Sequencing Genomes

Human Genome Project:
History, results and impact

MUDr. Jan Plátenník, PhD.

(December 2020)

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

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 workers, main coordinator: Francis Collins, NIH
- Publicly funded (about 3 billion USD)
- Approach: *clone-by-clone*
- Results: duty to upload on internet within 24 hours (the Bermuda rule).

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 of employees
- Approach: *whole-genome shotgun* + utilised publicly shared data from HGP.
- 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

February 2001:

- **International Human Genome Sequencing Consortium publishes draft of human genome in Nature (Feb. 15th 2001)**
  - Draft: 90 % euchromatin (2.95 Gbp, whole genome 3.2 Gbp). 25 % definitive.
- **Celera Genomics, Inc. publishes human genome sequence in Science (Feb. 16th 2001)**
  - Sequence of euchromatin (2.91 Gbp)

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...

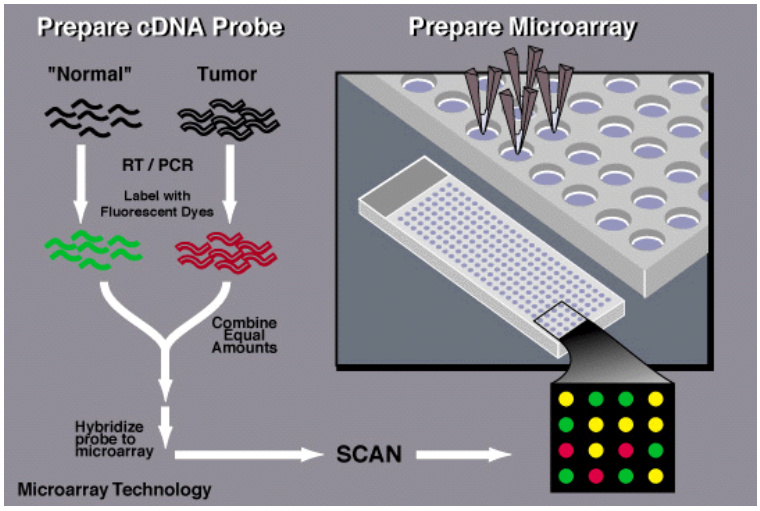
- **Human genome now:** Definitive version announced 14/4/2003 …50 years since DNA double helix. The reference sequence still being updated.
- **Fugu rubripes:** draft of genome in August 2002
- **Mouse:**
  - Celera Genomics: draft in June 2001
  - Mouse Genome Sequencing Consortium: Nature, December 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.

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”)



Single Nucleotide Polymorphism (SNP)

A G A G T T C T G C T C G

A G G G T T C T G C G C G

- 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

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**

Human genetic variation

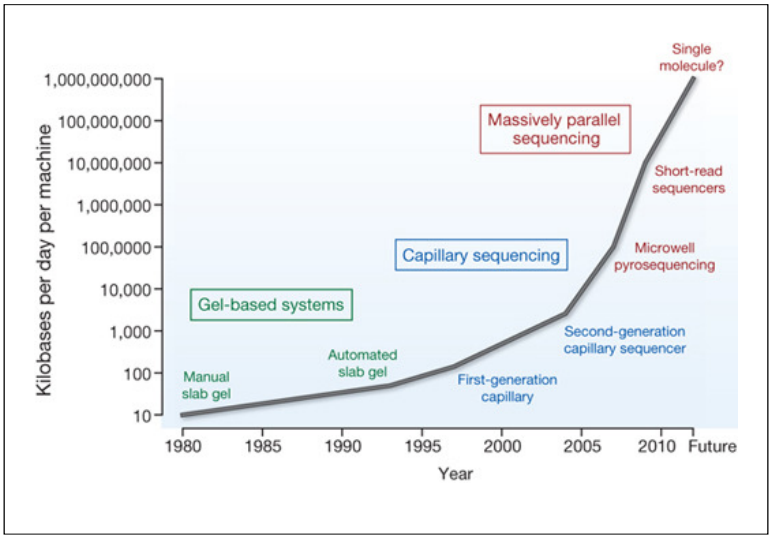
- Two unrelated humans have 99.5% of genome identical
  - Single Nucleotide Polymorphisms: 0.1%
  - Copy number variation (insertions, deletions, duplications): 0.4%
- Variable number tandem repeats (…DNA fingerprinting in forensics)
- Epigenetics (methylation)

Second-generation sequencers:

E.g. Illumina Co., XII/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 $
- First individual human genomes sequenced:
  - 2007: Craig Venter, James Watson – both genomes published in the internet

... and third-generation sequencers



Graph: Nature 458, 719-724 (2009).
Obtained from http://genome.wellcome.ac.uk

Next-Generation Sequencing (NGS)

- 454 pyrosequencing
- Sequencing by synthesis (Illumina)
- SOLiD sequencing by ligation
- Ion Torrent semiconductor sequencing
- DNA nanoball sequencing
- Heliscope single molecule sequencing
- Single molecule real time (SMRT) sequencing
- Nanopore DNA sequencing

Archon X Prize for Genomics

$ 10,000,000

Announced in 2006.

For the first team that succeeds in sequencing of 100 individual human genomes within 30 days in certain requested quality and cost below $1,000 per one genome.

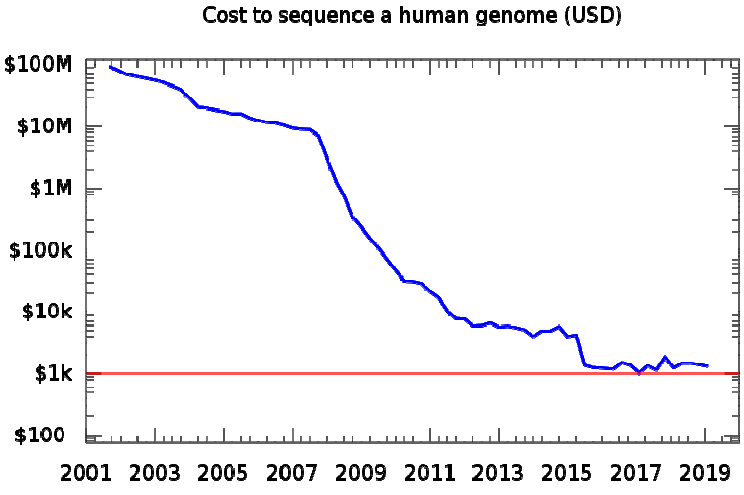
Archon X Prize for Genomics

$ 10,000,000

Announced in 2006.

For the first te… …eeds in sequenci… …ividual human gen… …50 days in certain …quality and cost below …50 per one genome.

Prize cancelled 22/8/2013
„Outpaced by innovation“



(Cost to sequence a human genome according to NHGRI, Wikimedia Commons)

Next-Generation Sequencing (NGS)

Current technology, e.g Illumina HiSeq 3000 or HiSeq 4000:

Sequencing by synthesis (SBS)

Up to 400 Gb/day …12 human genomes or 96 exomes in 3.5 days

www.illumina.com

https://www.youtube.com/watch?v=womKfikWlxM

EARTH BIOGENOME PROJECT

- Sequencing all cca 1.5 million of known eukaryotic species on Earth (… all plants, animals, protozoa and fungi, so far <0.2 % sequenced)
- Launched 1/11/2018, expected to take 10 years, 4.7 billion USD.

Human Genome Project: Results

The Human Genome

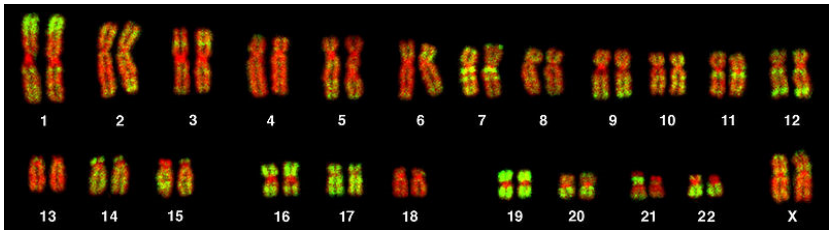
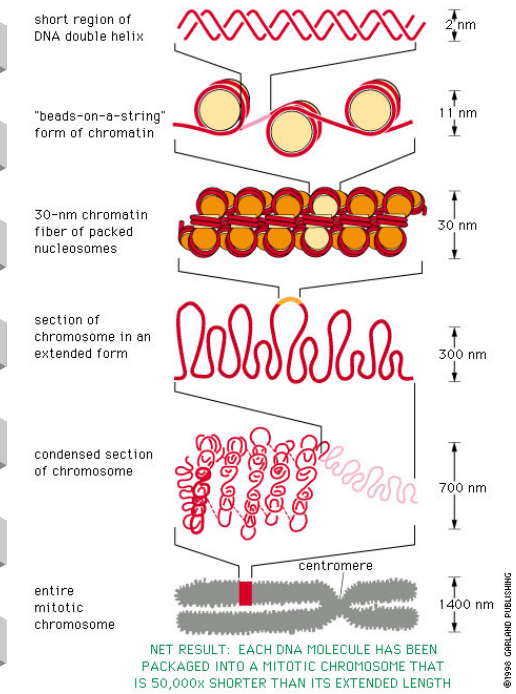


Fig. from Bolzer et al. 2005, PLoS Biol. 3(5): e157 DOI: 10.1371/journal.pbio.0030157

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)



DNA in cell nucleus

Nucleus of typical human cell has diameter 5-8 µm and contains 2 m of DNA

Comparable to a tennis ball into which 20 km of thin thread has been neatly packed.

Classification of eukaryotic genomic DNA:

- Degree of condensation:
  - Euchromatin
  - Heterochromatin (cca 10%, sequencing difficult)
- Repetitivity:
  - Highly repetitive
  - Moderately repetitive
  - Non-repetitive (single-copy)
- Function:
  - Structural (centromeres, telomeres)
  - Coding protein
  - Transcribed to noncoding RNA (introns, rRNA, tRNA, miRNA etc.)
  - Transposons
  - Regulatory sequences
  - Junk…?

Experiments with denaturation & reassociation of DNA:

Rapid reassociation (10-15%):
- highly repetitive DNA
Intermediate reassociation (25- 40%):
- moderately repetitive DNA
Slow reassociation (50- 60%):
- non-repetitive (single copy) DNA

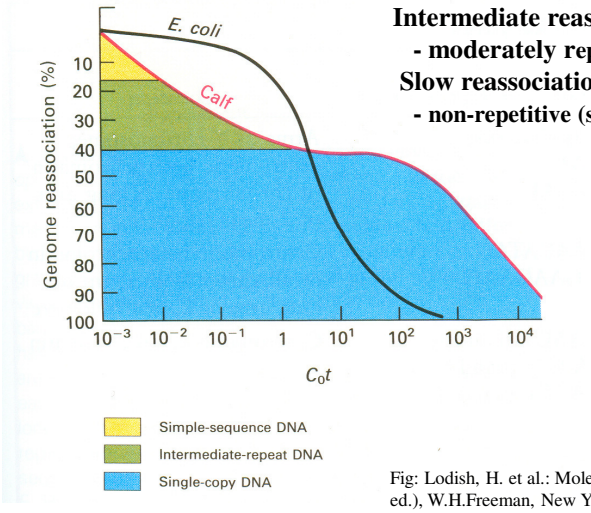


Fig: Lodish, H. et al.: Molecular Cell Biology (3rd ed.), W.H.Freeman, New York 1995.

Classification of eukaryotic genomic DNA:

- **Highly repetitive (simple-sequence DNA):**
  - **All heterochromatin** (centromeres, telomeres, 8% of genome, yet unsequenced)
  - **Minisatellites** (3% of euchromatin)
- **Moderately repetitive:**
  - **Tandemly repeated genes for rRNA, tRNA and histones** (more identical copies to achieve high transcription efficiency, e.g. rRNA genes in eukaryotes >100 copies)
  - **Transpozons**
- **Non-repetitive:**
  - **Protein genes**
  - **Genes for noncoding RNA**
  - **Regulatory sequences**

Eukaryotic GENE

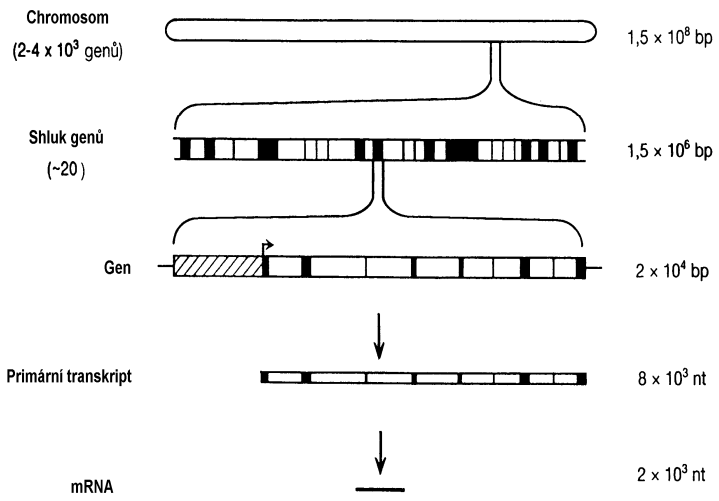
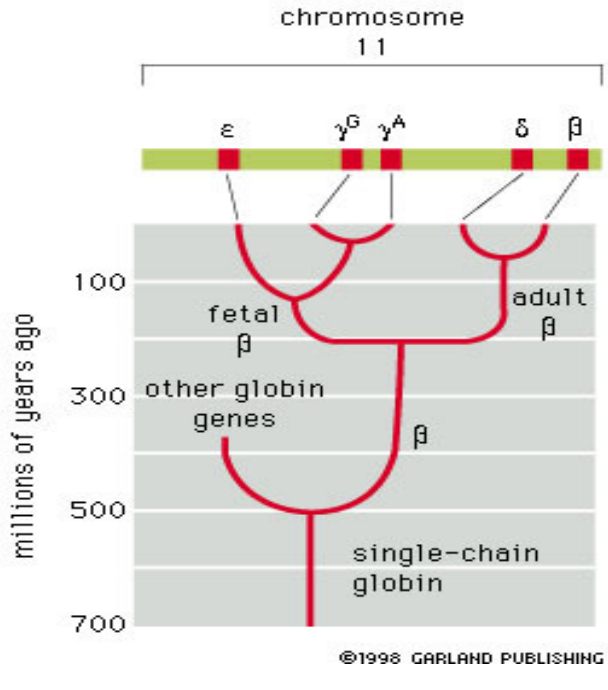


Fig: Murray, R.K. et al.: Harperova biochemie, Appleton & Lange 1993, in Czech H&H 2002.

Genes are not placed evenly in genome

- Big differences among chromosomes:
  - chromosome 1: 2,057 coding genes
  - chromosome Y: 65 coding genes
- Regions rich in genes (“cities”)
  - more C and G
- Regions poor in genes (“deserts”)
  - more A and T, up to 3 Mb!
- 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)
- **Tandemly repeated genes for rRNA, tRNA, histones**
- **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 ("molecular fossil")
- **Processed pseudogene:**
  - originated from reverse transcription of mRNA and integration to genome ("dead on arrival")



Number of genes in human genome

- **Coding genes: 20,448**
- **Non coding genes: 23,997**
  - **Small non coding genes** (<200 bp, rRNA, miRNA, ncRNA, snRNA, snoRNA …): **4,867**
  - **Long non coding genes** (>200 bp, various non coding RNA): **16,909**
  - **Misc. non coding genes**: **2,221**
- **Pseudogenes: 15,217**
- **Gene transcripts in total: 232,186**

Ensembl release 102, Nov. 2020 (www.ensembl.org)

Protein genes in human genome

cca 20 400

About 25% genome transcribed to pre-mRNA,

From this only 5% are exons

…Human EXOME: cca 1.5 % of genome

Number of genes does not reflect organism complexity?!

Sacch. cerevisiae	6,600 genes
C. elegans	20,191 genes
Drosophila	13,931 genes
Arabidopsis thaliana	27,655 genes

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

- low gene density, longer introns

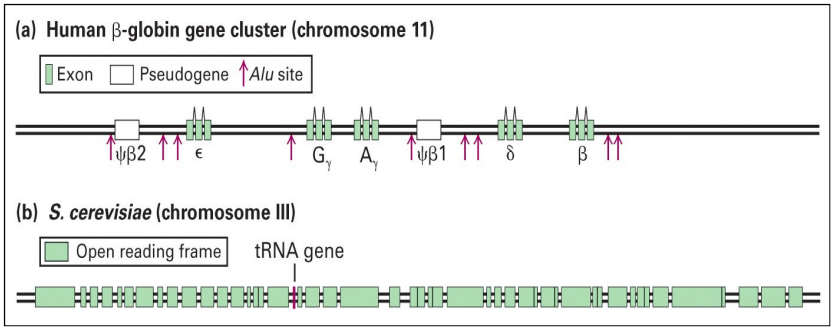


Figure from: Lodish, H. et al.: Molecular Cell Biology (5th ed.), W.H.Freeman, New York 2004.

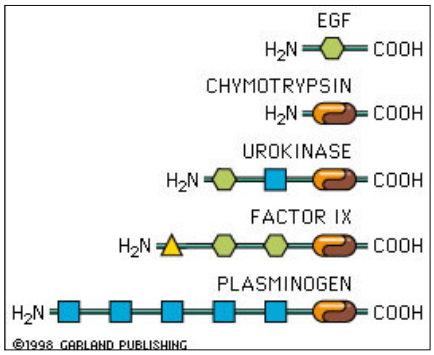
How to find genes in genomes:

- **Bacteria, yeast:**
  - open reading frames (ORFs)
- **Higher organisms:**
  - comparison with transcriptome (RNA-seq)
  - by similarity with other known genes
  - prediction of recognition sites for splicing
  - comparison with genomes of other organisms

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

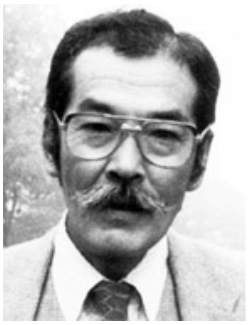
- expansion of gene families /new families related to:
  - blood clotting
  - acquired (specific) immunity
  - nervous system
  - intra- and intercellular communication
  - regulation of gene expression
  - programmed cell death (apoptosis)

- only few 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 95 %

Susumu Ohno, 1972



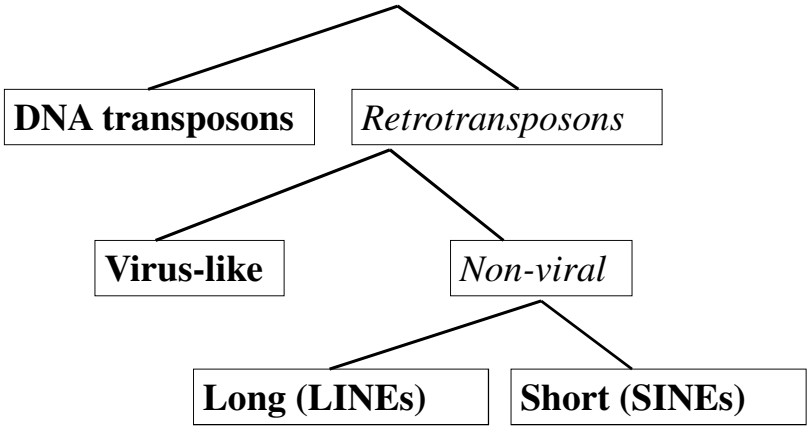
Susumu Ohno
Feb. 1, 1928 - Jan 13, 2000

- Because of mutation load the human haploid genome cannot afford to keep more than about 30,000 gene loci.
- Most of DNA is redundant … junk!

http://www.junkdna.com/ohno.html

Mobile DNA elements (transposons)

Autonomous DNA sequences, capable to copy themselves, represent 44 % of genome



DNA transposons

2-3 kb (or shorter), encode transposase, cut & paste in genome without RNA intermediate

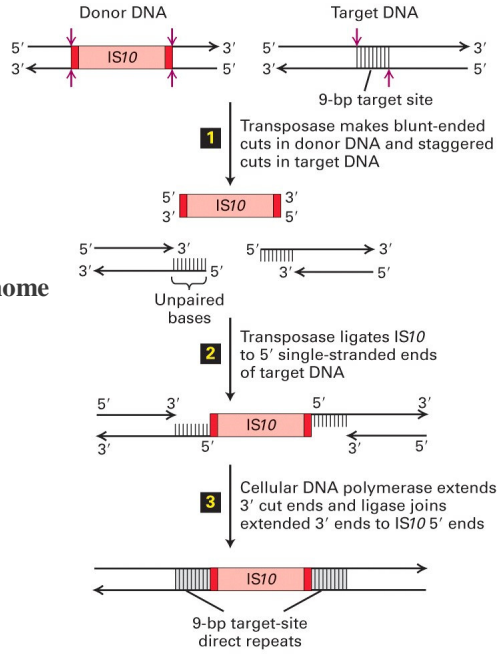


Fig: Lodish, H. et al.: Molecular Cell Biology (5th ed.), W.H.Freeman, New York 2004.

Mobile elements (transposons):

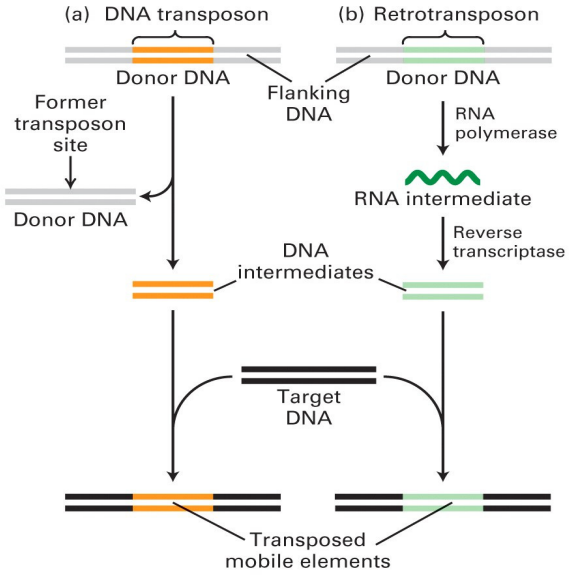


Fig.: Lodish, H. et al.: Molecular Cell Biology (5th ed.), W.H.Freeman, New York 2004.

Mobile (parasitic) elements in mammalian genome:

- **DNA transposons**
  - **2-3 kb (or shorter), encode transposase, cut & paste or copy & paste in genome without RNA intermediate**
- **Virus-like retrotransposons**
  - **6-11 kb (or shorter), retroviruses without gene for protein envelope (env)**
- **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 noncoding cellular 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**

- **Mostly mutated and/or incomplete copies, only small part (<0,05%) still active:**
  - **LINEs: 80-100 L1**
  - **SINEs: 2000-3000 Alu, <100 SVA**
  - **Retrovirus-like: ?** *(HERV-K…really extinct?)*
  - **DNA transposons: 0**
- **Mouse genome contains much more functional transposons** *(...why?)*

Significance of transposons in human genome

- Transposition in germinal cells is a rare event (approx. 1 new insertion per 20 live births, mostly Alu)
- Still a significant source of human genetic variability
- Can inactivate genes – documented as a rare cause of inherited diseases
- In somatic cells can result in mosaicism
  - role of L1 in neurogenesis?

- Transposons facilitate recombination …driving force of evolution!

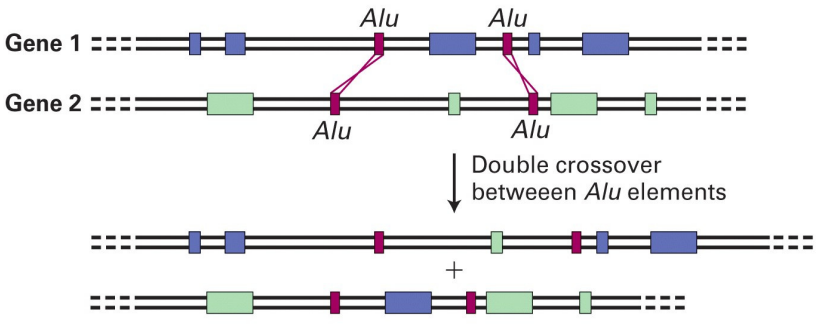


Fig.: Lodish, H. et al.: Molecular Cell Biology (5th ed.), W.H.Freeman, New York 2004.

Non-classified "spacer" DNA:
non-repetitive, noncoding, >1/2 genome … likely also dead transposons, too mutated to be recognizable

Project ENCODE, 2012: no junk DNA!

- Up to 80% of genome has biological function
- Up to 75% of genome is at least some time and somewhere transcribed to RNA
- Despite the fact that only 20% of genome at best is under evolutionary constraint

….?????.....

Human Genome Project: Impact

Benefits of genome sequencing

- Facilitates research into molecular basis of diseases
- Study of human evolution and migration
- What the genome determines ("nature vs. nurture") and how genetic variation causes differences among people
- Genomic medicine, pharmacogenomics, personalized medicine….

Genomic medicine

- **1) Diagnostics at the gene level**
  - **Rare monogenic diseases**
  - **Shift to earlier diagnostics**
    - **Possibility of diagnosis before disease appears**
    - **Newborn screening**
    - **Noninvasive prenatal testing**
    - **Preconception carrier testing, preimplantation genetic analysis in IVF**
  - **Genomic-based analysis of tumors enables effective targeted therapies**
  - **In common complex diseases with polygenic predispositions (diabetes, coronary disease etc.) still difficult**

Personal Genomics: 23andME

- Saliva sample sent by DHL, genotyping cca 700 000 SNPs
- DNA relatives
- Ancestry:
  - Ancestry Composition
  - Paternal (Y chromosome haplogroup)
  - Maternal (mitochondrial DNA haplogroup)
  - Per cent Neanderthal DNA
- Health

Personal Genomics: 23andME

- Saliva sample sent by DHL, genotyping cca 700 000 SNPs
- DNA relatives
- Ancestry
- Health:
  - Disease risk: 123 (31 high confidence)
  - Drug response: 25 (12 high confidence)
    Inherited conditions: 53 (all high confidence)
  - Traits: 63 (15 high confidence)

Genome-Wide Association Studies (GWAS)

- Phenotype (trait or disease) + genotyping of SNPs
- Association analysed by statistical tests
- High power needed to achieve significance (over 10,000 participants)
- Common variants: numerous, individual effects small, but together most of heritability
- Rare variants: uncommon, often de novo, have big effect if present

Why analysis of SNPs does not say more?

- Common SNPs not sufficient – necessary to find individual (rare) polymorphisms
- SNPs are not the main source of human genetic variability – duplications/deletions and insertions of transposons more significant
- Trait controlled by a single gene is probably rather uncommon condition – phenotype is result of interplay of numerous genes
- Expression of genes (how genome is used) is what decides
  - Polymorphisms in noncoding regulatory DNA
  - Epigenetics (DNA methylation etc.) – also heritable!

Genomic medicine

- 2) Pharmacogenomics
  - Targeted therapy of tumors directed by genetic analysis
    - E.g.: antibody against HER-2 only in breast tumors that express this protein
  - Genomic-based tests predict drug efficacy, occurrence of adverse side effects, or help to optimize dosage.
    - E.g.: treatment of chronic hepatitis C, HIV, possibly dosage of warfarin

… personalized medicine

Genomic medicine

- 2) Farmacogenomics – example:

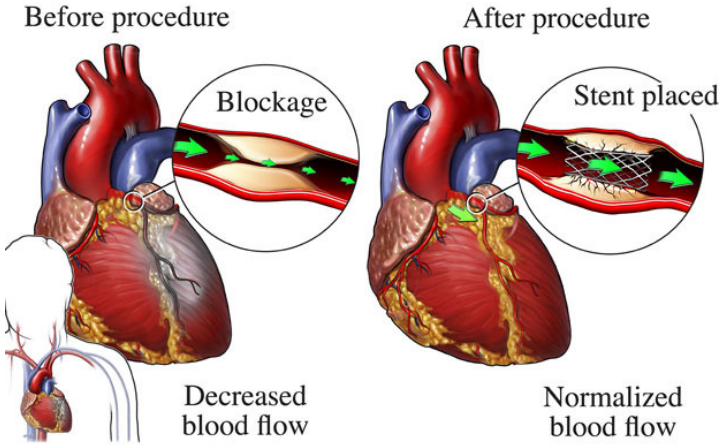


Fig.: https://www.telegraph.co.uk/news/uknews/theroyalfamily/8977422/Coronary-stenting-how-does-it-work.html

Genomic medicine

- **2) Farmacogenomics – example:**
  - **Clopidogrel:**
    - Anticoagulans (blocks ADP receptor on thrombocytes)
    - Pro-drug: Requires metabolic activation by microsomal hydroxylases (cytochrome P450 2C19)
    - Up to 30% patients have genetically decreased or absent level of the activating enzyme 2C19
  - **Study in patients with coronary stents:**
    - Choice of anticoagulant therapy guided by genotypization of CYP2C19: clopidogrel or ticagrelol/prasugrel
    - Control group: ticagrelol/prasugrel
    - Result: In genotype-guided group the therapy equally effective, but lower incidence of bleeding

Claasens et al. N. Engl. J. Med. 2019, 381:1621

Genomic medicine

- **3) Microorganisms:**
  - **Pathogenic:**
    - **Rapid diagnostics of infectious disease by pathogen sequencing – especially relevant in tracing new epidemic outbreaks (SARS, MRSA…)**
  - **Nonpathogenic: Human Microbiome**
    - **E.g. human gut bacteria – metabolic activity comparable to liver, individually different spectrum, relationships to inflammatory bowel disease, atherosclerosis, obesity…**

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**
- **Designer babies**
- **Behavioral genetics:**
  - how genes determine human behaviour, possible fall into genetic determinism and loss of responsibility for one's own behaviour
- **GMO**
- **Gene patenting**

References:

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

Lodish, H. et al.: Molecular Cell Biology, W.H.Freeman, New York 1995, 2004 ("Darnell").

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

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

Trends in Genetics 2007: 23, pp.183-191.

Nature 2009: 458, 719-724.

FEBS Letters 2011: 585; pp. 1589-1594.

Science Translational Medicine 2013: 5, 189sr4.

PNAS 2014: 111, pp. 6131-6138

N. Engl. J. M. 2019: 381, pp. 1621-1631.

Lecture by Eric Lander, 1.LF UK, 18.2.2020.

www.ncbi.nlm.nih.gov

genomics.energy.gov

en.wikipedia.org

www.ensembl.org

www.illumina.com

www.earthbiogenome.org

www.23andme.com

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