Геномика
Лекция №22

Понятние геномики, структурная геномика, функциональная геномика, фармакогеномика, популяционная геномика, сравнительная геномика, эволюционная геномика. История геномных технологий. Структуры эукариотических и прокариотических геномов. Ортология, паралогия, синтения, COGs (clusters of orthologous groups). Геном человека. Генетические вариации. Понятие гаплотипа и гаплогруппы. Геномные проекты: геном человека, 1000 Genomes, Epigenomics Roadmap, ENCODE, 4D nucleome, TCGA, ICGC. GWAS исследования, связь генотип-фенотип. Заболевания наследуемые по Менделю и комплексные заболевания. Базы данных dbSNP, OMIM, ClinVar. Геномные браузеры. Демонстрация браузера Ensembl.

Алексей Константинович Шайтан, к.ф.-м.н.

Сайт курса: http://intbio.org/bioinf2018-2019

---

15 апреля 2019

Геномика

- **Гено́м** — совокупность наследственного материала, заключенного в клетке организма.
- **Геномика** – изучение строения, работы, функций генов и геномов.
- Геномика тесно связана с биоинформатикой и технологиями секвенирования.
- **Functional genomics** is a field of molecular biology that attempts to make use of the vast wealth of data given by genomic and transcriptomic projects (such as genome sequencing projects and RNA sequencing) to describe gene (and protein) functions and interactions.
- **Comparative genomics** is a field of biological research in which the genomic features of different organisms are compared.
- **Population genomics** is the large-scale comparison of DNA sequences of populations.
- **Metagenomics** is the study of genetic material recovered directly from environmental samples
- **Pharmacogenomics** is the study of how genes affect a person's response to drugs.
- **Structural genomics** seeks to describe the 3-dimensional structure of every protein encoded by a given genome.

Омиксные технологии

- Транскриптомика
- Протеомика
- Метаболомика
- Эпигеномика
- Липидомка
- Гликомика

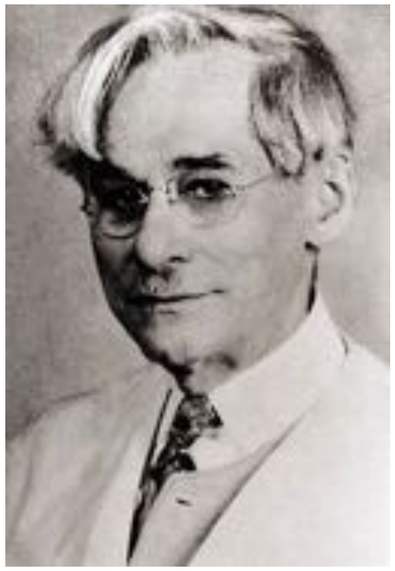
- Omics aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms.

Развитие представлений о ДНК

1869

Friedrich Miescher

~1919



Phoebus Levene

1927

Николай Константинович Кольцов

1935

Николай Владимирович Тимофеев-Ресовский

Max Delbrück
Karl Zimmer

1943

Oswald Avery

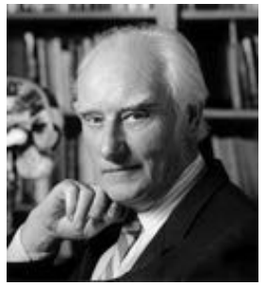
1944

Erwin Schrödinger

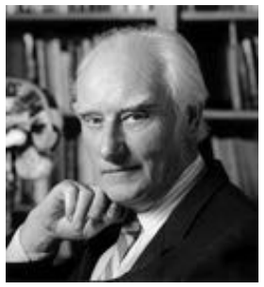
1947

Erwin Chargaff

1953



1957



1958

1966

https://www.dna-worldwide.com/resource/160/history-dna-timeline

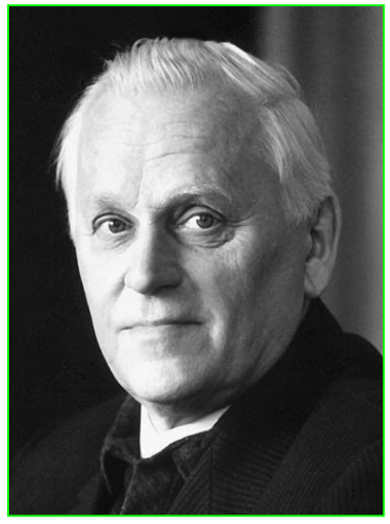
Секвенирование ДНК/РНК

alanine tRNA

1964

Robert Holley

1967



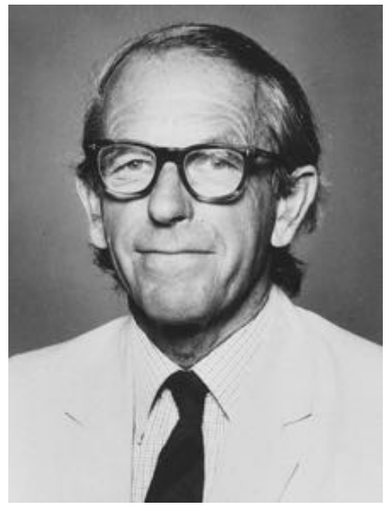
А.А. Баев (1904 - 1994)

геном бактериофага φX174 (5386bp)

1975

«plus and minus method» for DNA sequencing

1977



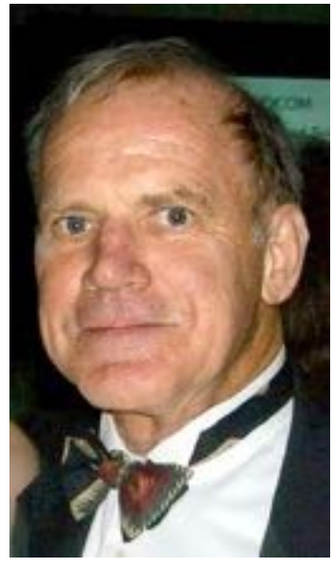
Frederick Sanger

1977

Метод Сэнгера для секвенирования ДНК

1983

PCR

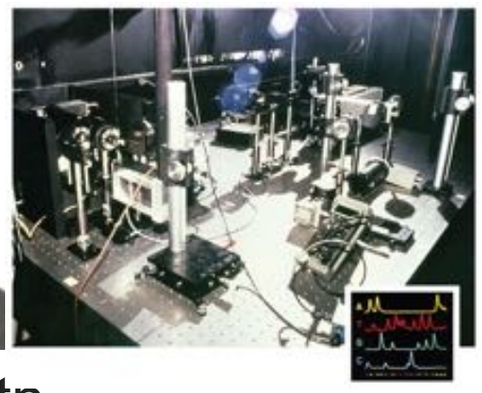


Kary Mullis

Баев А.А., Венкстерн Т.В., Мирзабеков А.Д., Крутилина А.И., Ли Л., Аксельрод В.Д. 1967. Первичная структура валиновой транспортной РНК1 пекарских дрожжей. Молекулярная биология, 1(5), 754

Прогресс в секвенировании

1986



Lloyd Smitn

Applied biosystems 370A DNA sequencer

Dye-terminator method

1988

NCBI

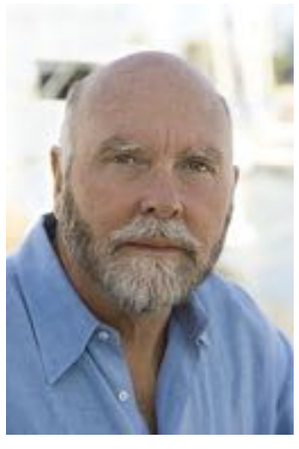
1990

Начало проекта **геном человека**.
План: 15 лет, $3 млрд
А также:
M. capricolum
E. coli
C. elegans
S. cerevisiae

Francis Collins

1995

геном *Haemophilus Influenzae* *1,830,137bp*



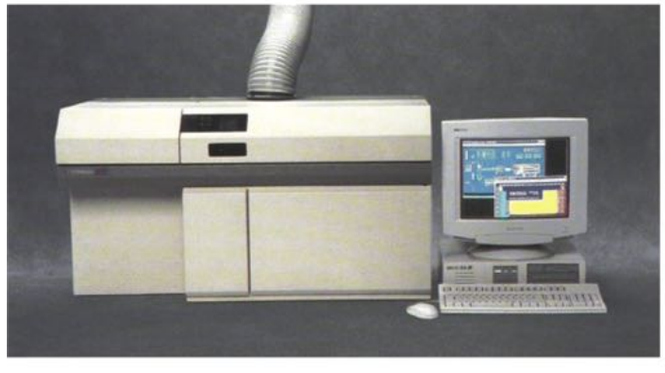
Craig Venter

1996

BERMUDA PRINCIPLES

1998

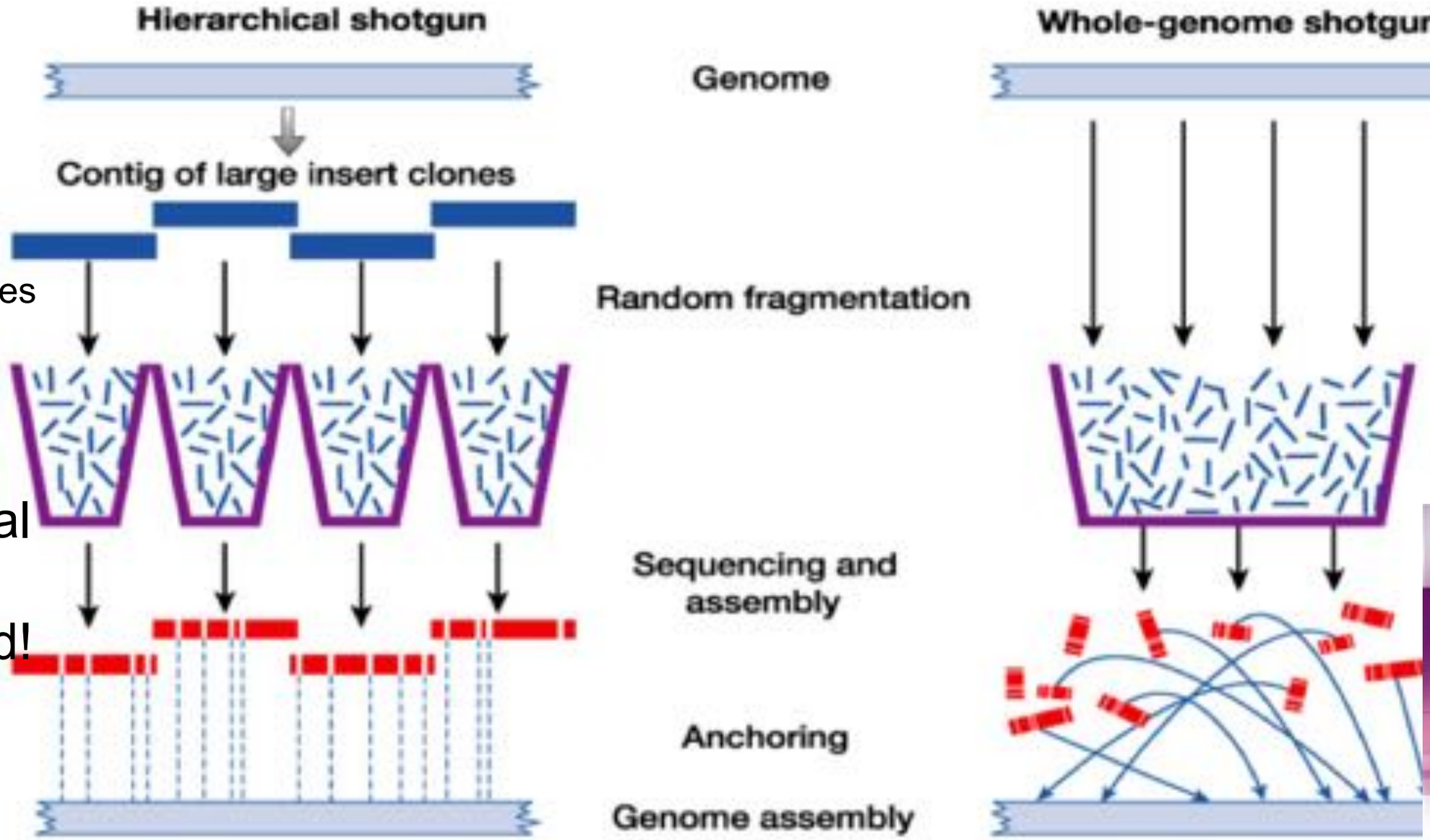
Celera Genomics enters genome race



ABI PRISM 3700
96 образцов*16 раз за день

Проект геном человека: методы и подходы

'BAC-by-BAC' approach



BACs= bacterial artificial chromosomes ~150 Kb

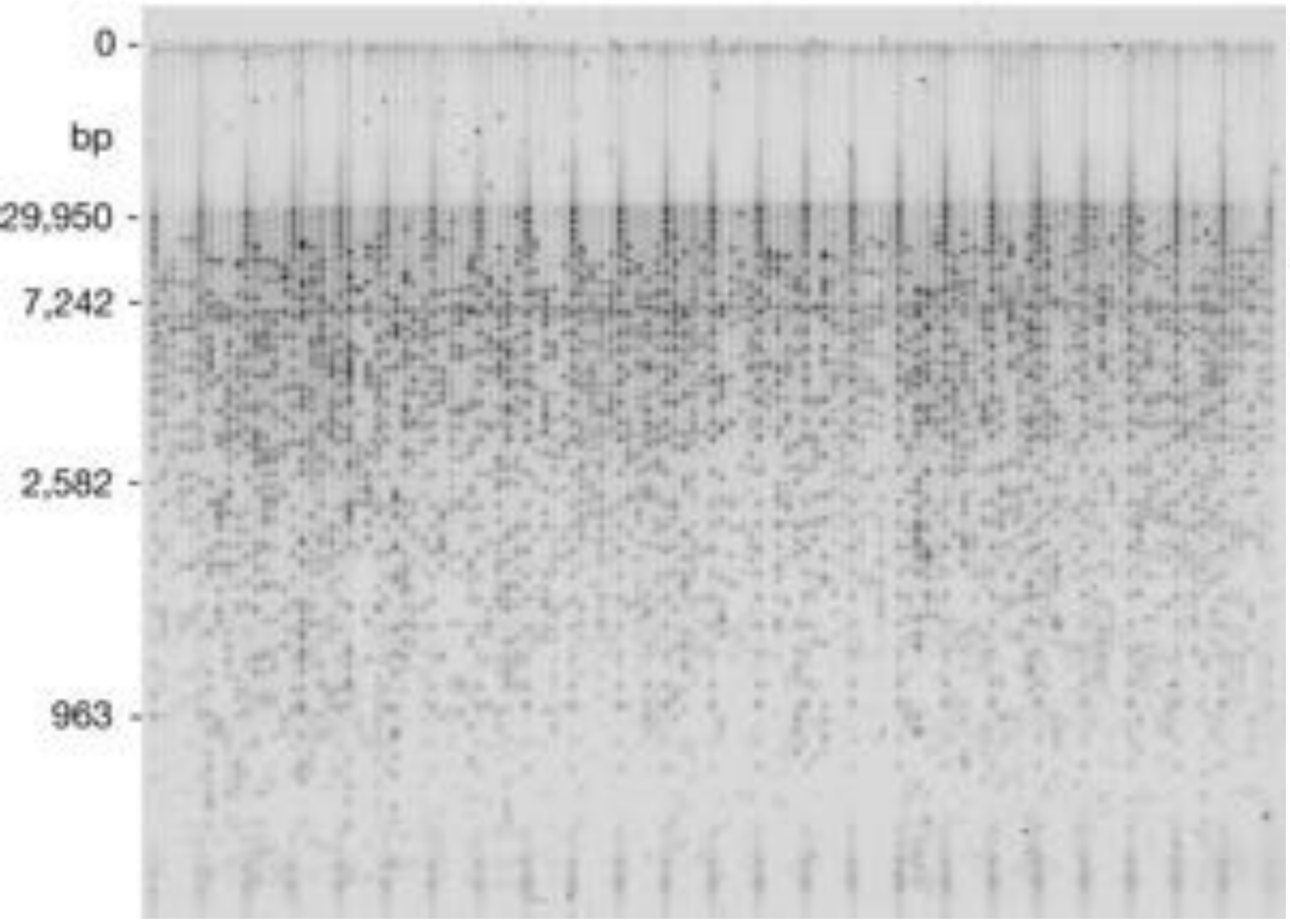
Physical map Needed!

https://www.nature.com/scitable/topicpage/dna-sequencing-technologies-key-to-the-human-828

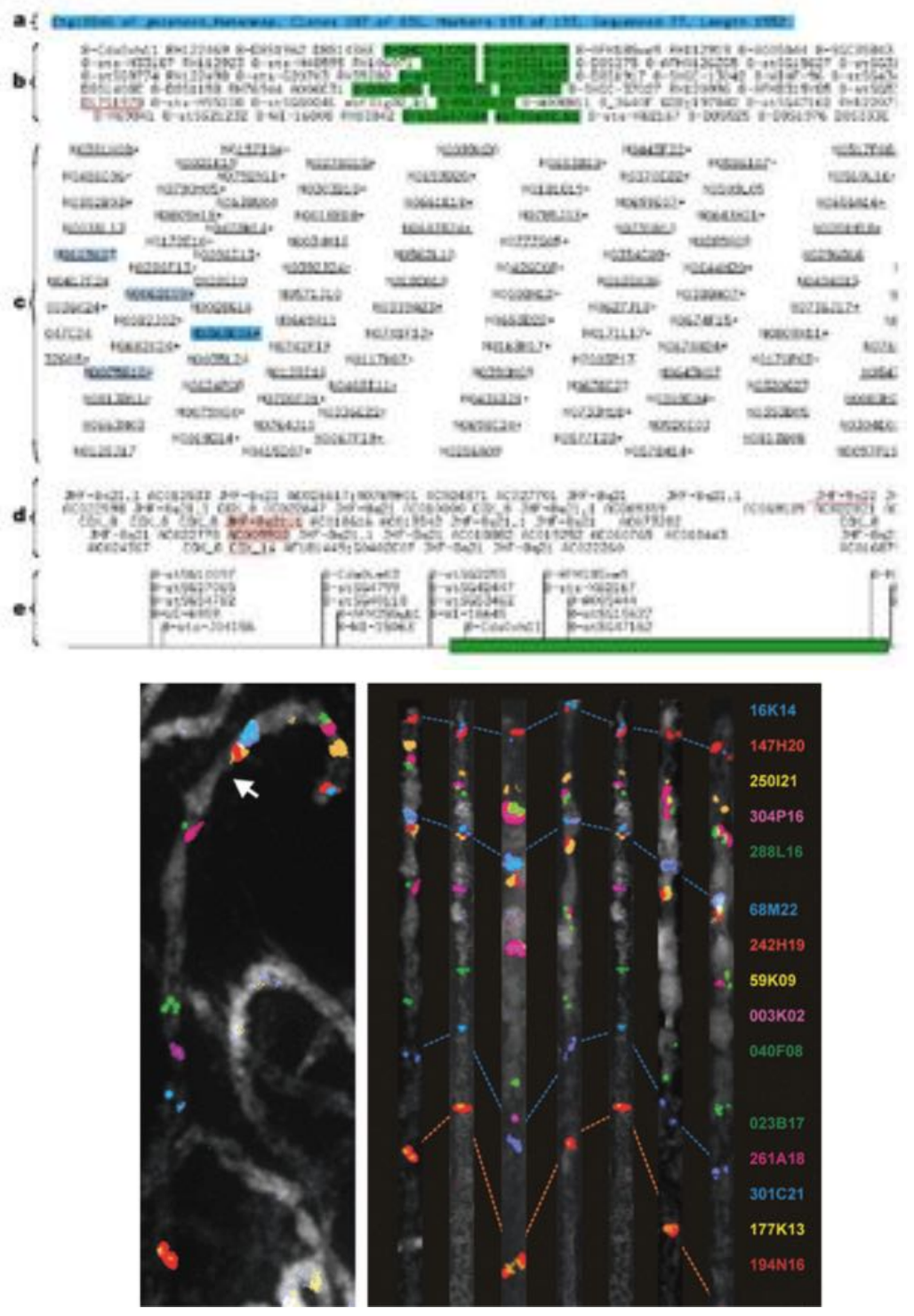
Проект геном человека: методы и подходы

Создание физической карты генома

Fingerprinting BACs



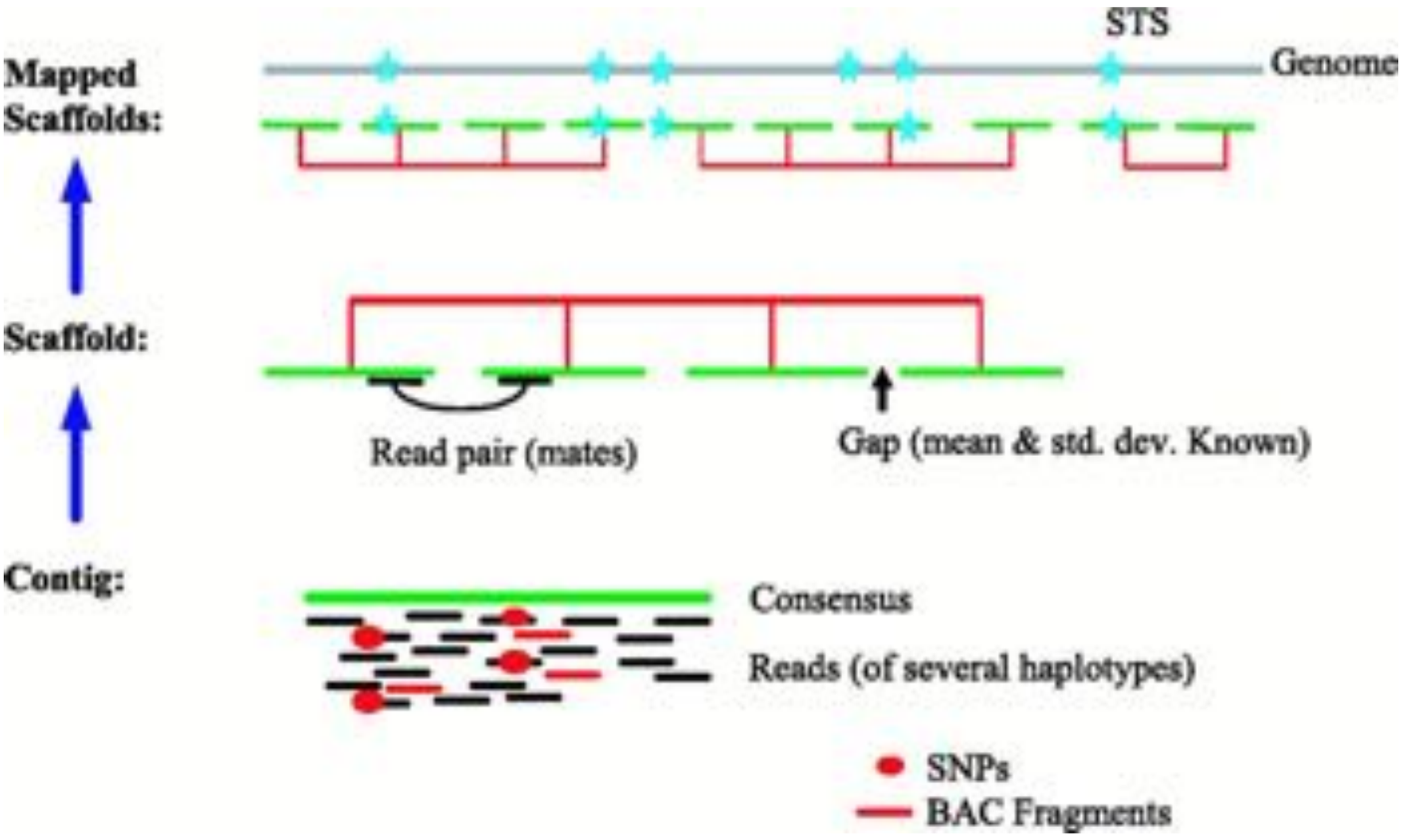
BAC DNAs are digested with *Hind*III and visualized on a SYBR-green-stained 1% agarose gel. Every fifth lane contains a mixture of marker DNAs; the sizes of selected marker fragments are indicated. 0, origin of fragment migration.



FISH

https://www.nature.com/articles/35057157

Проект геном человека: методы и подходы

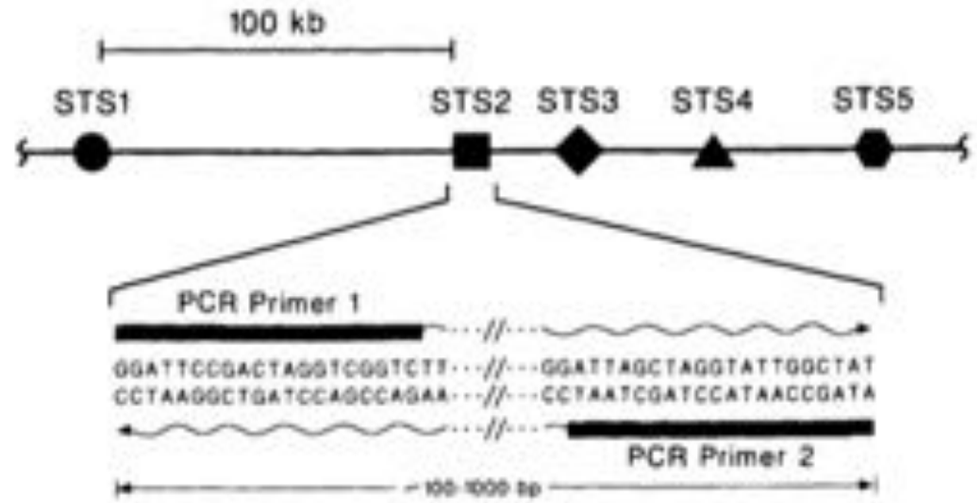


Проект геном человека: методы и подходы

Sequence-Tagged Sites (STS)

https://www.ncbi.nlm.nih.gov/dbSTS/

STS is a relatively short, easily PCR-amplified sequence (200 to 500 bp) which can be specifically amplified by PCR and detected in the presence of all other genomic sequences and whose location in the genome is mapped.



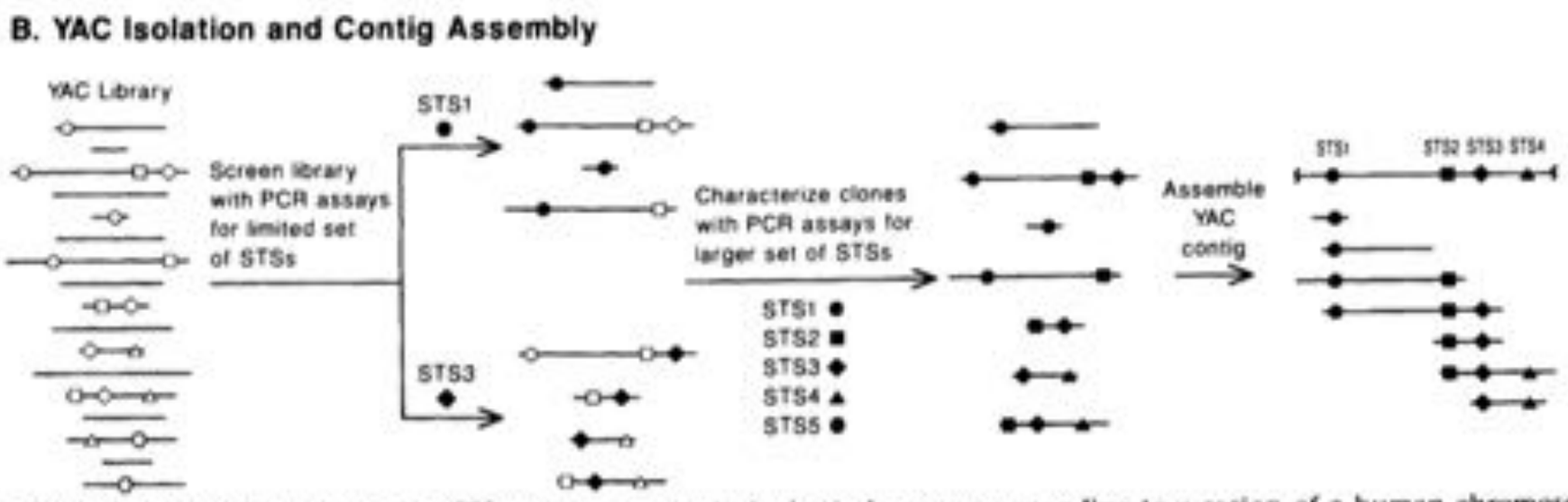


FIGURE 1 General strategy for constructing STS-content maps. (A) A physical map corresponding to a region of a human chromosome is

Olson M et al. A common language for physical mapping of the human genome. Science. 1989 Sep 29;245(4925):1434-5

https://genome.cshlp.org/content/1/2/77.full.pdf

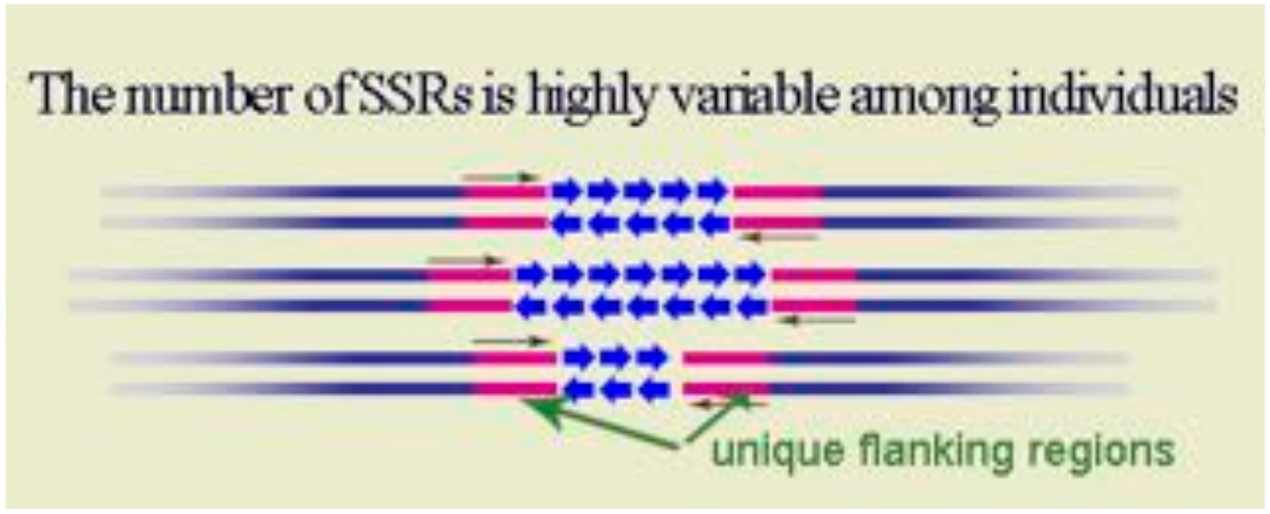
Генетические маркеры

https://www.ncbi.nlm.nih.gov/probe

- When STS loci contain genetic polymorphisms, they become valuable genetic markers, i.e. loci which can be used to distinguish individuals.
- A **genetic marker** is a gene or DNA sequence with a known location on a chromosome that can be used to identify individuals or species.

Some commonly used types of genetic markers are:

- RFLP (or Restriction fragment length polymorphism)
- SSLP (or Simple sequence length polymorphism)
- AFLP (or Amplified fragment length polymorphism)
- RAPD (or Random amplification of polymorphic DNA)
- VNTR (or Variable number tandem repeat)
- SSR Microsatellite polymorphism, (or Simple sequence repeat)
- SNP (or Single nucleotide polymorphism)
- STR (or Short tandem repeat)
- SFP (or Single feature polymorphism)
- DArT (or Diversity Arrays Technology)
- RAD markers (or Restriction site associated DNA markers)



Генетические маркеры
ДНК-дактилоскопия (DNA profiling)

Twenty CODIS Core Loci

In early 2015, the FBI announced that the validation project for additional CODI
had been completed and that an additional seven loci would be added to the C
effective January 1, 2017.[3] The additional seven loci—D1S1656, D2S441, D2S
D10S1248, D12S391, D19S433 and D22S1045—along with the original 13 loci
comprise the new CODIS Core Loci. Below is a listing of the 20 CODIS Core Lo

- CSF1PO
- D3S1358
- D5S818
- D7S820
- D8S1179
- D13S317
- D16S539
- D18S51
- D21S11
- FGA
- TH01
- TPOX
- vWA
- D1S1656 (effective January 1, 2017)
- D2S441 (effective January 1, 2017)
- D2S1338 (effective January 1, 2017)
- D10S1248 (effective January 1, 2017)
- D12S391 (effective January 1, 2017)
- D19S433 (effective January 1, 2017)
- D22S1045 (effective January 1, 2017)

Probe | Probe | CSF1PO

Create alert Limits Advanced

Display Settings: Full Report Send to:

Pr012387263

STS probe GDB:212649 for colony stimulating factor 1 receptor (CSF1R) and 3 more genes

Synopsis

Field Name	Values
Name	GDB:212649
Alias	CSF1PO
Type	STS
Application	
Source organism	
Source sequence	
Target organism	Homo sapiens
Target genes	CSF1R; CSF1R; CSF1R; CSF1R

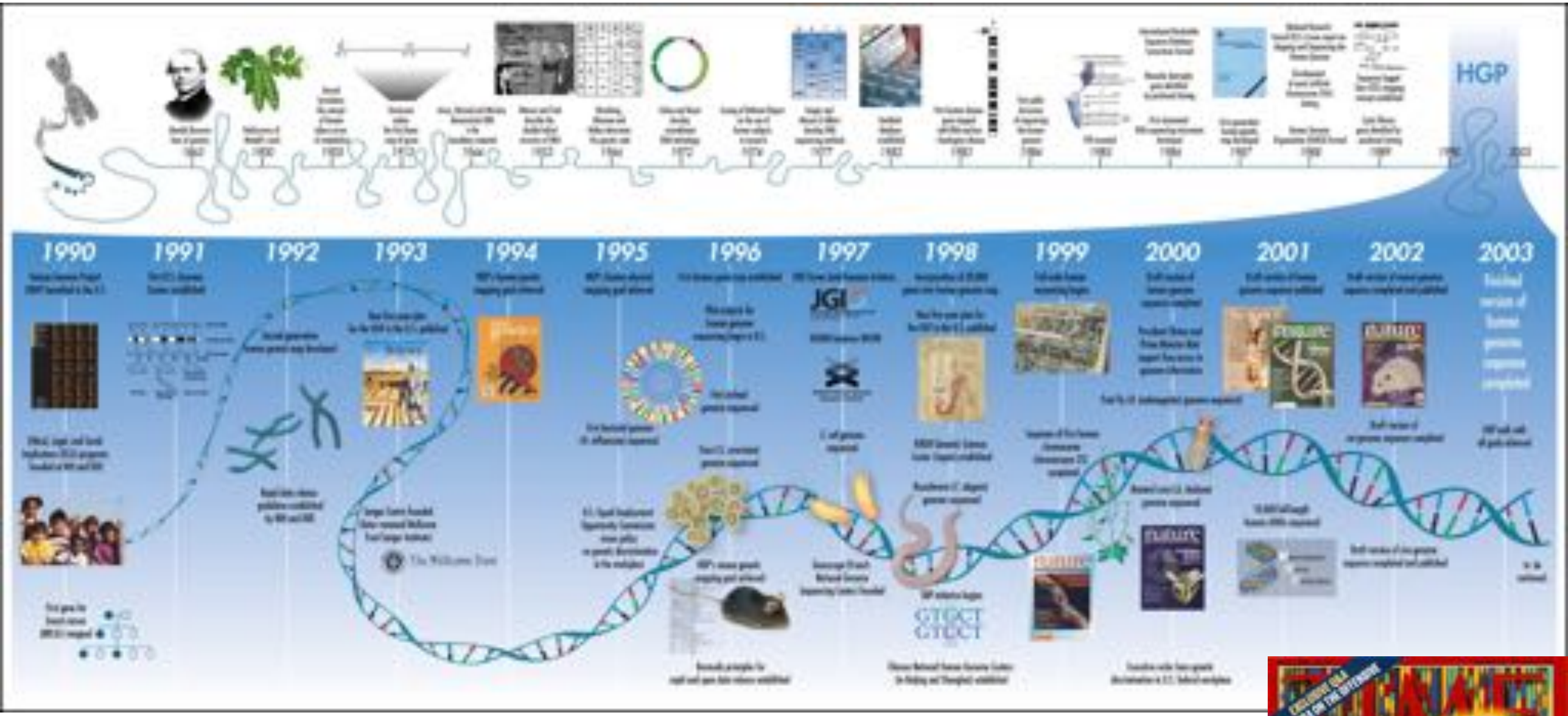
Sequences

>Probe|12387263|PRIMERF Forward PCR primer (outermost) (24b)
AACCTGAGTCTGCCAAGGACTAGC

>Probe|12387263|PRIMERR Reverse PCR primer (outermost) (24b)
TTCCACACACCACTGGCCATCTTC

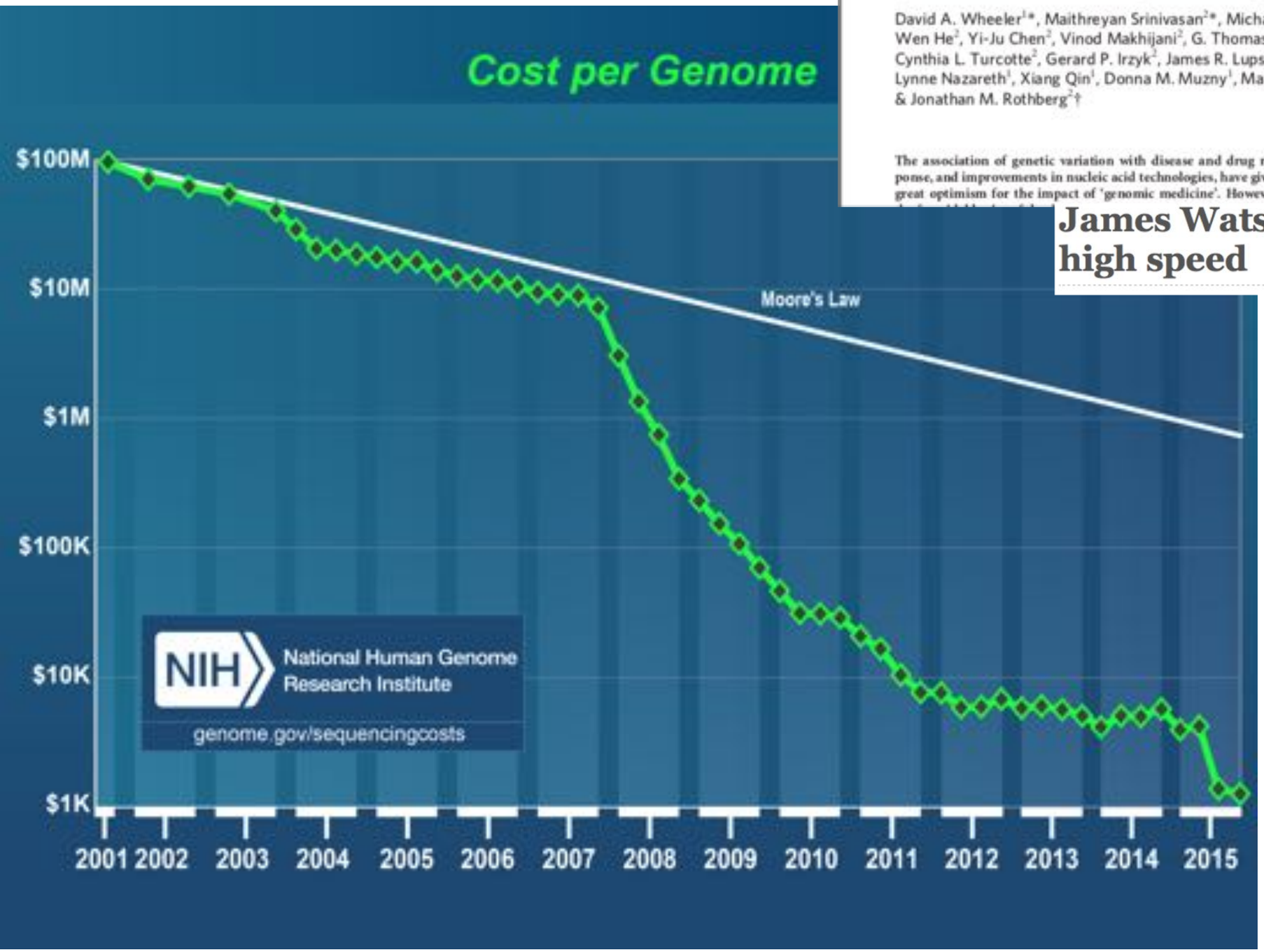
В России 3 декабря 2008 года Госдума приняла Федеральный закон «О государственной геномной регистрации в Российской Федерации»[19]. По этому закону создана федеральная база данных ДНК, содержащая информацию об осуждённых за тяжкие и особо тяжкие преступления, за преступления против половой неприкосновенности, а также о неопознанных трупах и о биологических следах, изъятых с мест совершения преступлений. Оператором базы данных является МВД России.

Проект геном человека



https://www.genome.gov/11006943/human-genome-project-completion-frequently-asked-questions/

NGS starts



nature Vol 452|17 April 2008|doi:10.1038/nature06884

LETTERS

The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2]†, Faheem Niazi[2], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2]†

The association of genetic variation with disease and drug response, and improvements in nucleic acid technologies, have given great optimism for the impact of 'genomic medicine'. However,

subject's DNA, including single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), and copy number variation (CNV).

James Watson's genome sequenced at high speed

2008

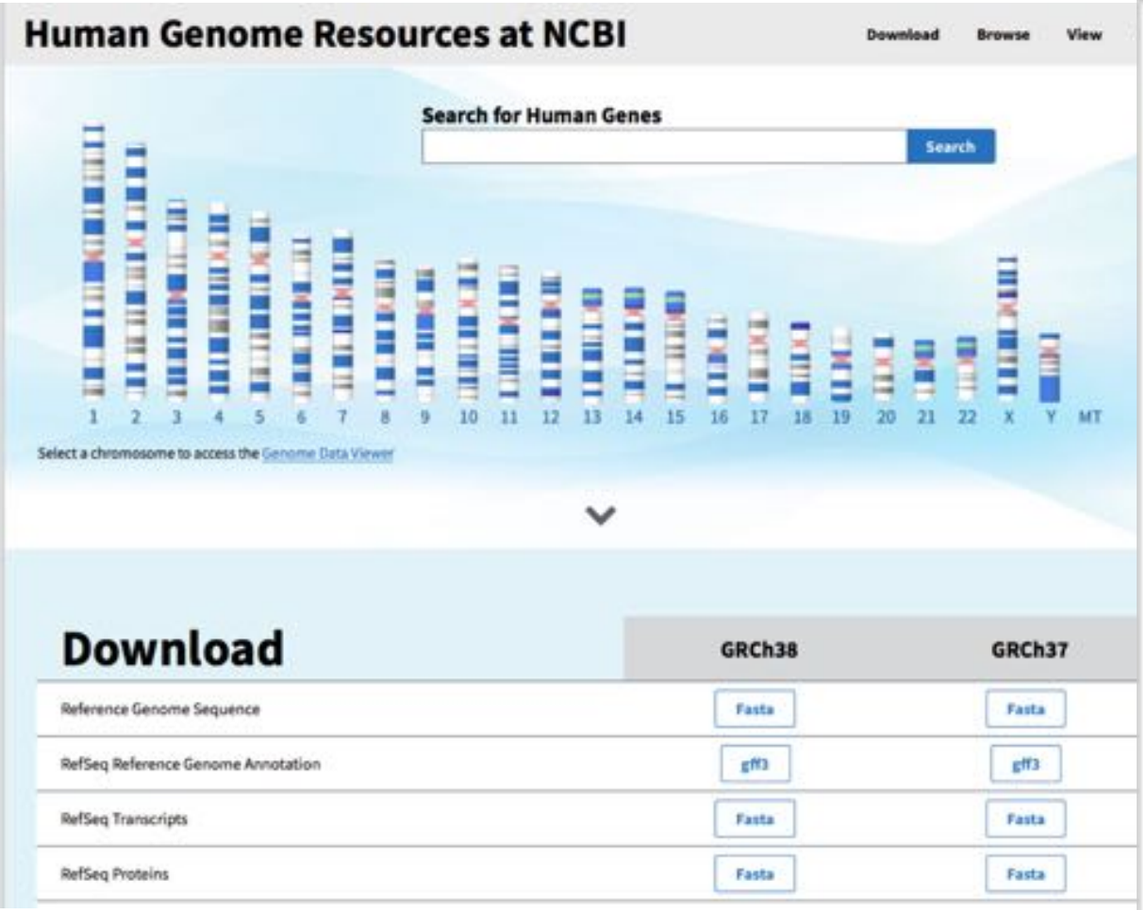
454 Life Sciences

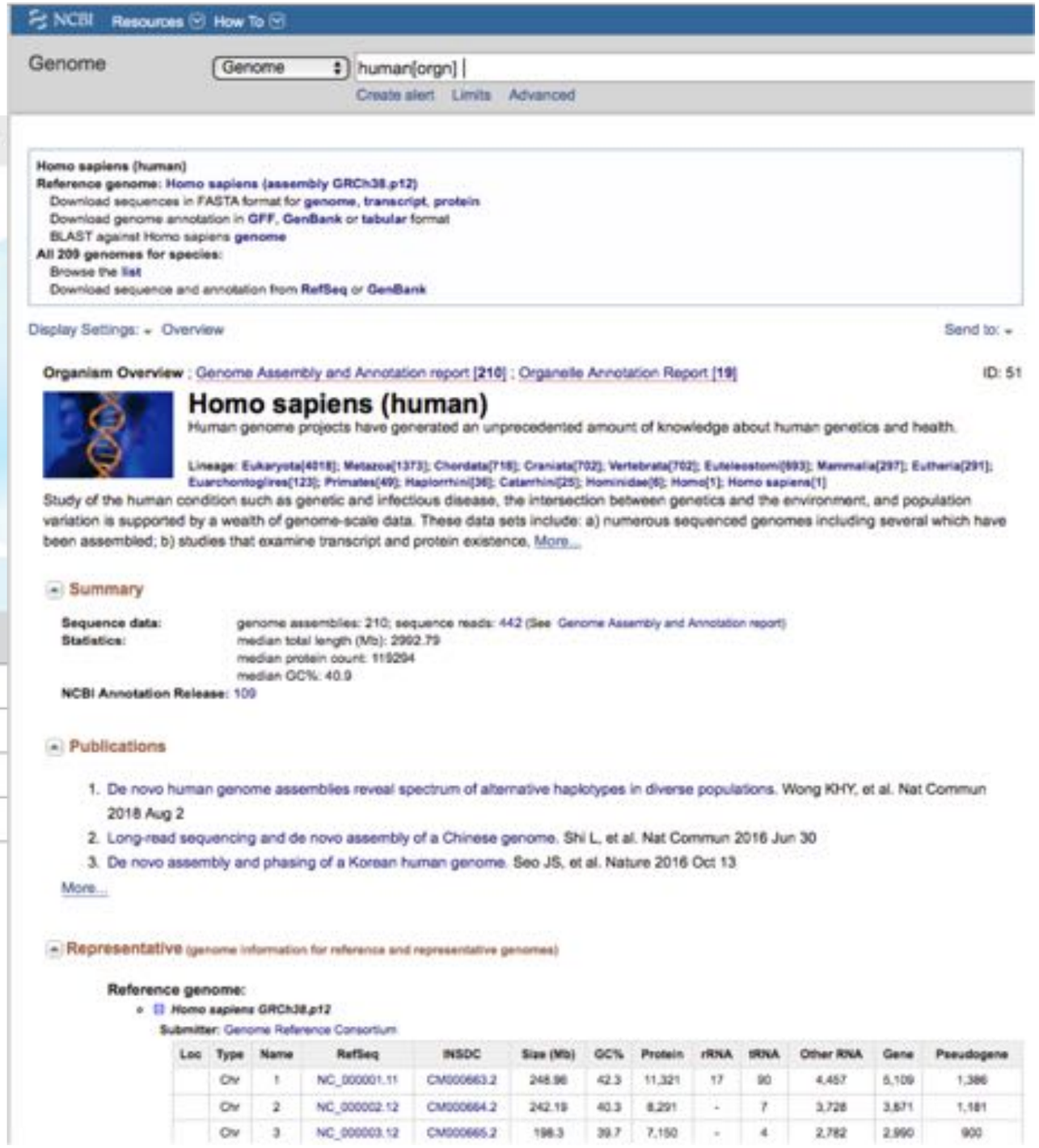
4 months

$1.5 mln

Где посмотреть на геном?

https://www.ncbi.nlm.nih.gov/genome/





https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/

Где посмотреть на геном?

- *Homo sapiens GRCh38.p12*

Submitter: Genome Reference Consortium

Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
	Chr	1	NC_000001.11	CM000663.2	248.96	42.3	11,321	17	90	4,457	5,109	1,386
	Chr	2	NC_000002.12	CM000664.2	242.19	40.3	8,291	-	7	3,728	3,871	1,181
	Chr	3	NC_000003.12	CM000665.2	198.3	39.7	7,150	-	4	2,782	2,990	900
	Chr	4	NC_000004.12	CM000666.2	190.22	38.3	4,599	-	1	2,193	2,441	803
	Chr	5	NC_000005.10	CM000667.2	181.54	39.5	4,729	-	17	2,194	2,592	778
	Chr	6	NC_000006.12	CM000668.2	170.81	39.6	5,522	-	138	2,453	3,005	882
	Chr	7	NC_000007.14	CM000669.2	159.35	40.7	5,112	-	22	2,330	2,792	911
	Chr	8	NC_000008.11	CM000670.2	145.14	40.2	4,199	-	4	2,011	2,165	671
	Chr	9	NC_000009.12	CM000671.2	138.4	42.3	4,699	-	3	2,222	2,270	706
	Chr	10	NC_000010.11	CM000672.2	133.8	41.6	5,429	-	3	2,133	2,179	640
	Chr	11	NC_000011.10	CM000673.2	135.09	41.6	6,394	-	13	2,336	2,924	829
	Chr	12	NC_000012.12	CM000674.2	133.28	40.8	5,975	-	9	2,457	2,526	691
	Chr	13	NC_000013.11	CM000675.2	114.36	40.2	2,056	-	4	1,243	1,385	476
	Chr	14	NC_000014.9	CM000676.2	107.04	42.2	3,501	-	18	1,704	2,065	585
	Chr	15	NC_000015.10	CM000677.2	101.99	43.4	3,623	-	9	1,810	1,824	554
	Chr	16	NC_000016.10	CM000678.2	90.34	45.1	4,625	-	27	1,761	1,938	469
	Chr	17	NC_000017.11	CM000679.2	83.26	45.3	6,226	-	33	2,243	2,450	556
	Chr	18	NC_000018.10	CM000680.2	80.37	39.8	2,029	-	1	996	984	296
	Chr	19	NC_000019.10	CM000681.2	58.62	47.9	6,750	-	6	1,877	2,499	523
	Chr	20	NC_000020.11	CM000682.2	64.44	43.9	2,904	-	-	1,308	1,358	338
	Chr	21	NC_000021.9	CM000683.2	46.71	42.2	1,297	12	1	707	777	207
	Chr	22	NC_000022.11	CM000684.2	50.82	47.7	2,582	-	-	1,014	1,189	354
	Chr	X	NC_000023.11	CM000685.2	156.04	39.6	3,801	-	4	1,265	2,186	875
	Chr	Y	NC_000024.10	CM000686.2	57.23	45.4	324	-	-	311	580	392
		MT	NC_012920.1	J01415.2	0.02	44.4	13	2	22	-	37	-
	Un	-	.	-	169.03	44.3	6,143	17	161	3,437	6,543	1,878

https://www.ncbi.nlm.nih.gov/genome/51

Где посмотреть на геном?

NCBI Resources How To

Nucleotide [Nucleotide]

Advanced

The Nucleotide database will include EST and GSS sequences in early 2019. Read more.

GenBank ▾ Send to: ▾

Due to the size of this record, annotated features are not shown by default. Use "Customize view" section to change the display

Homo sapiens chromosome 1, GRCh38.p12 Primary Assembly

NCBI Reference Sequence: NC_000001.11

FASTA Graphics

Go to:

```
LOCUS       NC_000001           248956422 bp    DNA     linear   CON 26-MAR-2018
DEFINITION  Homo sapiens chromosome 1, GRCh38.p12 Primary Assembly.
ACCESSION   NC_000001
VERSION     NC_000001.11
DBLINK      BioProject: PRJNA168
            Assembly: GCF_000001405.38
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
FEATURES             Location/Qualifiers
     source          1..248956422
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="1"
CONTIG      join(gap(10000),NT_077402.3:1..197666,gap(50000),
            NT_187170.1:1..40302,gap(50000),NT_077912.2:1..188020,gap(50000),
            NT_032977.10:1..121390471,gap(50000),NT_187171.1:1..198076,
            gap(100),NT_187172.1:1..278512,gap(100),NT_187173.1:1..2282185,
            gap(100),NT_187174.1:1..63597,gap(100),NT_187175.1:1..83495,
            gap(100),NT_187176.1:1..251763,gap(18000000),
            NT_004487.20:1..80374348,gap(50000),NT_167186.2:1..25337487,
            gap(10000))
//
```

Центромерная область

Homo sapiens chromosome 1, GRCh38.p12 Primary As

NCBI Reference Sequence: NC_000001.11

GenBank Graphics

```
>NC_000001.11 Homo sapiens chromosome 1, GRCh38.p12 Primary Assembly
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

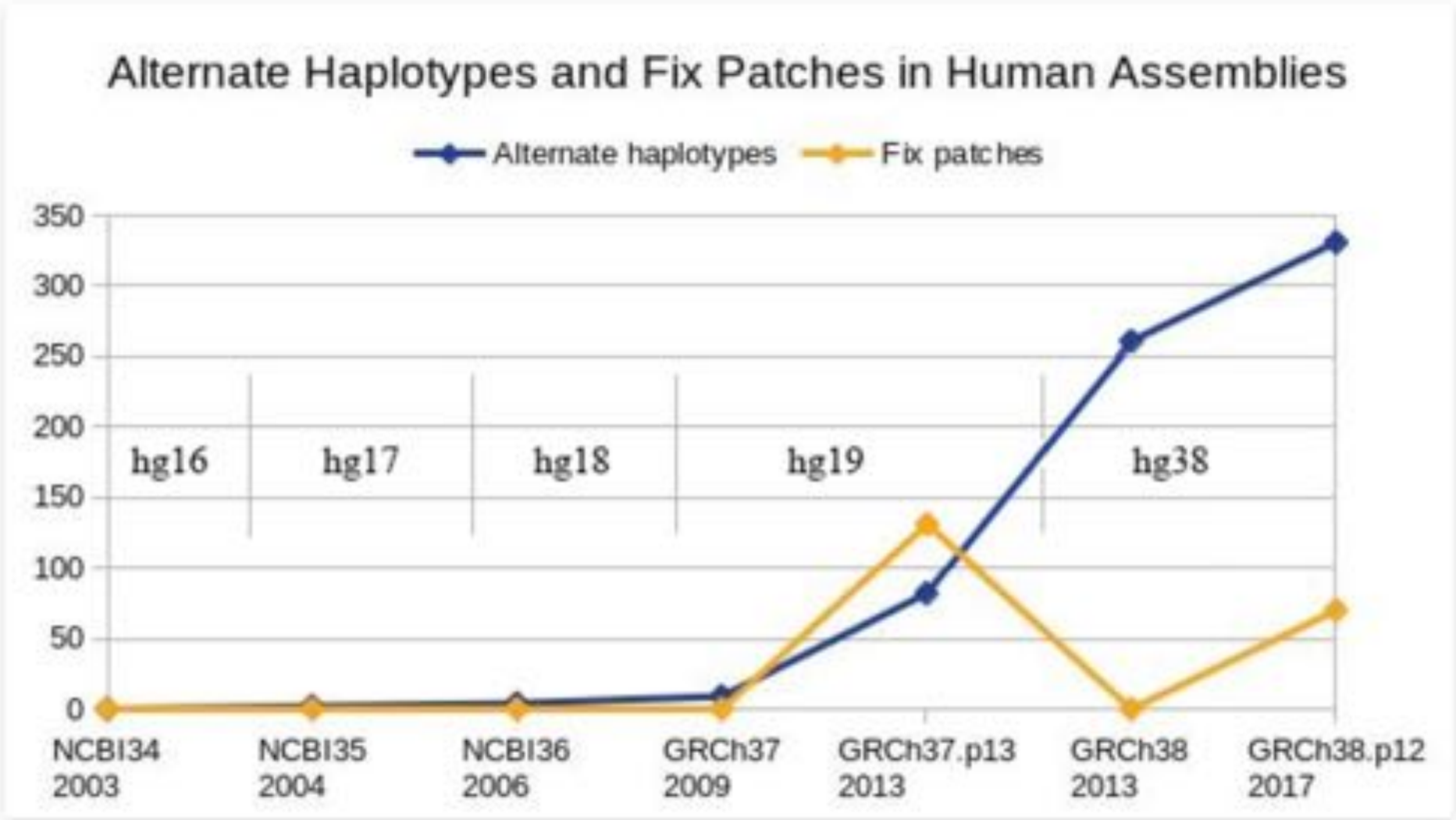
Понятие Сборки генома и версии

Human	hg38	Dec. 2013	Genome Reference Consortium GRCh38	Available
	hg19	Feb. 2009	Genome Reference Consortium GRCh37	Available
	hg18	Mar. 2006	NCBI Build 36.1	Available
	hg17	May 2004	NCBI Build 35	Available
	hg16	Jul. 2003	NCBI Build 34	Available
	hg15	Apr. 2003	NCBI Build 33	Archived
	hg13	Nov. 2002	NCBI Build 31	Archived
	hg12	Jun. 2002	NCBI Build 30	Archived
	hg11	Apr. 2002	NCBI Build 29	Archived (data only)
	hg10	Dec. 2001	NCBI Build 28	Archived (data only)
	hg8	Aug. 2001	UCSC-assembled	Archived (data only)
	hg7	Apr. 2001	UCSC-assembled	Archived (data only)
	hg6	Dec. 2000	UCSC-assembled	Archived (data only)
	hg5	Oct. 2000	UCSC-assembled	Archived (data only)
	hg4	Sep. 2000	UCSC-assembled	Archived (data only)
	hg3	Jul. 2000	UCSC-assembled	Archived (data only)
	hg2	Jun. 2000	UCSC-assembled	Archived (data only)
	hg1	May 2000	UCSC-assembled	Archived (data only)

Последняя версия GRCh38.p13

p=patch
Patch releases do not change chromosome coordinates

Понятие Сборки генома и версии



Понятие Сборки генома

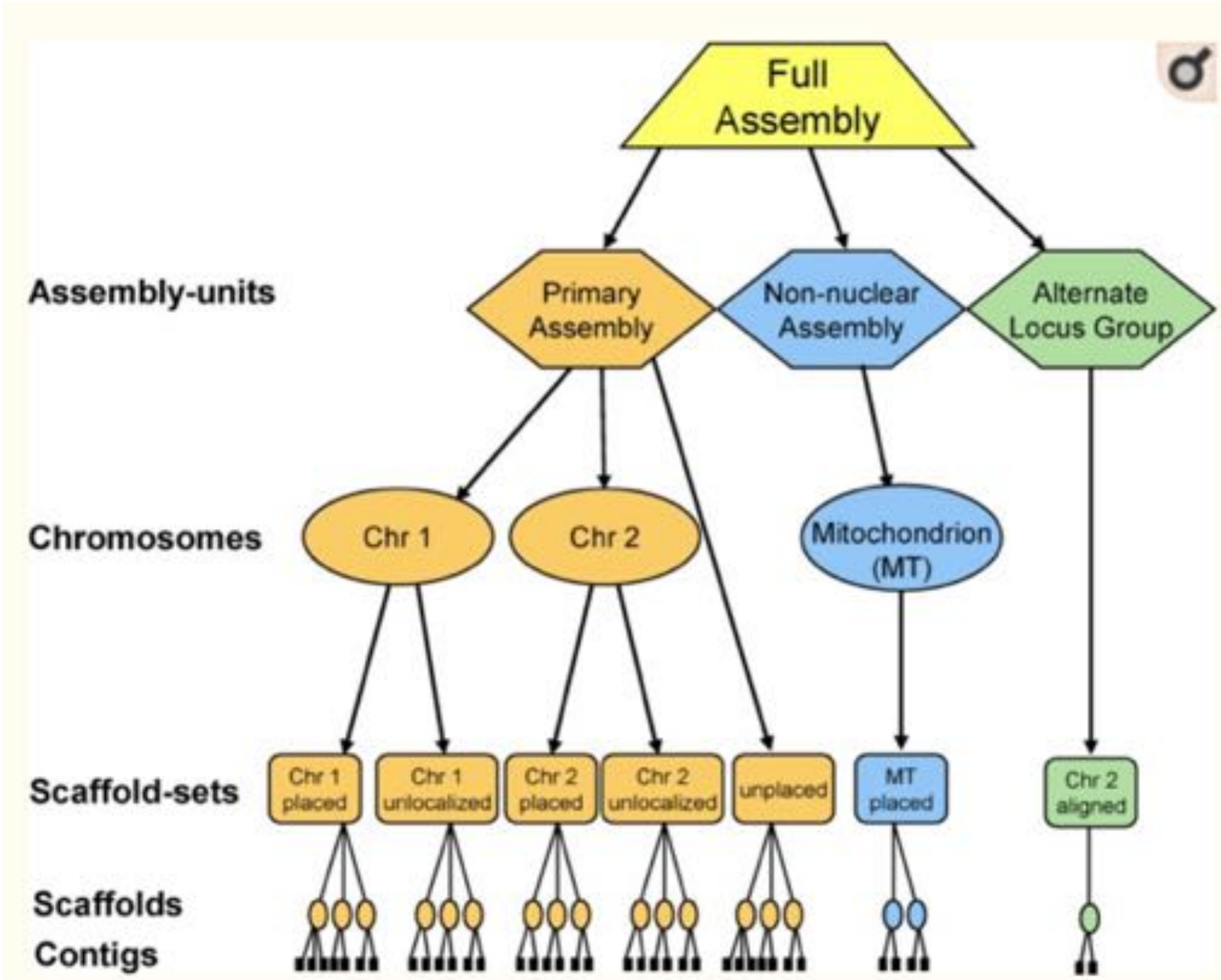


Figure 2.

The NCBI genome assembly model. The diagram depicts the assembly organization for a eukaryote with two

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702866/
https://www.ncbi.nlm.nih.gov/assembly/help/

Где посмотреть и скачать сборку генома?

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

NCBI Resources How To Sign in to NCBI

Assembly Assembly Search

Advanced Browse by organism Help

Full Report Send to:

GRCh38.p12

⚠ **This assembly has been updated.** See current version

Description: Genome Reference Consortium Human Build 38 patch release 12 (GRCh38.p12)
Organism name: Homo sapiens (human)
BioProject: PRJNA31257
Submitter: Genome Reference Consortium
Date: 2017/12/21
Assembly type: haploid-with-alt-loci
Release type: patch
Assembly level: Chromosome
Genome representation: full
GenBank assembly accession: GCA_000001405.27 (replaced)

See Genome information for **Homo sapiens**

There are 223 assemblies for this organism

See more

Access the data

Browse in Genome Data Viewer
View the Annotation Report
Download the RefSeq assembly
Download the GenBank assembly
BLAST search the assembly
Download the full sequence report
Download the statistics report
Download the regions report

Number of regions with alternate loci or patches	317
Total sequence length	3,099,706,404
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
Contig N50	57,879,411
Contig L50	18
Total number of chromosomes and plasmids	24
Number of component sequences (WGS or clone)	35,613

Где посмотреть и скачать сборку генома?

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

Assembly Definition | Assembly Statistics

Global assembly definition

Download the full sequence report

Click on the table row to see sequence details in the table to the right

Assembly Unit Name
Primary Assembly
PATCHES
ALT_REF_LOCI_1
ALT_REF_LOCI_2
ALT_REF_LOCI_3
ALT_REF_LOCI_4
ALT_REF_LOCI_5
ALT_REF_LOCI_6
ALT_REF_LOCI_7

Assembly Unit: Primary Assembly (GCF_000001405.15)

Molecule name	GenBank sequence		RefSeq sequence	Unlocalized sequences count
Chromosome 1	CM000663.2	=	NC_000001.11	9
Chromosome 2	CM000664.2	=	NC_000002.12	2
Chromosome 3	CM000665.2	=	NC_000003.12	1
Chromosome 4	CM000666.2	=	NC_000004.12	1
Chromosome 5	CM000667.2	=	NC_000005.10	1
Chromosome 6	CM000668.2	=	NC_000006.12	0
Chromosome 7	CM000669.2	=	NC_000007.14	0
Chromosome 8	CM000670.2	=	NC_000008.11	0
Chromosome 9	CM000671.2	=	NC_000009.12	4
Chromosome 10	CM000672.2	=	NC_000010.11	0
Chromosome 11	CM000673.2	=	NC_000011.10	1

Где посмотреть и скачать сборку генома?

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

Molecule	Sequence Role	Total Length	Scaffold Count	Ungapped Length	Scaffold N50	Spanned Gaps	Unspanned Gaps
All	**Assembled molecule**	3,099,706,404	472	2,948,583,725	67,794,873	526	349
	All	249,698,942	21	231,223,641	121,390,471	64	13
Chromosome 1	**Assembled molecule**	248,956,422	12	230,481,121	121,390,471	64	13
	Unlocalized scaffolds	742,520	9	742,520	127,682	0	0
	All	242,508,799	9	240,863,511	147,687,514	14	8
Chromosome 2	**Assembled molecule**	242,193,529	7	240,548,241	147,687,514	14	8
	Unlocalized scaffolds	315,270	2	315,270	161,471	0	0

Где посмотреть и скачать сборку генома?

AGP-файлы описывают сборку

```
# AGP dumped from Genomic-Collection: Assembly GRCh38.p12 Primary_Assembly
# (Assembly accession = GCF_000001305.15, asm_id = 265453)
# Chromosomes-from-Scaffolds (placed)
##agp-version 2.0
# Format: object object_beg object_end part_number component_type component_id component_beg component_end  orientation
#   Gaps: object object_beg object_end part_number    N          gap_length   gap_type      linkage        evidence
NC_000001.11    1   10000   1   N   10000   telomere    no  na
NC_000001.11    10001   207666  2   F   NT_077402.3 1   197666  +
NC_000001.11    207667  257666  3   N   50000   contig  no  na
NC_000001.11    257667  297968  4   F   NT_187170.1 1   40302   +
NC_000001.11    297969  347968  5   N   50000   contig  no  na
NC_000001.11    347969  535988  6   F   NT_077912.2 1   188020  +
NC_000001.11    535989  585988  7   N   50000   contig  no  na
NC_000001.11    585989  121976459   8   F   NT_032977.10    1   121390471   +
NC_000001.11    121976460   122026459   9   N   50000   contig  no  na
NC_000001.11    122026460   122224535   10  F   NT_187171.1 1   198076  +
NC_000001.11    122224536   122224635   11  N   100 contig  no  na
NC_000001.11    122224636   122503147   12  F   NT_187172.1 1   278512  +
NC_000001.11    122503148   122503247   13  N   100 contig  no  na
NC_000001.11    122503248   124785432   14  F   NT_187173.1 1   2282185 +
NC_000001.11    124785433   124785532   15  N   100 contig  no  na
NC_000001.11    124785533   124849129   16  F   NT_187174.1 1   63597   +
NC_000001.11    124849130   124849229   17  N   100 contig  no  na
NC_000001.11    124849230   124932724   18  F   NT_187175.1 1   83495   +
NC_000001.11    124932725   124932824   19  N   100 contig  no  na
NC_000001.11    124932825   125184587   20  F   NT_187176.1 1   251763  +
NC_000001.11    125184588   143184587   21  N   18000000    heterochromatin no  na
NC_000001.11    143184588   223558935   22  F   NT_004487.20    1   80374348    +
NC_000001.11    223558936   223608935   23  N   50000   contig  no  na
NC_000001.11    223608936   248946422   24  F   NT_167186.2 1   25337487    +
NC_000001.11    248946423   248956422   25  N   10000   telomere    no  na
```

Был ли геном секвенирован на 100%?

- Только эухроматическая часть
- Проблемы с повторами
- Проблемы со структурным полиморфизмом

nature
International journal of science

Article | Published: 21 October 2004

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium

Nature **431**, 931–945 (2004) | Download Citation ⭳

this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work

3,234.83 Mb (Mega-basepairs) per haploid genome

GRCh38.p12

Total sequence length		3,099,706,404
Total ungapped length		2,948,583,725
unplaced	**Assembled molecule**	4,457,764

https://www.ncbi.nlm.nih.gov/grc/help/faq/ https://doi.org/10.1371/journal.pcbi.1003628

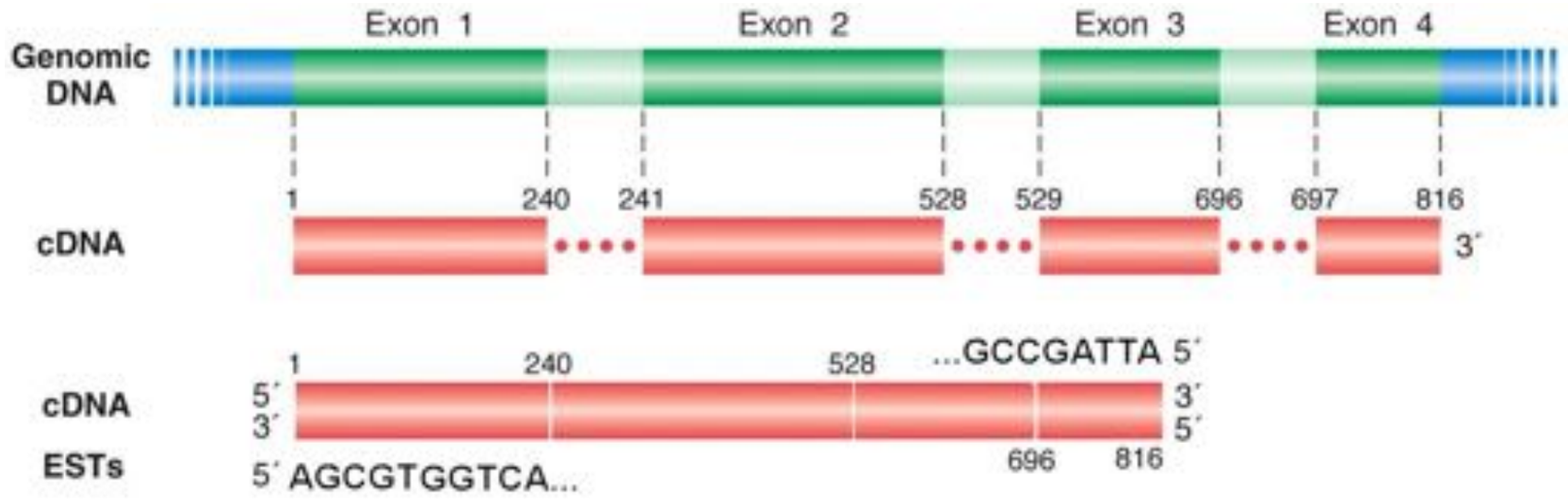
Аннотация геномов

- Структурная аннотация – разбивка генома на гены.
- Функциональная аннотация – функции генов, экспрессия, регуляция.
- Возможны чисто вычислительные алгоритмы поиска генов.
- Поиск открытых рамок считывания (ORF), поиск гомологов
- Важную роль играли/играют Expressed Sequence Tags (EST) и данные RNAseq
- Понятие complementary DNA, cDNA, кДНК.

dbEST
Expressed
Sequence Tags
database

↕

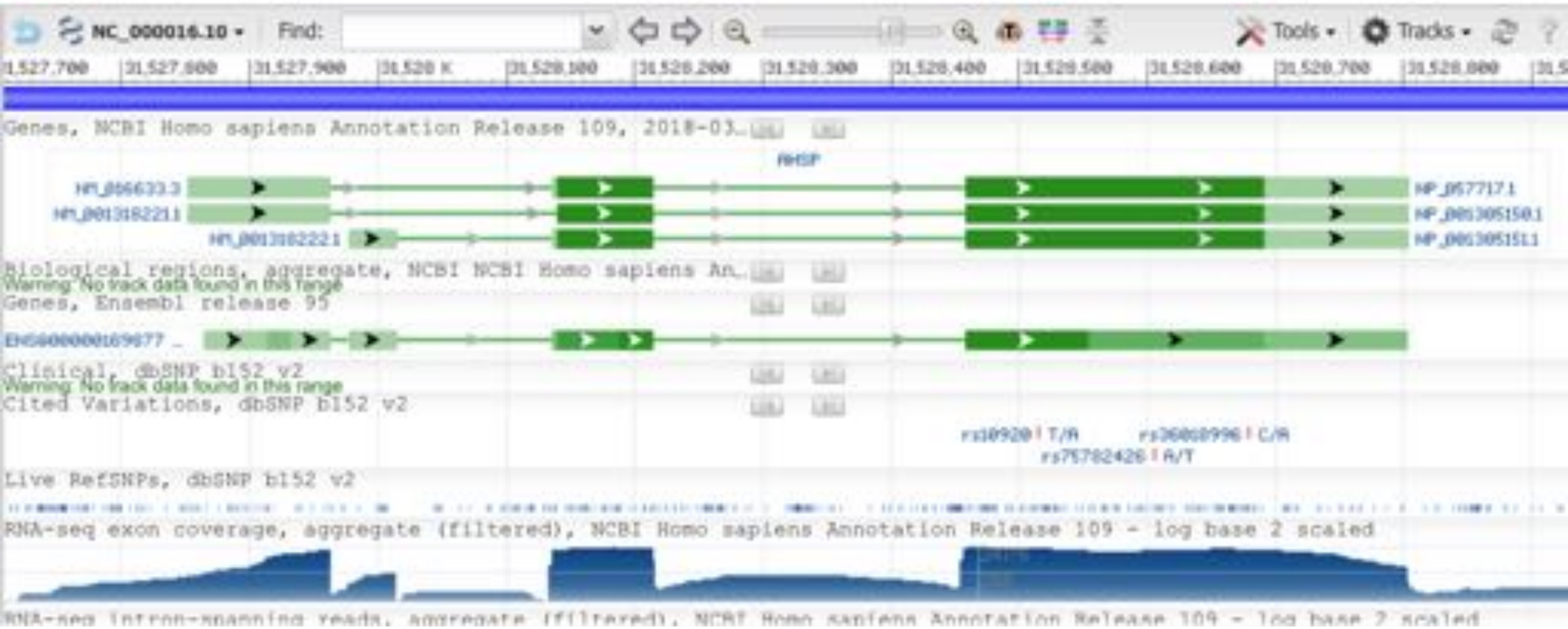
1992



https://www.ncbi.nlm.nih.gov/dbEST/

Аннотация геномов

- RNAseq данные помогают в аннотации геномов



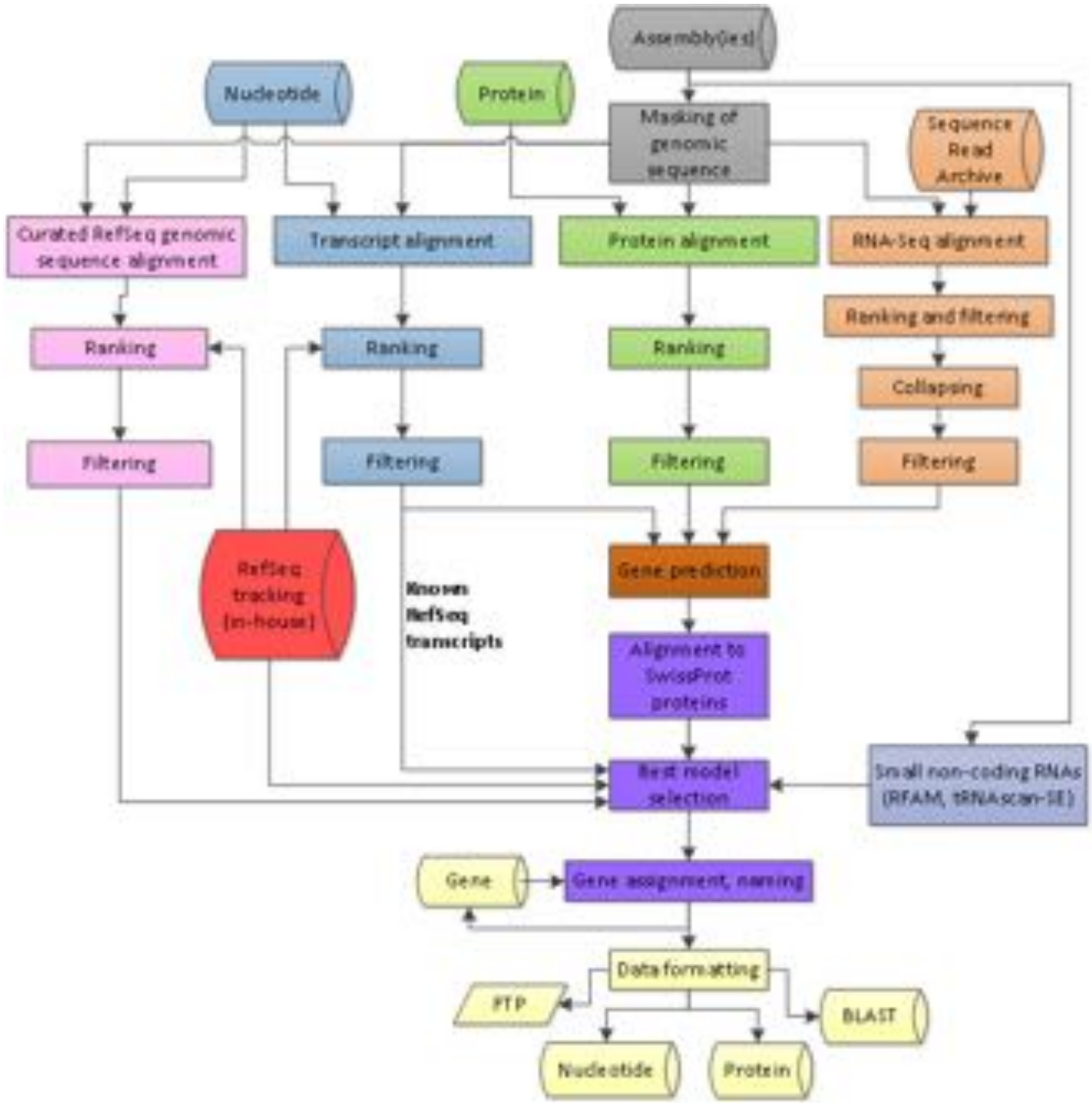
Аннотация геномов

- Формат GFF, GFF3

```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build GRCh38.p12
#!genome-build-accession NCBI_Assembly:GCF_000001405.38
#!annotation-source NCBI Homo sapiens Annotation Release 109
##sequence-region NC_000001.11 1 248956422
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=9606
NC_000001.11    RefSeq  region  1       248956422       .       +       .       ID=NC_000001.11:1..248956422;Dbx
NC_000001.11    BestRefSeq      pseudogene      11874   14409   .       +       .       ID=gene-DDX11L1;Dbxref=G
NC_000001.11    BestRefSeq      transcript      11874   14409   .       +       .       ID=rna-NR_046018.2;Paren
NC_000001.11    BestRefSeq      exon    11874   12227   .       +       .       ID=exon-NR_046018.2-1;Parent=rna
NC_000001.11    BestRefSeq      exon    12613   12721   .       +       .       ID=exon-NR_046018.2-2;Parent=rna
NC_000001.11    BestRefSeq      exon    13221   14409   .       +       .       ID=exon-NR_046018.2-3;Parent=rna
NC_000001.11    BestRefSeq      pseudogene      14362   29370   .       -       .       ID=gene-WASH7P;Dbxref=Ge
rue
NC_000001.11    BestRefSeq      transcript      14362   29370   .       -       .       ID=rna-NR_024540.1;Paren
.1
NC_000001.11    BestRefSeq      exon    29321   29370   .       -       .       ID=exon-NR_024540.1-1;Parent=rna
NC_000001.11    BestRefSeq      exon    24738   24891   .       -       .       ID=exon-NR_024540.1-2;Parent=rna
NC_000001.11    BestRefSeq      exon    18268   18366   .       -       .       ID=exon-NR_024540.1-3;Parent=rna
NC_000001.11    BestRefSeq      exon    17915   18061   .       -       .       ID=exon-NR_024540.1-4;Parent=rna
NC_000001.11    BestRefSeq      exon    17606   17742   .       -       .       ID=exon-NR_024540.1-5;Parent=rna
NC_000001.11    BestRefSeq      exon    17233   17368   .       -       .       ID=exon-NR_024540.1-6;Parent=rna
NC_000001.11    BestRefSeq      exon    16858   17055   .       -       .       ID=exon-NR_024540.1-7;Parent=rna
NC_000001.11    BestRefSeq      exon    16607   16765   .       -       .       ID=exon-NR_024540.1-8;Parent=rna
NC_000001.11    BestRefSeq      exon    15796   15947   .       -       .       ID=exon-NR_024540.1-9;Parent=rna
NC_000001.11    BestRefSeq      exon    14970   15038   .       -       .       ID=exon-NR_024540.1-10;Parent=rn
NC_000001.11    BestRefSeq      exon    14362   14829   .       -       .       ID=exon-NR_024540.1-11;Parent=rn
NC_000001.11    BestRefSeq      gene    17369   17436   .       -       .       ID=gene-MIR6859-1;Dbxref=GeneID:
NC_000001.11    BestRefSeq      primary_transcript      17369   17436   .       -       .       ID=rna-NR_106918
ipt_id=NR_106918.1
NC_000001.11    BestRefSeq      exon    17369   17436   .       -       .       ID=exon-NR_106918.1-1;Parent=rna
NC_000001.11    BestRefSeq      miRNA   17369   17391   .       -       .       ID=rna-MIR6859-1;Parent=rna-NR_1
```

Аннотация геномов, автоматические пайплайны

NCBI Eukaryotic Genome Annotation Pipeline



Аннотация геномов

Основные ресурсы/проекты по аннотации

EMBL-EBI

wellcome sanger institute

Ensembl

havana
human and vertebrate analysis and annotation

GENCODE

NCBI

RefSeq

Consensus Coding Sequence (CCDS) Project

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

UCSC annotation

SGD Saccharomyces GENOME DATABASE

HGNC
HUGO Gene Nomenclature Committee

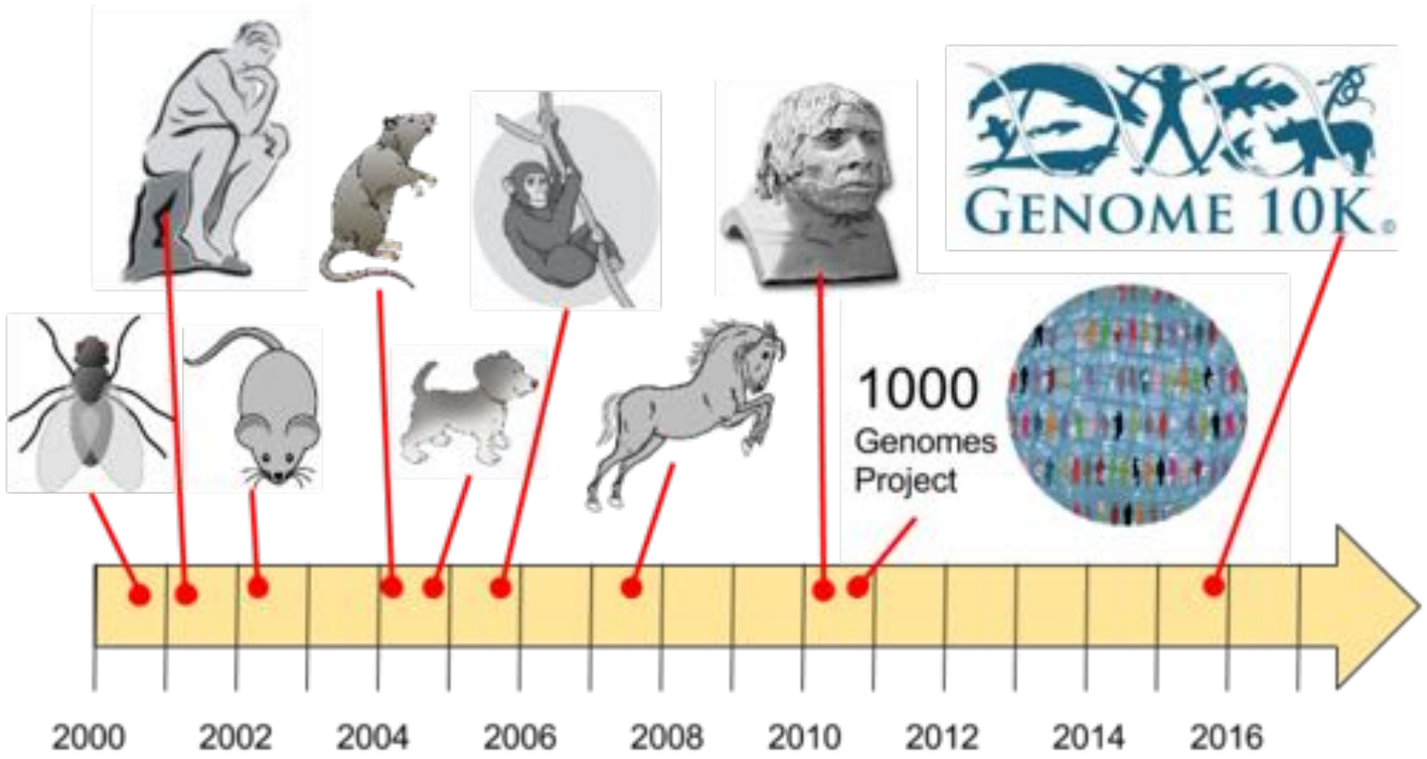
HGVS
HUMAN GENOME VARIATION SOCIETY

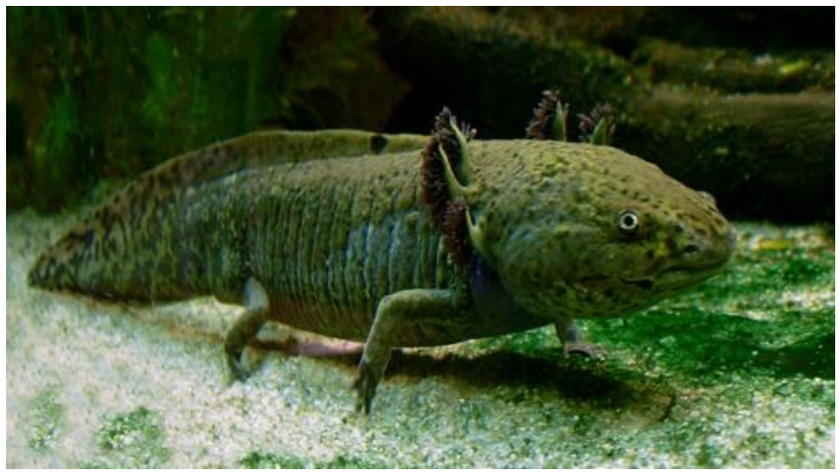
FlyBase

WormBase

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4339237/

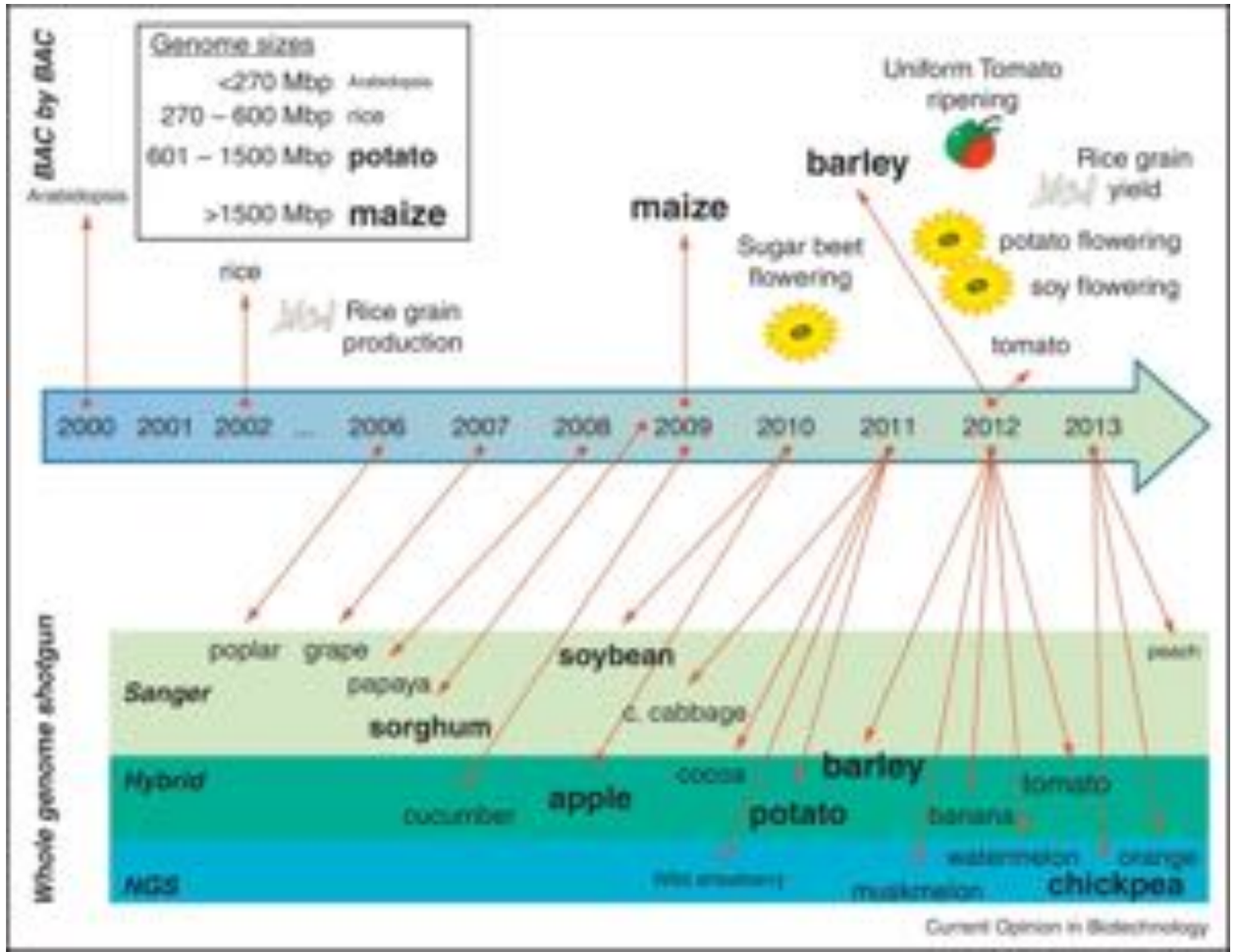
Геномы других организмов





Аксолотль мексиканской амбистомы

32 Гб – секвенирован в 2018



Геномы других организмов

Ambystoma mexicanum (axolotl)

Ambystoma mexicanum strain:DD151 Genome sequencing and assembly

Lineage: Eukaryota[4034]; Metazoa[1378]; Chordata[721]; Craniata[705]; Vertebrata[705]; Euteleo]; Ambystomatidae[1]; Ambystoma[1]; Ambystoma mexicanum[1]

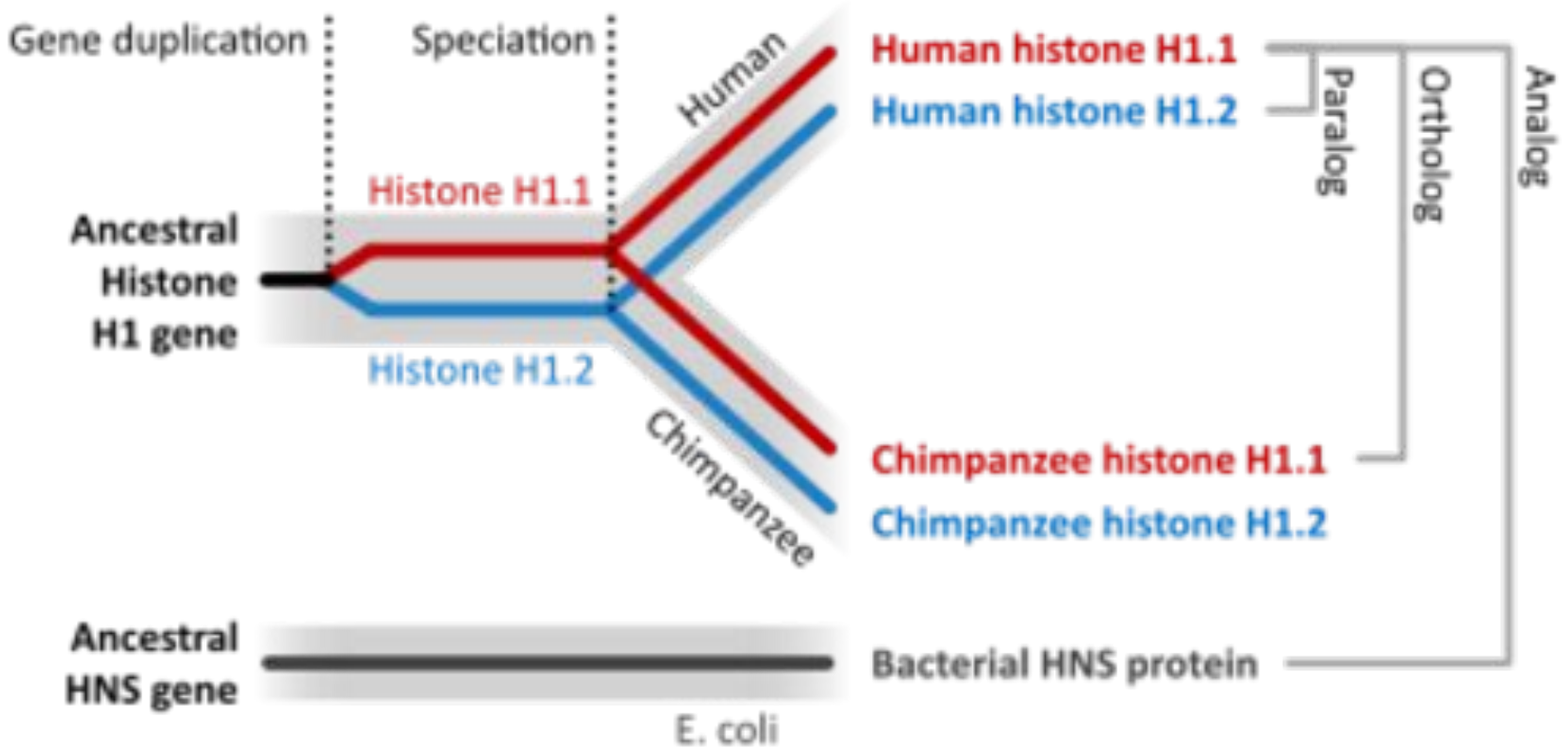
Summary

Submitter:	Max Planck Society/University of Kentucky
Assembly level:	Chromosome
Assembly:	GCA_002915635.2 ASM291563v2 **scaffolds:** 98,070 **contigs:** 891,205 **N50:** 216,366 **L50:** 35,791
BioProjects:	PRJNA378970
Whole Genome Shotgun (WGS):	INSDC: PGSH00000000.1
Statistics:	total length (Mb): 32396.4

Replicon Info

Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%
	Chr	1P	-	CM010927.1	1,477.09	46.7
	Chr	1Q	-	CM010928.1	1,479.73	46.5
	Chr	2P	-	CM010929.1	1,412.62	46.6
	Chr	2Q	-	CM010930.1	1,511.87	46.5
	Chr	3P	-	CM010931.1	1,240.32	46.7
	Chr	3Q	-	CM010932.1	1,258.74	46.7
	Chr	4P	-	CM010933.1	1,160.65	46.5
	Chr	7	-	CM010939.1	2,030.16	46.3
	Chr	4Q	-	CM010934.1	1,294.5	46.4
	Chr	8	-	CM010940.1	1,711.68	46.5
	Chr	5P	-	CM010935.1	1,291.88	46.3
	Chr	9	-	CM010941.1	1,496.29	46.6
	Chr	5Q	-	CM010936.1	1,339.52	46.5
	Chr	10	-	CM010942.1	1,640.17	46.5
	Chr	6P	-	CM010937.1	1,551.79	46.4
	Chr	11	-	CM010943.1	1,437.31	46.5
	Chr	6Q	-	CM010938.1	1,566.49	46.3
	Chr	12	-	CM010944.1	1,211.25	46.5
	Chr	13	-	CM010945.1	719.86	47.1
	Chr	14	-	CM010946.1	658.39	47.0

Сравнительная/эволюционная геномика
Ортологи, паралоги, COGs, синтения

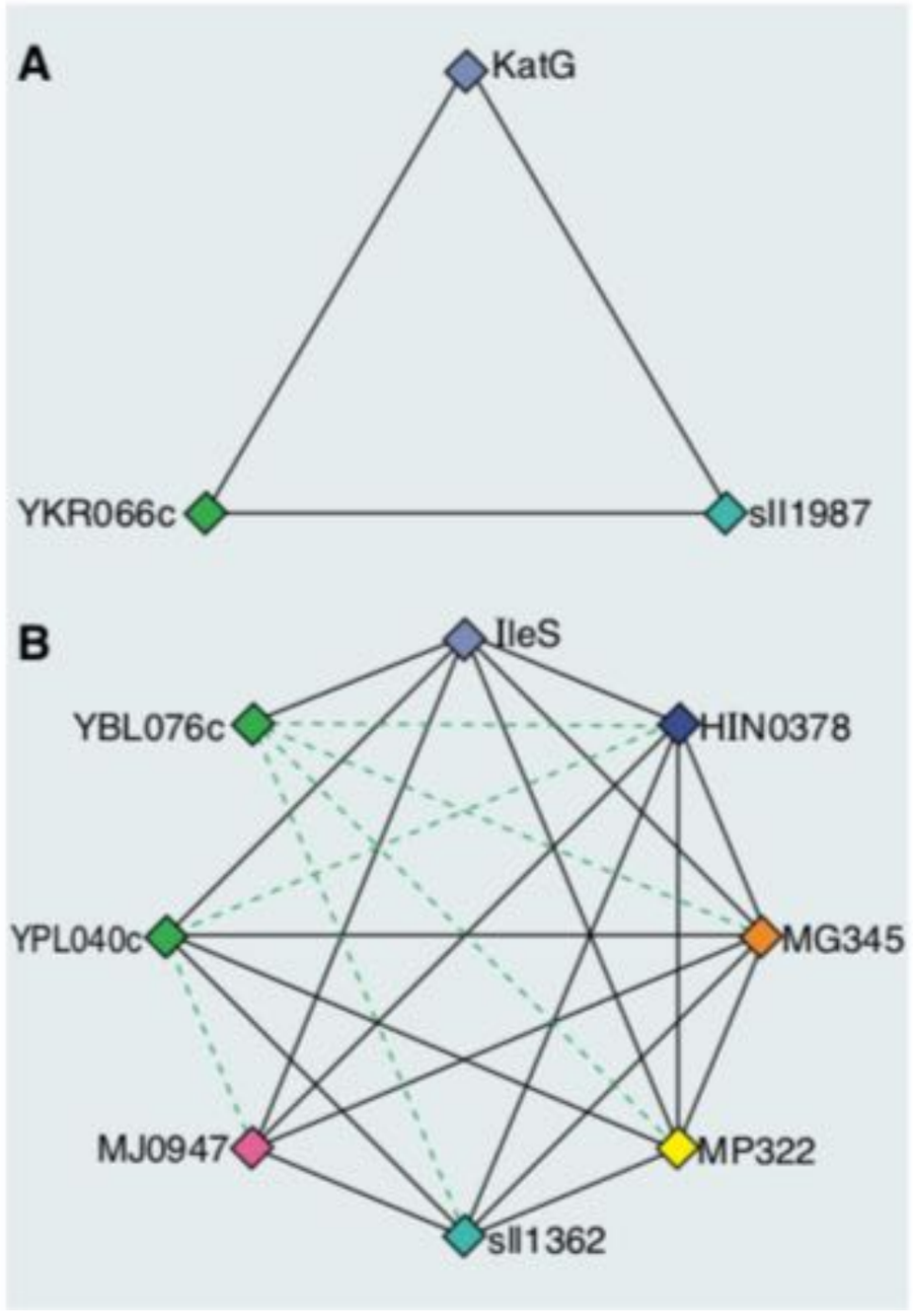


Сравнительная/эволюционная геномика

COGs=Clusters of Orthologous Groups

COGs

Phylogenetic classification of proteins encoded in complete genomes



A Genomic Perspective on Protein Families

Roman L. Tatusov, Eugene V. Koonin,* David J. Lipman

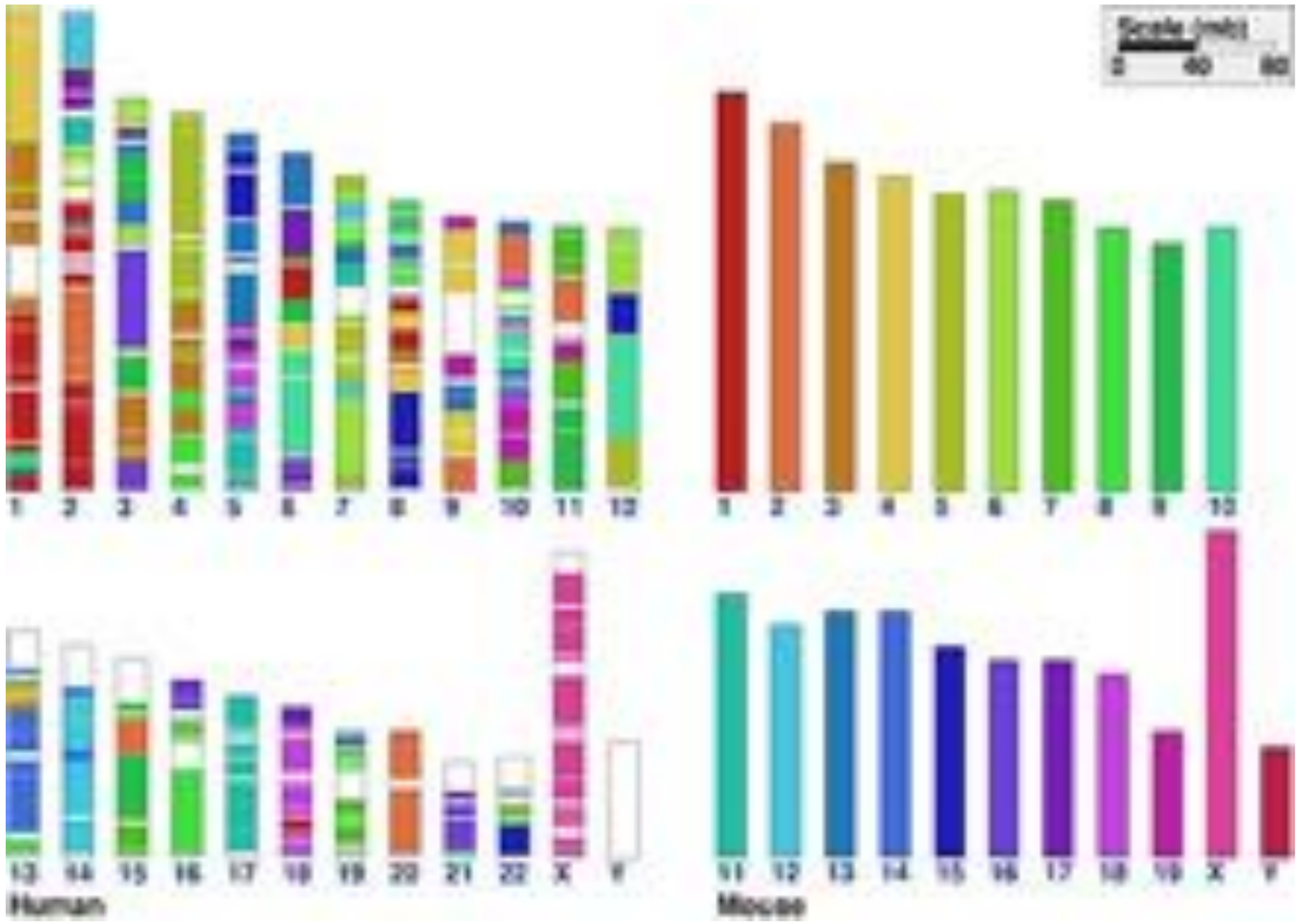
Science 24 Oct 1997:
Vol. 278, Issue 5338, pp. 631-637
DOI: 10.1126/science.278.5338.631

Fig. 1. Examples of COGs. Solid lines show symmetrical BeTs. Broken lines show asymmetrical BeTs, with color corresponding to the species for which the BeT is observed. Genes from the same species are adjacent; otherwise the gene names are positioned arbitrarily. A unique COG ID is indicated in the upper left corner. (**A**) Congruent BeTs form a triangle, the minimal COG. Origin of the proteins: KatG, *E. coli*; sll1987, *Synechocystis* sp.; and YKR066c, *S. cerevisiae*. Note that all the BeTs are symmetrical. (**B**) A simple COG with two

Сравнительная/эволюционная геномика

Синтения

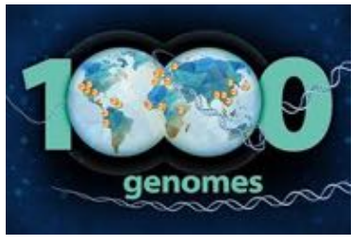
synteny -- the conservation of blocks of order within two sets of chromosomes that are being compared with each other.



Synteny between human and mouse chromosomes. Colors indicate homologous regions. For instance, sequences homologous to mouse chromosome 1 are primarily on human chromosomes 1 and 2, but also 6,8, and 18. The X chromosome is almost completely syntenic in both species

Проекты постгеномной эры

Вариация ДНК в популяции



2008-2012

dpSNP

Связь генотипа, фенотипа и заболеваний

dbGAP
GWAS
Персонифицированная медицина

OMIM
ClinVar

Что значит ДНК и как она работает?

2003-

Соматические мутации

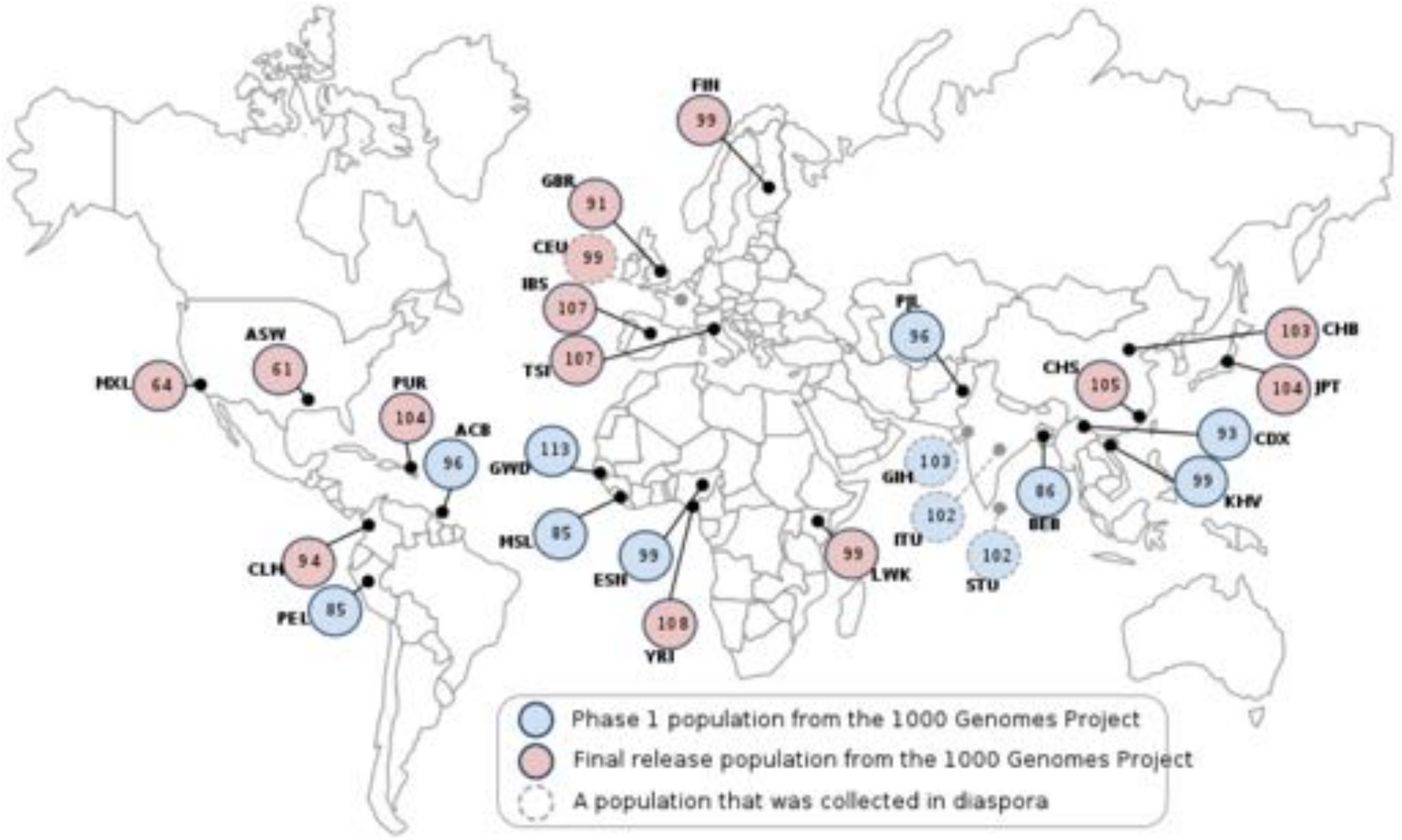
Секвенирование живых организмов

Метагеномика

Проекты постгеномной эры

Вариация ДНК в популяции

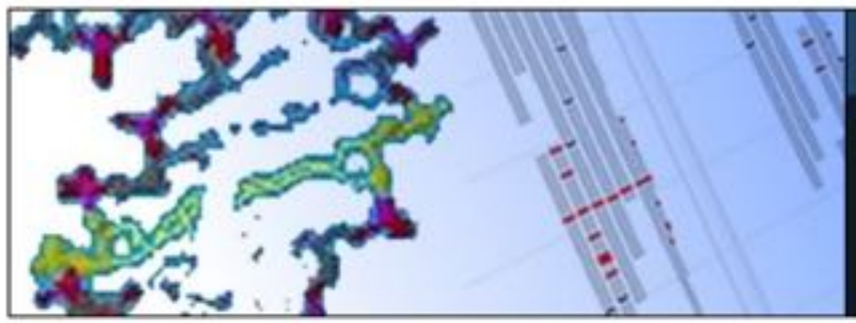
2008-2012



dbSNP

Вариация ДНК в популяции

https://www.ncbi.nlm.nih.gov/snp



dbSNP

dbSNP contains human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations.

TAS2R38 gene:

Linked to the taste of bitter in broccoli, Brussels sprouts, cabbage, watercress, chard, ethanol, and PROP. [ref]

rs713598(G;G)

Possibly unable to taste bitter in some foods. Depending on your other SNPs, this might cause the inability to taste the bitterness of phenylthiocarbamide (PTC) and similar molecules in foods (like cabbage and raw broccoli) or drinks (like coffee and dark beers). That would make cabbage etc. taste horribly bland and boring. On the other hand, fruit from the tropical bignay tree would taste unpleasantly bitter to you instead of sweet. You might eat less healthily.

dbSNP

Вариация ДНК в популяции

https://www.ncbi.nlm.nih.gov/snp

Reference SNP (refSNP) Cluster Report: rs713598 ** With drug-response allele **

RefSNP	
Organism:	human (Homo sapiens)
Molecule Type:	Genomic
Created/Updated in build:	86/151
Map to Genome Build:	108/Weight 1
Validation Status:	
Citation:	PubMed LitVar
Association:	NHGRI GWAS PheGenI

Allele	
Variation Class:	SNV: single nucleotide variation
RefSNP Alleles:	C/G (REV)
Allele Origin:	C:germline G:germline
Ancestral Allele:	C
Variation Viewer:	
Clinical Significance:	With drug-response allele [ClinVar]
MAF/MinorAlleleCount:	G=0.4459/54099 (ExAC) C=0.4952/2480 (1000 Genomes) G=0.4306/5600 (GO-ESP) G=0.4628/58111 (TOPMED)

HGVS Names	Links
CM000669.2:g.141973545C>G NC_000007.13:g.141673345C>G NC_000007.14:g.141973545C>G NG_016141.1:g.5229G>C NM_176817.4:c.145G>C NP_789787.4:p.Ala49Pro NW_003571040.1:g.115496C>G	

SNP Details are organized in the following sections:

dbSNP

Вариация ДНК в популяции

Study: The 1000 Genomes Project (phase 3	
Go to Selection / Scroll Region	141,673,345 rs713598
Populations / Samples	C=0.4952 G=0.5048
ACB African Carribbeans in ...	C=0.5365 G=0.4635
ASW Americans of African A...	C=0.4426 G=0.5574
BEB Bengali from Bangladesh	C=0.6802 G=0.3198
CDX Chinese Dai in Xishuan...	C=0.2849 G=0.7151
CEU Utah Residents (CEPH) ...	C=0.6162 G=0.3838
CHB Han Chinese in Bejing, ...	C=0.3252 G=0.6748
CHS Southern Han Chinese	C=0.3190 G=0.6810
CLM Colombians from Medell...	C=0.4362 G=0.5638
ESN Esan in Nigeria	C=0.5202 G=0.4798
FIN Finnish in Finland	C=0.6263 G=0.3737
GBR British in England and S...	C=0.6099 G=0.3901
GIH Gujarati Indian from Hou...	C=0.5971 G=0.4029
GWD Gambian in Western Di...	C=0.5841 G=0.4159
IBS Iberian Population in Spain	C=0.5514 G=0.4486
ITU Indian Telugu from the UK	C=0.6814 G=0.3186

ClinVAR

NCBI Resources How To

ClinVar ClinVar Search ClinVar for gene symbols, HGVS expressions, conditions, and n

Advanced

Home About Access Help Submit Statistics FTP

NEW Click here to see the new Variation Report design!

NM_176817.4(TAS2R38):c.145G>C (p.Ala49Pro)

Variation ID: 2904

Review status: ☆ ☆ ☆ ☆ (0/4) no assertion criteria provided

Interpretation

Go to:

Clinical significance: drug response

Last evaluated: Dec 30, 2010

Number of submission(s): 1

Condition(s): Phenylthiocarbamide tasting [MedGen]

See supporting ClinVar records

Assertion and evidence details

Go to:

Clinical assertions | Summary evidence | Supporting observations

Germline

Filter:

Clinical significance (Last evaluated)	Review status (Assertion method)	Collection method	Condition(s) (Mode of inheritance)	Origin	Citations	Submitter - Study name	Submission accession
drug response (Dec 30, 2010)	no assertion criteria provided	literature only	Phenylthiocarbamide tasting [MedGen]	germline	· PubMed (1) [See all records that cite this PMID]	OMIM	SCV000023196.1

OMIM

50 YEARS OMIM
Human Genetics Knowledge for the World

*607751
Table of Contents

- Title
- Gene-Phenotype Relationships
- Text
  - Description
  - Cloning and Expression
  - Gene Structure
  - Mapping
  - Gene Function
  - Molecular Genetics
  - Evolution
- Allelic Variants
  - Table View
- References
- Contributors
- Creation Date
- Edit History

* 607751

TASTE RECEPTOR, TYPE 2, MEMBER 38; TAS2R38

Alternative titles; symbols

TRANSFORMING GROWTH FACTOR BETA-STIMULATED CLONE 22; TSC22
PTC
T2R61

HGNC Approved Gene Symbol: TAS2R38

Cytogenetic location: 7q34 ***Genomic coordinates (GRCh38): 7:141,972,630-141,973,772*** (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
7q34	[Phenylthiocarbamide tasting]	171200	AD	3

PheneGene Graphics

▼ ALLELIC VARIANTS (3 Selected Examples):

Table View | ClinVar

.0001 PHENYLTHIOCARBAMIDE TASTING

TAS2R38, ALA49PRO dbSNP:rs713598 RCV000003038
Within the PTC gene, Kim et al. (2003) found 3 common polymorphisms that influence the ability to taste phenylthiocarbamide (see 171200). One was a 145G-C transversion, resulting in an ala49-to-pro (A49P) substitution (rs713598).

.0002 PHENYLTHIOCARBAMIDE TASTING

TAS2R38, VAL262ALA dbSNP:rs1726866 RCV000003039
Certain haplotypes of polymorphisms within the PTC gene account for the ability to taste or not taste phenylthiocarbamide (see 171200). Kim et al. (2003) found one of these to be a 785T-C transition, resulting in a val262-to-ala (V262A) substitution (rs1726866).

.0003 PHENYLTHIOCARBAMIDE TASTING

TAS2R38, ILE296VAL dbSNP:rs10246939 RCV000003040
Kim et al. (2003) identified an 886A-G transition in the PTC gene, resulting in an ile296-to-val (I296V) substitution (rs10246939). This polymorphism, in conjunction with other SNPs in the gene, give rise to the ability to taste or not taste phenylthiocarbamide (see 171200).

TEXT

▼ Description

TAS2R38 belongs to the large TAS2R receptor family. TAS2Rs are expressed on the surface of taste receptor cells and mediate the perception of bitterness through a G protein-coupled second messenger pathway (summary by Conte et al., 2002). For further information on the TAS2R gene family, see 604791.

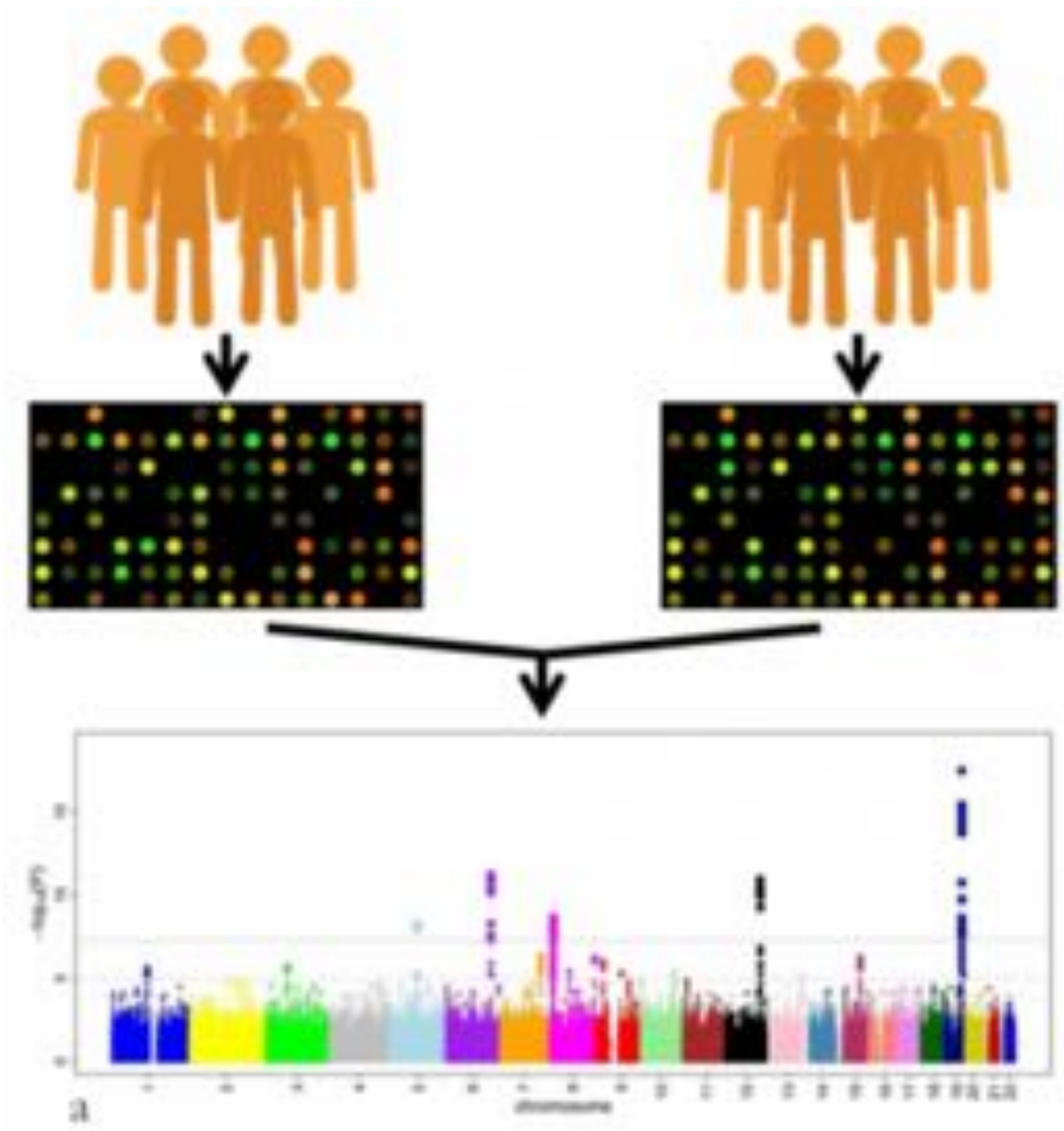
Гаплотипы

A **haplotype** (haploid genotype) is a group of alleles in an organism that are inherited together from a single parent.

Genetic linkage is the tendency of DNA sequences that are close together on a chromosome to be inherited together during the meiosis phase of sexual reproduction.

GWAS

Genome-wide association study



Manhattan plot

Trait
Sample size
Ancestry

Genotyping platform
Number of SNPs analysed

Results
- variant
- risk allele
- p-value
- effect size

dpGAP

https://www.ncbi.nlm.nih.gov/gap

CIDR: Collaborative Study on the Genetics of Alcoholism Case Control Study

dbGaP Study Accession: phs000125.v1.p1

Request Access

Show BioProject list

Study | Variables | Documents | Analyses | Datasets | Molecular Data

Analysis Name and Accession

Name: GWAS for alcohol dependence in European-Americans
Accession: pha002892.1

View association results in Genome Browser

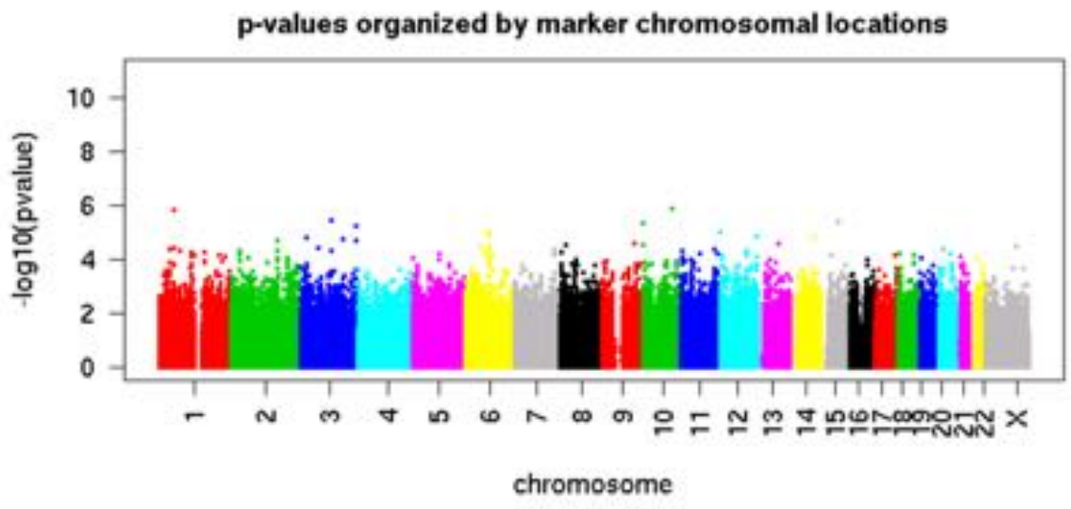
Analysis Description

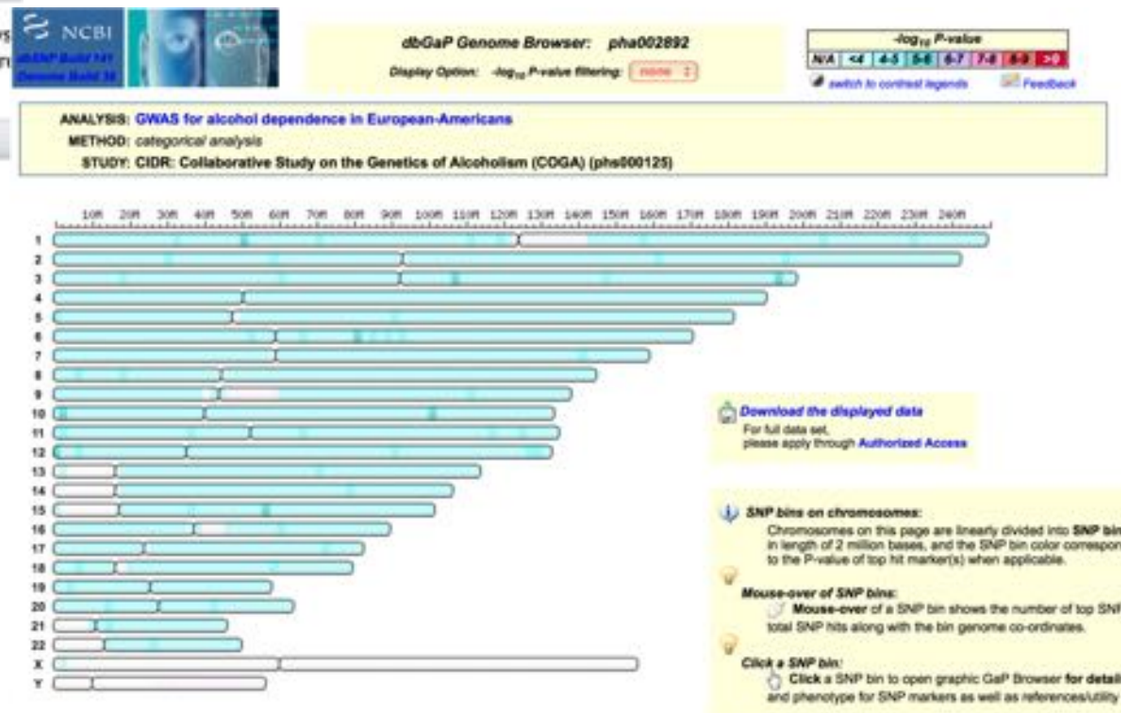
To identify common variants underlying alcohol-dependence, probands were ascertained through alcohol treatment programs and evalu
and their relatives were administered a validated poly-diagnostic instrument, the Semi- Structured Assessment for the Genetics of Alco
performed by the Center for Inherited Disease Research (CIDR). DNA sources included blood (n = 1453) and lymphoblastoid cell lines

Analysis Methods

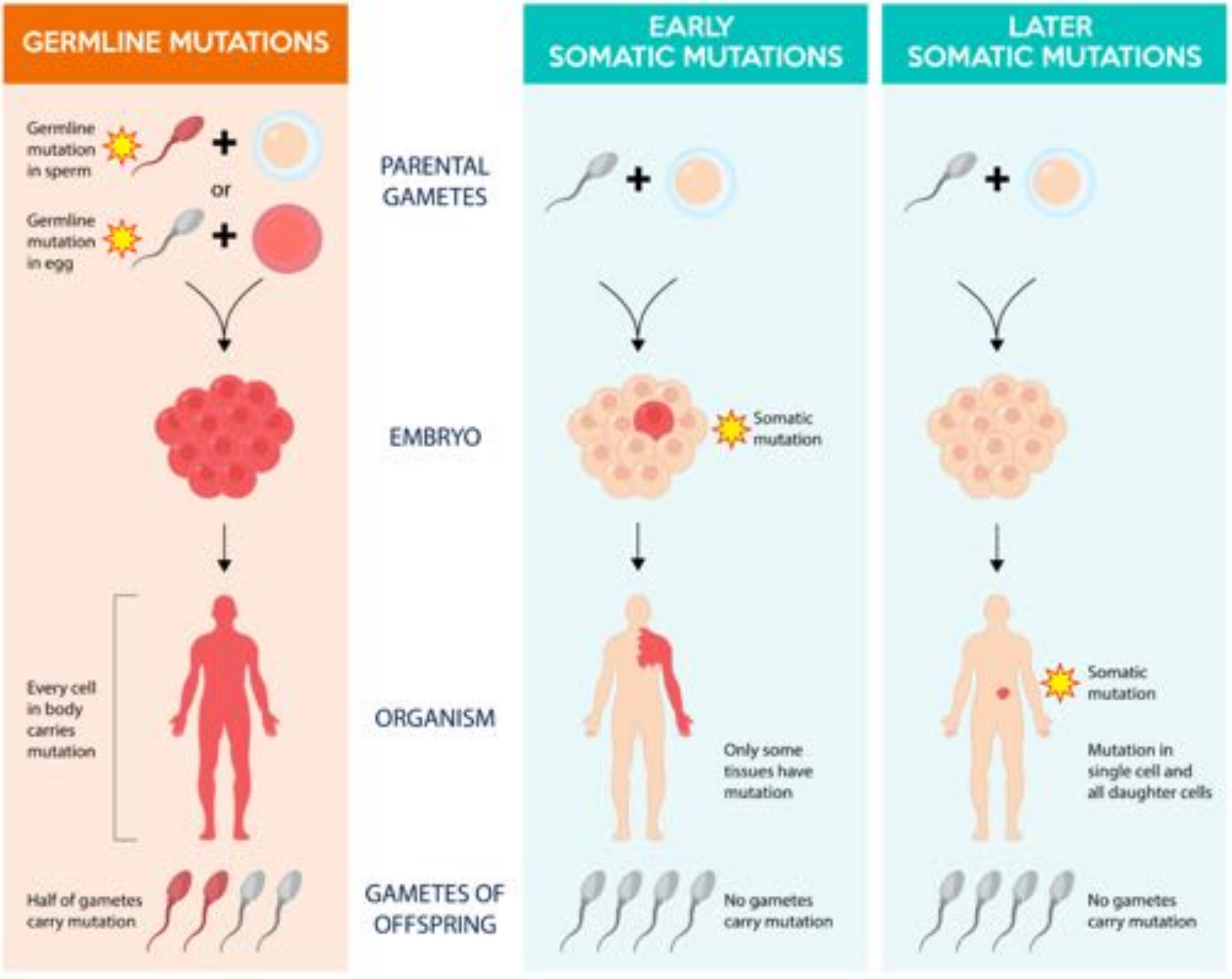
Sample QC filters consisted of 98% genotyping completeness, unrelatedness of subjects (n = 9). A principal component-based analys
either the European or African-American groups and were excluded. EA and AA SNP QC filters (applied separately in each group) wer
was performed in PLINK adjusting for sex.

Analysis Plots

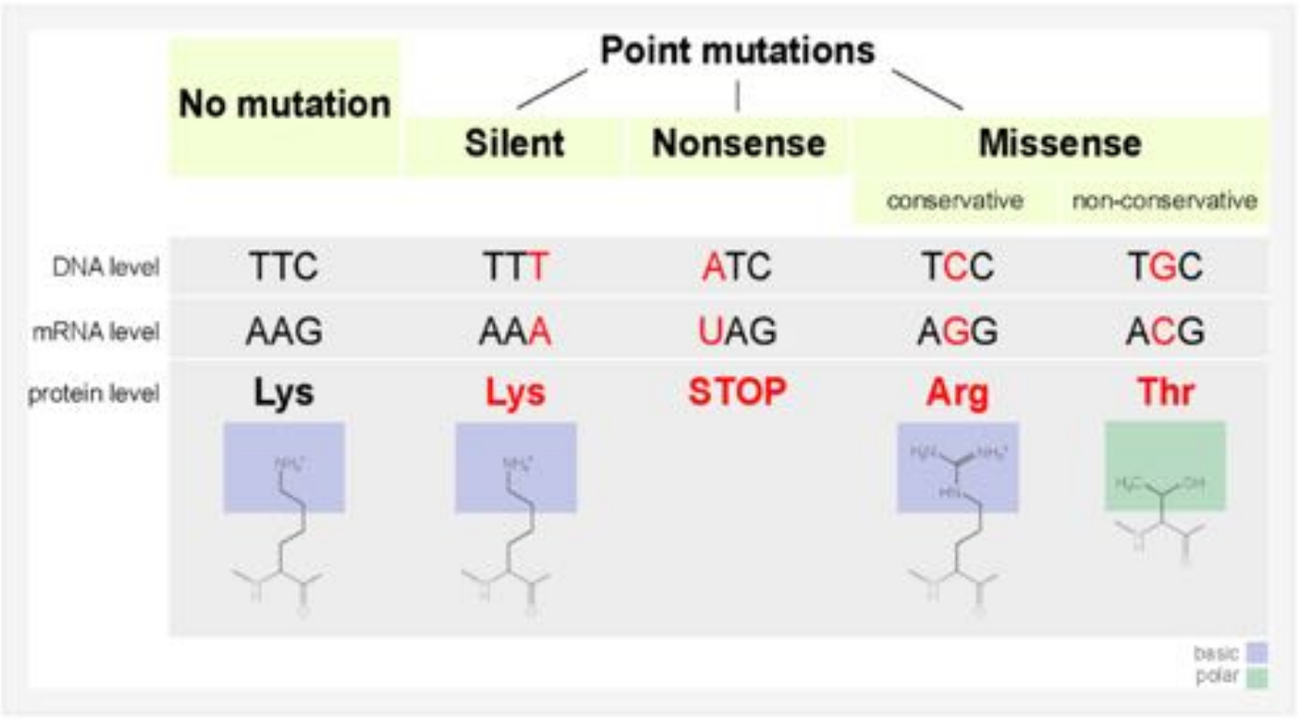


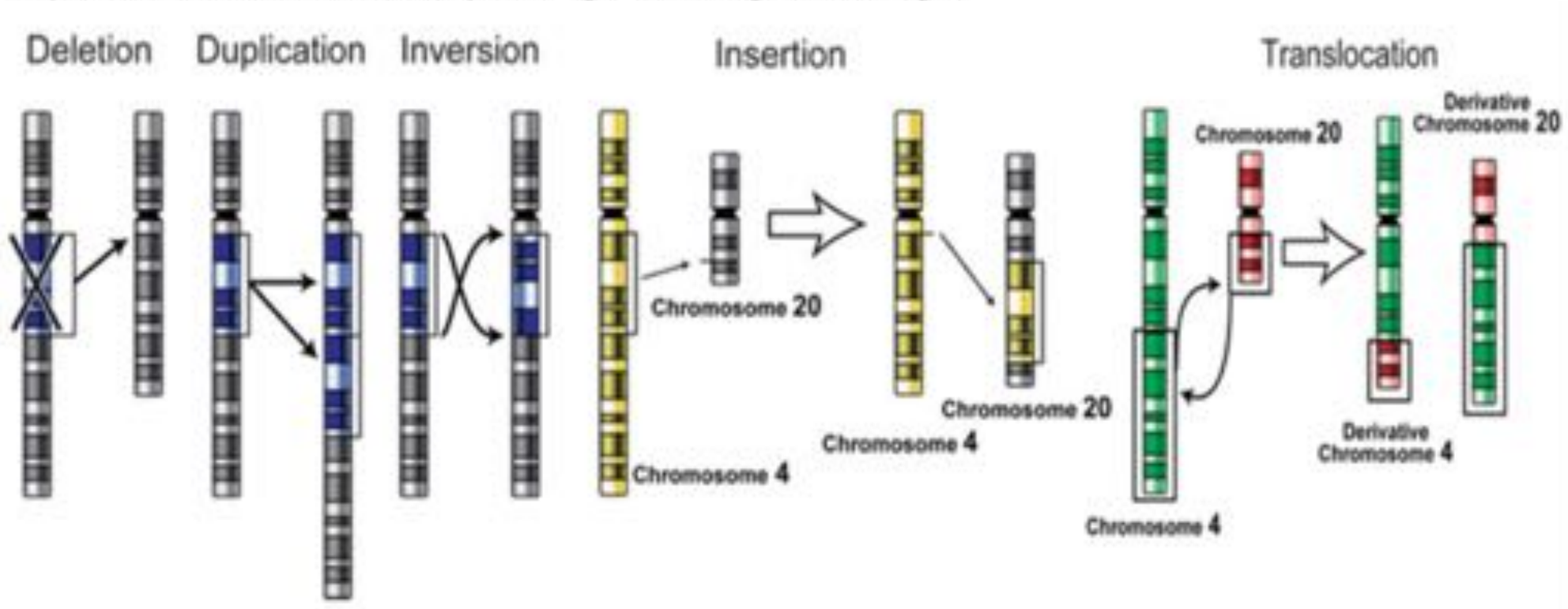


Мутации


GERMLINE MUTATIONS
EARLY SOMATIC MUTATIONS
LATER SOMATIC MUTATIONS
Germline mutation in sperm
or
Germline mutation in egg
PARENTAL GAMETES
EMBRYO
Somatic mutation
Every cell in body carries mutation
ORGANISM
Only some tissues have mutation
Somatic mutation
Mutation in single cell and all daughter cells
Half of gametes carry mutation
GAMETES OF OFFSPRING
No gametes carry mutation
No gametes carry mutation

Мутации


Point mutations
No mutation
Silent
Nonsense
Missense
conservative
non-conservative
DNA level
mRNA level
protein level
TTC
TTT
ATC
TCC
TGC
AAG
AAA
UAG
AGG
ACG
Lys
Lys
STOP
Arg
Thr
basic
polar


Deletion
Duplication
Inversion
Insertion
Translocation
Chromosome 20
Chromosome 4
Chromosome 20
Chromosome 4
Chromosome 20
Chromosome 4
Derivative Chromosome 20
Derivative Chromosome 4

TCGA, ICGC, COSMIC

TCGA, ICGC, COSMIC

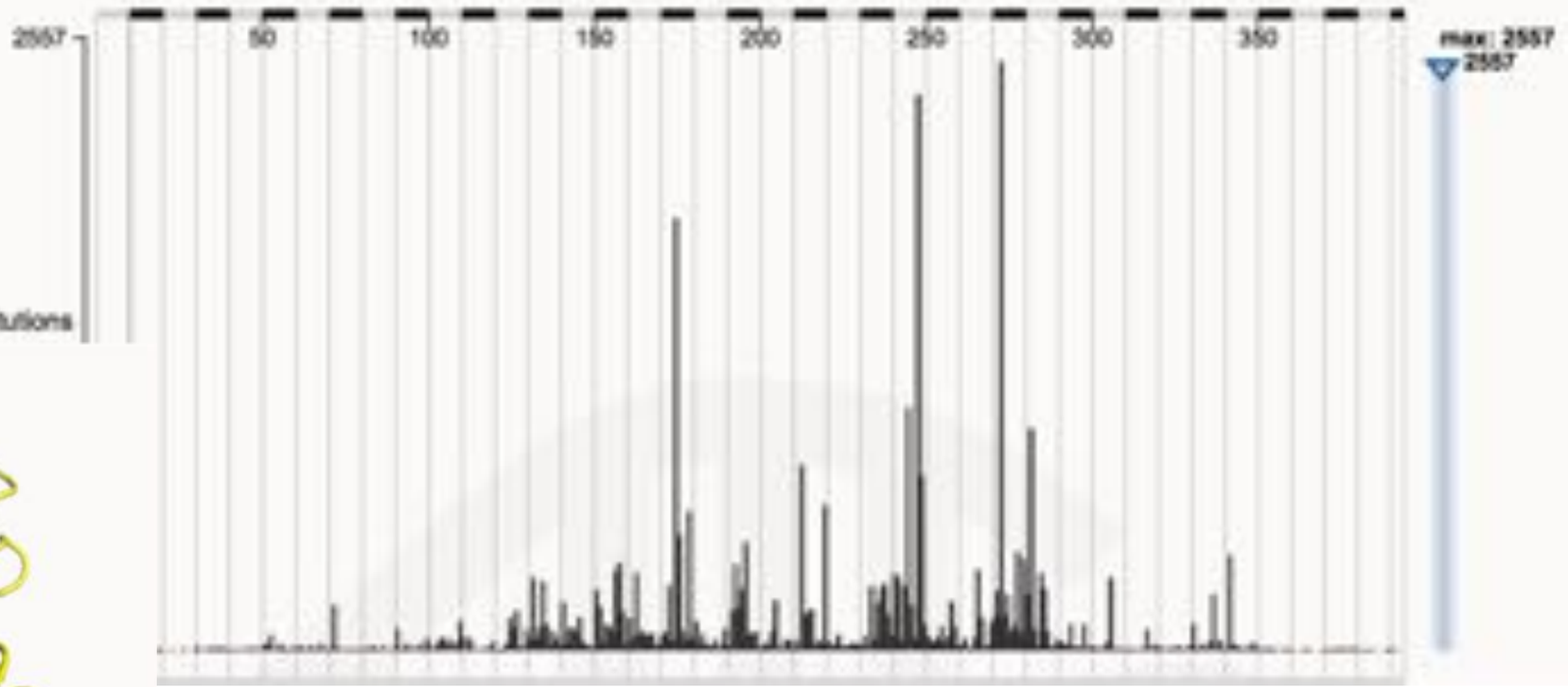
Gene

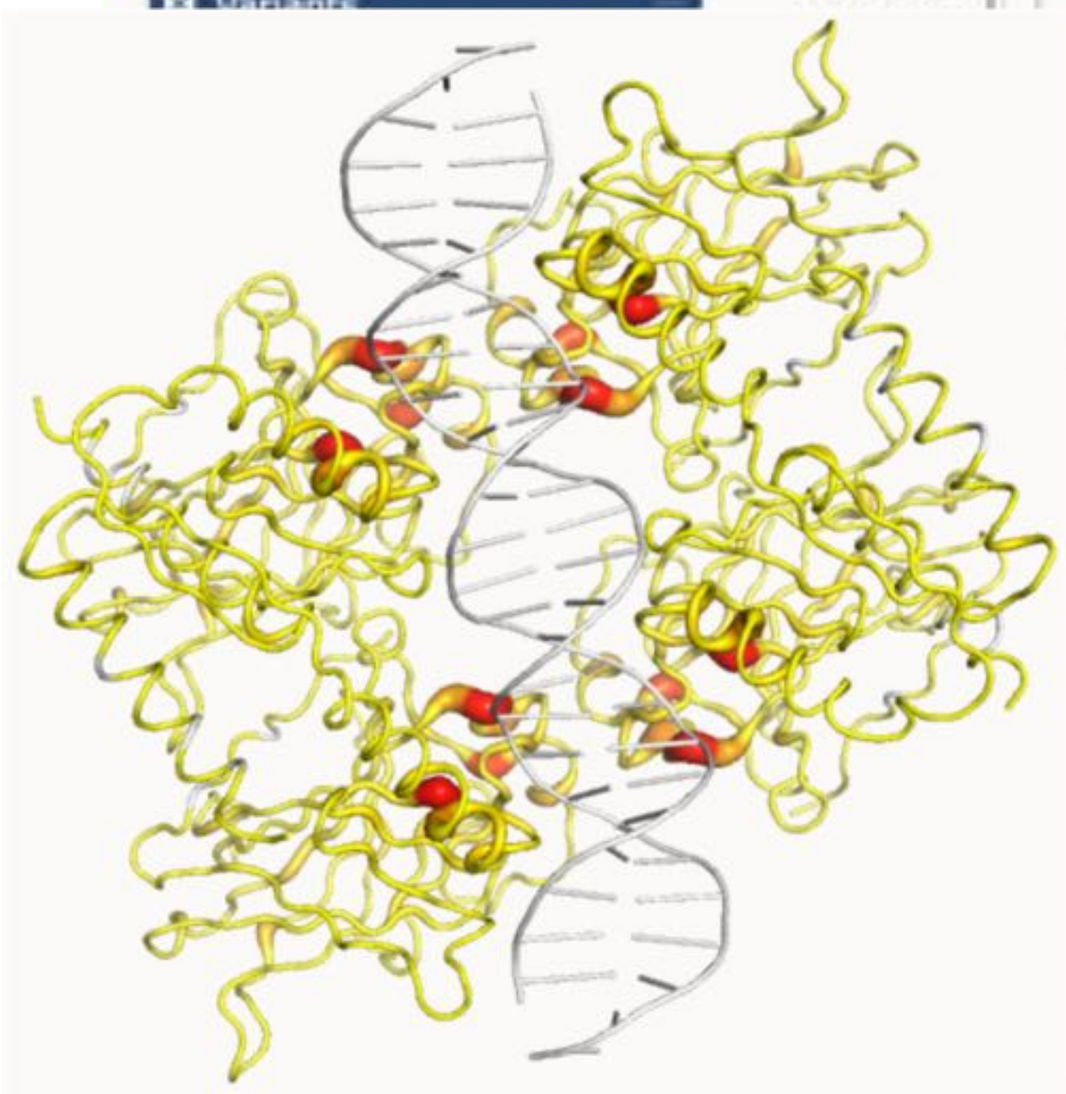
TP53

- Gene view
- Overview
- External links
- Drug resistance
- Tissue distribution
- Genome browser
- Mutation distribution
- Variants

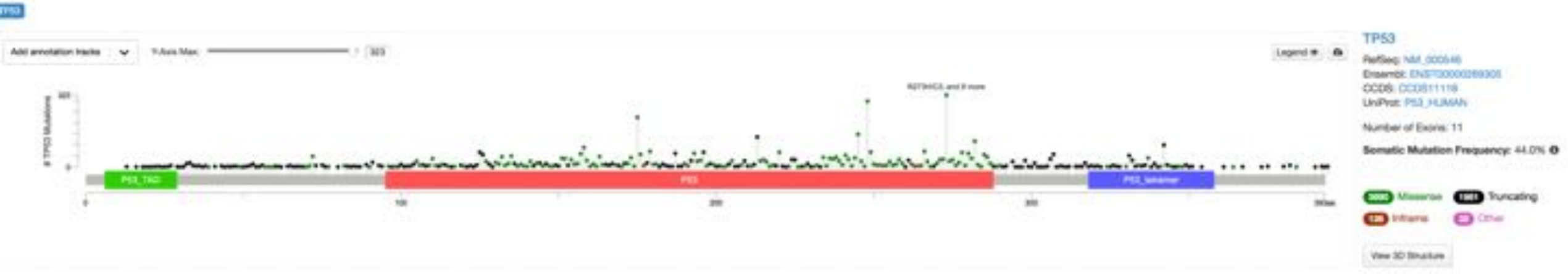
Gene view

The gene view histogram is a graphical view of mutations across TP53. These mutations are displayed
the histogram to highlight the region of interest, or by using the sliders in the filters panel to the left.





cBioPortal


TP53
Add annotation tracks
Y-Axis Max:
Legend
TP53
Number of Exons: 11
Somatic Mutation Frequency: 44.0%
Missense
Truncating
Inframe
Other
View 3D Structure

SNPedia

SNPedia

Talk Contributions Create account Log in

Search SNPedia

Navigation

Main page Discussion Edit History

Have questions? Visit https://www.reddit.com/r/SNPedia

SNPedia

SNPedia is a wiki investigating human genetics. We share information about the effects of variations in DNA, citing peer-reviewed scientific publications. It is used by Promethease to create a personal report linking your DNA variations to the information published about them. Please see the SNPedia:FAQ for answers to common questions.

Help! [edit]

- look at the example rs1234
- learn more about SNPs
- browse
  - genes
  - genosets
  - genotypes
  - medicines
  - medical conditions
  - topics

Popular [edit]

- rs53576 in the oxytocin receptor influences social behavior and personality
- rs1815739 muscle performance
- rs7412 and rs429358 can raise the risk of Alzheimer's disease by more than 10x
- rs6152 can influence baldness
- rs333 resistance to HIV
- rs1800497 in a dopamine receptor may influence the sense of pleasure

Предсказание эффекта мутаций

PolyPhen-2 prediction of functional effects of human nsSNPs

Home | About | Help | Downloads | Batch query | WHESS.db

Sunyaev Lab

PolyPhen-2 report for P59533 A49P (rs713598)

Query

Protein Acc	Position	AA_1	AA_2	Description
P59533	49	A	P	Canonical; RecName: Full=Taste receptor type 2 member 38; Short=T2R38; AltName: Full=PTC bitter taste receptor; AltName: Full=Taste receptor type 2 member 61; Short=T2R61; Length: 333

Results

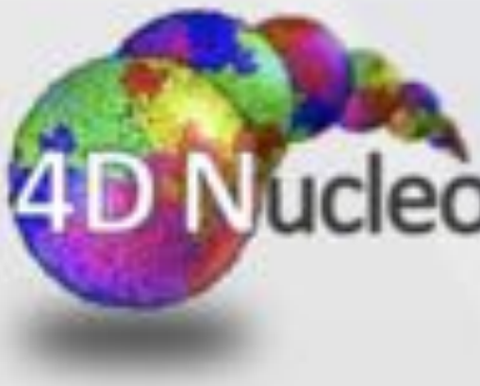
Prediction/Confidence

HumDiv

This mutation is predicted to be **BENIGN** with a score of **0.001** (sensitivity: **0.99**; specificity: **0.15**)

HumVar

Finishing the Job:
Understanding Genome Organization

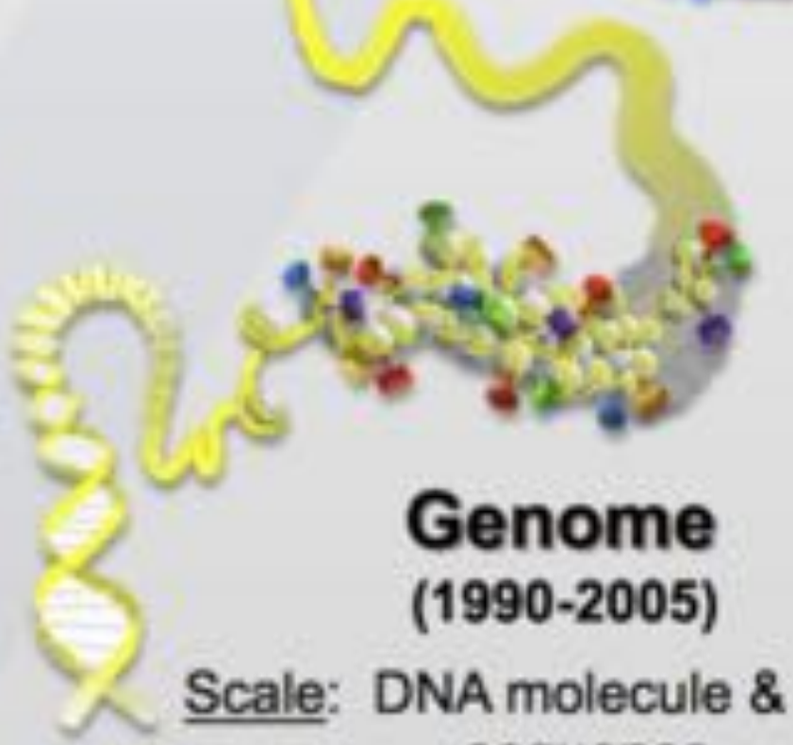


3D Nucleome
(2015-2022?)

Scale: cell nucleus & chromatin domains

ENCODE

ROADMAP epigenomics PROJECT



Epigenome
(2005-2015)

Scale: nucleosome & epigenetic marks

Human Genome Project

Genome
(1990-2005)

Scale: DNA molecule & sequence

NIH National Institutes of Health
Office of Strategic Coordination - The Common Fund

GENOMICS

ENCODE Project Writes Eulogy For Junk DNA

SCIENCE VOL 337 7 SEPTEMBER 2012

ENCODE By the Numbers

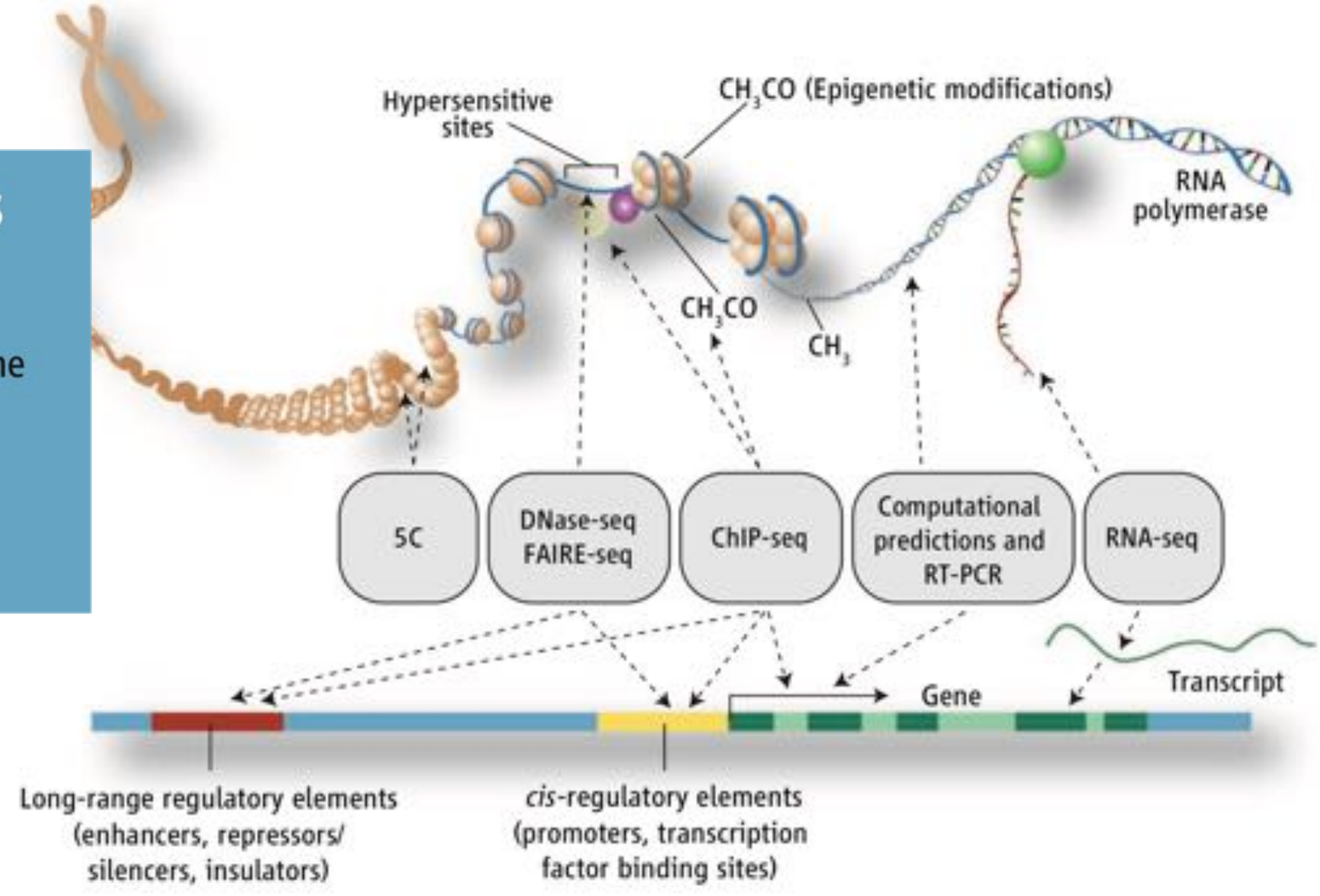
147 cell types studied

80% functional portion of human genome

20,687 protein-coding genes

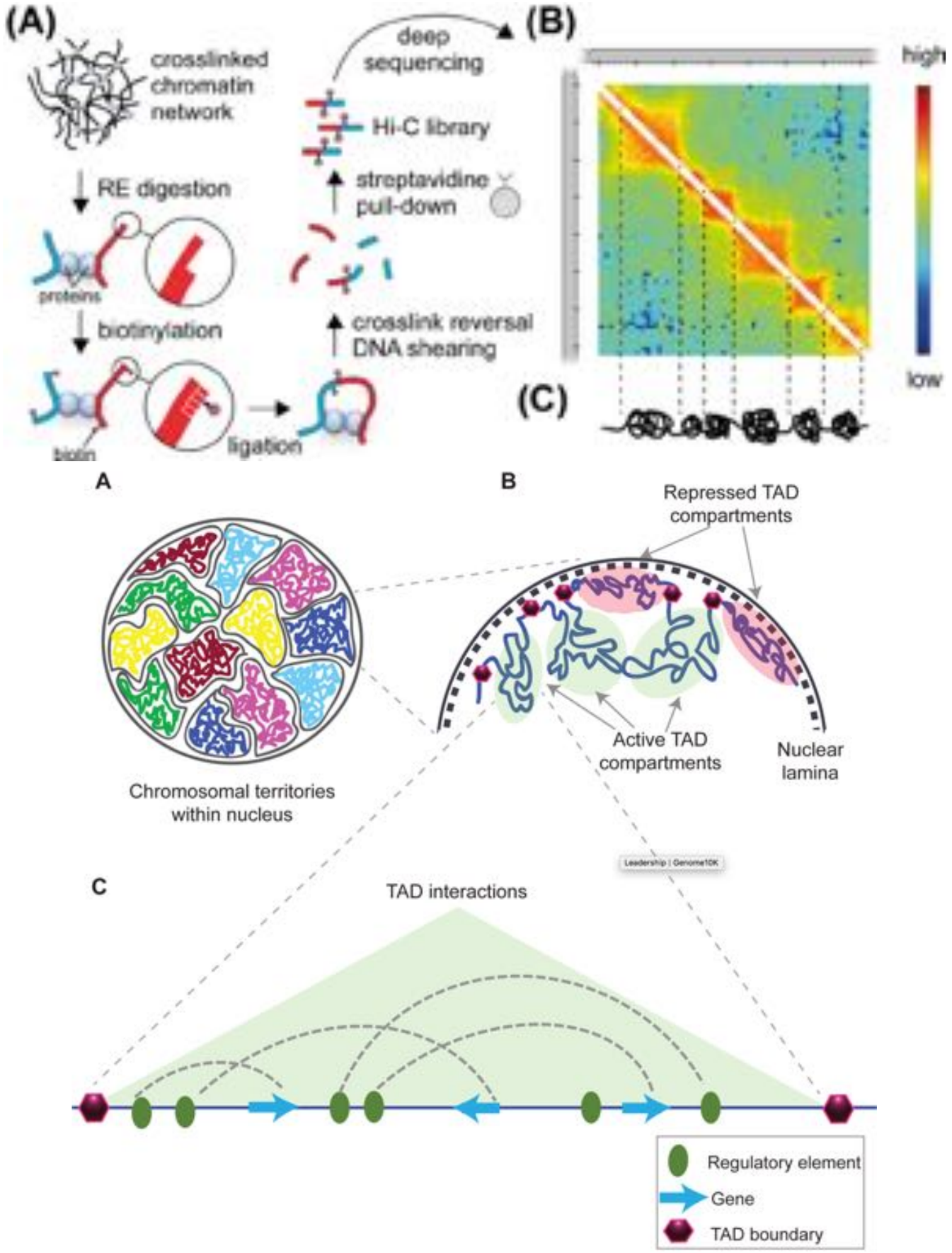
18,400 RNA genes

1640 data sets

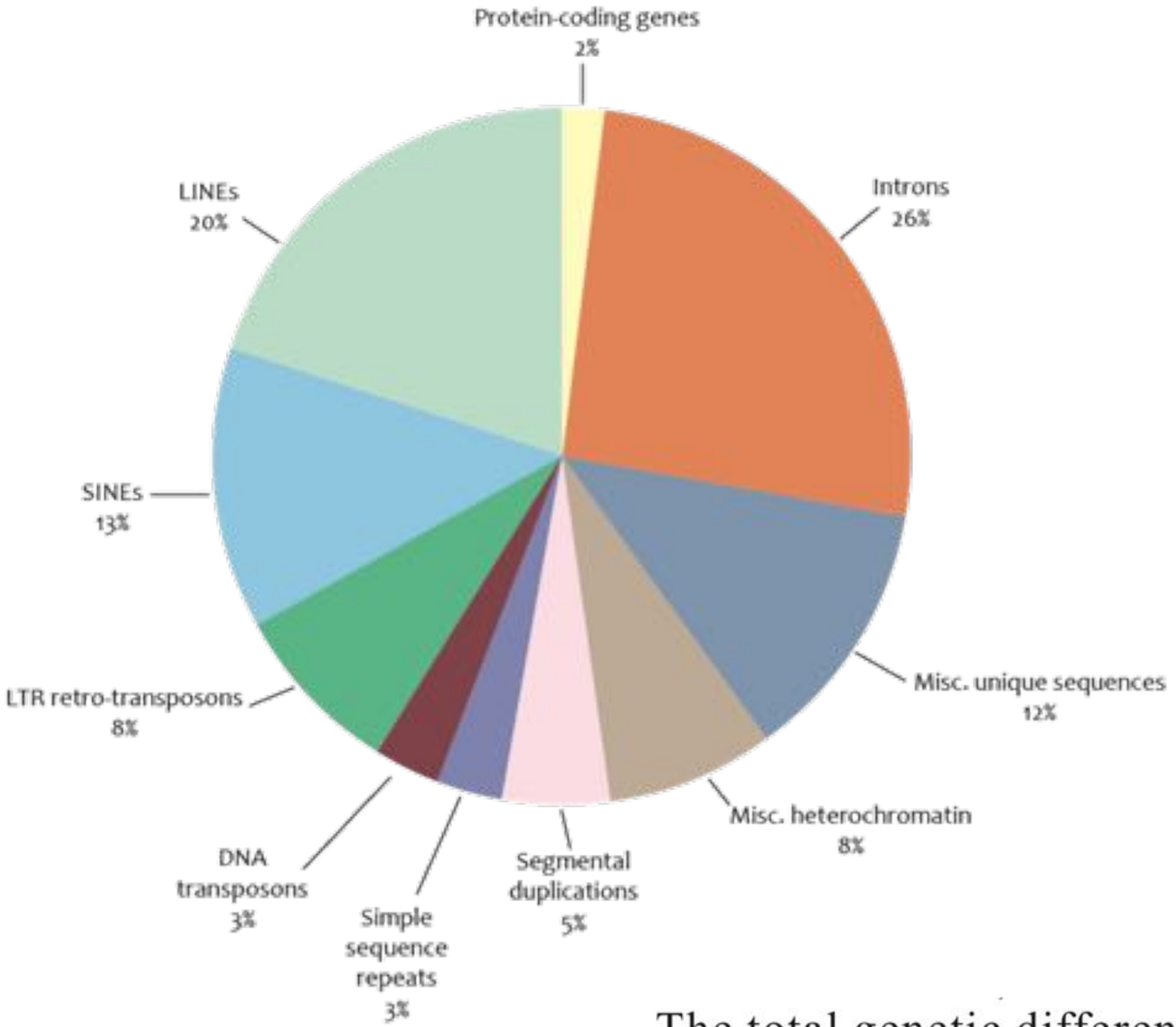


https://science.sciencemag.org/content/337/6099/1159

3D геномика, методы 3C, Hi-C и др.



Структура генома человека и вариации



4 to 5 million SNPs in a person's genome

The total genetic difference between humans and chimps, in terms of number of bases, sums to about 4% of the genome. That

2 SEPTEMBER 2005 VOL 309 SCIENCE

99% identity of the aligned sequence
96% identity between whole genomes

Геномные браузеры

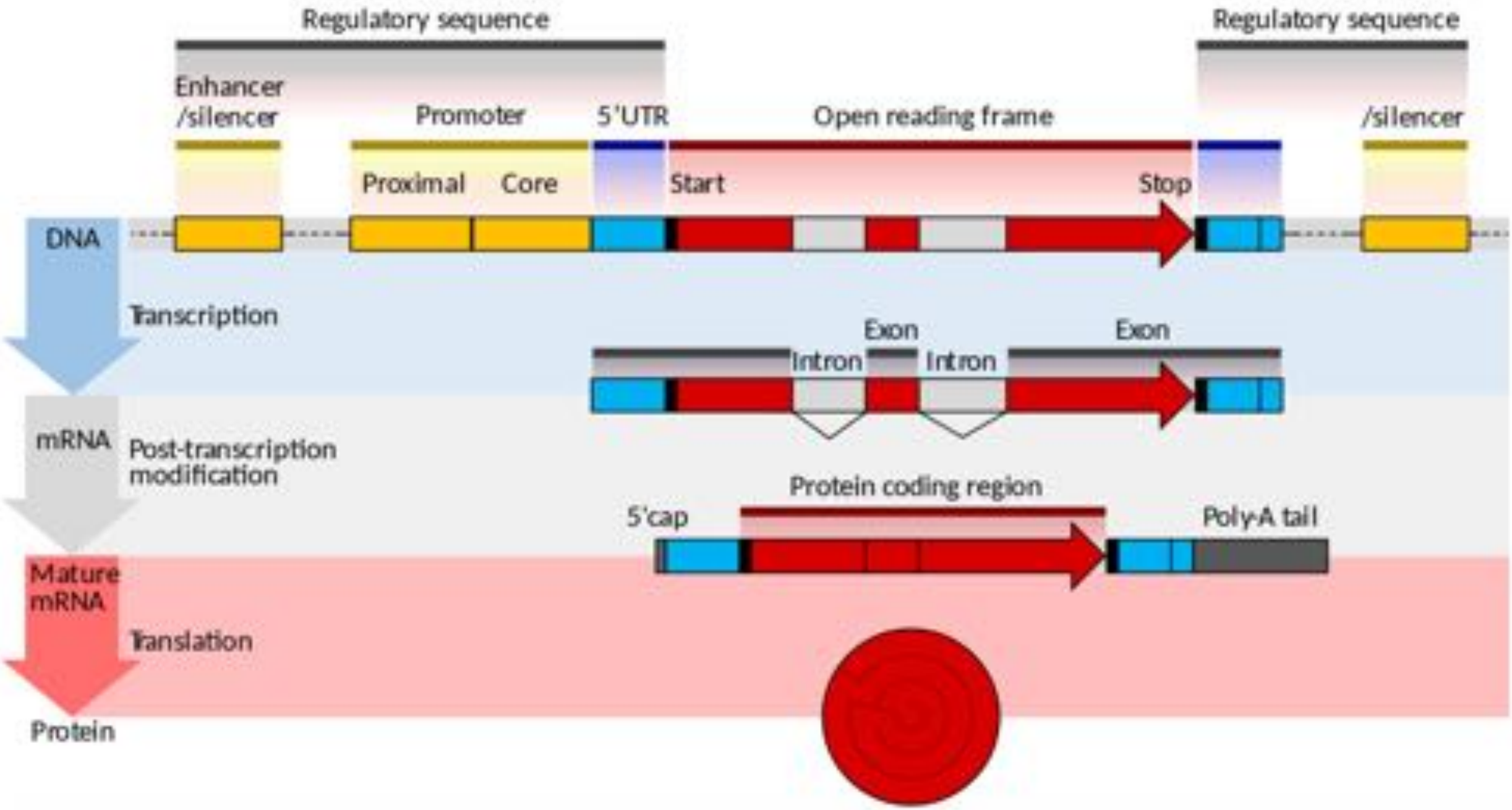
http://www.ensembl.org

http://ensemblgenomes.org

https://www.ncbi.nlm.nih.gov/genome/gdv/

https://genome.ucsc.edu

Структура гена, понятие транскрипта, кДНК



Демонстрация ENSEMBL

Рецепторы горького вкуса в капусте
ген TAS2R38

Рецептор вкуса умами
ген TAS1R3