L2.1

Genomes - Epigenomes

In the previous Chapter we have seen how the Genome is physically organized in different states, through both sub-nuclear compartmentalization and histone and nonhistone proteins modulation.

These different organization states are not distributed randomly, nor in a «pulverized», gene-by-gene fashion.

Instead, genomic domains of variable size are homogeneously organized as «constitutive heterochromatic», «facultative heterochromatic», or «euchromatic» (…just widely speaking).

Questions:

1. what are the determinants of these states ?

2. what is the mechanism leading to «domain» uniformity and boundaries ?

3. is the chromatin status inheritable mitotically ?

These are fundamental questions in Biology:

the Genome is organized during development and differentiation and this organization is conservatively propagated through cell division. Thus, the genome is «programmed» through the **establishment** and **maintenance** of epigenetic information.

Finally, we will see how certain characters that are not passed on following classical Mendelian rules depend upon trans-generational «epigenetic» inheritance

In the second Chapter of the course, we will examine first which are the **determinants** of chromatic states (information input), second how this information in managed **locally** (i.e. within the chromatin domain); third how this information is **passed on** during cell division or between generations.

Definition of Epigenetics:

Any kind of information that is not contained in the DNA sequence itself, which can be transmitted mitotically.

Epigenetic inheritance: (Campos-textbook)

The inheritance of a phenotype in a manner that is independent of the DNA sequence and that remains self-perpetuating in the absence of the initial stimulus that determined the phenotype in the parental cell or organism.

1. Does **DNA sequence** influence chromatin organization ?

Evidence 1:
Certain organization of chromatin is intrinsic and depends on the DNA sequence itself. Repetitive DNA is an example of this.

Evidence 2:
We said that chromatin organization follows a «domain» rule. How are domain ends determined ? Again, specific sequences called «insulators» do the job (through interaction with specific proteins).

Evidence 3:
How does a «signal» reach chromatin and drive its organization ? This job is made by Transcription Factors, sequence-specific DNA binding proteins that bring Writers, Readers and Erasers (WRE) to specific locations of Genome. Therefore, the Transcription Factor Binding Site are DNA sequence determinants of chromatin states.

DNA sequence

HGP

NGS

Human Variation

Databases

Composition

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

?

1990-1998 - 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

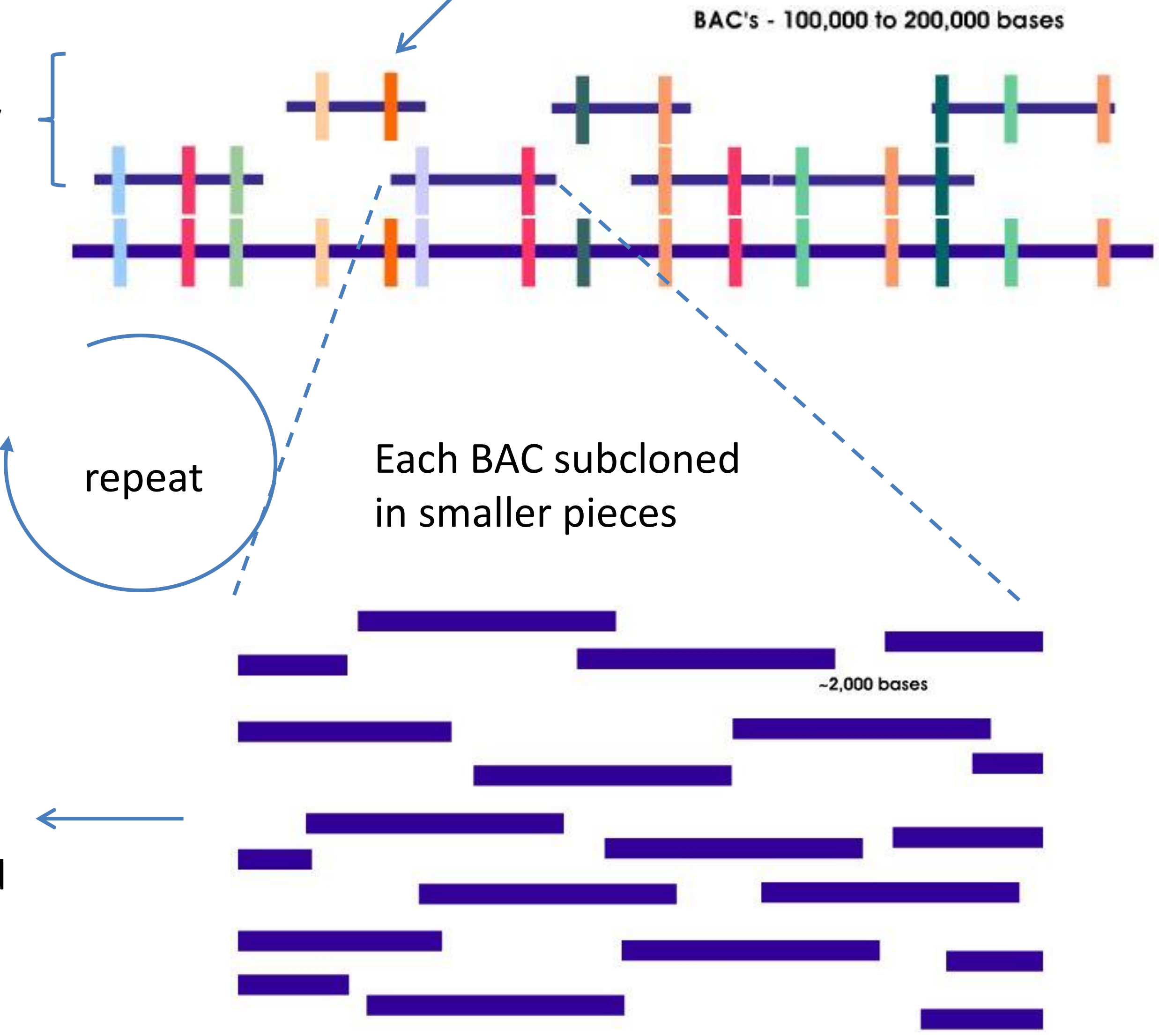
- BACs, PACs – 100-200 Kb
- Cosmids and other phage-derived vectors (20-40Kb)
- Plasmids – 2-3 Kb

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

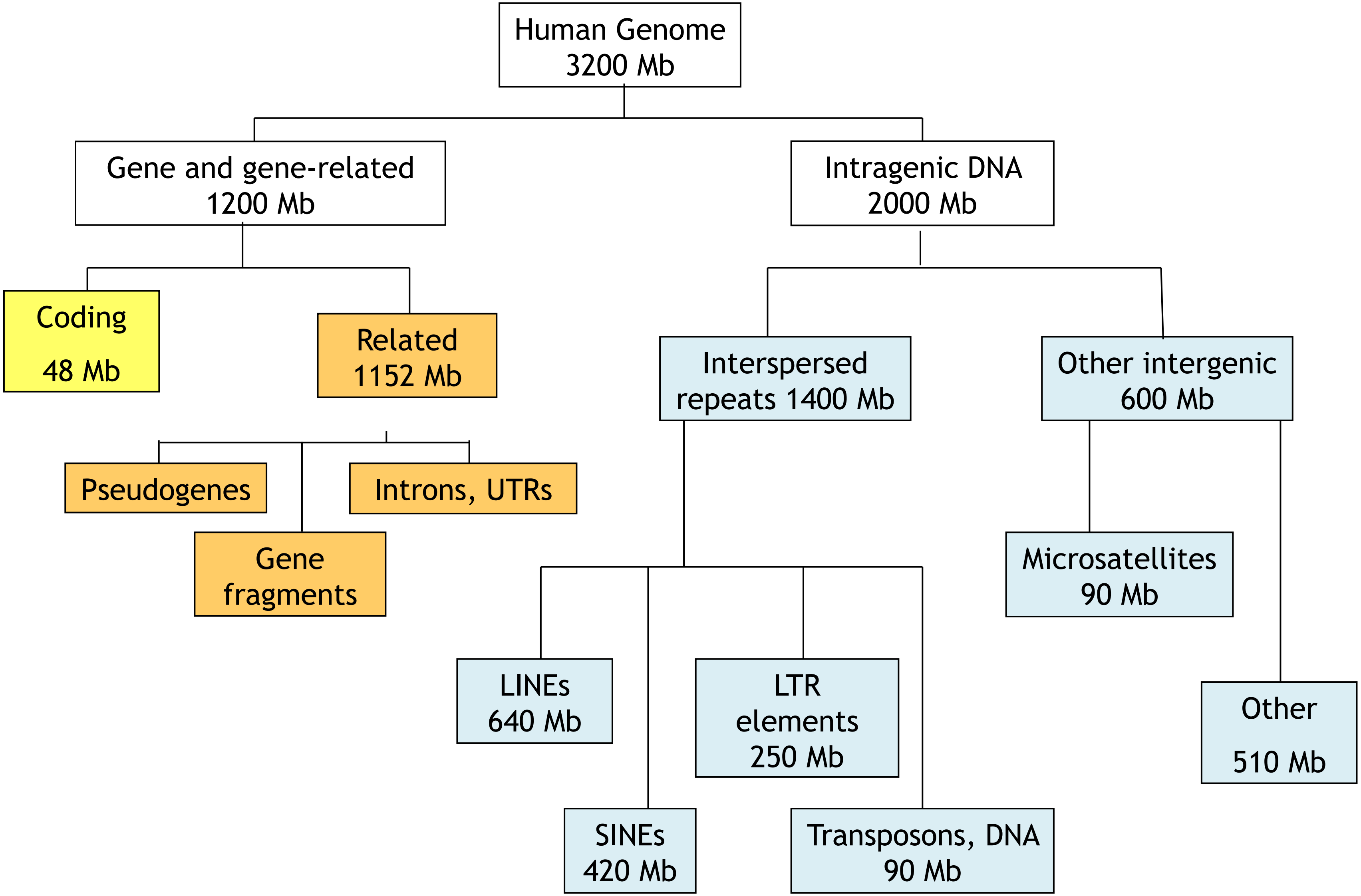
Sanger sequencing

Composition of the Human Genome

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

- Sequence similarity (homology)
- Correspondence to 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

Gene fragments

Introns, UTRs

Interspersed repeats 1400 Mb

Other intergenic
600 Mb

LINEs
640 Mb

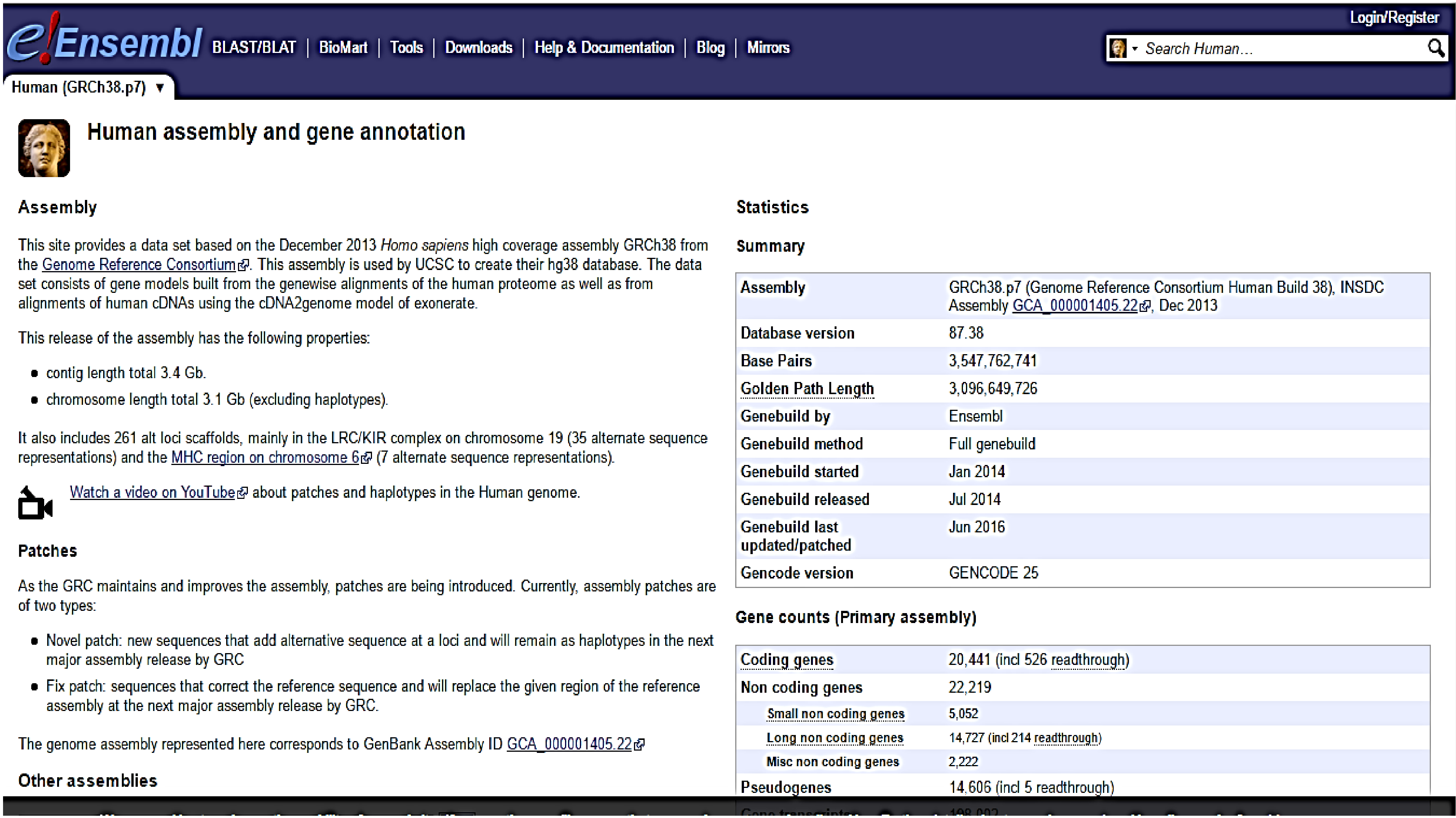
SINEs
420 Mb

LTR elements
250 Mb

Transposons, DNA
90 Mb

Microsatellites
90 Mb

Other
510 Mb


e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login/Register
Search Human…
Human (GRCh38.p7) ▼

Human assembly and gene annotation

Assembly

This site provides a data set based on the December 2013 *Homo sapiens* high coverage assembly GRCh38 from the Genome Reference Consortium. This assembly is used by UCSC to create their hg38 database. The data set consists of gene models built from the genewise alignments of the human proteome as well as from alignments of human cDNAs using the cDNA2genome model of exonerate.

This release of the assembly has the following properties:

- contig length total 3.4 Gb.
- chromosome length total 3.1 Gb (excluding haplotypes).

It also includes 261 alt loci scaffolds, mainly in the LRC/KIR complex on chromosome 19 (35 alternate sequence representations) and the MHC region on chromosome 6 (7 alternate sequence representations).

Watch a video on YouTube about patches and haplotypes in the Human genome.

Patches

As the GRC maintains and improves the assembly, patches are being introduced. Currently, assembly patches are of two types:

- Novel patch: new sequences that add alternative sequence at a loci and will remain as haplotypes in the next major assembly release by GRC
- Fix patch: sequences that correct the reference sequence and will replace the given region of the reference assembly at the next major assembly release by GRC.

The genome assembly represented here corresponds to GenBank Assembly ID GCA_000001405.22

Other assemblies

Statistics

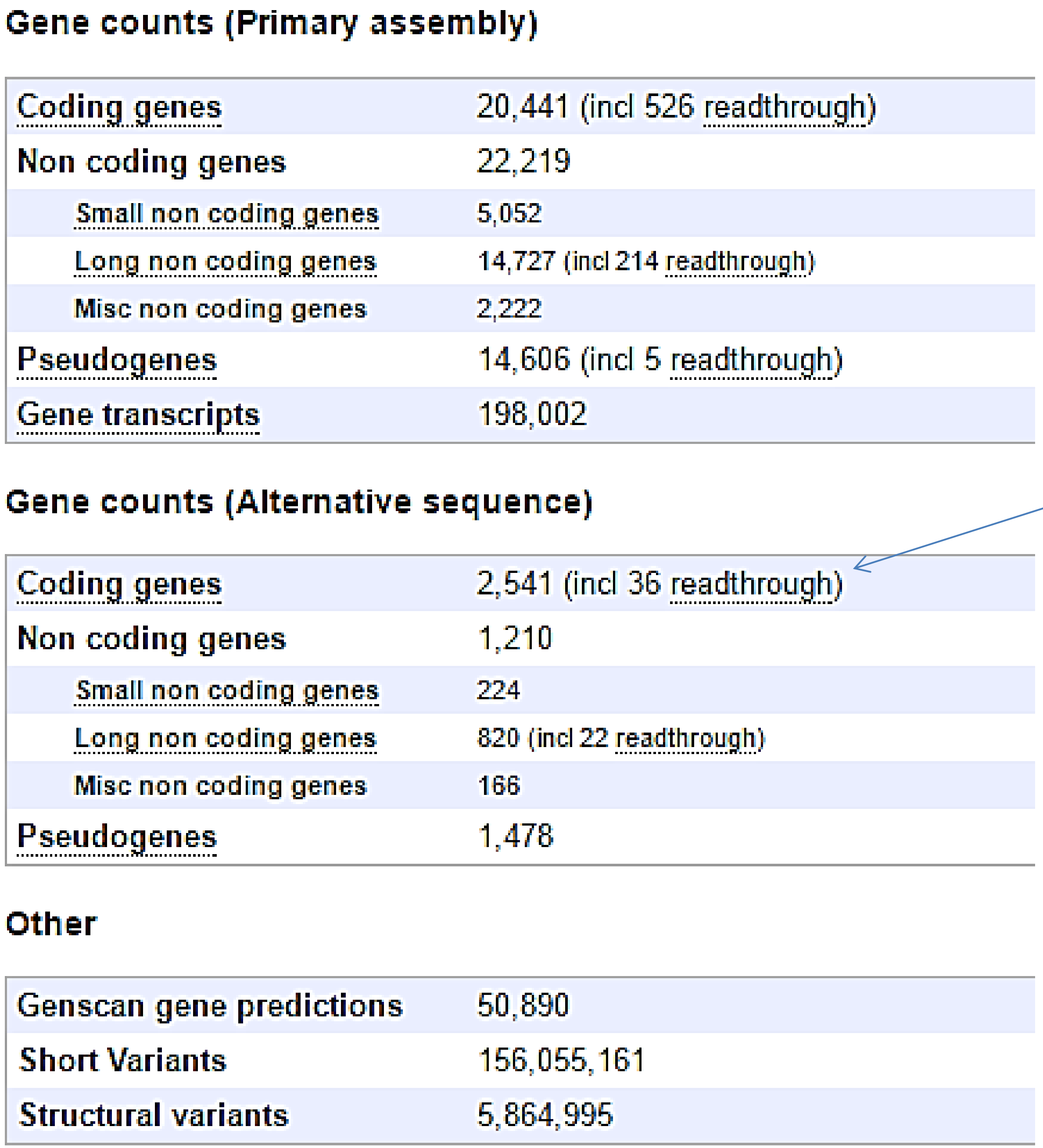
Summary

Assembly	GRCh38.p7 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.22, Dec 2013
Database version	87.38
Base Pairs	3,547,762,741
Golden Path Length	3,096,649,726
Genebuild by	Ensembl
Genebuild method	Full genebuild
Genebuild started	Jan 2014
Genebuild released	Jul 2014
Genebuild last updated/patched	Jun 2016
Gencode version	GENCODE 25

Gene counts (Primary assembly)

Coding genes	20,441 (incl 526 readthrough)
Non coding genes	22,219
Small non coding genes	5,052
Long non coding genes	14,727 (incl 214 readthrough)
Misc non coding genes	2,222
Pseudogenes	14,606 (incl 5 readthrough)

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



Gene counts (Primary assembly)

Coding genes	20,441 (incl 526 readthrough)
Non coding genes	22,219
Small non coding genes	5,052
Long non coding genes	14,727 (incl 214 readthrough)
Misc non coding genes	2,222
Pseudogenes	14,606 (incl 5 readthrough)
Gene transcripts	198,002

Gene counts (Alternative sequence)

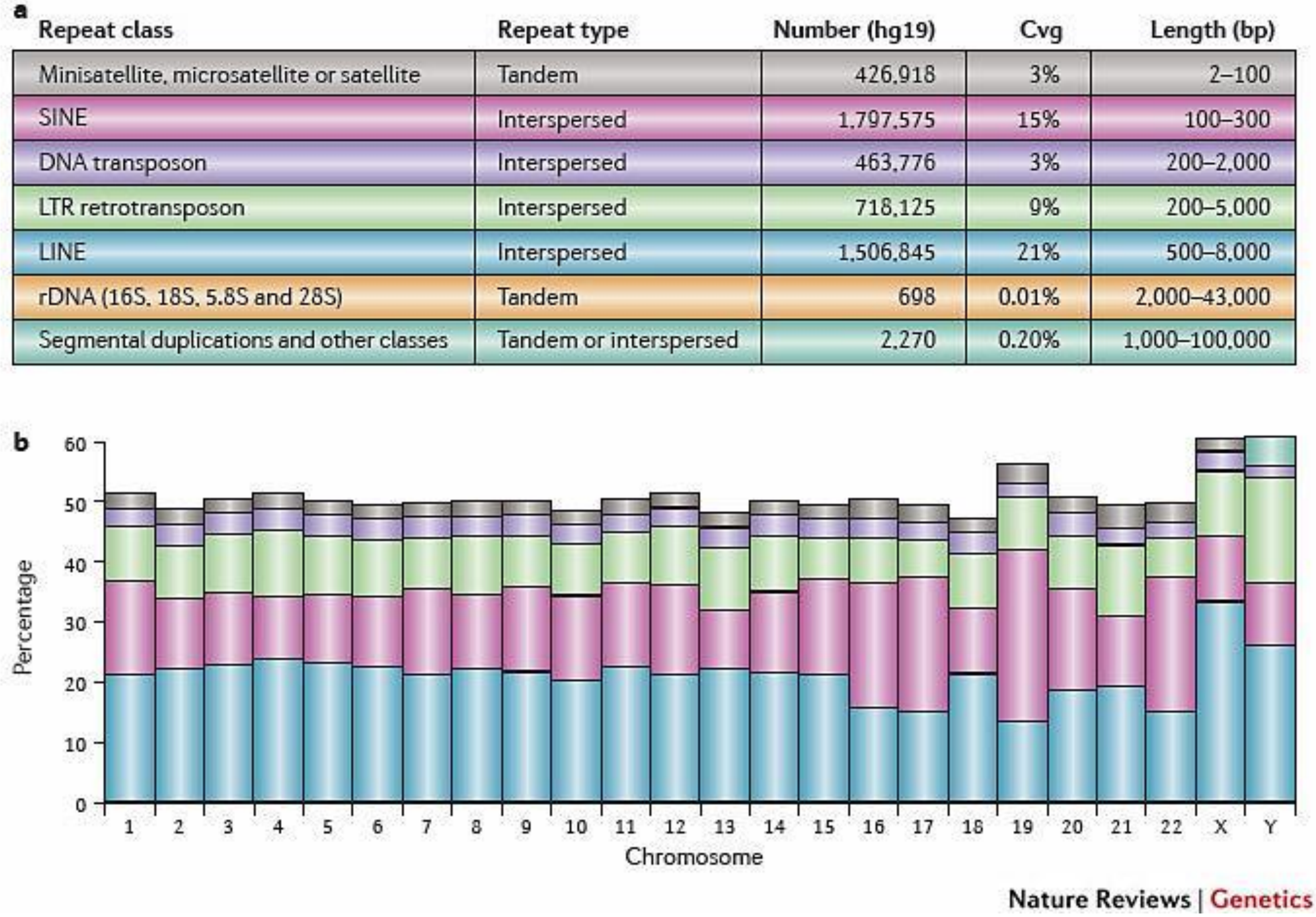
Coding genes	2,541 (incl 36 readthrough)
Non coding genes	1,210
Small non coding genes	224
Long non coding genes	820 (incl 22 readthrough)
Misc non coding genes	166
Pseudogenes	1,478

Other

Genscan gene predictions	50,890
Short Variants	156,055,161
Structural variants	5,864,995

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

Nature Reviews | Genetics

From: Treangen & Salzberg, 2012

Interspersed repetitive elements - Mobile genetic elements

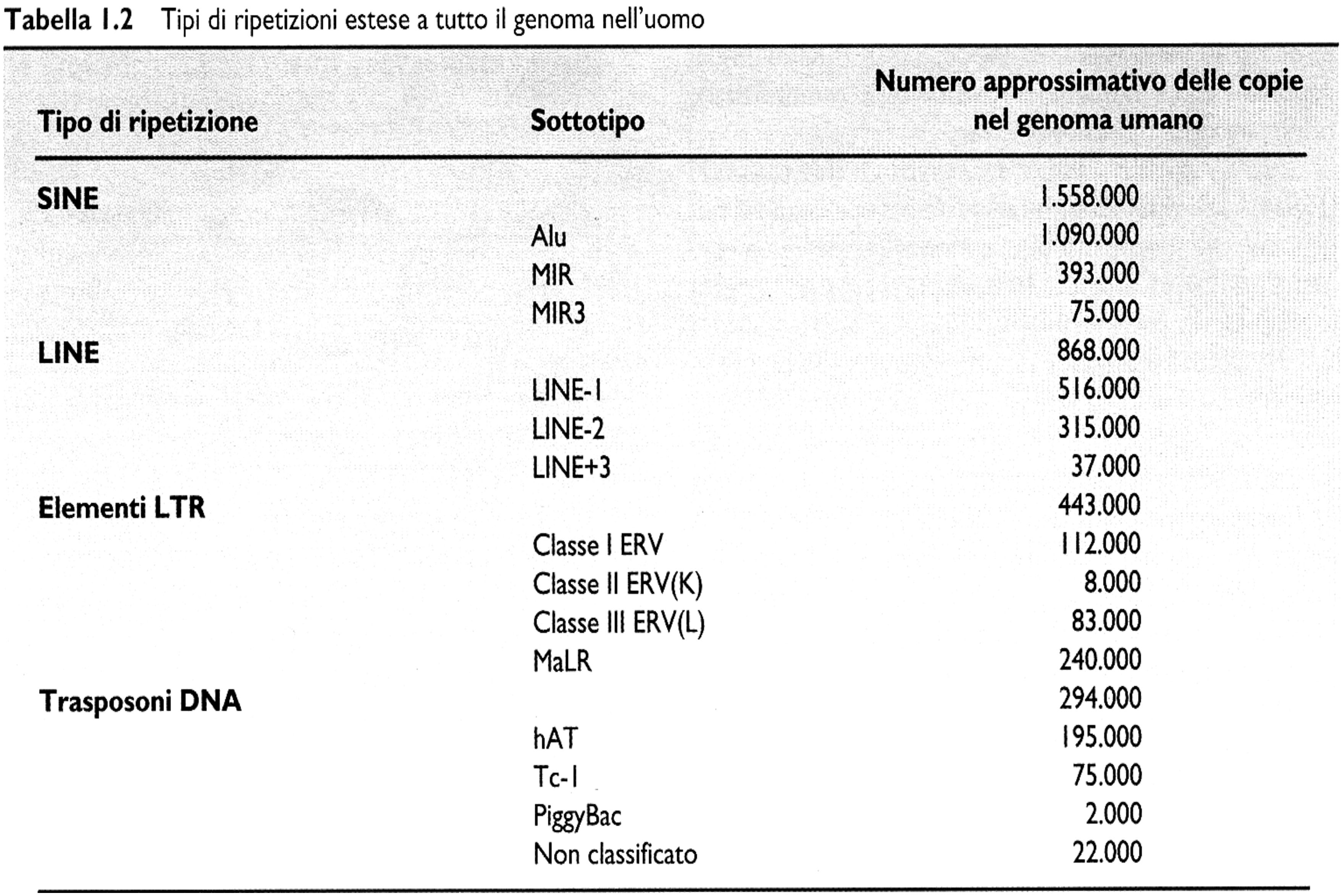
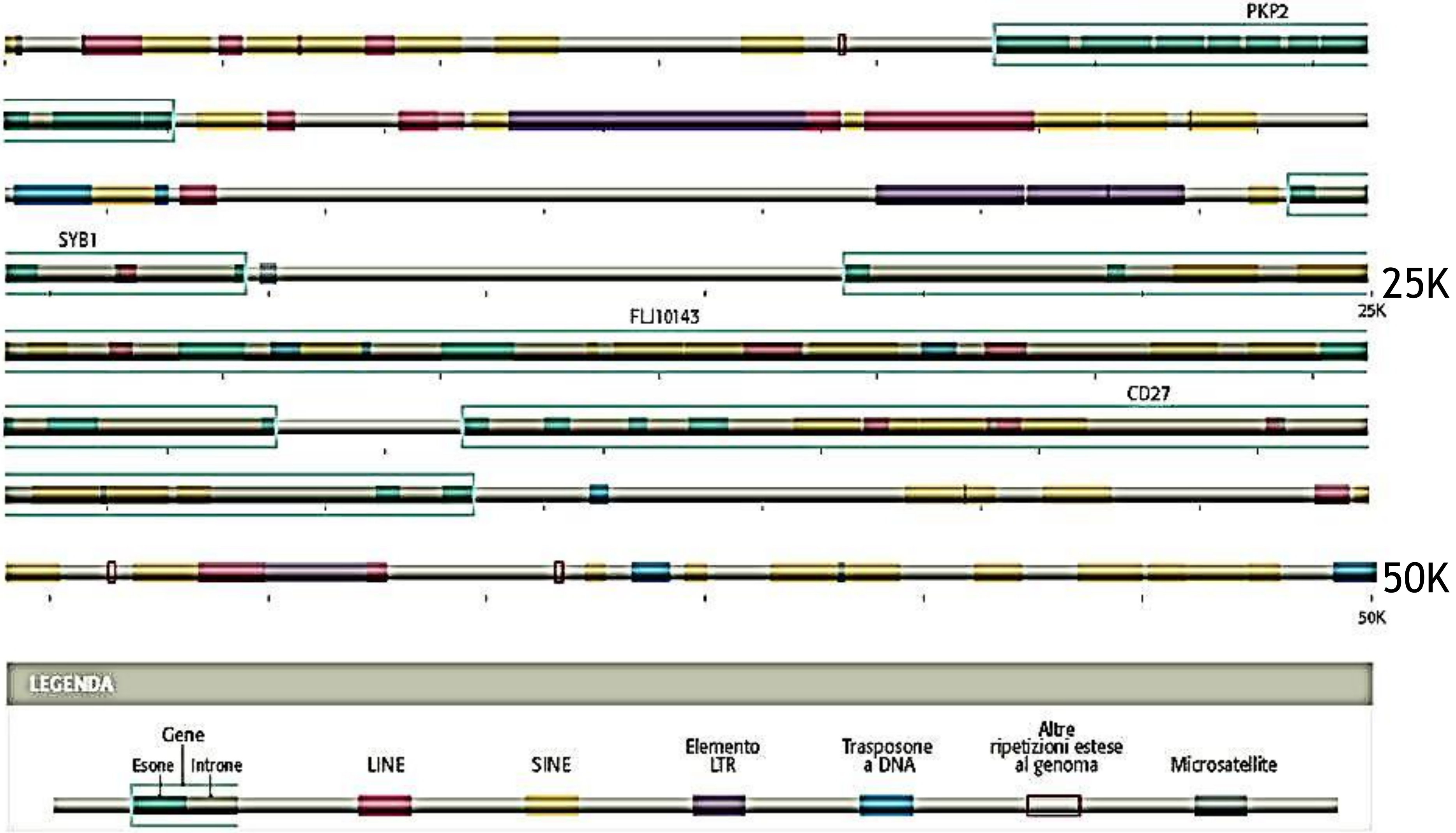


Tabella 1.2 Tipi di ripetizioni estese a tutto il genoma nell'uomo

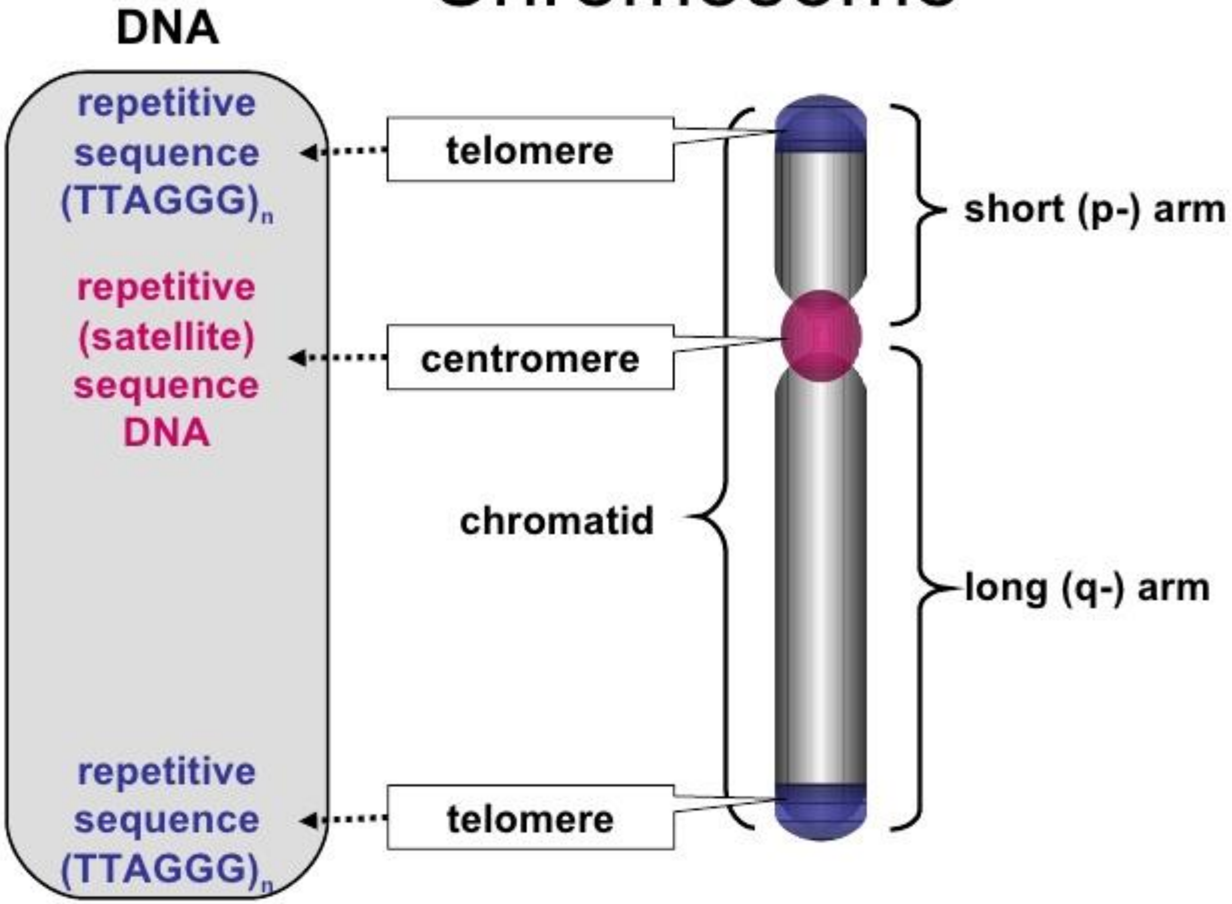
Tipo di ripetizione	Sottotipo	Numero approssimativo delle copie nel genoma umano
SINE		1.558.000
	Alu	1.090.000
	MIR	393.000
	MIR3	75.000
LINE		868.000
	LINE-1	516.000
	LINE-2	315.000
	LINE+3	37.000
Elementi LTR		443.000
	Classe I ERV	112.000
	Classe II ERV(K)	8.000
	Classe III ERV(L)	83.000
	MaLR	240.000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

Figura 7.12 Un tratto del genoma umano. Questa mappa mostra la posizione dei geni, dei segmenti genici, delle ripetizioni estese all'intero genoma e dei microsatelliti in un segmento da 50 kb del cromosoma 12 umano.

A 50 Kb tract of the Human genome



Chromosome



DNA

repetitive sequence $(TTAGGG)_n$

repetitive (satellite) sequence DNA

repetitive sequence $(TTAGGG)_n$

telomere

centromere

telomere

chromatid

short (p-) arm

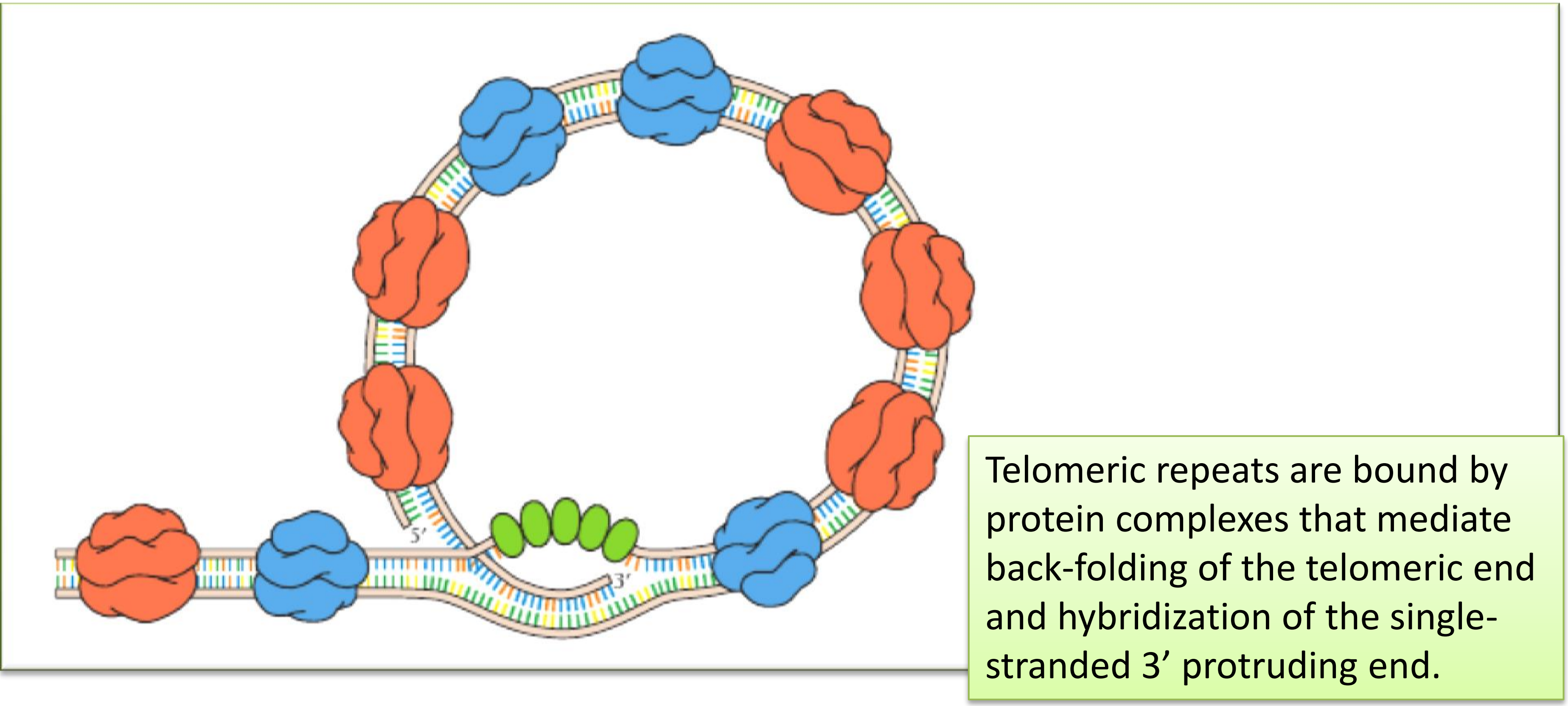
long (q-) arm

Telomers and telomerase

Short tandem sequence repeats at telomeres.

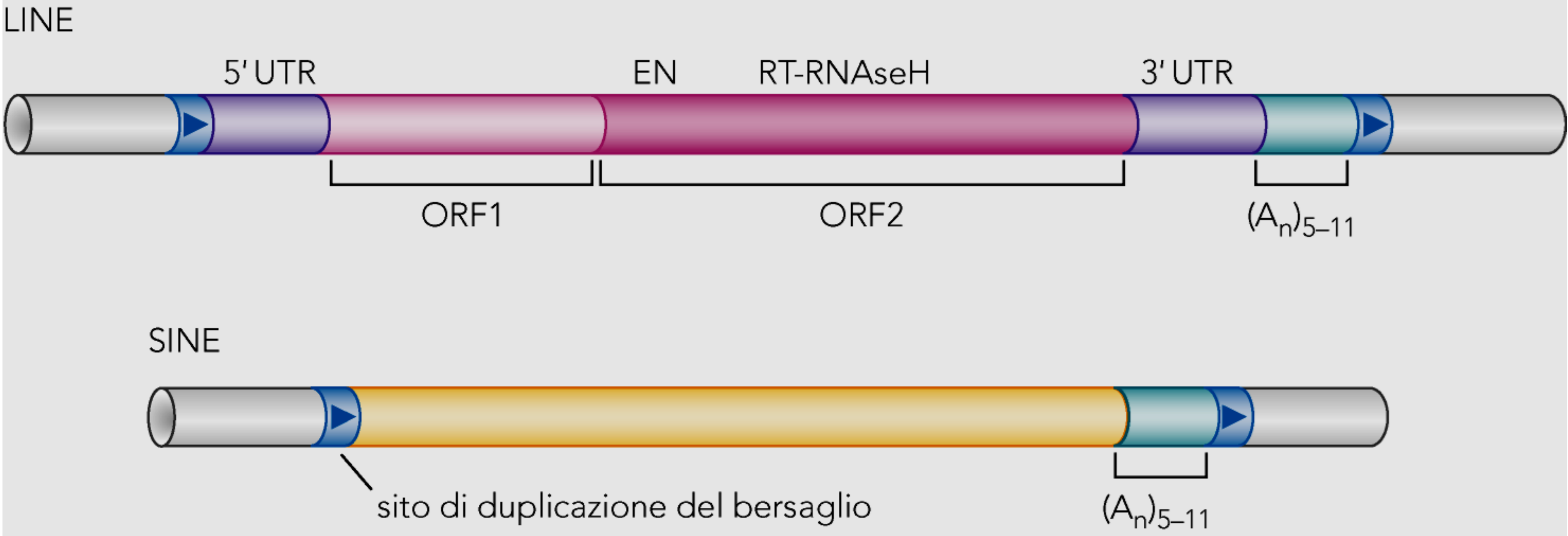
In H. sapiens: TTAGGG (2,500 repeats)

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.

Trasposizione e retrotrasposizione



pseudogenes

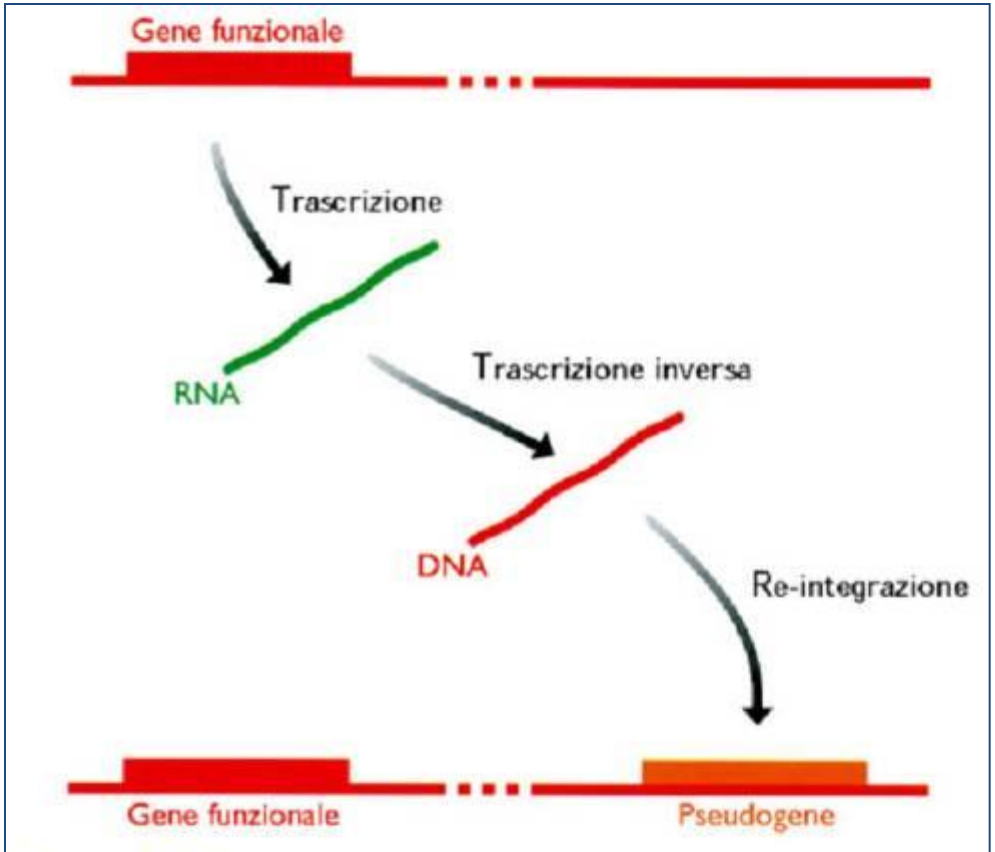
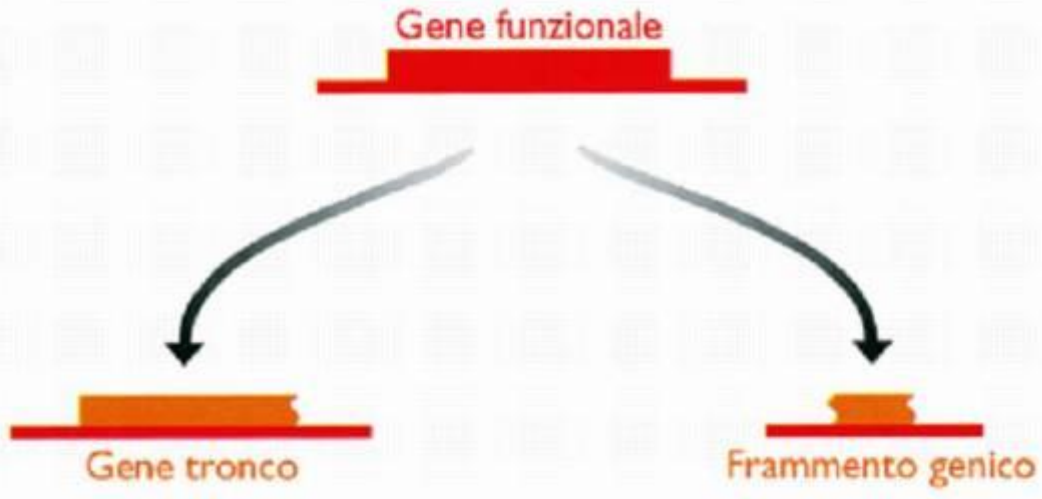


Figura 1.19 L'origine di uno pseudogene maturato.

Retrotranscription-insertion

Second class of pseudogenes are gene copies inactivated by multiple mutations, or:



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** :

NCBI Genomes Genomic Data

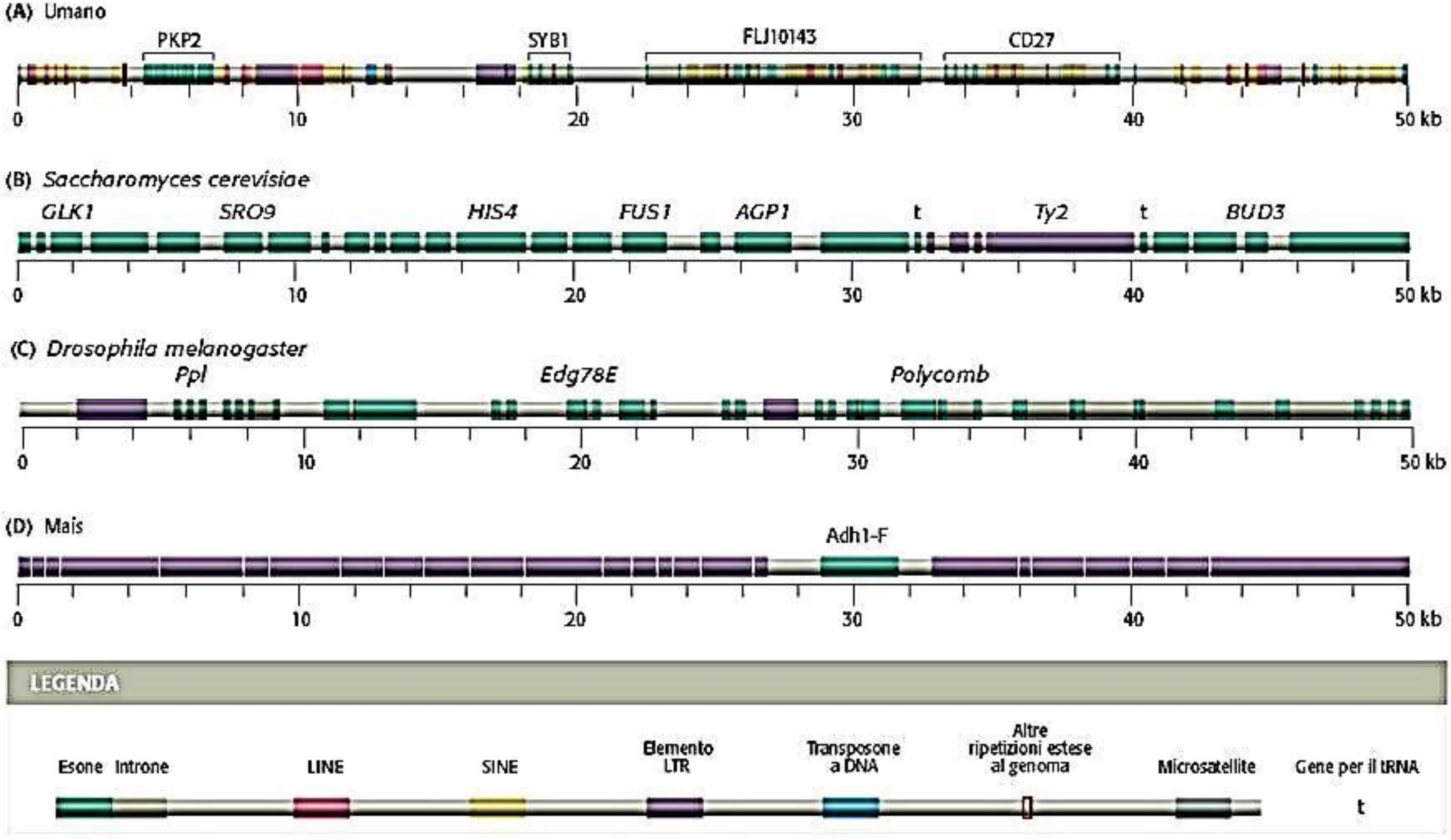
Eukaryotic (Mammals)

T.A. Brown
Genomi, III Ed.
EdiSES

Comparative:

- Human
- Yeast
- Drosophila
- Mais

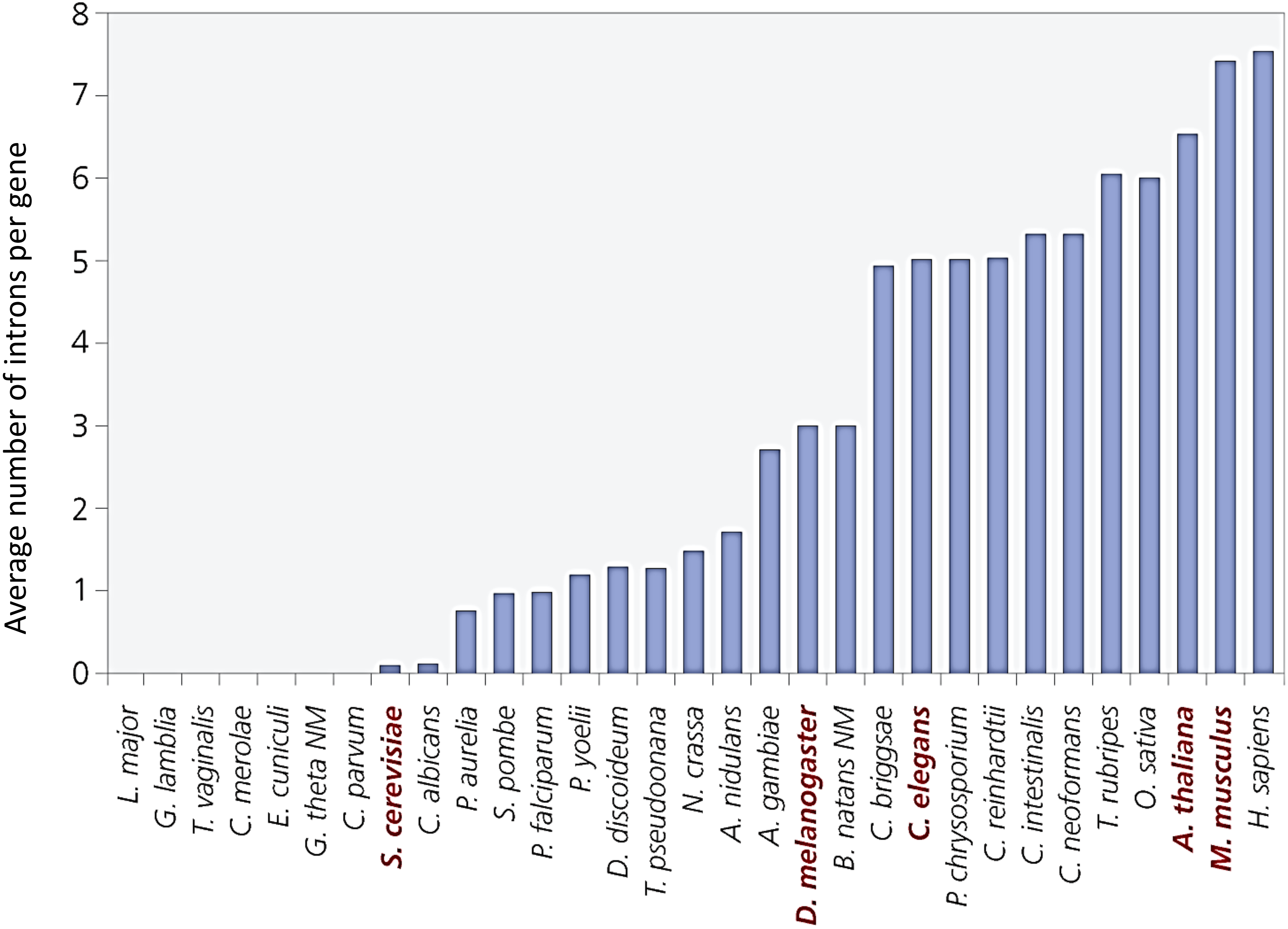
Figura 7.15 Confronto tra genoma umano, di lievito, del moscerino della frutta e di mais. (A) Il segmento di 50 kb del cromosoma 12 umano mostrato precedentemente, è confrontato con segmenti di 50 kb derivanti da genomi di (B) *S. cerevisiae*; (C) *Drosophila melanogaster*; (D) mais.



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

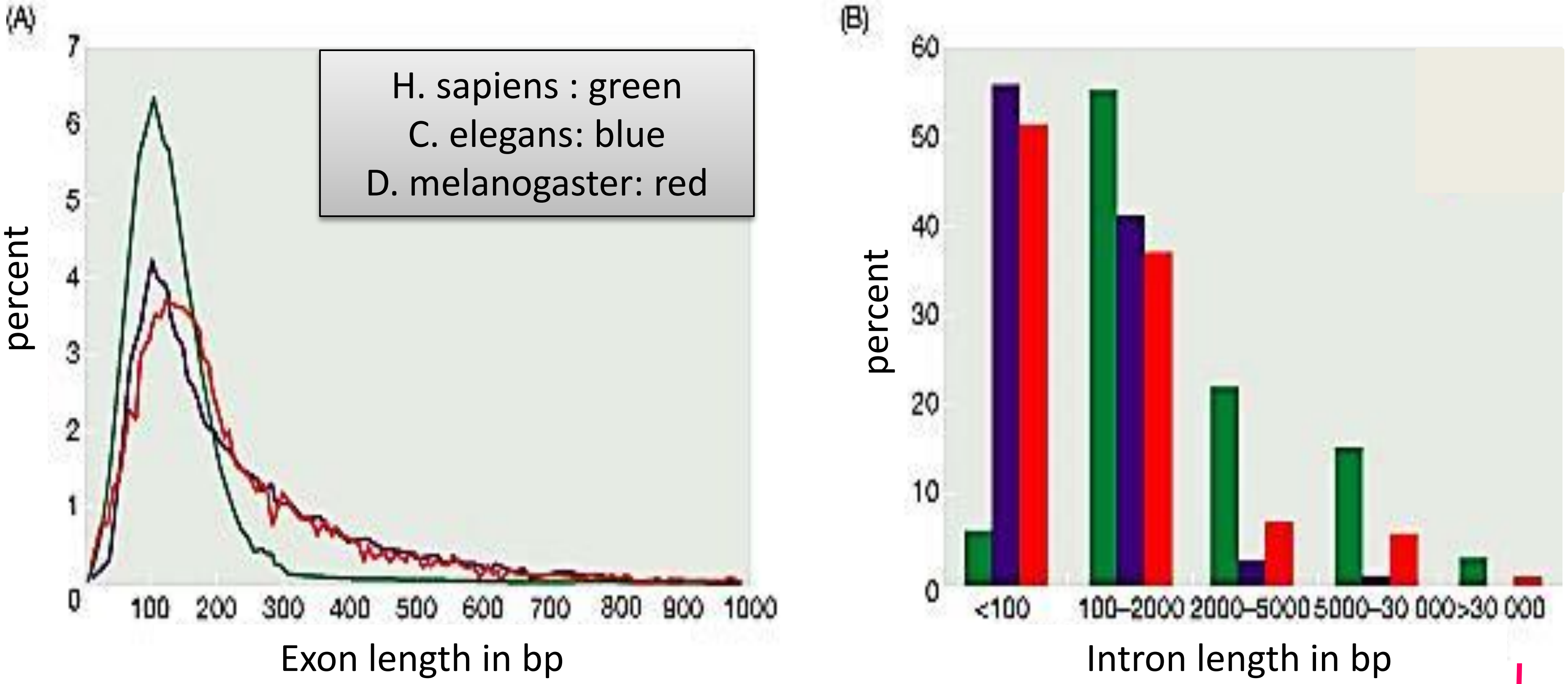

Average number of introns per gene
0
1
2
3
4
5
6
7
8
L. major
G. lamblia
T. vaginalis
C. merolae
E. cuniculi
G. theta NM
C. parvum
S. cerevisiae
C. albicans
P. aurelia
S. pombe
P. falciparum
P. yoelii
D. discoideum
T. pseudoonana
N. crassa
A. nidulans
A. gambiae
D. melanogaster
B. natans NM
C. briggsae
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

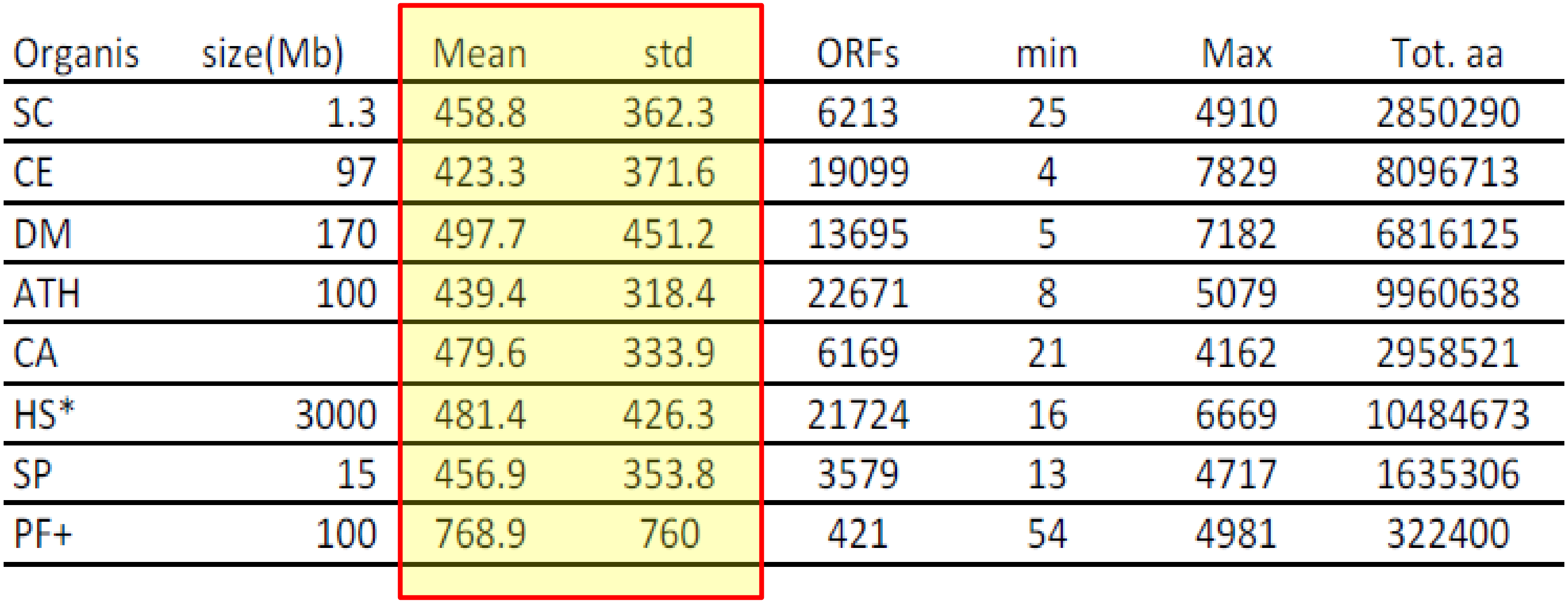


one intron in the human neurexin gene is approx. 480,000 nt !

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.

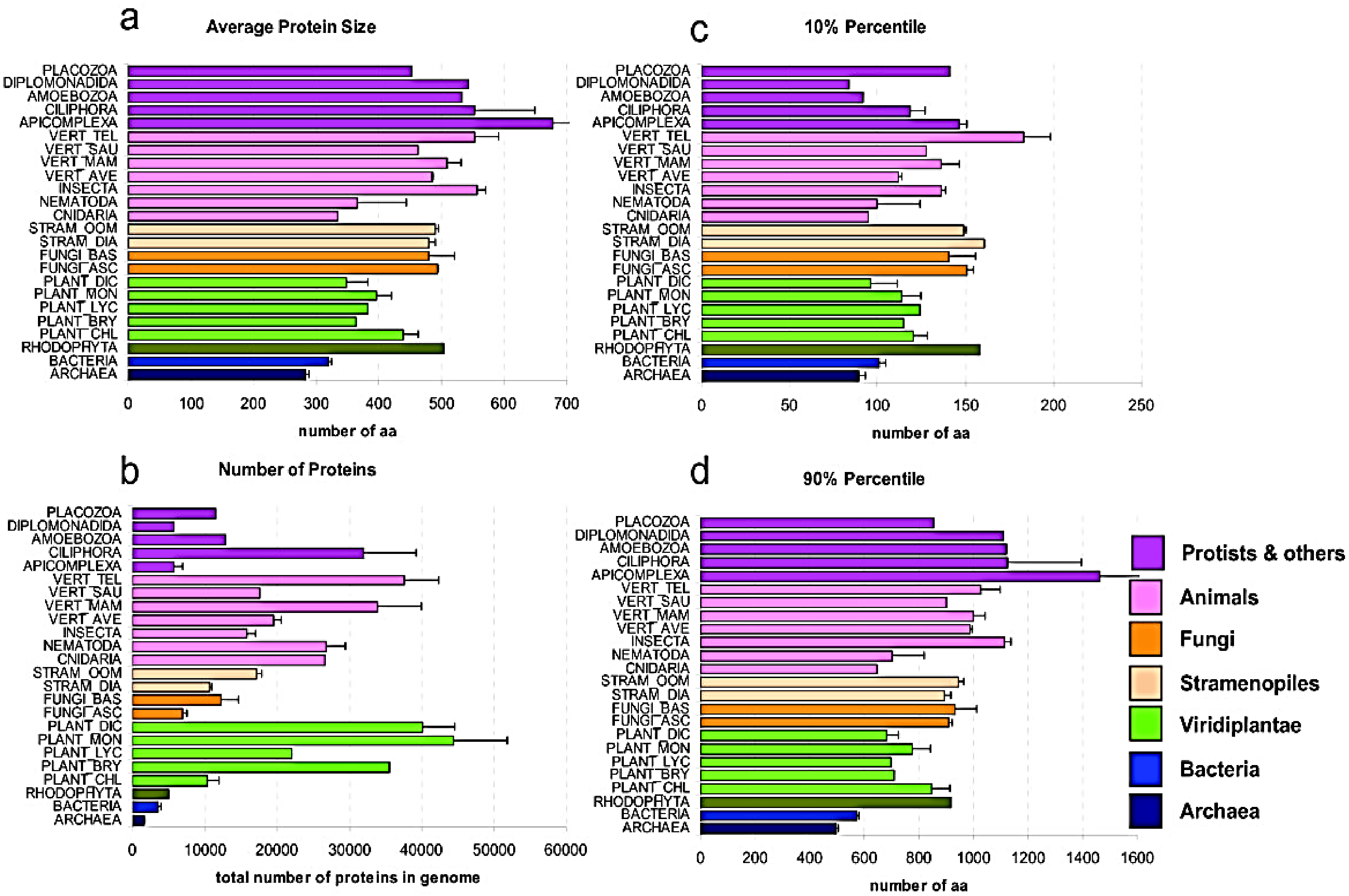
Note that the absolute number of genes does not follow organism complexity.

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

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)

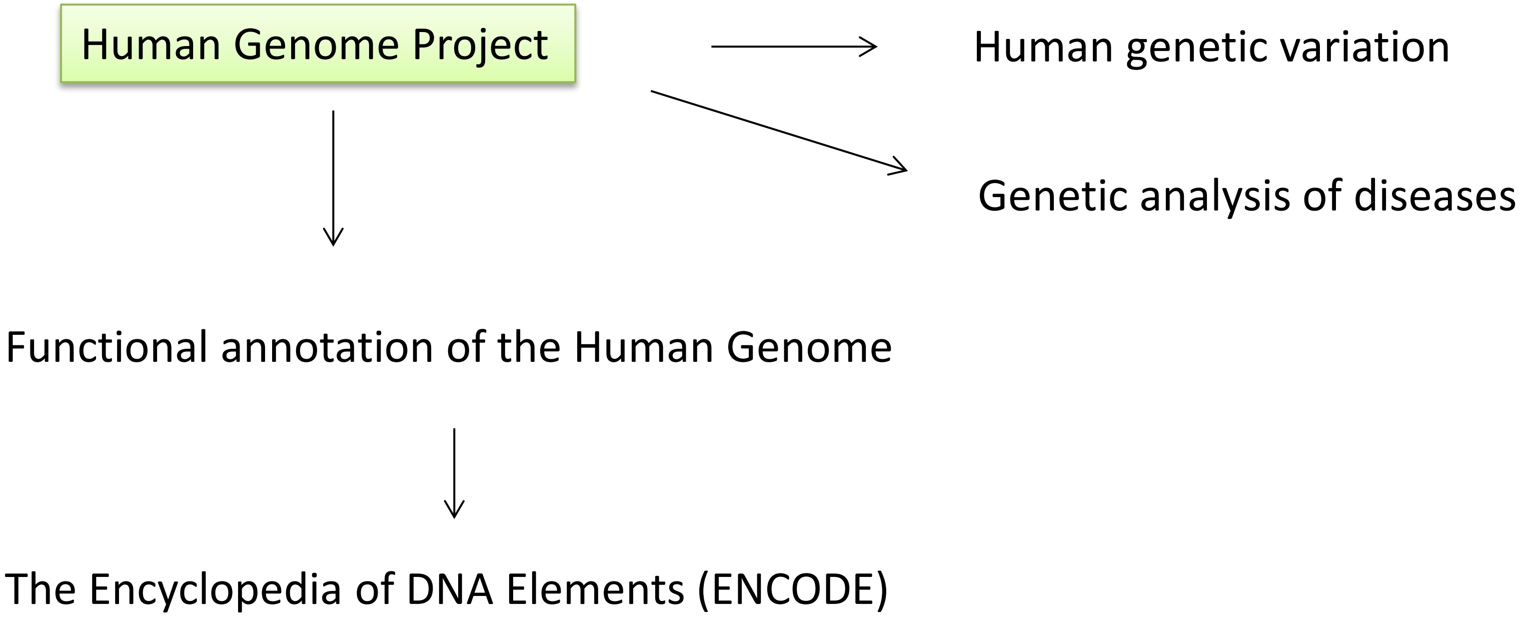
Human Genetic Variation (1000 Human Genomes - HapMap)

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

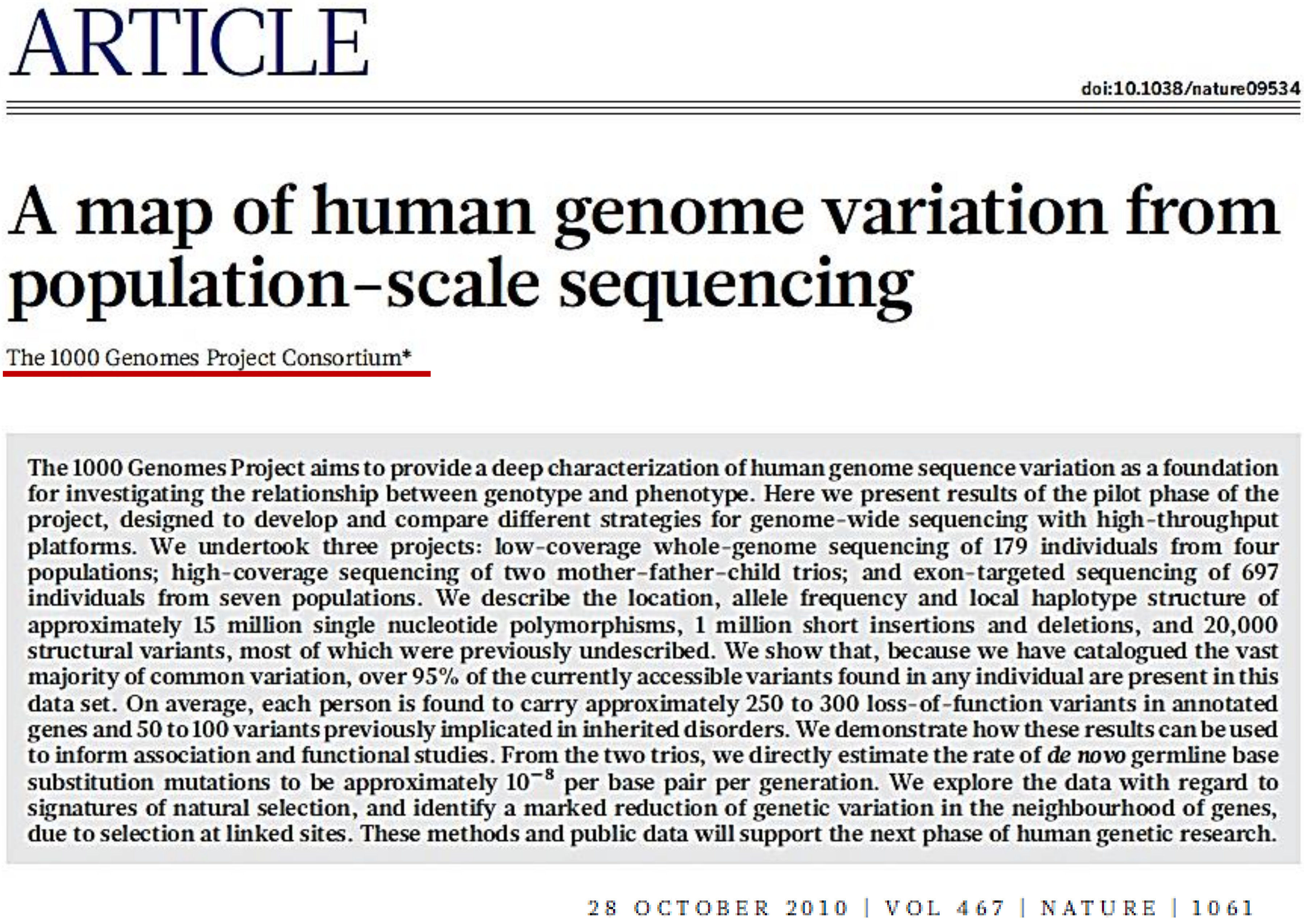
CNV – copy number variations

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)



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

16 | VOL.5 NO.1 | JANUARY 2008 | NATURE METHODS

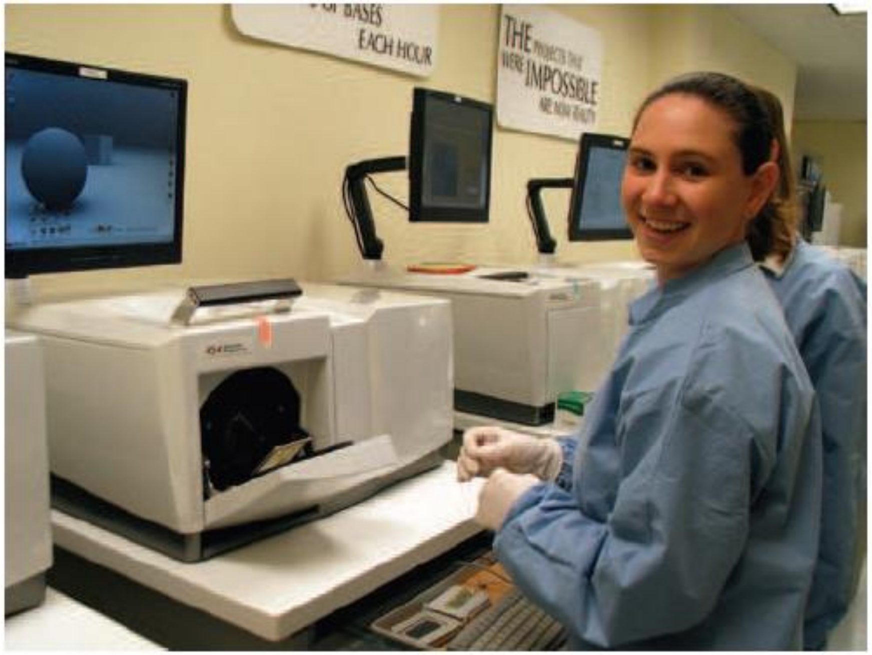
Next-generation sequencing transforms today's biology

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

High-parallel sequencing of fragments

No cloning step required

Next Generation Sequencing

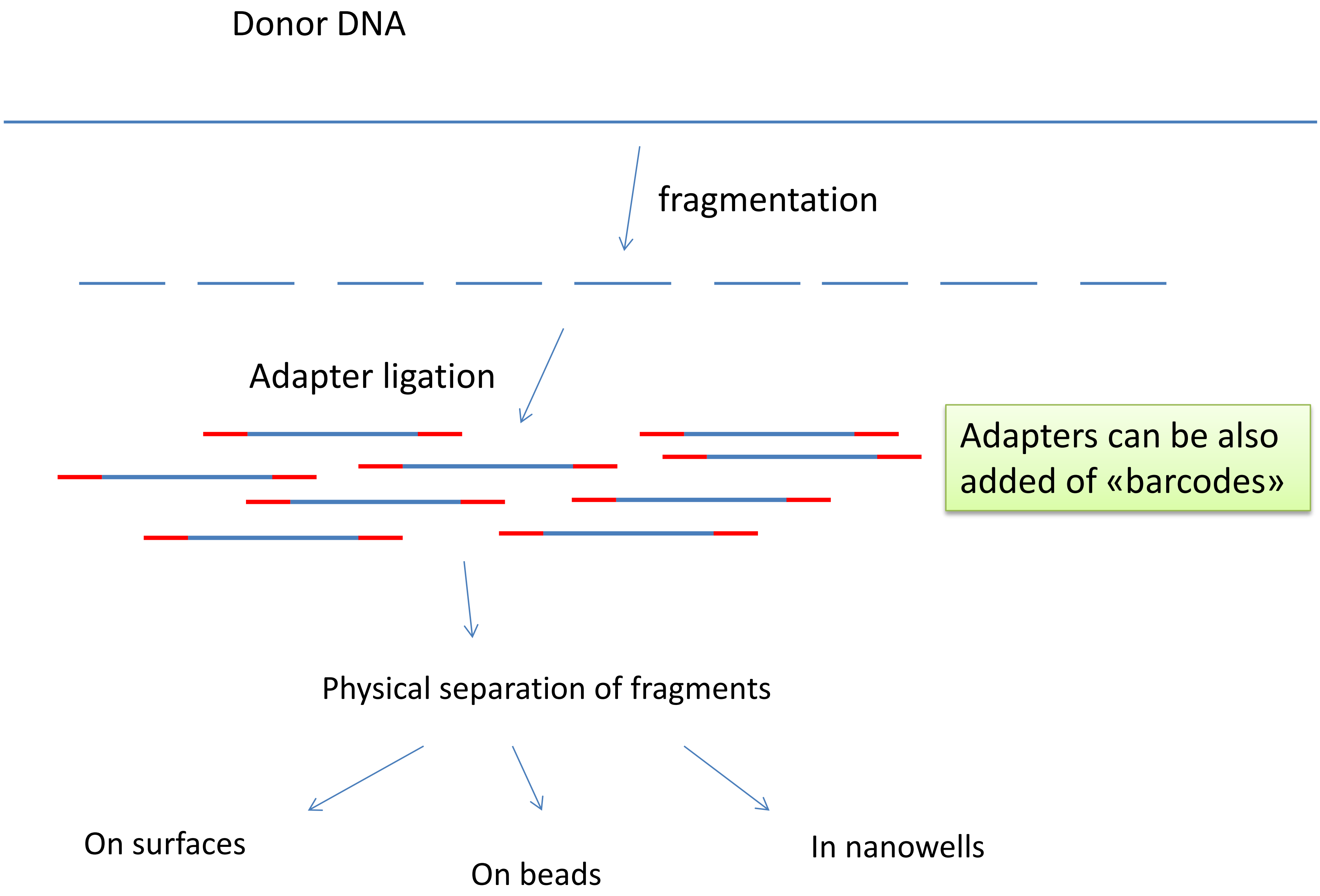
(deep-sequencing / mass sequencing)

✓generation of "DNA-nanoclones" on distinct solid surfaces by PCR or single-molecule isolation

✓highly parallel in situ sequencing

✓record read-out i.e. millions or short sequences ("**reads**")

✓align reads on genomes or assembly



Donor DNA

fragmentation

Adapter ligation

Adapters can be also added of «barcodes»

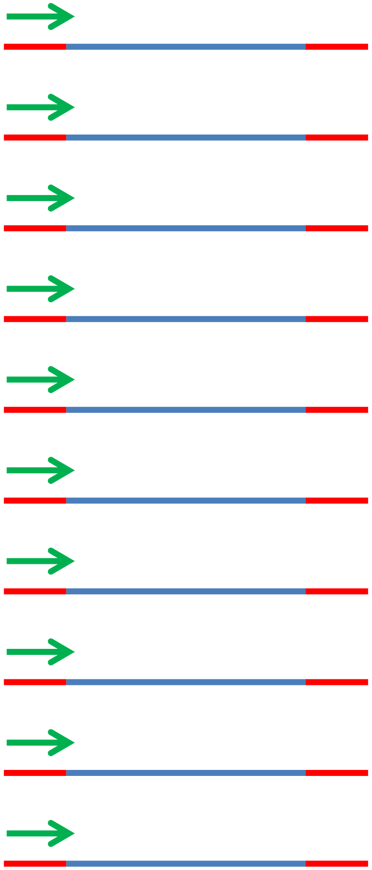
Physical separation of fragments

On surfaces

On beads

In nanowells

Adapter-primed sequencing by synthesis in high-parallel mode



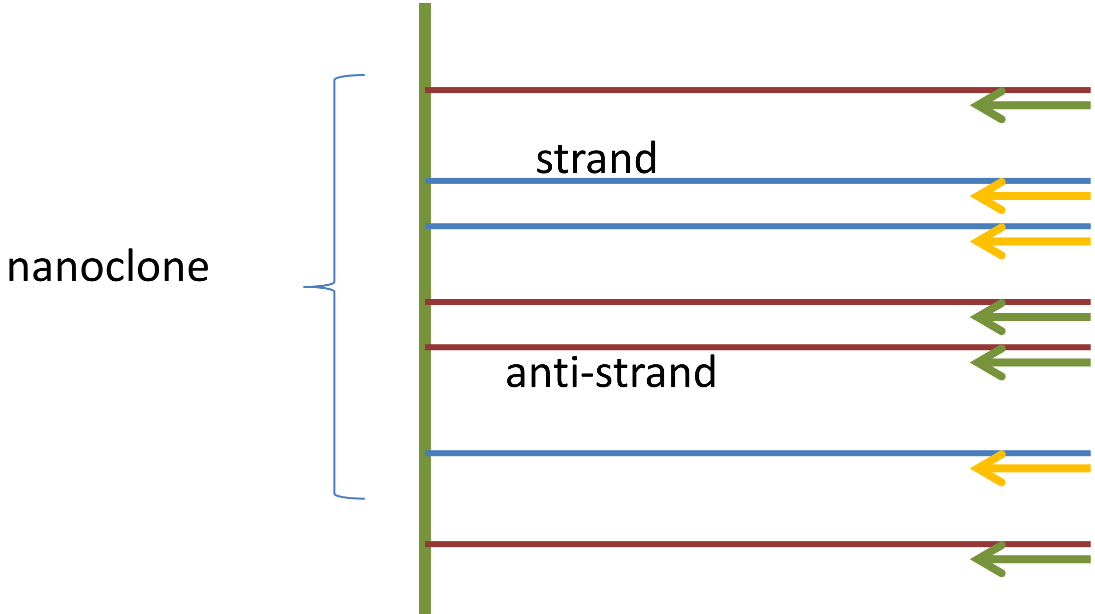
Millions of sequence reads
In parallel

Read length is defined by the number of synthesis cycles you perform

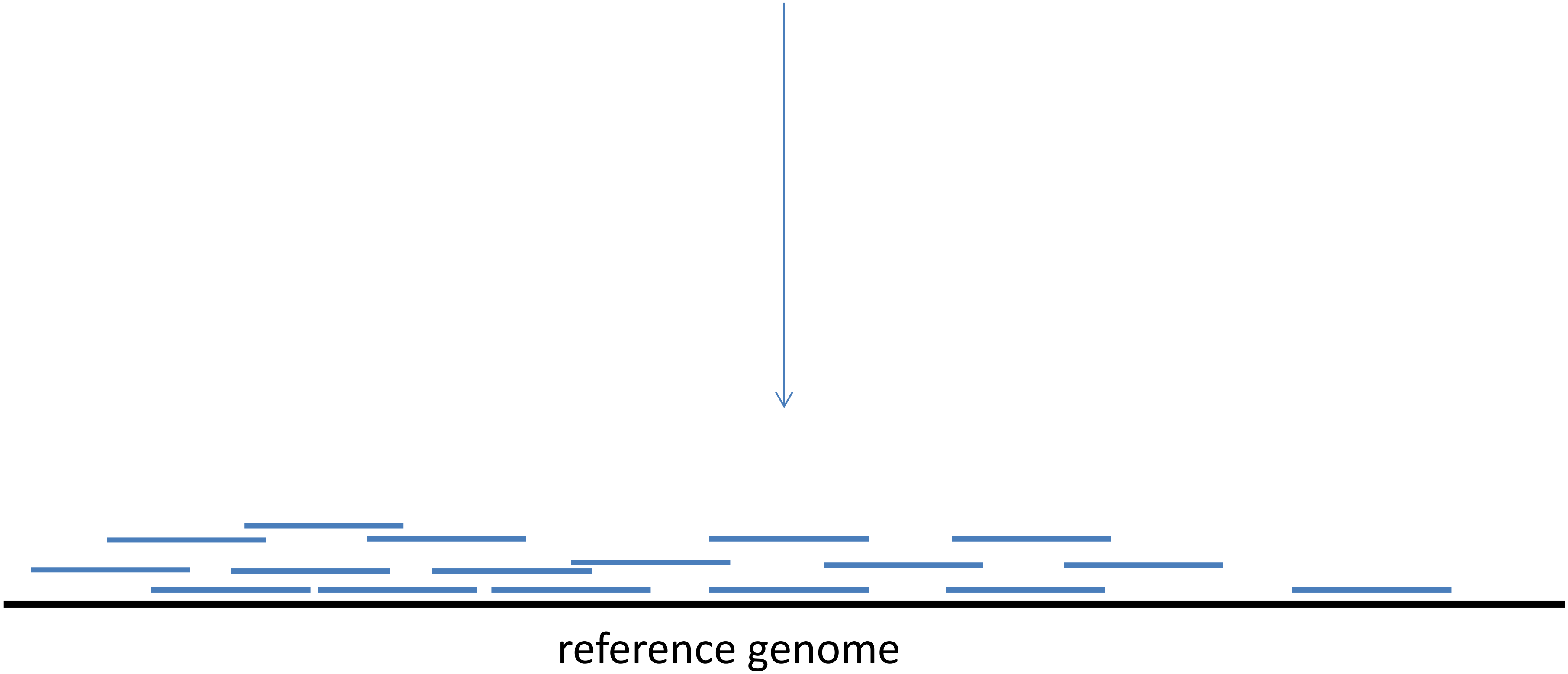
To increase information and reducing costs, often the Paired-end (PE) method is used

This means that we run sequencing using one primer, then everything is washed away and sequencing from the opposite primer is performed.

Since the length of fragments in the library can be controlled, we know that the two sequences should be «paired» i.e. Close to each other in the reference genome.



Reads are mapped to the 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

Next generation sequencing methods:

Number of molecules per sequence

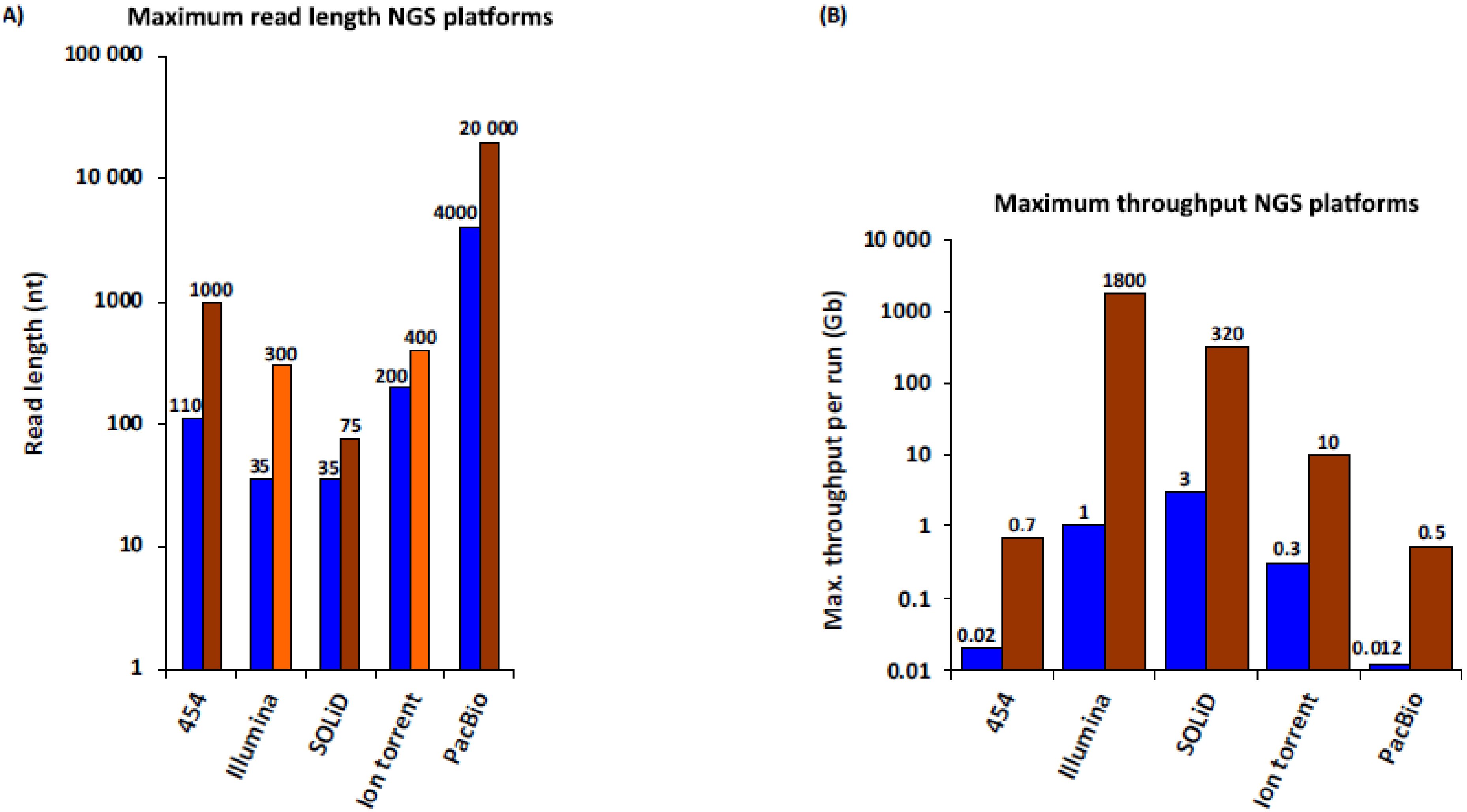
- Amplification
- Single-molecule

Biochemical measurement

- Sequencing by synthesis (Sanger is synthesis + termination)
    - Nucleotide chemistry
    - Associated chemistry
- Sequencing by annealing and ligation
- Sequencing by direct physico-chemical measurements

Detection

- Optical detection
- Ion or conductance detection



In blue the first version of the instruments

From Van Dijk et al., 2014 (Textbook)

NGS - mapping

«**Coverage**» (or depth of coverage or depth).

Definition: Number of reads (mappable) x read length / size of target genome

Example: we ran a NGS of 50 bp – length, obtaining 200*10^6 reads using whole human genomic DNA. 50*200*10^6 / 3.2*10^9 = 3.125

Is this a good coverage for human genome sequencing ?

We can also consider the «depth» as the number of times that a single bp of the target genome is represented in sequencing reads.

Due to random fragmentation and random efficiency of preparation steps, some parts can not be represented at all. Increasing the coverage depth will increase the probability that every base has been «confidently» sequenced.

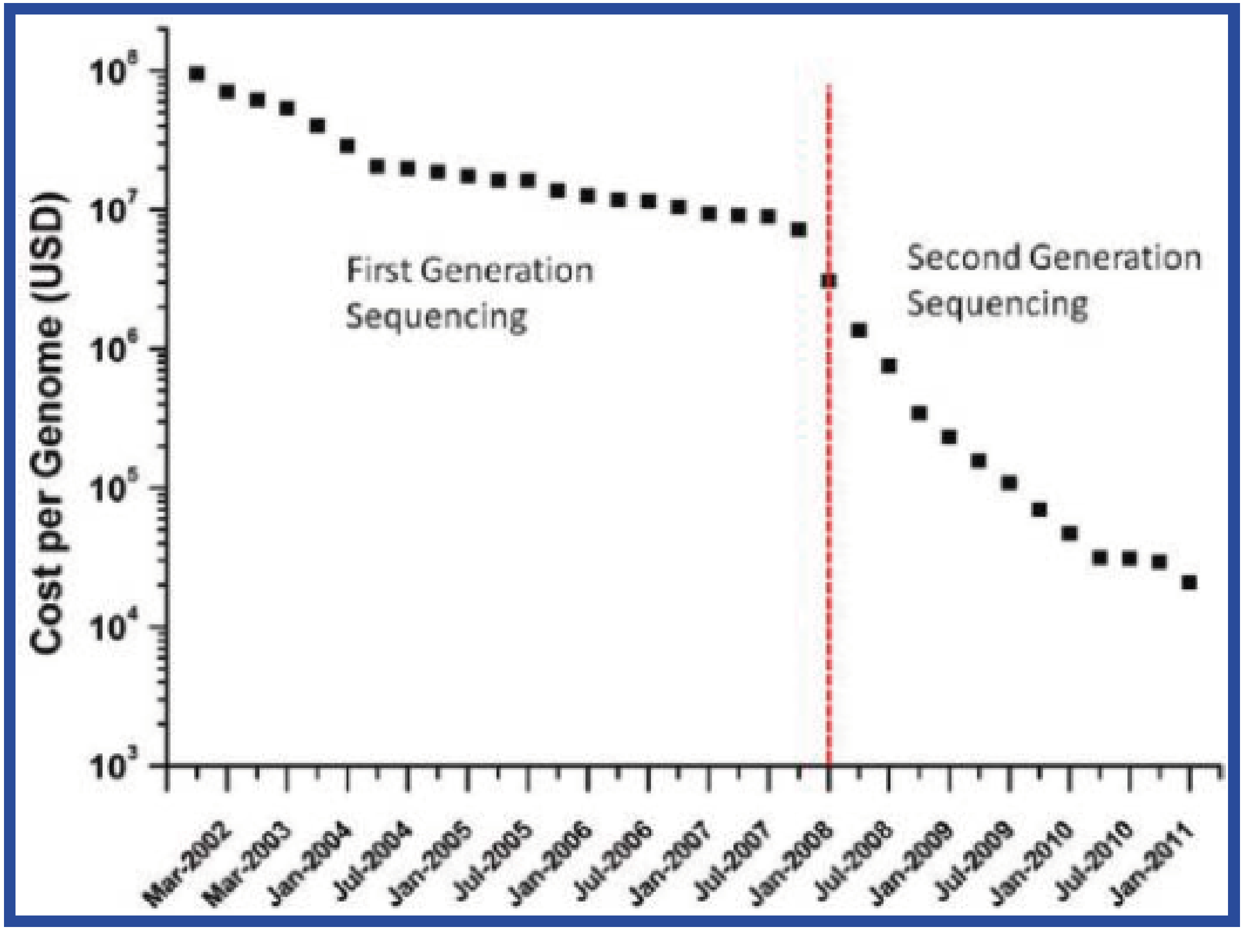


Figure 1. Estimated cost required to sequence a complete human genome based on data generated from NHGRI-funded large-scale DNA sequencing centers.[28]

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)

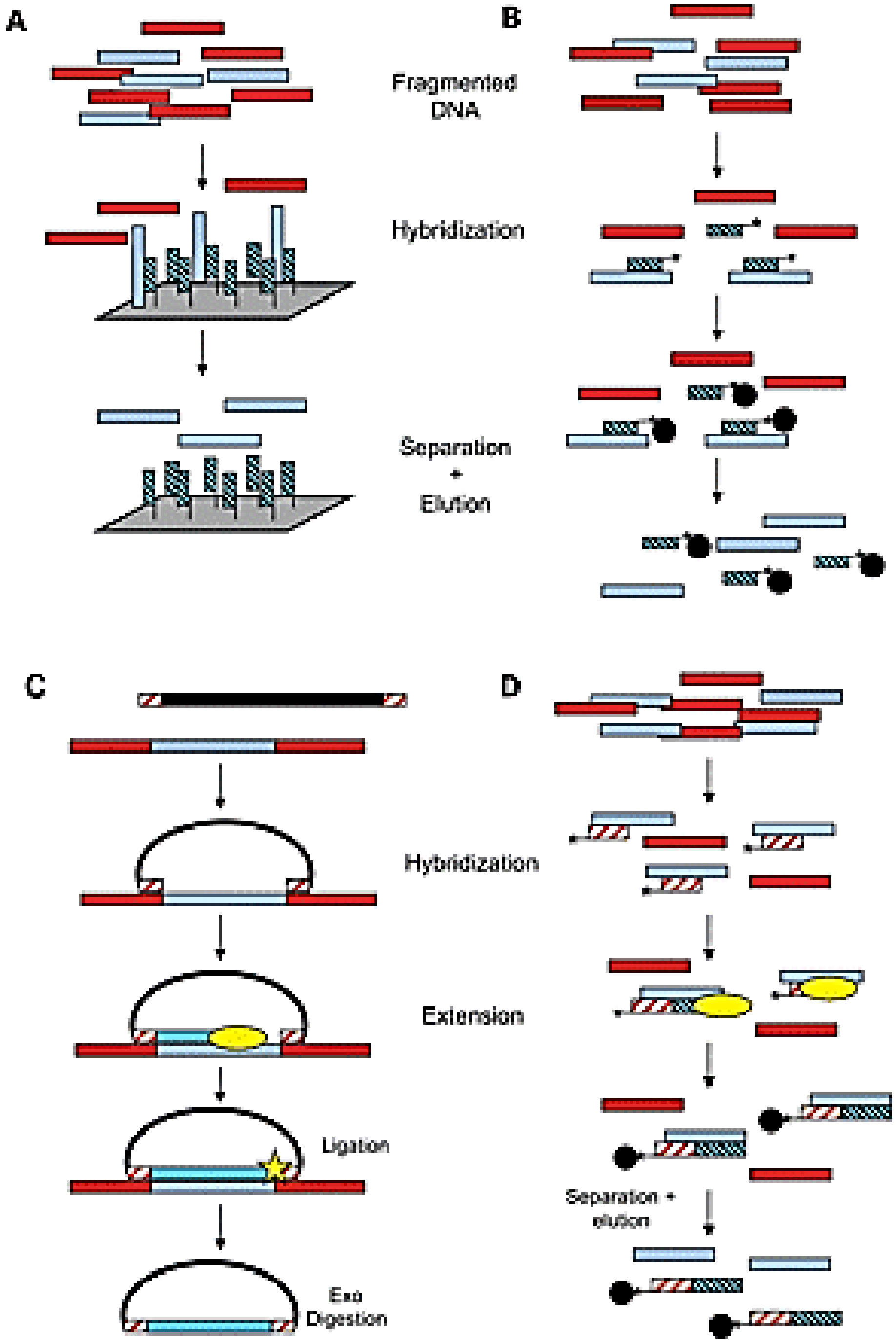
TCGA – The Cance Genome Atlas
Sequencing of tumor cell DNA to evidence mutations occurring in tumors.

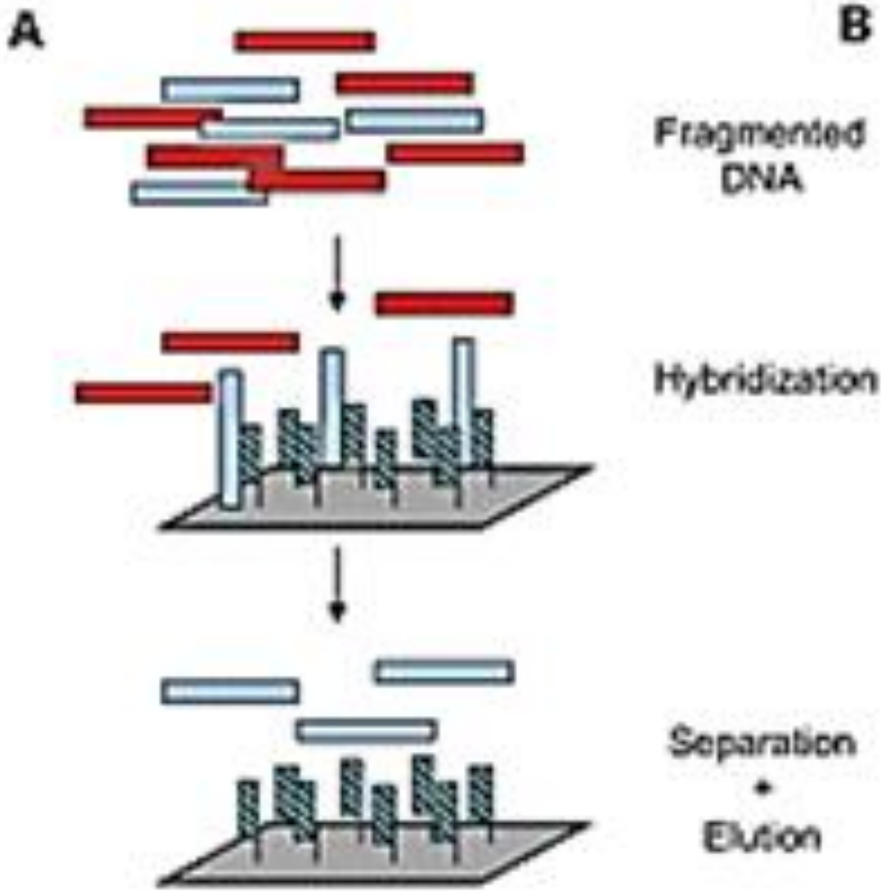
Exome sequencing

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

How can this all have anything to do with our story of Genomic Regulation ?

Variations in DNA sequence can affect regulation in many ways:

- TFBS can be altered
- Noncoding transcripts can be altered
- TSS, splicing sites, 3’ UTR variants
- Sites demarkating the border between chromatin domains
- Copy number alteration in regulatory regions
- Translocation of regulatory elements