EKSPERIMENTALNE METODE V BIOKEMIJI IN MOLEKULARNI BIOLOGIJI

Določanje nukleotidnih zaporedij

Sangerjeva metoda, metode druge in tretje generacije

Genomika

Podatkovne zbirke nukleotidnih zaporedij: GenBank

EST kloni

Projekt človeški genom

HUMAN GENOME

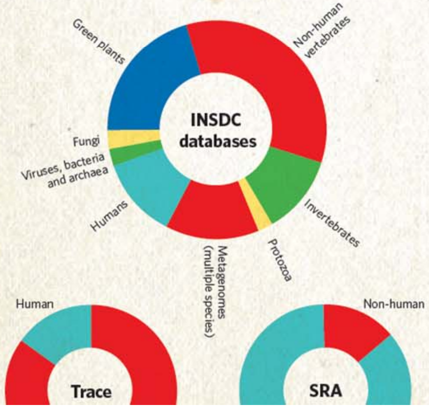
THE SEQUENCE EXPLOSION

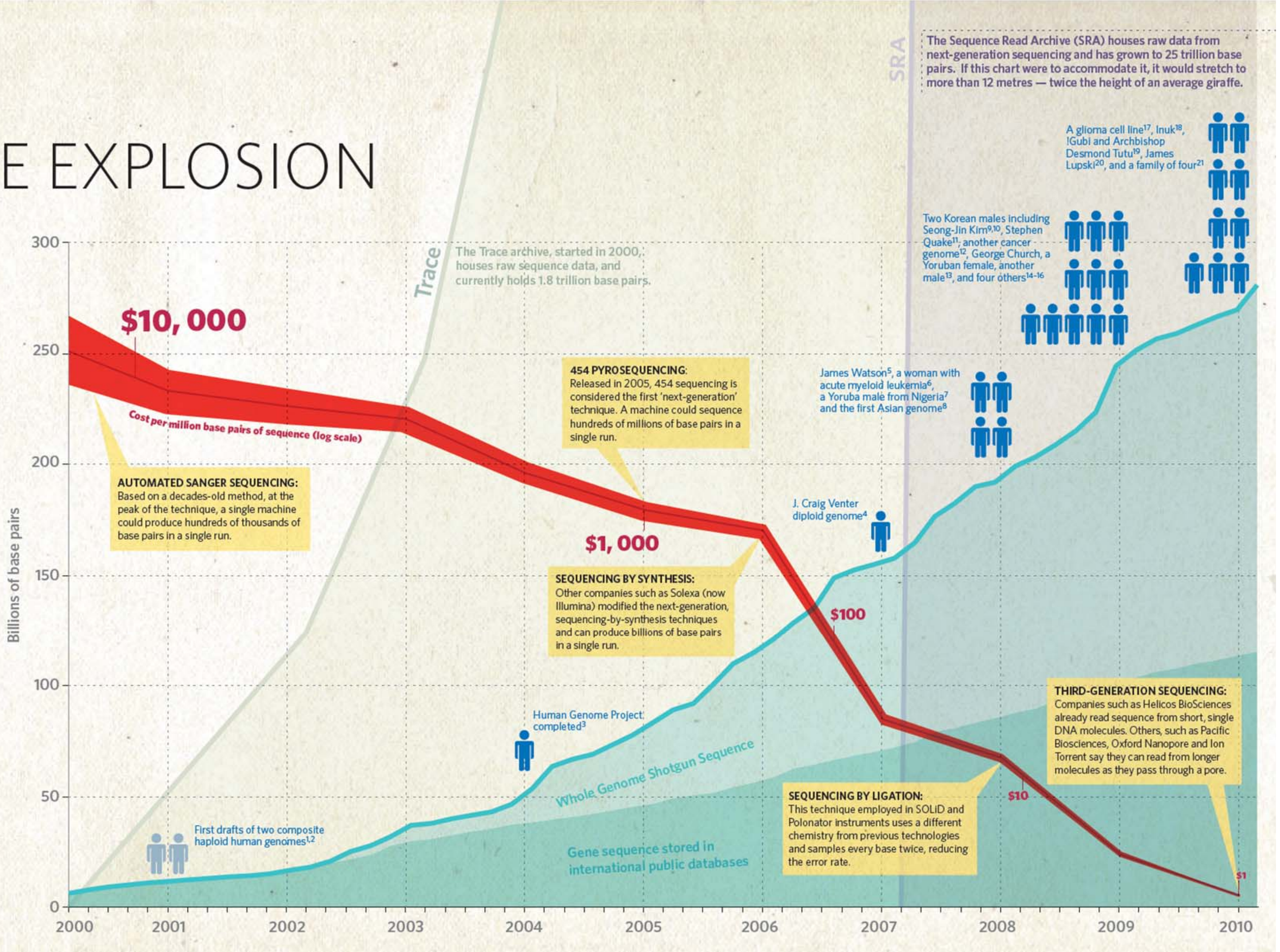
At the time of the announcement of the first drafts of the human genome in 2000, there were 8 billion base pairs of sequence in the three main databases for 'finished' sequence: GenBank, run by the US National Center for Biotechnology Information; the DNA Databank of Japan; and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database. The databases share their data regularly as part of the International Nucleotide Sequence Database Collaboration (INSDC). In the subsequent first post-genome decade, they have added another 270 billion bases to the collection of finished sequence, doubling the size of the database roughly every 18 months. But this number is dwarfed by the amount of raw sequence that has been created and stored by researchers around the world in the Trace archive and Sequence Read Archive (SRA).

See Editorial, page 649, and human genome special at www.nature.com/humangenome

DNA SEQUENCES BY TAXONOMY

International Nucleotide Sequence Database Collaboration: The main repositories of 'finished' sequence span a wide range of organisms, representing the many priorities of scientists worldwide.





The Sequence Read Archive (SRA) houses raw data from next-generation sequencing and has grown to 25 trillion base pairs. If this chart were to accommodate it, it would stretch to more than 12 metres — twice the height of an average giraffe.

The Trace archive, started in 2000, houses raw sequence data, and currently holds 1.8 trillion base pairs.

AUTOMATED SANGER SEQUENCING: Based on a decades-old method, at the peak of the technique, a single machine could produce hundreds of thousands of base pairs in a single run.

454 PYROSEQUENCING: Released in 2005, 454 sequencing is considered the first 'next-generation' technique. A machine could sequence hundreds of millions of base pairs in a single run.

SEQUENCING BY SYNTHESIS: Other companies such as Solexa (now Illumina) modified the next-generation, sequencing-by-synthesis techniques and can produce billions of base pairs in a single run.

SEQUENCING BY LIGATION: This technique employed in SOLiD and Polonator instruments uses a different chemistry from previous technologies and samples every base twice, reducing the error rate.

THIRD-GENERATION SEQUENCING: Companies such as Helicos BioSciences already read sequence from short, single DNA molecules. Others, such as Pacific Biosciences, Oxford Nanopore and Ion Torrent say they can read from longer molecules as they pass through a pore.

First drafts of two composite haploid human genomes[1,2]

Human Genome Project completed[3]

J. Craig Venter diploid genome[4]

James Watson[5], a woman with acute myeloid leukemia[6], a Yoruba male from Nigeria[7] and the first Asian genome[8]

Two Korean males including Seong-Jin Kim[9,10], Stephen Quake[11], another cancer genome[12], George Church, a Yoruban female, another male[13], and four others[14–16]

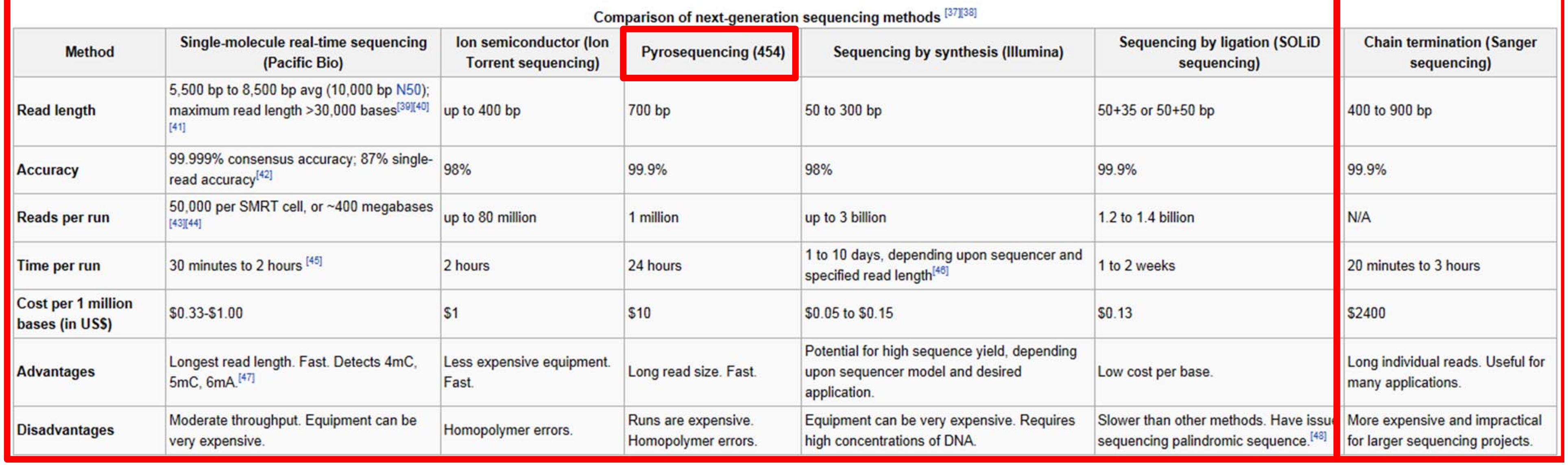
A glioma cell line[17], Inuk[18], !Gubi and Archbishop Desmond Tutu[19], James Lupski[20], and a family of four[21]

Nature, 1. 4. 2010

Primerjava metod za določanje nukleotidnih zaporedij

Metode 2. generacije

Metoda 1. generacije



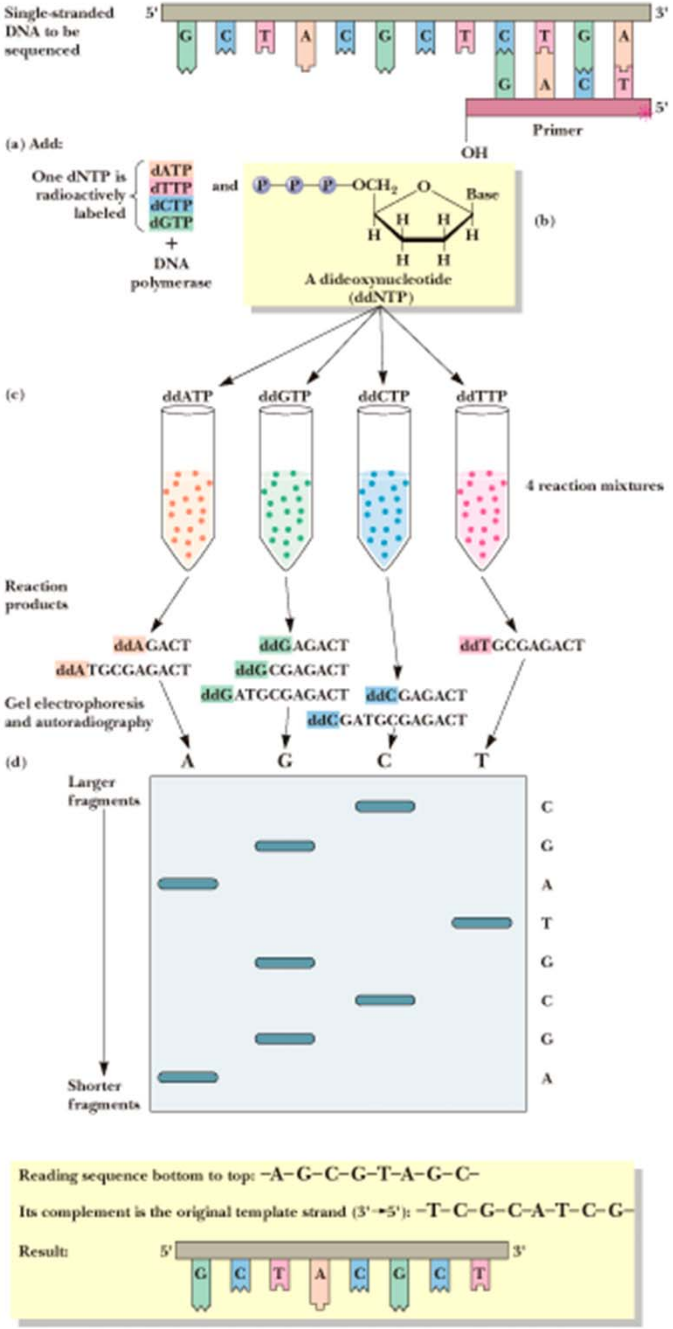
Comparison of next-generation sequencing methods [37][38]

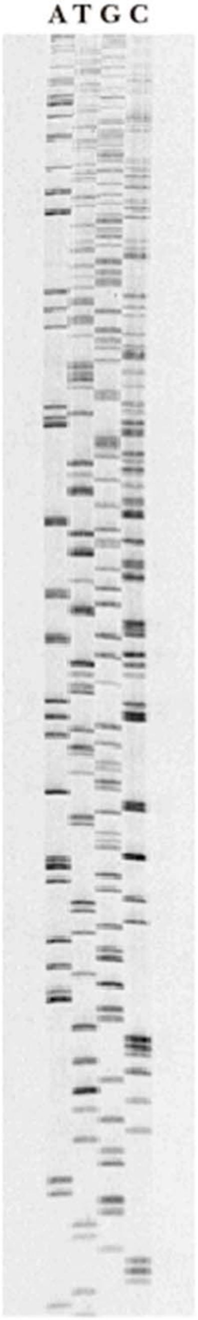
Method	Single-molecule real-time sequencing (Pacific Bio)	Ion semiconductor (Ion Torrent sequencing)	Pyrosequencing (454)	Sequencing by synthesis (Illumina)	Sequencing by ligation (SOLiD sequencing)	Chain termination (Sanger sequencing)
Read length	5,500 bp to 8,500 bp avg (10,000 bp N50); maximum read length >30,000 bases[39][40][41]	up to 400 bp	700 bp	50 to 300 bp	50+35 or 50+50 bp	400 to 900 bp
Accuracy	99.999% consensus accuracy; 87% single-read accuracy[42]	98%	99.9%	98%	99.9%	99.9%
Reads per run	50,000 per SMRT cell, or ~400 megabases [43][44]	up to 80 million	1 million	up to 3 billion	1.2 to 1.4 billion	N/A
Time per run	30 minutes to 2 hours [45]	2 hours	24 hours	1 to 10 days, depending upon sequencer and specified read length[46]	1 to 2 weeks	20 minutes to 3 hours
Cost per 1 million bases (in US$)	$0.33-$1.00	$1	$10	$0.05 to $0.15	$0.13	$2400
Advantages	Longest read length. Fast. Detects 4mC, 5mC, 6mA.[47]	Less expensive equipment. Fast.	Long read size. Fast.	Potential for high sequence yield, depending upon sequencer model and desired application.	Low cost per base.	Long individual reads. Useful for many applications.
Disadvantages	Moderate throughput. Equipment can be very expensive.	Homopolymer errors.	Runs are expensive. Homopolymer errors.	Equipment can be very expensive. Requires high concentrations of DNA.	Slower than other methods. Have issu sequencing palindromic sequence.[48]	More expensive and impractical for larger sequencing projects.

Metode 3. generacije: uporaba nanopor

Wikipedia: DNA sequencing

„Sangerjeva reakcija“ 1. generacija določanja zaporedij („chain terminating“)





The Nobel Prize in Chemistry 1980

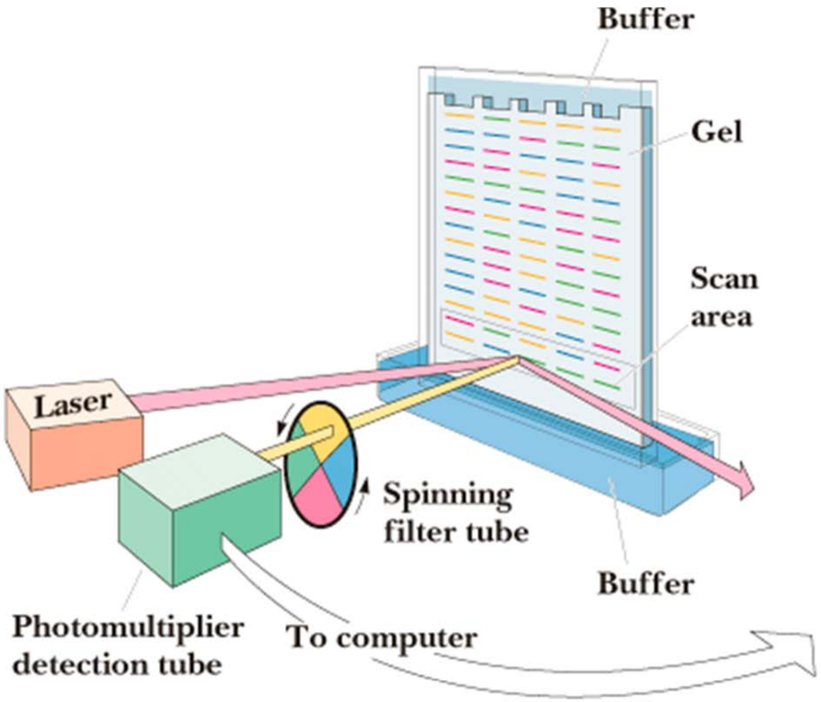
Walter Gilbert

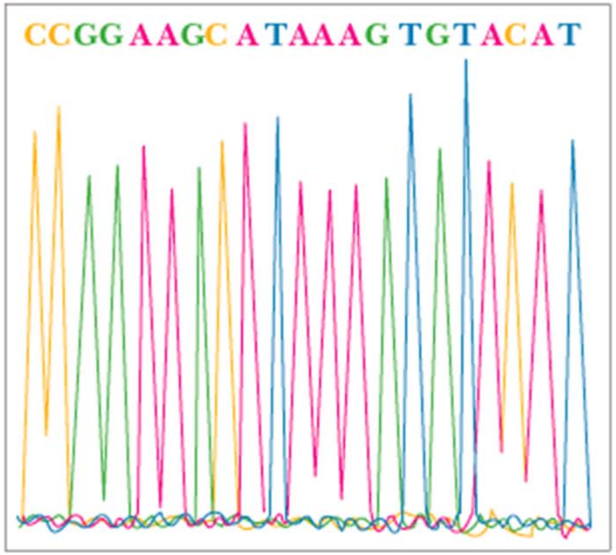
Frederick Sanger

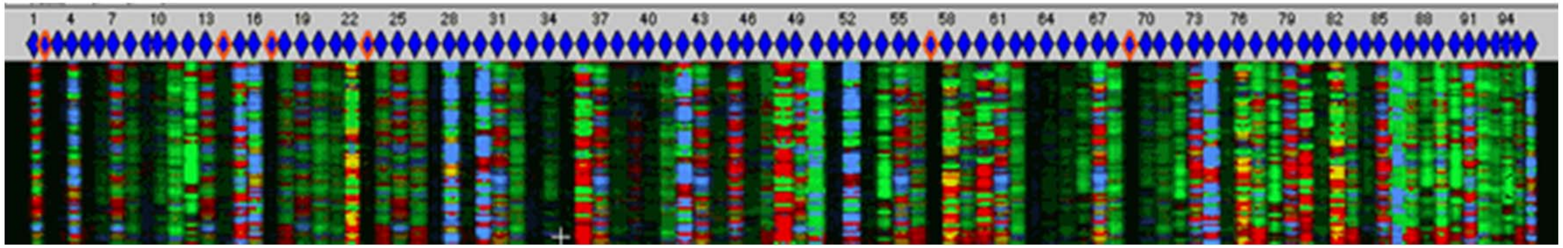
"for their contributions concerning the determination of base sequences in nucleic acids"

Sanger is responsible for the first complete determination of the sequence of a DNA molecule. He has established the sequence of the 5375 building blocks in DNA from a bacterial virus called phi-X174.

Sequence investigations with the methods of Gilbert and Sanger together with the recombinant-DNA technique make excellent tools for continued investigations of the structure and function of the genetic material.





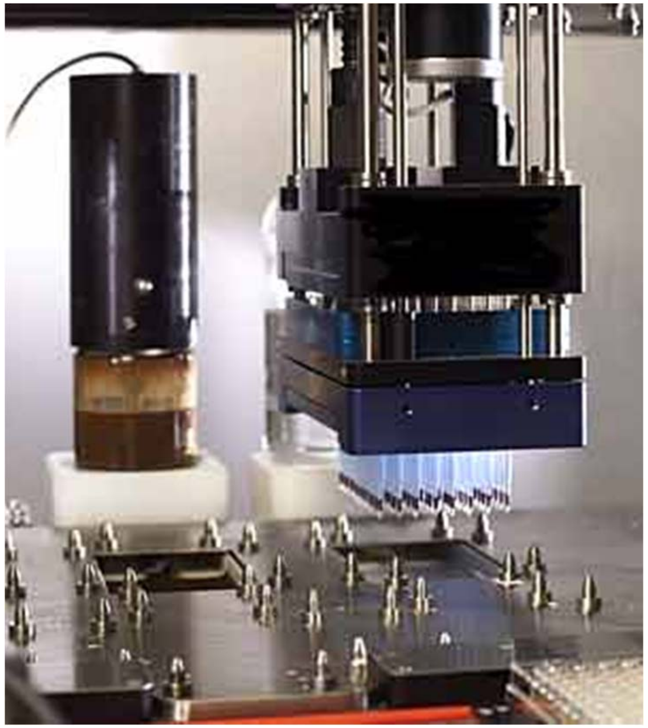


1986

Avtomatizacija določanja nukleotidnih zaporedij in uporaba fluorescentnih barvil

Leroy Hood in Mike Hunkapiller

GENEMACHINES

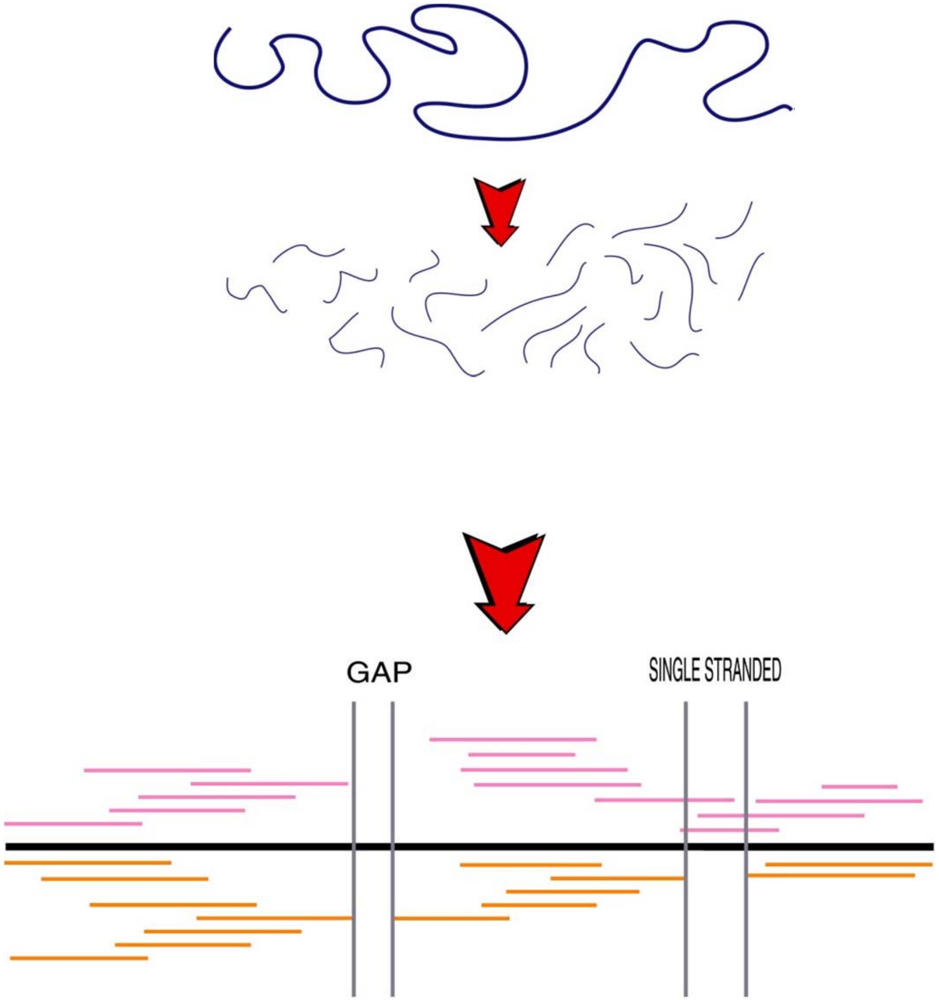


CELERA

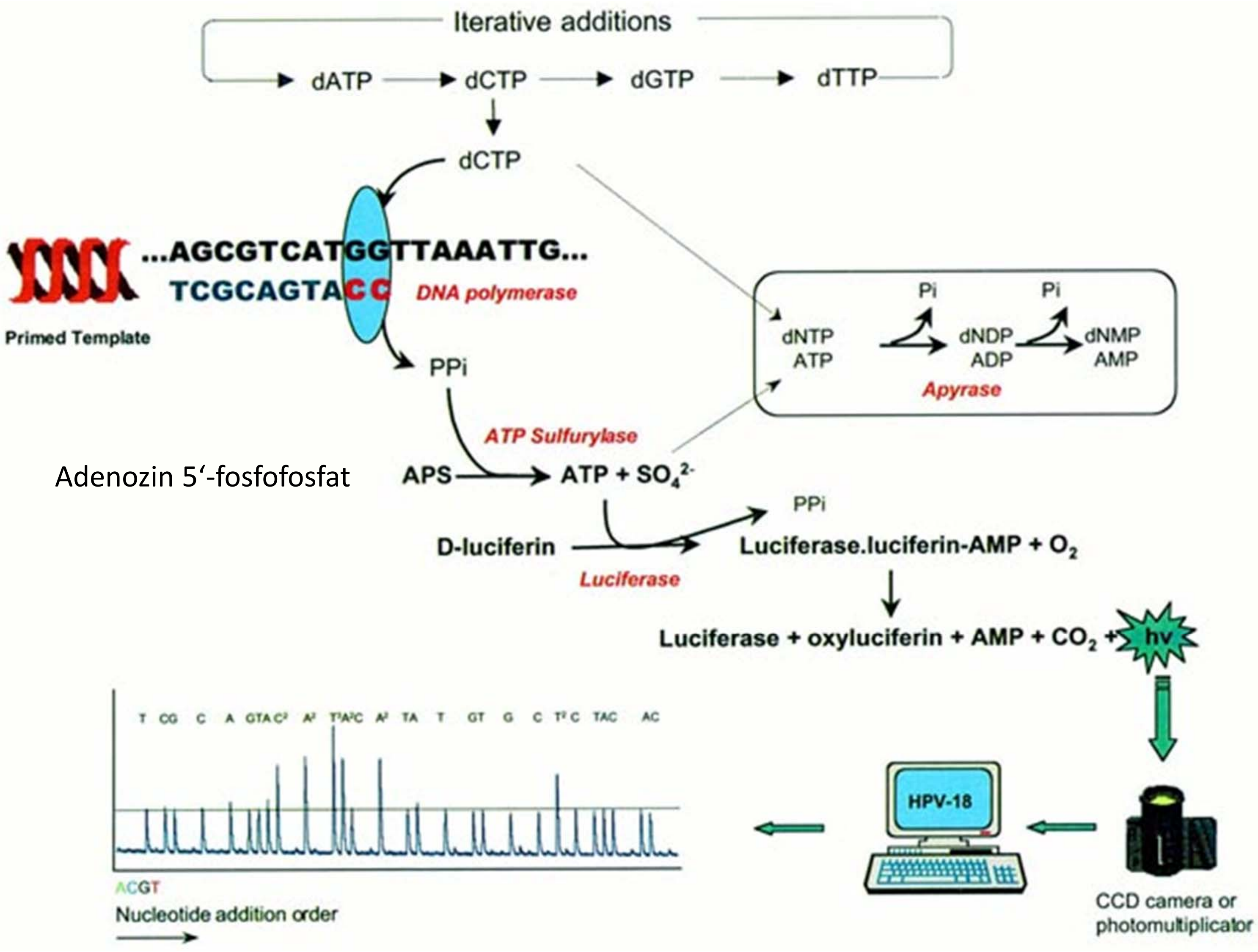
an Applera Corporation Business

300 ABI PRISM® 3700 DNA Analyzers, Applied Biosystems

800 povezanih Compaq Alpha-based 64-bit postaj, vsaka sposobna več kot 250 bilijonov primerjav zaporedij na uro.

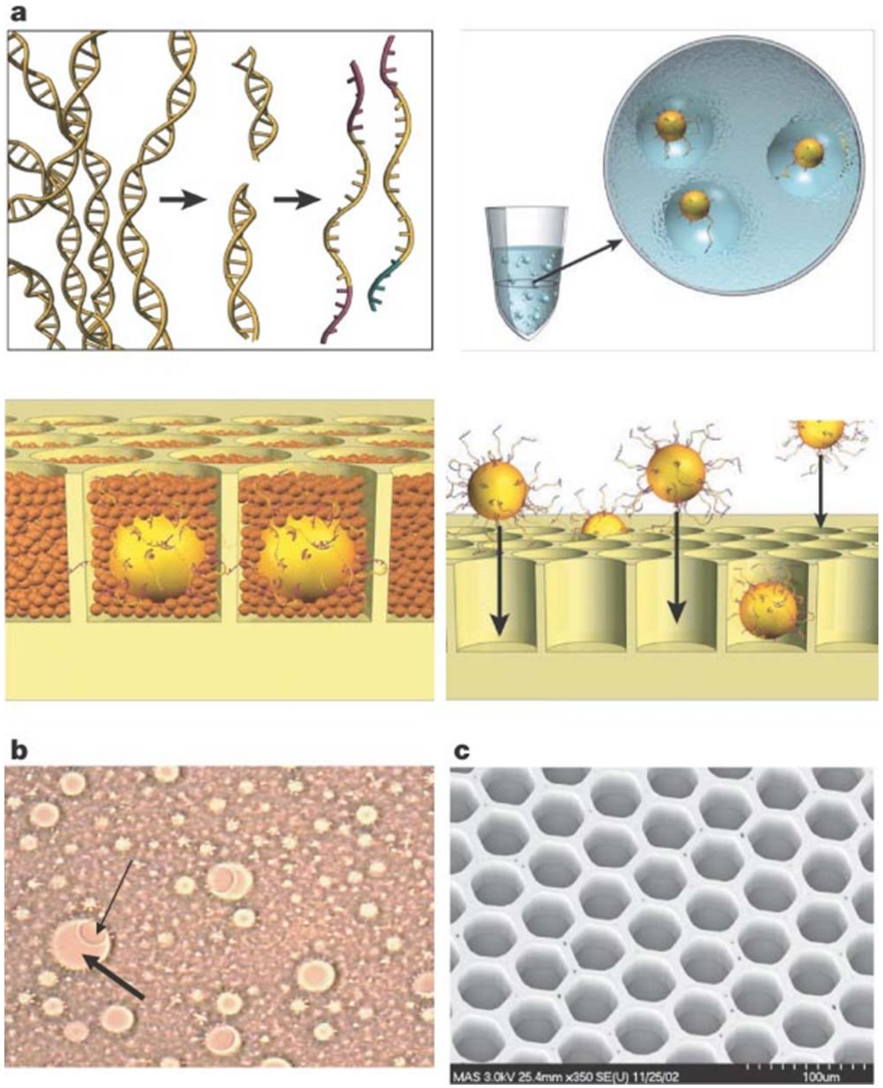


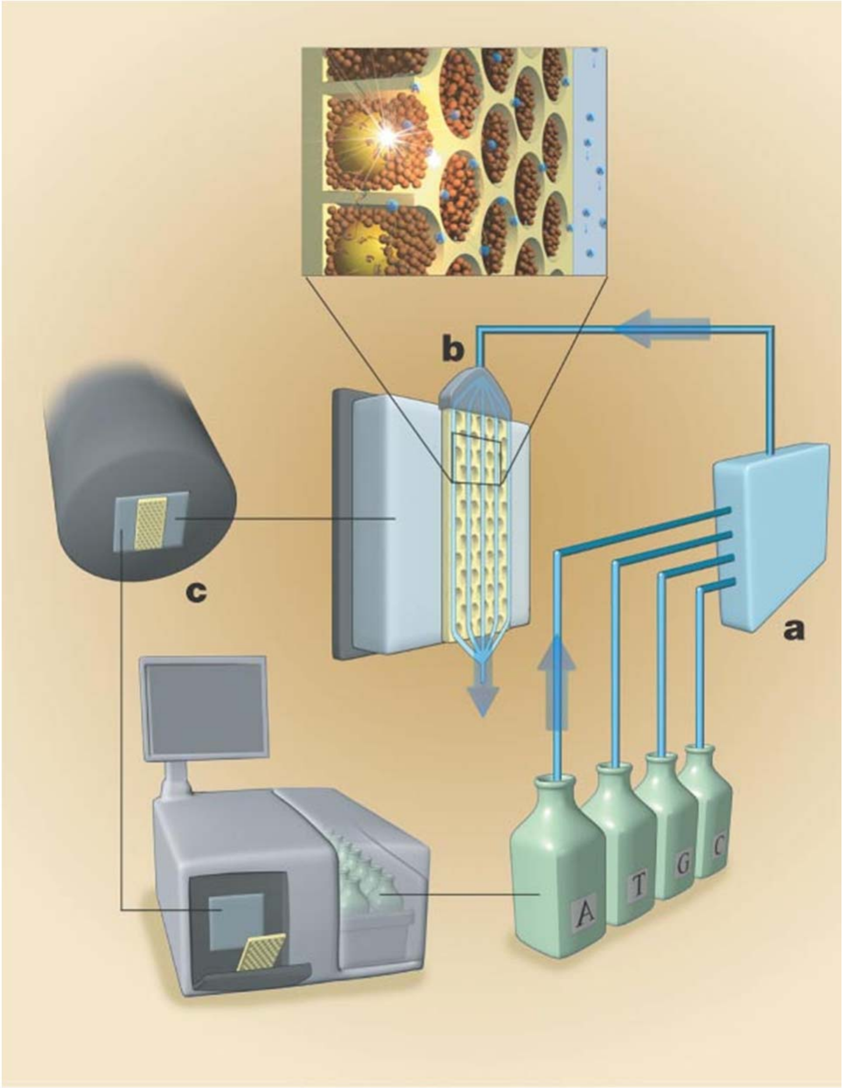
„Pyrosequencing“ 2. generacija določanja zaporedij („by synthesis“)



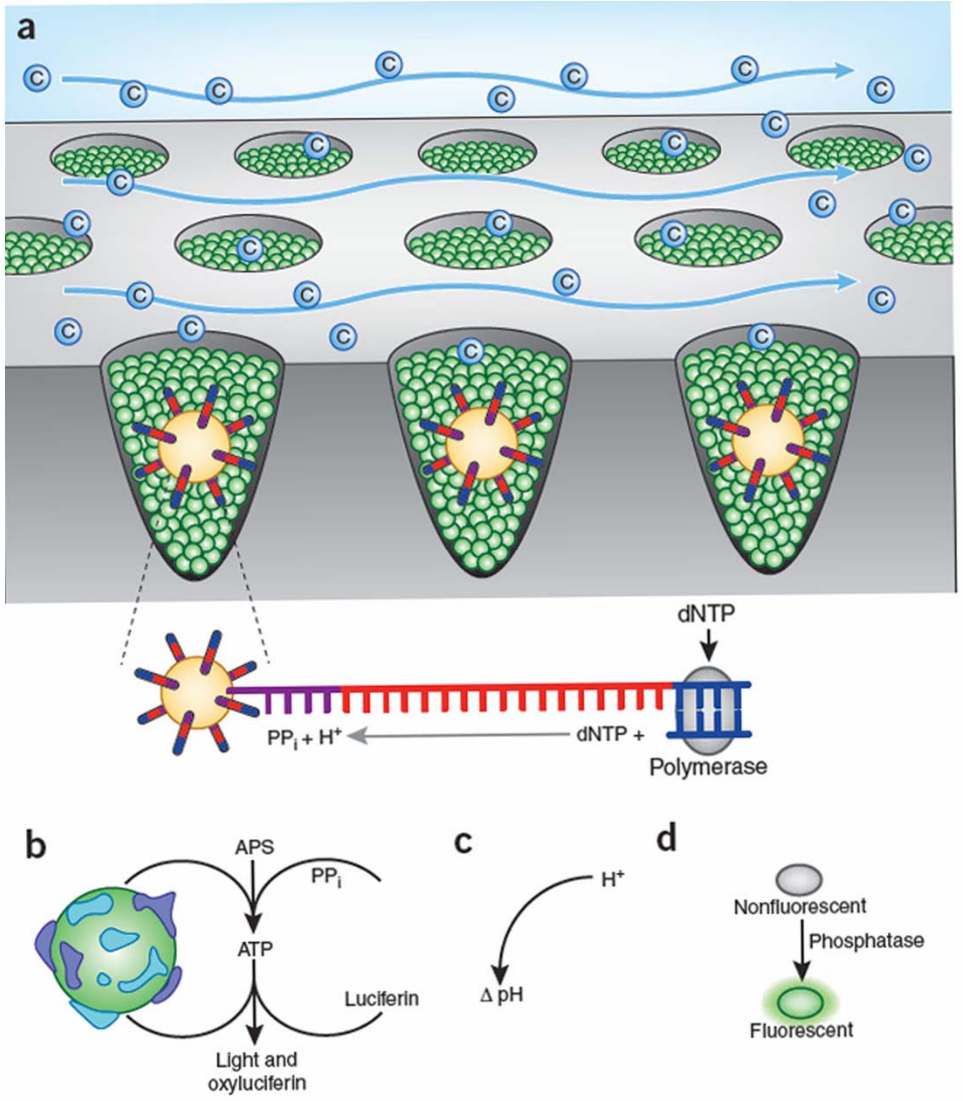
Adenozin 5‘-fosfofosfat

Gharizedah B et al. (2001) Lab. Investig. 81:673–679





Margulies M et al. (2005) Nature 437, 376-380



Tehnologija vzporednega določanja zaporedij

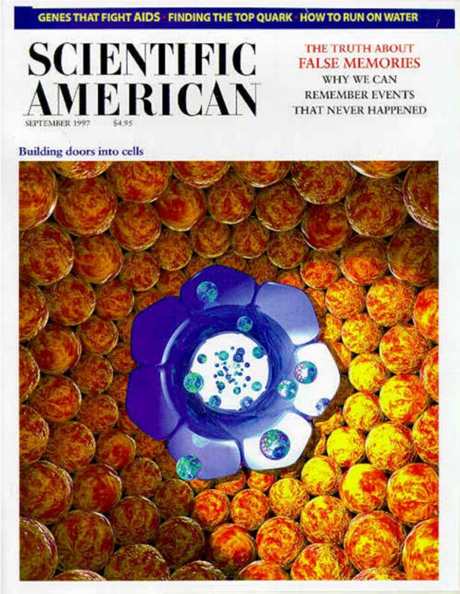
- Masovno določanje zaporedij
- Hitro
- Kloniranje ni potrebno
- „Emulsion PCR“
- Cenejše
- Bolj zanesljivo določanje zaporedij

Npr. sistem 454 Roche

Steen JA and Cooper MA (2011) Nature Methods 8,548–549

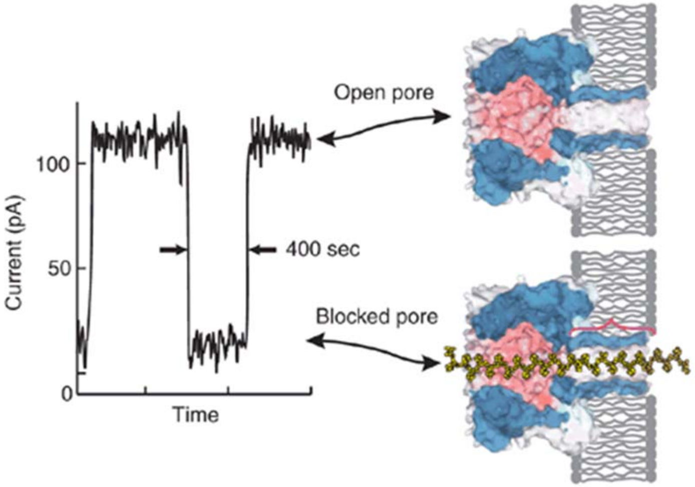
„Nanopore sequencing“ 3. generacija določanja zaporedij

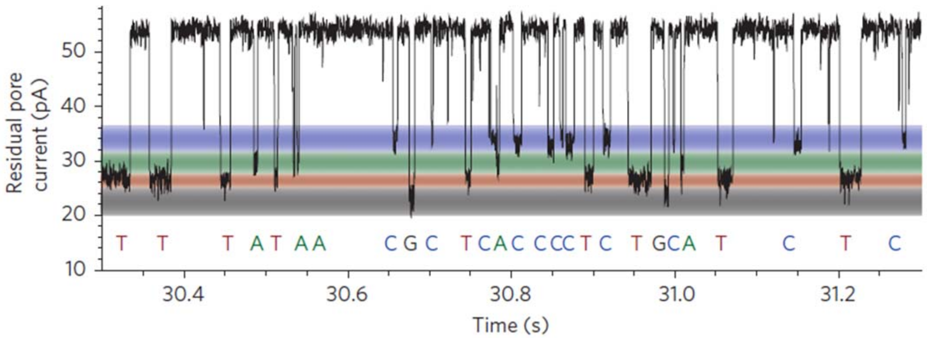
nature biotechnology
VOLUME 18 NUMBER 10 • OCTOBER 2000
http://biotech.nature.com
Protein sensor on a tether
Hybrid Bt rice
Killer Abs on commensals
The FACS about enzyme evolution


GENES THAT FIGHT AIDS · FINDING THE TOP QUARK · HOW TO RUN ON WATER
SCIENTIFIC AMERICAN
SEPTEMBER 1997 $4.95
THE TRUTH ABOUT FALSE MEMORIES
WHY WE CAN REMEMBER EVENTS THAT NEVER HAPPENED
Building doors into cells

Volume 2
Number 2
February 1995
Chemistry & Biology
Editors: Stuart L Schreiber and KC Nicolaou
Crosstalk
• Signaling by lipids
Minireviews
• Ribozyme substrate recognition
• Rapid enzyme evolution
• How NF-κB binds DNA
Research Papers
• Designed zinc-finger proteins
• Calcium effects on fibrillin domain folding
• Zn²⁺-dependent pore formation
• Fingerprinting proteins to predict ligand affinity
CURRENT BIOLOGY LIMITED
Zn²⁺
Prepore trapped by Zn²⁺ switch

On the web at http://biomednet.com/cbiology/cmb
Chemistry & Biology
Editors: Stuart L Schreiber and KC Nicolaou
Vol 4 No 7 July 1997
Crosstalk
Linked nucleic acid replication and coded peptide synthesis
Minireviews
Taming phosphatases
Angiogenesis inhibitors
Research Papers
Designed protein pores as components for biosensors
A limitation of two-state analysis
The role of metal cofactors in hairpin RNA catalysis
Stretching and breaking duplex DNA
Conversion of a toxin to its glycoside by yew
Current Biology Ltd
Protein pore with a designed analyte binding site





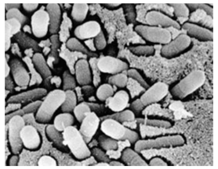
Branton D *et al.* (2008) *Nat. Biotechnol.* 26: 1146 – 1153.

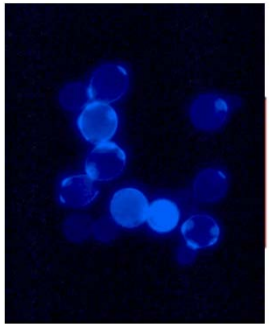
Tehnologija določanja zaporedij s pomočjo proteinskih nanopor

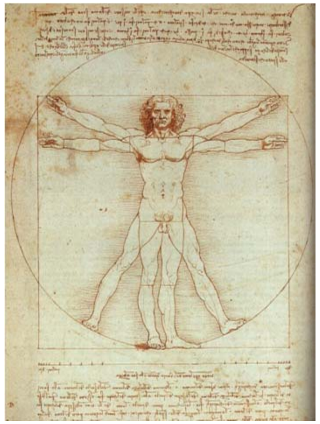
- Masovno določanje zaporedij
- Hitro
- Kloniranje in PCR ni potrebno
- Določitev veliko daljših zaporedij, npr. kbp

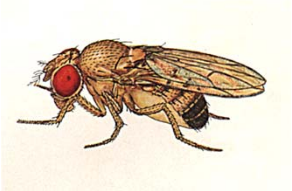
Npr. Oxford Nanopore Technologies

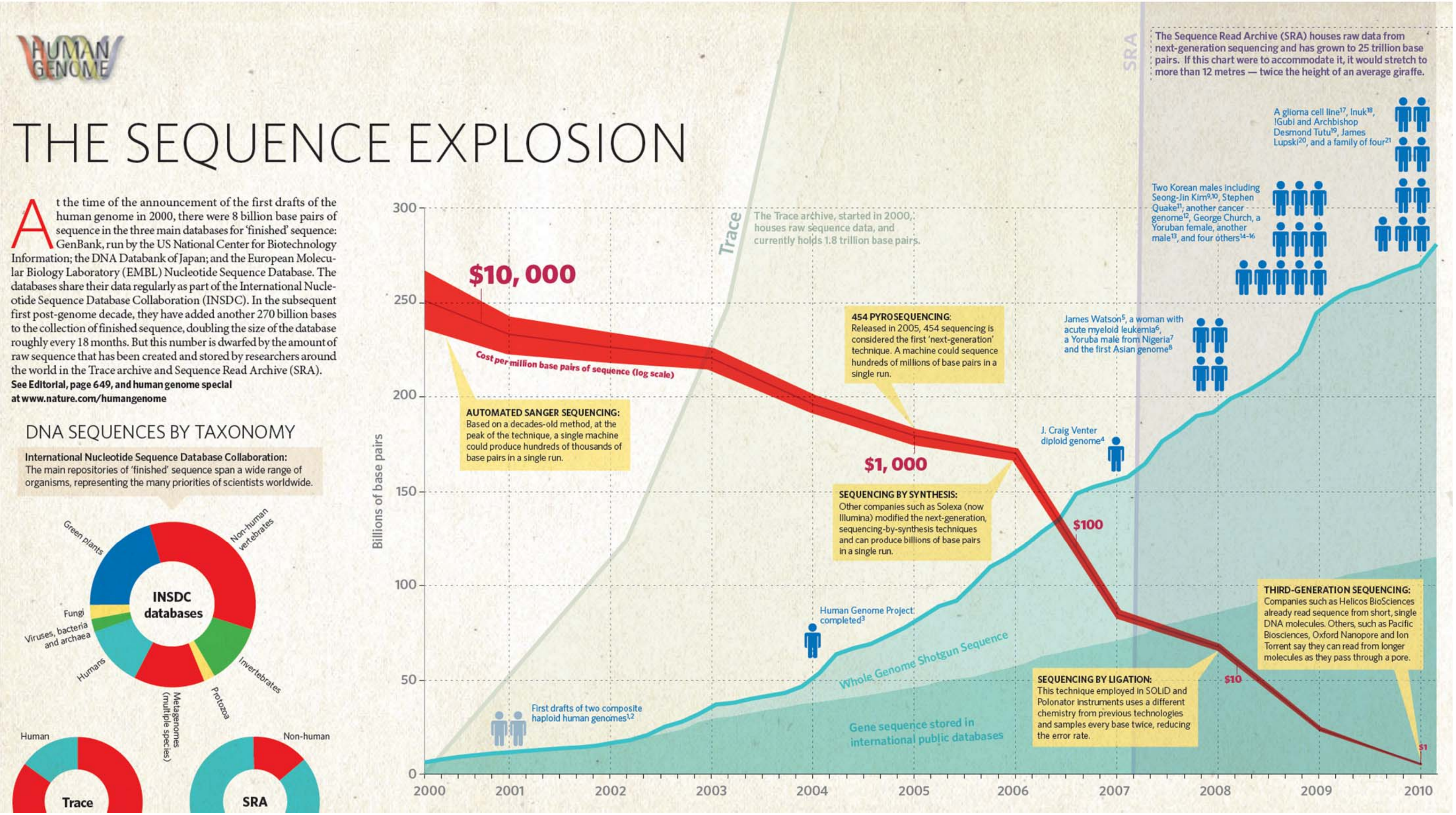
	Število baznih parov $\times 10^6$	Število genov	Določen
Bakteriofag ϕX174	0.005	10	1977
Mycoplasma genitalium	0.58	483	1995
Hemophilus influenzae	1.83	1738	1995
M. tuberculosis	4.41	3959	1998
Escherichia coli	4.6	4377	1997
Saccharomyces cerevisiae	12.00	5885	1996
Caenorhabditis elegans	95.50	19.820	1998
Drosophila melanogaster	180.00	13.601	2000
Arabidopsis thaliana	117.00	25.498	2000
Človek	3300.00	≈ 34.000	2001











THE SEQUENCE EXPLOSION

At the time of the announcement of the first drafts of the human genome in 2000, there were 8 billion base pairs of sequence in the three main databases for 'finished' sequence: GenBank, run by the US National Center for Biotechnology Information; the DNA Databank of Japan; and the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database. The databases share their data regularly as part of the International Nucleotide Sequence Database Collaboration (INSDC). In the subsequent first post-genome decade, they have added another 270 billion bases to the collection of finished sequence, doubling the size of the database roughly every 18 months. But this number is dwarfed by the amount of raw sequence that has been created and stored by researchers around the world in the Trace archive and Sequence Read Archive (SRA).

See Editorial, page 649, and human genome special at www.nature.com/humangenome

DNA SEQUENCES BY TAXONOMY

International Nucleotide Sequence Database Collaboration: The main repositories of 'finished' sequence span a wide range of organisms, representing the many priorities of scientists worldwide.

The Trace archive, started in 2000, houses raw sequence data, and currently holds 1.8 trillion base pairs.

The Sequence Read Archive (SRA) houses raw data from next-generation sequencing and has grown to 25 trillion base pairs. If this chart were to accommodate it, it would stretch to more than 12 metres — twice the height of an average giraffe.

AUTOMATED SANGER SEQUENCING: Based on a decades-old method, at the peak of the technique, a single machine could produce hundreds of thousands of base pairs in a single run.

454 PYROSEQUENCING: Released in 2005, 454 sequencing is considered the first 'next-generation' technique. A machine could sequence hundreds of millions of base pairs in a single run.

SEQUENCING BY SYNTHESIS: Other companies such as Solexa (now Illumina) modified the next-generation, sequencing-by-synthesis techniques and can produce billions of base pairs in a single run.

SEQUENCING BY LIGATION: This technique employed in SOLiD and Polonator instruments uses a different chemistry from previous technologies and samples every base twice, reducing the error rate.

THIRD-GENERATION SEQUENCING: Companies such as Helicos BioSciences already read sequence from short, single DNA molecules. Others, such as Pacific Biosciences, Oxford Nanopore and Ion Torrent say they can read from longer molecules as they pass through a pore.

Nature, 1. 4. 2010

NUKLEOTIDNE PODATKOVNE ZBIRKE

Dnevna izmenjava zaporedij!

DDBJ *DNA Data Bank of Japan*

Zaporedja javno dostopnih virov in večjih genomskih projektov

EMBL

Podatkovna zbirka EBI. Vsebuje direktno vpisana zaporedja, rezultate določevanj zaporedij genomov, zaporedja iz literature in patentov.

Iskanje in primerjave zaporedij preko vmesnikov

GenBank

Zaporedja javno dostopnih virov in večjih genomskih projektov.

Dnevna izmenjava podatkov z DDBJ in EMBL.

Razdeljena na posamezne odseke (nivo organizmov, EST, PAT, STS, GSS, HTG).

157.943.793.171 baz v **171.123.749** zaporedjih (14. 2. 2014)

iz več kot **165 000** organizmov

Dostop preko Entreza (Medline ali BLAST)

dbEST podatkovna zbirka oznak izraženih zaporedij

Ostale specializirane podatkovne zbirke

GSDB (The Genome Sequence DataBase), SGD (*Saccharomyces* Genome Database), UniGene, TDB (TIGR podatkovna zbirka), ACeDB (A *C. elegans* DataBase)

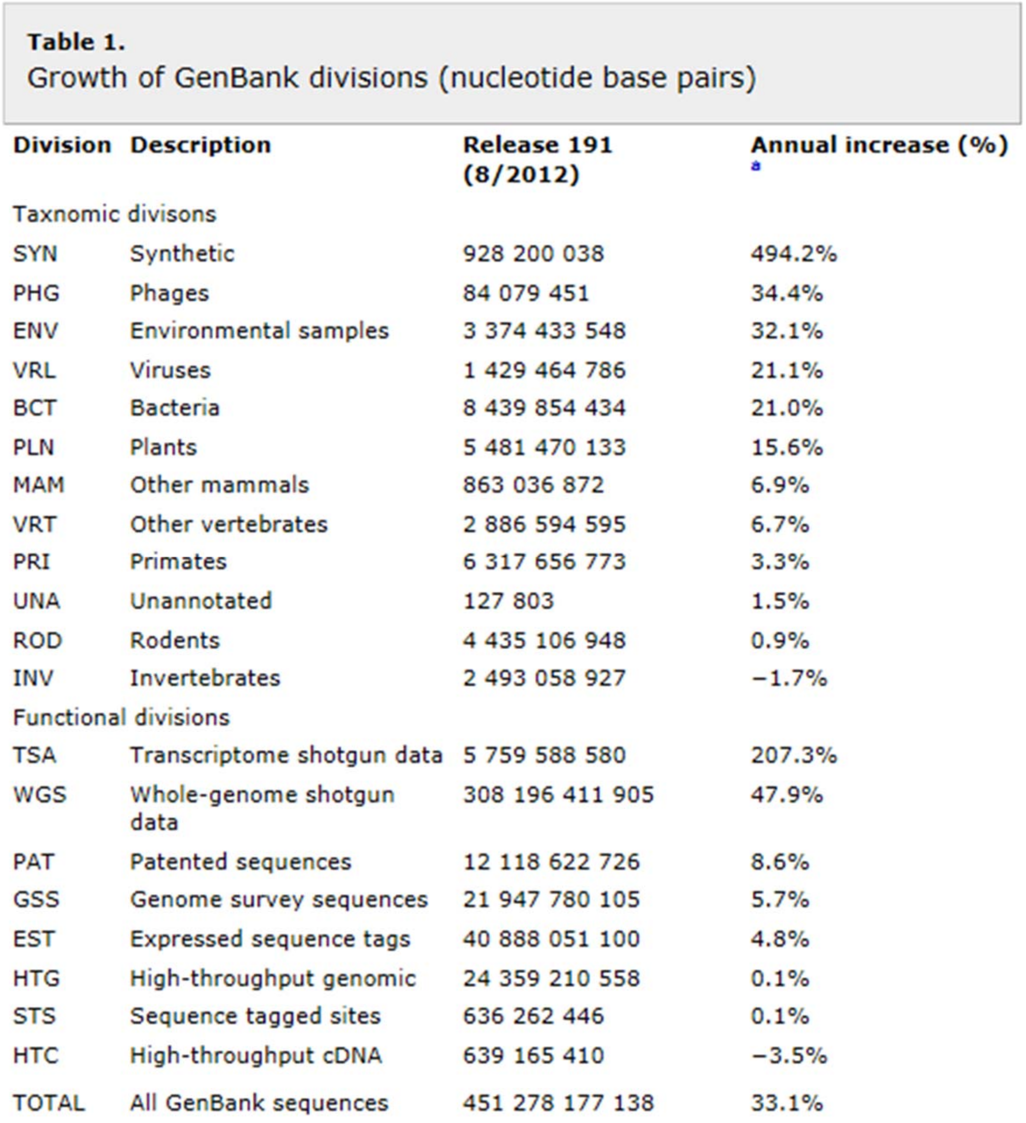
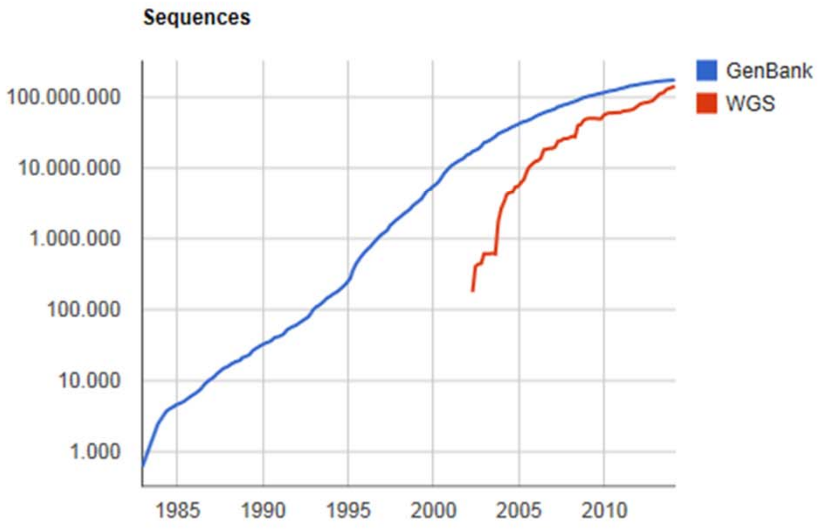


Table 1.
Growth of GenBank divisions (nucleotide base pairs)

Division	Description	Release 191 (8/2012)	Annual increase (%)[a]
Taxnomic divisons			
SYN	Synthetic	928 200 038	494.2%
PHG	Phages	84 079 451	34.4%
ENV	Environmental samples	3 374 433 548	32.1%
VRL	Viruses	1 429 464 786	21.1%
BCT	Bacteria	8 439 854 434	21.0%
PLN	Plants	5 481 470 133	15.6%
MAM	Other mammals	863 036 872	6.9%
VRT	Other vertebrates	2 886 594 595	6.7%
PRI	Primates	6 317 656 773	3.3%
UNA	Unannotated	127 803	1.5%
ROD	Rodents	4 435 106 948	0.9%
INV	Invertebrates	2 493 058 927	−1.7%
Functional divisions			
TSA	Transcriptome shotgun data	5 759 588 580	207.3%
WGS	Whole-genome shotgun data	308 196 411 905	47.9%
PAT	Patented sequences	12 118 622 726	8.6%
GSS	Genome survey sequences	21 947 780 105	5.7%
EST	Expressed sequence tags	40 888 051 100	4.8%
HTG	High-throughput genomic	24 359 210 558	0.1%
STS	Sequence tagged sites	636 262 446	0.1%
HTC	High-throughput cDNA	639 165 410	−3.5%
TOTAL	All GenBank sequences	451 278 177 138	33.1%

Benson DA et al. (2013) Nucl. Acids Res. 41: D36-D42



Iskanje preko Entrez Nucleotide

CoreNucleotide (the main collection)- osnovna zbirka

dbEST (Expressed Sequence Tags)- zbirka izraženih oznak zaporedij

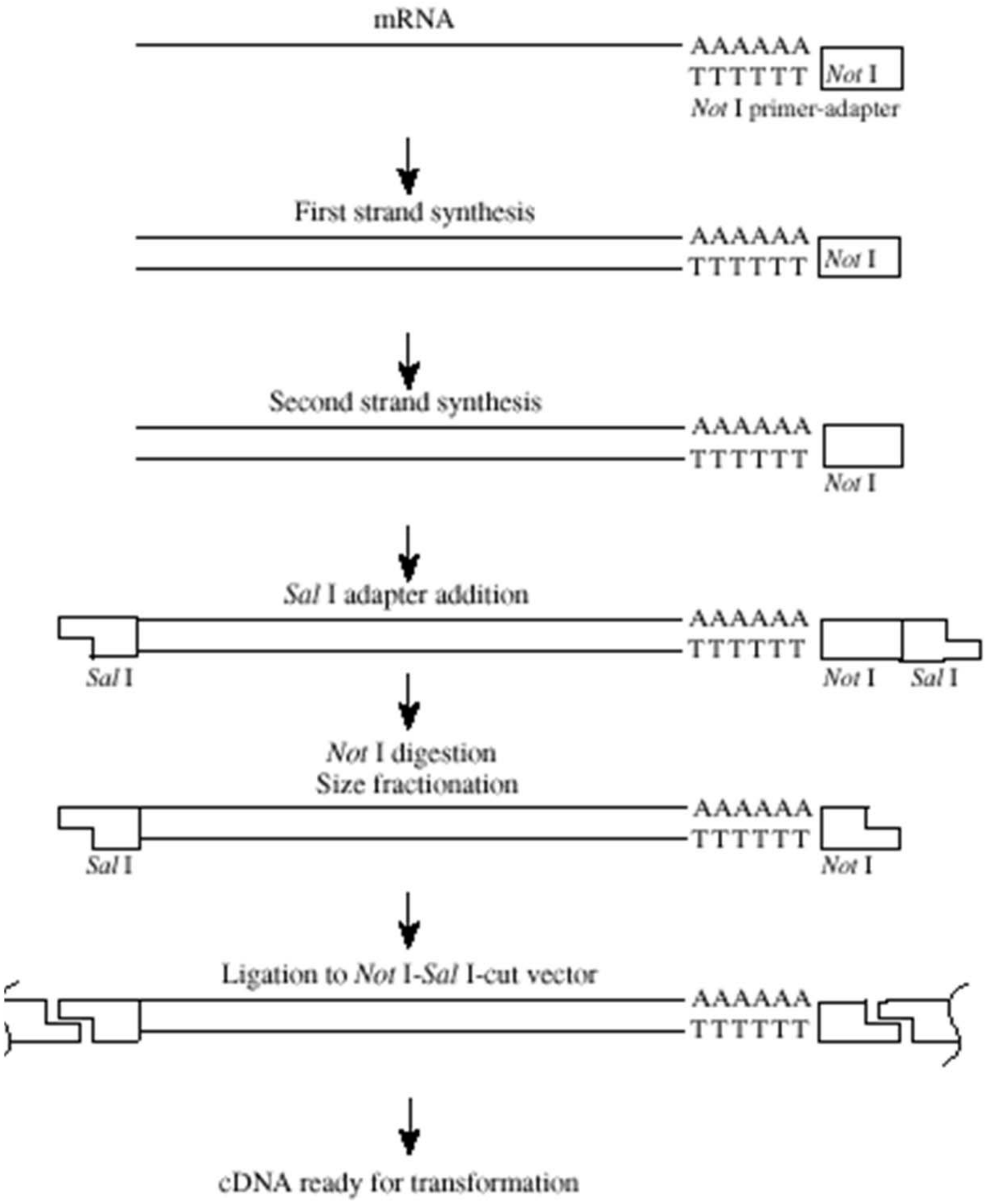
dbGSS (Genome Survey Sequences)- neanotirana zaporedja („single-read“)

EST kloni

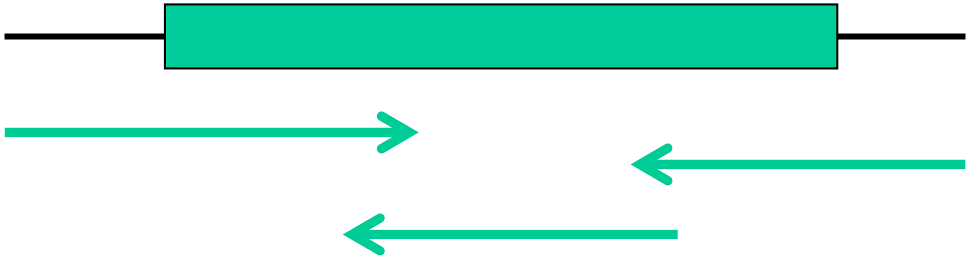
cDNA knjižnica kot osnova

- •Iskanje novih genov
- •Pomoč pri določanju genov v genomih
- •Kvantifikacija izražanja genov
- •Primerjava med celicami/tkivi

Z modernimi metodami določanja zaporedij ni več aktualno- **RNASeq**



EST kloni kot pomoč pri določanju genov



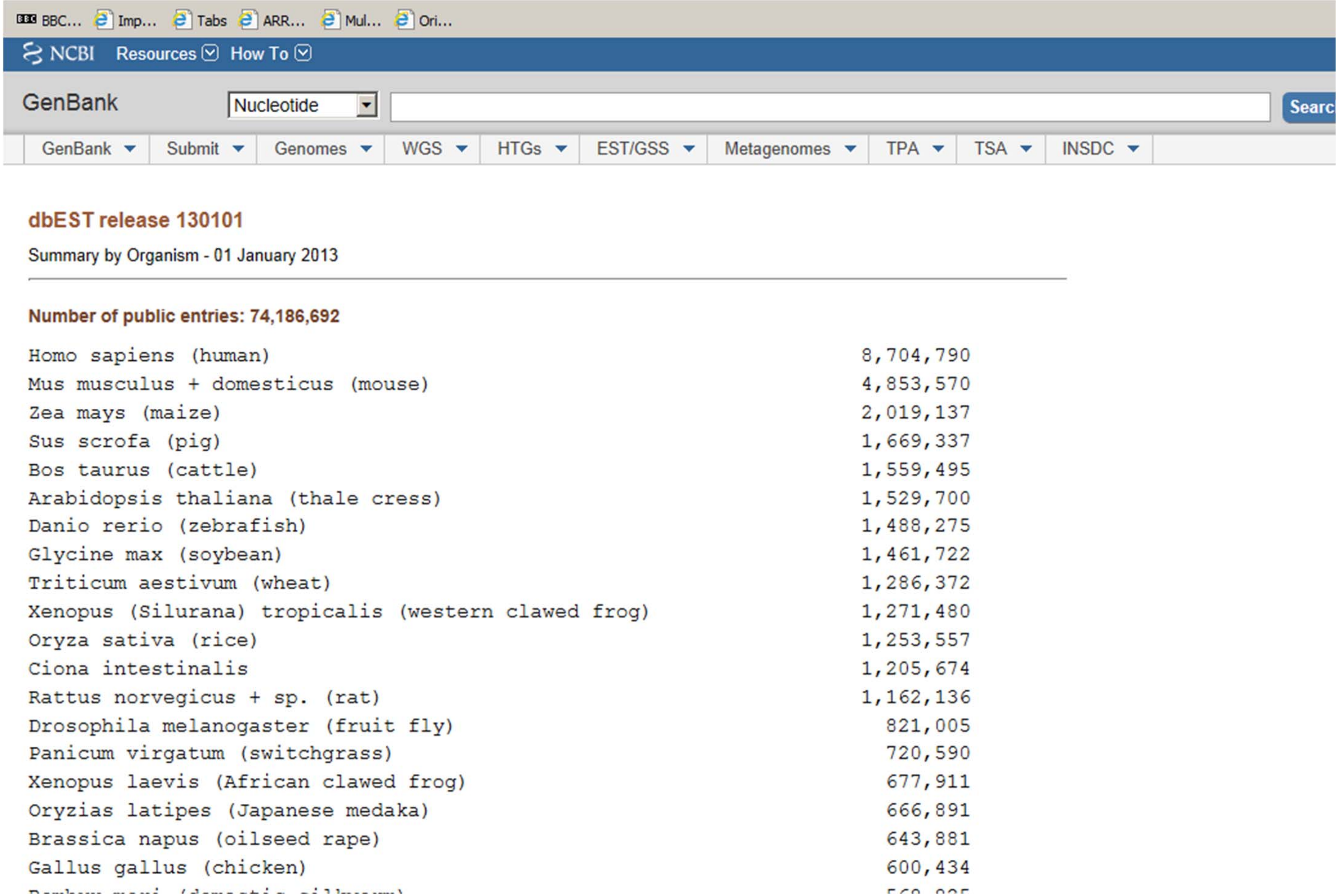
TIGR
The **I**nstitute for **G**enomic **R**esearch
http://www.tigr.org/

Tudi ostali:
Merck/IMAGE, Incyte....

dbEST

dbEST (Nature Genetics 4:332-3;1993) is a division of GenBank that contains sequence data and other information on "single-pass" cDNA sequences, or "Expressed Sequence Tags", from a number of organisms. A brief account of the history of human ESTs in GenBank is available (Trends Biochem. Sci. 20:295-6;1995). Also, consult the special "Genome Directory" issue of Nature (vol. 377, issue 6547S, 28 September 1995).

Adams MD *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. Science. 1991 21;252(5013):1651-6.



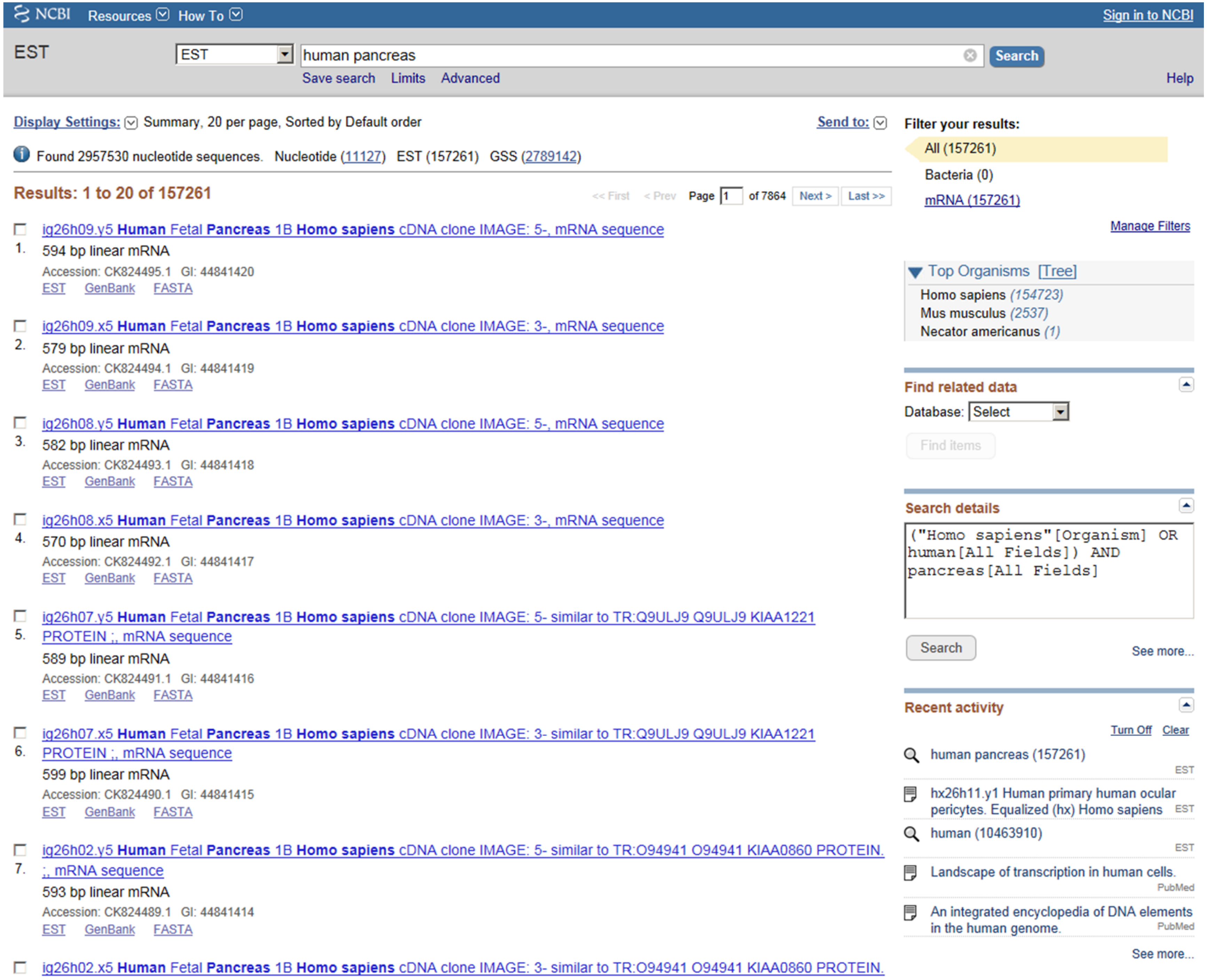
BBC… Imp… Tabs ARR… Mul… Ori…
NCBI Resources How To
GenBank
Nucleotide
Searc
GenBank Submit Genomes WGS HTGs EST/GSS Metagenomes TPA TSA INSDC

dbEST release 130101

Summary by Organism - 01 January 2013

Number of public entries: 74,186,692

Homo sapiens (human)	8,704,790
Mus musculus + domesticus (mouse)	4,853,570
Zea mays (maize)	2,019,137
Sus scrofa (pig)	1,669,337
Bos taurus (cattle)	1,559,495
Arabidopsis thaliana (thale cress)	1,529,700
Danio rerio (zebrafish)	1,488,275
Glycine max (soybean)	1,461,722
Triticum aestivum (wheat)	1,286,372
Xenopus (Silurana) tropicalis (western clawed frog)	1,271,480
Oryza sativa (rice)	1,253,557
Ciona intestinalis	1,205,674
Rattus norvegicus + sp. (rat)	1,162,136
Drosophila melanogaster (fruit fly)	821,005
Panicum virgatum (switchgrass)	720,590
Xenopus laevis (African clawed frog)	677,911
Oryzias latipes (Japanese medaka)	666,891
Brassica napus (oilseed rape)	643,881
Gallus gallus (chicken)	600,434
Bombyx mori (domestic silkworm)	569,925



human pancreas

Display Settings: Summary, 20 per page, Sorted by Default order

Found 2957530 nucleotide sequences. Nucleotide (11127) EST (157261) GSS (2789142)

Results: 1 to 20 of 157261

□ 1. ig26h09.y5 **Human** Fetal **Pancreas** 1B **Homo sapiens** cDNA clone IMAGE: 5-, mRNA sequence
594 bp linear mRNA
Accession: CK824495.1 GI: 44841420
EST GenBank FASTA

□ 2. ig26h09.x5 **Human** Fetal **Pancreas** 1B **Homo sapiens** cDNA clone IMAGE: 3-, mRNA sequence
579 bp linear mRNA
Accession: CK824494.1 GI: 44841419
EST GenBank FASTA

□ 3. ig26h08.y5 **Human** Fetal **Pancreas** 1B **Homo sapiens** cDNA clone IMAGE: 5-, mRNA sequence
582 bp linear mRNA
Accession: CK824493.1 GI: 44841418
EST GenBank FASTA

□ 4. ig26h08.x5 **Human** Fetal **Pancreas** 1B **Homo sapiens** cDNA clone IMAGE: 3-, mRNA sequence
570 bp linear mRNA
Accession: CK824492.1 GI: 44841417
EST GenBank FASTA

□ 5. ig26h07.y5 **Human** Fetal **Pancreas** 1B **Homo sapiens** cDNA clone IMAGE: 5- similar to TR:Q9ULJ9 Q9ULJ9 KIAA1221 PROTEIN ;, mRNA sequence
589 bp linear mRNA
Accession: CK824491.1 GI: 44841416
EST GenBank FASTA

□ 6. ig26h07.x5 **Human** Fetal **Pancreas** 1B **Homo sapiens** cDNA clone IMAGE: 3- similar to TR:Q9ULJ9 Q9ULJ9 KIAA1221 PROTEIN ;, mRNA sequence
599 bp linear mRNA
Accession: CK824490.1 GI: 44841415
EST GenBank FASTA

□ 7. ig26h02.y5 **Human** Fetal **Pancreas** 1B **Homo sapiens** cDNA clone IMAGE: 5- similar to TR:O94941 O94941 KIAA0860 PROTEIN. ;, mRNA sequence
593 bp linear mRNA
Accession: CK824489.1 GI: 44841414
EST GenBank FASTA

□ ig26h02.x5 **Human** Fetal **Pancreas** 1B **Homo sapiens** cDNA clone IMAGE: 3- similar to TR:O94941 O94941 KIAA0860 PROTEIN.

Filter your results:
All (157261)
Bacteria (0)
mRNA (157261)

Top Organisms [Tree]
Homo sapiens *(154723)*
Mus musculus *(2537)*
Necator americanus *(1)*

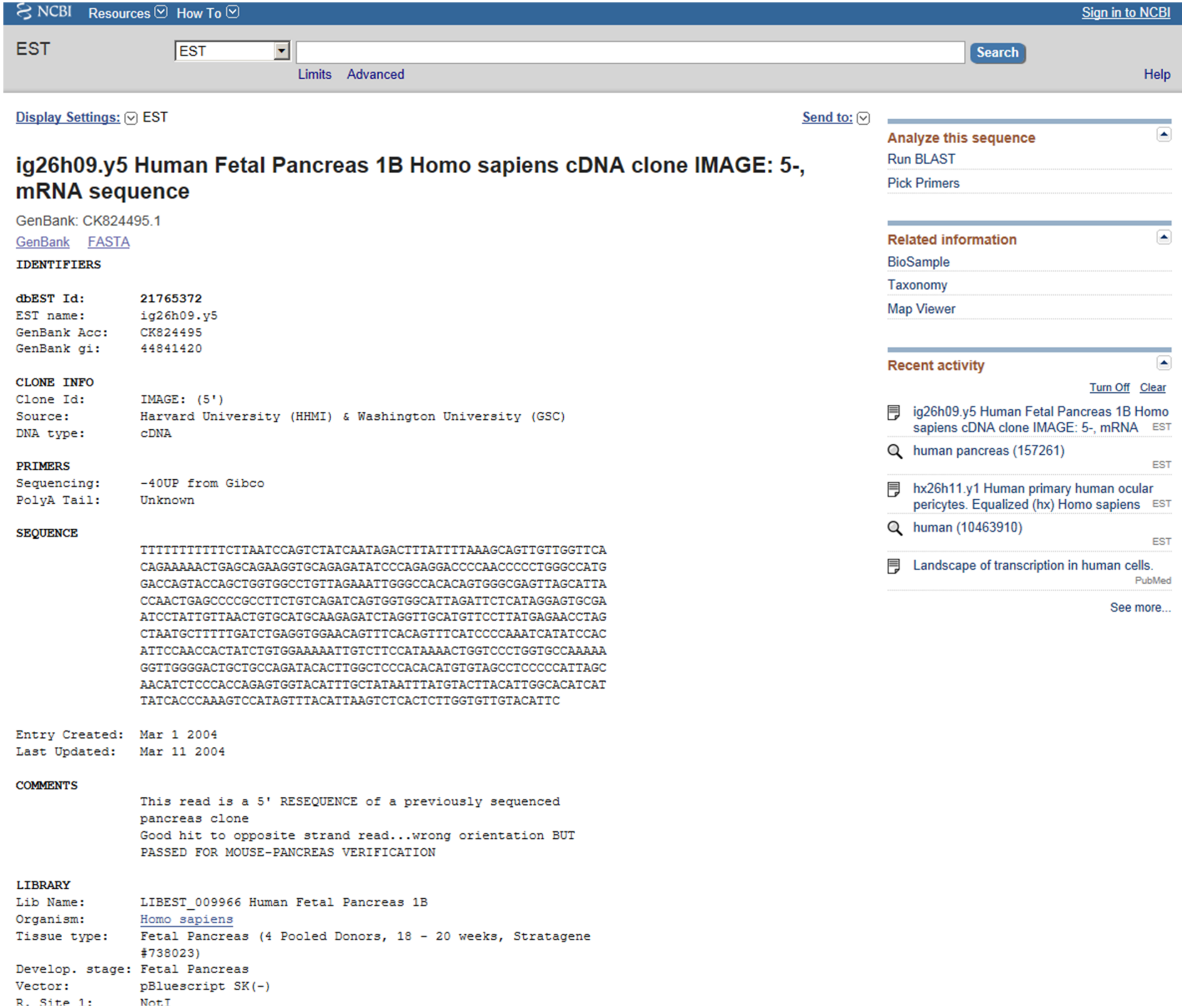
Find related data
Database: Select

Search details

```
("Homo sapiens"[Organism] OR
human[All Fields]) AND
pancreas[All Fields]
```

Recent activity

human pancreas (157261) EST
hx26h11.y1 Human primary human ocular pericytes. Equalized (hx) Homo sapiens EST
human (10463910) EST
Landscape of transcription in human cells. PubMed
An integrated encyclopedia of DNA elements in the human genome. PubMed



ig26h09.y5 Human Fetal Pancreas 1B Homo sapiens cDNA clone IMAGE: 5-, mRNA sequence

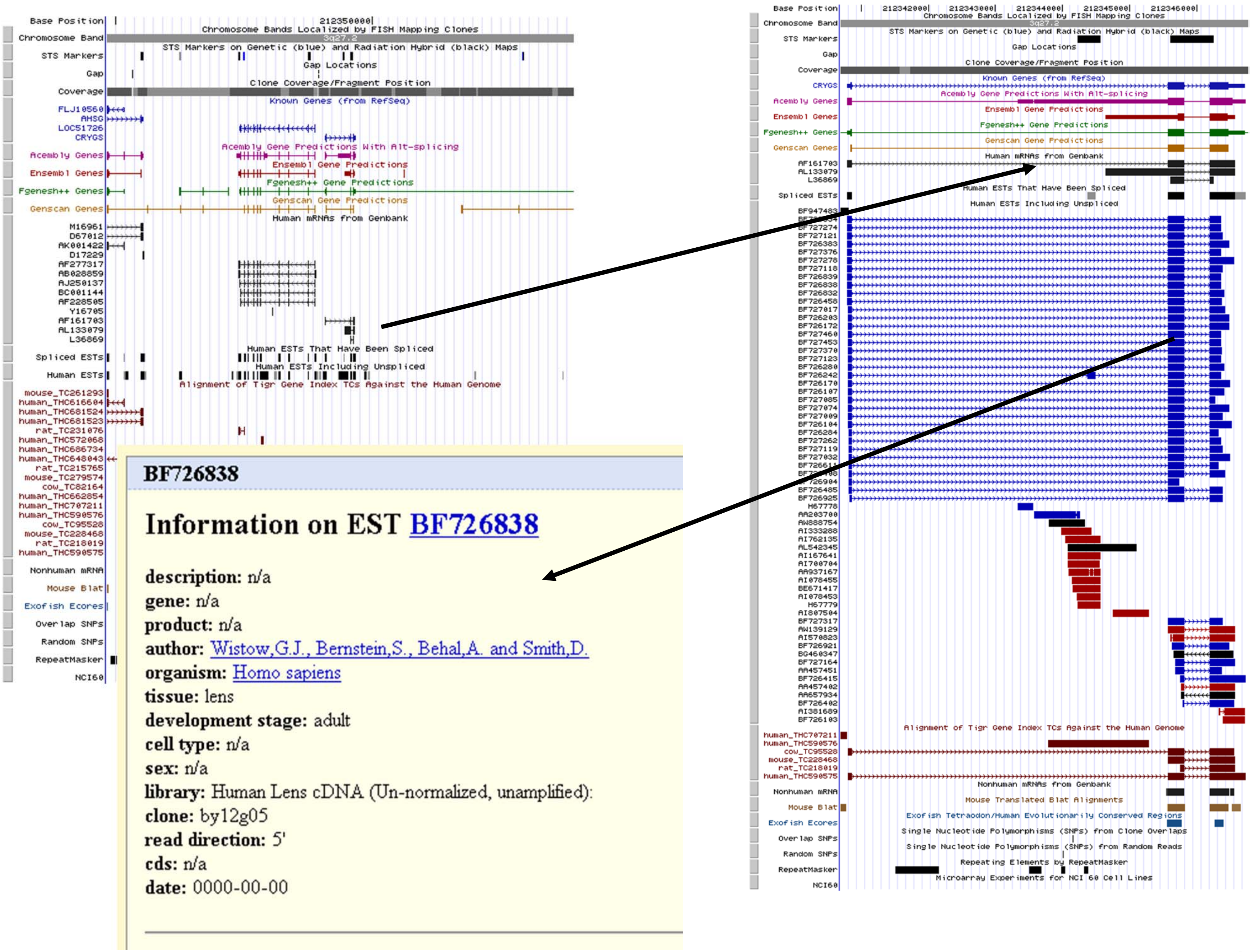
GenBank: CK824495.1

GenBank FASTA

```
IDENTIFIERS

dbEST Id:        21765372
EST name:        ig26h09.y5
GenBank Acc:     CK824495
GenBank gi:      44841420

CLONE INFO
Clone Id:        IMAGE: (5')
Source:          Harvard University (HHMI) & Washington University (GSC)
DNA type:        cDNA

PRIMERS
Sequencing:      -40UP from Gibco
PolyA Tail:      Unknown

SEQUENCE
                 TTTTTTTTTTTCTTAATCCAGTCTATCAATAGACTTTATTTTAAAGCAGTTGTTGGTTCA
                 CAGAAAAACTGAGCAGAAGGTGCAGAGATATCCCAGAGGACCCCAACCCCCTGGGCCATG
                 GACCAGTACCAGCTGGTGGCCTGTTAGAAATTGGGCCACACAGTGGGCGAGTTAGCATTA
                 CCAACTGAGCCCCGCCTTCTGTCAGATCAGTGGTGGCATTAGATTCTCATAGGAGTGCGA
                 ATCCTATTGTTAACTGTGCATGCAAGAGATCTAGGTTGCATGTTCCTTATGAGAACCTAG
                 CTAATGCTTTTTGATCTGAGGTGGAACAGTTTCACAGTTTCATCCCCAAATCATATCCAC
                 ATTCCAACCACTATCTGTGGAAAAATTGTCTTCCATAAAACTGGTCCCTGGTGCCAAAAA
                 GGTTGGGGACTGCTGCCAGATACACTTGGCTCCCACACATGTGTAGCCTCCCCCATTAGC
                 AACATCTCCCACCAGAGTGGTACATTTGCTATAATTTATGTACTTACATTGGCACATCAT
                 TATCACCCAAAGTCCATAGTTTACATTAAGTCTCACTCTTGGTGTTGTACATTC

Entry Created:   Mar 1 2004
Last Updated:    Mar 11 2004

COMMENTS
                 This read is a 5' RESEQUENCE of a previously sequenced
                 pancreas clone
                 Good hit to opposite strand read...wrong orientation BUT
                 PASSED FOR MOUSE-PANCREAS VERIFICATION

LIBRARY
Lib Name:        LIBEST_009966 Human Fetal Pancreas 1B
Organism:        Homo sapiens
Tissue type:     Fetal Pancreas (4 Pooled Donors, 18 - 20 weeks, Stratagene
                 #738023)
Develop. stage:  Fetal Pancreas
Vector:          pBluescript SK(-)
R. Site 1:       NotI
```

Analyze this sequence
Run BLAST
Pick Primers
Related information
BioSample
Taxonomy
Map Viewer
Recent activity
Turn Off
Clear
ig26h09.y5 Human Fetal Pancreas 1B Homo sapiens cDNA clone IMAGE: 5-, mRNA
EST
human pancreas (157261)
EST
hx26h11.y1 Human primary human ocular pericytes. Equalized (hx) Homo sapiens
EST
human (10463910)
EST
Landscape of transcription in human cells.
PubMed
See more...

```
>gi|44841420|gb|CK824495.1|CK824495 ig26h09.y5 Human Fetal Pancreas 1B Homo
sapiens cDNA clone IMAGE: 5', mRNA sequence
TTTTTTTTTTTCTTAATCCAGTCTATCAATAGACTTTATTTTAAAGCAGTTGTTGGTTCACAGAAAAACT
GAGCAGAAGGTGCAGAGATATCCCAGAGGACCCCAACCCCCTGGGCCATGGACCAGTACCAGCTGGTGGC
CTGTTAGAAATTGGGCCACACAGTGGGCGAGTTAGCATTACCAACTGAGCCCCGCCTTCTGTCAGATCAG
TGGTGGCATTAGATTCTCATAGGAGTGCGAATCCTATTGTTAACTGTGCATGCAAGAGATCTAGGTTGCA
TGTTCCTTATGAGAACCTAGCTAATGCTTTTTGATCTGAGGTGGAACAGTTTCACAGTTTCATCCCCAAA
TCATATCCACATTCCAACCACTATCTGTGGAAAAATTGTCTTCCATAAAACTGGTCCCTGGTGCCAAAAA
GGTTGGGGACTGCTGCCAGATACACTTGGCTCCCACACATGTGTAGCCTCCCCCATTAGCAACATCTCCC
ACCAGAGTGGTACATTTGCTATAATTTATGTACTTACATTGGCACATCATTATCACCCAAAGTCCATAGT
TTACATTAAGTCTCACTCTTGGTGTTGTACATTC
```


Base Position
Chromosome Band
212350000
3q27.2
Chromosome Bands Localized by FISH Mapping Clones
STS Markers
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
Gap
Gap Locations
Coverage
Clone Coverage/Fragment Position
Known Genes (from RefSeq)
FLJ10560
AHSG
LOC51726
CRYGS
Acembly Genes
Acembly Gene Predictions With Alt-splicing
Ensembl Genes
Ensembl Gene Predictions
Fgenesh++ Genes
Fgenesh++ Gene Predictions
Genscan Genes
Genscan Gene Predictions
Human mRNAs from Genbank
M16961
D67012
AK001422
D17229
AF277317
AB028859
AJ250137
BC001144
AF228505
Y16705
AF161703
AL133079
L36869
Spliced ESTs
Human ESTs That Have Been Spliced
Human ESTs
Human ESTs Including Unspliced
Alignment of Tigr Gene Index TCs Against the Human Genome
Nonhuman mRNA
Mouse Blat
Exofish Ecores
Overlap SNPs
Random SNPs
RepeatMasker
NCI60
212342000
212343000
212344000
212345000
212346000
Nonhuman mRNAs from Genbank
Mouse Translated Blat Alignments
Exofish Tetraodon/Human Evolutionarily Conserved Regions
Single Nucleotide Polymorphisms (SNPs) from Clone Overlaps
Single Nucleotide Polymorphisms (SNPs) from Random Reads
Repeating Elements by RepeatMasker
Microarray Experiments for NCI 60 Cell Lines

BF726838

Information on EST BF726838

description: n/a

gene: n/a

product: n/a

author: Wistow,G.J., Bernstein,S., Behal,A. and Smith,D.

organism: Homo sapiens

tissue: lens

development stage: adult

cell type: n/a

sex: n/a

library: Human Lens cDNA (Un-normalized, unamplified):

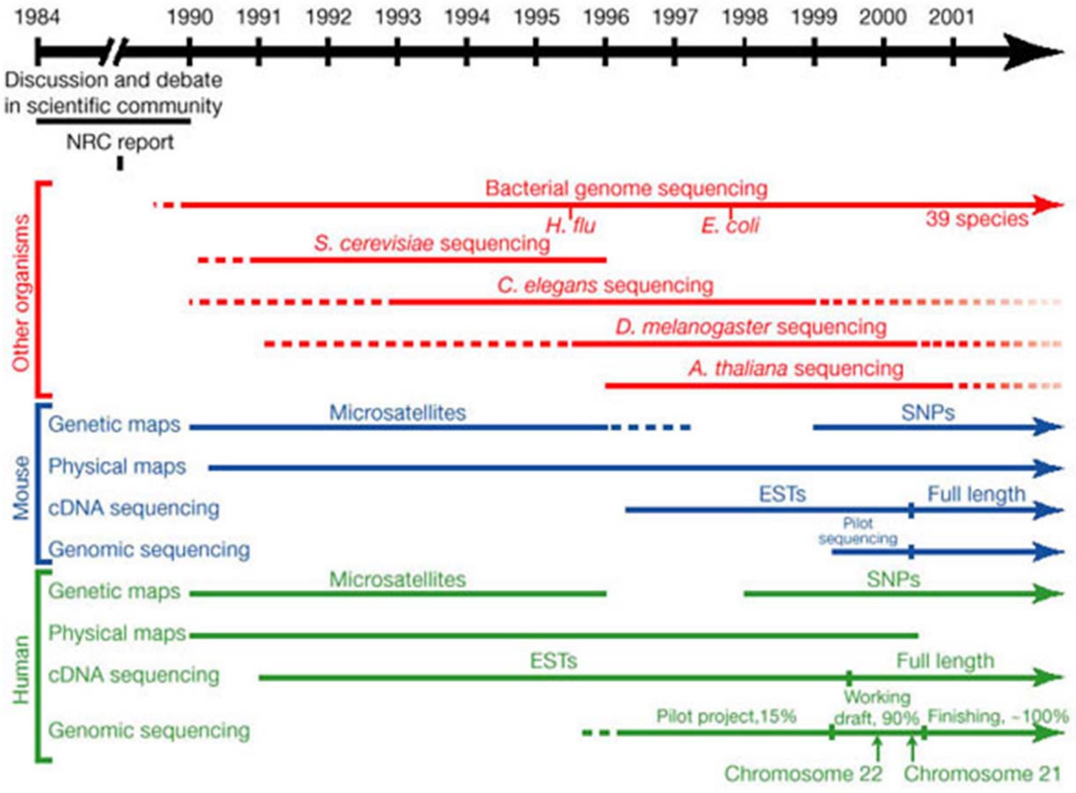
clone: by12g05

read direction: 5'

cds: n/a

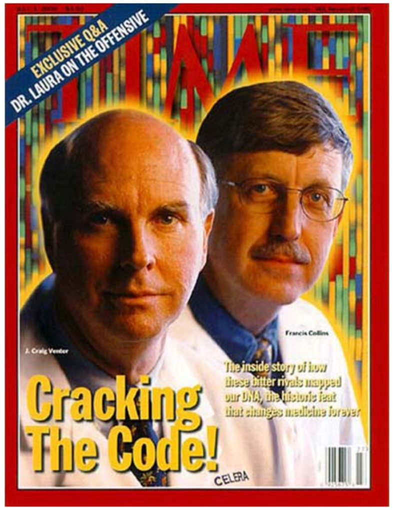
date: 0000-00-00

UCSC Genome browser
http://genome.ucsc.edu/goldenPath/aug2001Tracks.html

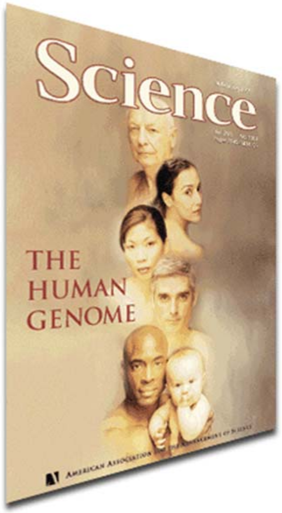


International Human Genome Sequencing Consortium (2001) *Initial sequencing and analysis of the human genome*. Nature 409, 860-921

Craig Venter


TIME
EXCLUSIVE Q&A
DR. LAURA ON THE OFFENSIVE
J. Craig Venter
Francis Collins
Cracking The Code!
The inside story of how these bitter rivals mapped our DNA, the historic feat that changes medicine forever
CELERA

Francis Collins



- International Human Genome Sequencing Consortium (2001) *Initial sequencing and analysis of the human genome*. Nature 409, 860-921
- J. Craig Venter et al. (2001) *The Sequence of the Human Genome*. Science, 1304-1135

Prve verziji človeškega genoma.

- The Human Genome 10th Anniversary: http://www.sciencemag.org/site/extra/genomeanniversary/

Tematska številka ob 10. obletnici objave človeškega genoma.

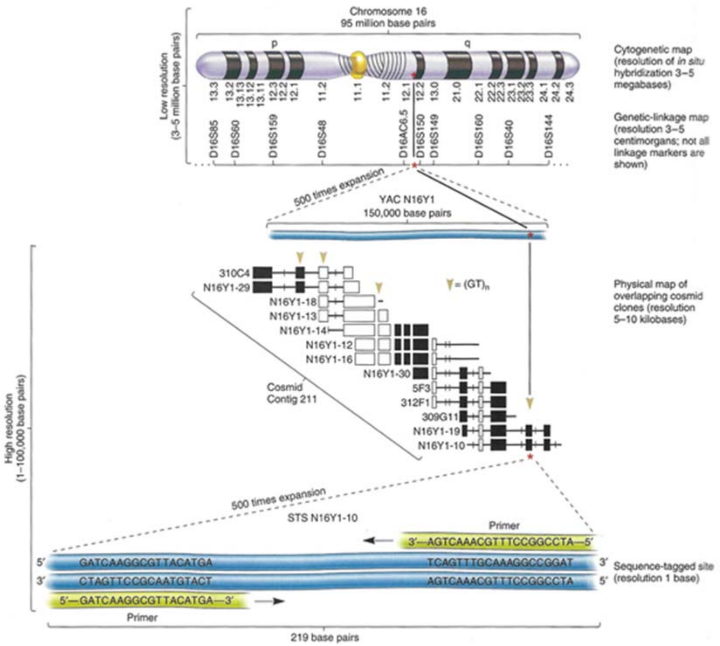
- The 1000 Genomes Project Consortium (2012) *An integrated map of genetic variation from 1,092 human genomes*. Nature 491, 56-65

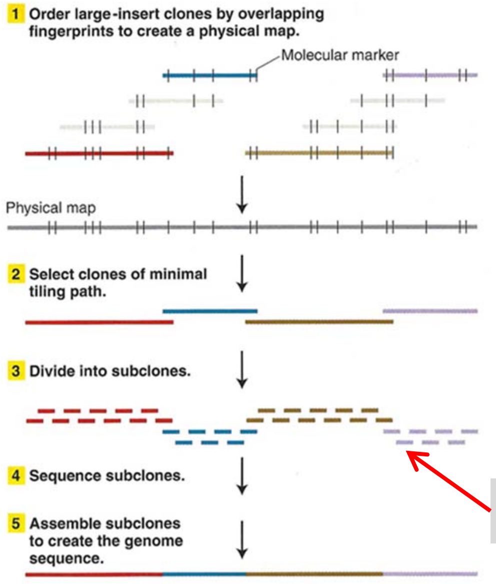
By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. **Here we describe the genomes of 1,092 individuals from 14 populations**, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of **38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions**. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98 % of accessible single nucleotide polymorphisms at a frequency of 1 % in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

- The ENCODE Project Consortium (2012) *An integrated encyclopedia of DNA elements in the human genome*. Nature 489, 57–74

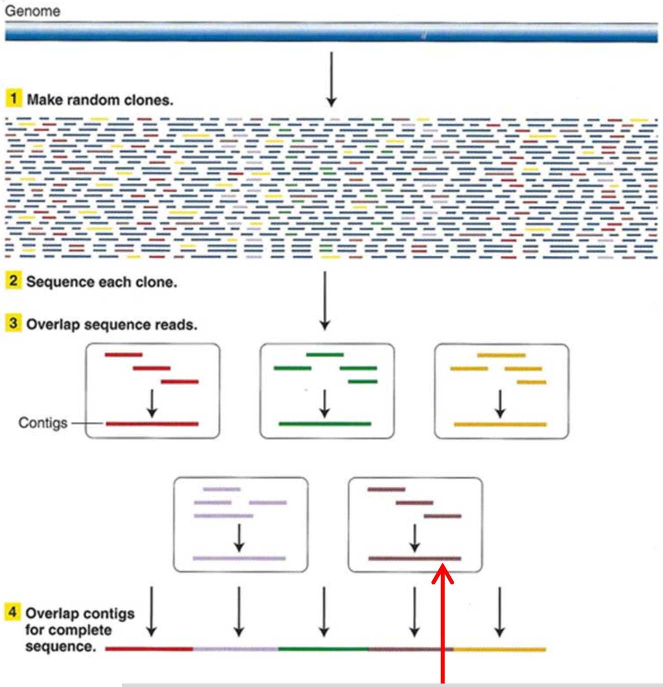
The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

Map-based genome sequencing


Low resolution (3–5 million base pairs)
Chromosome 16
95 million base pairs
p
q
Cytogenetic map (resolution of in situ hybridization 3–5 megabases)
13.3 13.2 13.13 13.12 13.11 12.3 12.2 12.1 11.2 11.1 11.2 12.1 12.2 13.0 21.0 22.1 22.2 22.3 23.1 23.2 23.3 24.1 24.2 24.3
D16S85 D16S60 D16S159 D16S48 D16AC6.5 D16S150 D16S149 D16S160 D16S40 D16S144
Genetic-linkage map (resolution 3–5 centimorgans; not all linkage markers are shown)
500 times expansion
YAC N16Y1
150,000 base pairs
High resolution (1–100,000 base pairs)
310C4
N16Y1-29
N16Y1-18
N16Y1-13
N16Y1-14
N16Y1-12
N16Y1-16
N16Y1-30
5F3
312F1
309G11
N16Y1-19
N16Y1-10
= (GT)n
Cosmid Contig 211
Physical map of overlapping cosmid clones (resolution 5–10 kilobases)
500 times expansion
STS N16Y1-10
Primer
3′—AGTCAAACGTTTCCGGCCTA—5′
5′ GATCAAGGCGTTACATGA TCAGTTTGCAAAGGCCGGAT 3′
3′ CTAGTTCCGCAATGTACT AGTCAAACGTTTCCGGCCTA 5′
5′—GATCAAGGCGTTACATGA—3′
Primer
Sequence-tagged site (resolution 1 base)
219 base pairs

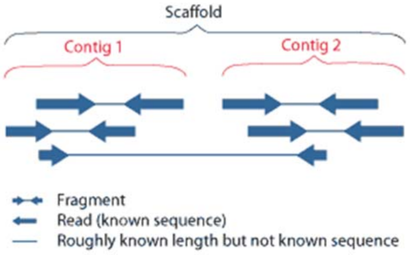

1 Order large-insert clones by overlapping fingerprints to create a physical map.
Molecular marker
Physical map
2 Select clones of minimal tiling path.
3 Divide into subclones.
4 Sequence subclones.
5 Assemble subclones to create the genome sequence.

Shotgun genome sequencing


Genome
1 Make random clones.
2 Sequence each clone.
3 Overlap sequence reads.
Contigs
4 Overlap contigs for complete sequence.

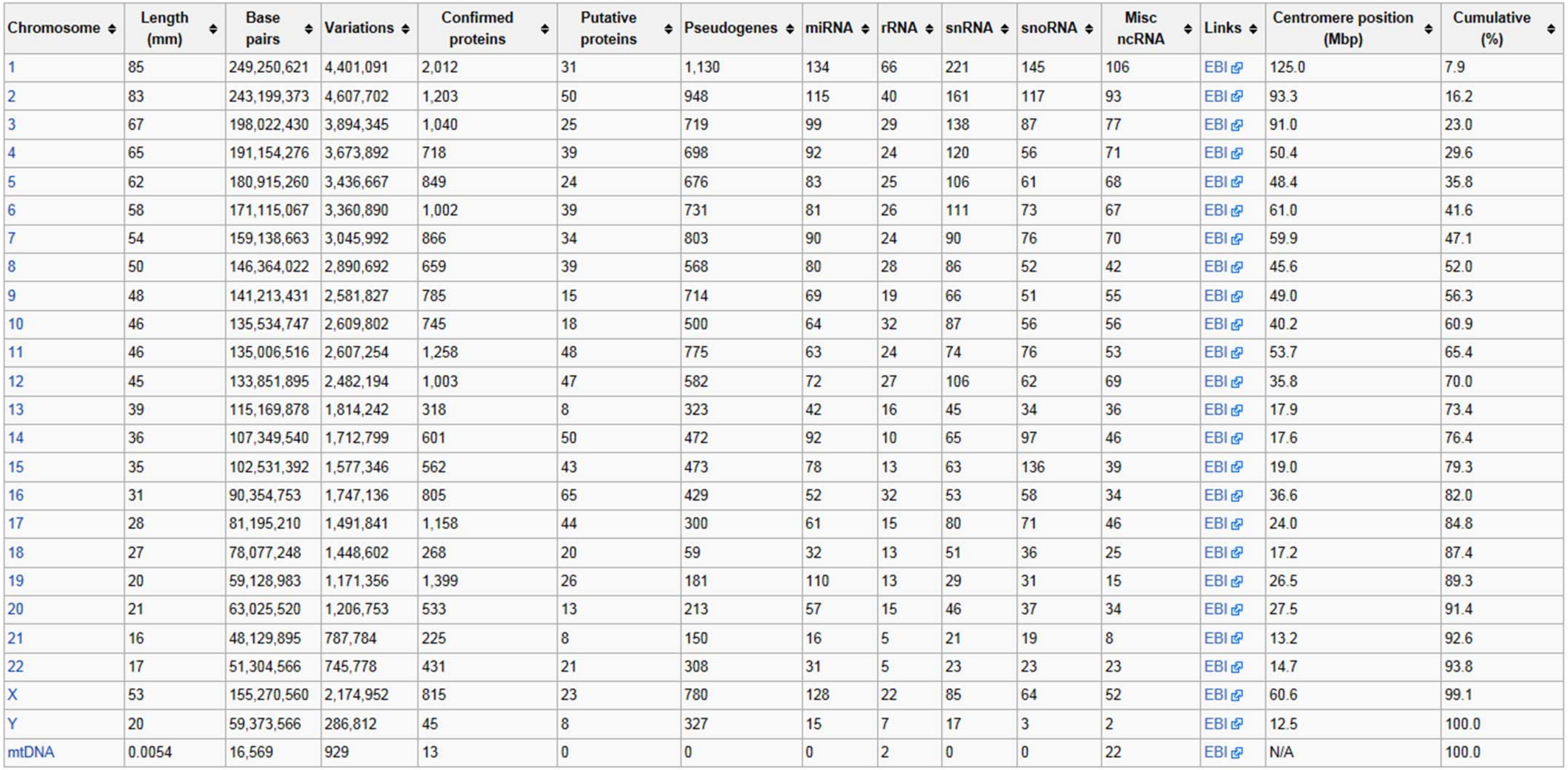
Prekrito zaporedje (“contig“)- prebrana zaporedja, ki se prekrivajo brez vrzeli in imajo visoko zanesljivost določitve.

Ogrodje (“scaffold“)- del genoma rekonstruiran iz odčitkov. Vsebuje prekrita zaporedja in vrzeli. Navadno se uporablja parne odčitke (“pair-end reads“)



Odčitek (“read“)- eno prebrano zaporedje

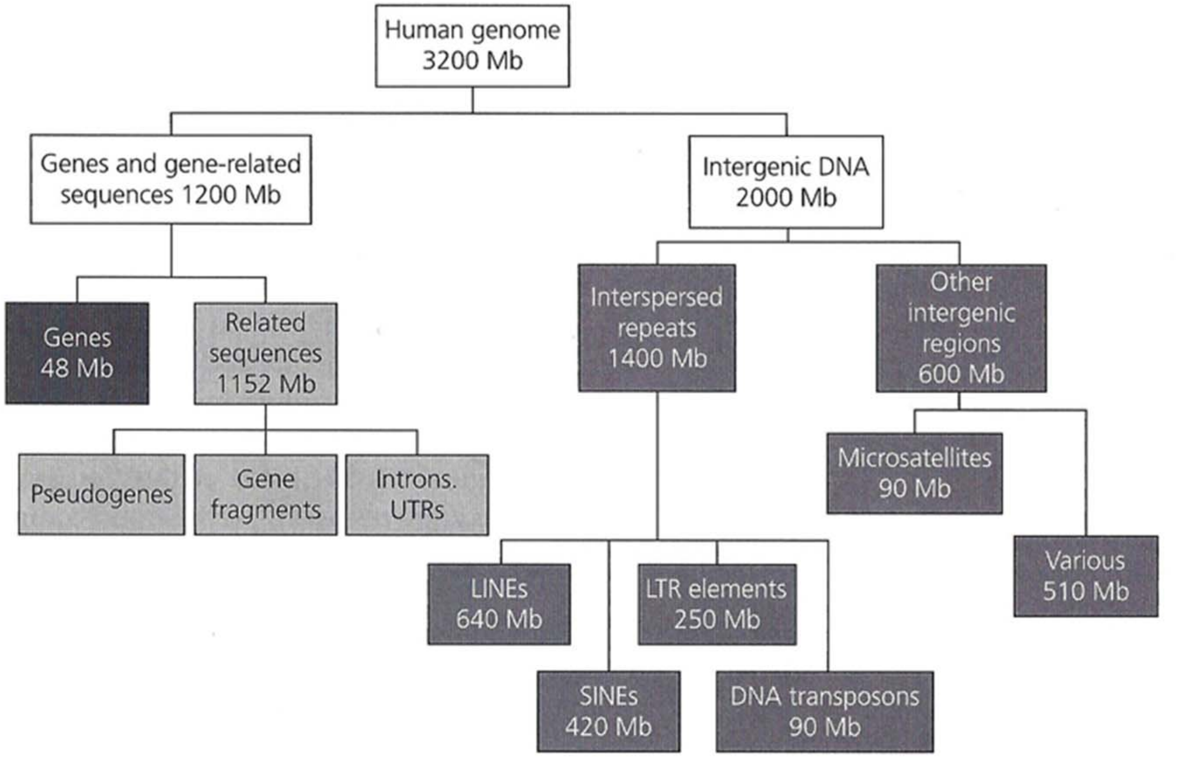
Vir: http://www.discoveryandinnovation.com/BIOL202



Chromosome	Length (mm)	Base pairs	Variations	Confirmed proteins	Putative proteins	Pseudogenes	miRNA	rRNA	snRNA	snoRNA	Misc ncRNA	Links	Centromere position (Mbp)	Cumulative (%)
1	85	249,250,621	4,401,091	2,012	31	1,130	134	66	221	145	106	EBI	125.0	7.9
2	83	243,199,373	4,607,702	1,203	50	948	115	40	161	117	93	EBI	93.3	16.2
3	67	198,022,430	3,894,345	1,040	25	719	99	29	138	87	77	EBI	91.0	23.0
4	65	191,154,276	3,673,892	718	39	698	92	24	120	56	71	EBI	50.4	29.6
5	62	180,915,260	3,436,667	849	24	676	83	25	106	61	68	EBI	48.4	35.8
6	58	171,115,067	3,360,890	1,002	39	731	81	26	111	73	67	EBI	61.0	41.6
7	54	159,138,663	3,045,992	866	34	803	90	24	90	76	70	EBI	59.9	47.1
8	50	146,364,022	2,890,692	659	39	568	80	28	86	52	42	EBI	45.6	52.0
9	48	141,213,431	2,581,827	785	15	714	69	19	66	51	55	EBI	49.0	56.3
10	46	135,534,747	2,609,802	745	18	500	64	32	87	56	56	EBI	40.2	60.9
11	46	135,006,516	2,607,254	1,258	48	775	63	24	74	76	53	EBI	53.7	65.4
12	45	133,851,895	2,482,194	1,003	47	582	72	27	106	62	69	EBI	35.8	70.0
13	39	115,169,878	1,814,242	318	8	323	42	16	45	34	36	EBI	17.9	73.4
14	36	107,349,540	1,712,799	601	50	472	92	10	65	97	46	EBI	17.6	76.4
15	35	102,531,392	1,577,346	562	43	473	78	13	63	136	39	EBI	19.0	79.3
16	31	90,354,753	1,747,136	805	65	429	52	32	53	58	34	EBI	36.6	82.0
17	28	81,195,210	1,491,841	1,158	44	300	61	15	80	71	46	EBI	24.0	84.8
18	27	78,077,248	1,448,602	268	20	59	32	13	51	36	25	EBI	17.2	87.4
19	20	59,128,983	1,171,356	1,399	26	181	110	13	29	31	15	EBI	26.5	89.3
20	21	63,025,520	1,206,753	533	13	213	57	15	46	37	34	EBI	27.5	91.4
21	16	48,129,895	787,784	225	8	150	16	5	21	19	8	EBI	13.2	92.6
22	17	51,304,566	745,778	431	21	308	31	5	23	23	23	EBI	14.7	93.8
X	53	155,270,560	2,174,952	815	23	780	128	22	85	64	52	EBI	60.6	99.1
Y	20	59,373,566	286,812	45	8	327	15	7	17	3	2	EBI	12.5	100.0
mtDNA	0.0054	16,569	929	13	0	0	0	2	0	0	22	EBI	N/A	100.0

Vir: Wikipedia

Overview of human genome



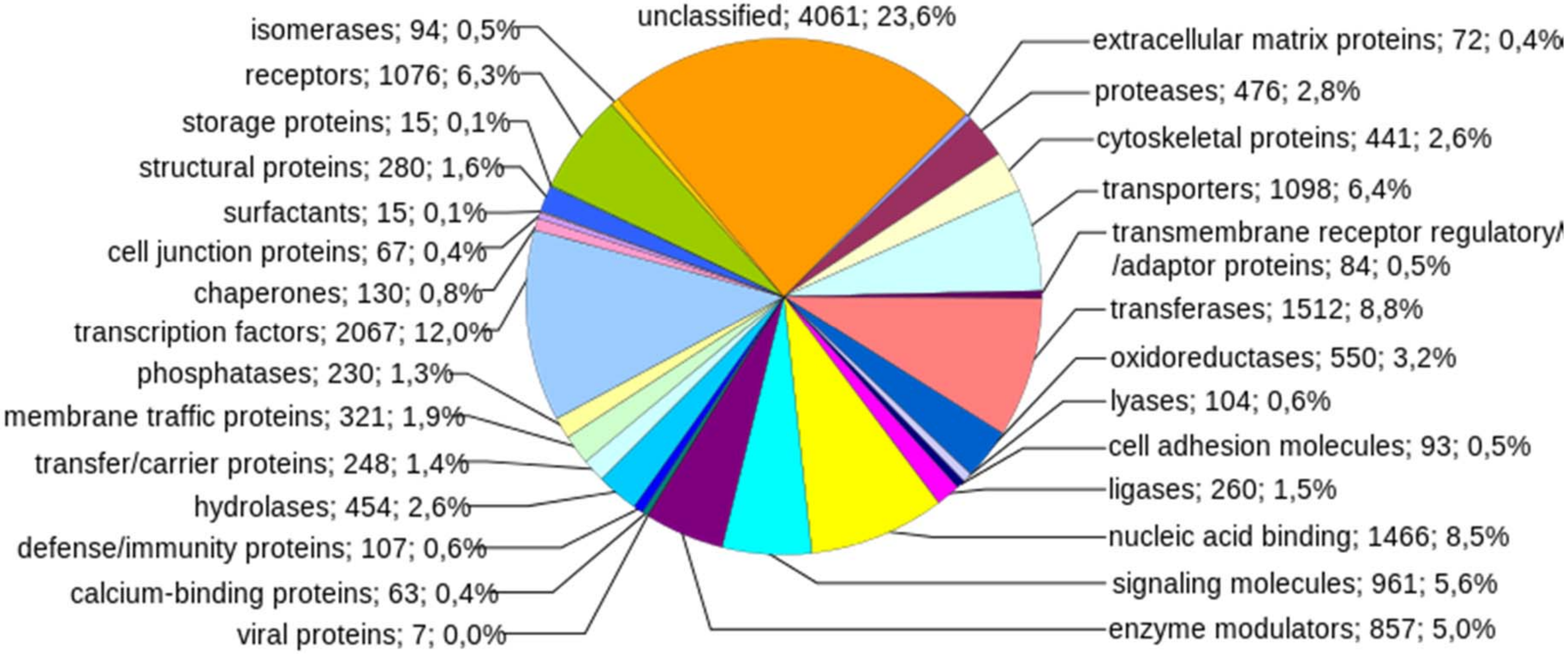
Ponovitve zaporedij v genomu:

- transpozoni
- enostavne ponovitve, npr. 3-100, zaporedja nukleotidov (mikrosateliti). 2-6 nukleotidov, npr. $(CA)_n$
- večje ponovitve (10-300 kb)

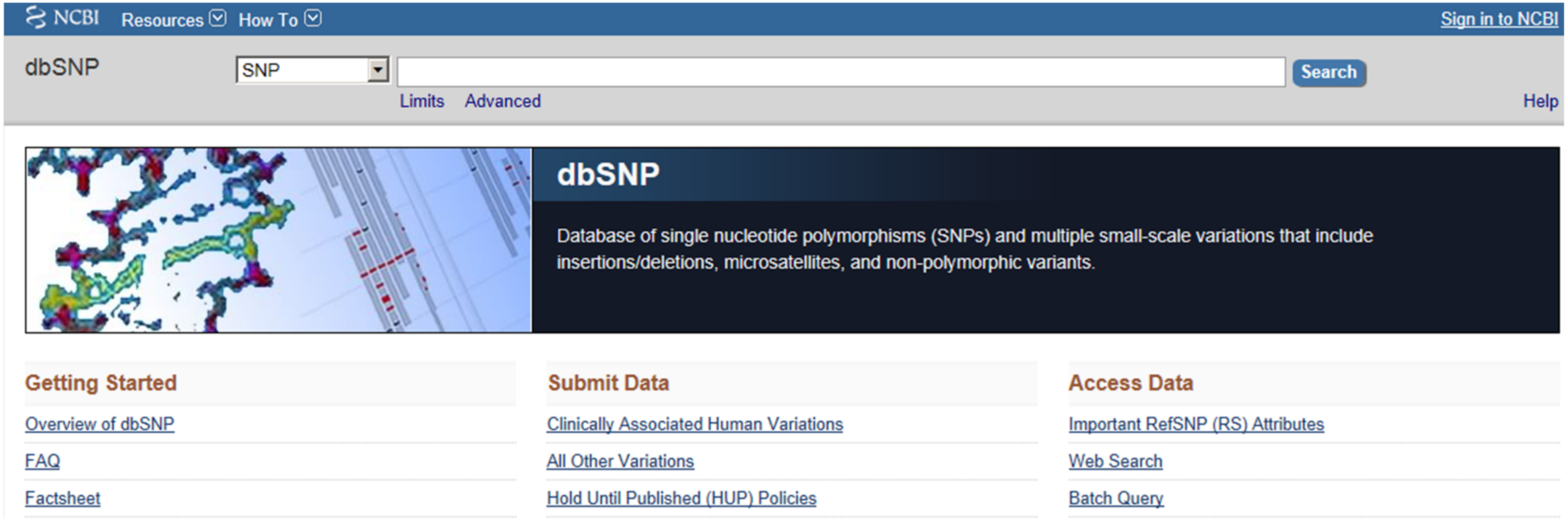
Vir: http://www.discoveryandinnovation.com/BIOL202

20.687 genov, ki kodirajo proteine

Funkcijska delitev človeških genov. Največje družine: vezavni proteini za nukleinske kisline transkripcijski faktorji, encimi, transporterji, receptorji...



Vir: Wikipedia



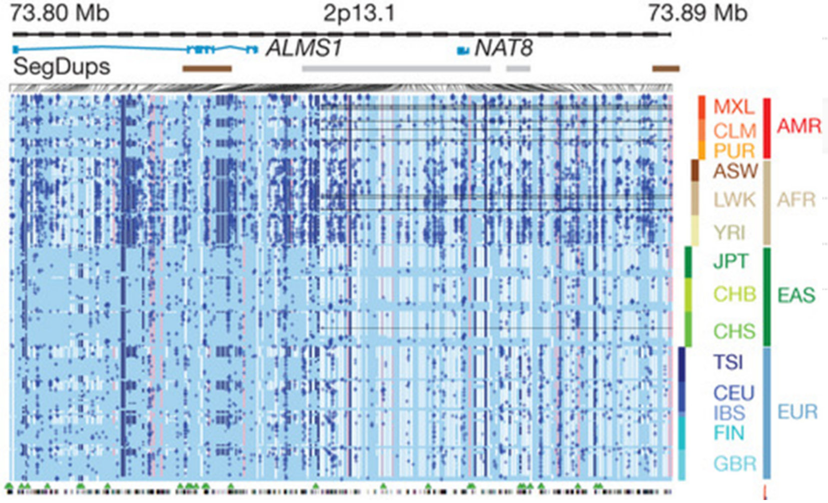
dbSNP

dbSNP

Database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations that include insertions/deletions, microsatellites, and non-polymorphic variants.

Getting Started	Submit Data	Access Data
Overview of dbSNP	Clinically Associated Human Variations	Important RefSNP (RS) Attributes
FAQ	All Other Variations	Web Search
Factsheet	Hold Until Published (HUP) Policies	Batch Query

The NCBI Short Genetic Variations (SNV) database, also known as dbSNP, catalogs short variations in nucleotide sequences from a wide range of organisms. **These variations include single nucleotide variations, short nucleotide insertions and deletions, short tandem repeats and microsatellites.**
Some of these rare human entries have additional information associated with them, including disease associations, genotype information and allele origin, as some variations are somatic rather than germline events.

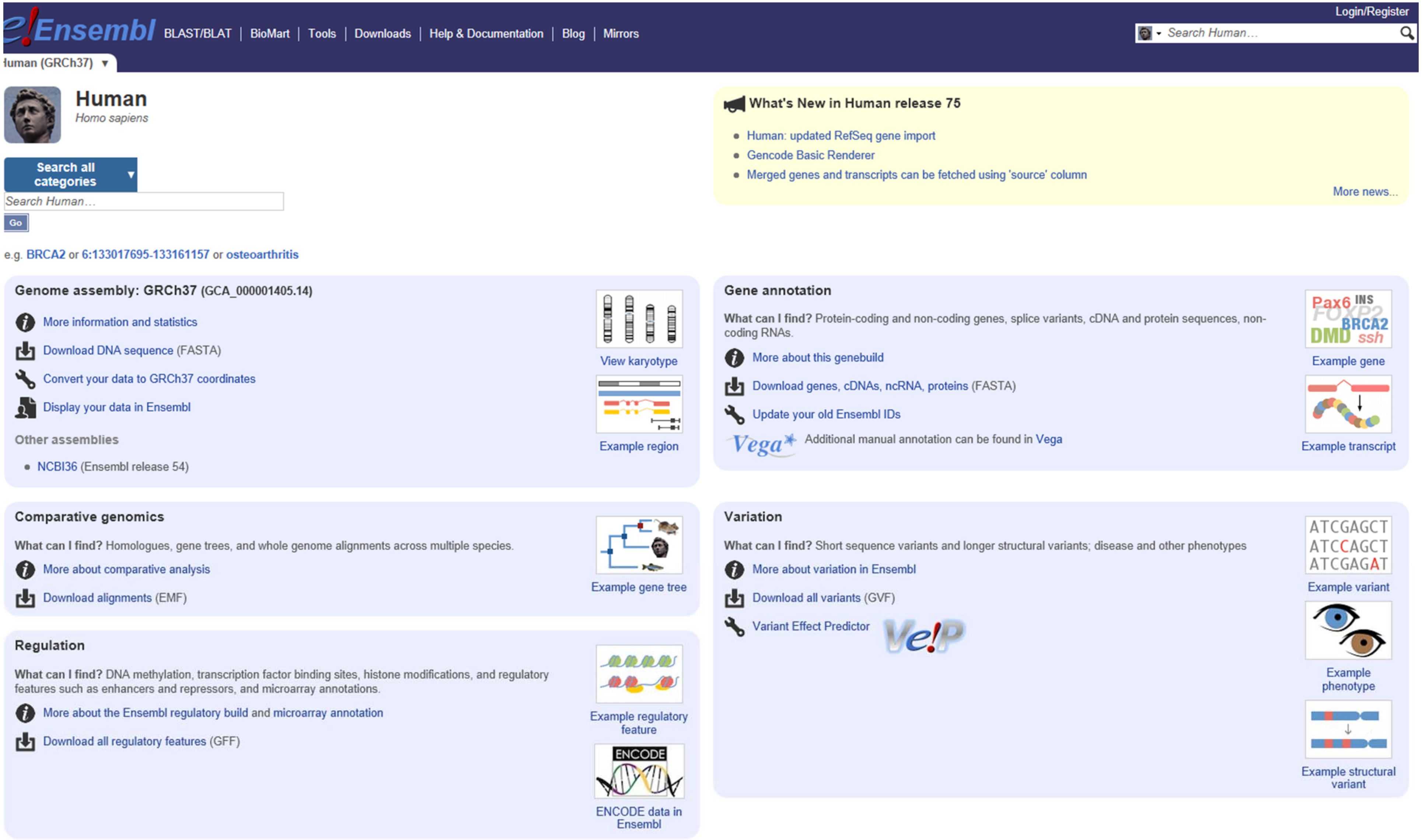


The 1000 Genomes Project Consortium (2012) *An integrated map of genetic variation from 1,092 human genomes*. Nature 491, 56-65

...By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of **38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions**. ...

Genomski brskalnik- Ensembl

http://www.ensembl.org/Homo_sapiens



Human

Homo sapiens

Search Human…

e.g. BRCA2 or 6:133017695-133161157 or osteoarthritis

What's New in Human release 75

- Human: updated RefSeq gene import
- Gencode Basic Renderer
- Merged genes and transcripts can be fetched using 'source' column

More news...

Genome assembly: GRCh37 (GCA_000001405.14)

More information and statistics

Download DNA sequence (FASTA)

Convert your data to GRCh37 coordinates

Display your data in Ensembl

Other assemblies

- NCBI36 (Ensembl release 54)

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

More about this genebuild

Download genes, cDNAs, ncRNA, proteins (FASTA)

Update your old Ensembl IDs

Vega Additional manual annotation can be found in Vega

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

More about comparative analysis

Download alignments (EMF)

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

More about variation in Ensembl

Download all variants (GVF)

Variant Effect Predictor

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations.

More about the Ensembl regulatory build and microarray annotation

Download all regulatory features (GFF)