that increases the number of times that the short DNA sequence is repeated. This type of mutation can cause the resulting protein to function improperly.

FRAMESHIFT MUTATION

Изменение рамки

СЧИТЫВАНИЯ



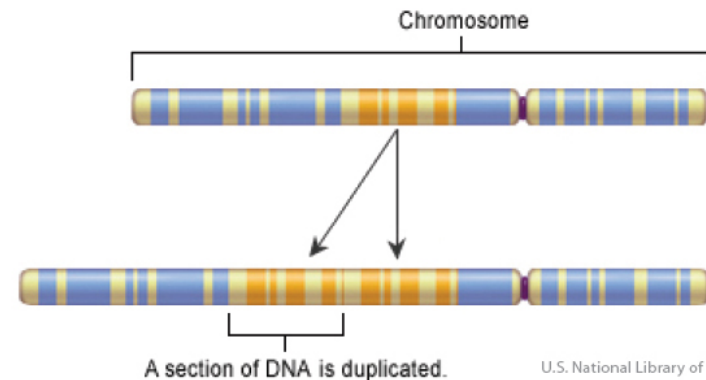
U.S. National Library of Medicine

This type of mutation occurs when the addition or loss of DNA bases changes a gene's reading frame. A reading frame consists of groups of 3 bases that each code for one amino acid. A frameshift mutation shifts the grouping of these bases and changes the code for amino acids. The resulting protein is usually nonfunctional. Insertions, deletions, and duplications can all be frameshift mutations.

DUPLICATION

Дублирование

A duplication consists of a piece of DNA that is abnormally copied one or more times. This type of mutation may alter the function of the resulting protein.



U.S. National Library of Medicine

To be continued ...

