Ch. 1

Genomes – Epigenomes – Nuclear organization

L1.1

Genomes

the Human Genome

DNA sequence

Human Genome Project (1991-2003)

Next Generation Sequencing

Databases

Composition

Human Variation

Background Help added in the last section of the Moodle Course site

The Human Genome Project

Animated tutorials on the Human Genome Project:

http://www.genome.gov/Pages/EducationKit/

(free downloads or on-line view)

HGP (see book, moodle site) 1990-2003

?

1991-1998 - Physical mapping period (**EST, SST, known genes**)

1998-2003 - Cloning, sequencing (Sanger) and assembly

The principle: «hierarchical cloning»

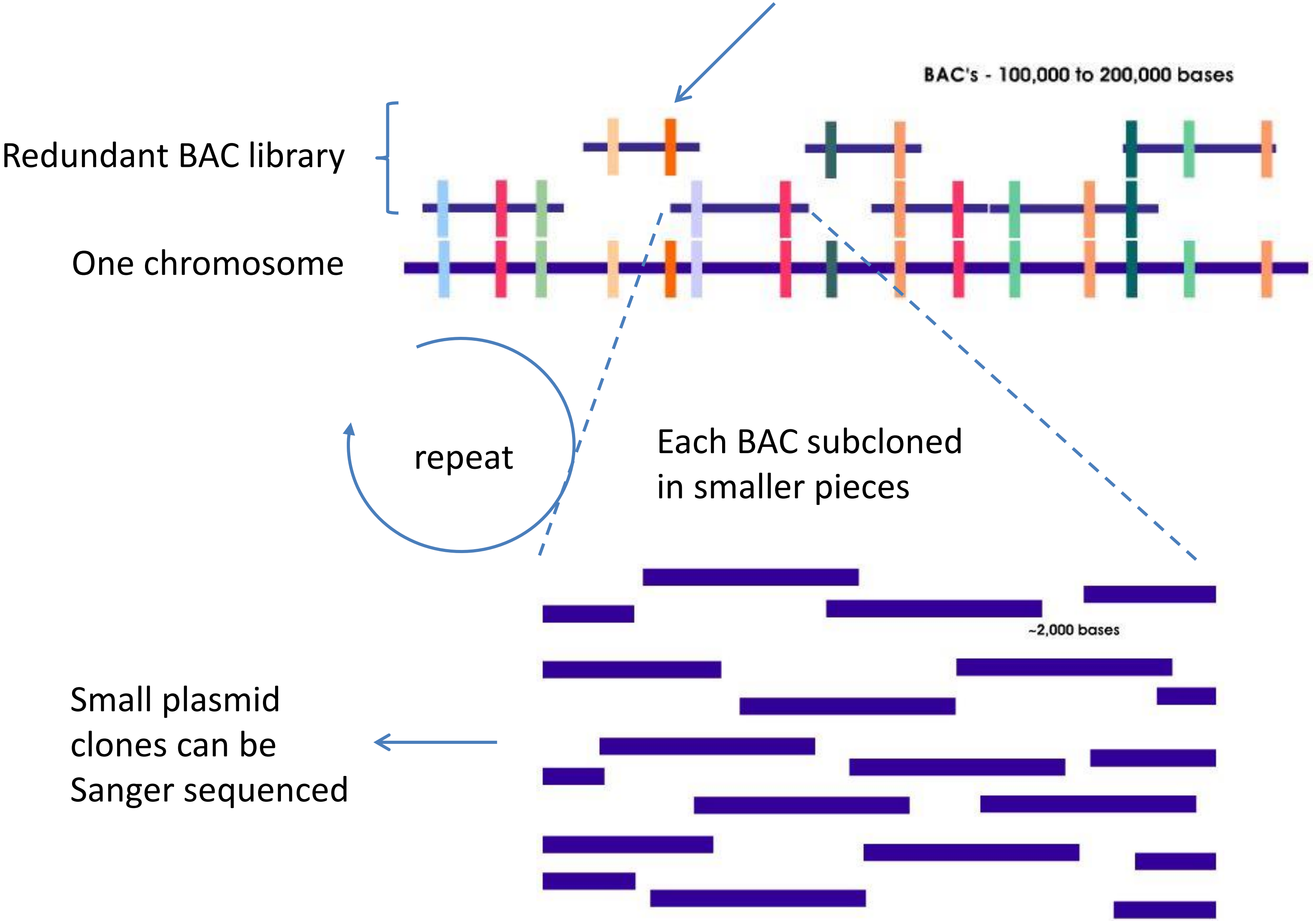
Stocastic: the process required super-extensive and highly redundant cloning

Hierarchical cloning vectors:

- BACs, PACs – 100-200 Kb
- Cosmids and other phage-derived vectors (20-40Kb)
- Plasmids – 2-3 Kb

HGP hierarchical strategy

These are «landmarks»
derived from physical mapping



Redundant BAC library

One chromosome

repeat

Each BAC subcloned
in smaller pieces

Small plasmid
clones can be
Sanger sequenced

Sanger di-deoxy-nucleotide terminator method

- Requires isolated DNA fragments (cloned)
- Requires known primer sequence
- Intrinsically limited to 5-700 bp

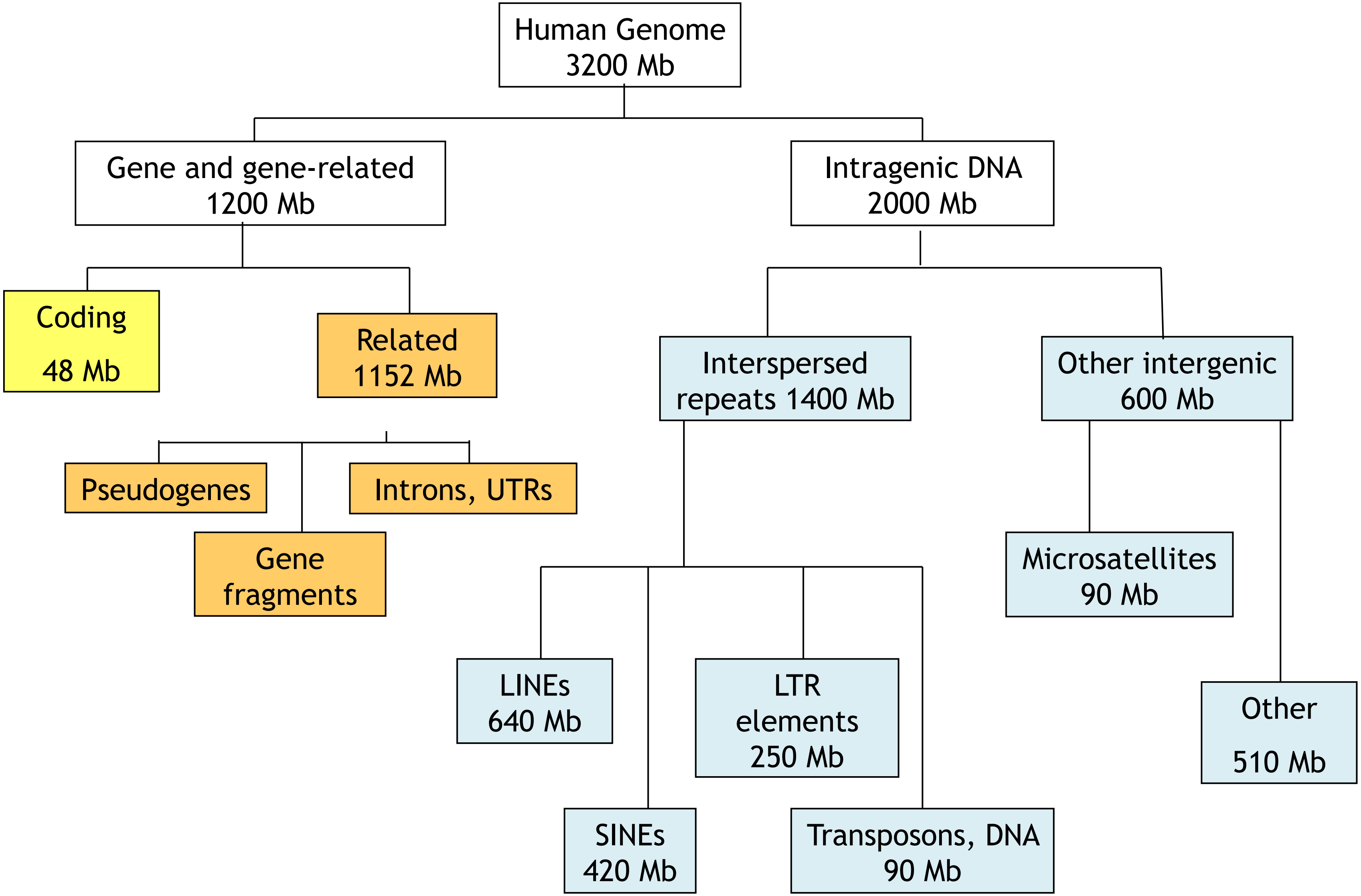
Sanger sequencing

Composition of the Human Genome

Sequence identity was progressively **annotated** in the Human Genome by extensive bioinformatic analysis

- Sequence similarity (homology)
- Correspondence to known RNA / proteins
- Repeated sequence comparison with known genetic elements
- Knowledge on genomes of different organisms

Genome composition - H. Sapiens (the 2003 version).



Human Genome
3200 Mb

Gene and gene-related
1200 Mb

Intragenic DNA
2000 Mb

Coding
48 Mb

Related
1152 Mb

Pseudogenes

Introns, UTRs

Gene
fragments

Interspersed
repeats 1400 Mb

Other intergenic
600 Mb

Microsatellites
90 Mb

LINEs
640 Mb

LTR
elements
250 Mb

SINEs
420 Mb

Transposons, DNA
90 Mb

Other
510 Mb

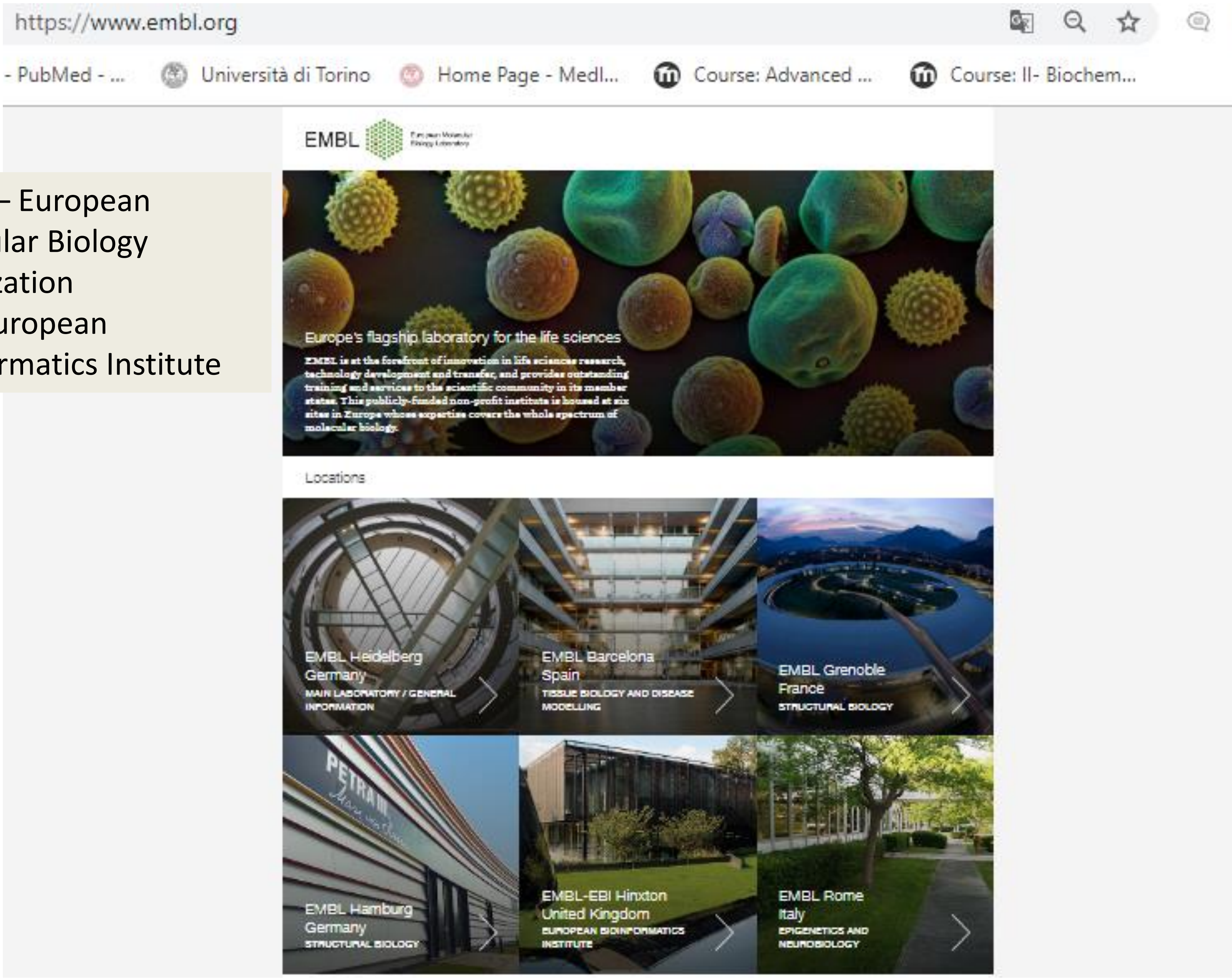
Sequences of the Reference Human Genome have been since continuously adjourned, revised, completed, maintained

Different versions released timely by the HGP Consortium

Sequences and annotations are conserved and available from biological **Databases**

Several organizations to maintain and run public databases. They are paid by Agencies and other public organizations

EMBO – European Molecular Biology Organization
EBI – European Bioinformatics Institute



Europe's flagship laboratory for the life sciences

EMBL is at the forefront of innovation in life sciences research, technology development and transfer, and provides outstanding training and services to the scientific community in its member states. This publicly-funded non-profit institute is housed at six sites in Europe whose expertise covers the whole spectrum of molecular biology.

Locations

EMBL Heidelberg
Germany
MAIN LABORATORY / GENERAL INFORMATION

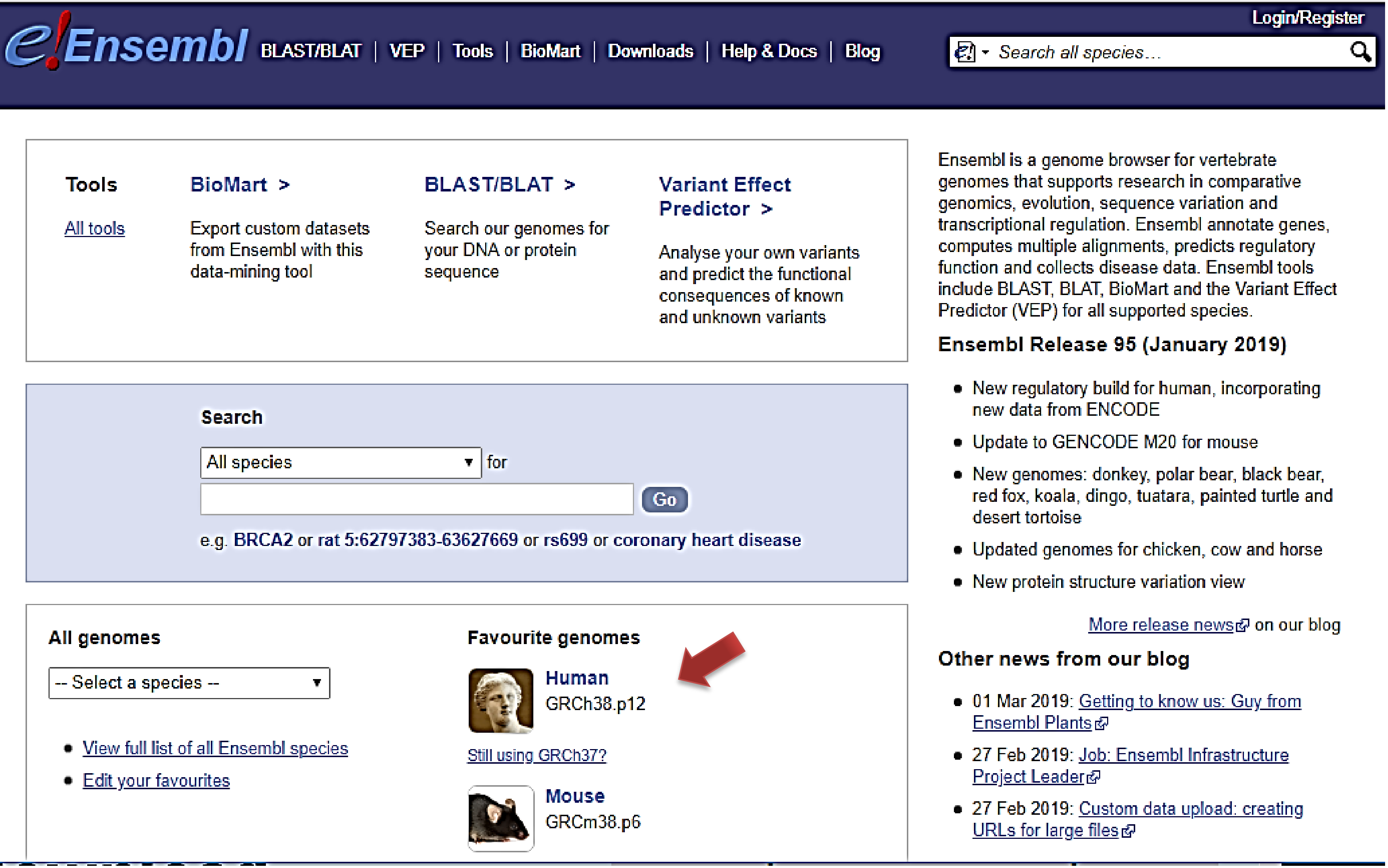
EMBL Barcelona
Spain
TISSUE BIOLOGY AND DISEASE MODELLING

EMBL Grenoble
France
STRUCTURAL BIOLOGY

EMBL Hamburg
Germany
STRUCTURAL BIOLOGY

EMBL-EBI Hinxton
United Kingdom
EUROPEAN BIOINFORMATICS INSTITUTE

EMBL Rome
Italy
EPIGENETICS AND NEUROBIOLOGY


Ensembl
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
Login/Register
Search all species…
Go

Tools

All tools

BioMart >

Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >

Search our genomes for your DNA or protein sequence

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

Search

All species ▾ for

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

All genomes

-- Select a species -- ▾

- View full list of all Ensembl species
- Edit your favourites

Favourite genomes

Human
GRCh38.p12

Still using GRCh37?

Mouse
GRCm38.p6

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 95 (January 2019)

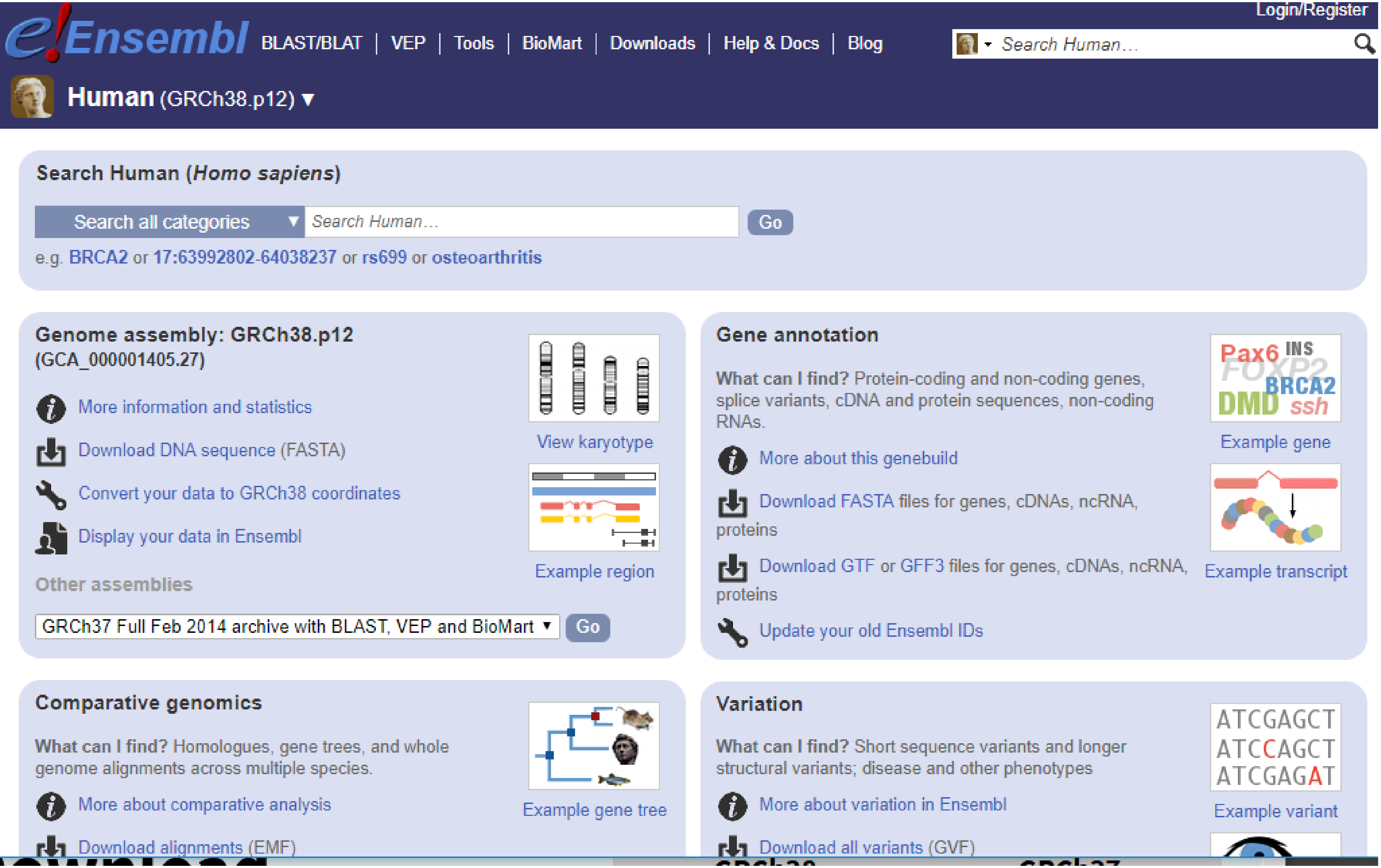
- New regulatory build for human, incorporating new data from ENCODE
- Update to GENCODE M20 for mouse
- New genomes: donkey, polar bear, black bear, red fox, koala, dingo, tuatara, painted turtle and desert tortoise
- Updated genomes for chicken, cow and horse
- New protein structure variation view

More release news on our blog

Other news from our blog

- 01 Mar 2019: Getting to know us: Guy from Ensembl Plants
- 27 Feb 2019: Job: Ensembl Infrastructure Project Leader
- 27 Feb 2019: Custom data upload: creating URLs for large files

http://www.ensembl.org/index.html


e!Ensembl
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
Login/Register
Search Human…
Human (GRCh38.p12) ▾

Search Human (*Homo sapiens*)

Search all categories ▾ | Search Human… | Go

e.g. BRCA2 or 17:63992802-64038237 or rs699 or osteoarthritis

Genome assembly: GRCh38.p12 (GCA_000001405.27)

- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to GRCh38 coordinates
- Display your data in Ensembl

View karyotype

Example region

Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart ▾ Go

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 FASTA files for genes, cDNAs, ncRNA, proteins
- Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins
- Update your old Ensembl IDs

Example gene

Example transcript

Comparative genomics

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

- More about comparative analysis
- Download alignments (EMF)

Example gene tree

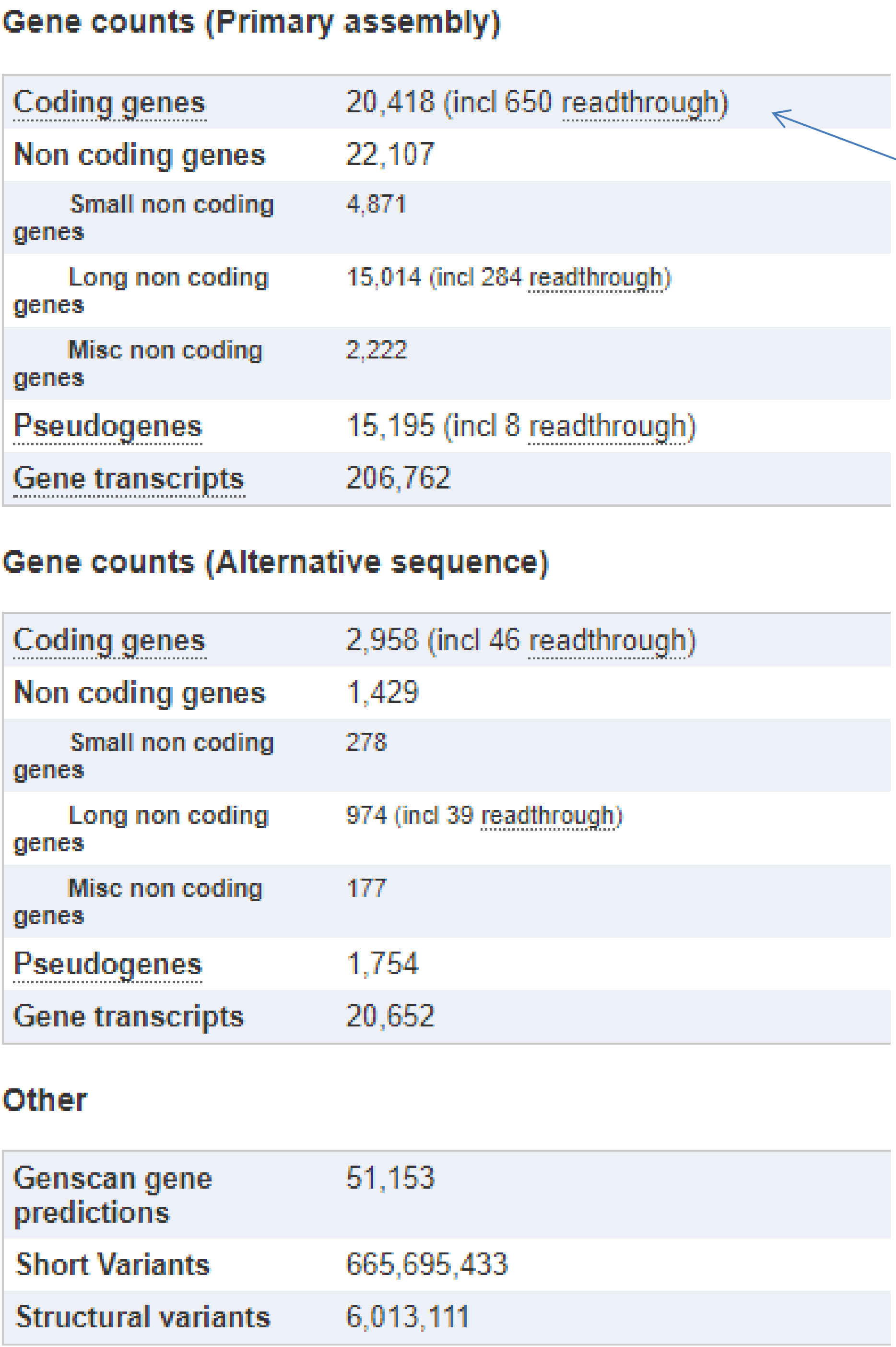
Variation

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

- More about variation in Ensembl
- Download all variants (GVF)

Example variant

https://www.ensembl.org/Homo_sapiens/Info/Index



Gene counts (Primary assembly)

Coding genes	20,418 (incl 650 readthrough)
Non coding genes	22,107
Small non coding genes	4,871
Long non coding genes	15,014 (incl 284 readthrough)
Misc non coding genes	2,222
Pseudogenes	15,195 (incl 8 readthrough)
Gene transcripts	206,762

Gene counts (Alternative sequence)

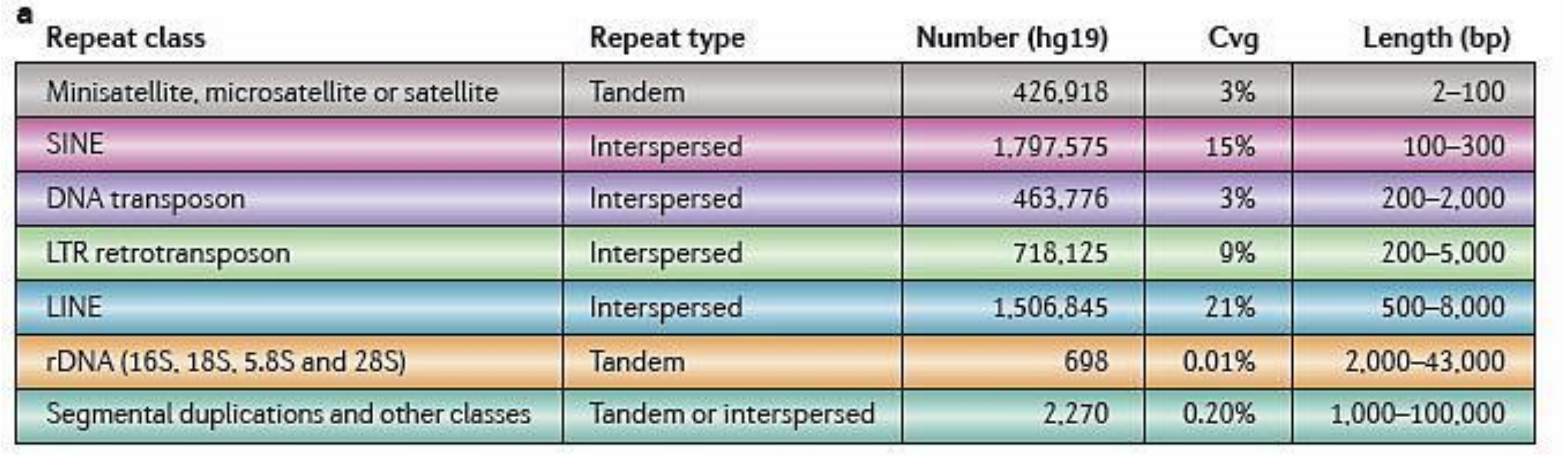
Coding genes	2,958 (incl 46 readthrough)
Non coding genes	1,429
Small non coding genes	278
Long non coding genes	974 (incl 39 readthrough)
Misc non coding genes	177
Pseudogenes	1,754
Gene transcripts	20,652

Other

Genscan gene predictions	51,153
Short Variants	665,695,433
Structural variants	6,013,111

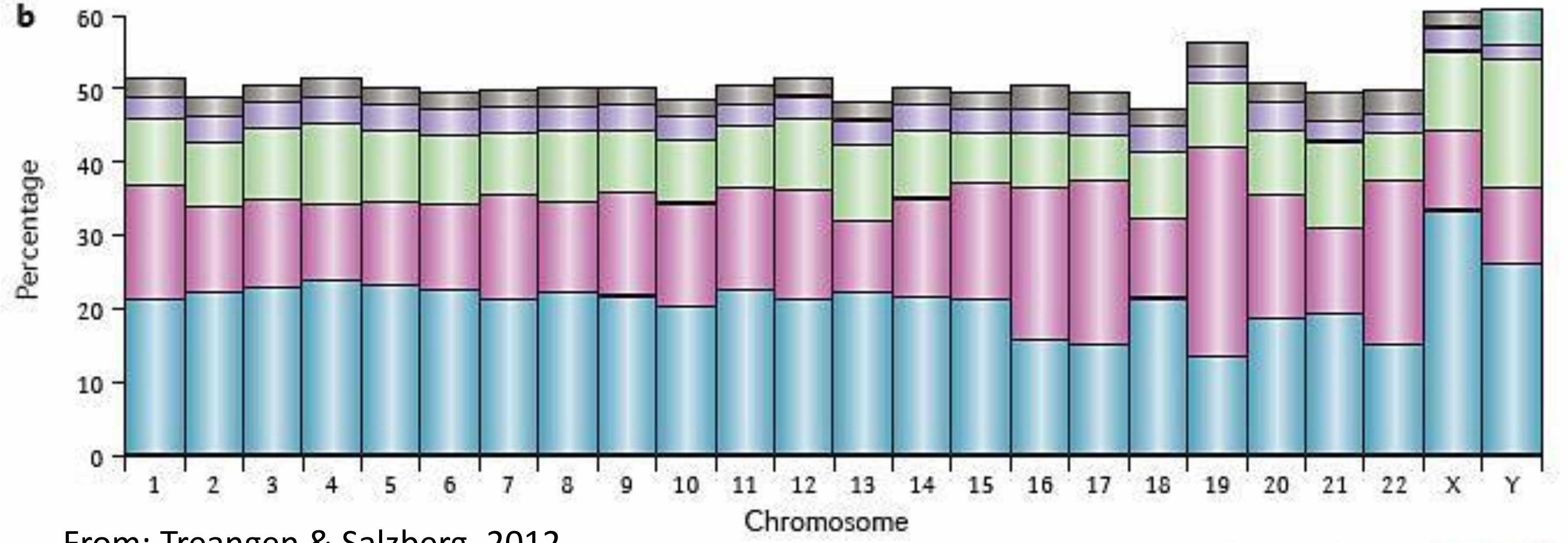
What is «readthrough» ?

Repetitive sequences cover nearly half of the Human Genome



a

Repeat class	Repeat type	Number (hg19)	Cvg	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	426,918	3%	2–100
SINE	Interspersed	1,797,575	15%	100–300
DNA transposon	Interspersed	463,776	3%	200–2,000
LTR retrotransposon	Interspersed	718,125	9%	200–5,000
LINE	Interspersed	1,506,845	21%	500–8,000
rDNA (16S, 18S, 5.8S and 28S)	Tandem	698	0.01%	2,000–43,000
Segmental duplications and other classes	Tandem or interspersed	2,270	0.20%	1,000–100,000



From: Treangen & Salzberg, 2012

Nature Reviews | Genetics

Interspersed repetitive elements - Mobile genetic elements

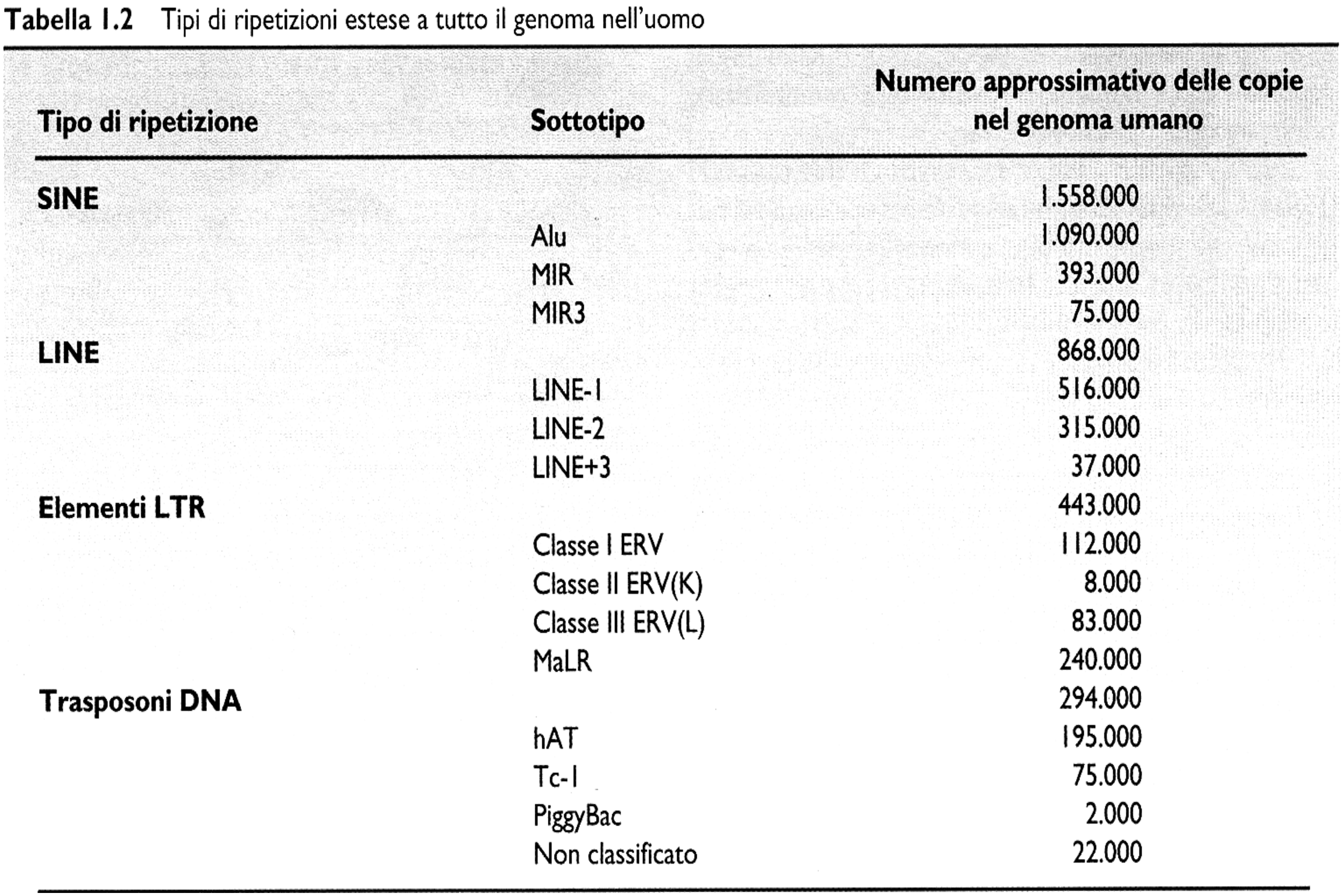
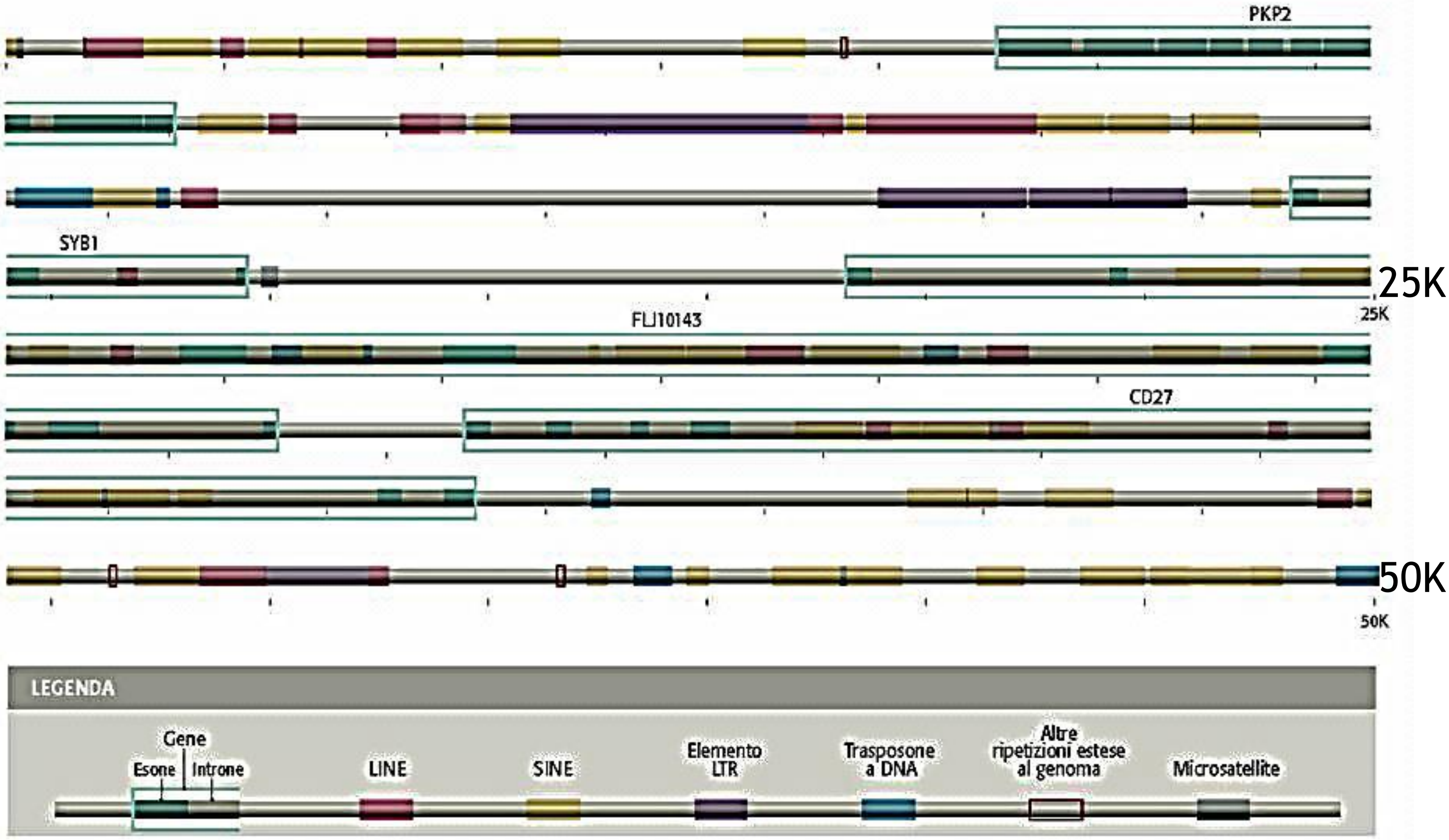


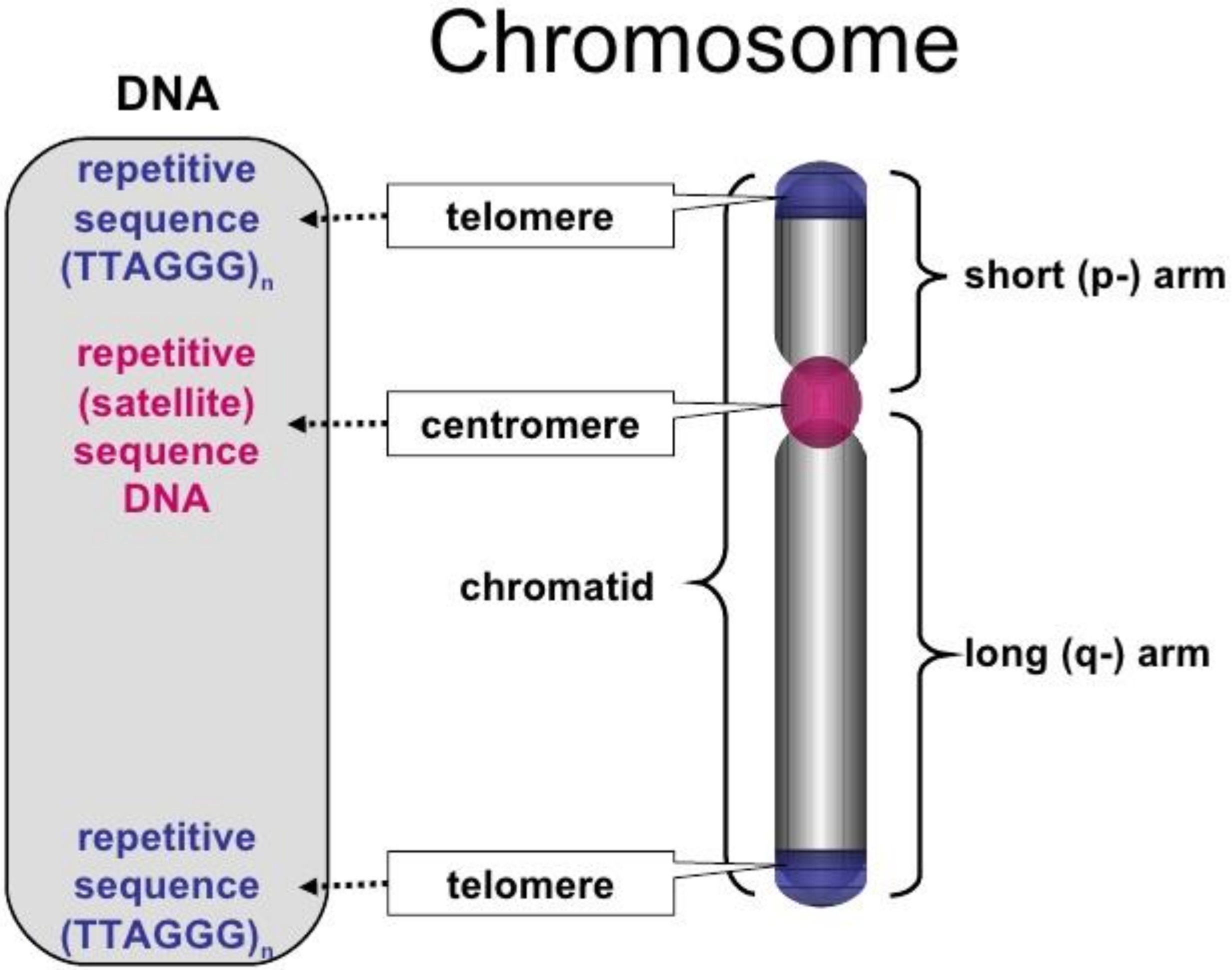
Tabella 1.2 Tipi di ripetizioni estese a tutto il genoma nell'uomo

Tipo di ripetizione	Sottotipo	Numero approssimativo delle copie nel genoma umano
SINE		1.558.000
	Alu	1.090.000
	MIR	393.000
	MIR3	75.000
LINE		868.000
	LINE-1	516.000
	LINE-2	315.000
	LINE+3	37.000
Elementi LTR		443.000
	Classe I ERV	112.000
	Classe II ERV(K)	8.000
	Classe III ERV(L)	83.000
	MaLR	240.000
Trasposoni DNA		294.000
	hAT	195.000
	Tc-1	75.000
	PiggyBac	2.000
	Non classificato	22.000

A 50 Kb tract of the Human genome showing gene position, repeats, microsatellites (taken from Chr. 12)



tandem repeats


Chromosome
DNA
repetitive sequence (TTAGGG)n
repetitive (satellite) sequence DNA
repetitive sequence (TTAGGG)n
telomere
centromere
chromatid
telomere
short (p-) arm
long (q-) arm

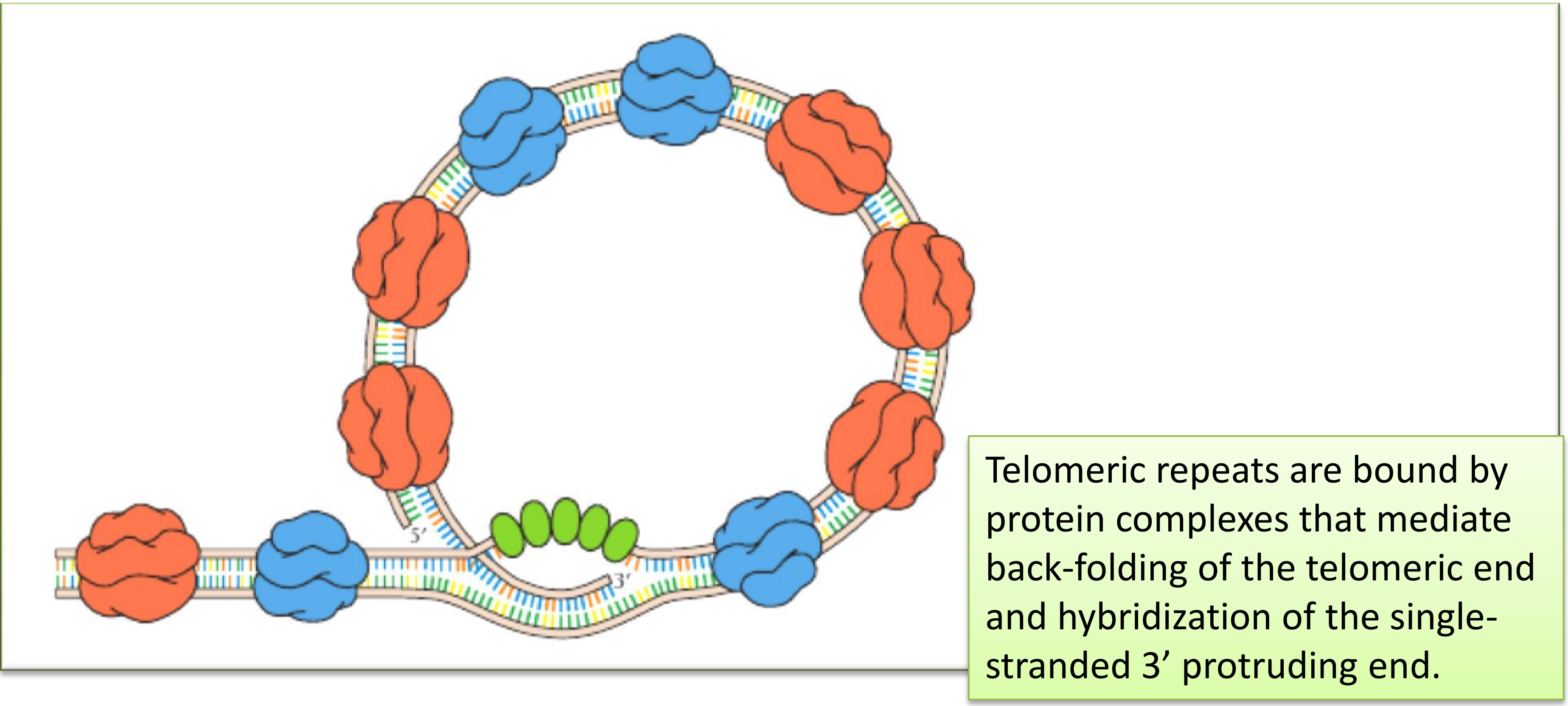
microsatellite

Telomers and telomerase

Short tandem sequence repeats at telomeres.

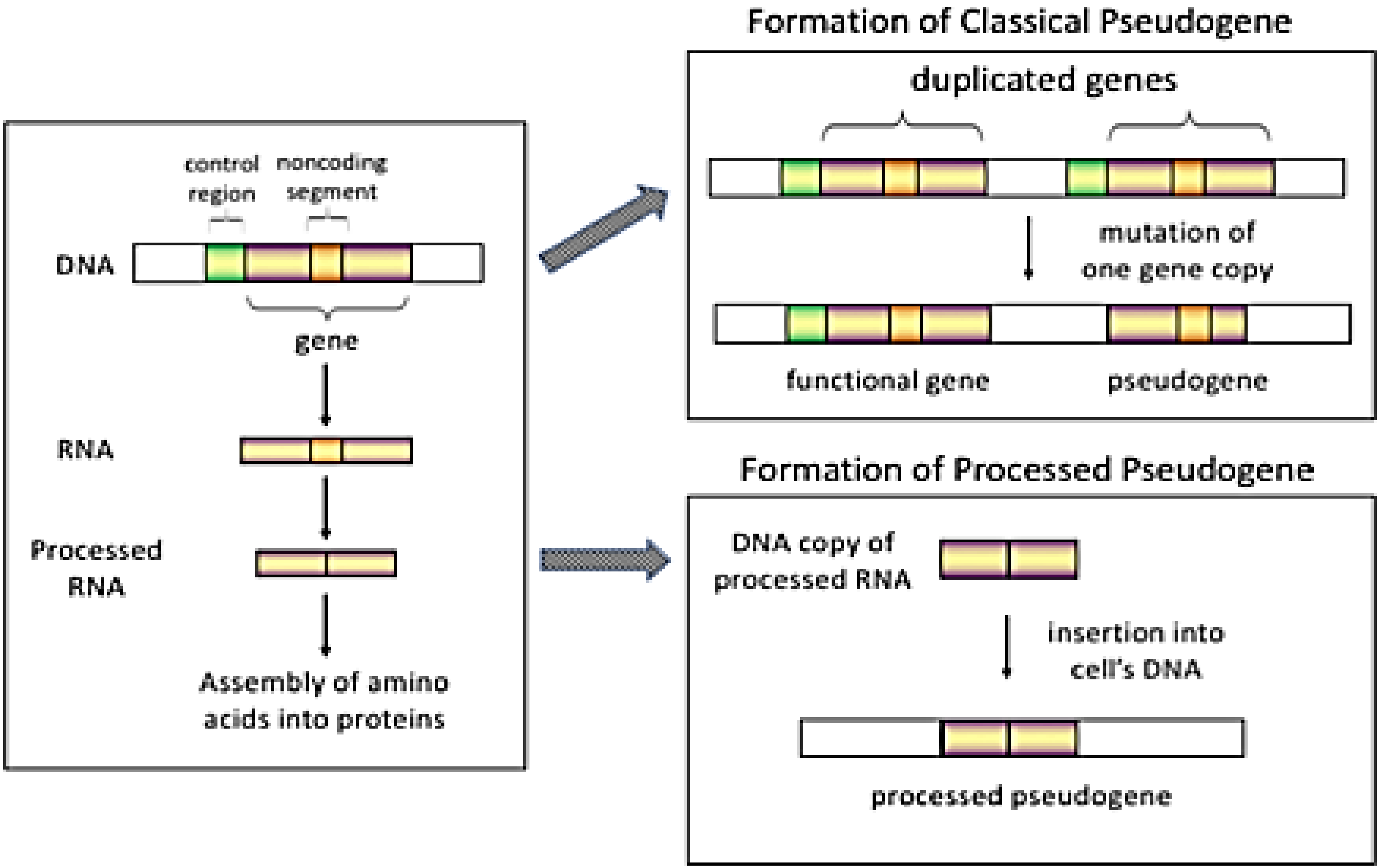
In H. sapiens: TTAGGG (2,500 repeats)

Repeat sequence differs in different organisms



Telomeric repeats are bound by protein complexes that mediate back-folding of the telomeric end and hybridization of the single-stranded 3′ protruding end.

pseudogenes



control region

noncoding segment

DNA

gene

RNA

Processed RNA

Assembly of amino acids into proteins

Formation of Classical Pseudogene

duplicated genes

mutation of one gene copy

functional gene

pseudogene

Formation of Processed Pseudogene

DNA copy of processed RNA

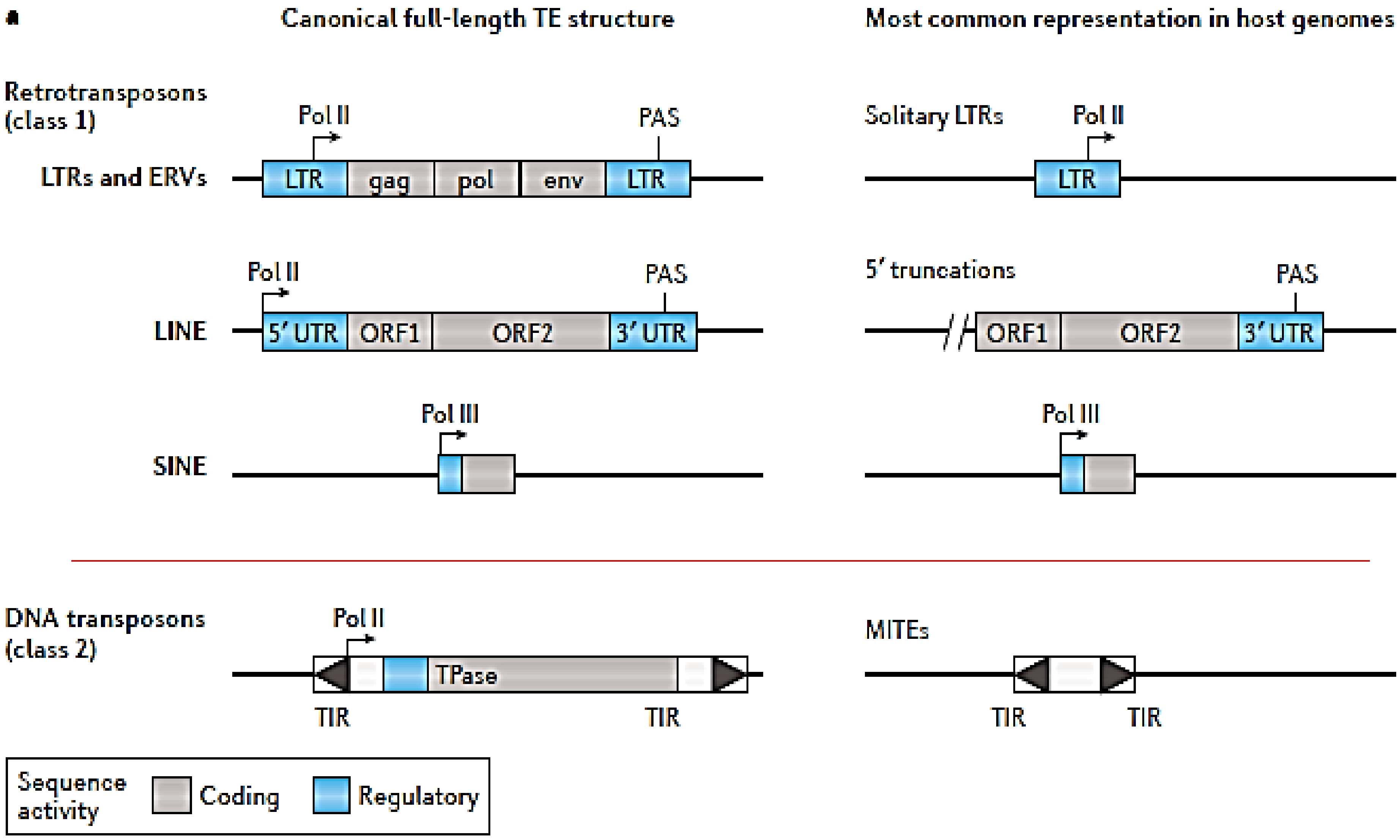
insertion into cell's DNA

processed pseudogene

Transposable Elements (TE)

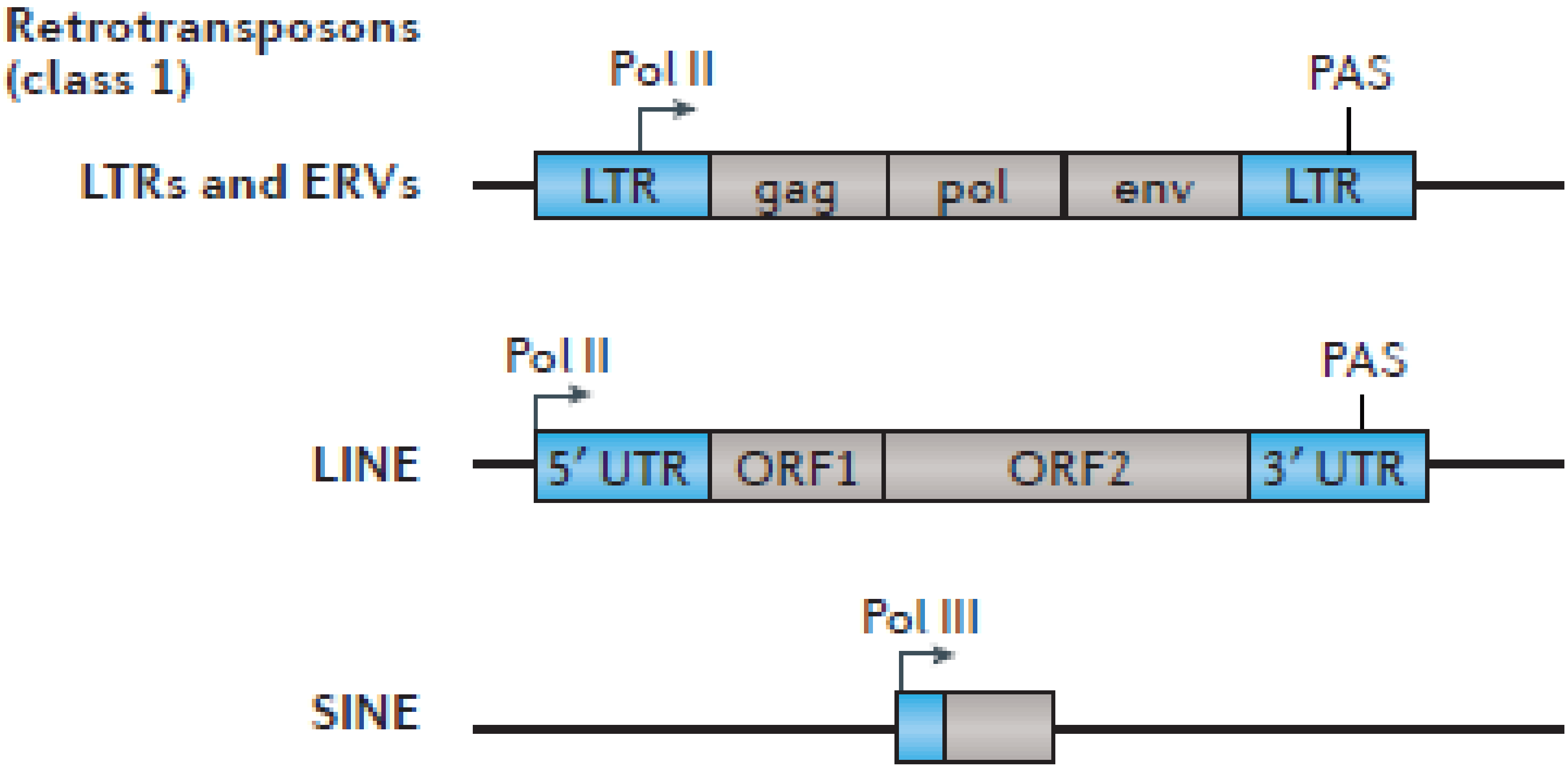
«mobile» elements deriving from either retrovirus infection or retrocopies of cell genes or even autonomousy replicating DNA tracts that can move to different genomic positions from the original.

DNA transposons
Retrotransposons
LTR
LINEs
SINEs



TPase = Transposase

from Chuong et al., 2017, Nat Rev Genet



Pol II & Pol III

Conclusions

In **2003**, only a thiny fraction of the Human Genome sequence could be attributed with a function.

Most of the sequence was thought to be redundant, repetitive and essentially «junk» DNA.

This conclusion, though, was adversed by scientists that studied the phylogenetic conservation, showing that many regions with no apparent function are indeed extremely conserved between organisms (the «dark matter» theory).

For this reason, scientists started several projects to systhematically analyze every regions of the Human (and mouse) genomes to unravel any possible functional role.

Comparative

Many other genomes sequenced completely or partially

Most of sequencing projects are publicly funded, results are open in databases

Many other are run by private funding and results are not open. They include many vegetables, bacteria, fungi.

Public **databases** :

ENSEMBL species

NCBI Genomes Genomic Data

NCBI is National Center for Biotechnological Information

is based in the National Library of Medicine at NIH
(National Institutes of Health)

USA – It is a public domain (still.... Trump permitting)

The National Institutes of Health

https://www.nih.gov/

The National Library of Medicine

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

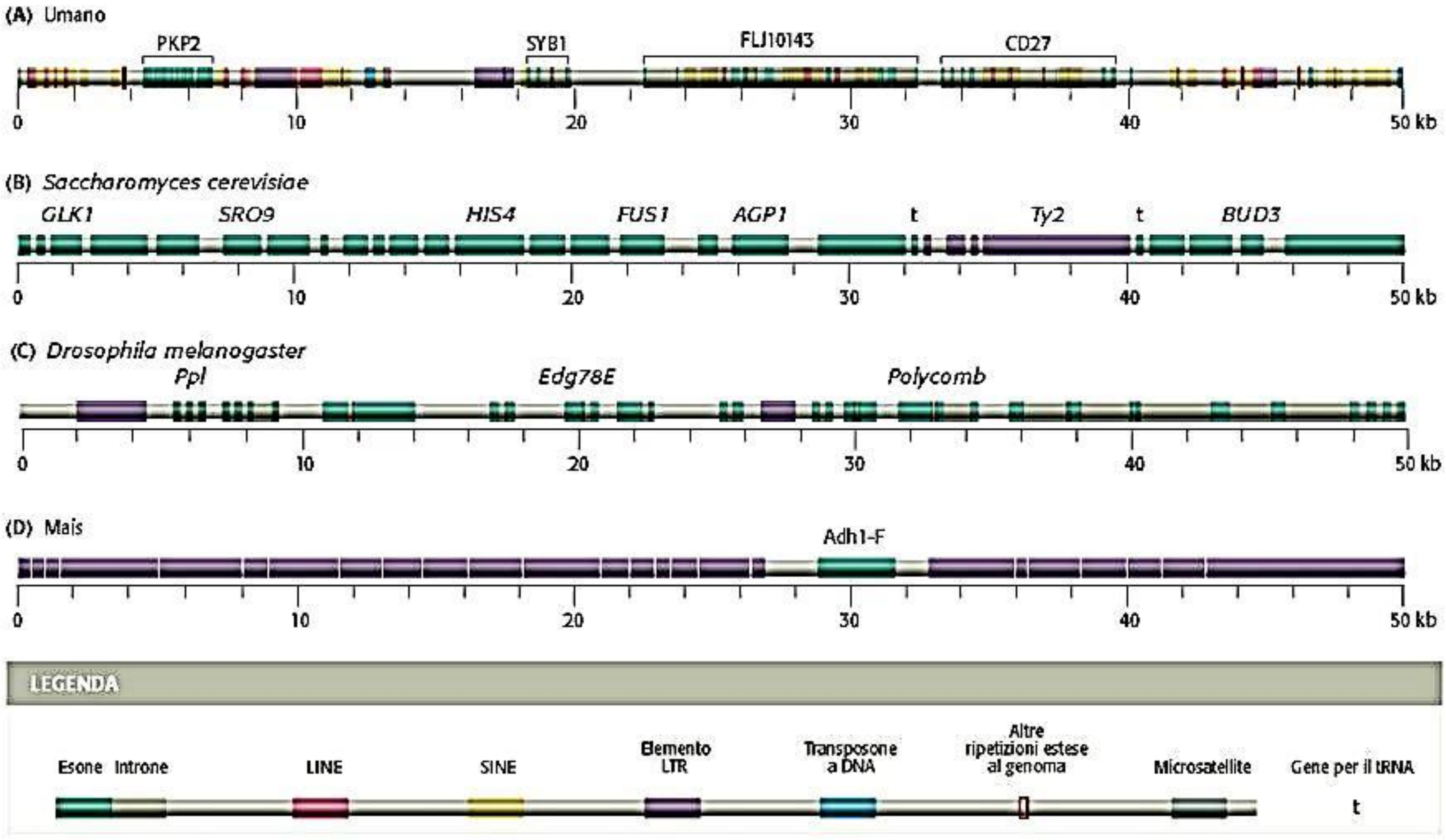
The National Center for Biotechnological Information

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

T.A. Brown
Genomi, III Ed.
EdiSES

Comparative:
- Human
- Yeast
- Drosophila
- Mais

Figura 7.15 Confronto tra genoma umano, di lievito, del moscerino della frutta e di mais. (A) Il segmento di 50 kb del cromosoma 12 umano mostrato precedentemente, è confrontato con segmenti di 50 kb derivanti da genomi di (B) *S. cerevisiae*; (C) *Drosophila melanogaster*; (D) mais.

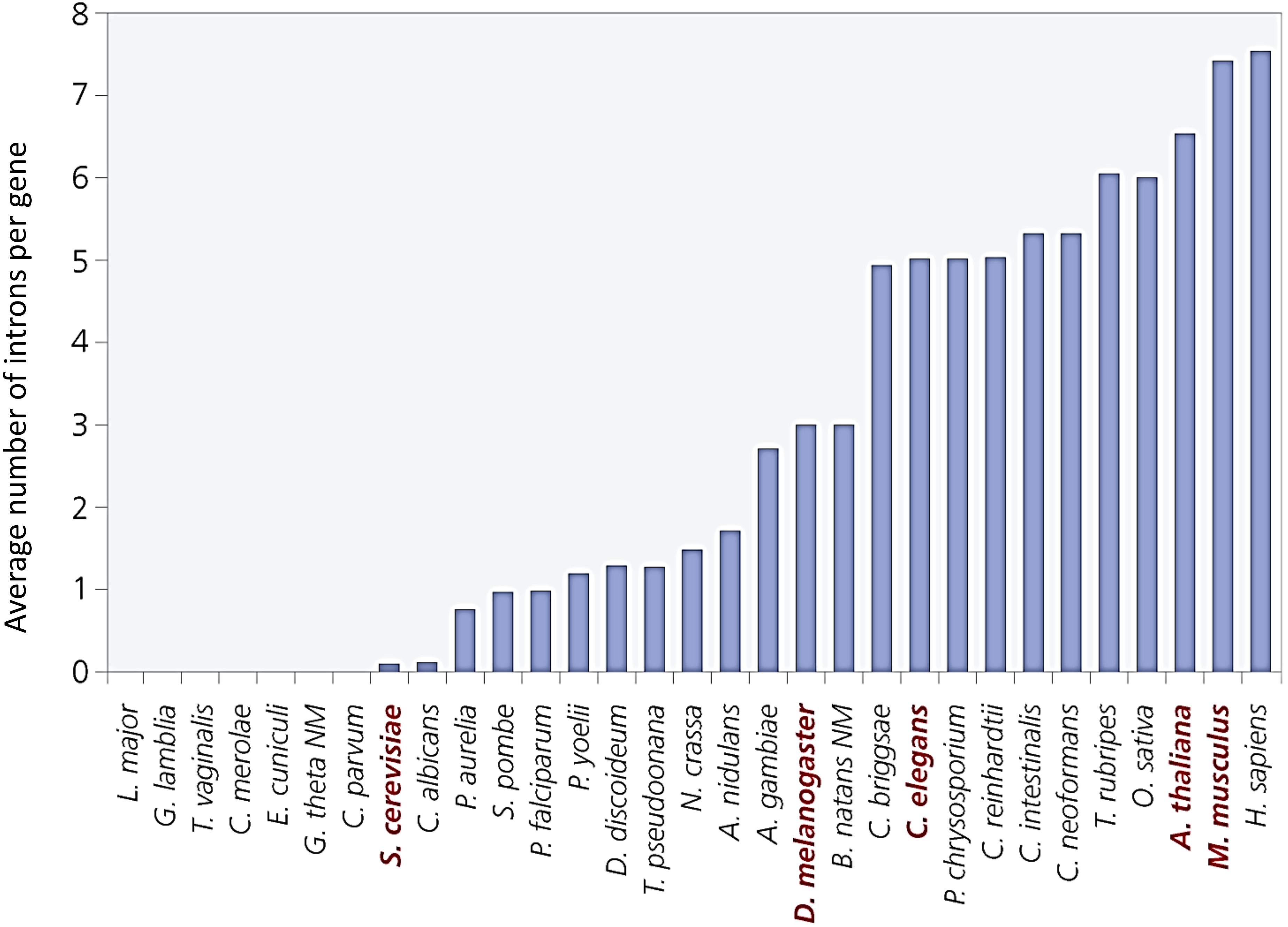


Gene structure

Exon-Intron structure is present in all Eukaryotes

Hower the average number of introns, as well as the lenght of introns and central exons, varies considerably

Are introns an evolutionary feature ?

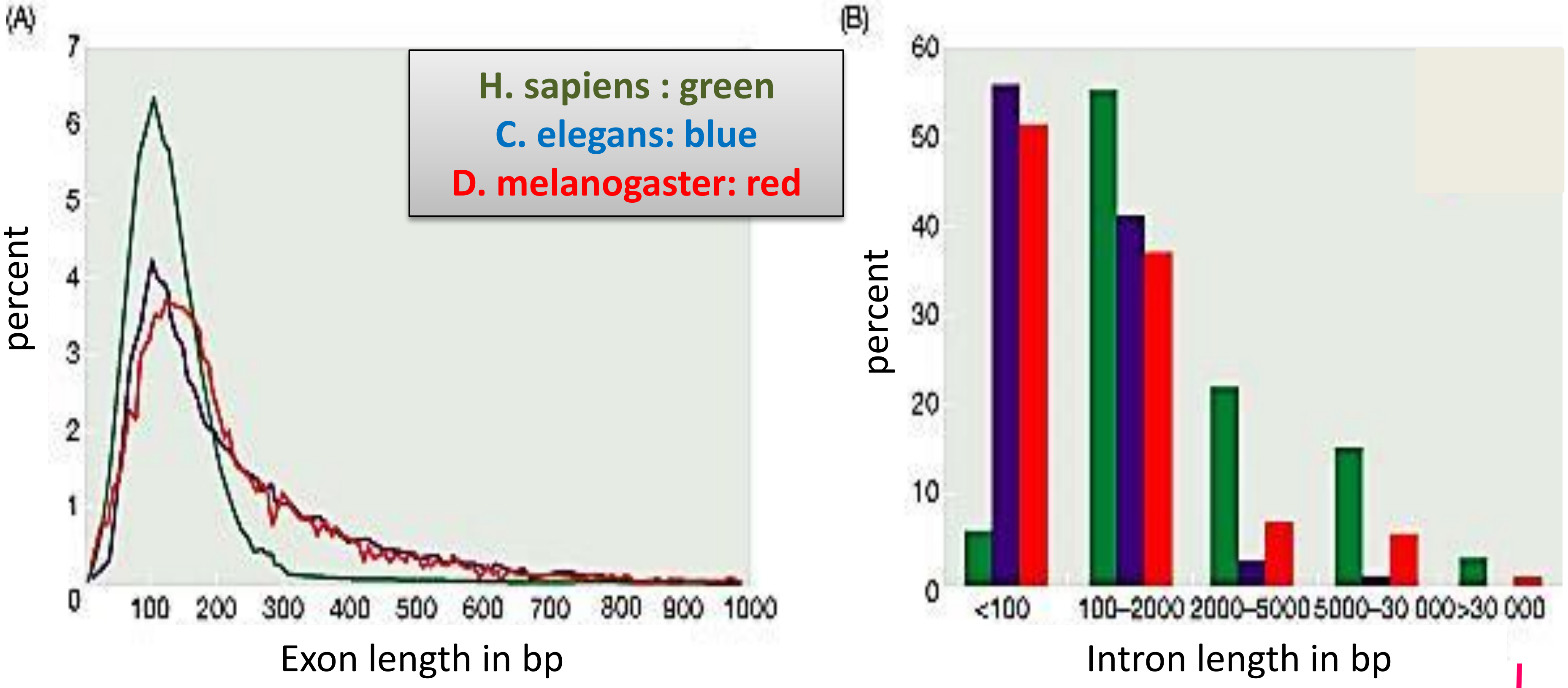

Average number of introns per gene
0
1
2
3
4
5
6
7
8
L. major
G. lamblia
T. vaginalis
C. merolae
E. cuniculi
G. theta NM
C. parvum
S. cerevisiae
C. albicans
P. aurelia
S. pombe
P. falciparum
P. yoelii
D. discoideum
T. pseudoonana
N. crassa
A. nidulans
A. gambiae
D. melanogaster
B. natans NM
C. briggsae
C. elegans
P. chrysosporium
C. reinhardtii
C. intestinalis
C. neoformans
T. rubripes
O. sativa
A. thaliana
M. musculus
H. sapiens

Averages in Human Genome: protein coding genes

Number of exons	8.8
Exon length	170 bp (quite narrow range, 85%<200bp)
Intron length	5420 bp (large range 20bp to 100Kb)

Range:
Intron =0 (3350 single-exon genes)
Max number of Introns = 147 (NEB gene).

How exons and introns changed during evolution

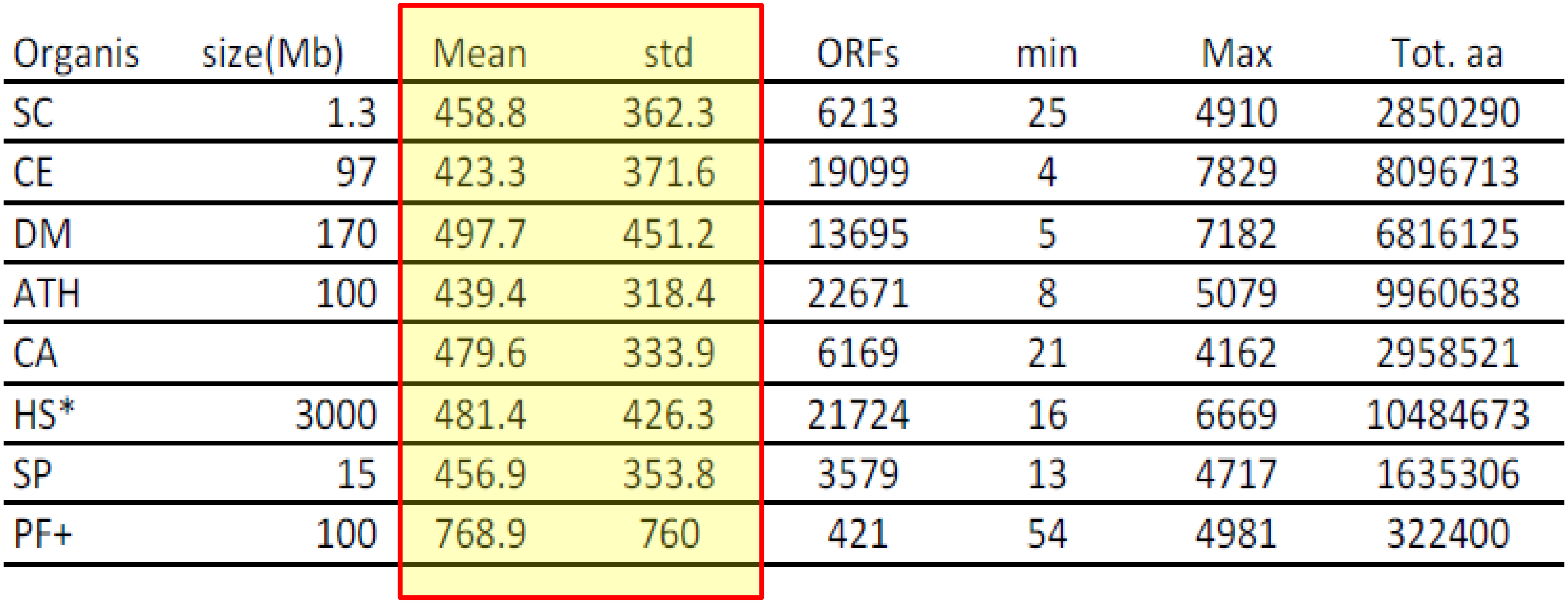


H. sapiens : green
C. elegans: blue
D. melanogaster: red

one intron in the human neurexin gene is approx. 480,000 nt !

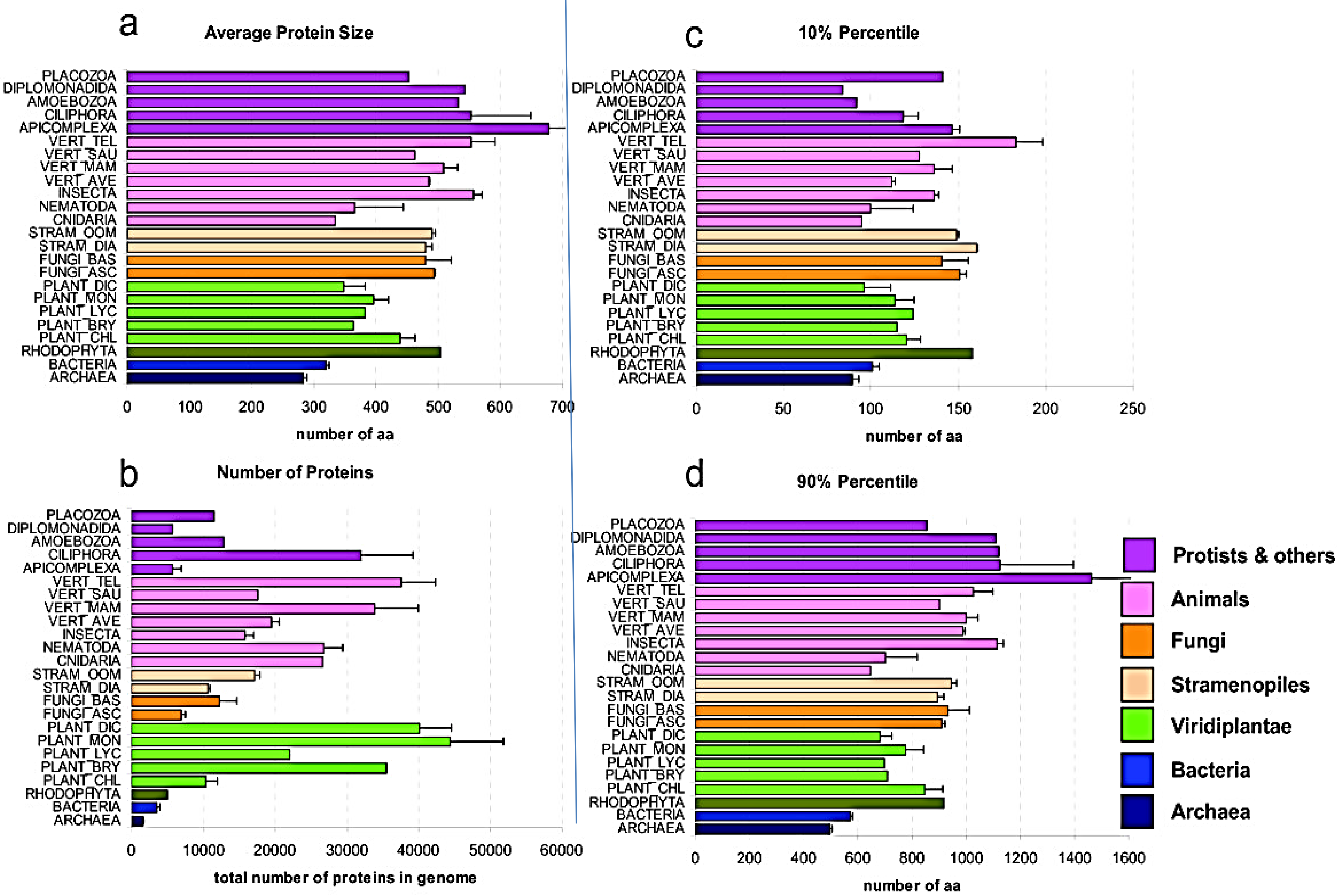
While genes vary enormously in size from bacteria to mammals, due to intronic prevalence, **coding regions** (ORF) are quite uniform, possibly due to protein structural constraints.
Note that the absolute number of genes does not follow organism complexity.

Predicted ORF products mean size in completely sequenced organisms



Organis	size(Mb)	Mean	std	ORFs	min	Max	Tot. aa
SC	1.3	458.8	362.3	6213	25	4910	2850290
CE	97	423.3	371.6	19099	4	7829	8096713
DM	170	497.7	451.2	13695	5	7182	6816125
ATH	100	439.4	318.4	22671	8	5079	9960638
CA		479.6	333.9	6169	21	4162	2958521
HS*	3000	481.4	426.3	21724	16	6669	10484673
SP	15	456.9	353.8	3579	13	4717	1635306
PF+	100	768.9	760	421	54	4981	322400

Average a.a. ~ 128 Da   in peptides: 110 Da



Summary of protein number and protein size (set 1). Comparison of the protein length attributes in species from different phylogenetic groups. Species were grouped as indicated in Table 1. a) Average protein size. b) Total number of proteins in genome. c) Average of the 10% percentiles. d) Average of the 90% percentiles. Bars indicate mean values ± standard error (SE). In panels acd the x axis indicates the number of amino acids (aa), whereas in panel b it gives the average number of proteins in those species. Tiessen *et al. BMC Research Notes* 2012 **5**:85

Introduction to GENOME BROWSERS

Genomic database have developed a way to «see» genes and sequences

ENSEMBL -

NCBI - Gene (https://www.ncbi.nlm.nih.gov/gene/?term=)

UCSC -

UWASH

and others

Other background from Genetics

Genes «families»

Similarity in «parts» of the proteins, called «domains»:

Paralogy and Orthology

Mechanisms of evolution

evolution

Post-genomics

Genetics

Comparative (phylogenetic conservation indicates conserved function)

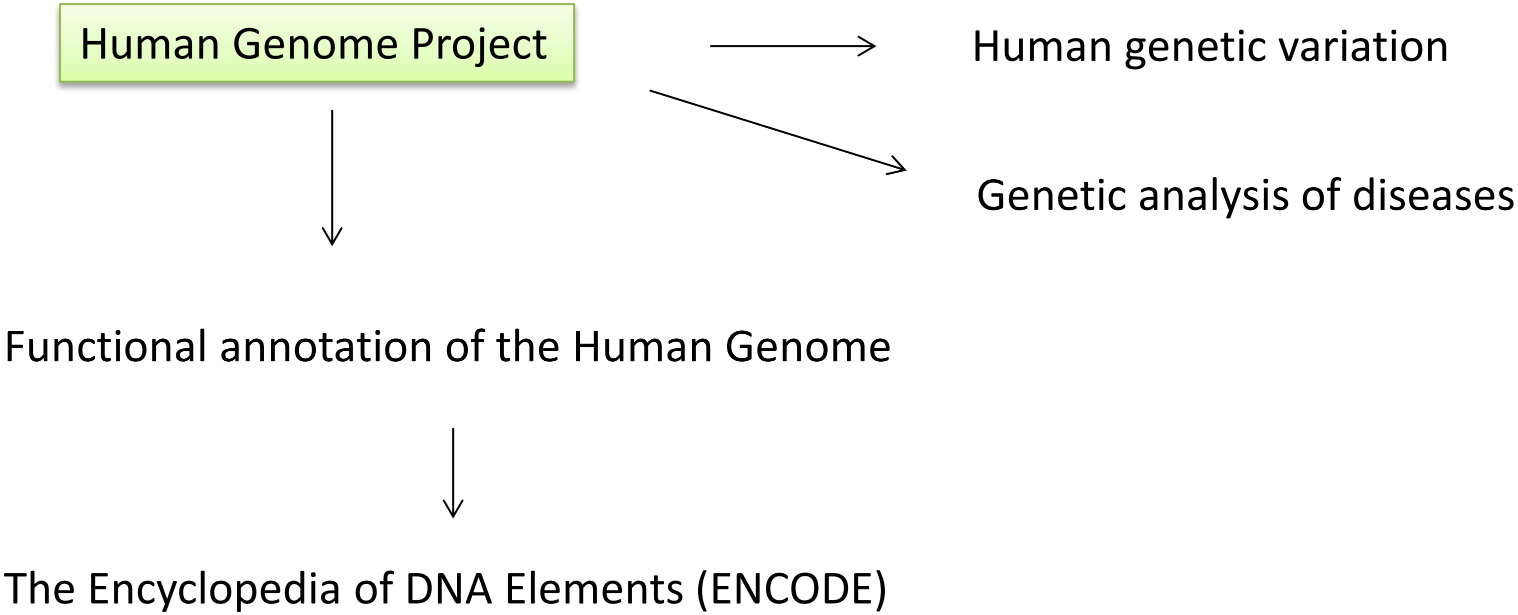
Human Genetic Variation (1000 Human Genomes - HapMap)

GWAS – Genome variations – phenotype correlation

Gene expression and phenotype

Functional Genomics (ENCODE – FANTOM)

Epigenomics:	CpG methylation
	Histone modifications (PTMs)
	Chromatin status
	Protein-DNA mapping (e.g. transcription factors
Transcriptomics:	Coding and noncoding RNAs



Human Genome Project

Human genetic variation

Genetic analysis of diseases

Functional annotation of the Human Genome

The Encyclopedia of DNA Elements (ENCODE)

The idea was to obtain functional information for every single nucleotide of the human genome

Started in 2000 using automated Sanger sequencing on 1% human genome (ca. 30 Mb), completed in 2006

With the advent of Next Generation Sequencing Technology, first draft completed in 2012

Genetics

Individual genomes display **variants**

SNP – single nucleotide polymorphisms

Indels – insertions and deletions

CNV – copy number variations

Variants are associated to more or less evident **phenotypes**

Some variants are clearly associated to specific **pathologies**.

Other variants are associated only weakly with a phenotype but require other variants (often in other loci) to become significantly associated (combinatorial association).

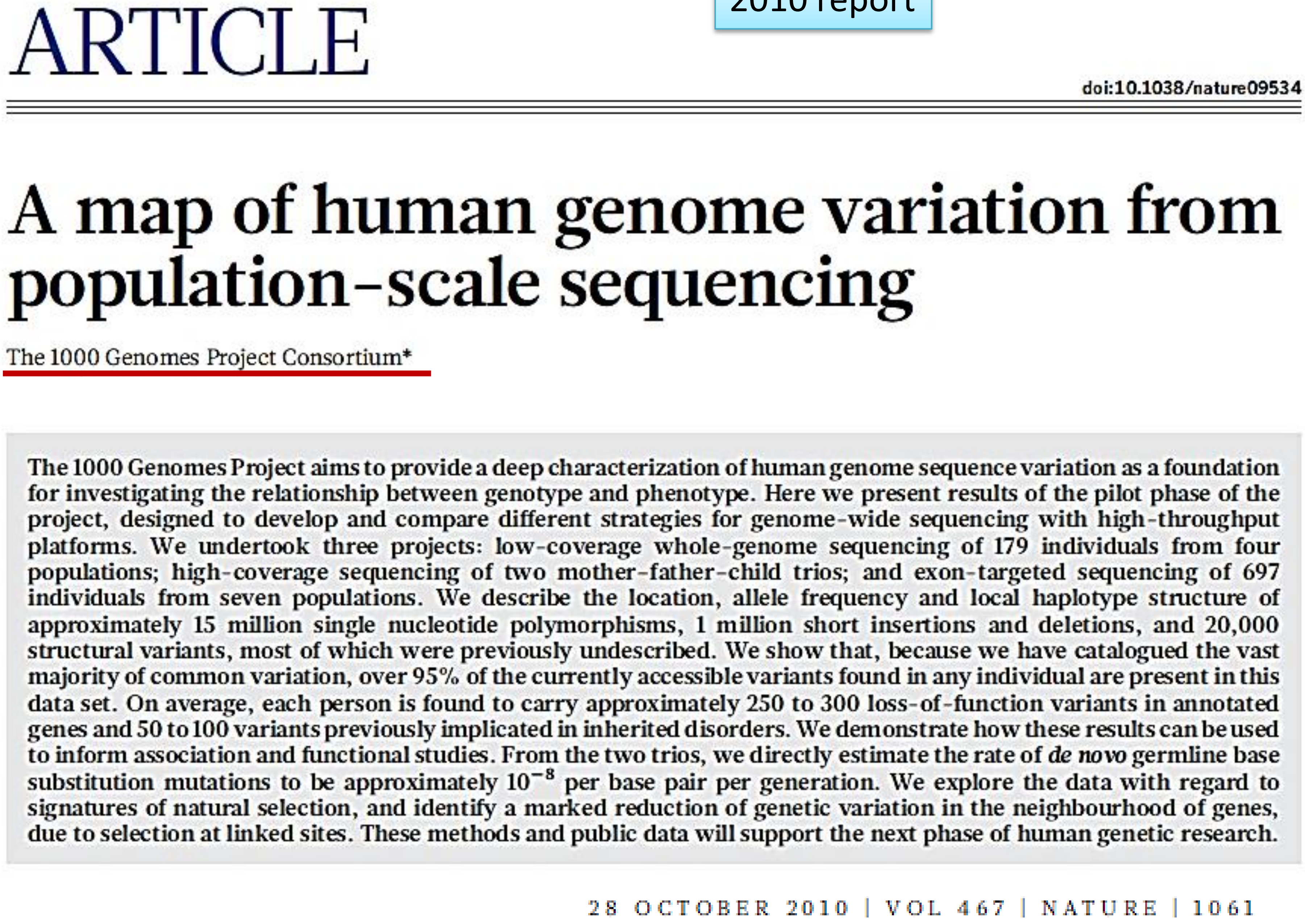
Projects are under way to describe all variants associated to risk of disease (GWAS: Genome Wide Association Studies)

The 1000 Genomes Project

http://www.internationalgenome.org/

Started immediately after the HGP but it has been dramatically accelerated by introduction of NGS

2010 report



ARTICLE

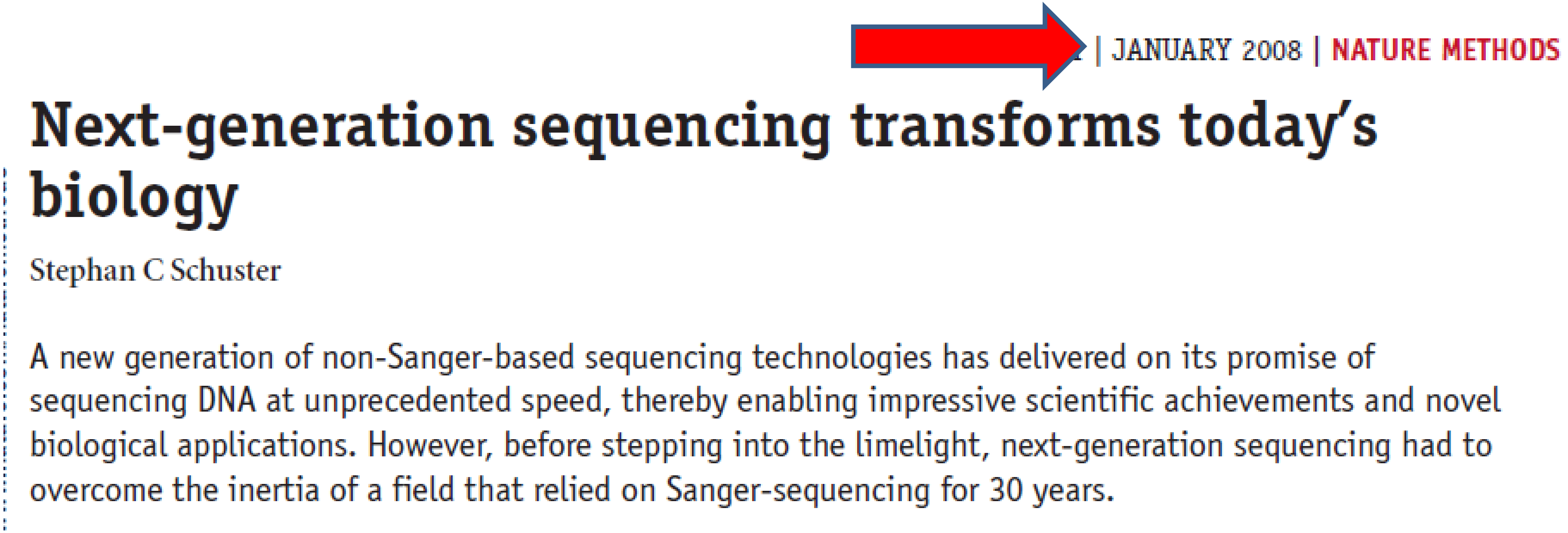
doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

The 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. Here we present results of the pilot phase of the project, designed to develop and compare different strategies for genome-wide sequencing with high-throughput platforms. We undertook three projects: low-coverage whole-genome sequencing of 179 individuals from four populations; high-coverage sequencing of two mother–father–child trios; and exon-targeted sequencing of 697 individuals from seven populations. We describe the location, allele frequency and local haplotype structure of approximately 15 million single nucleotide polymorphisms, 1 million short insertions and deletions, and 20,000 structural variants, most of which were previously undescribed. We show that, because we have catalogued the vast majority of common variation, over 95% of the currently accessible variants found in any individual are present in this data set. On average, each person is found to carry approximately 250 to 300 loss-of-function variants in annotated genes and 50 to 100 variants previously implicated in inherited disorders. We demonstrate how these results can be used to inform association and functional studies. From the two trios, we directly estimate the rate of *de novo* germline base substitution mutations to be approximately 10^{-8} per base pair per generation. We explore the data with regard to signatures of natural selection, and identify a marked reduction of genetic variation in the neighbourhood of genes, due to selection at linked sites. These methods and public data will support the next phase of human genetic research.

28 OCTOBER 2010 | VOL 467 | NATURE | 1061



| JANUARY 2008 | NATURE METHODS

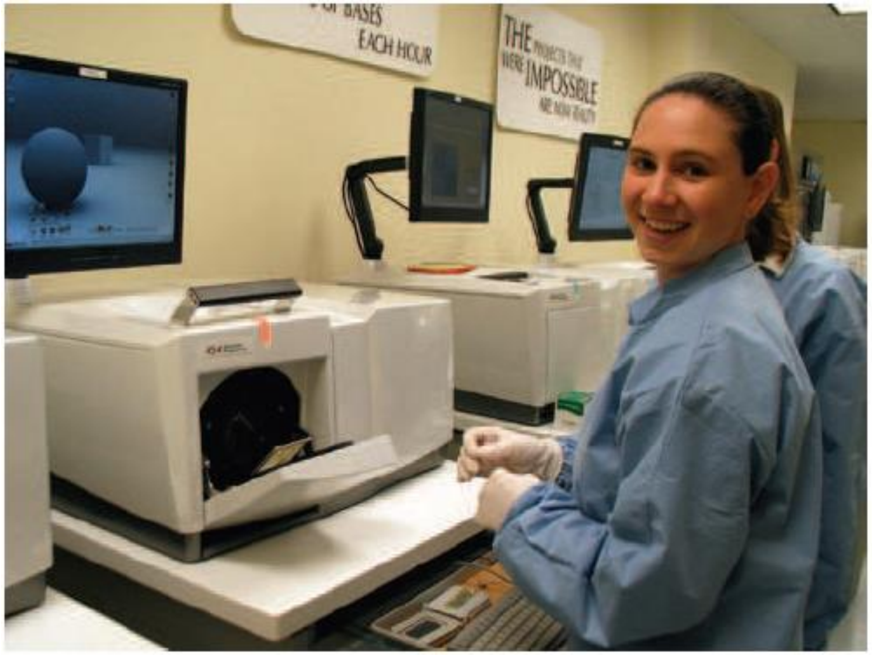
Next-generation sequencing transforms today's biology

Stephan C Schuster

A new generation of non-Sanger-based sequencing technologies has delivered on its promise of sequencing DNA at unprecedented speed, thereby enabling impressive scientific achievements and novel biological applications. However, before stepping into the limelight, next-generation sequencing had to overcome the inertia of a field that relied on Sanger-sequencing for 30 years.

Post-Genome projects started in the early 2Ks with the same Sanger tech used for HGP, i.e. cutting-cloning-sequencing.

Projects were greatly accelerated by introduction in 2005-2006 of NGS (Next Generation Sequencing) technologies



The latest next-generation sequencing instruments can generate as much data in 24 h as several hundred Sanger-type DNA capillary sequencers, but are operated by a single person.

NGS

Fragment the DNA (or RNA) to be sequenced in smaller pieces

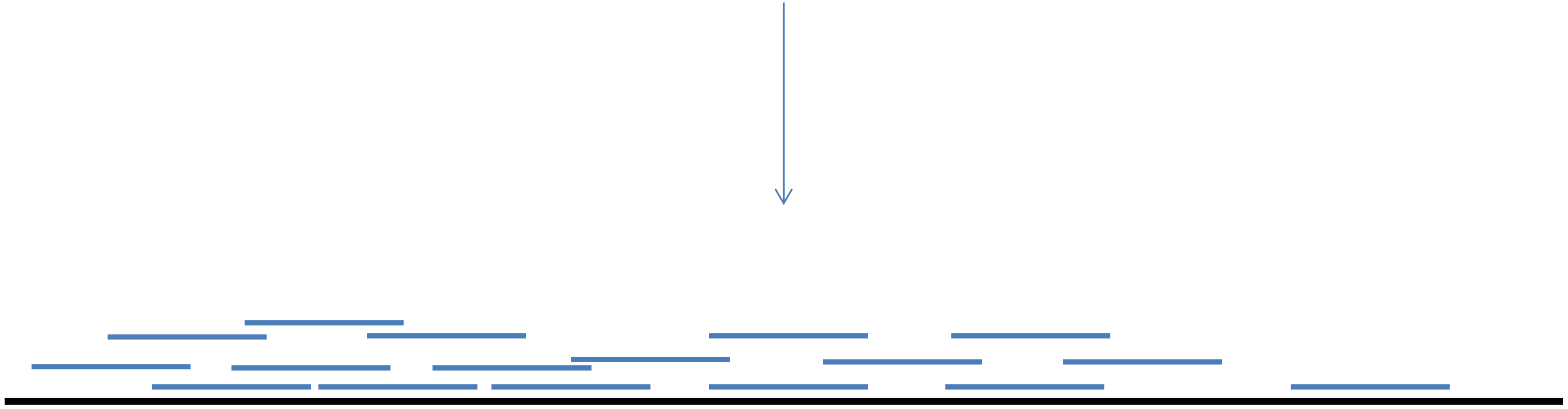
Physically separate the fragments

Highly-parallel sequencing of fragments, high-throughput

No cloning step required

NGS sequencing produces hundreds of millions of short «reads» per run

Reads are mapped to the reference genome



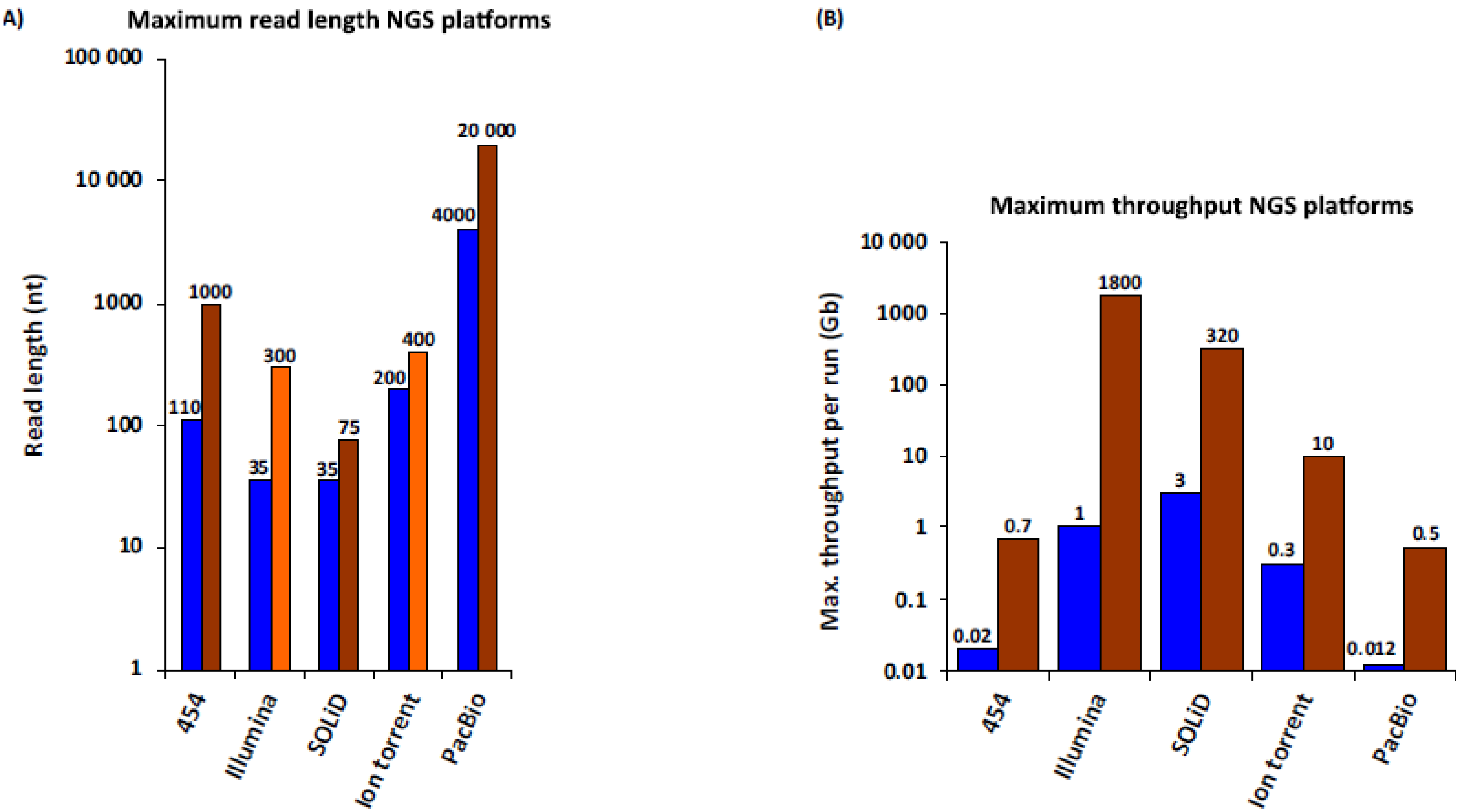
reference genome

In NGS sequencing, the number of independent sequences (called «reads») is more important than lenght

The % of reference genome that is represented in «reads» is the «**coverage**».

Other essential aspects:

1) speed
2) cost
3) error-to-depth ratio



In blue the first version of the instruments

From Van Dijk et al., 2014 (Textbook)

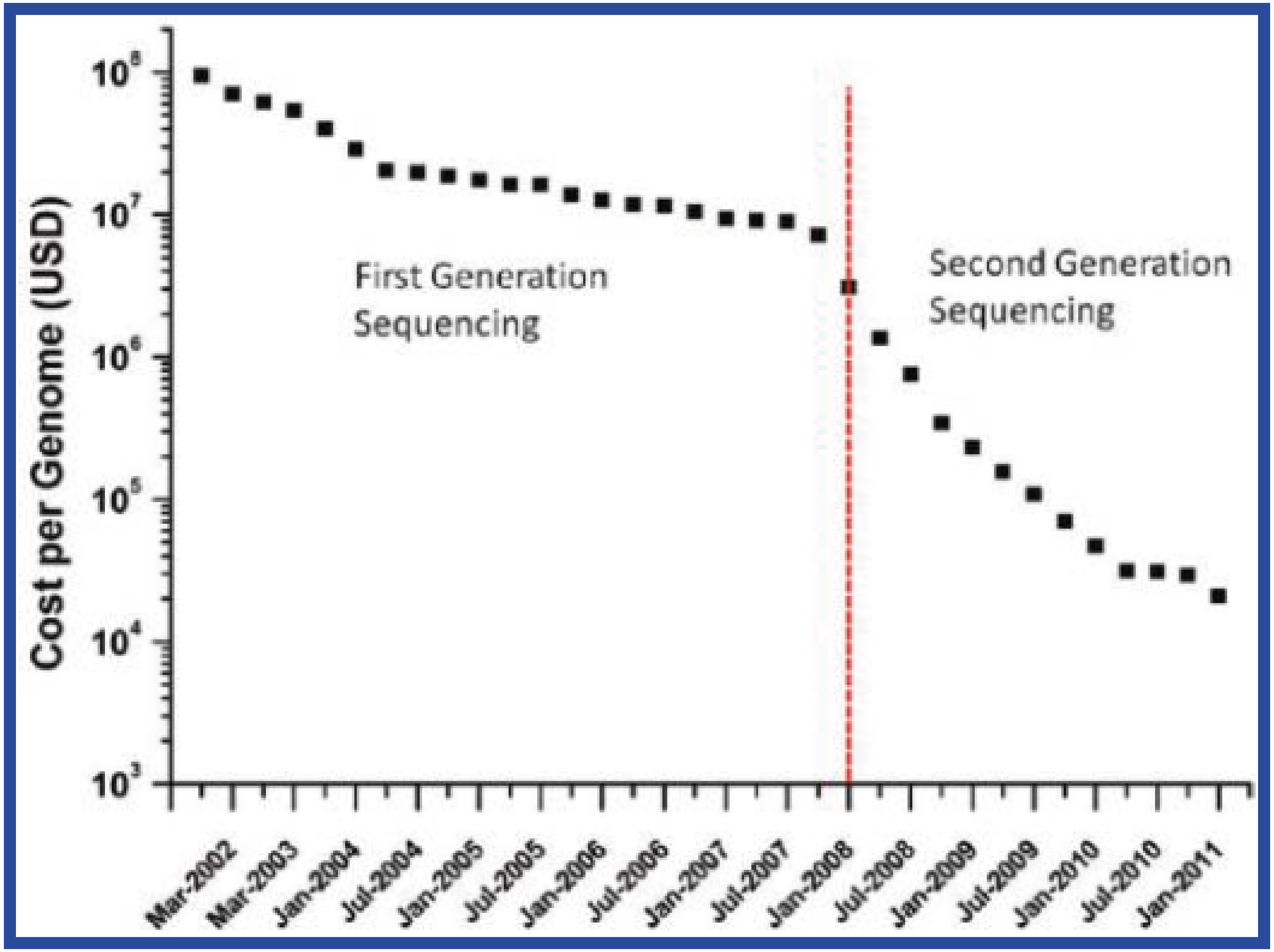


Figure 1. Estimated cost required to sequence a complete human genome based on data generated from NHGRI-funded large-scale DNA sequencing centers.[28]

From Niedringhaus et al., 2011

Post-genomics

Genetics

Comparative (phylogenetic conservation indicates conserved function)

Human Genetic Variation (1000 Human Genomes - HapMap)

GWAS – Genome variations – phenotype correlation

Gene expression and phenotype

Functional Genomics

Epigenomics:	CpG methylation
	Histone modifications (PTMs)
	Chromatin status
	Protein-DNA mapping (e.g. transcription factors
Transcriptomics:	Coding and noncoding RNAs

1000 Human Genomes, HapMap project
Describing variations among genomes of individuals

GWAS
Genome-wide association studies
Variations (SNPs, CNV, indels) studied in individuals as related to the occurence of a phenotype (pathology, risks, other features)

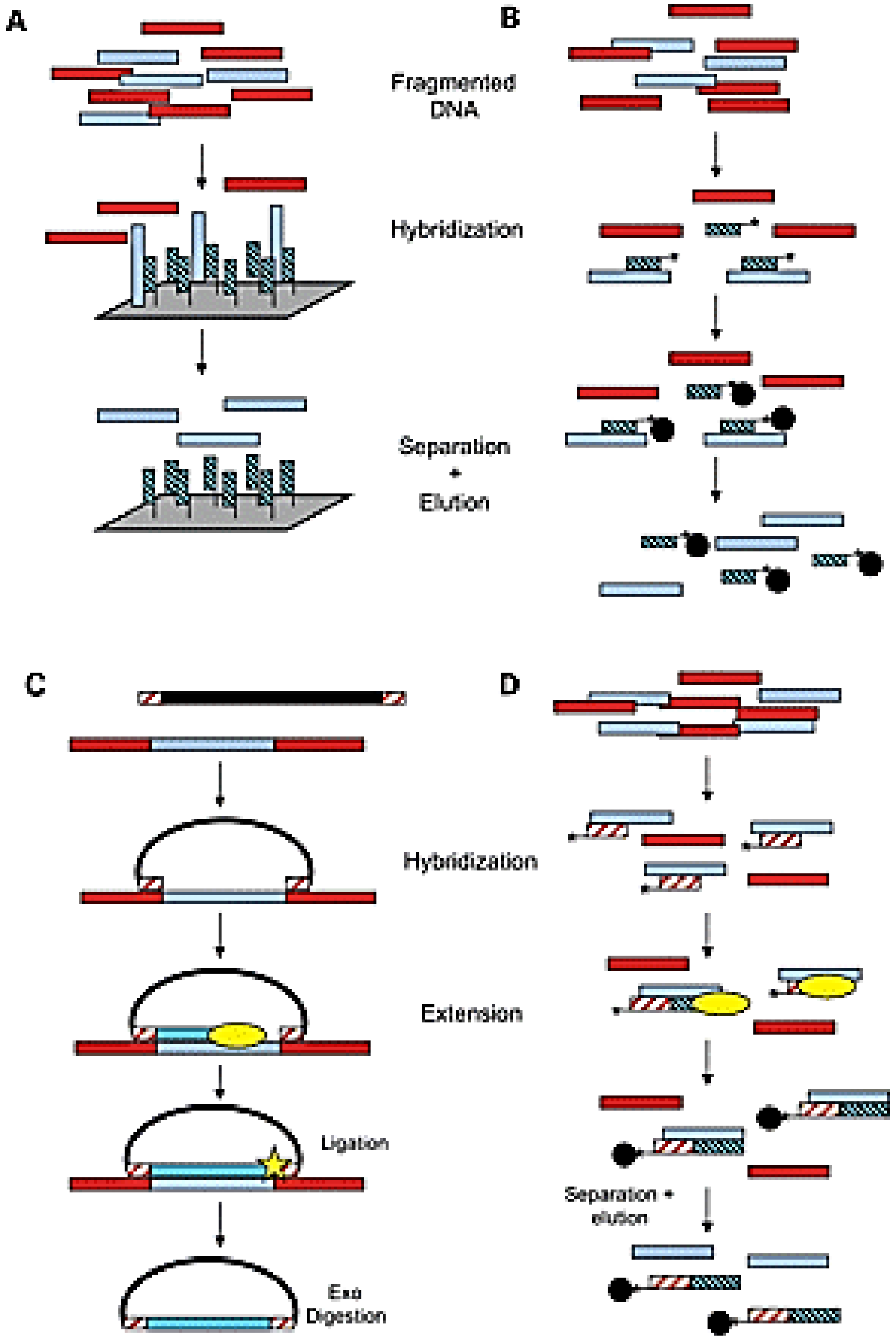
TCGA – The Cance Genome Atlas
Sequencing of tumor cell DNA to evidence mutations occurring in tumors.

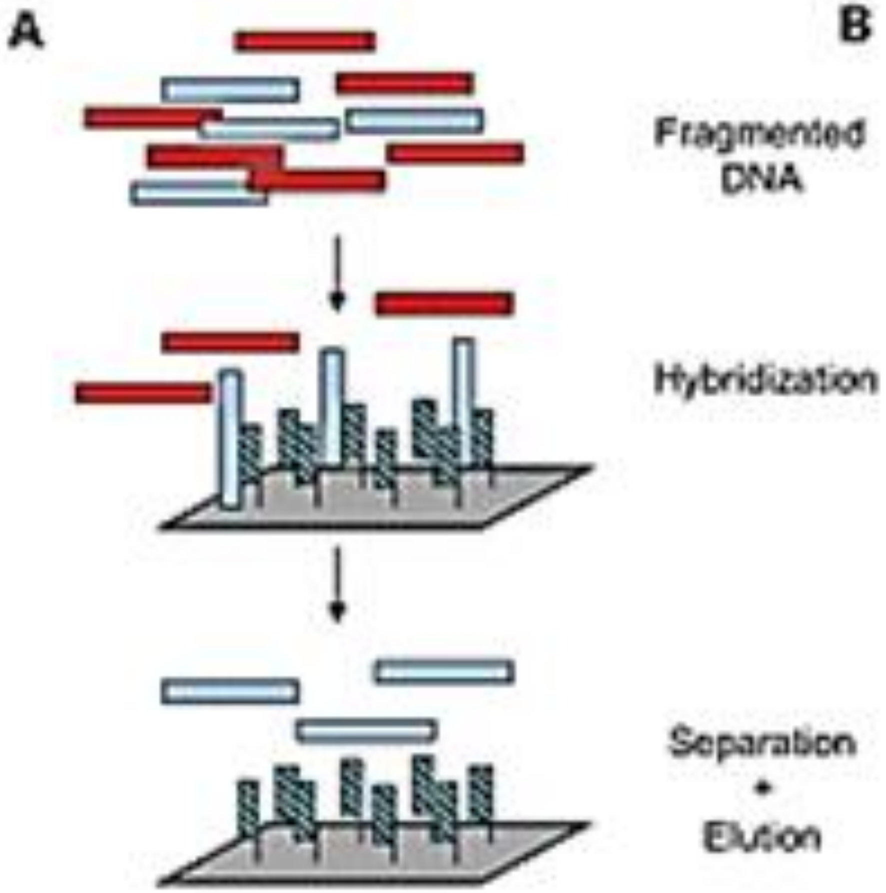
Exome sequencing

Due to elevated costs, many studies were limited to the «**exome**»
Exome is the set of sequences that make up all known mRNAs.

Requires enrichment of exon sequences from a genomic DNA. This is obtained using different methods, as exemplified in these schemes.

From: Teer and Mullikin, 2010. Hum Mol Genet. 9(R2):R145-51




A
B
Fragmented DNA
Hybridization
Separation
+
Elution

see one gene variants using NCBI or ENSEMBL

No class tomorrow

Readings:

Textbook - Geyer_2011_nuclear_organization

Research Paper – Reddy et al, 2008