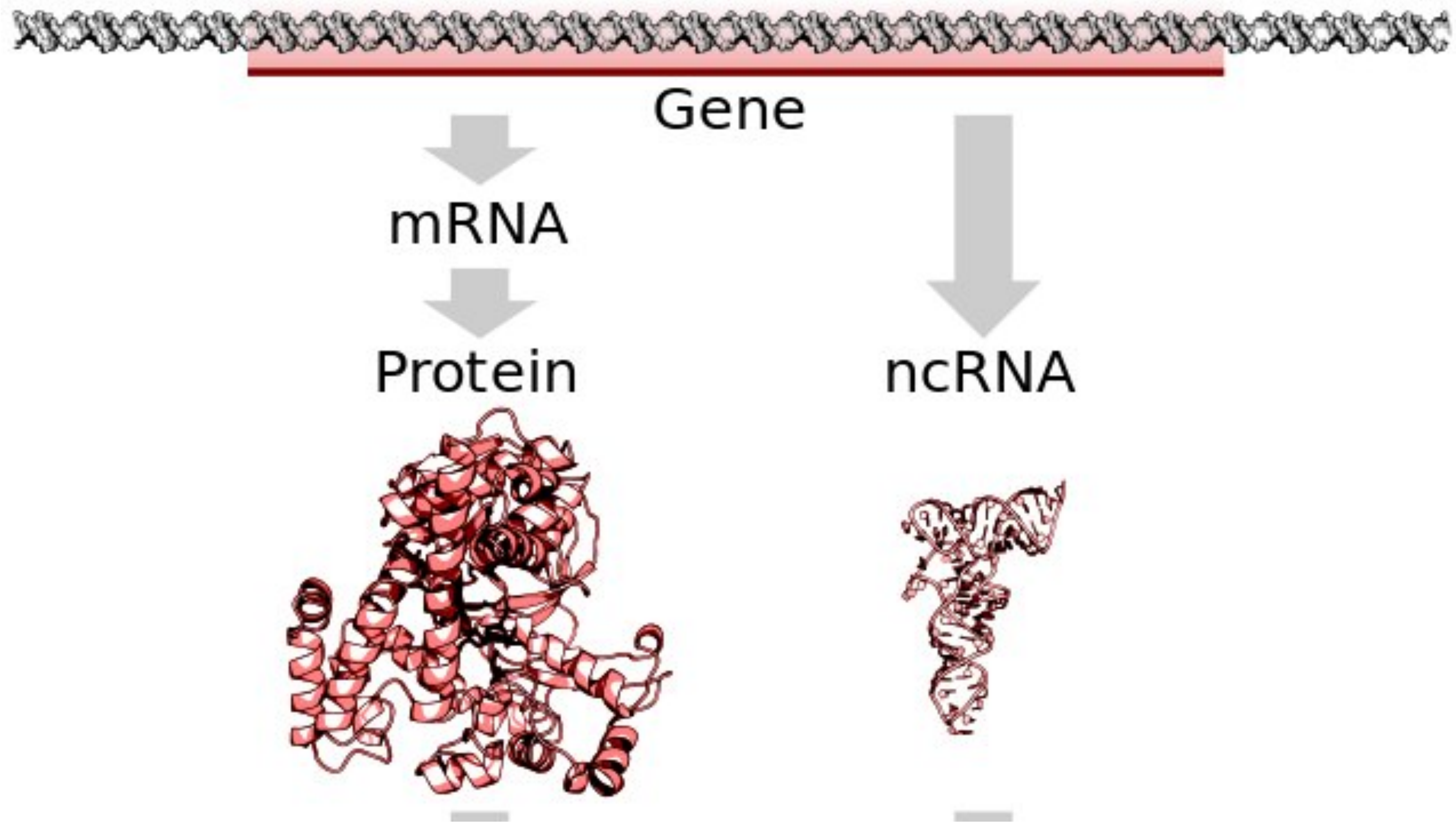
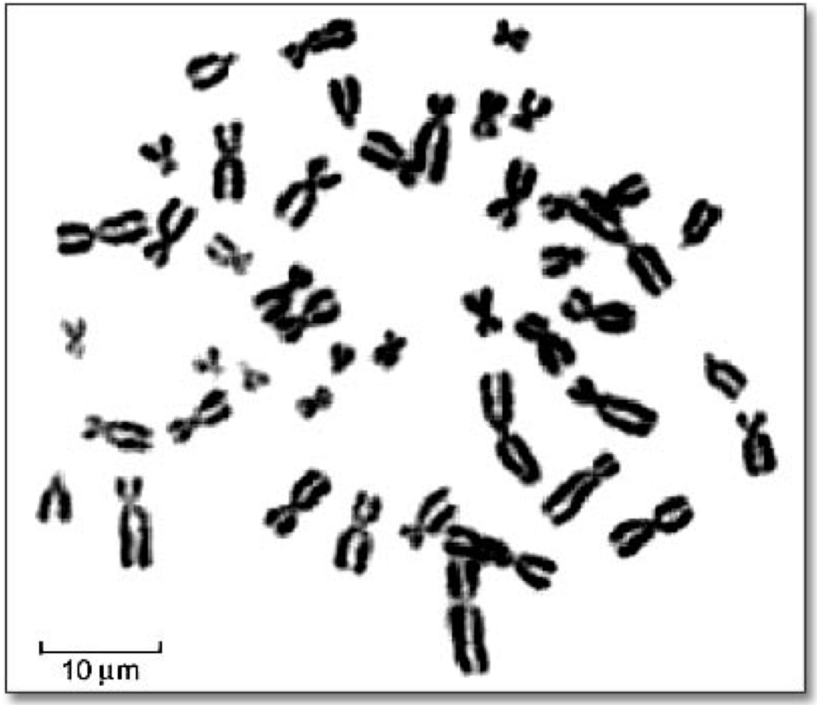
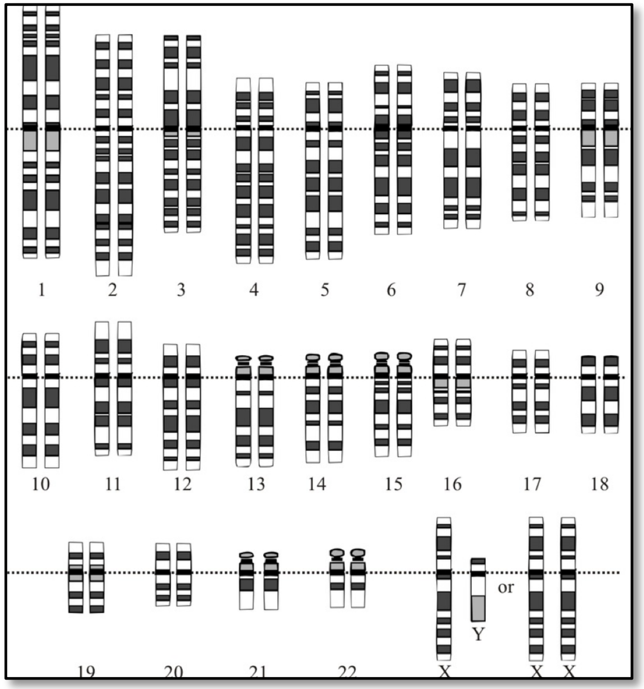
Non coding RNA Biology

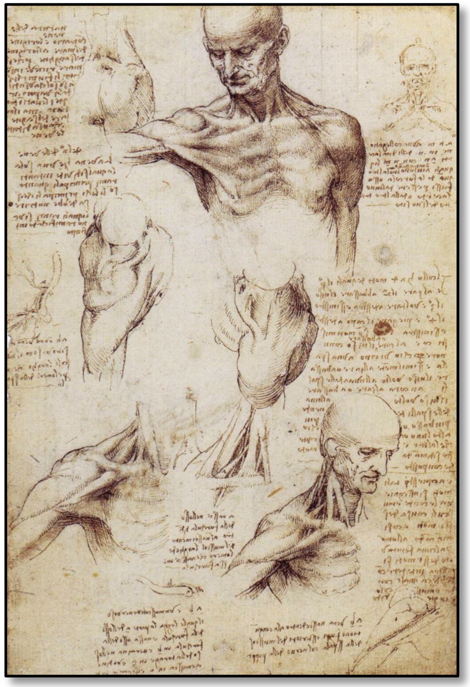
AA 2021/2022



The human genome is highly structured







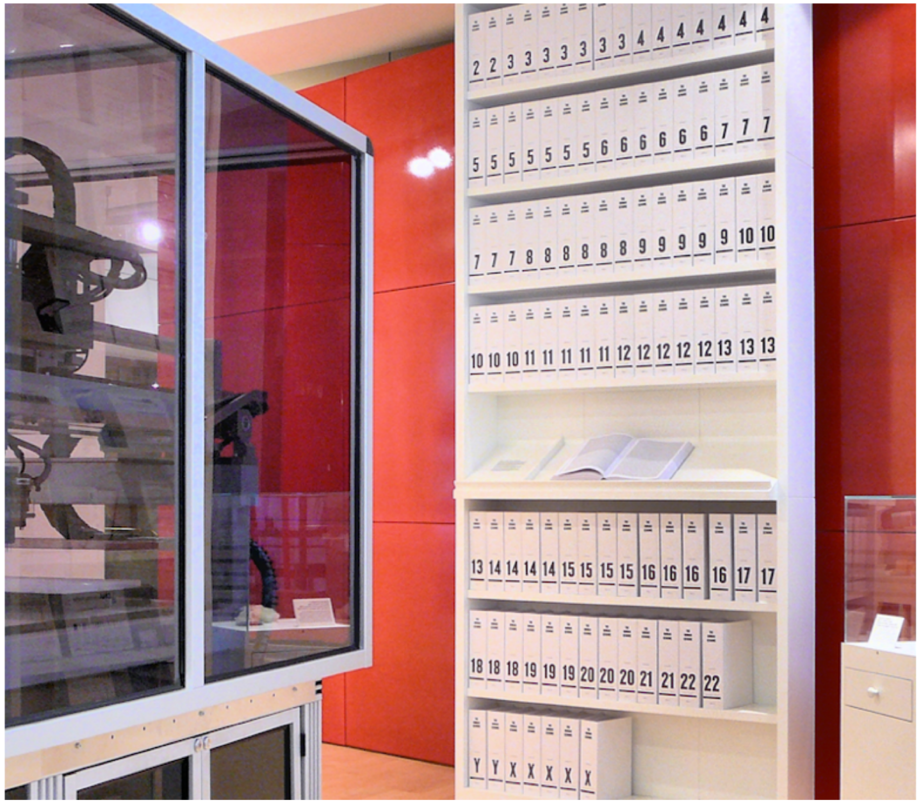
The human genome:

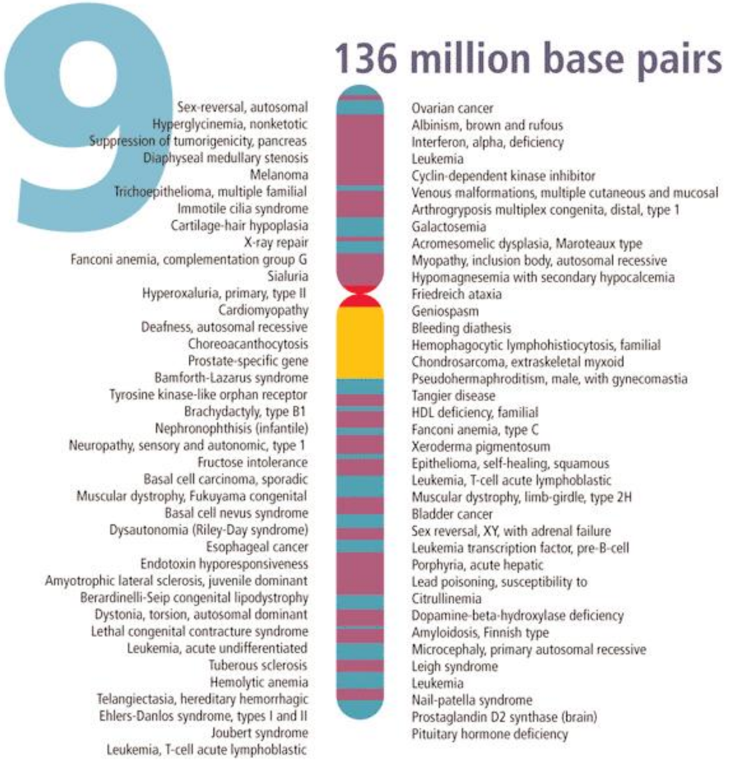
22 autosome paires

2   Sex chromosome pairs (XX o XY)

Total haploid genome $3x10^9$

The human genome is highly structured




9
136 million base pairs
Sex-reversal, autosomal
Hyperglycinemia, nonketotic
Suppression of tumorigenicity, pancreas
Diaphyseal medullary stenosis
Melanoma
Trichoepithelioma, multiple familial
Immotile cilia syndrome
Cartilage-hair hypoplasia
X-ray repair
Fanconi anemia, complementation group G
Sialuria
Hyperoxaluria, primary, type II
Cardiomyopathy
Deafness, autosomal recessive
Choreoacanthocytosis
Prostate-specific gene
Bamforth-Lazarus syndrome
Tyrosine kinase-like orphan receptor
Brachydactyly, type B1
Nephronophthisis (infantile)
Neuropathy, sensory and autonomic, type 1
Fructose intolerance
Basal cell carcinoma, sporadic
Muscular dystrophy, Fukuyama congenital
Basal cell nevus syndrome
Dysautonomia (Riley-Day syndrome)
Esophageal cancer
Endotoxin hyporesponsiveness
Amyotrophic lateral sclerosis, juvenile dominant
Berardinelli-Seip congenital lipodystrophy
Dystonia, torsion, autosomal dominant
Lethal congenital contracture syndrome
Leukemia, acute undifferentiated
Tuberous sclerosis
Hemolytic anemia
Telangiectasia, hereditary hemorrhagic
Ehlers-Danlos syndrome, types I and II
Joubert syndrome
Leukemia, T-cell acute lymphoblastic
Ovarian cancer
Albinism, brown and rufous
Interferon, alpha, deficiency
Leukemia
Cyclin-dependent kinase inhibitor
Venous malformations, multiple cutaneous and mucosal
Arthrogryposis multiplex congenita, distal, type 1
Galactosemia
Acromesomelic dysplasia, Maroteaux type
Myopathy, inclusion body, autosomal recessive
Hypomagnesemia with secondary hypocalcemia
Friedreich ataxia
Geniospasm
Bleeding diathesis
Hemophagocytic lymphohistiocytosis, familial
Chondrosarcoma, extraskeletal myxoid
Pseudohermaphroditism, male, with gynecomastia
Tangier disease
HDL deficiency, familial
Fanconi anemia, type C
Xeroderma pigmentosum
Epithelioma, self-healing, squamous
Leukemia, T-cell acute lymphoblastic
Muscular dystrophy, limb-girdle, type 2H
Bladder cancer
Sex reversal, XY, with adrenal failure
Leukemia transcription factor, pre-B-cell
Porphyria, acute hepatic
Lead poisoning, susceptibility to
Citrullinemia
Dopamine-beta-hydroxylase deficiency
Amyloidosis, Finnish type
Microcephaly, primary autosomal recessive
Leigh syndrome
Leukemia
Nail-patella syndrome
Prostaglandin D2 synthase (brain)
Pituitary hormone deficiency

Haploid human genome: 3.2 x 10^9 bp (3200000000 bp)

→ 22 autosomes
→ 2 sex chromosomes (X ed Y)
→ 19797 protein coding genes (ca 20.000)

Chromosome dimensions: 45-275 Mb;
→ 3,2 x 10^9 bp: haploid chromosome set

Usage of genetic information:

5.000-10.000 geni espressi da ogni cellula
** 100.000 different proteins (post- translational modifactions per cell)**
** 10^8 total protein spcecies**

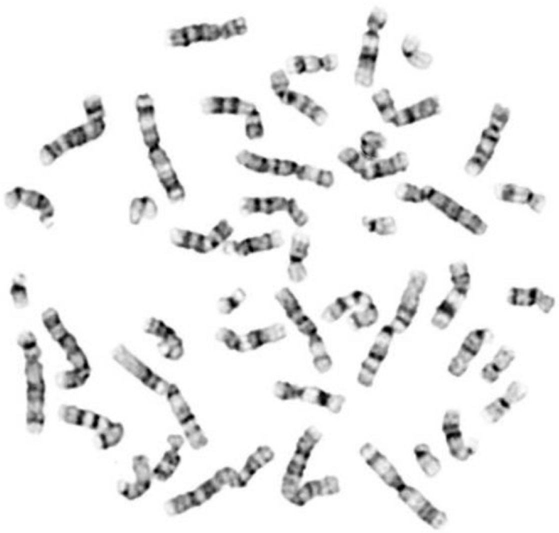
ENORMOUSE COMPLEXITY

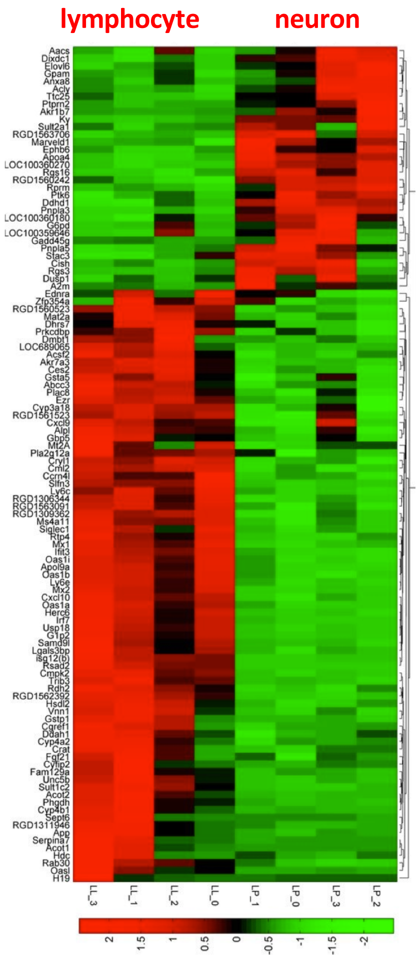
The human genome encodes information that underlies cell specification in multi-cellular organisms

GENOMA
coding and
non-coding genes

Specific gene expression programs

Cell function





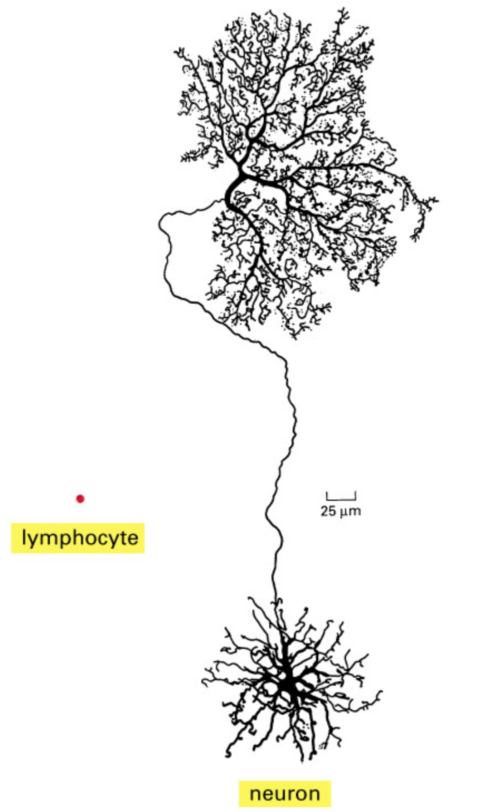


Figure 7–1. Molecular Biology of the Cell, 4th Edition.

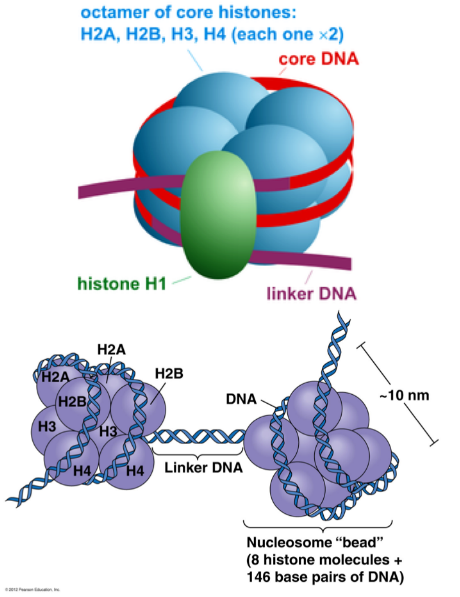
Genetic information must be highly organized

The human genome is highly structured

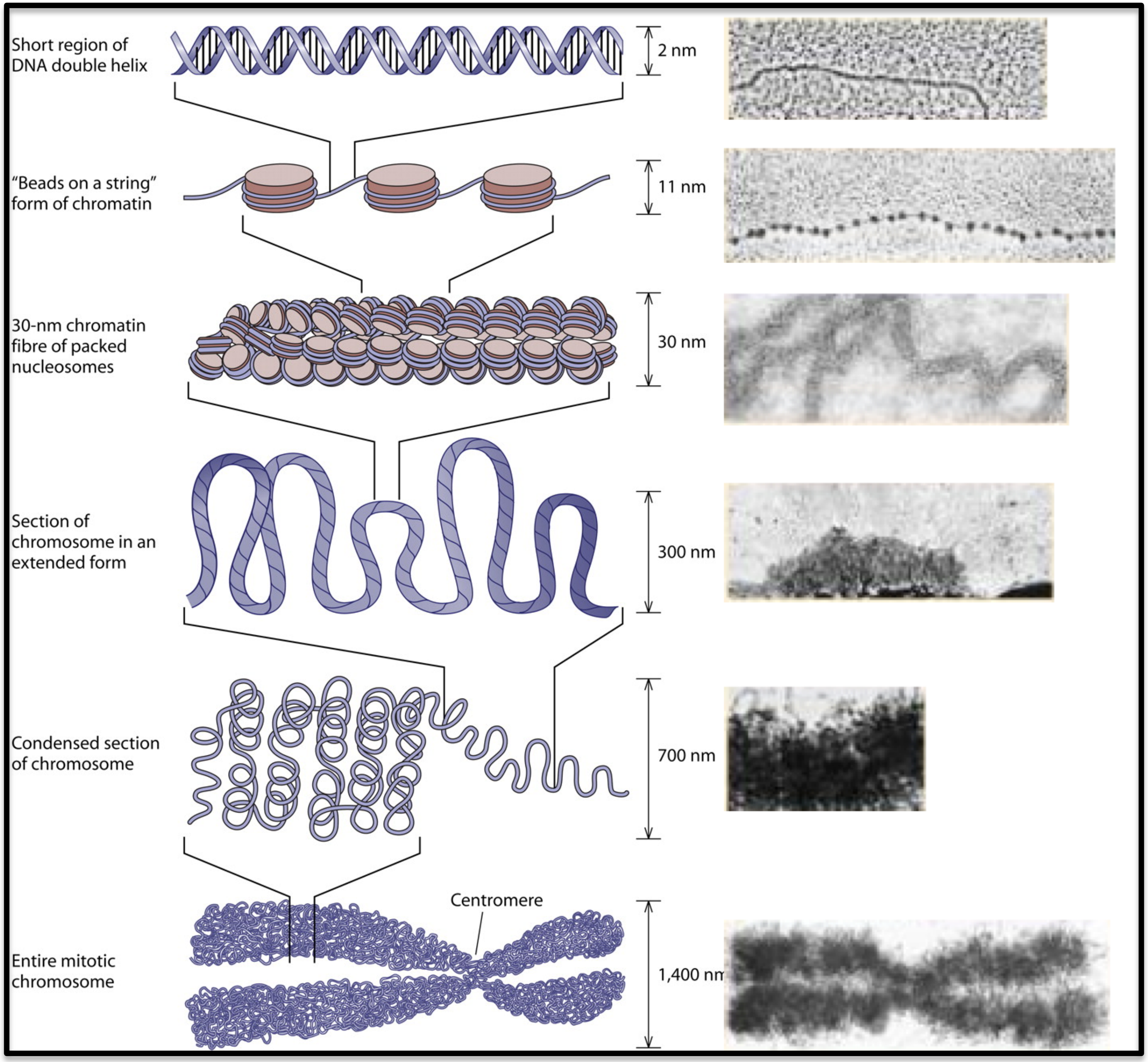
Chromatin: DNA + protein in nucleus
Organisation of genetic information

Function:

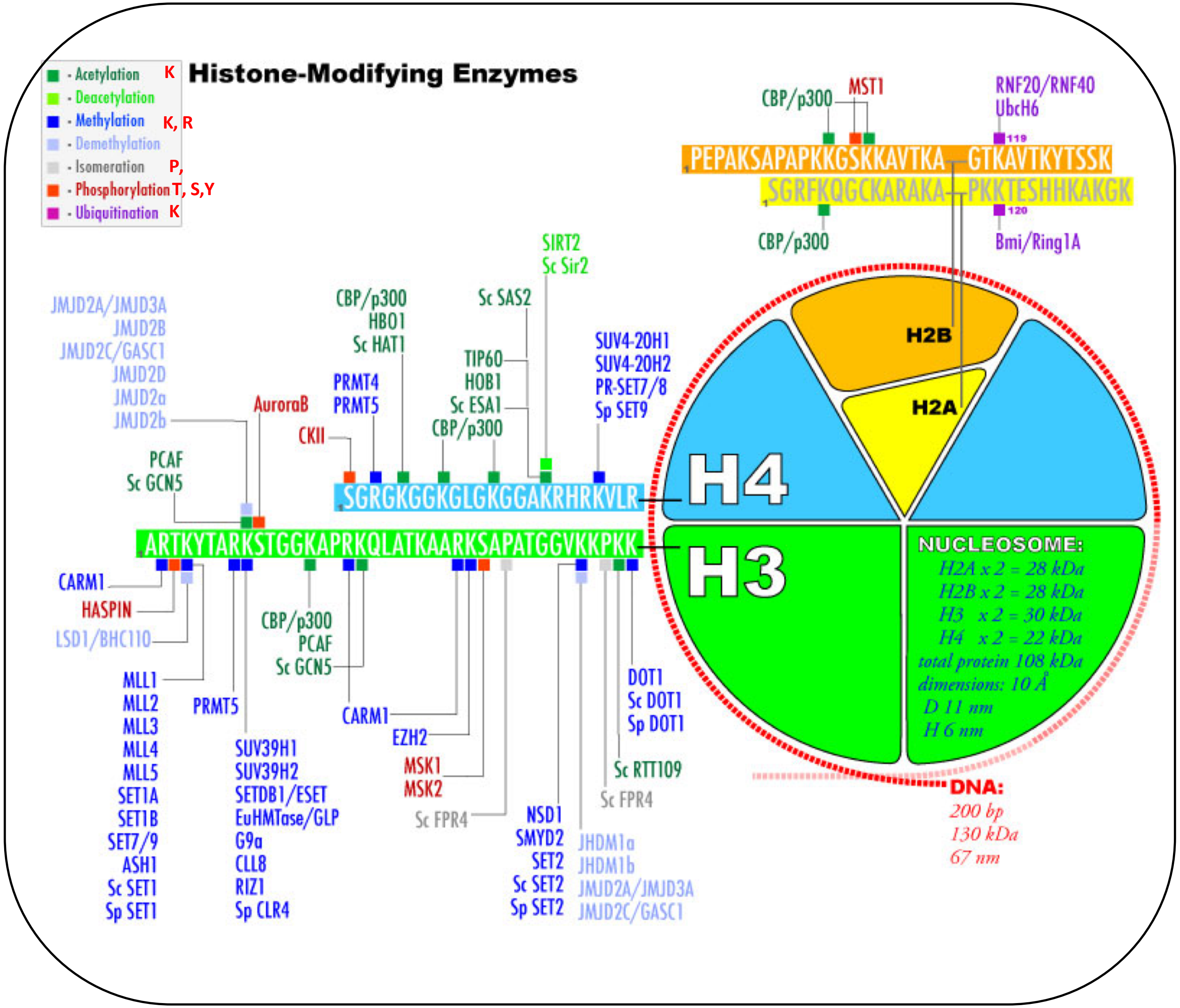
Packaging of DNA
Compaction of DNA
Definition of reagions of gene
Expression (euchromatin) or repression (heterochromatin)
-Increasing stability of DNA
-Prevention of damage
-Control of replication, gene expression
-Cell cycle



Nucleosome:
8 histone proteins
2 turns of DNA(146 nt)


Short region of DNA double helix
2 nm
"Beads on a string" form of chromatin
11 nm
30-nm chromatin fibre of packed nucleosomes
30 nm
Section of chromosome in an extended form
300 nm
Condensed section of chromosome
700 nm
Entire mitotic chromosome
Centromere
1,400 nm

POST-TRANSLATIONAL HISTONE MODIFICATIONS



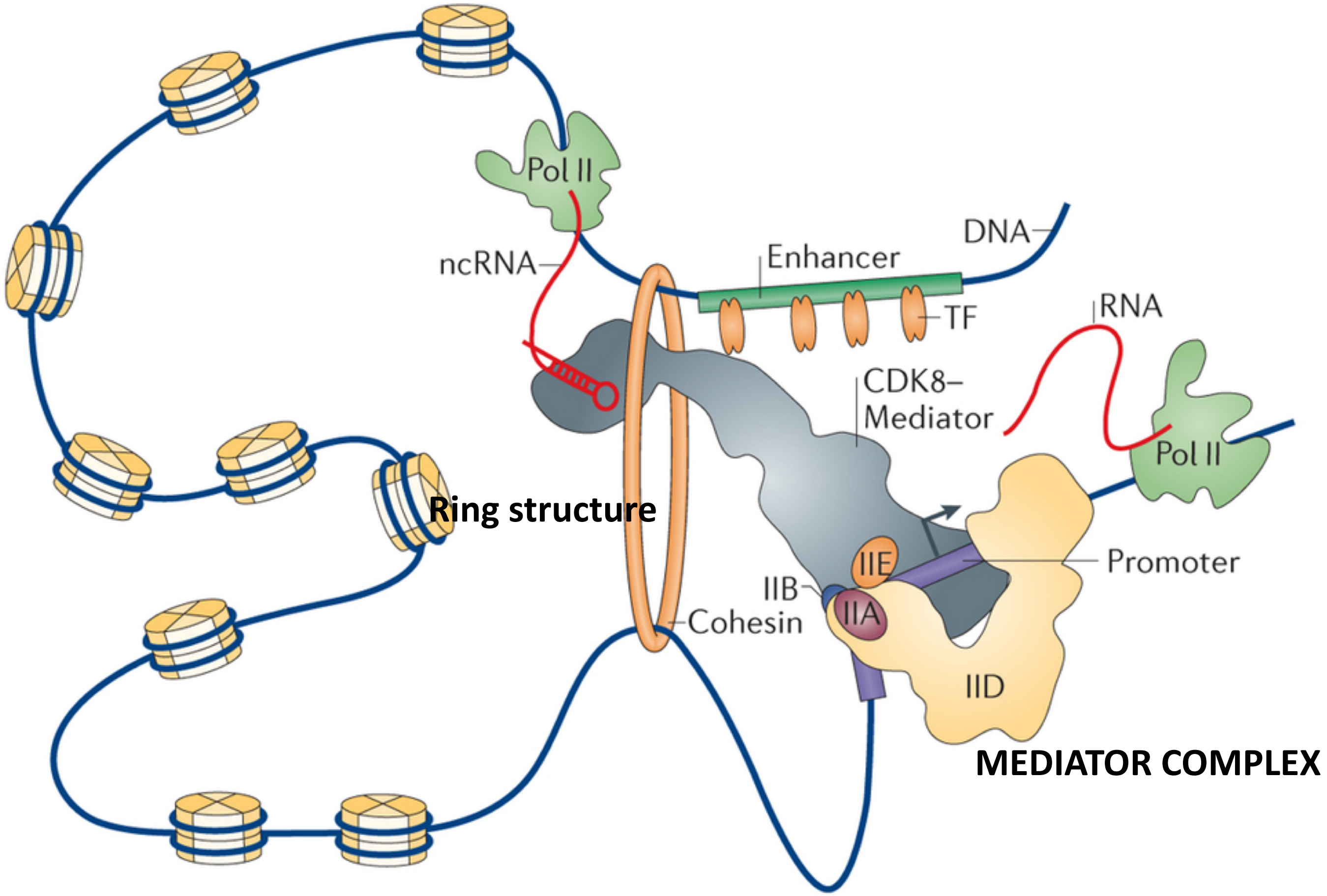
Gene expression
Control by post-translational
histone modifications

→Activate transcription
(H3K9 acetylation, …)
→Repress transcription
(H3K27 trimethylation)
can be cell type specific

**Sum of all modifications
= HISTONE CODE**

Specific histone
+modifications at promoters
Enhancers, along active
Genes, site of termination

The human genome is highly structured



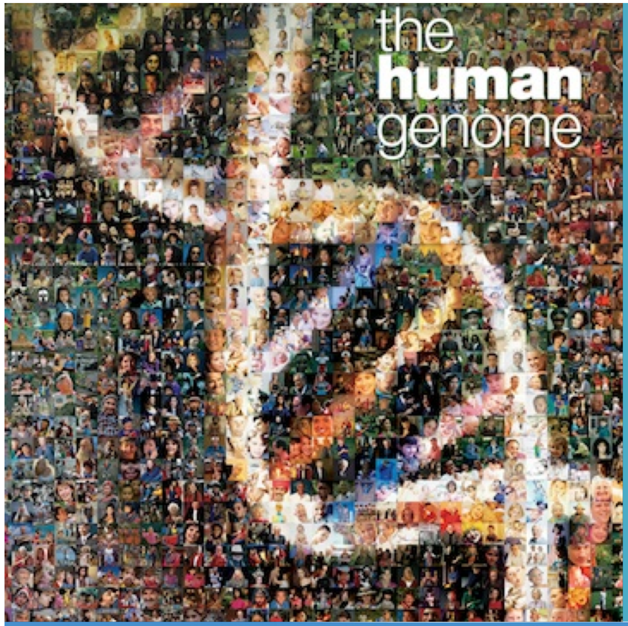
Nature Reviews | Molecular Cell Biology

Specific transcription factors can bind promoters and enhancers

RNAs can support the use enhancers

Enhancers are brought In vicinity to promoters and other gene regulatory Elements

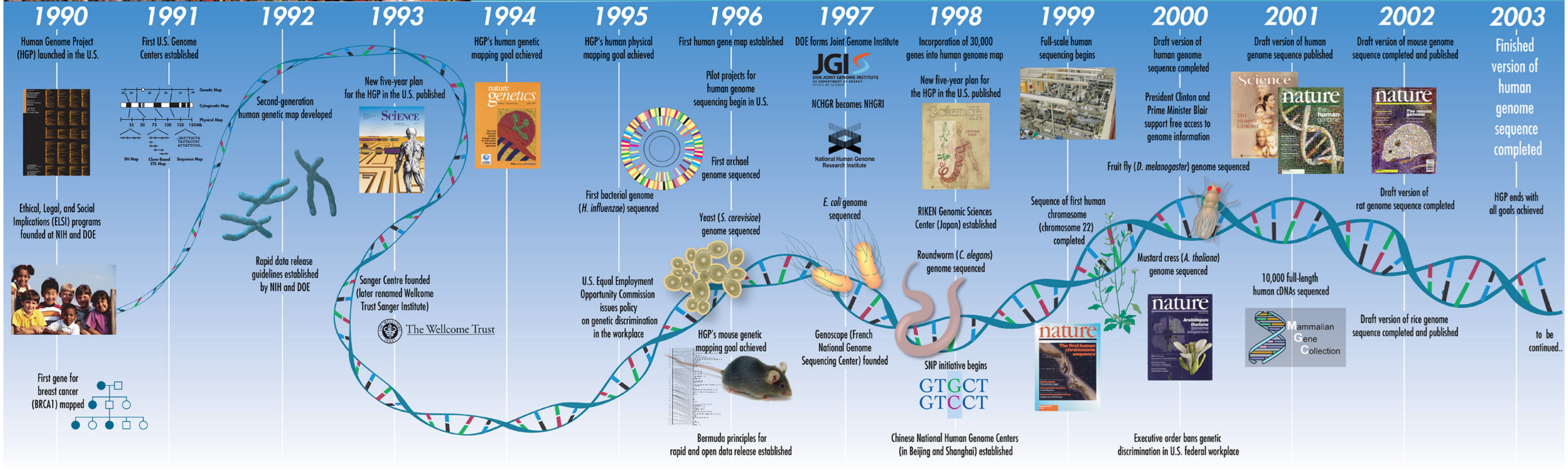
→ SPECIFIC 3 DIMENTSIONAL STRUCTURE



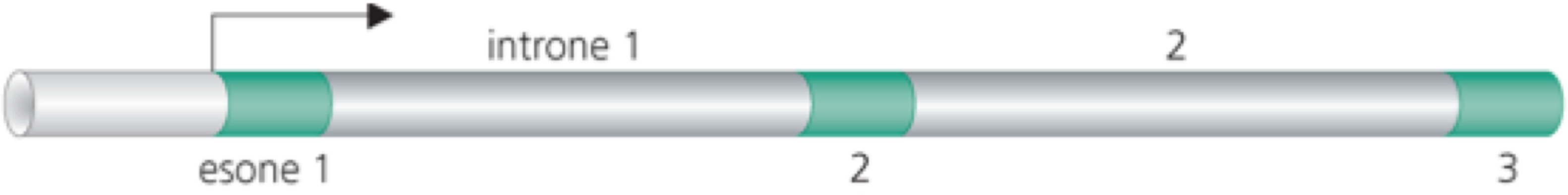
THE GENOME OF MANY ORGANSIMS IS ALREADY SEQUENCED

THE HUMAN GENOME PROJECT

SEQEUNCING GENOMIC DNA


1990
Human Genome Project (HGP) launched in the U.S.
Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE
First gene for breast cancer (BRCA1) mapped
1991
First U.S. Genome Centers established
Genetic Map
Cytogenetic Map
Physical Map
25 50 75 100 125 150 Mb
...GATCTGCTA TACTACCGC ATTATTCCG...
RH Map
Clone-Based STS Map
Sequence Map
1992
Second-generation human genetic map developed
Rapid data release guidelines established by NIH and DOE
1993
New five-year plan for the HGP in the U.S. published
SCIENCE
Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)
The Wellcome Trust
1994
HGP's human genetic mapping goal achieved
nature genetics
1995
HGP's human physical mapping goal achieved
First bacterial genome (H. influenzae) sequenced
U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace
1996
First human gene map established
Pilot projects for human genome sequencing begin in U.S.
First archael genome sequenced
Yeast (S. cerevisiae) genome sequenced
HGP's mouse genetic mapping goal achieved
Bermuda principles for rapid and open data release established
1997
DOE forms Joint Genome Institute
JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE
NCHGR becomes NHGRI
National Human Genome Research Institute
E. coli genome sequenced
Genoscope (French National Genome Sequencing Center) founded
1998
Incorporation of 30,000 genes into human genome map
New five-year plan for the HGP in the U.S. published
Science
RIKEN Genomic Sciences Center (Japan) established
Roundworm (C. elegans) genome sequenced
SNP initiative begins
GTGCT
GTCCT
Chinese National Human Genome Centers (in Beijing and Shanghai) established
1999
Full-scale human sequencing begins
Sequence of first human chromosome (chromosome 22) completed
nature
The first human chromosome sequence
2000
Draft version of human genome sequence completed
President Clinton and Prime Minister Blair support free access to genome information
Fruit fly (D. melanogaster) genome sequenced
Mustard cress (A. thaliana) genome sequenced
nature
Arabidopsis thaliana genome sequence
Executive order bans genetic discrimination in U.S. federal workplace
2001
Draft version of human genome sequence published
Science
THE HUMAN GENOME
nature
the human genome
10,000 full-length human cDNAs sequenced
Mammalian Gene Collection
2002
Draft version of mouse genome sequence completed and published
nature
The mouse genome
Draft version of rat genome sequence completed
Draft version of rice genome sequence completed and published
2003
Finished version of human genome sequence completed
HGP ends with all goals achieved
to be continued..

ISOLATE LARGE PIECES OF DNA AND SEQEUNCE!



Dideoxy (Sanger) sequencing

Principle:

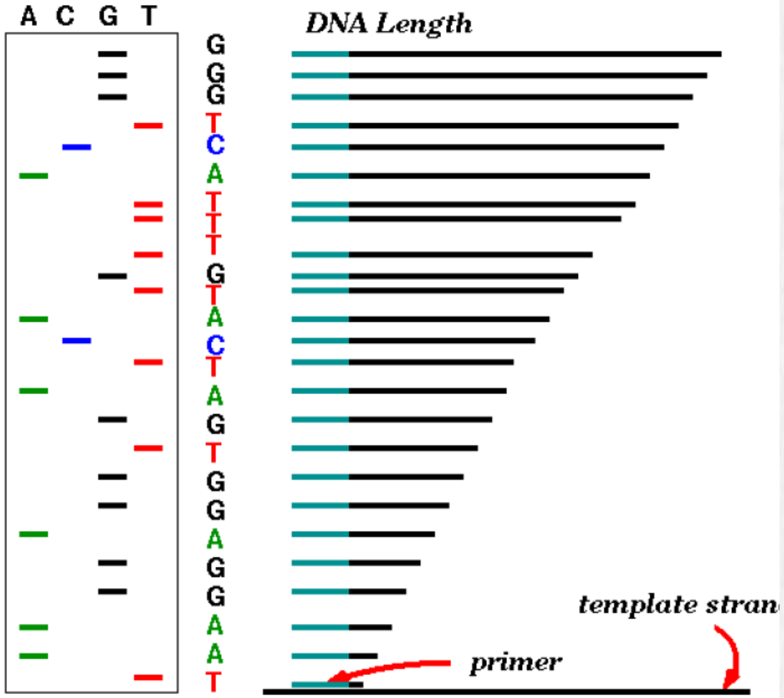
Gel electrophoresis: discrimination of 1 bp: size range below 300 bp in the lab

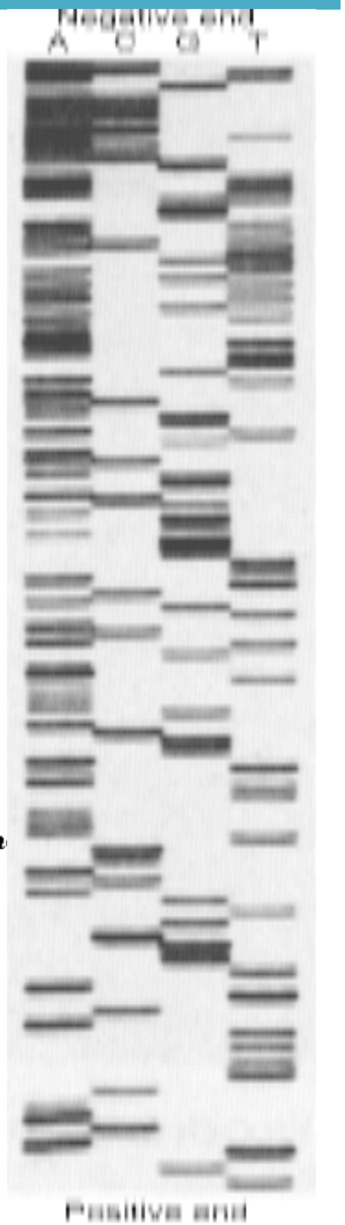
DNA template + 32P-labelled sequencing oligo

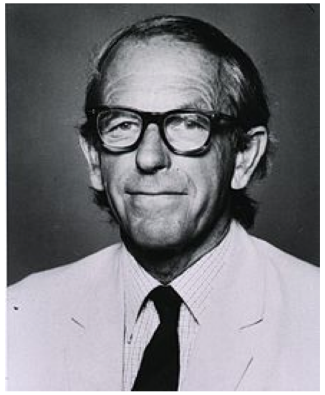
4 parallel seqeuncing reactions:

1. dATP, dCTP, dGTP, dTTP + ddATP (low conc)
2. dATP, dCTP, dGTP, dTTP + ddCTP (low conc)
3. dATP, dCTP, dGTP, dTTP + ddGTP (low conc)
4. dATP, dCTP, dGTP, dTTP + ddTTP (low conc)

Synthesis: starts with a32-P labeled DNA oligo
stops after incorporating a (marked) ddNTP







Frederic Sanger
Nobel Prize 1980

Dideoxy (Sanger) sequencing with Dye termination

Principle:

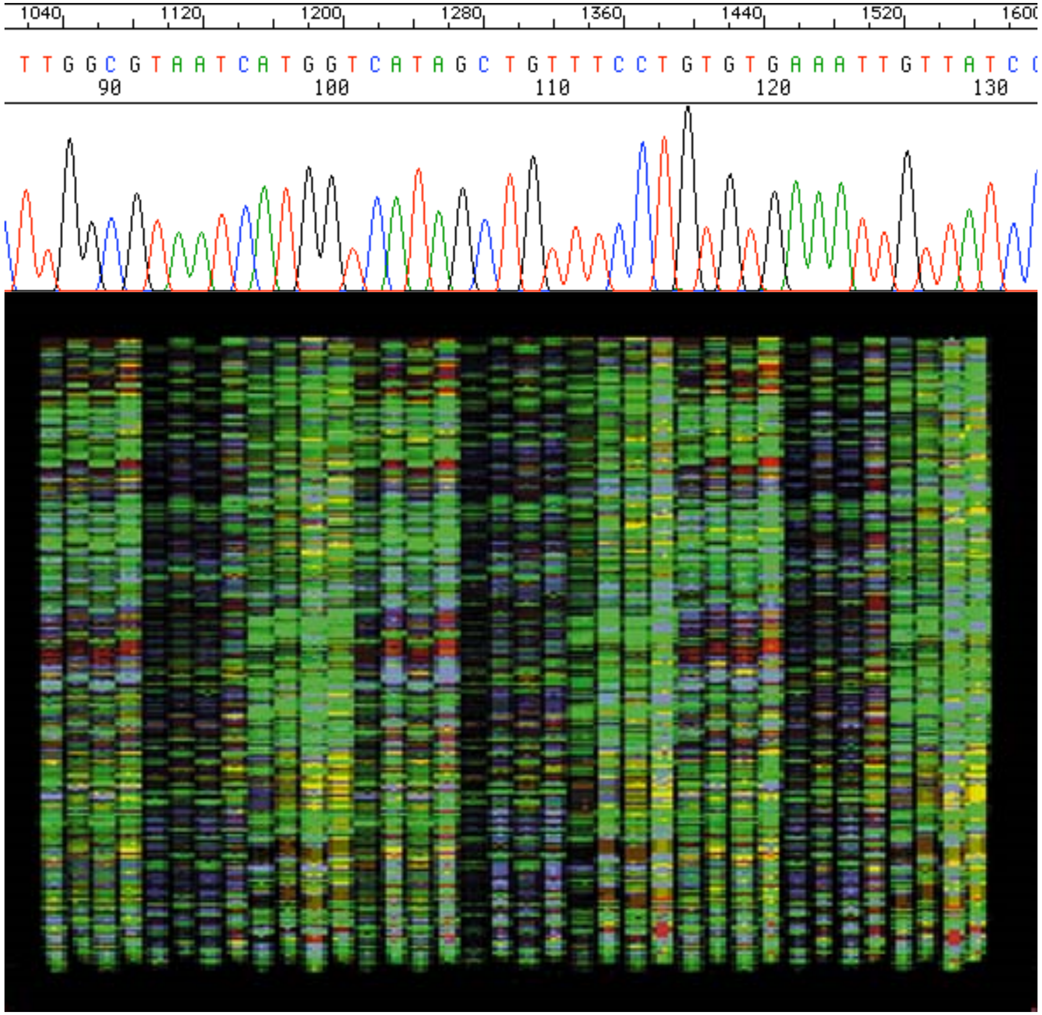
Gel electrophoresis: discrimination of 1 bp: size range below ~1000 bp

DNA template + sequencing oligo

1 seqeuncing reaction:

1. dATP, dCTP, dGTP, dTTP + ddATP-Dye1, ddCTP-Dye2, + ddGTP-Dye3+ddTTP-Dye4 (low conc)

Synthesis: starts with DNA oligo
stops after incorporating a (marked) ddNTP

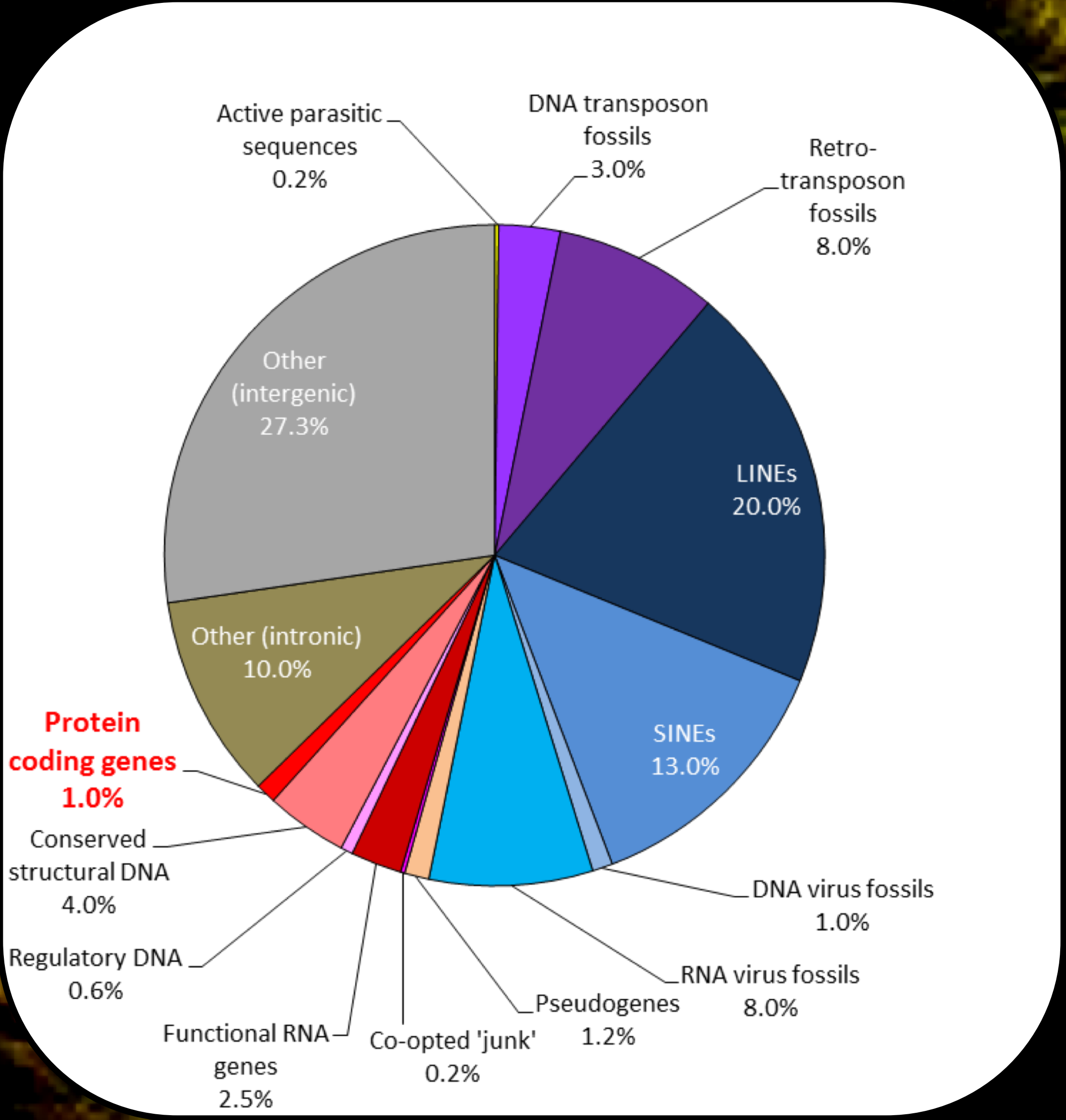


98% OF GENOMIC DNA DOES NOT ENCODE FOR PROTEINS

ca 50% transposable elements

1-2% protein coding genes

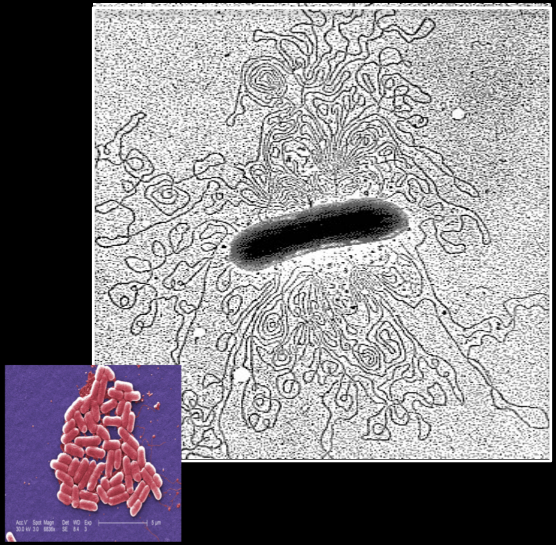
0.5-1% pseudogenes



Almost all genomic sequences are subjected to transcription

THE NUMBER OF PROTEIN CODING GENES IS RELATVLY LOW

E.coli



C. elegans

H. sapiens

Genome	$5x10^6$ bp	$1x10^8$ bp	$3x10^9$ bp
Chromosomes	1	6	23
Coding genes	6692	20541	21995
ncDNA			
non-coding RNA genes			
miRNAs			
pseudogenes			

??????????????

WHAT INFORMATION INCREASES ORGNAISMAL COMPLEXITY

ncDNA derived information?

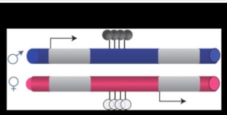
Why to study ncRNAs
1. There are things proteins cannot do

AGING-CANCER

Telomere function

hTR lncRNA

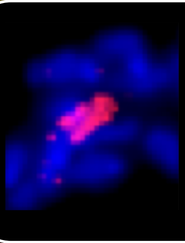
DEVELOPMENTAL DEFECTS



mono-allelic Gene expression

Airn lncRNA

Xist lncRNA



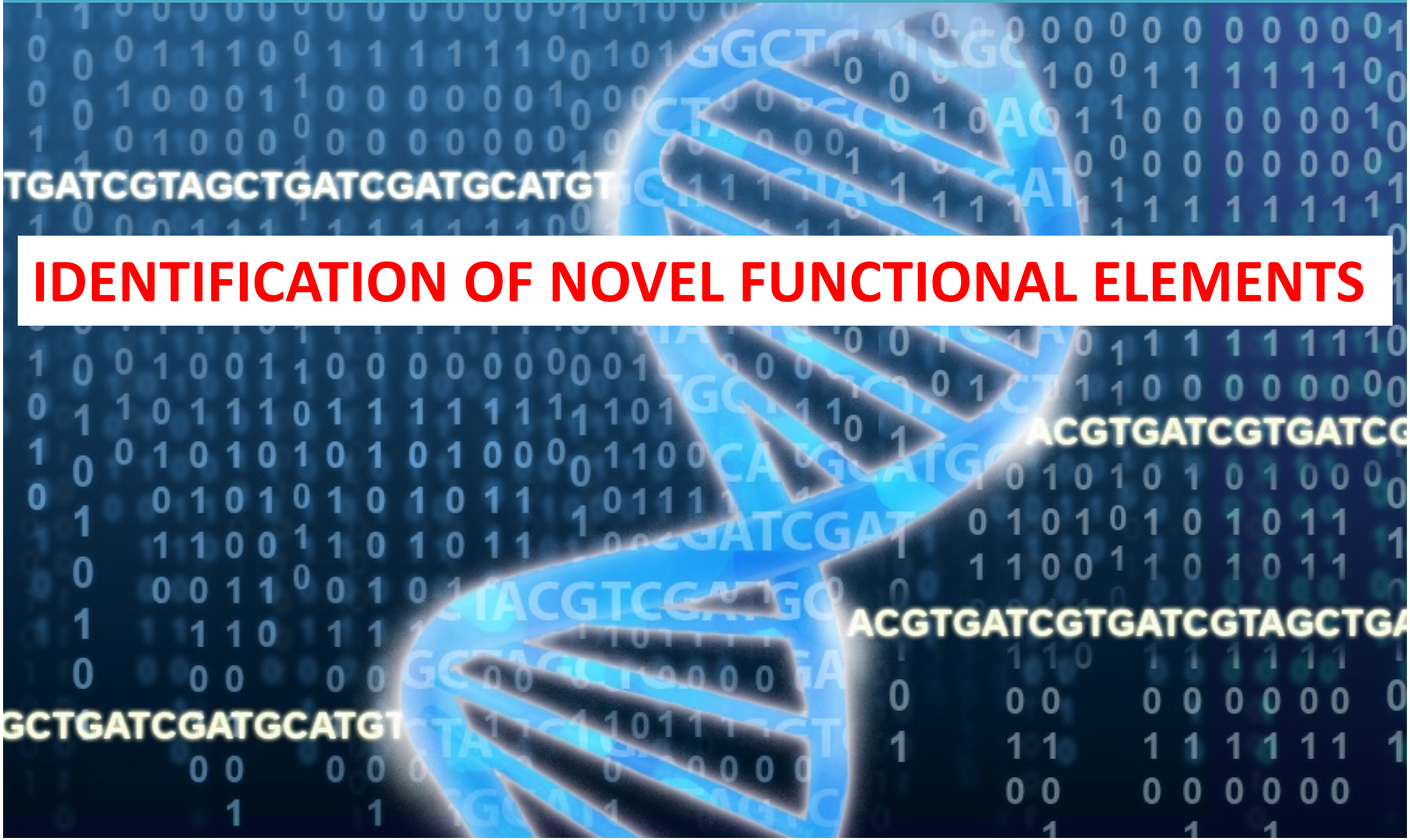
X inactivation Silencing 1000 genes on the Xi

LETHAL

2. they have high relevance for development and pathology

Classic Sanger sequencing is inefficient and slow:
→Establishement of massive parallel sequencing

NEXT GENERATION SEQEUNCING OF DNA AND RNA



IDENTIFICATION OF NOVEL FUNCTIONAL ELEMENTS

NEXT GENERATION SEQEUNCING OF DNA AND RNA

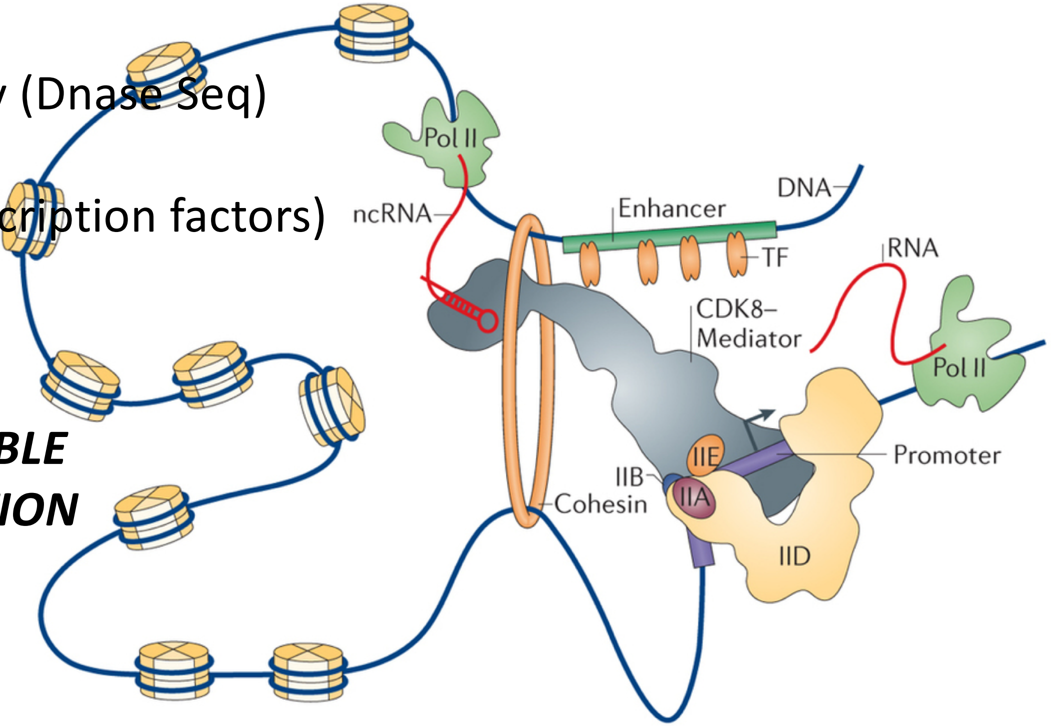
→IDENTIFICATION OF ALL GENES
→ IDENTIFICATION OF ALL CODING AND NON-CODING TRANSCRIPTS
→IDENTIFICATION OF REGUALTORY ELEMENTS

HOW CAN "NEW" = ***FUNCTIONAL ELEMENTS*** - (GENES/TRANSCRIPTS) BE IDENTIFIED?

1. DNA Seqeuncing (Human genome project, DNA-Seq)
2. Landscape of transcription: Seqeuncing of RNA (total RNA, small/large RNA, CAGE)
3. DNA methylation: High representation reduced representation bisulfite sequencing (RRBS)
4. Local chromatin structure:

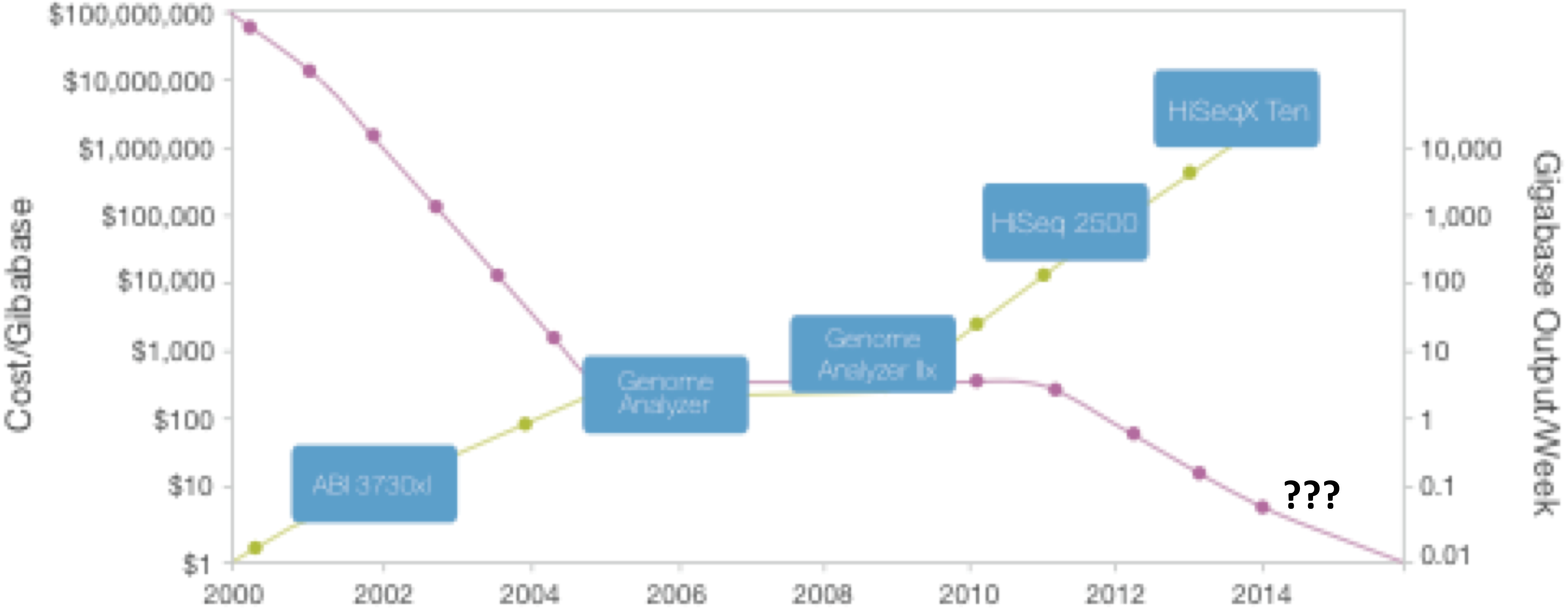
- determination of DNAseI hypersensitivity (Dnase Seq)
- nucelosome occupancy (MNase-seq)
- ChIP-seq (chromatin modifications, transcription factors)
- 3 Dimensional space interaction

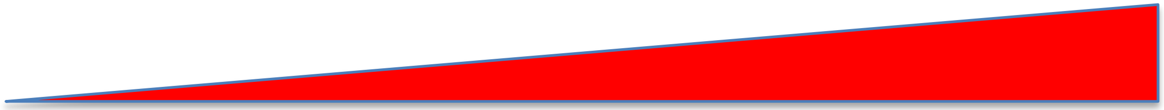
GENE REGUALTION AS INDICATOR OF POSSIBLE FUNCTIONAL RELEVANCE OF lncRNA FUNCTION



Nature Reviews | Molecular Cell Biology

PROGRESS IN SEQUENCING POWER





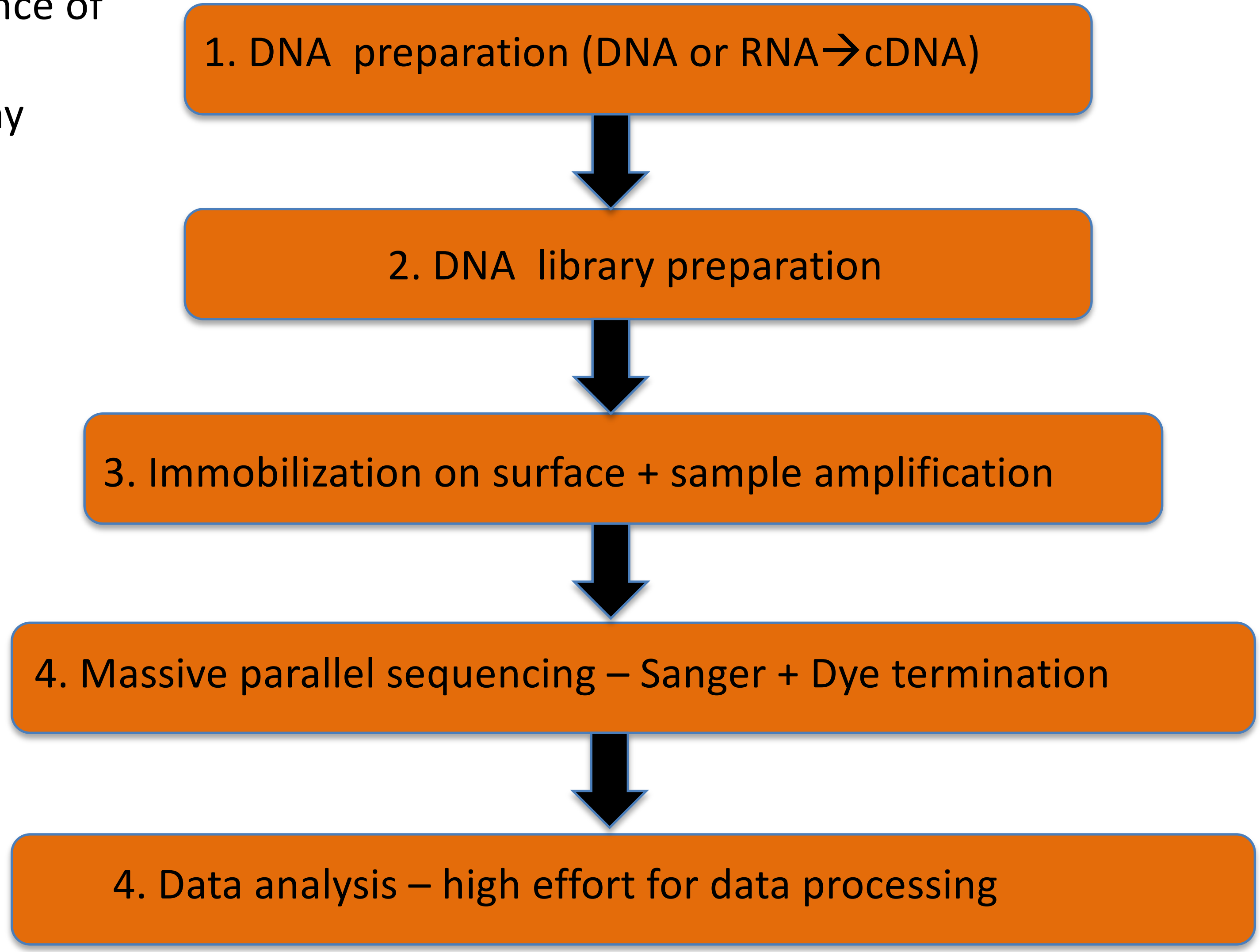
BIOINFORMATICS EFFORT = PROCESING OF DATA

Next generation sequencing:

MASSIVE PARALLEL SEQUENCING (ILLUMINA)

DNA SEQ – genome sequence of many organisms
RNA SEQ – all RNAs of many organisms – also at low anbunance

ChiP seq…..



1. DNA preparation (DNA or RNA→cDNA)

2. DNA library preparation

3. Immobilization on surface + sample amplification

4. Massive parallel sequencing – Sanger + Dye termination

4. Data analysis – high effort for data processing

Illumina: massive parallel sequencing Genomic DNA

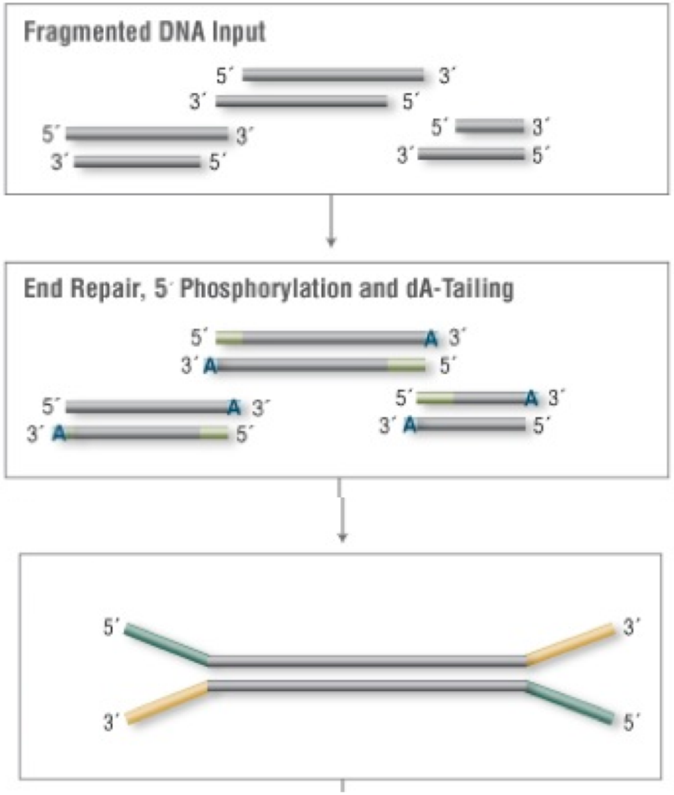
Generation of DNA libraries:

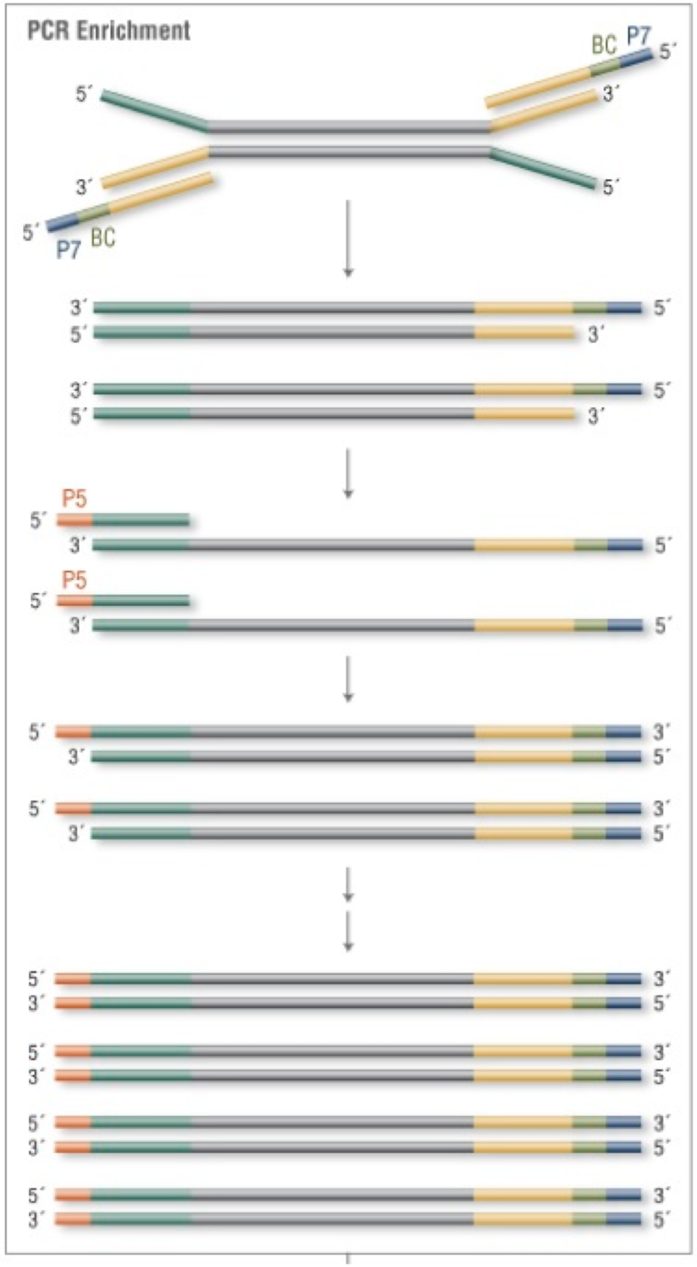
Application:

ChIP Seq

Genome Seq

Methyl Seq



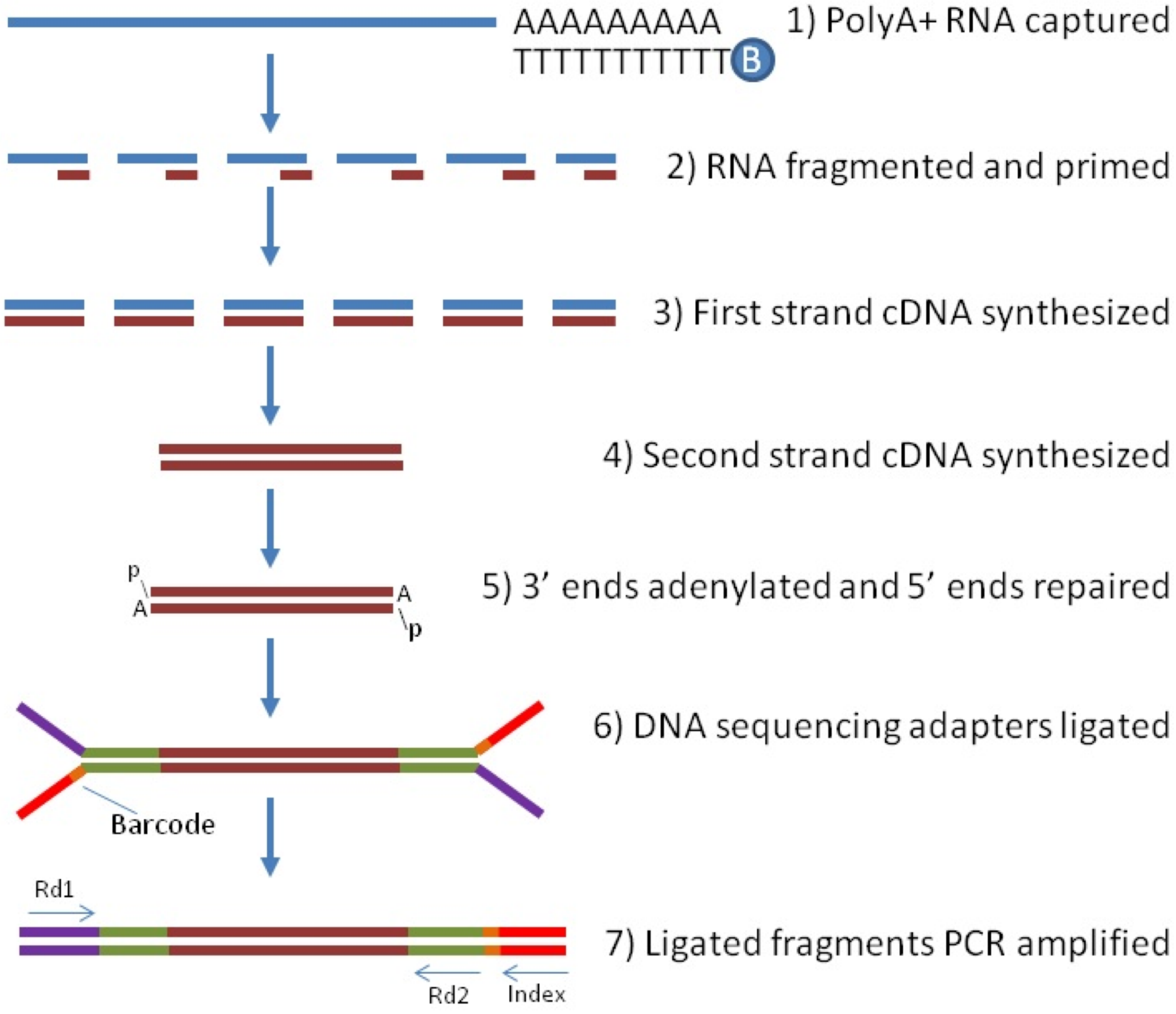


Illumina: massive parallel sequencing: ALL TRANSCRIPTS

Generation of RNA libraries:

Application:
RNA Seq

Important:
Involves cDNA synthesis


AAAAAAAAA
TTTTTTTTTTT B
1) PolyA+ RNA captured
2) RNA fragmented and primed
3) First strand cDNA synthesized
4) Second strand cDNA synthesized
p
A
A
p
5) 3' ends adenylated and 5' ends repaired
6) DNA sequencing adapters ligated
Barcode
Rd1
7) Ligated fragments PCR amplified
Rd2
Index

Illumina: massive parallel sequencing:

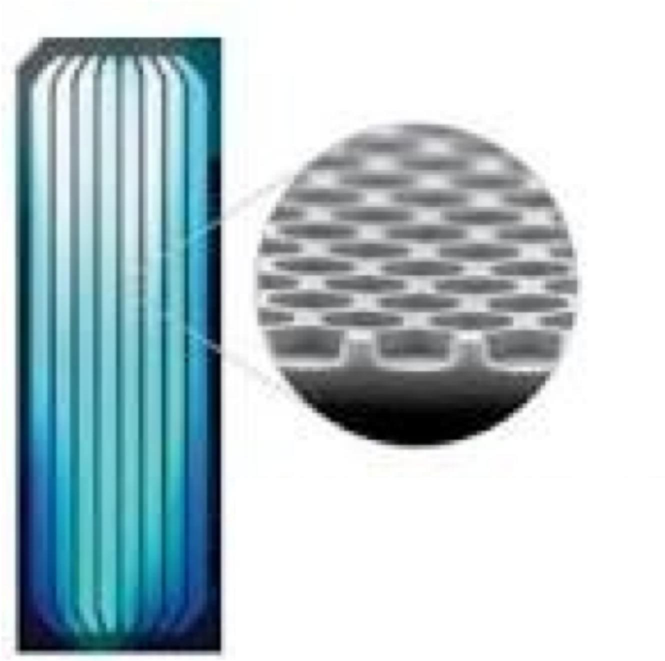
Illumina Massively Parallel Sequencing

https://www.illumina.com/company/video-hub/pfZp5Vgsbw0.html

The heart of the Illumina Massive Parallel Sequencer is the "FLOW-CELL". A surface with millions of small wells that allow thousands of Sanger-sequencing reaction In parallel = "massive parallel sequencing". In each well a SINGLE MOLECULE of DNA Is amplified and sequenced

Illumina offers the most potent massive sequencing instruments – leader on the market

Illumina: massive parallel sequencing:



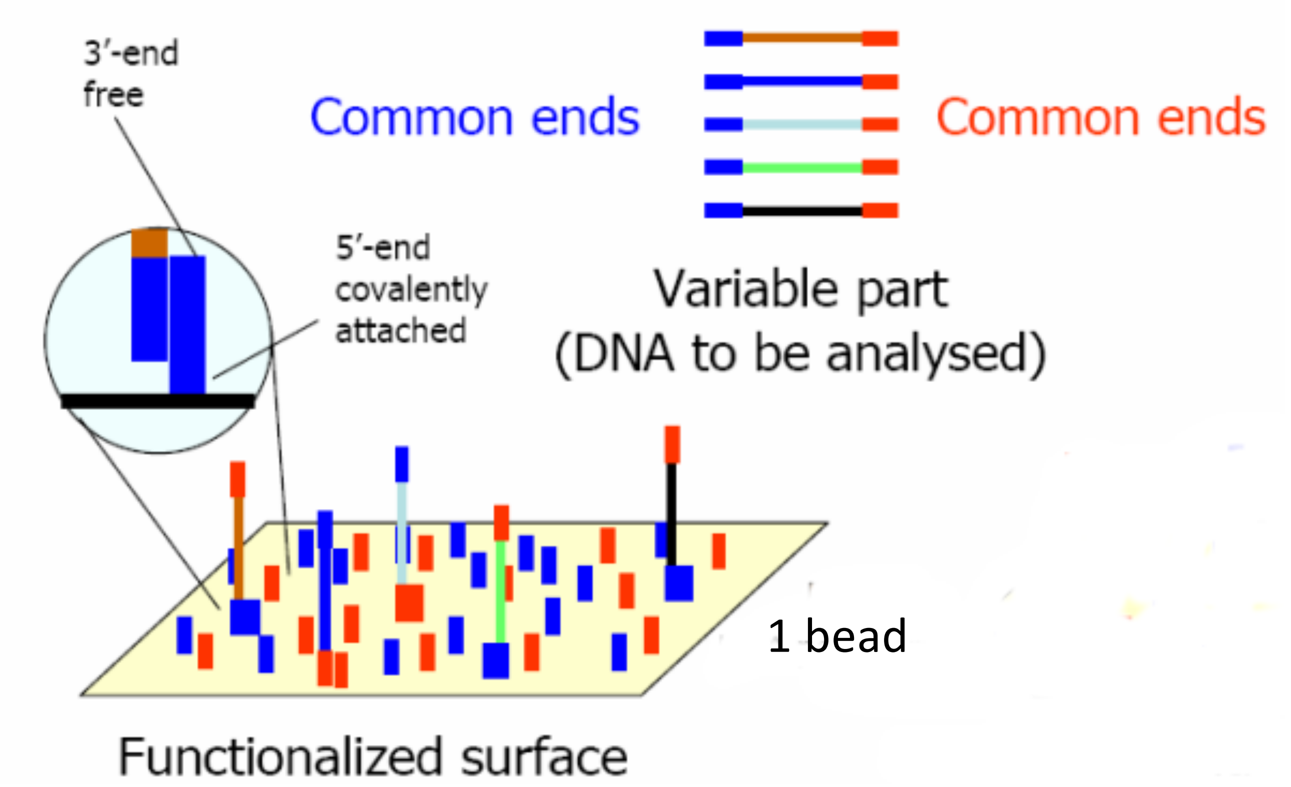
Flow cell contains surface with millions of wells

→Each well contains beads mounted with 2 species of oligonucleotides that hybridize with adaptor oligos of DNA library

→DNA library will be loaded onto the flow cell in a determined concentration:
ONLY ONE MOLECULE PER WELL

Illumina: massive parallel sequencing:

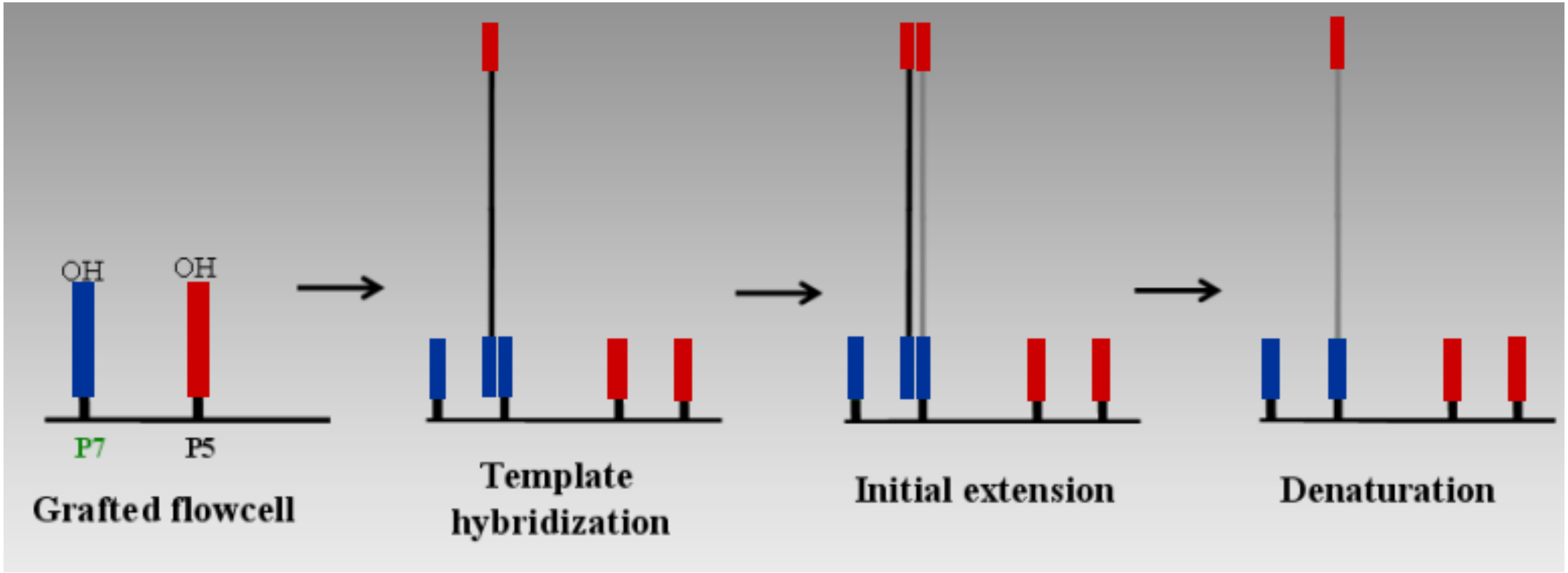
-making DNA library (~300bp fragments)
-ligation of adapters **A** and **B** to the fragments



-binding the ssDNA randomly to the flow cell surface
-***complementary*** primers are ligated to the surface

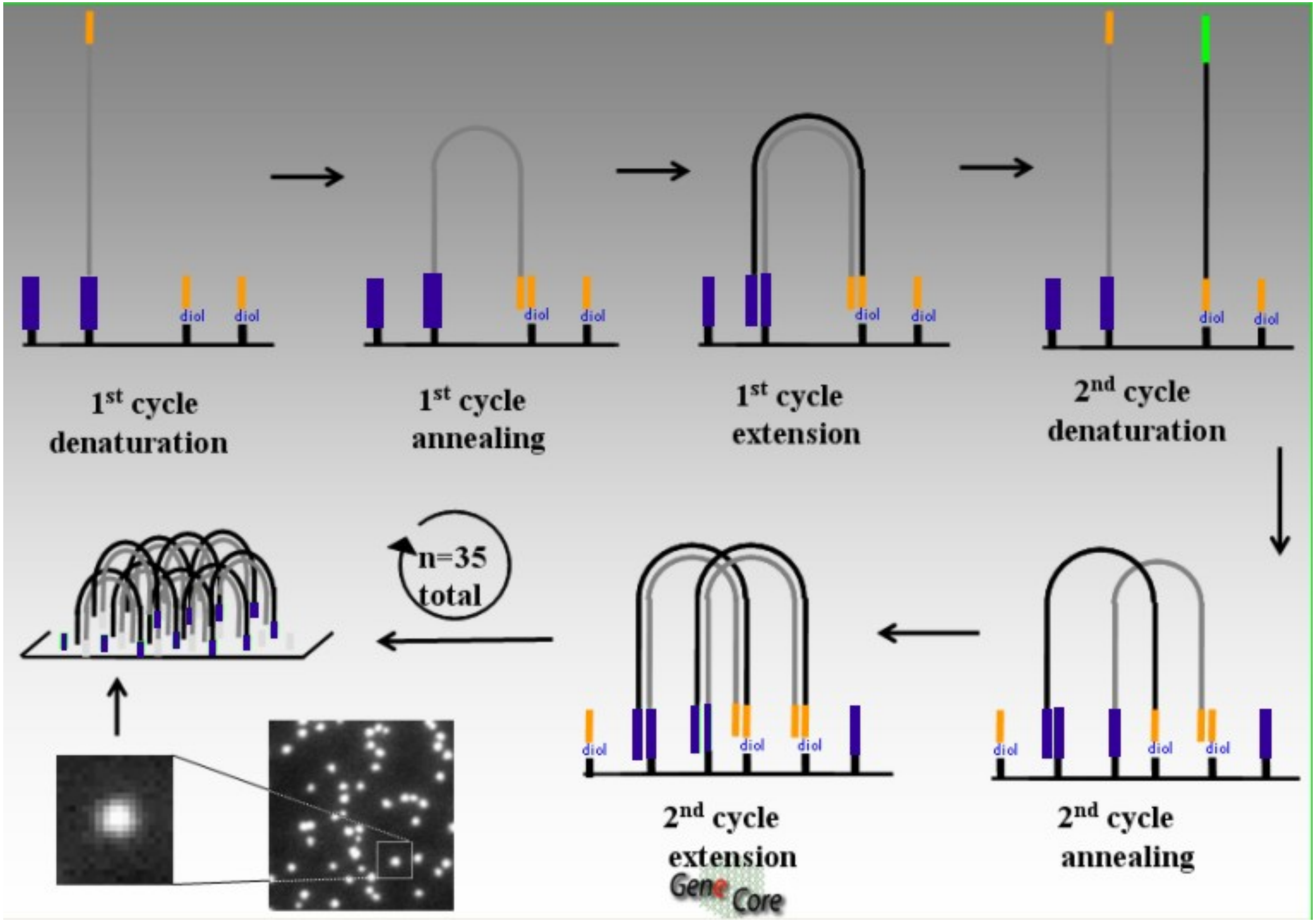
Illumina: massive parallel sequencing:

Bridge amplification:
initiation



On the surface: complementary oligos

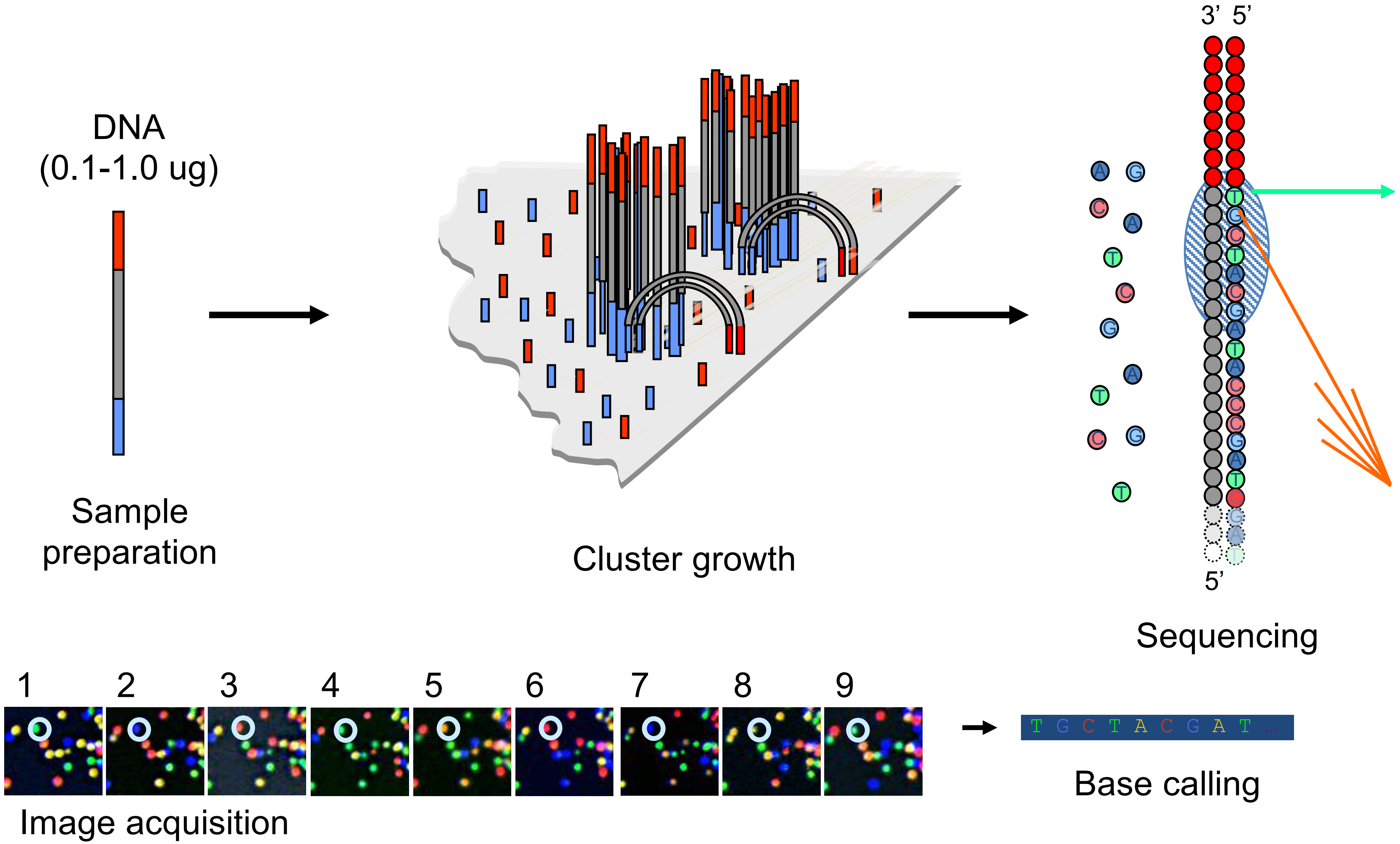
Illumina: massive parallel sequencing:



EMBL Gene Core

Illumina Sequencing Technology

Robust Reversible Terminator Chemistry Foundation



DNA (0.1-1.0 ug)

Sample preparation

Cluster growth

Sequencing

Image acquisition

Base calling

Why to study ncRNAs

Genomic stability

RNA maturation

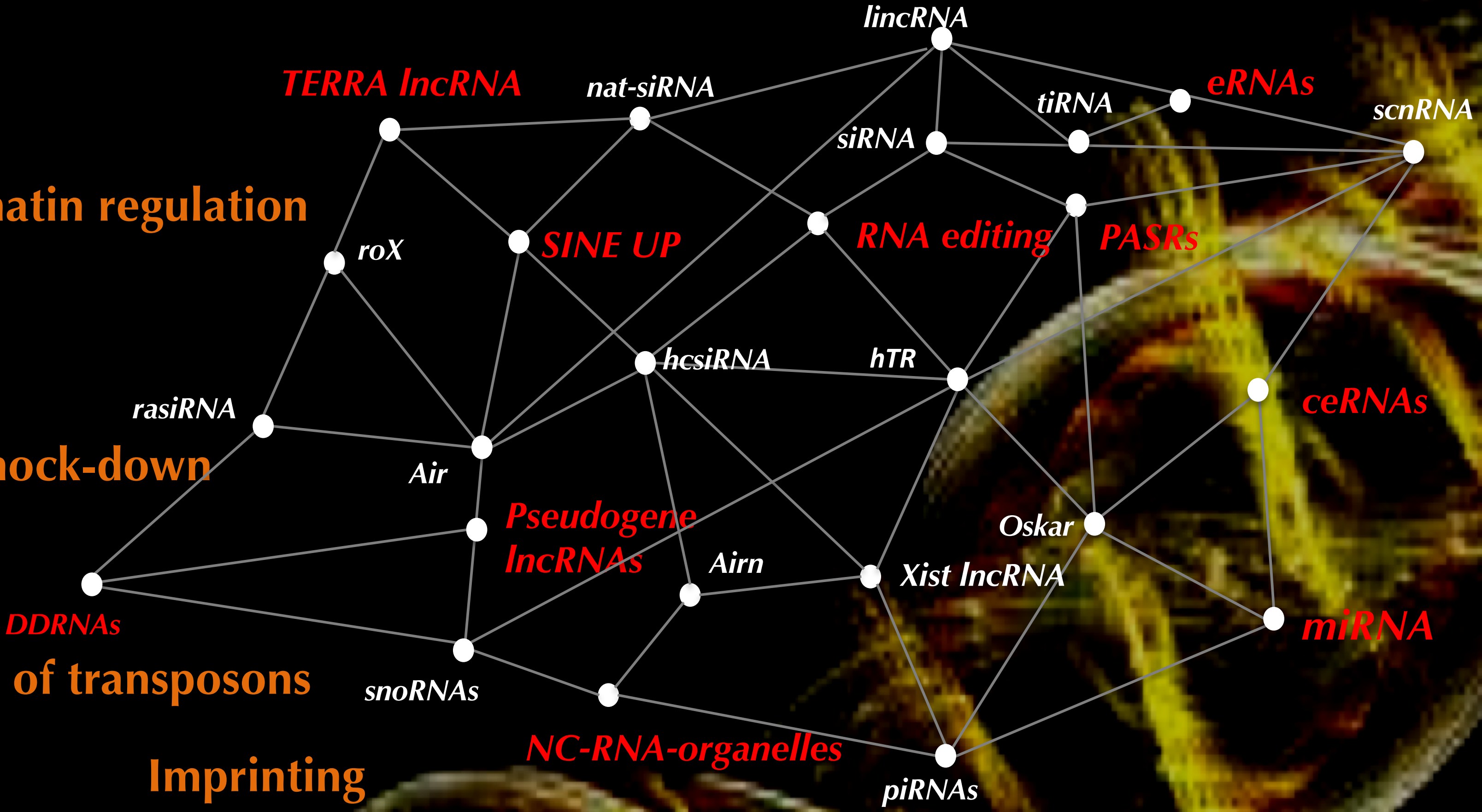
Translational regulation

Chromatin regulation

RNA knock-down

Control of transposons

Imprinting



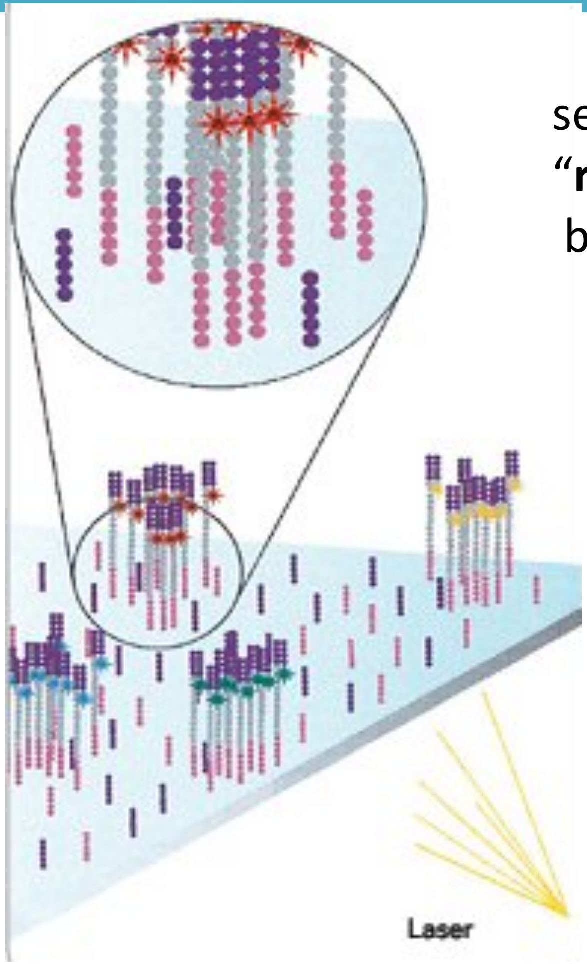
Aging

Cancer

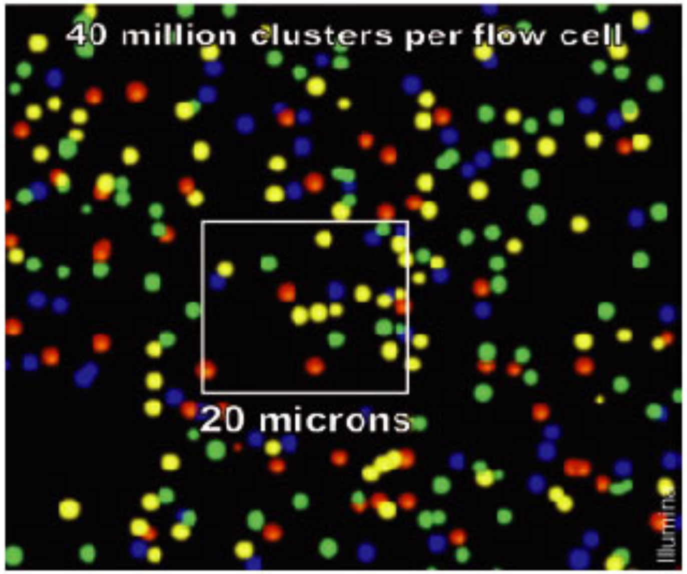
Degenerative diseases

Stem cells

Illumina: massive parallel sequencing:



sequencing by synthesis:
“**reverible terminator**” nucleotides
blocked + fluorescently labeled

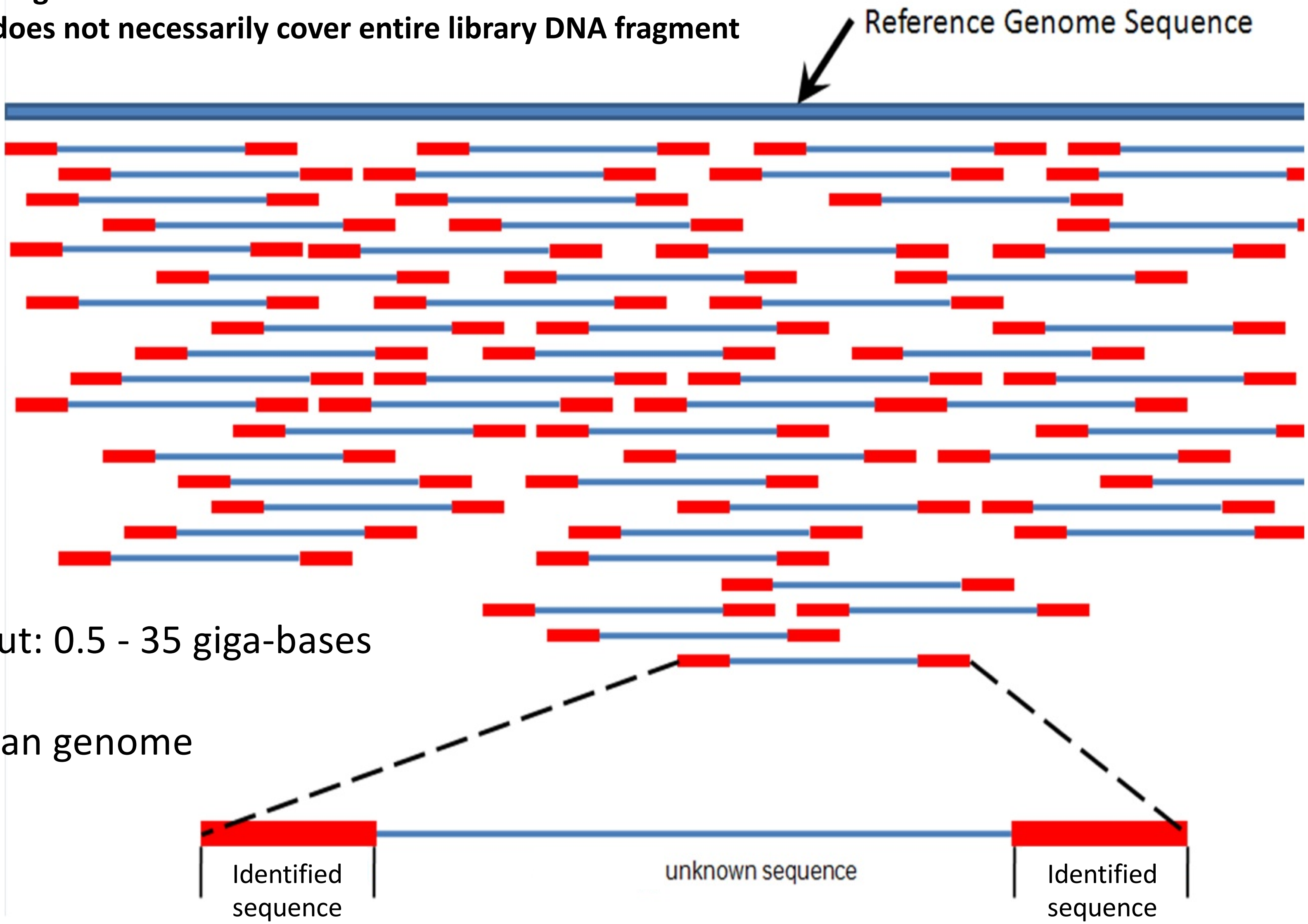


1. Synthesis = incorporation of fluorescent nucleotide: blocking synthesis
2. dye cleavage + elimination
3. wash step
4. Scanning of fluorescent signal
1. Synthesis = incorporation of fluorescent nucleotide: blocking synthesis
READ LENGTH: ca: 150nt from each primer (2x150nt = 300nt)

illumina.com

Data analysis: obtained sequence reads are aligned along genomic DNA sequence → high number of reads necessary to obtain full sequence coverage

Read length: 50 – max. 300 nt
Read does not necessarily cover entire library DNA fragment

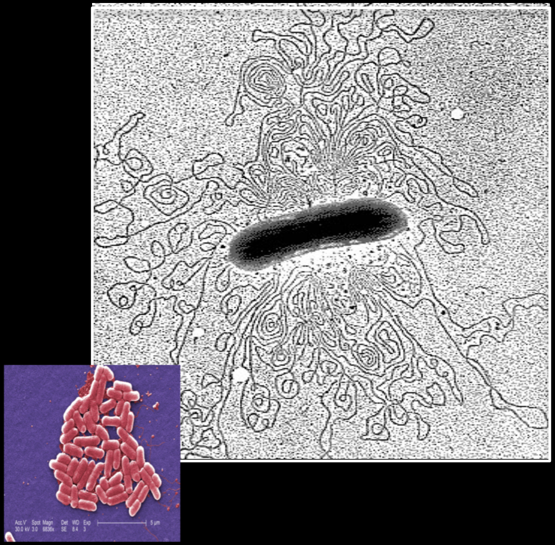


Max. output: 0.5 - 35 giga-bases
$=3.5*10^{10}$
= 10x human genome

Sequence derived from one amplified cluster

Reason 1:
The non-coding genome (r)evolution

E.coli



C. elegans

H. sapiens

Genome	5x10^6 bp	1x10^8 bp	3x10^9 bp
Chromosomes	1	6	23
Coding genes	6692	20541	21995
ncDNA	5%	60%	**98%**
non-coding RNA genes	15	23136	ca. 40000
miRNAs	0	224	4274
pseudogenes	21	1522	10616

The ENCODE PROJECT: IDENTIFCATION OF ALL FUNCTIONAL ELEMENTS IN THE REMAINING 98% OF THE HUMAN GENOME (2003)

The Encyclopedia of DNA Elements (ENCODE) is a public research project launched by the US National Human Genome Research Institute (NHGRI) in September 2003.

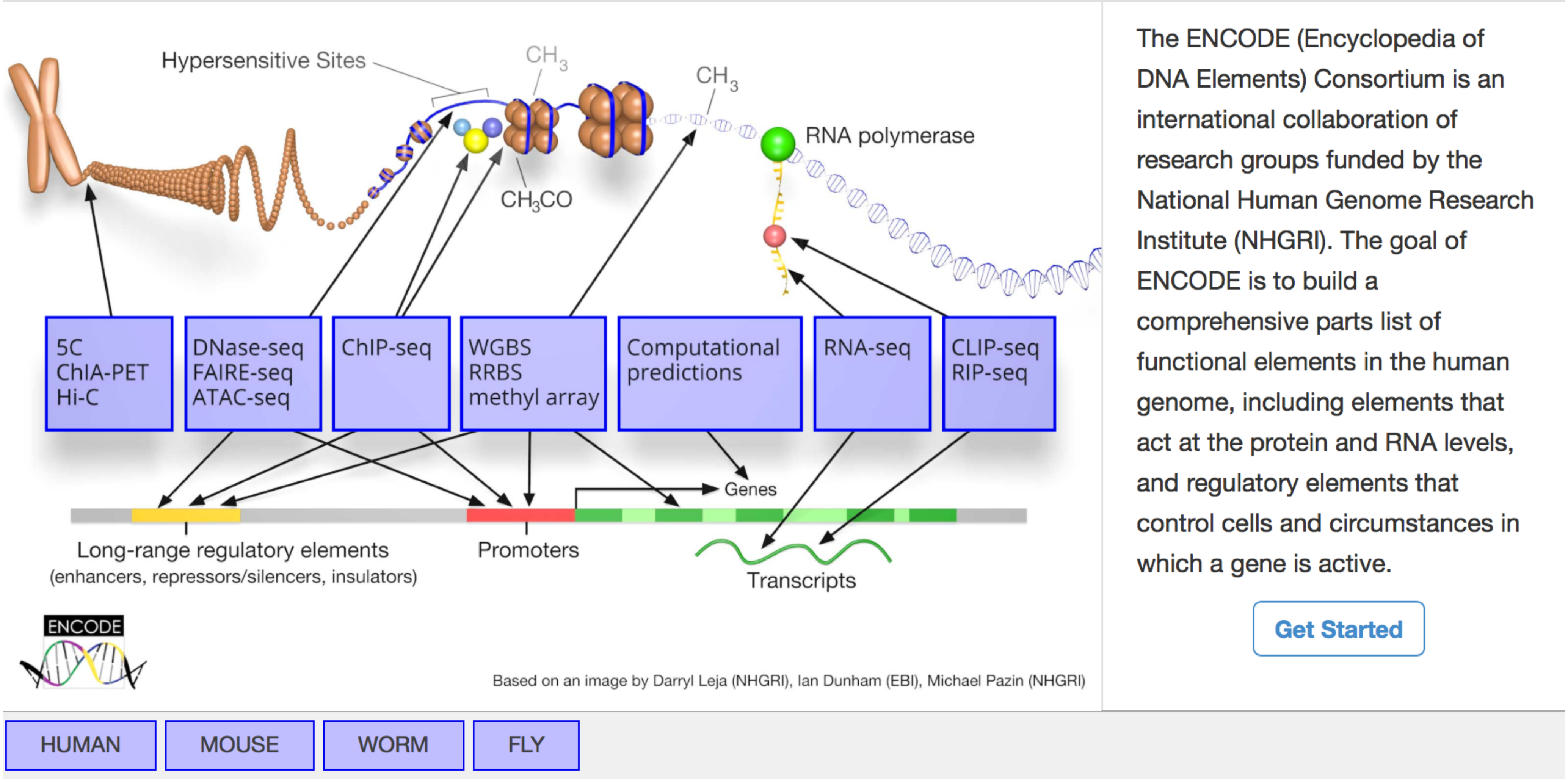
Intended as a follow-up to the Human Genome Project (Genomic Research), the ENCODE project aims to identify all functional elements in the human genome.

The project involves a worldwide consortium of research groups, and data generated from this project can be accessed through public databases.

NCODE is implemented in three phases: the pilot phase, the technology development phase and the production phase.

Along the pilot phase, the ENCODE Consortium evaluated strategies for identifying various types of genomic elements. The goal of the pilot phase was to identify a set of procedures that, in combination, could be applied cost-effectively and at high-throughput to accurately and comprehensively characterize large regions of the human genome. The pilot phase had to reveal gaps in the current set of tools for detecting functional sequences, and was also thought to reveal whether some methods used by that time were inefficient or unsuitable for large-scale utilization. Some of these problems had to be addressed in the ENCODE technology development phase (being executed concurrently with the pilot phase), which aimed to devise new laboratory and computational methods that would improve our ability to identify known functional sequences or to discover new functional genomic elements. The results of the first two phases determined the best path forward for analysing the remaining 99% of the human genome in a cost-effective and comprehensive production phase.

ENCODE: Encyclopedia of DNA Elements



Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

https://www.encodeproject.org

NEXT GENERATION SEQEUNCING OF DNA AND RNA

→IDENTIFICATION OF ALL GENES

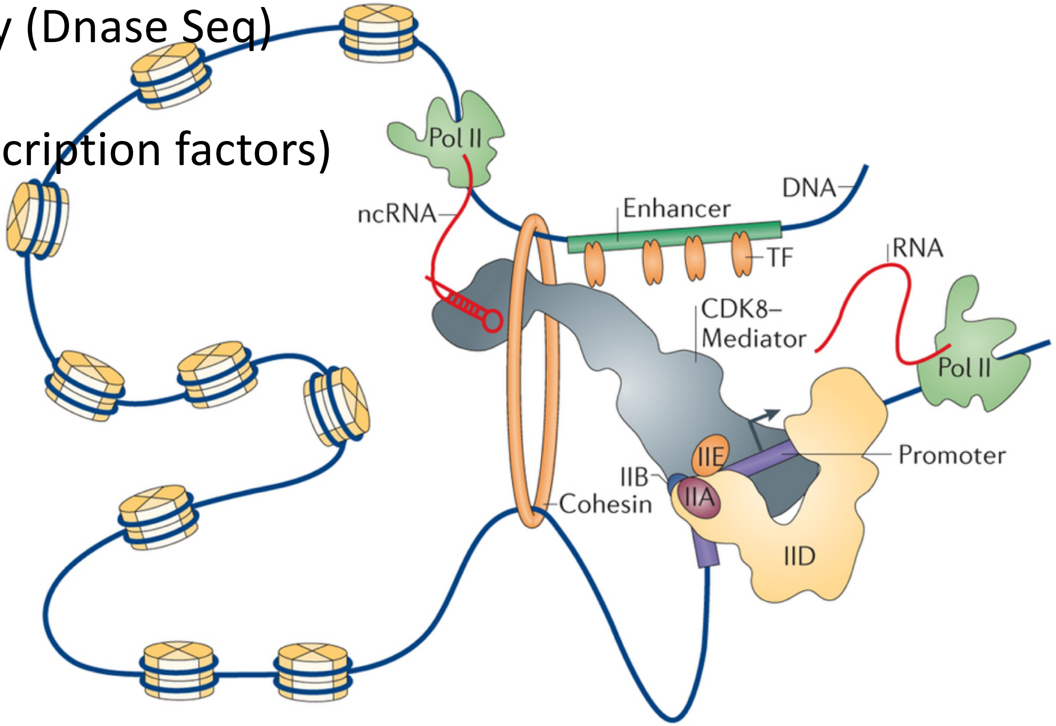
→ IDENTIFICATION OF ALL CODING AND NON-CODING TRANSCRIPTS

HOW CAN GENES/TRANSCRIPTS BE DEFINED?

1. DNA Seqeuncing (Human genome project, DNA-Seq)
2. Landscape of transcription: Sequencing of RNA (total RNA, small/large RNA, CAGE)
3. DNA methylation: High representation reduced representation bisulfite sequencing (RRBS)
4. Local chromatin structure:

- determination of DNAseI hypersensitivity (Dnase Seq)
- nucelosome occupancy (MNase-seq)
- ChIP-seq (chromatin modifications, transcription factors)
- 3 Dimensional space interaction

chromatin structure is combined with RNA expression data and DNA sequence to identify all genes/functional elements The presence of regulated chromatin indicates the presence of a real functional element



Nature Reviews | Molecular Cell Biology

ENCODE MASSIVE EXPERIMENTAL INPUT

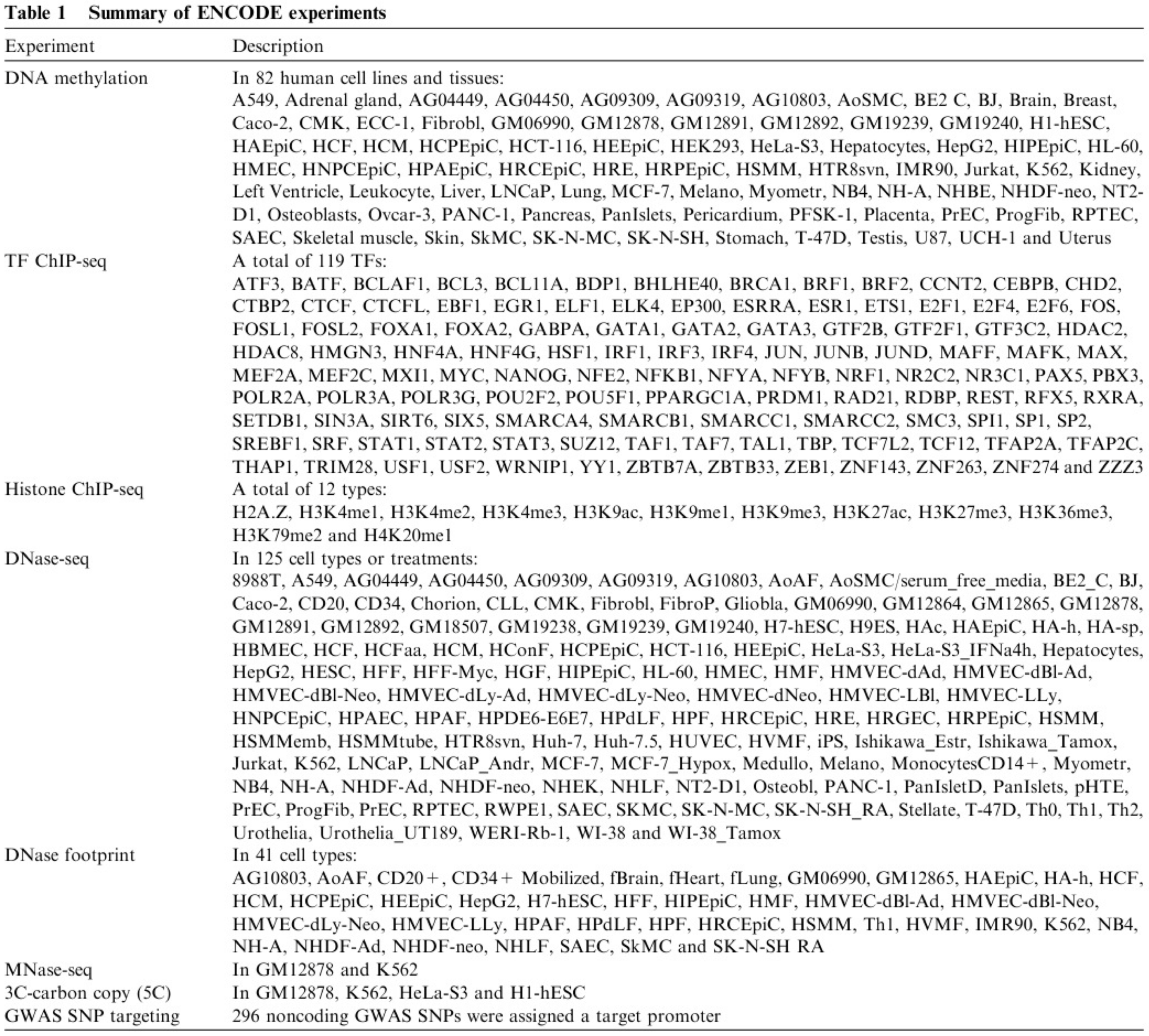


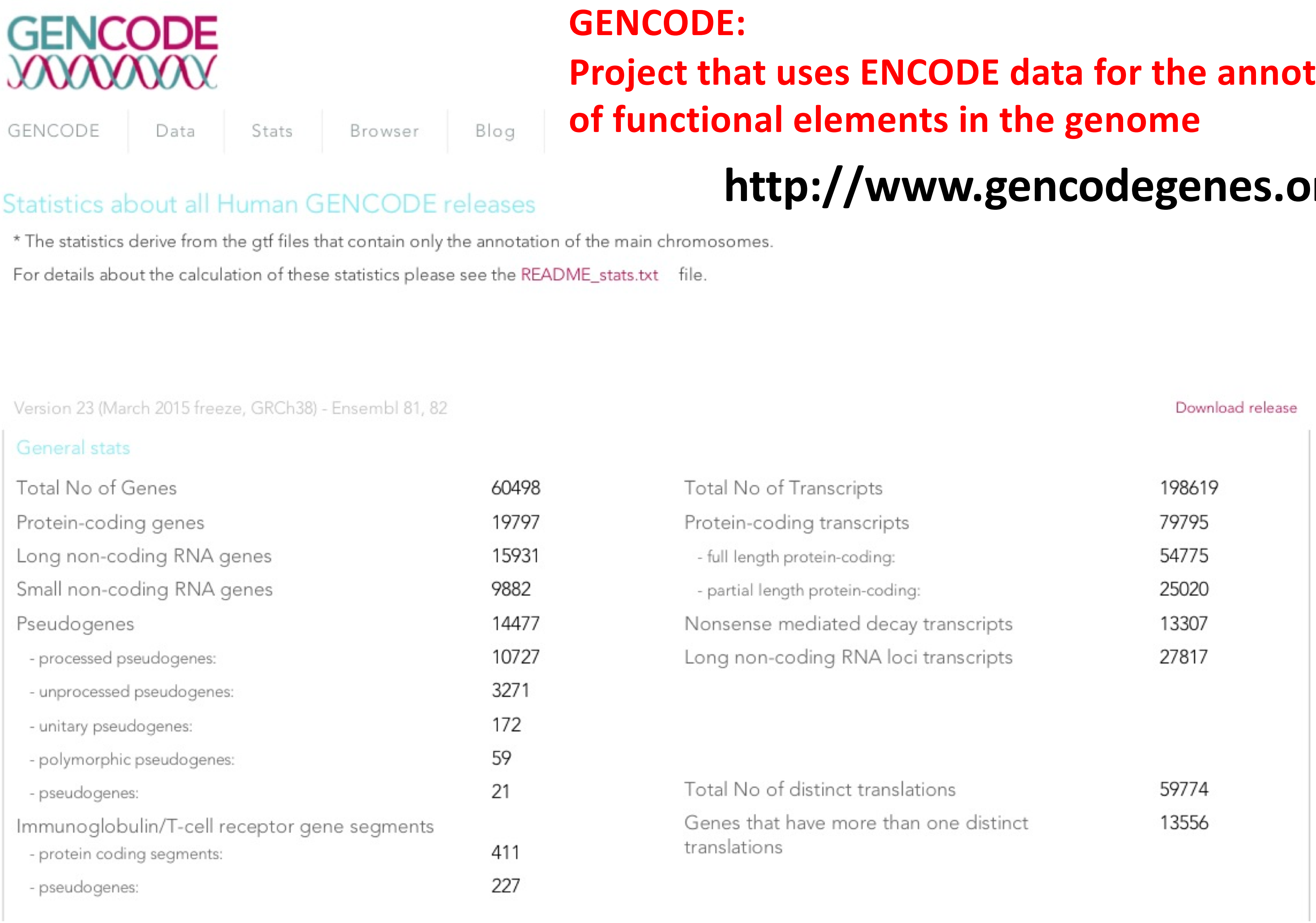
Table 1 Summary of ENCODE experiments

Experiment	Description
DNA methylation	In 82 human cell lines and tissues: A549, Adrenal gland, AG04449, AG04450, AG09309, AG09319, AG10803, AoSMC, BE2 C, BJ, Brain, Breast, Caco-2, CMK, ECC-1, Fibrobl, GM06990, GM12878, GM12891, GM12892, GM19239, GM19240, H1-hESC, HAEpiC, HCF, HCM, HCPEpiC, HCT-116, HEEpiC, HEK293, HeLa-S3, Hepatocytes, HepG2, HIPEpiC, HL-60, HMEC, HNPCEpiC, HPAEpiC, HRCEpiC, HRE, HRPEpiC, HSMM, HTR8svn, IMR90, Jurkat, K562, Kidney, Left Ventricle, Leukocyte, Liver, LNCaP, Lung, MCF-7, Melano, Myometr, NB4, NH-A, NHBE, NHDF-neo, NT2-D1, Osteoblasts, Ovcar-3, PANC-1, Pancreas, PanIslets, Pericardium, PFSK-1, Placenta, PrEC, ProgFib, RPTEC, SAEC, Skeletal muscle, Skin, SkMC, SK-N-MC, SK-N-SH, Stomach, T-47D, Testis, U87, UCH-1 and Uterus
TF ChIP-seq	A total of 119 TFs: ATF3, BATF, BCLAF1, BCL3, BCL11A, BDP1, BHLHE40, BRCA1, BRF1, BRF2, CCNT2, CEBPB, CHD2, CTBP2, CTCF, CTCFL, EBF1, EGR1, ELF1, ELK4, EP300, ESRRA, ESR1, ETS1, E2F1, E2F4, E2F6, FOS, FOSL1, FOSL2, FOXA1, FOXA2, GABPA, GATA1, GATA2, GATA3, GTF2B, GTF2F1, GTF3C2, HDAC2, HDAC8, HMGN3, HNF4A, HNF4G, HSF1, IRF1, IRF3, IRF4, JUN, JUNB, JUND, MAFF, MAFK, MAX, MEF2A, MEF2C, MXI1, MYC, NANOG, NFE2, NFKB1, NFYA, NFYB, NRF1, NR2C2, NR3C1, PAX5, PBX3, POLR2A, POLR3A, POLR3G, POU2F2, POU5F1, PPARGC1A, PRDM1, RAD21, RDBP, REST, RFX5, RXRA, SETDB1, SIN3A, SIRT6, SIX5, SMARCA4, SMARCB1, SMARCC1, SMARCC2, SMC3, SPI1, SP1, SP2, SREBF1, SRF, STAT1, STAT2, STAT3, SUZ12, TAF1, TAF7, TAL1, TBP, TCF7L2, TCF12, TFAP2A, TFAP2C, THAP1, TRIM28, USF1, USF2, WRNIP1, YY1, ZBTB7A, ZBTB33, ZEB1, ZNF143, ZNF263, ZNF274 and ZZZ3
Histone ChIP-seq	A total of 12 types: H2A.Z, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me1, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K79me2 and H4K20me1
DNase-seq	In 125 cell types or treatments: 8988T, A549, AG04449, AG04450, AG09309, AG09319, AG10803, AoAF, AoSMC/serum_free_media, BE2_C, BJ, Caco-2, CD20, CD34, Chorion, CLL, CMK, Fibrobl, FibroP, Gliobla, GM06990, GM12864, GM12865, GM12878, GM12891, GM12892, GM18507, GM19238, GM19239, GM19240, H7-hESC, H9ES, HAc, HAEpiC, HA-h, HA-sp, HBMEC, HCF, HCFaa, HCM, HConF, HCPEpiC, HCT-116, HEEpiC, HeLa-S3, HeLa-S3_IFNa4h, Hepatocytes, HepG2, HESC, HFF, HFF-Myc, HGF, HIPEpiC, HL-60, HMEC, HMF, HMVEC-dAd, HMVEC-dBl-Ad, HMVEC-dBl-Neo, HMVEC-dLy-Ad, HMVEC-dLy-Neo, HMVEC-dNeo, HMVEC-LBl, HMVEC-LLy, HNPCEpiC, HPAEC, HPAF, HPDE6-E6E7, HPdLF, HPF, HRCEpiC, HRE, HRGEC, HRPEpiC, HSMM, HSMMemb, HSMMtube, HTR8svn, Huh-7, Huh-7.5, HUVEC, HVMF, iPS, Ishikawa_Estr, Ishikawa_Tamox, Jurkat, K562, LNCaP, LNCaP_Andr, MCF-7, MCF-7_Hypox, Medullo, Melano, MonocytesCD14+, Myometr, NB4, NH-A, NHDF-Ad, NHDF-neo, NHEK, NHLF, NT2-D1, Osteobl, PANC-1, PanIsletD, PanIslets, pHTE, PrEC, ProgFib, PrEC, RPTEC, RWPE1, SAEC, SKMC, SK-N-MC, SK-N-SH_RA, Stellate, T-47D, Th0, Th1, Th2, Urothelia, Urothelia_UT189, WERI-Rb-1, WI-38 and WI-38_Tamox
DNase footprint	In 41 cell types: AG10803, AoAF, CD20+, CD34+ Mobilized, fBrain, fHeart, fLung, GM06990, GM12865, HAEpiC, HA-h, HCF, HCM, HCPEpiC, HEEpiC, HepG2, H7-hESC, HFF, HIPEpiC, HMF, HMVEC-dBl-Ad, HMVEC-dBl-Neo, HMVEC-dLy-Neo, HMVEC-LLy, HPAF, HPdLF, HPF, HRCEpiC, HSMM, Th1, HVMF, IMR90, K562, NB4, NH-A, NHDF-Ad, NHDF-neo, NHLF, SAEC, SkMC and SK-N-SH RA
MNase-seq	In GM12878 and K562
3C-carbon copy (5C)	In GM12878, K562, HeLa-S3 and H1-hESC
GWAS SNP targeting	296 noncoding GWAS SNPs were assigned a target promoter

Ca.
400 Mio $

GENCODE:
Project that uses ENCODE data for the annotation of functional elements in the genome

http://www.gencodegenes.org/



Statistics about all Human GENCODE releases

* The statistics derive from the gtf files that contain only the annotation of the main chromosomes.

For details about the calculation of these statistics please see the README_stats.txt file.

Version 23 (March 2015 freeze, GRCh38) - Ensembl 81, 82

Download release

General stats

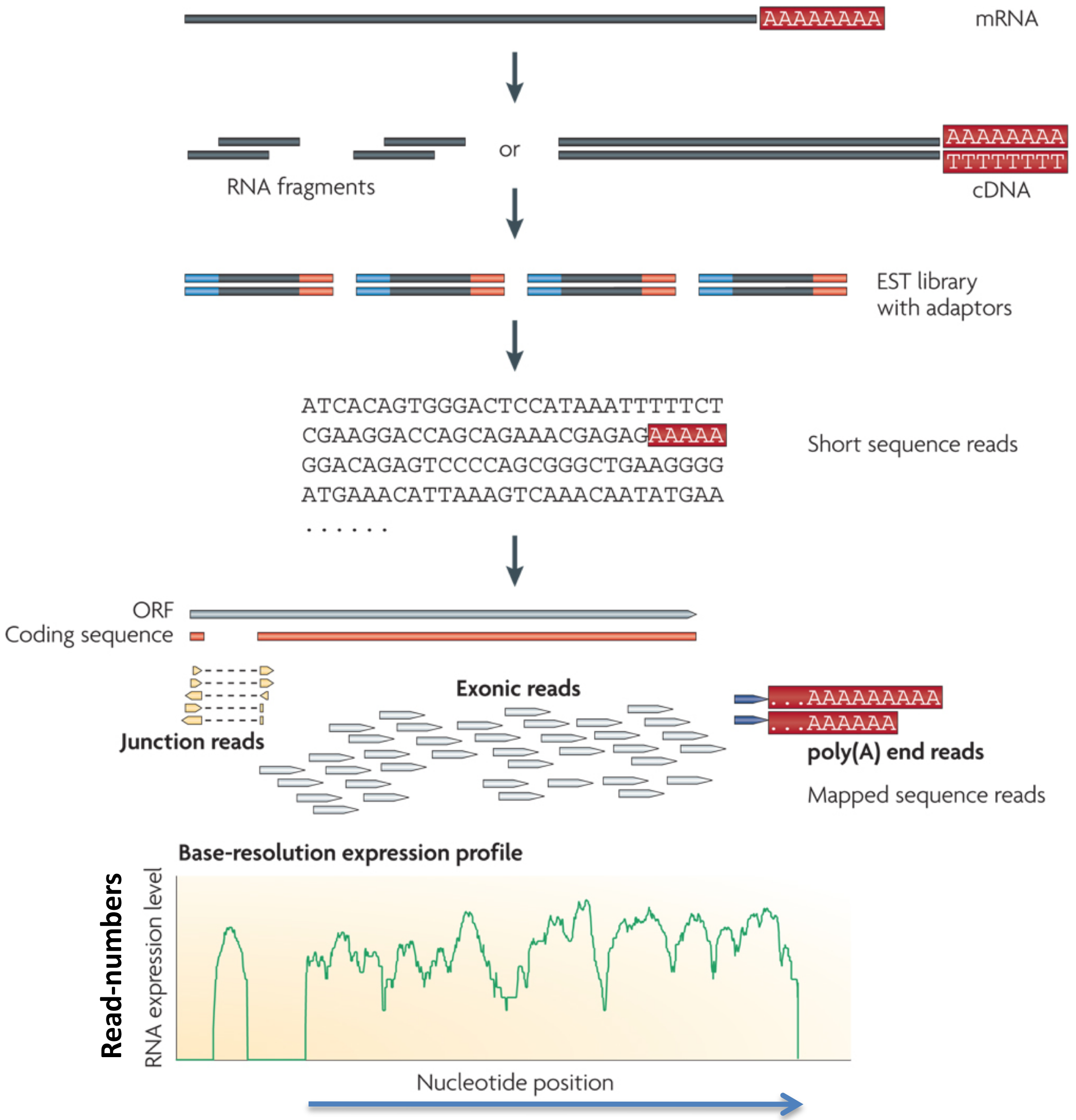
Total No of Genes	60498
Protein-coding genes	19797
Long non-coding RNA genes	15931
Small non-coding RNA genes	9882
Pseudogenes	14477
- processed pseudogenes:	10727
- unprocessed pseudogenes:	3271
- unitary pseudogenes:	172
- polymorphic pseudogenes:	59
- pseudogenes:	21
Immunoglobulin/T-cell receptor gene segments	
- protein coding segments:	411
- pseudogenes:	227

Total No of Transcripts	198619
Protein-coding transcripts	79795
- full length protein-coding:	54775
- partial length protein-coding:	25020
Nonsense mediated decay transcripts	13307
Long non-coding RNA loci transcripts	27817
Total No of distinct translations	59774
Genes that have more than one distinct translations	13556

2. RNA SEQ – TO IDENTIFY ALL SORTS OF TRANSCRIPTS

Serial Analysis of Gene Expression (SAGE, superSAGE)

Method can also be used
for all transcripts
When using a random
Primers for reverse
transcription

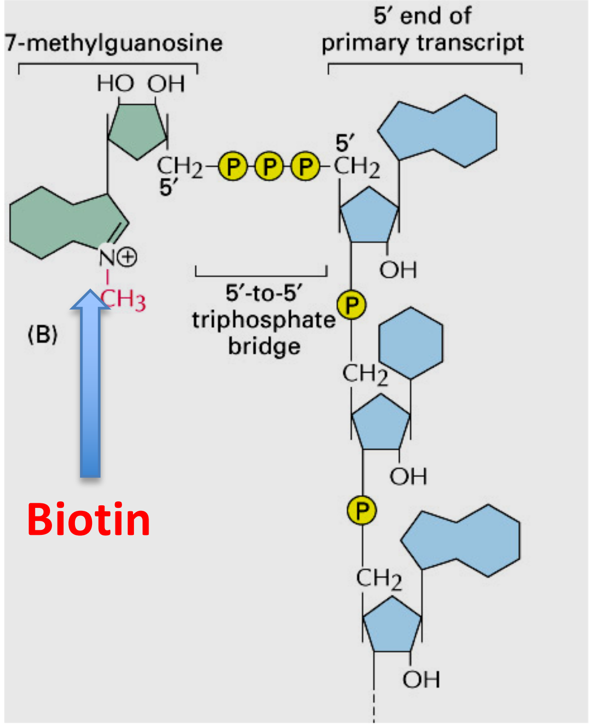


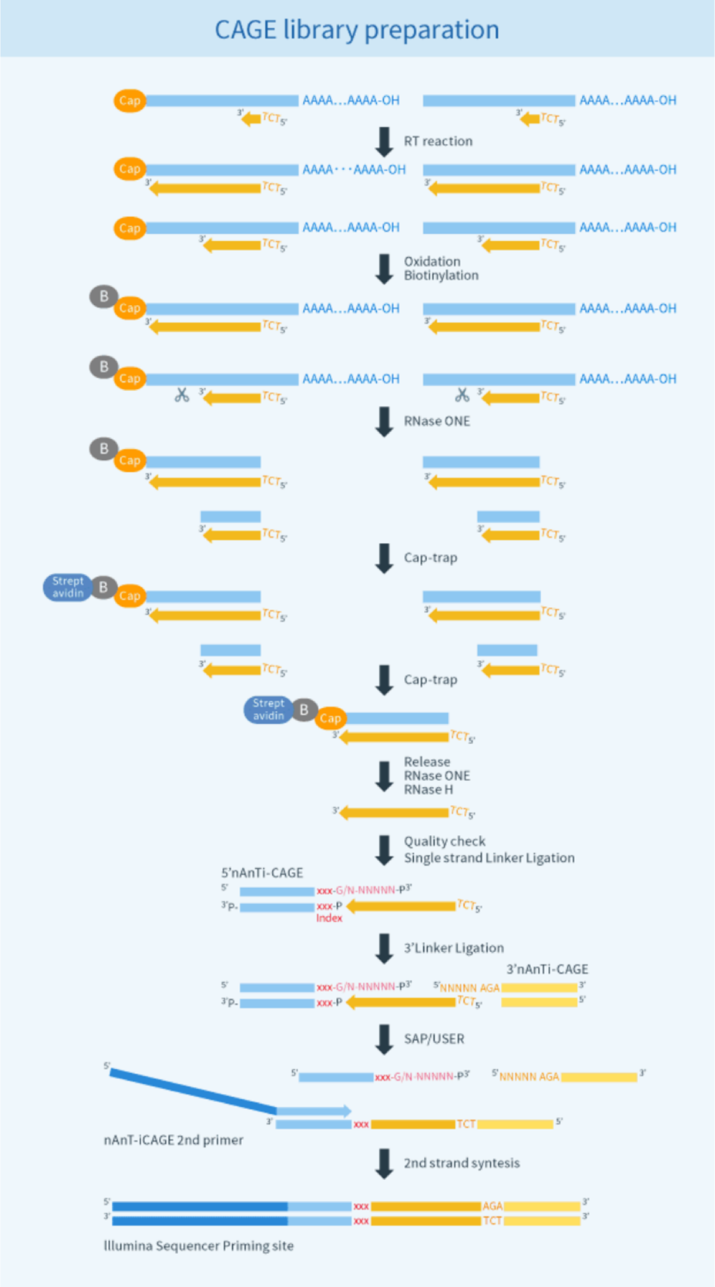
**DNA sequence corresponding
With identified RNAs**

2. RNA Seq variant technology: CAGE (Cap Analysis of Gene Expression)

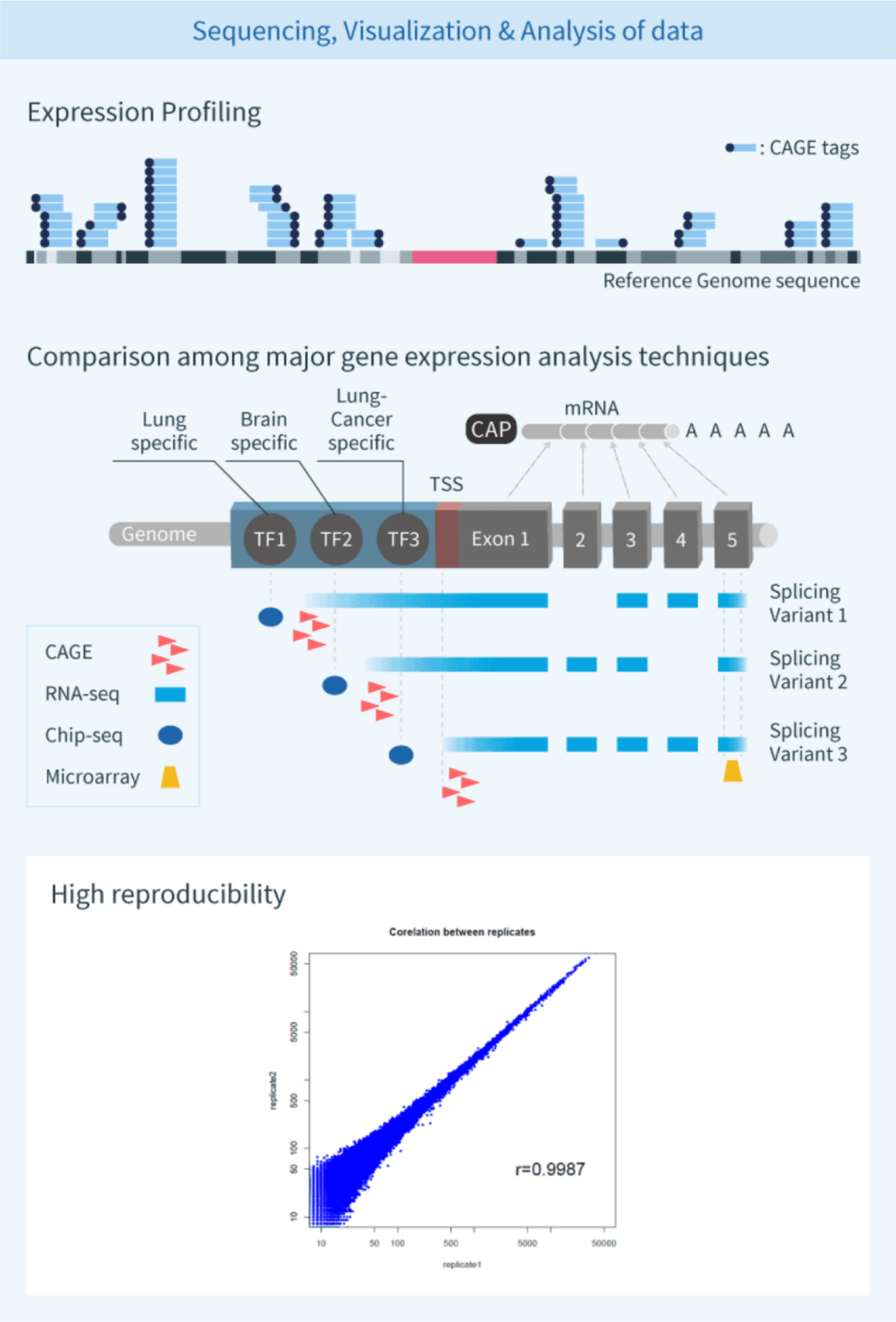
http://www.osc.riken.jp/english/activity/cage/basic/

Unlike a similar technique Serial Analysis of Gene Expression (SAGE, superSAGE) in which tags come from other parts of transcripts, CAGE is primarily used to locate an exact transcription start sites in the genome. This knowledge in turn allows a researcher to investigate promoter structure necessary for gene expression.




CAGE library preparation

2. RNA Seq variant technology: CAGE (Cap Analysis of Gene Expression)



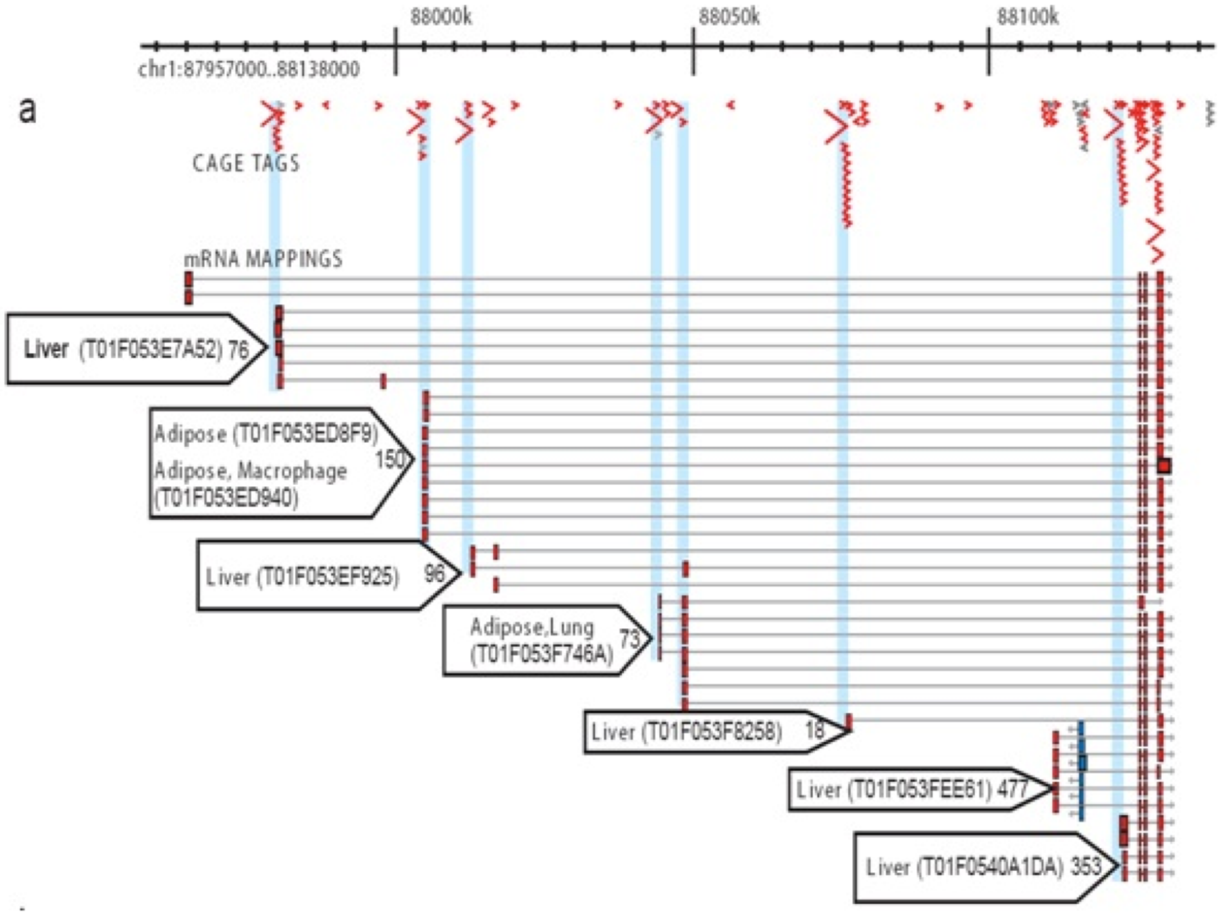
Sequencing, Visualization & Analysis of data

Expression Profiling

Comparison among major gene expression analysis techniques

High reproducibility

2. RNA Seq variant technology: CAGE (Cap Analysis of Gene Expression)



Excellent tool
To identify
transcriptional
start sites

Help to identify up-stream
regulatory sequences =
PROMOTERS RELEVANT CpG