Sequencing Genomes

Human Genome Project:
History, results and impact

MUDr. Jan Plátenník, PhD.

(December 2010)

Beginnings of sequencing

- **1965: Sequence of a yeast tRNA (80 bp) determined**
- **1977: Sanger's and Maxam & Gilbert's techniques invented**
- **1981: Sequence of human mitochondrial DNA (16.5 kbp)**
- **1983: Sequence of bacteriophage T7 (40 kbp)**
- **1984: Epstein & Barr's Virus (170 kbp)**

Homo sapiens

- 1985-1990: Discussion on human genome sequencing
  - "dangerous" - "meaningless" - "impossible to do"
- 1988-1990: Foundation of HUMAN GENOME PROJECT
  - International collaboration: HUGO (Human Genome Organisation)
  - Aims:
    - genetic map of human genome
    - physical map: marker every 100 kbp
    - sequencing of model organisms (E. coli, S. cerevisiae, C. elegans, Drosophila, mouse)
    - find all human genes (estim. 60-80 tisíc)
    - sequence all human genome (estim. 4000 Mbp) by 2005

Other genomes

- July 1995: Haemophilus influenzae (1.8 Mbp) ... First genome of independent organism
- October 1996: Saccharomyces cerevisiae (12 Mbp) ... First Eukaryota
- December 1998: Caenorhabditis elegans (100 Mbp) ... First Metazoa

May 1998:

- Craig Venter launches private biotechnology company CELERA GENOMICS, Inc. and announces intention to sequence whole human genome in just 3 years and 300 mil. USD using the *whole-genome shotgun* approach.
- The publicly funded HGP in that time: sequenced cca 4 % of the genome

March 2000:

- Celera Genomics & academic collaborators publish draft genome of Drosophila melanogaster (cca 2/3 from 180 Mbp)
- ... *whole-genome shotgun* is feasible for large genomes as well
- ... ... Human genome: competition between Human Genome Project and Celera Genomics

February 2001:

- International Human Genome Sequencing Consortium publishes draft of human genome in Nature (Feb. 15th 2001)
- Celera Genomics, Inc. publishes human genome sequence in Science (Feb. 16th 2001)

International Human Genome Sequencing Consortium (Human Genome Project, HGP)

- Open to co-operation from any country
- 20 laboratories from USA, Great Britain, Japan, France, Germany and China
- About 2800 co-workers, main coordinator: Francis Collins, NIH
- Publicly funded
- Approach: *clone-by-clone*
- Draft (Nature 15.2.2001): 90 % of euchromatin (2.95 Gbp, whole genome 3.2 Gbp). 25 % finished.
- Results: every sequence made publicly available without restriction within 24 hours (in the Internet), permanent updating,
- Final version: 14.4. 2003 (50 year-anniversary of discovery of DNA double helix).

Clone-by-clone

genomic DNA

↓

fragments cca 150,000 bp

↓

cloning in BAC (*bacterial artificial chromosome*)

↓

clones positioned in the genome using physical maps
(STS - *sequence tagged site*, *fingerprint* - cleavage by restrictases)

↓

digestion of every clone to short fragments cca 500 bp

↓

sequencing

↓

assembly of each clone sequence with computer

Celera Genomics, Inc.

- Private biotechnology company, based in Rockville, Maryland, USA. President Craig Venter.
- Investments into automation and computer processing, few dozens employees
- Approach: *whole-genome shotgun* + utilised publicly shared data from HGP.
- Publication in Science 16.2.2001: sequence of euchromatin (2.91 Gbp)
- Results: raw data temporarily available at company www site, but all other updates and annotations for commercial purpose.

Whole-genome shotgun

genomic DNA

↓

fragments 2, 10, 50 kbp

↓

cloned in plasmids E.coli

↓

sequencing

↓

sequence assembly using sophisticated computer algorithms

Advance in sequencing

1985: 500 bp /lab and day

- still the Sanger dideoxynucleotide technique, but
  - capillary electrophoresis instead of gel
  - fluorescence markers instead radioactivity
  - full automatisation & robotisation
  - computer power

2000: 175,000 bp /day (Celera)

1000 bp/sec. (HGP)

Sequencing continues...

- **Human genome now:** finished (99 % of euchromatin)
- **Fugu rubripes:** draft of genome in August 2002
- **Mouse:**
  - Celera Genomics: draft in June 2001
  - Mouse Genome Sequencing Consortium: Nature, December 2002
- **Laboratory rat:** draft in March 2004
- **Chimpanzee:** September 2005

- **... and many other genomes:** malaria (the cause Plasmodium falciparum and carrier Anopheles gambiae), zebrafish, rice, dog, cattle, sheep, pig, chicken, honeybee, mammoth etc.

Public databases of DNA/RNA sequences

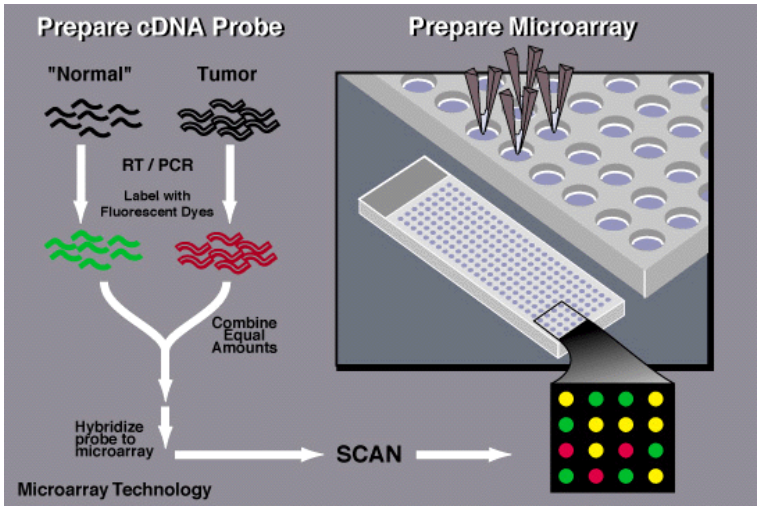
- GenBank, National Center for Biotechnology Information (NCBI), Bethesda, Maryland, USA
- EMBL-Bank, EMBL's European Bioinformatics Institute, Hinxton, UK
- DNA Data Bank of Japan, National Institute of Genetics, Mishima, Japan

- **Content of all three databases in August 2005 exceeded 100,000,000,000 base pairs (100 Gb) ... from genes/genomes of 165,000 species of organisms**

Research in "postgenomic" age

- **New approaches to study genes & proteins:**
  - **GENOMICS** ... analysis of whole genome and its expression
  - **PROTEOMICS** ... analysis of whole proteome, i.e. all proteins in given tissue or organism
  - **BIOINFORMATICS** ... processing, analysis and interpretation of large data sets (NA or protein sequences, gene arrays, 3D protein structures etc. Experiments ***in silico***
- **Rapid development of new technologies:**
  - e.g. **DNA Microarray** - expression of thousands of genes can be studied simultaneously

DNA Microarray (" DNA chip")



Human Genome Project: Results

The Human Genome

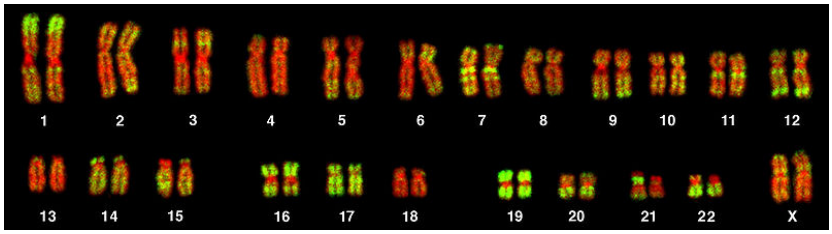


Fig. from Bolzer et al. 2005, PLoS Biol. 3(5): e157 DOI: 10.1371/journal.pbio.0030157

Haploid genome: 3 billion base pairs divided to 23 chromosomes

- 1 meter of DNA if extended
- 750 Mb (1 CD)
- 2 million standard printed pages

  (50 letters/line, 30 lines/page)

Classification of eukaryotic genomic DNA:

- degree of condensation:
  - euchromatin
  - heterochromatin
- repetitivity:
  - highly repetitive
  - moderately repetitive
  - non-repetitive (single-copy)
- function:
  - structural (centromers, telomers)
  - coding (protein or RNA)
  - non-coding
    - regulatory regions
    - spacer, “junk DNA”, “selfish DNA”

Experiments with DNA denaturation & reassociation:

- **Rapid reassociation (10-15%):**
  - **simple-sequence DNA:**
    - **heterochromatin in telomeres and centromeres**
    - **satellites, minisatellites.**
- **Intermediate reassociation (25- 40%):**
  - **intermediate-repeat DNA (mobile elements, transposons)**
  - **tandemly repeated genes for rRNA, histones**
- **Slow reassociation (50- 60%):**
  - **single-copy DNA**
    - **genes**
    - **other non-classified spacer DNA**

Classification of eukaryotic genomic DNA:

- Genes encoding proteins
- Tandemly repeated genes coding for rRNAs, tRNAs, and histones
- Repetitive DNA
- Spacer DNA

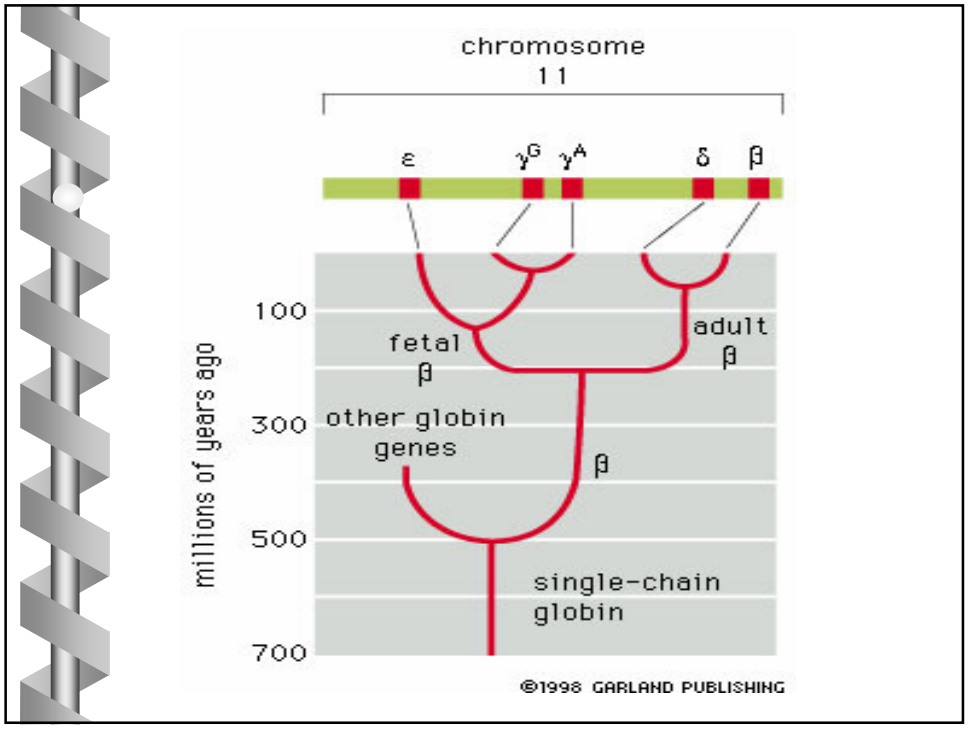
Classification of eukaryotic genomic DNA:

- Genes encoding proteins
  - about 25 % of genome, but from this only 5 % (1.4 % of genome) are exons
- Tandemly repeated genes coding for rRNAs, tRNAs, and histones
- Repetitive DNA
- Spacer DNA

Genes are not placed evenly in genome

- Big differences among chromosomes:
  - chromosome 1: 2968 genes
  - chromosome Y: 231 genes
- Regions rich in genes (“cities”)
  - more C and G
- Regions poor in genes (“deserts”)
  - more A and T
- CpG islands - “barriers between cities and deserts” ... regulation of gene activity

- **Solitary gene:**
  - **present as a single copy in the whole haploid genome (about half of genes)**
- **Gene family:**
  - **cluster of related genes that in evolution originated from a single ancestor, gradual diversification of sequence and function**
- **Pseudogene:**
  - **gene where mutations accumulated to an extent that it cannot be transcribed (in the whole genome > 20 000 !)**
- **Processed pseudogene:**
  - **pseudogene that originated from reverse transcription of mRNA and integration to genome**



Number of genes in human genome

20,000 – 25,000

Much less than expected!

... not related to organism complexity:

Sacch. cerevisiae	6,000 genes
C. elegans	18,000 genes
Drosophila	13,000 genes
Arabidopsis thaliana	26,000 genes

Comparison of human/mouse genome with genomes of lower organisms (C. elegans, Drosophila):

- low gene density, longer introns

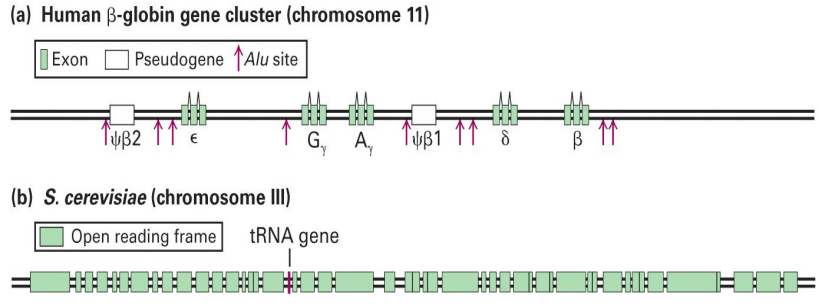
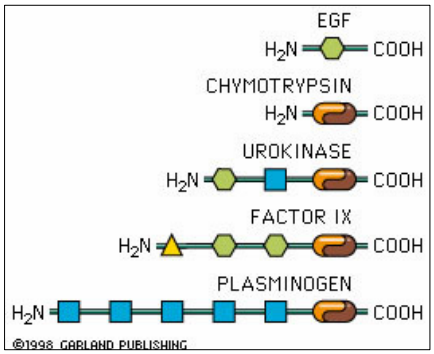


Figure from: Lodish, H. et al.: Molecular Cell Biology (5th ed.), W.H.Freeman, New York 2004.

How to find genes in genomes:

- **Bacteria, yeast:**
  - **open reading frames (ORFs)**
- **Higher organisms:**
  - **hybridisation/comparison with cDNA or EST (expressed sequence tag = part cDNA)**
  - **by similarity with other known genes**
  - **prediction of recognition sites for splicing**
  - **by similarity with genomes of other organisms**

- only about 7 % of protein domains entirely new in vertebrates, but
  - expansion of protein families
  - new combinations of domains; and proteins more complex (more domains per protein)



  - more proteins from one gene - alternative splicing in up to 60 %

Comparison of human/mouse genome with genomes of lower organisms (C. elegans, Drosophila):

- expansion of gene families /new families related to:
  - blood clotting
  - acquired (specific) immunity
  - nervous system
  - intra- and intercellular communication
  - regulation of gene expression
  - programmed cell death (apoptosis)

Classification of eukaryotic genomic DNA:

- Genes encoding proteins
- Tandemly repeated genes encoding rRNAs, tRNAs, and histones
  - can be considered as intermediate repeat DNA
  - number of identical copies of the same gene in tandem array, in order to increase productivity of transcription
  - genes for ribosomal RNA in eukaryotes: >100 copies
- Repetitive DNA
- Spacer DNA

Classification of eukaryotic genomic DNA:

- Genes encoding proteins
- Tandemly repeated genes coding for rRNAs, tRNAs, and histones
- Repetitive DNA
  - simple-sequence DNA: highly repetitive, 3% of euchromatin (minisatellites) + all heterochromatin (centromeres, telomeres, 8% of genome, yet unsequenced)
  - interspersed repeats (moderately repetitive DNA) = mobile elements (transposons), 45 % of euchromatin
- Spacer DNA

Mobile elements (transposons):

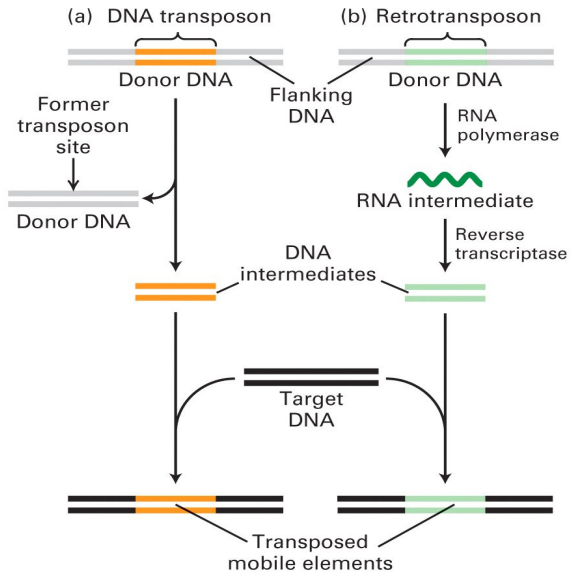


Figure from: Lodish, H. et al.: Molecular Cell Biology (5th ed.), W.H.Freeman, New York 2004.

Mobile (parasitic) elements in mammalian genome:

- LINEs (long-interspersed repeats),
  - 6-8 kb, e.g. L1, encode 2 proteins (one is reverse transcriptase)
- SINEs (short-interspersed repeats),
  - 100-300 bp, e.g. Alu, code no protein, proliferation depends on LINEs, origin: small non-coding cellular RNA
- Virus-like retrotransposons
  - 6-11 kb (or shorter), retroviruses without gene for protein envelope (env)
- DNA transposons
  - 2-3 kb (or shorter), encode transposase, cut & paste or copy & paste in genome without ever transcribed to RNA

Census of parasitic elements in human genome:

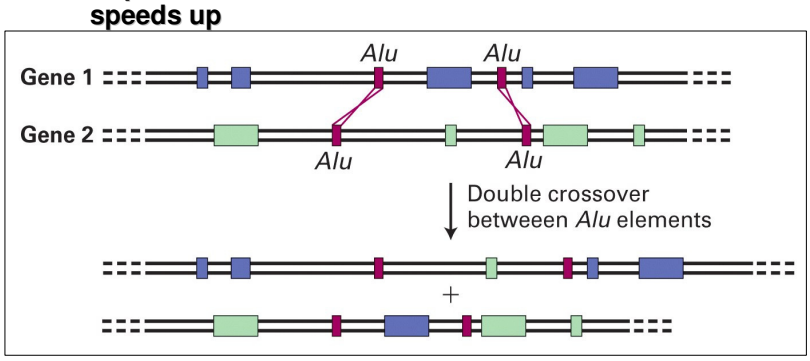
LINEs:	850,000x	21 % genome
SINEs:	1,500,000x	13 % genome
Retrovirus-like:	450,000x	8 % genome
DNA transposons:	300,000x	3 % genome

- In the whole genome only small number of active LINEs and endogenous retroviruses.
- All others: old, mutated, no longer active "molecular fossils".
- Mouse genome contains much more functional transposons (...why?).

Significance of transposons in human genome

- Expansion of these elements in genome occurred in several waves long time ago, some might be earlier than oligocene mammalian radiation
- Responsible for part of spontaneous mutations (1:600 in human)
- For an individual, activity of transposons is malign (inactivation of a gene)
  - ... transposition of L1 (LINEs) documented as a rare cause of inborn errors in humans

- **But for a species evolution it has positive effects:**
  - **Repeat elements -> more recombinations -> evolution speeds up**



  - **Reverse transcription of mRNA - "processed pseudogenes" (occasionally a functional gene arises)**
  - **Alu (SINEs) - more abundant in gene-enriched areas; may encode an RNA with function in the cell**

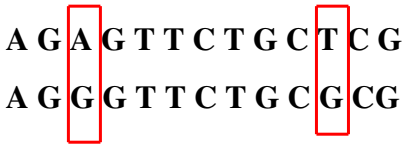
Figure from: Lodish, H. et al.: Molecular Cell Biology (5th ed.), W.H.Freeman, New York 2004.

Classification of eukaryotic genomic DNA:

- **Genes encoding proteins**
- **Tandemly repeated genes coding for rRNAs, tRNAs, and histones**
- **Repetitive DNA**
- **Spacer DNA**
  - **non-repetitive, non-coding, about quarter of genome. Probably old dead transposons mutated too much to be recognised by computer analysis**

Vast majority of our DNA has originated from reverse transcription of retrotransposons

Single Nucleotide Polymorphism (SNP)



Occurs on average in one base per 1000 bp, i.e. in 0.1 % of human genome

About 10 millions of SNPs with occurrence > 1%

Coding/non-coding

Protein structure changed/unchanged

Human genetic variation

- Two unrelated humans have 99.5% of genome identical
  - Single Nucleotide Polymorphisms: 0.1%
  - Copy number variation (insertions, deletions, duplications): 0.4%
- Epigenetics (methylation)
- Variable number tandem repeats (…DNA fingerprinting in forensics)

Human Genome Project: Impact

Benefits of genome sequencing

- **Milestone in the history of mankind**
- **Facilitates research into molecular basis of diseases**
  - "virtual cloning" *in silico* instead laborious positional cloning and sequencing
- **Molecular medicine:**
  - diagnosis at the gene level, early detection of susceptibility to certain diseases
  - gene therapy
  - individual sensitivity to risk factors in the environment (radiation, mutagens) can be assessed

Benefits of genome sequencing

- **New targets for pharmacotherapy**
  - all drugs known so far - interact with only about 500 protein targets !!!
- **Farmakogenomics:**
  - how individual genetic information affects response to medication ... "therapy "tailored" for every patient
- **Study of human evolution and migration**
- **What the genome determines ("nature vs. nurture") and which SNPs cause differences among people**

Benefits of genome sequencing: Where are we now?

- **Rapid development of technology for affordable personal genome sequencing**
- **Exploration of SNP data**

Sequencing must be faster and cheaper in order to become part of routine healthcare...

- **J. Craig Venter Science Foundation, 2003:**
  - **$ 500,000 Genomic Technology Prize** ... for an inventor of new technology capable of mammalian genome sequencing <$ 1000
- **X Prize Foundation, 2006:**

  $ 10,000,000 Archon X Prize for Genomics to

  "the first Team that can build a device and use it to sequence 100 human genomes within 10 days or less, with an accuracy of no more than one error in every 100,000 bases sequenced, with sequences accurately covering at least 98% of the genome, and at a recurring cost of no more than $10,000 per genome."

Second generation of sequencers:

E.g. Illumina Co., Dec. 2008:

- One run (3 days) of Genome Analyzer made by Illumina Inc. = 60 years of work of ABI 3730xl (used by Celera Genomics)
- Cost of one human genome sequencing: 40-50,000 $ (currently <20,000 $)

First individual human genomes sequenced:

2007: Craig Venter, James Watson – both genomes published in the internet

(since then several others)

Second generation of sequencers:

E.g. Illumina Co., Dec. 2008:

- One run (3 days) of Genome Analyzer made by Illumina Inc. = 60 years of work of ABI 3730xl (used by Celera Genomics)
- Cost of one human genome sequencing: 40-50,000 $ (currently <20,000 $)

Race to genome sequencing<1000 $:

- Illumina, Pacific Biosciences, Complete Genomics, Helicos Biosciences, IBM and several others struggle to become the first provider of affordable personal genome sequencer
- … available in 5 years?

Benefits of genome sequencing: Where are we now?

- **Development of technology for affordable personal genome sequencing**
- **Exploration of SNP data**

International HapMap Project

- Further international collaboration 2002-2009
- Genotyping and sequencing of DNA from 270 people from four different populations (USA, Nigeria, Japan, China)
- Aims at finding
  - all important human SNPs (about 10,000,000)
  - their stable combinations (haplotypes)
  - Tag SNP for each haplotype
- Data publicly available for further exploration

Personal/Recreational Genomics

- Analysis of personal SNPs (ancestry, inherited traits, disease risks…)
- Currently about 4 companies:
  - **Personal Genome Project**
  - **deCODE ME**
  - **Navigenics**
  - **23andME**

23andME

- Spit sample sent by DHL, genotyping of 580,000 SNPs
- Ancestry:
  - Paternal (Y chromosome haplogroup)
  - Maternal (mitochondrial DNA haplogroup)
  - Finding relatives, global similarity search
- Health:
  - Disease risk: 92 (25 high confidence)
  - Carrier status: 24
  - Response to drugs: 18 (8 high confidence)
  - Traits: 44 (13 high confidence)

Ethical, legislative and social issues

- **Gene privacy:**
  - who has the right of knowing someone else's genetic information and how it can be used, worries about discrimination by employer, health insurance company...
- **Gene testing**
- **Gene therapy**
- **Behavioural genetics:**
  - how genes determine human behaviour, possible fall into genetic determinism and loss of responsibility for one's own behaviour
- **GM food**
- **Gene patenting:**
  - about one fifth of human genes or their fragments have been patented, mostly by biotechnology companies…

References:

Nature 2001: 409 (6822, 15.2.2001); pp. 813-958

Science 2001: 291 (5507, 16.2.2001); pp.1177-1351

Lodish, H. et al.: Molecular Cell Biology (5th ed.), W.H.Freeman, New York 2004 ("Darnell").

Alberts, B. et al.: Essential Cell Biology, Garland Publishing, Inc., New York 1998

Lecture by dr. M. Lebl: Next generation DNA sequencing (1.LF UK, 1.12.2008)

http://www.ncbi.nlm.nih.gov

http://genomics.energy.gov

http://en.wikipedia.org

http://hapmap.ncbi.nlm.nih.gov

https://www.23andme.com

and internet in general…

Fig. "Human and DNA Shadow": Courtesy of U.S. Department of Energy's Joint Genome Institute, Walnut Creek, CA, http://www.jgi.doe.gov.