TRASCRITTOMICA

GF 2° year 1° Semester

Schedule lectures– AA 2015/2016

TOPICS OF THE COURSE (3 CFU)

1. **A non-coding RNA revolution; identification of ncRNA elements**

2. **Pseudogenes and gene regulation**

3. **miRNA regulatory pathways**

4. **competitors of endogenouse RNAs**

5. **DNA Damage Repair RNAs (siRNAs and DNA damage repair)**

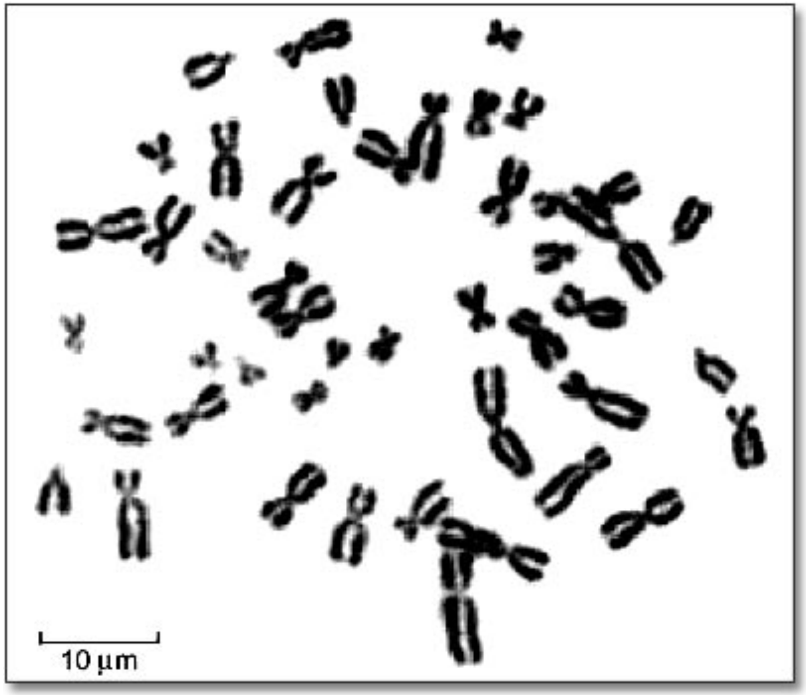
6. **Promoter and enhancer regulating ncRNAs**

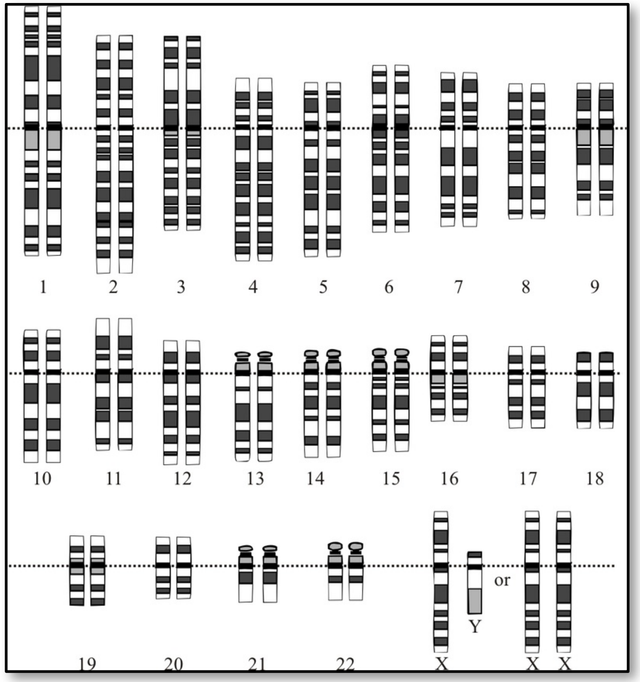
7. **RNA editing**

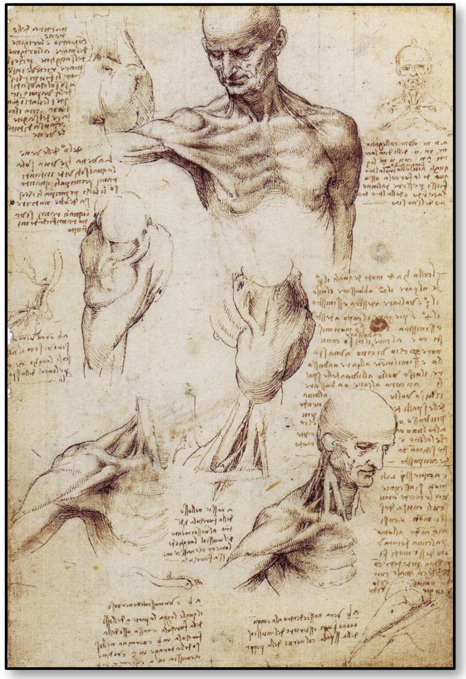
9. **RNA function in cis: RNA:DNA hybrids in Disease**

10. **RNA-Protein bodies – Cajal bodies, Paraspeckles**

The human genome is highly structured

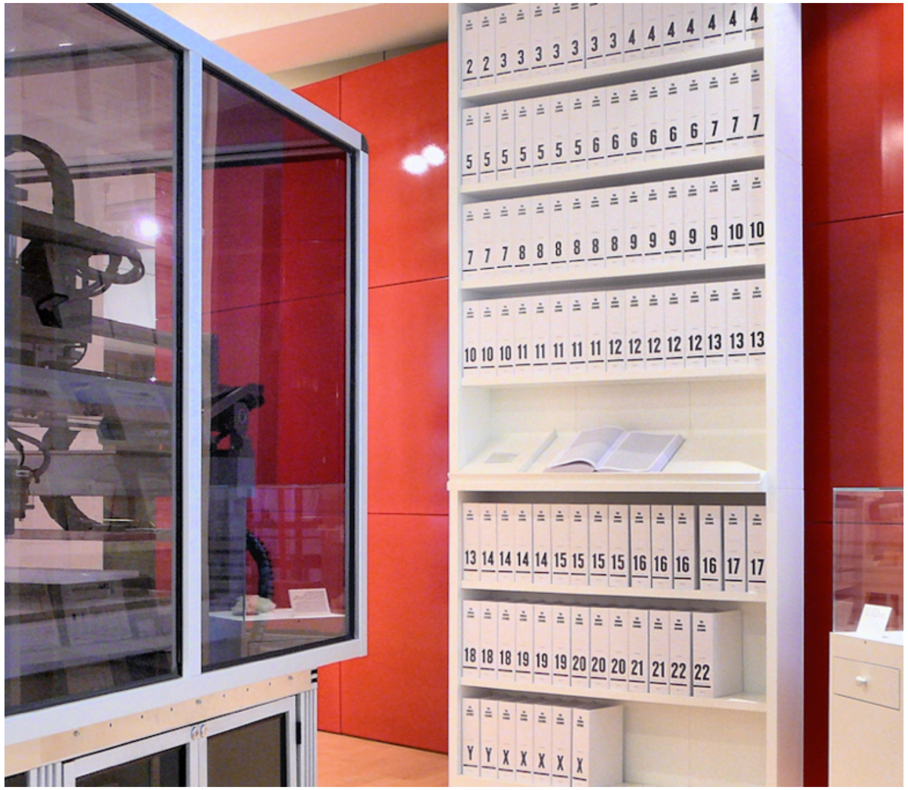


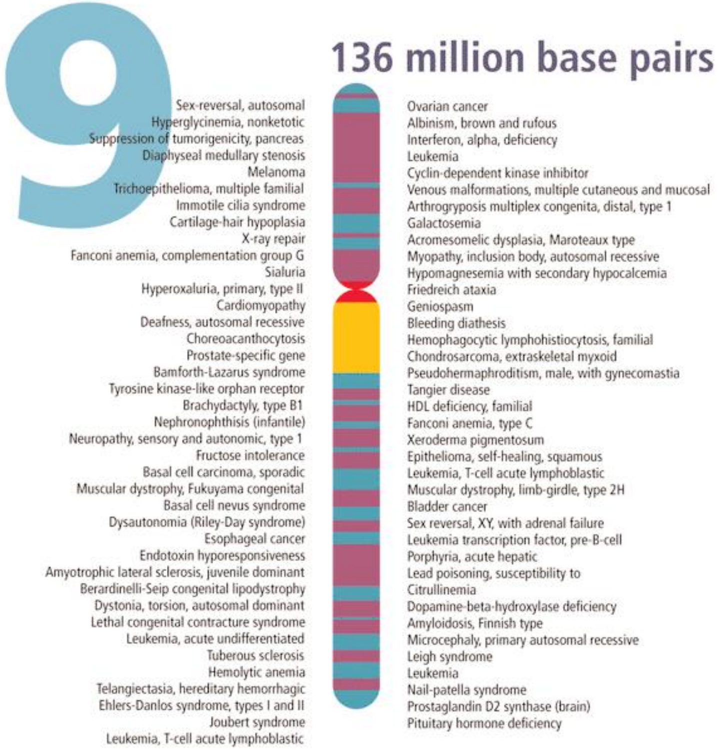




The human genome:
22 autosome paires
2 Sex chromosome pairs (XX o XY)
Total haploid genome 3×10^9

The human genome is highly structured





Genoma umano aploide: 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;
→ 2.9 x 10^9 bp: haploid chromosome set

Usage of genetic information:

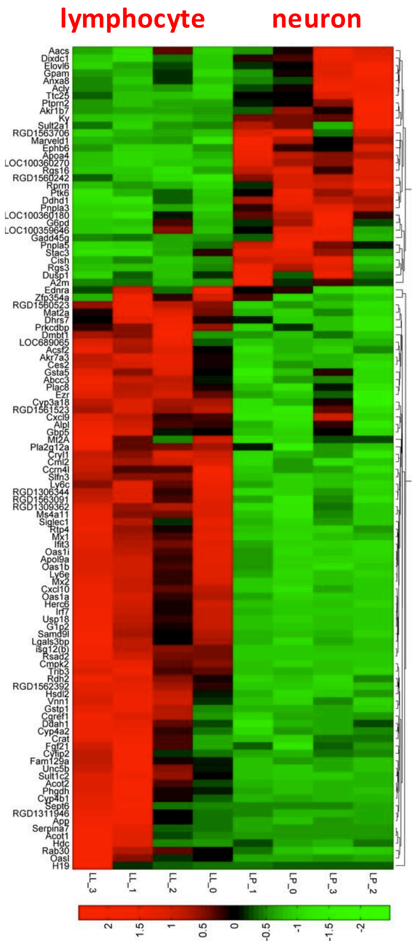
5.000-10.000 geni espressi da ogni cellula
- 100.000 different proteins (post-translational modifactions per cell)
- 10^8 total protein spcecies

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

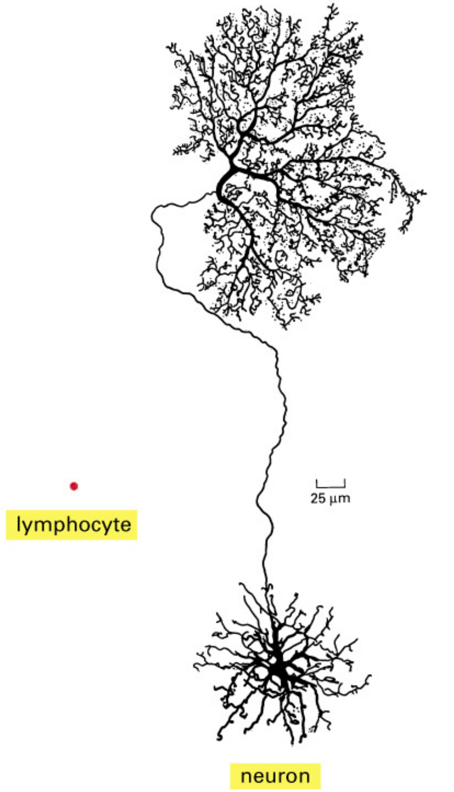


Figure 7–1. Molecular Biology of the Cell, 4th Edition.

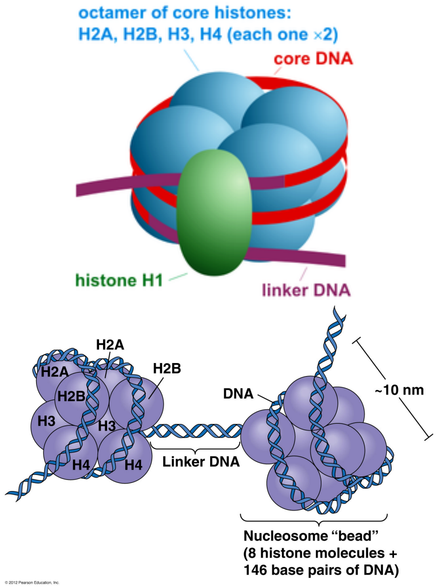
Genetic information must be highly organized

The human genome is highly structured

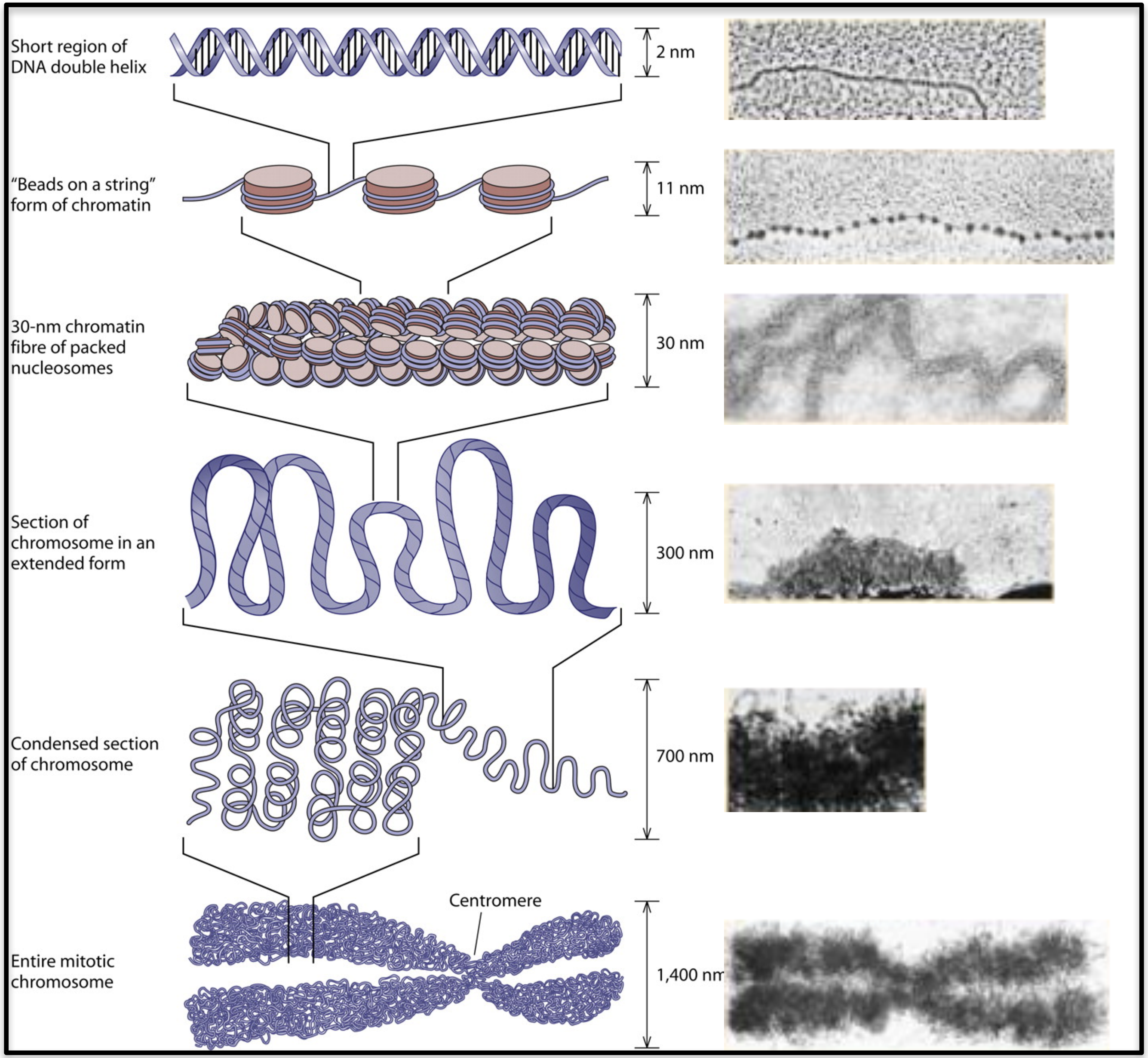
Chromatin: DNA + protein in nucleus
Organisation of genetic information

Function:

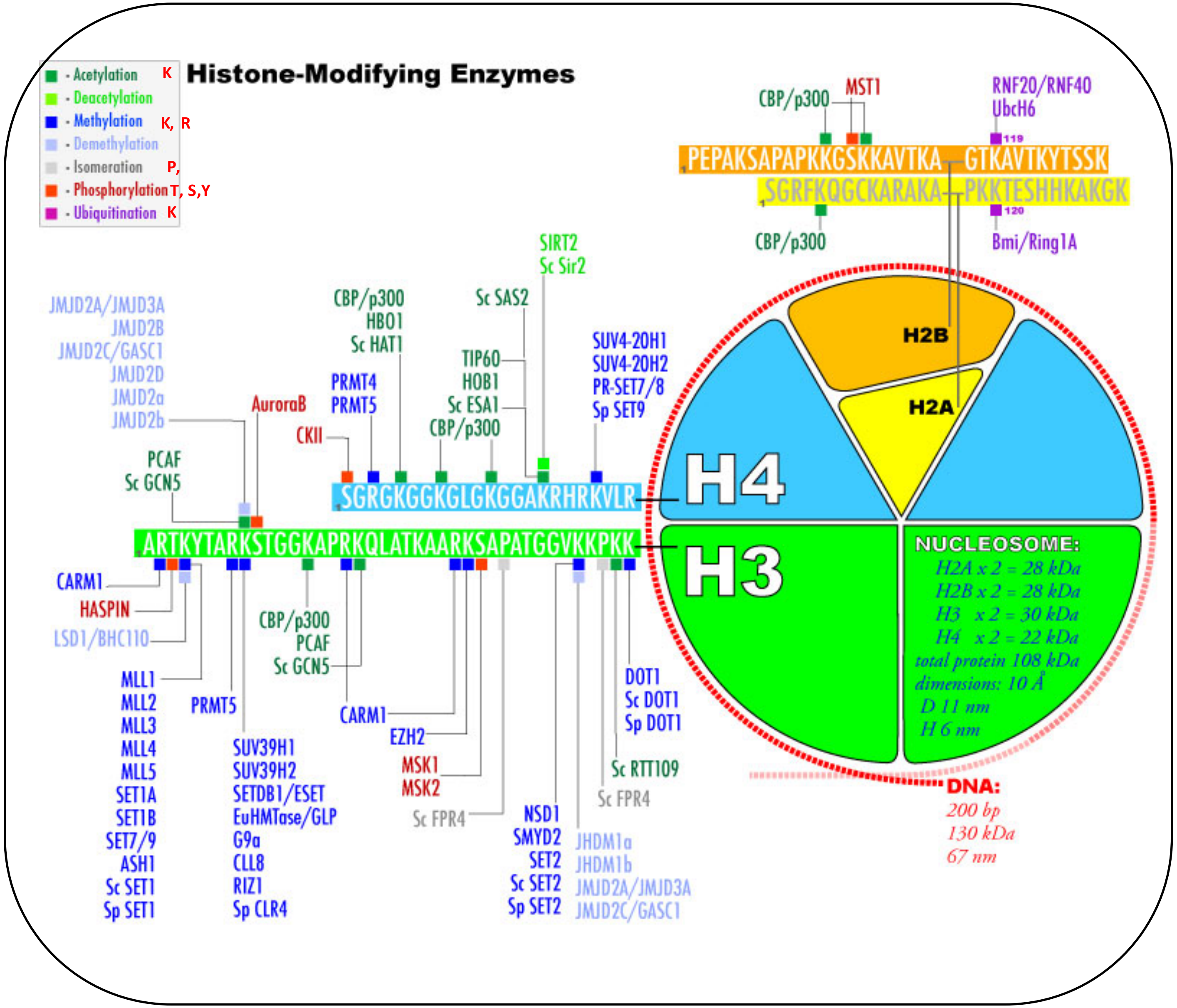
Packaging of DNA
Compaction of DNA
Definition of reagions of gene
Expression (euchromatin) or repression (heterochromatin)
-Increasing stability of DNA
-Prevention of damage
-Control of replication, gene expression
-Cell cycle



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



POST-TRANSLATIONAL HISTONE MODIFICATIONS



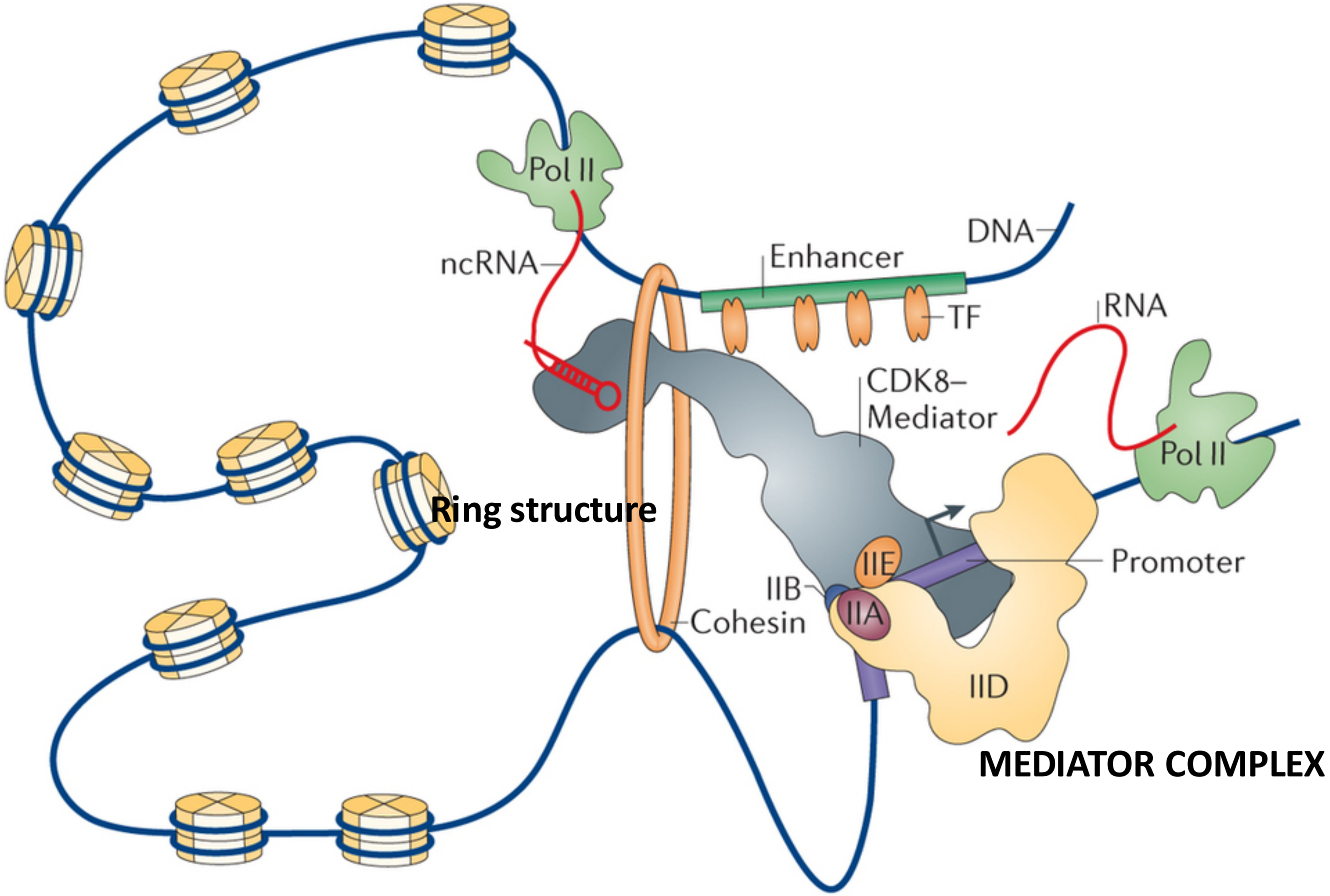
Gene expression
Control by post-
translational
histone modifications

→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

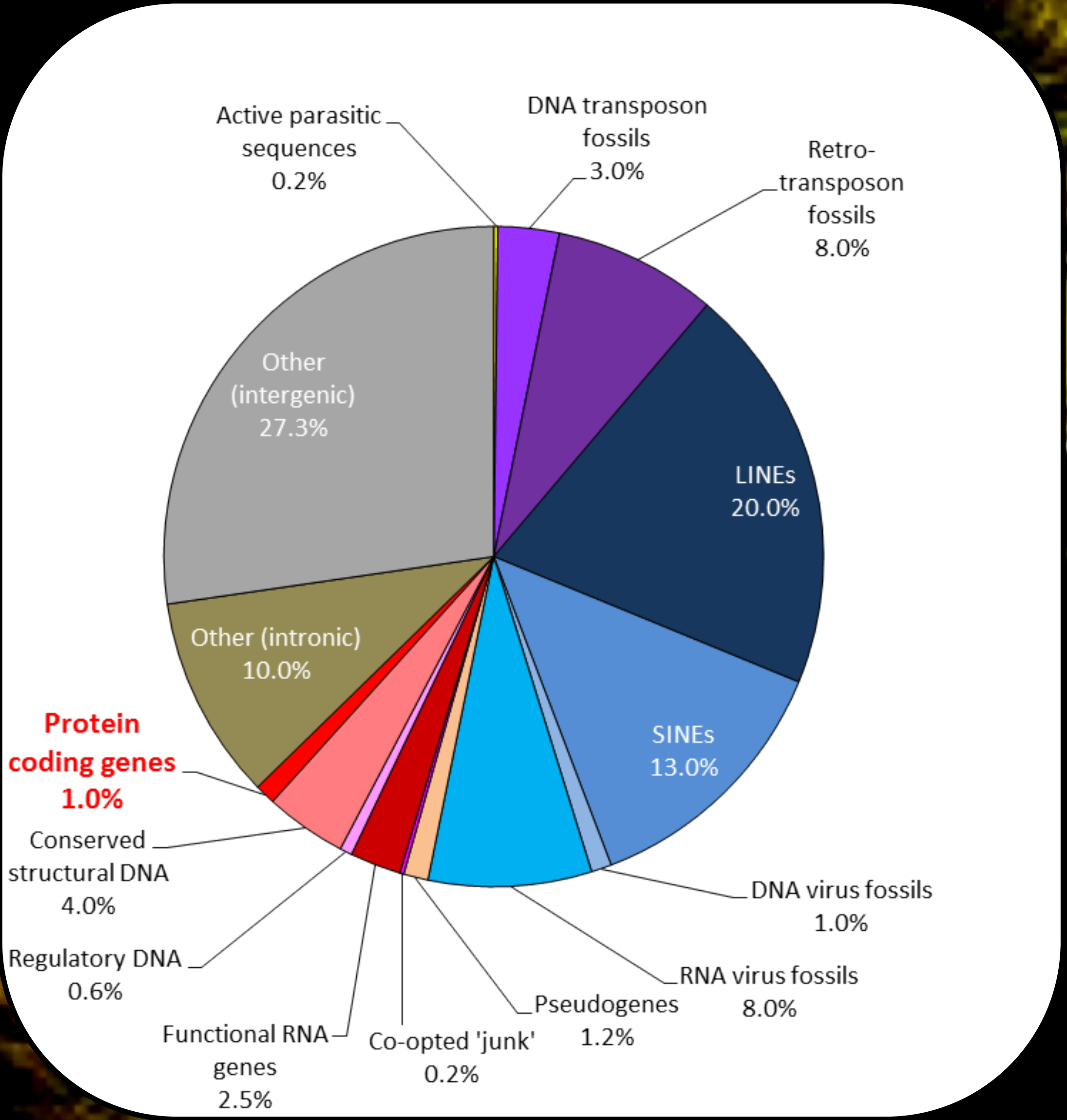
→ SPECIFIC 3 DIMENTSIONAL
STRUCTURE

98% OF GENOMIC DNA DOES NOT ENCODE FOR PROTEINS

ca 50% transposable elements

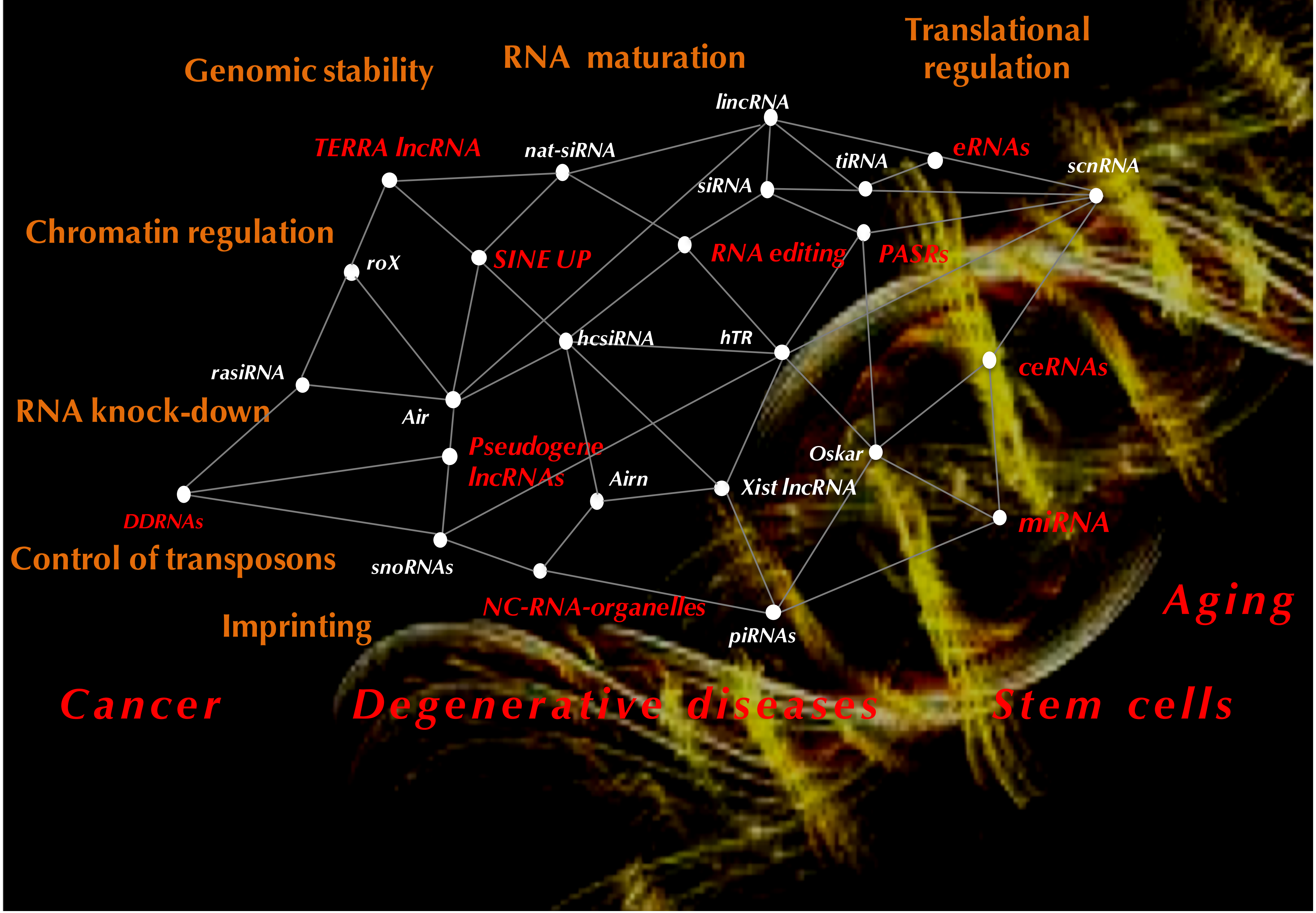
1-2% protein coding genes

0.5-1% pseudogenes



Almost all genomic sequences are subjected to transcription

Why to study ncRNAs



Genomic stability

RNA maturation

Translational regulation

Chromatin regulation

RNA knock-down

Control of transposons

Imprinting

lincRNA

TERRA lncRNA

nat-siRNA

tiRNA

eRNAs

scnRNA

siRNA

roX

SINE UP

RNA editing

PASRs

hcsiRNA

hTR

rasiRNA

Air

ceRNAs

Pseudogene lncRNAs

Oskar

Airn

Xist lncRNA

DDRNAs

miRNA

snoRNAs

NC-RNA-organelles

piRNAs

Aging

Cancer

Degenerative diseases

Stem cells

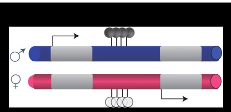
Why to study ncRNAs
1. There are things proteins cannot do

AGING-CANCER

Telomere function

hTR lncRNA

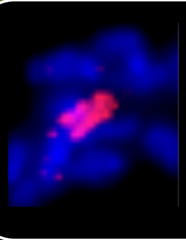
DEVELOPMENTAL DEFECTS



mono-allelic
Gene expression

Airn lncRNA

Xist lncRNA

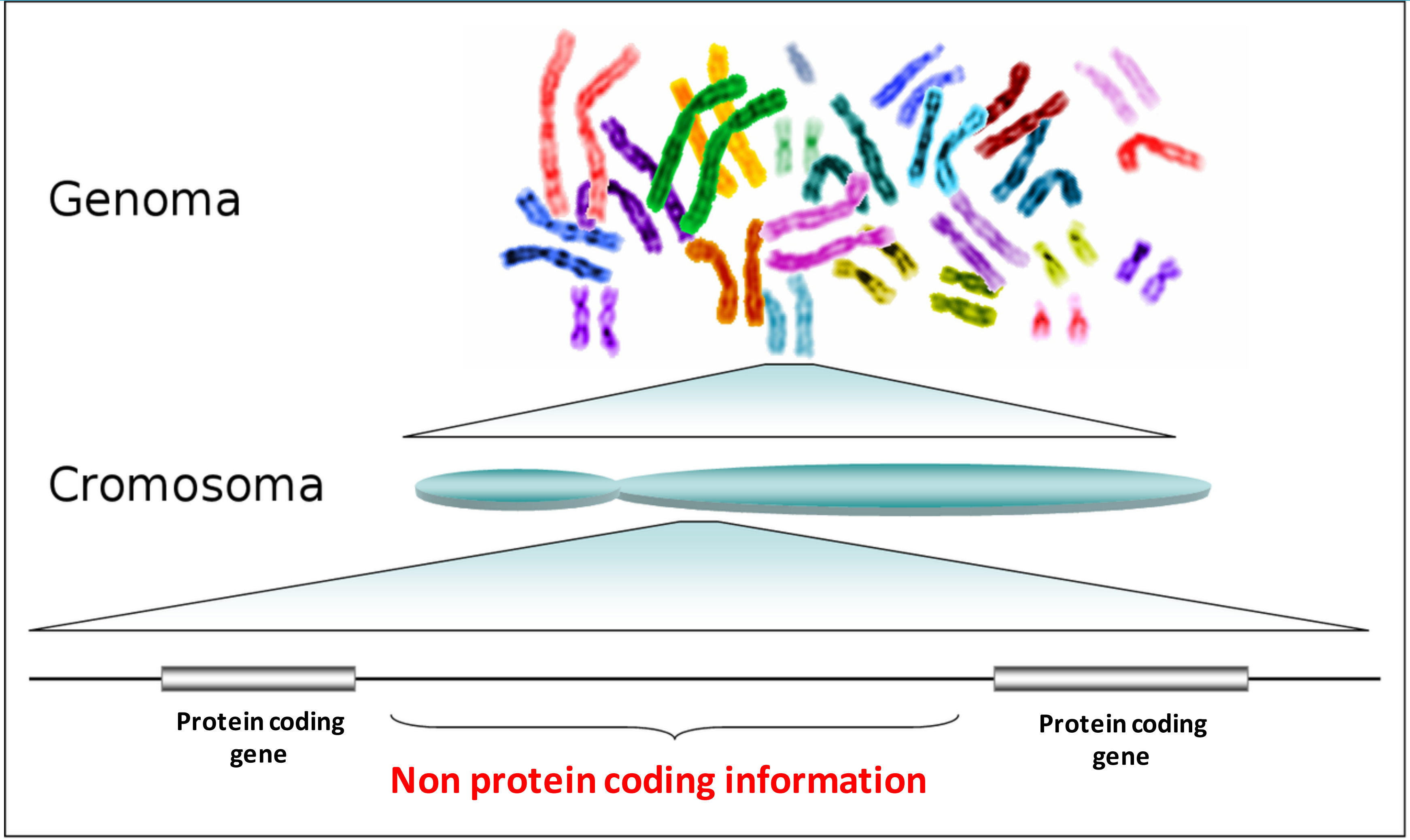


X inactivation
Silencing
1000 genes
on the Xi

LETHAL

2. they have high relevance for development and pathology

The human genome is highly structured

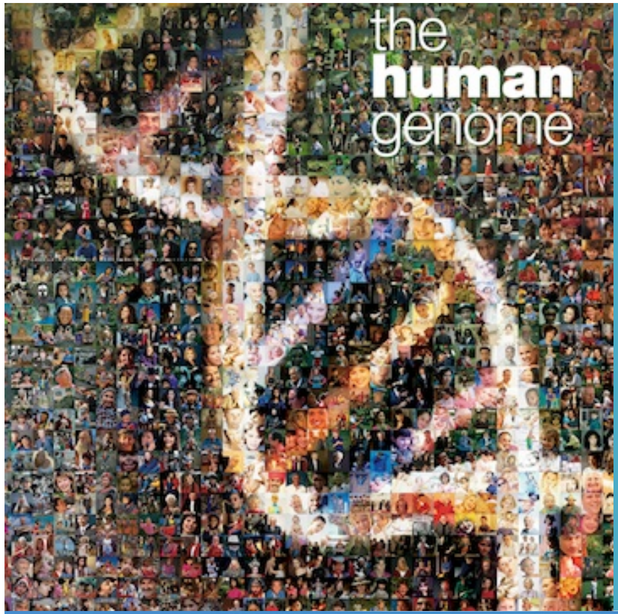

Genoma

Cromosoma

Protein coding gene

Protein coding gene

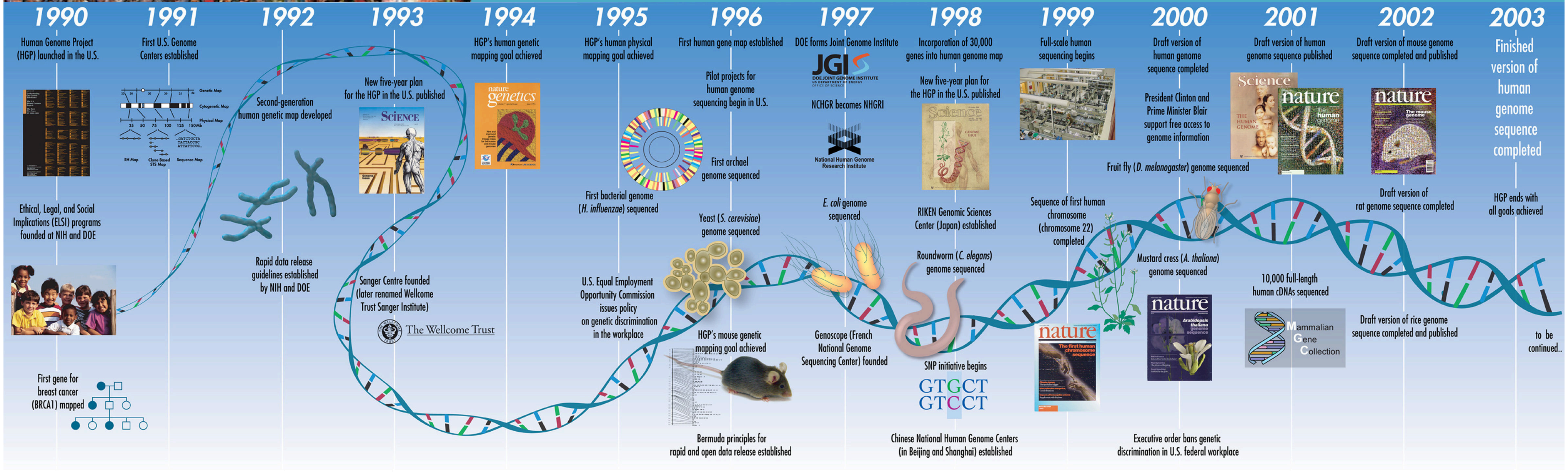
Non protein coding information



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

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
Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)

1994
HGP's human genetic mapping goal achieved

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
NCHGR becomes NHGRI
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
RIKEN Genomic Sciences Center (Japan) established
Roundworm (*C. elegans*) genome sequenced
SNP initiative begins
Chinese National Human Genome Centers (in Beijing and Shanghai) established

1999
Full-scale human sequencing begins
Sequence of first human chromosome (chromosome 22) completed

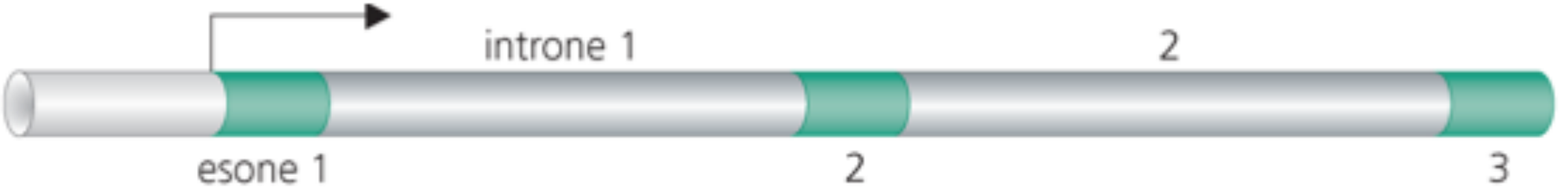
2000
Draft version of human genome sequence completed
President Clinton and Prime Minister Blair support free access to genome information
Fruit fly (*D. melanogaster*) genome sequenced
Mustard cress (*A. thaliana*) genome sequenced
Executive order bans genetic discrimination in U.S. federal workplace

2001
Draft version of human genome sequence published
10,000 full-length human cDNAs sequenced

2002
Draft version of mouse genome sequence completed and published
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:

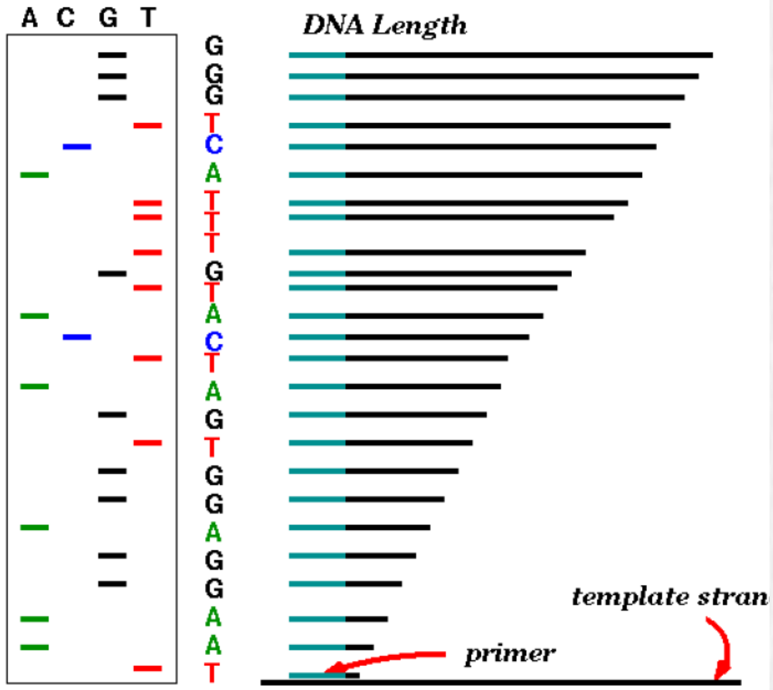
Gel electrophoresis: discrimination of 1 bp: size range below 300 bp in the lab

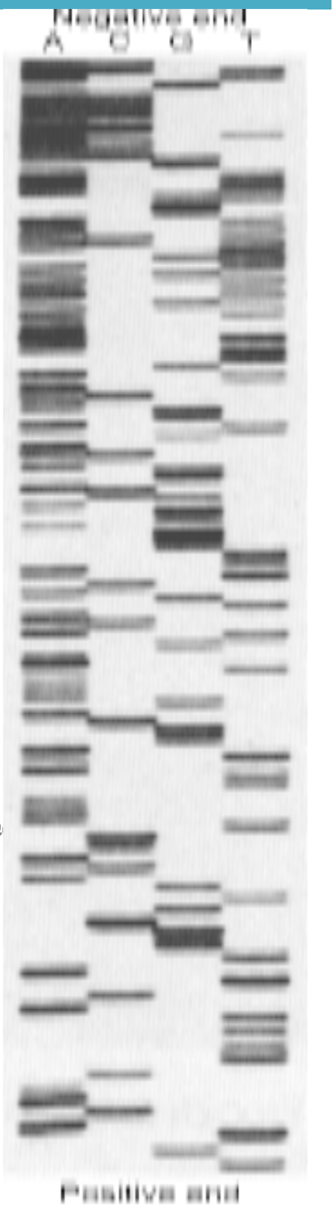
DNA template + 32P-labelled sequencing oligo

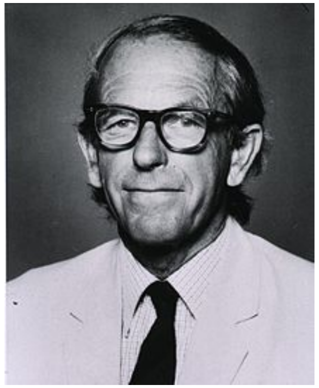
4 parallel seqeuncing reactions:

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:

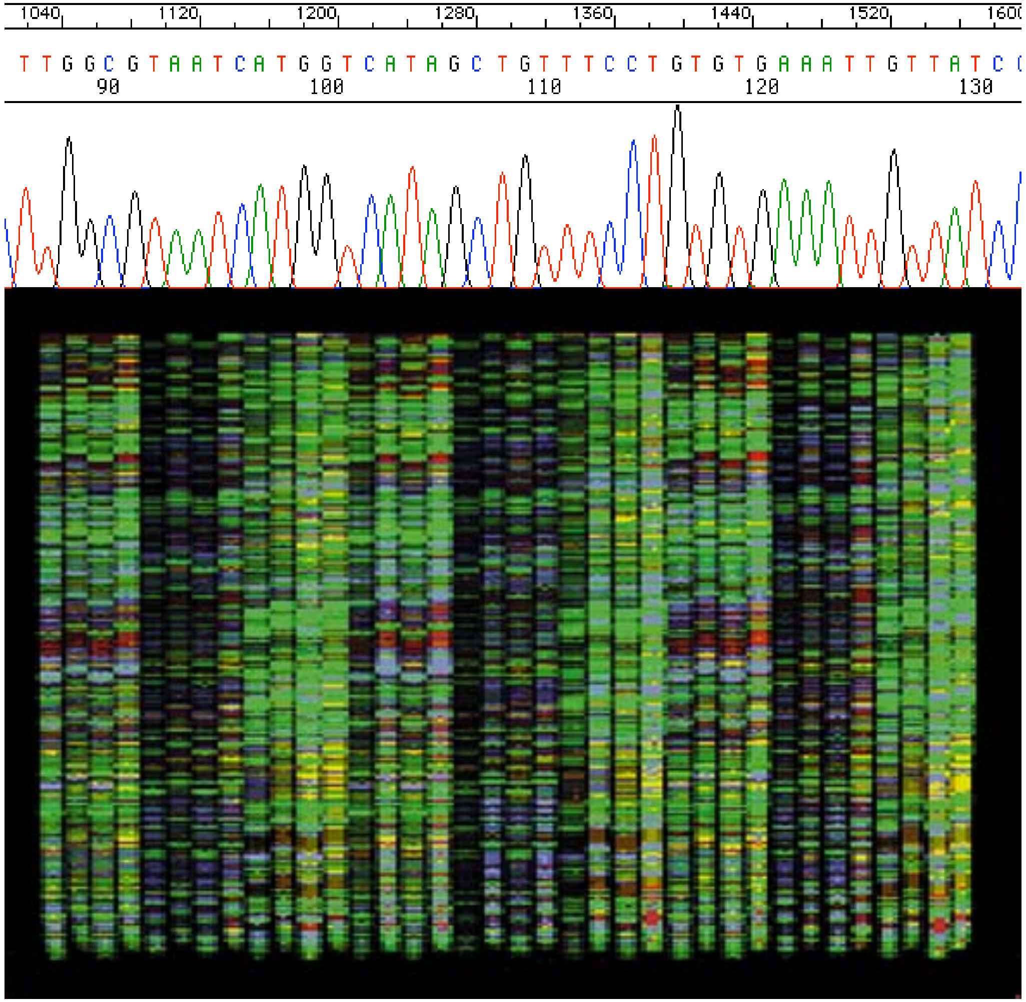
Gel electrophoresis: discrimination of 1 bp: size range below ~1000 bp

DNA template + sequencing oligo

1 seqeuncing reaction:

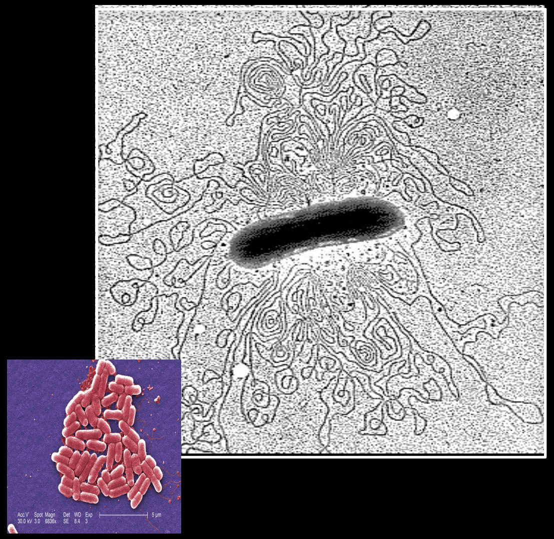
1. dATP, dCTP, dGTP, dTTP + ddATP-Dye1, ddCTP-Dye2, + ddGTP-Dye3+ddTTP-Dye4 (low conc)

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



THE NUMBER OF PROTEIN CODING GENES IS RELATVLY LOW

E.coli



C. elegans

H. sapiens

Genome	5×10^6 bp	1×10^8 bp	3×10^9 bp
Chromosomes	1	6	23
Coding genes	6692	20541	21995
ncDNA			
non-coding RNA genes			
miRNAs			
pseudogenes			

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

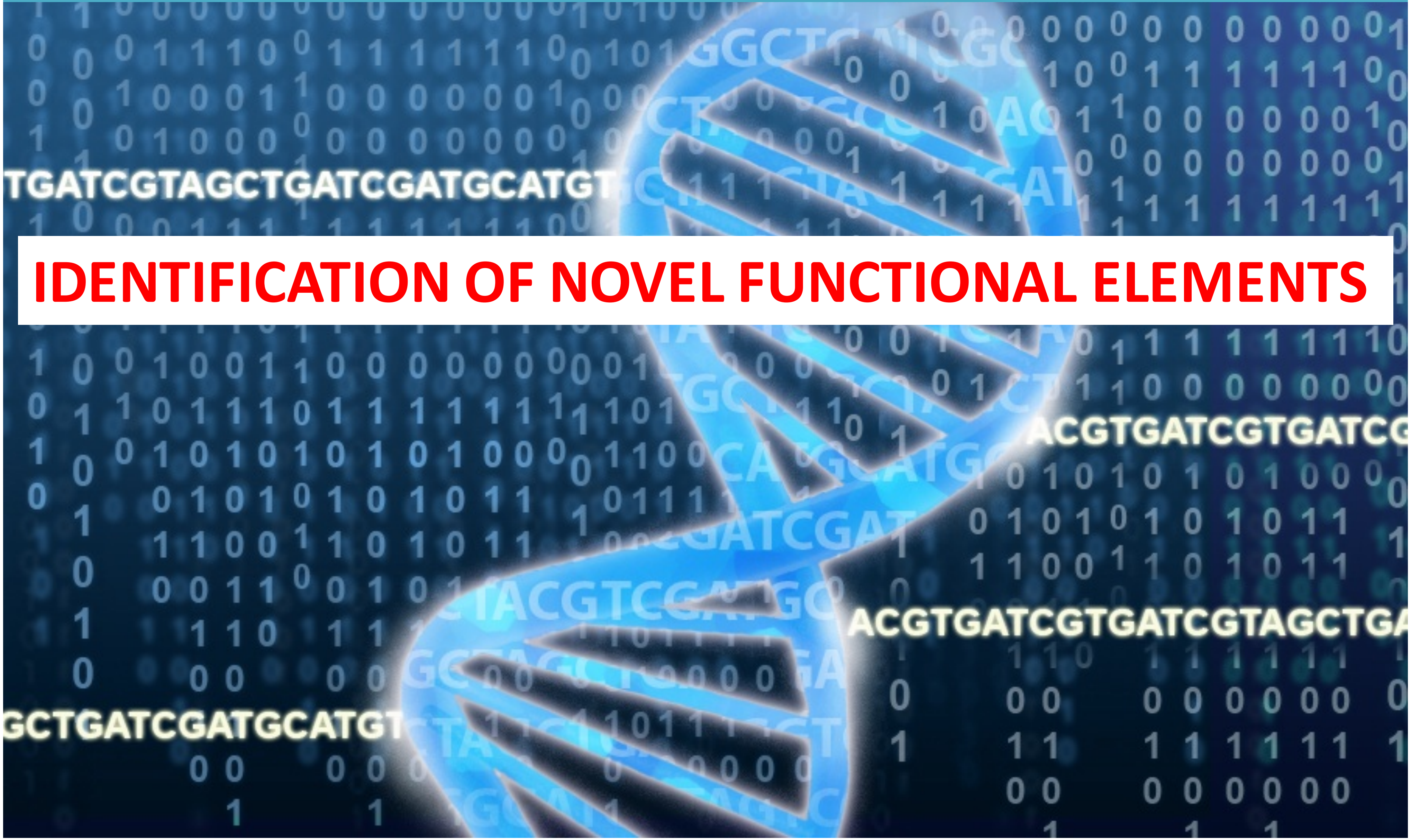
ENSEMBL 11/2014

WHAT INFORMATION INCREASES ORGNAISMAL COMPLEXITY

ncDNA derived information?

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

NEXT GENERATION SEQEUNCING OF DNA AND RNA



IDENTIFICATION OF NOVEL FUNCTIONAL ELEMENTS

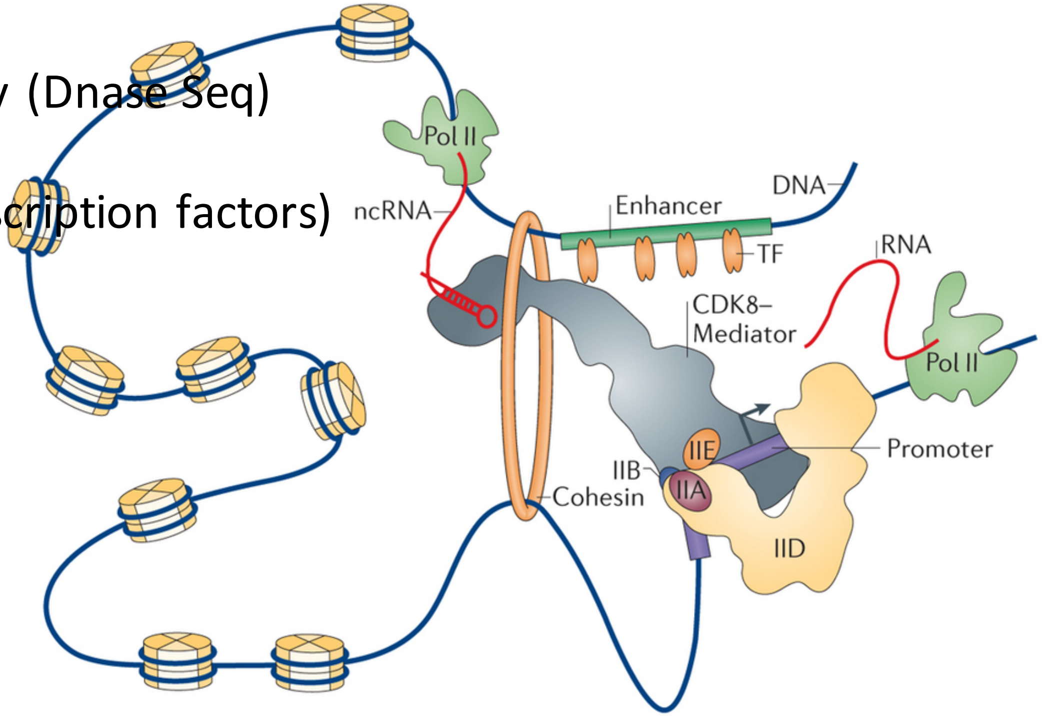
NEXT GENERATION SEQEUNCING OF DNA AND RNA

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

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

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:

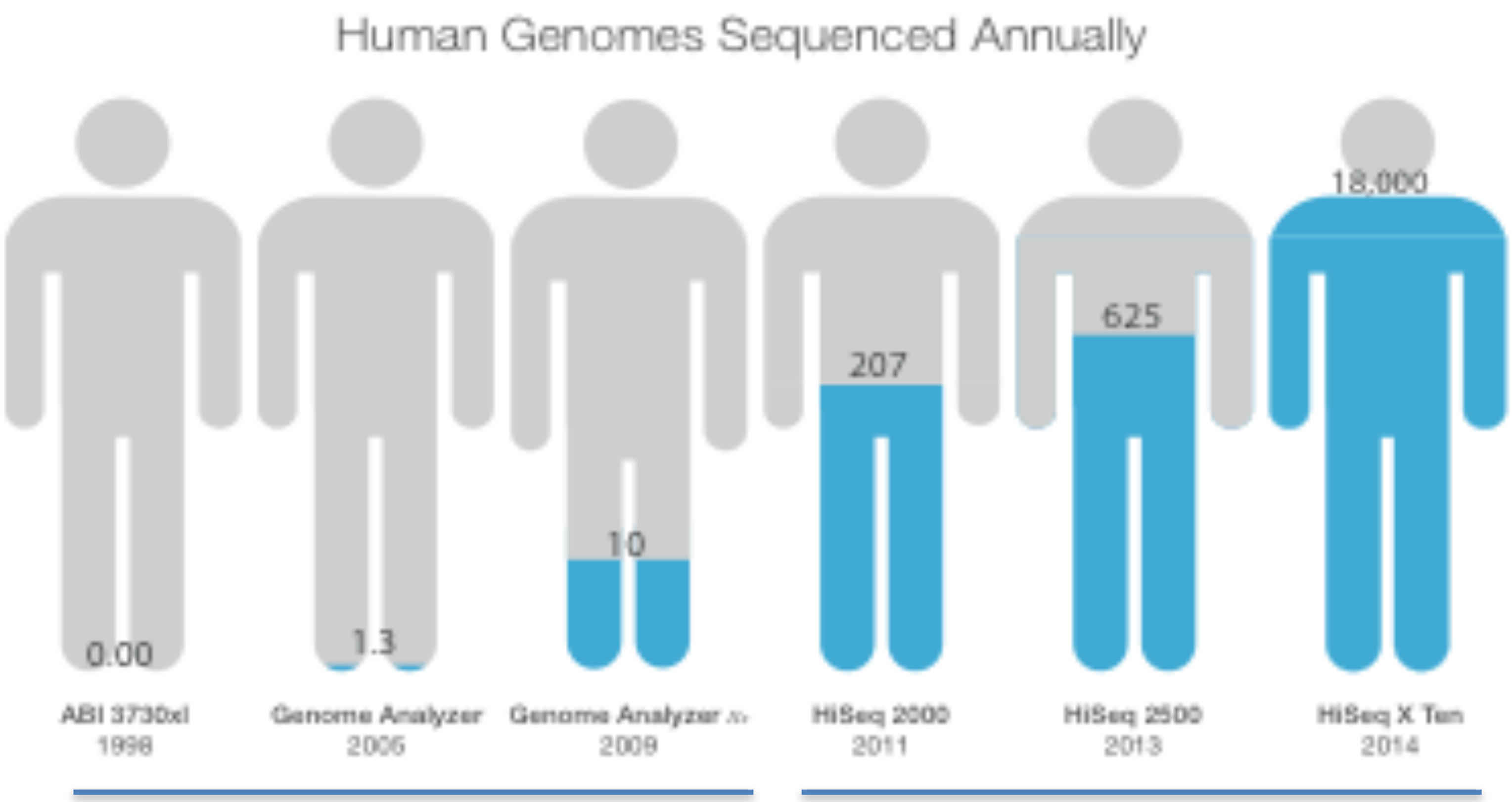
- determination of DNAseI hypersensitivity (Dnase Seq)
- nucelosome occupancy (MNase-seq)
- ChIP-seq (chromatin modifications, transcription factors)
- 3 Dimensional space interaction



Nature Reviews | Molecular Cell Biology

1990: TO UNDERSTAND LIFE WE NEED TO IDENTIFY ALL RELEVANT GENETIC INFORMATION → LETS SEQEUNCE THE GENOME

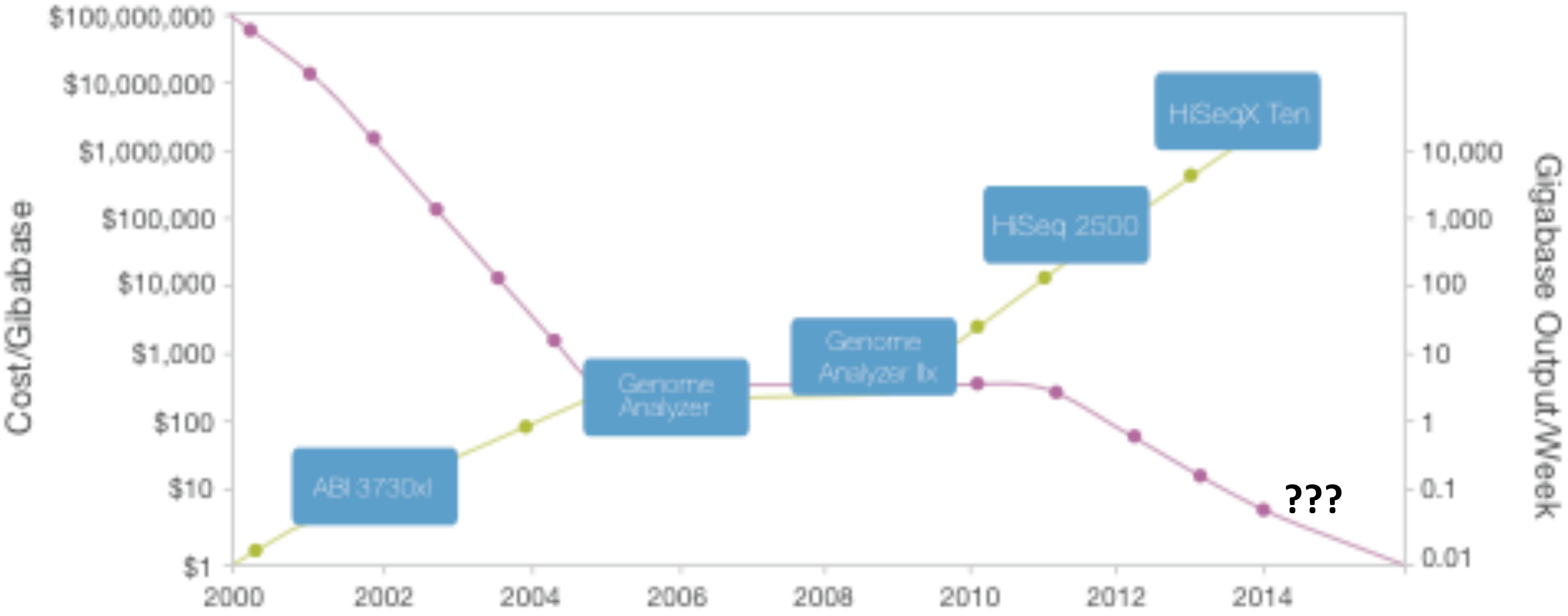
2003: HUMAN GENOME SEQUENCED

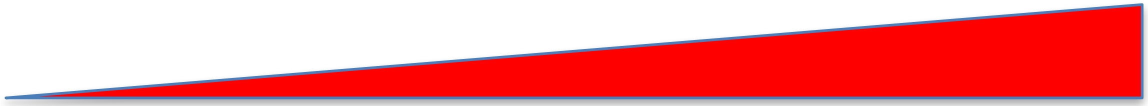


”classic sequencing”

”NGS sequencing”

PROGRESS IN SEQUENCING POWER





BIOINFORMATICS EFFORT
= PROCESING OF DATA

Next generation sequencing:

MASSIVE PARALLEL SEQUENCING (ILLUMINA)

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

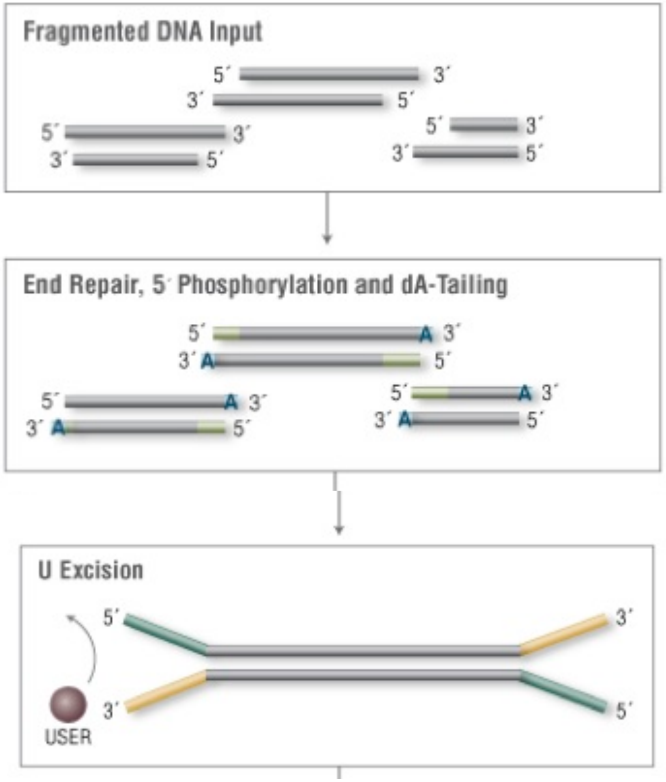
Generation of DNA libraries:

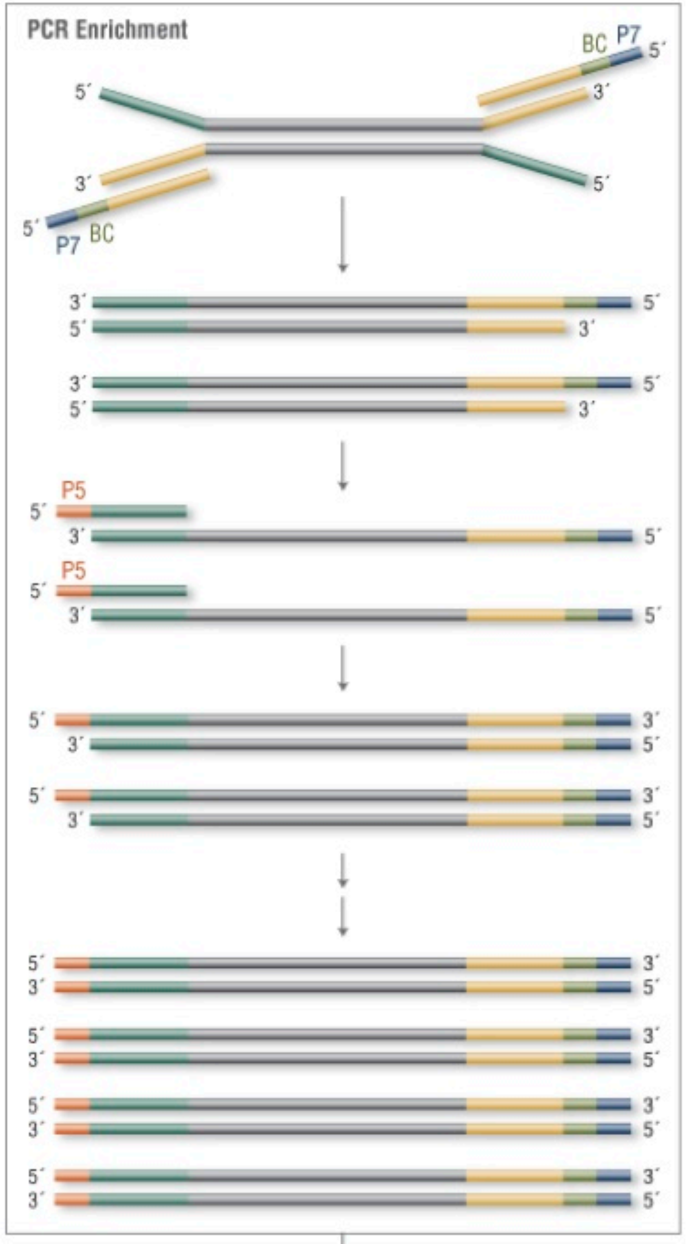
Application:

ChIP Seq

Genome Seq

Methyl Seq





Illumina: massive parallel sequencing: Genomic DNA

Generation of RNA libraries:

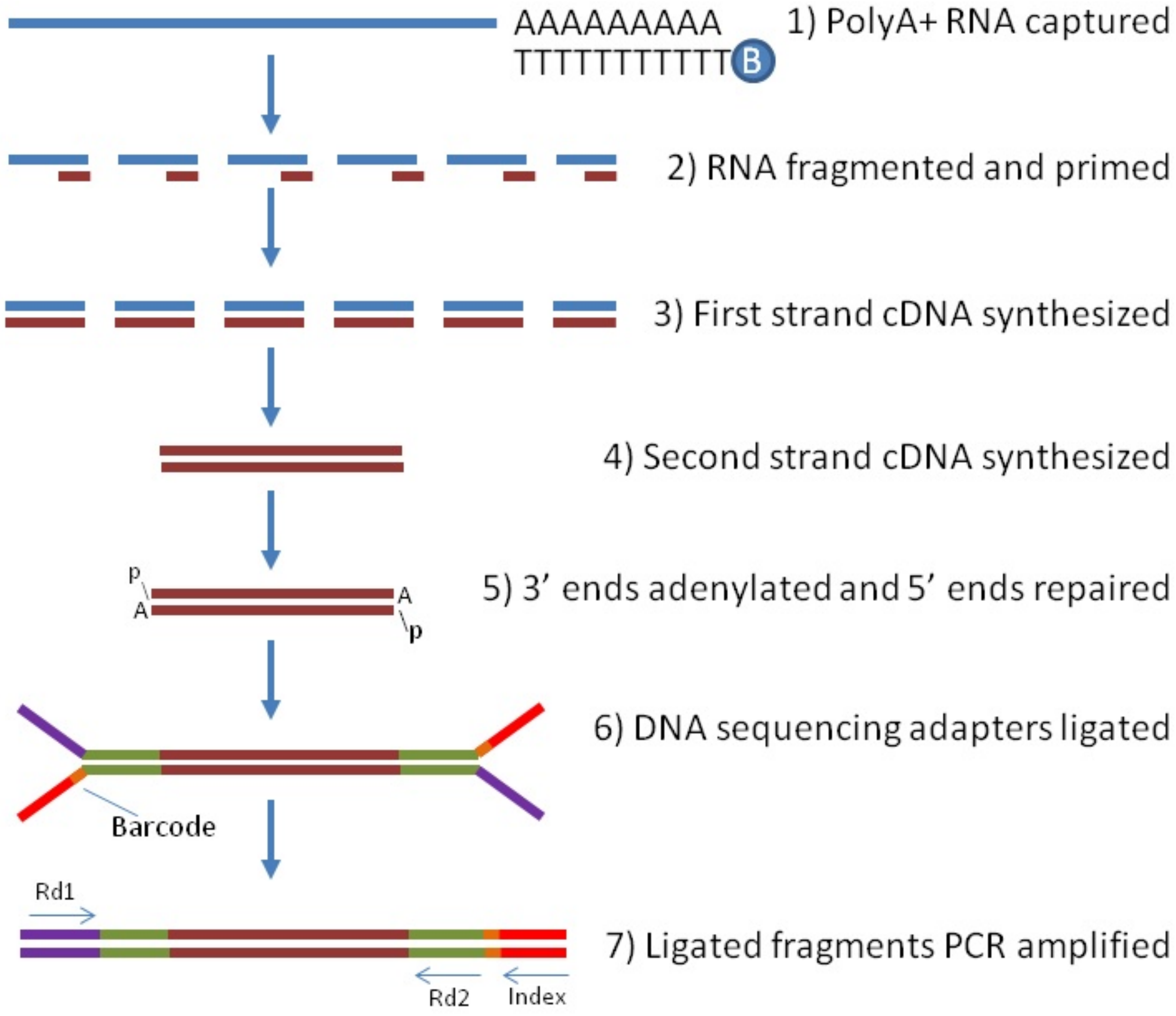
Application:

RNA Seq

Exon Seq

Important:

Involves cDNA synthesis



1) PolyA+ RNA captured

2) RNA fragmented and primed

3) First strand cDNA synthesized

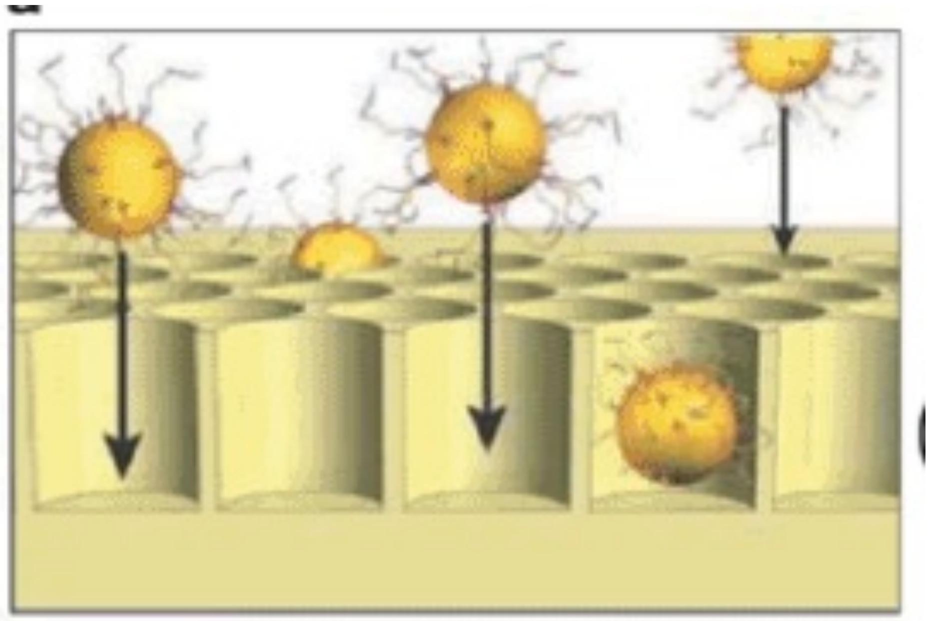
4) Second strand cDNA synthesized

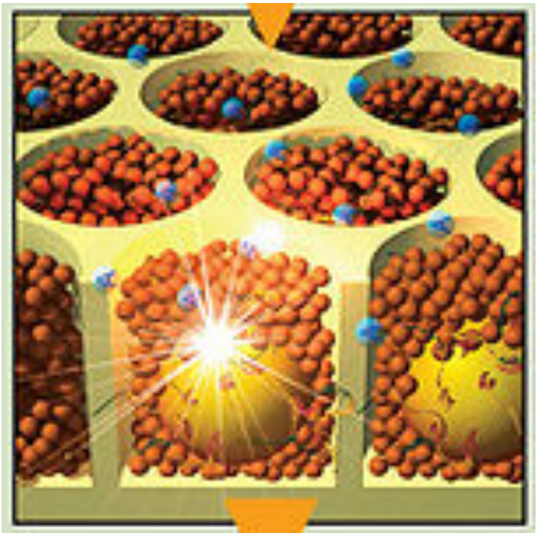
5) 3’ ends adenylated and 5’ ends repaired

6) DNA sequencing adapters ligated

7) Ligated fragments PCR amplified

CLUSTER AMPLIFICATION:





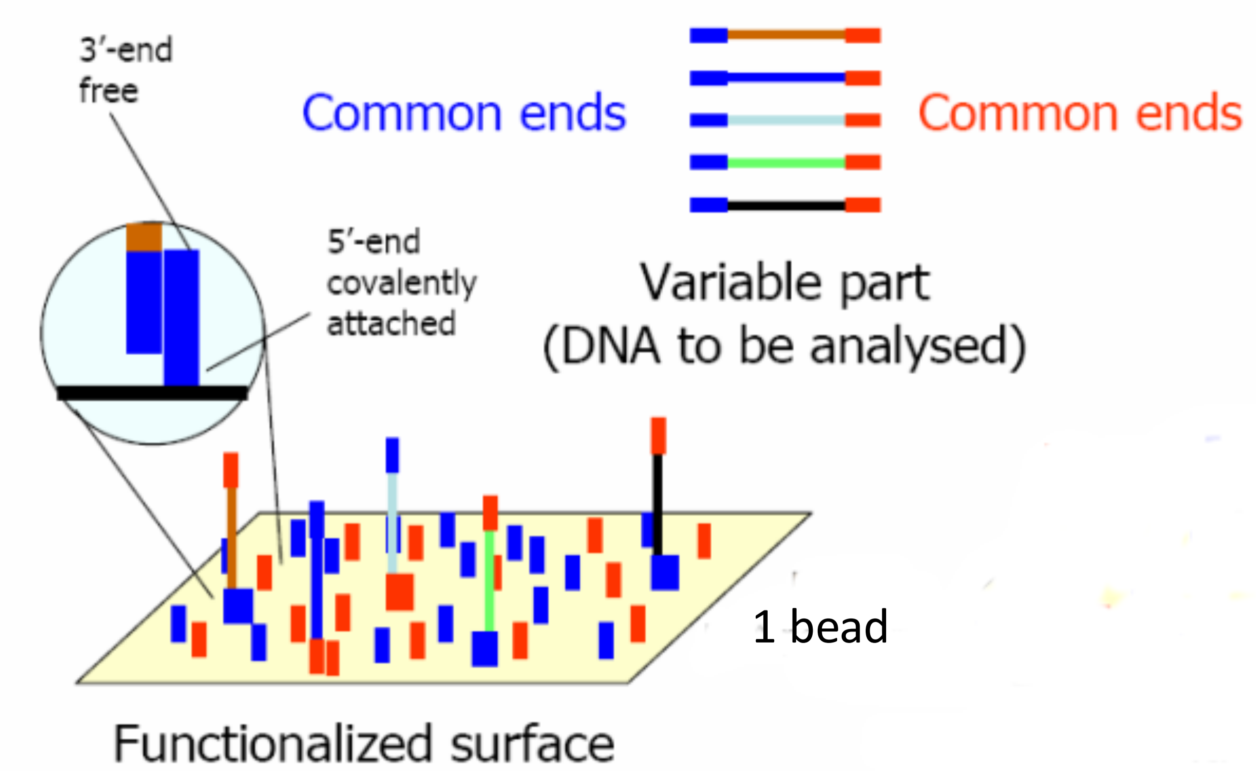
Flow cell contains surface with millions of wells

→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
(1 molecule per bead)

Illumina: massive parallel sequencing:

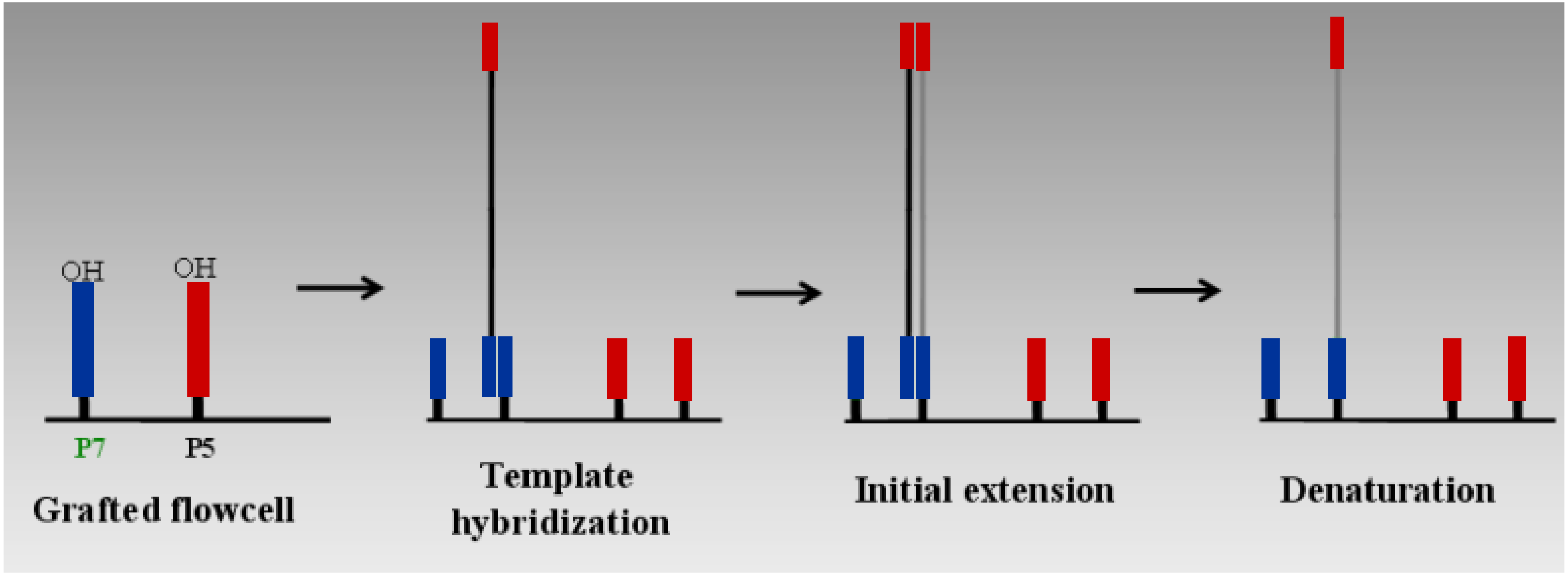
-making DNA library (~300bp fragments)
-ligation of adapters **A** and **B** to the fragments



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

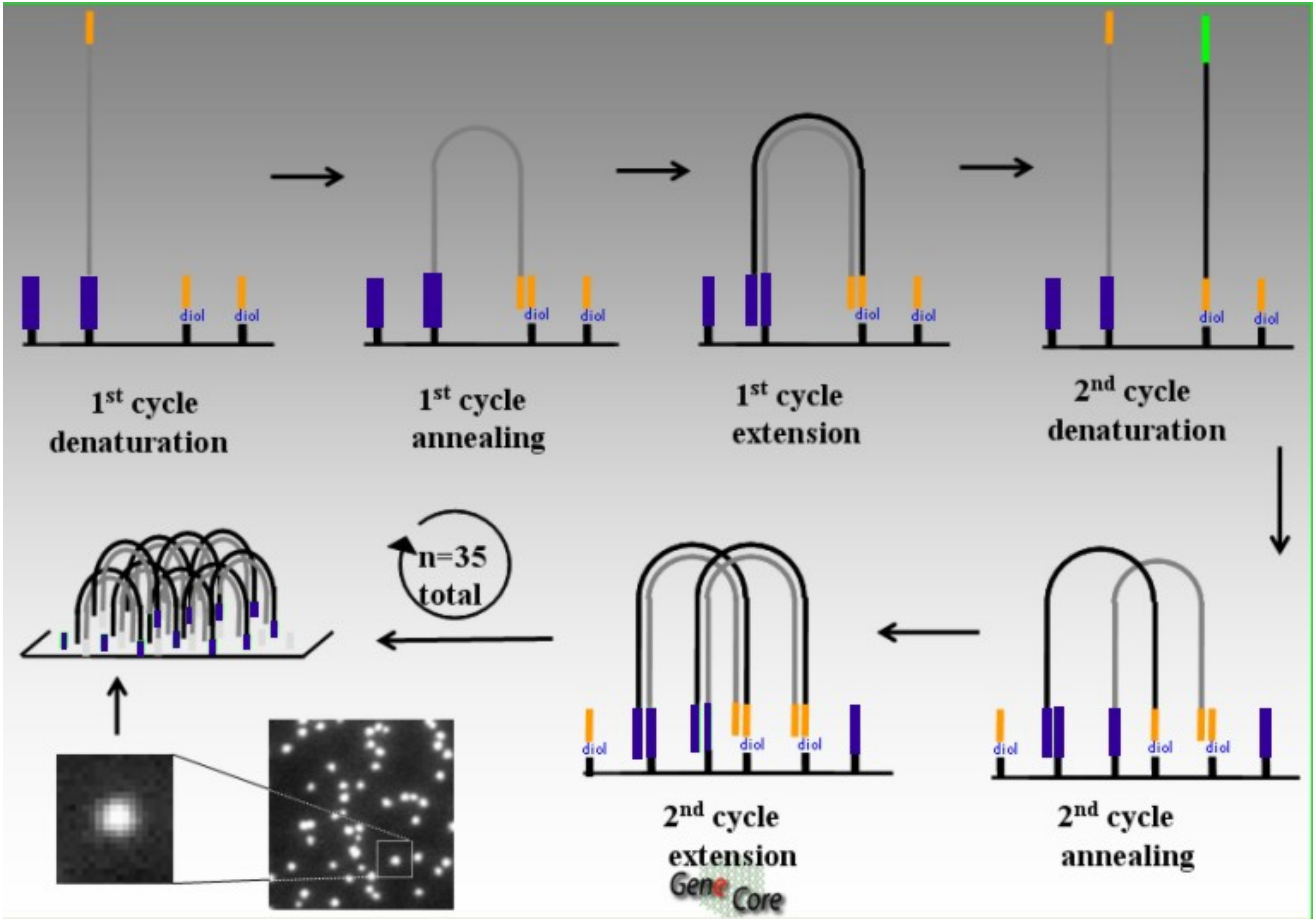
Illumina: massive parallel sequencing:

Bridge amplification:
initiation



On the surface: complementary oligos

Illumina: massive parallel sequencing:

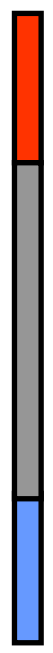


EMBL Gene Core

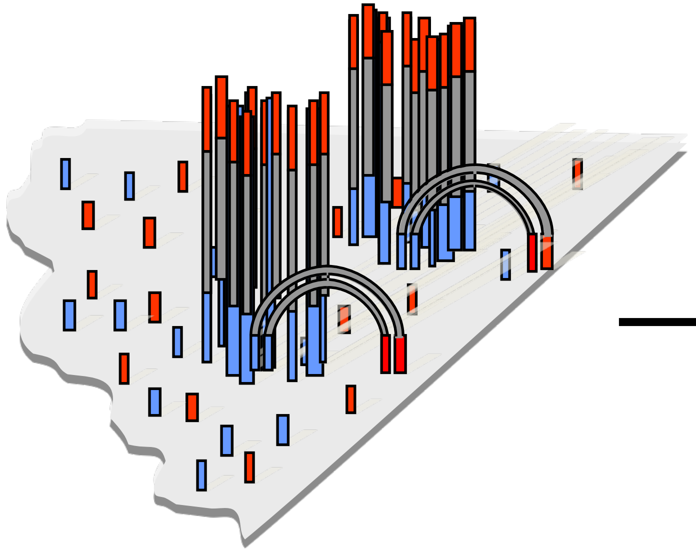
Illumina Sequencing Technology

Robust Reversible Terminator Chemistry Foundation

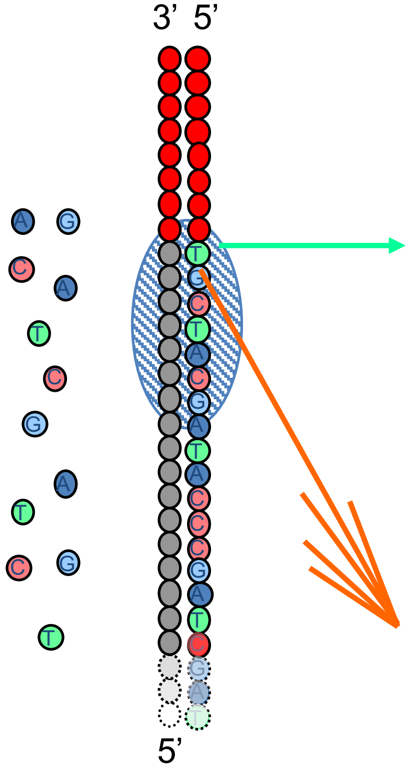
DNA
(0.1-1.0 ug)



Sample
preparation



Cluster growth



Sequencing

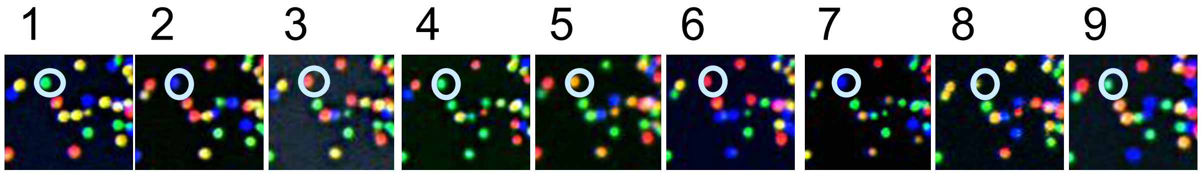
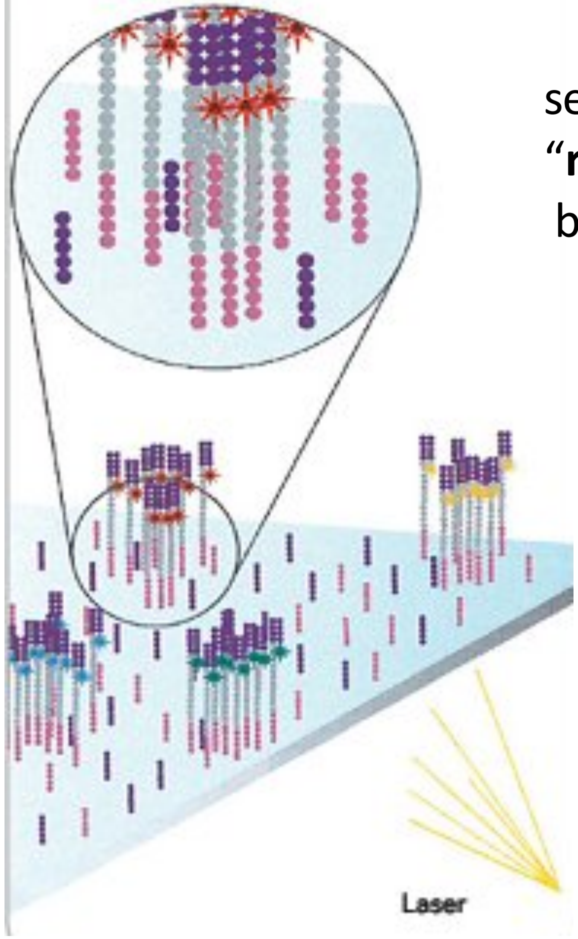


Image acquisition

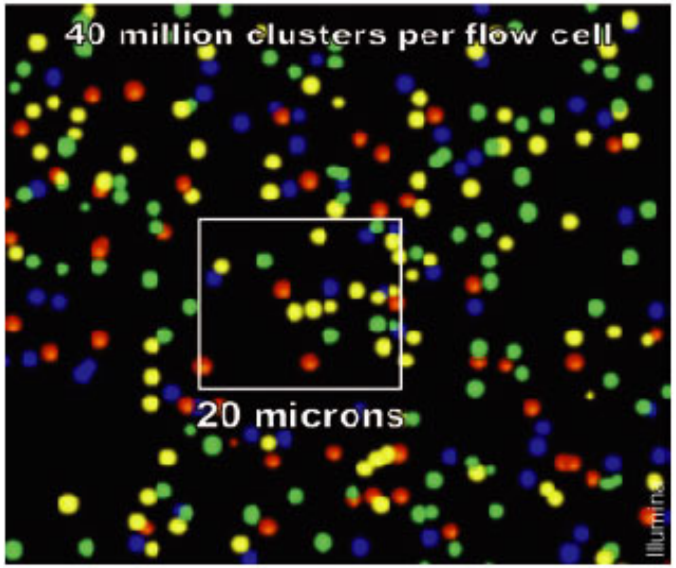
T G C T A C G A T …

Base calling

Illumina: massive parallel sequencing:



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

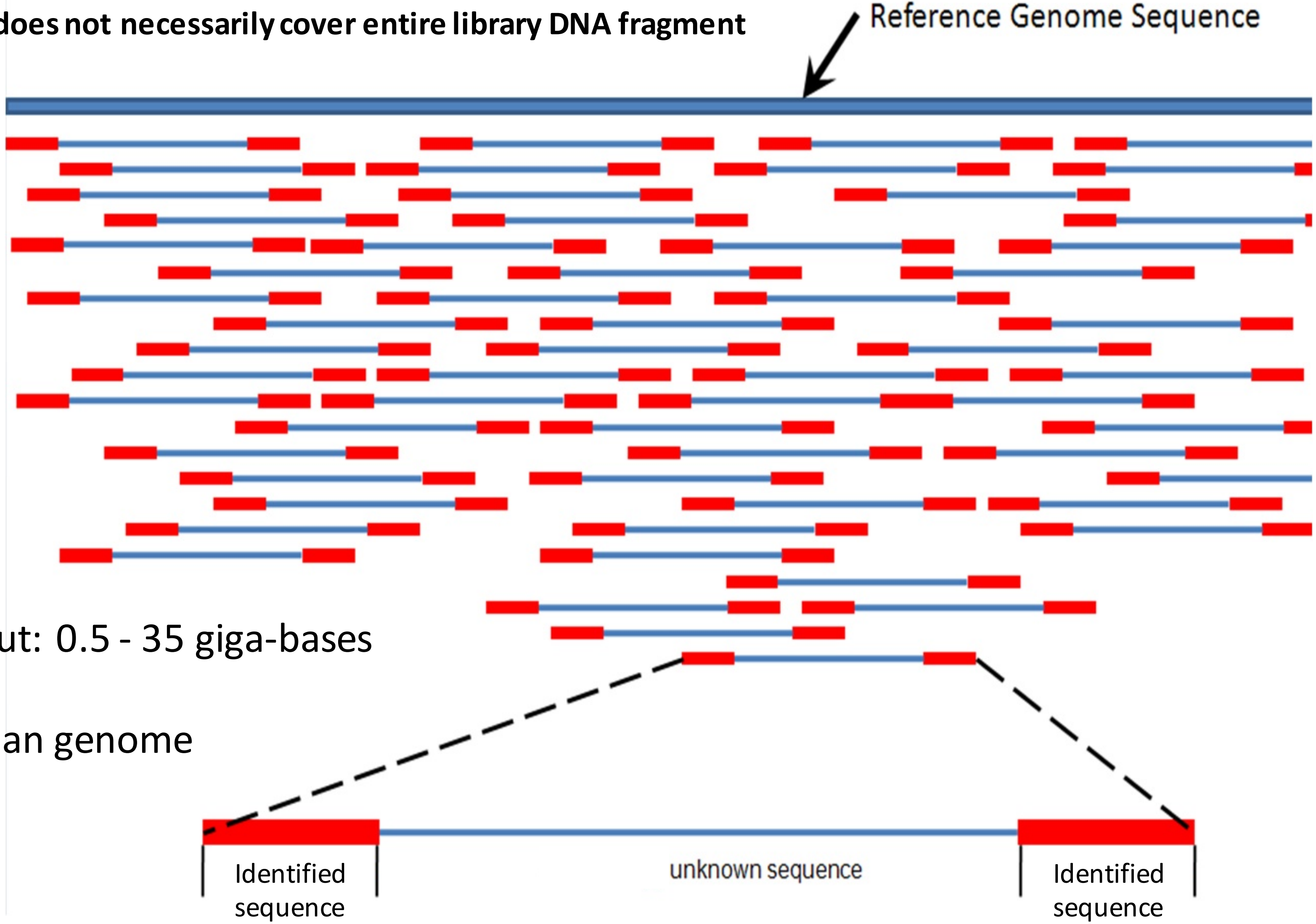


1. Synthesis = incorporation of fluorescent nucleotide: blocking synthesis
2. dye cleavage + elimination
3. wash step
4. Scanning of fluorescent signal
1. Synthesis = incorporation of fluorescent nucleotide: blocking synthesis
READ LENGTH: ca: 150nt from each primer (2x150nt = 300nt)

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

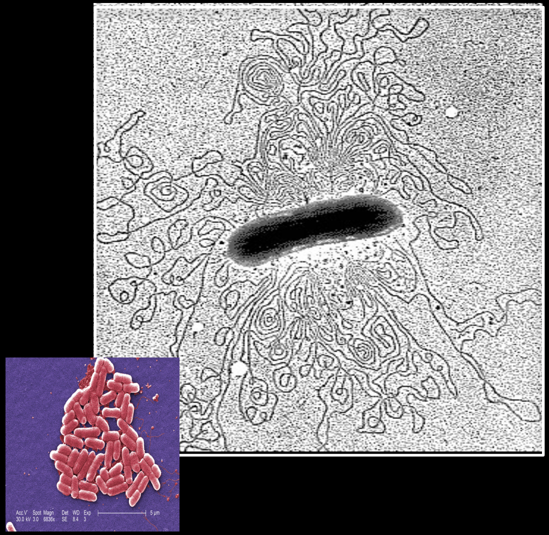


Max. output: 0.5 - 35 giga-bases
$=3.5*10^{10}$
= 10x human genome

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

ENSEMBL 11/2014

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.

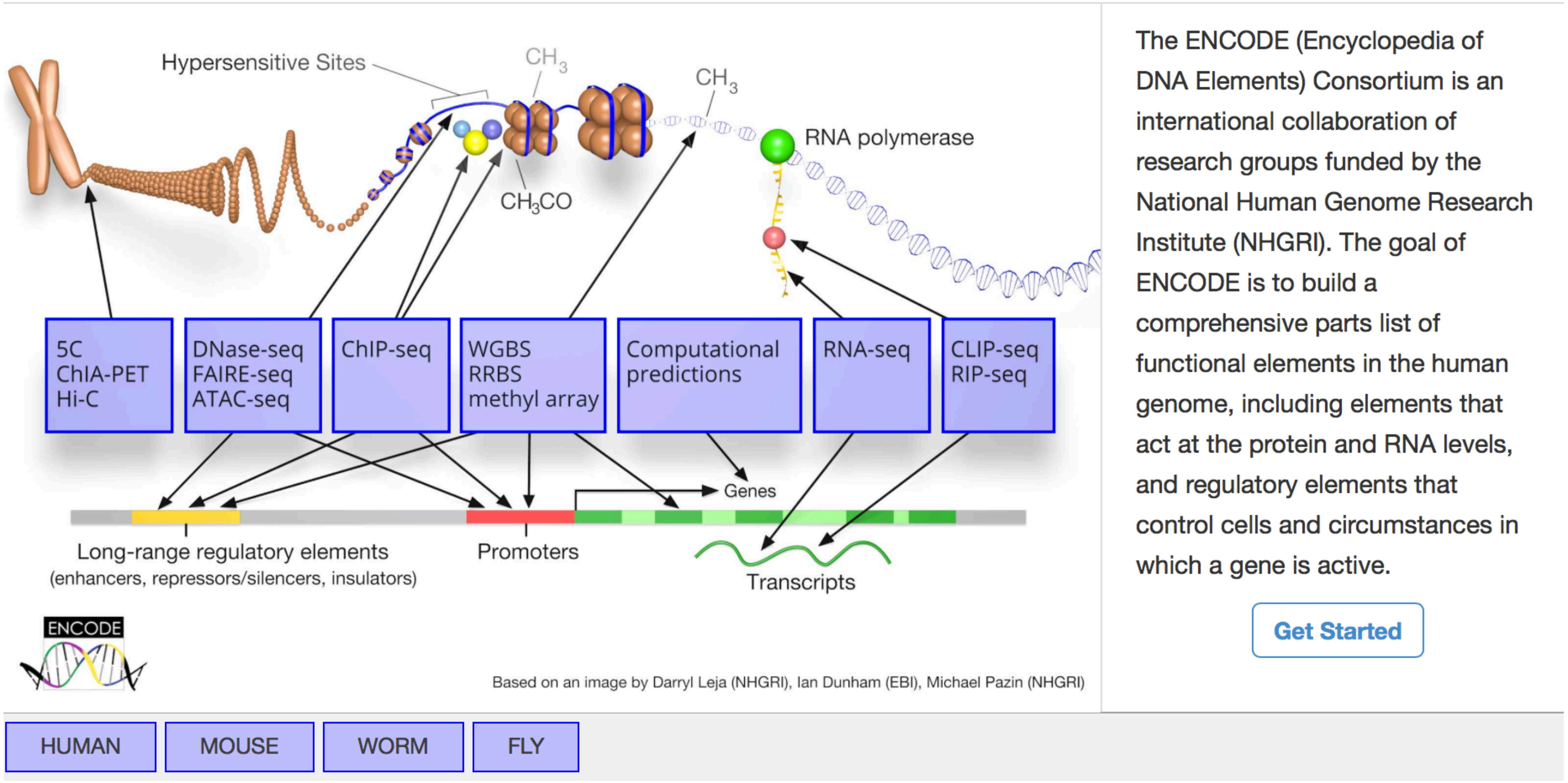
Intended as a follow-up to the Human Genome Project (Genomic Research), the ENCODE project aims to identify all functional elements in the human genome.

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

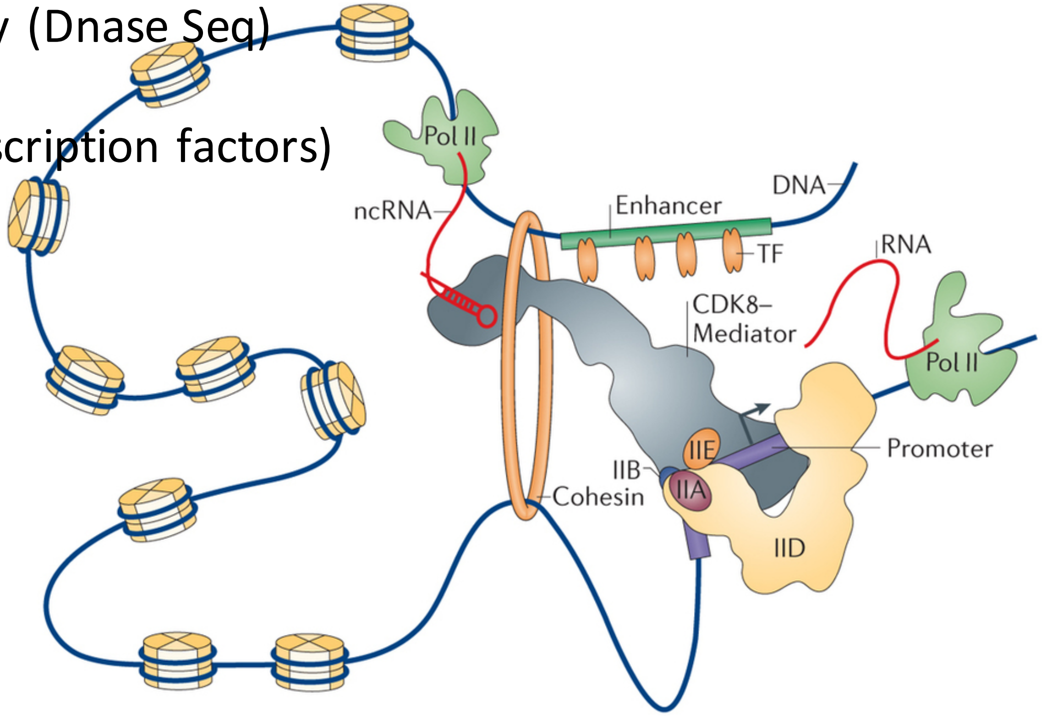
→ IDENTIFICATION OF ALL GENES

→ 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

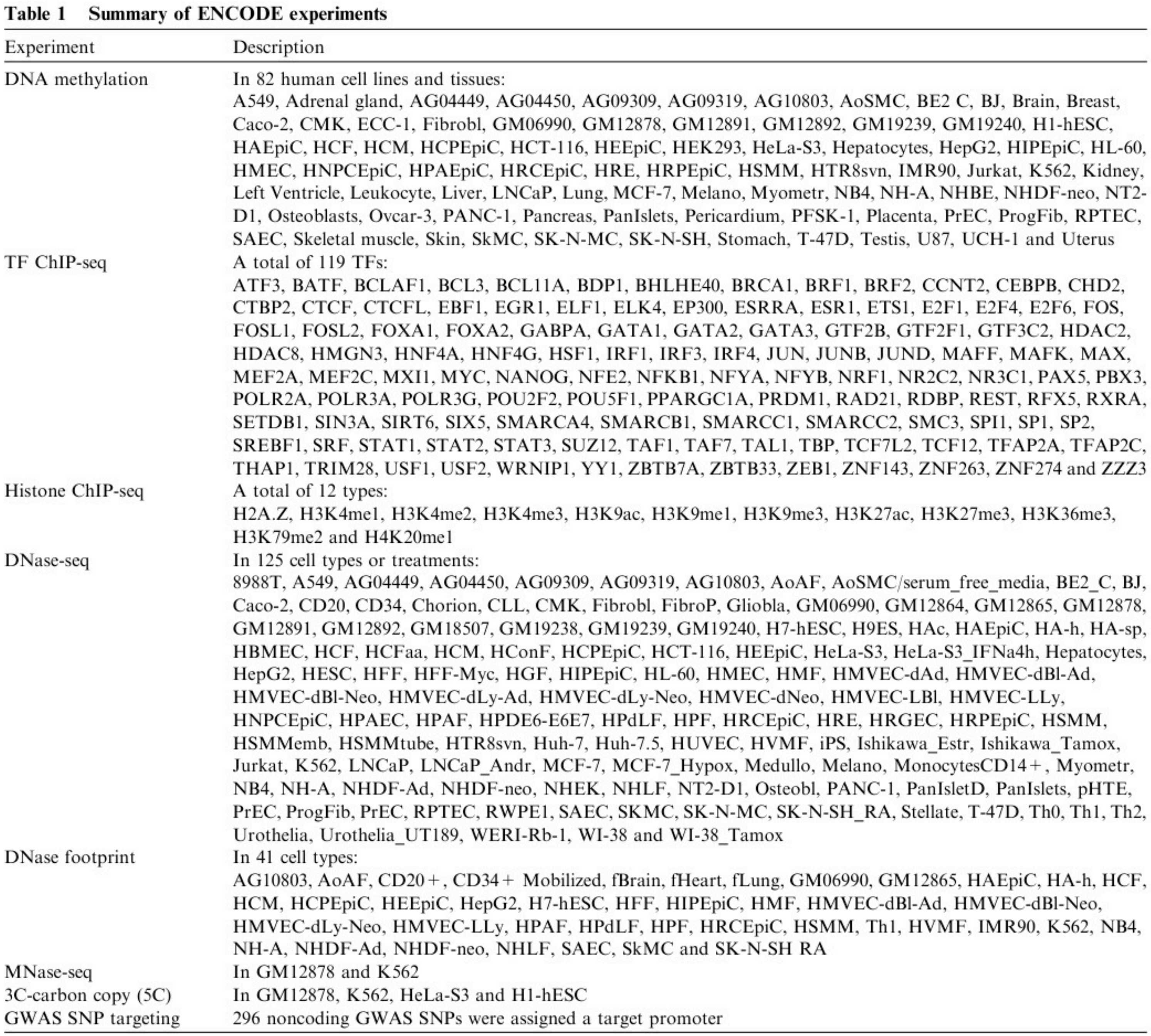


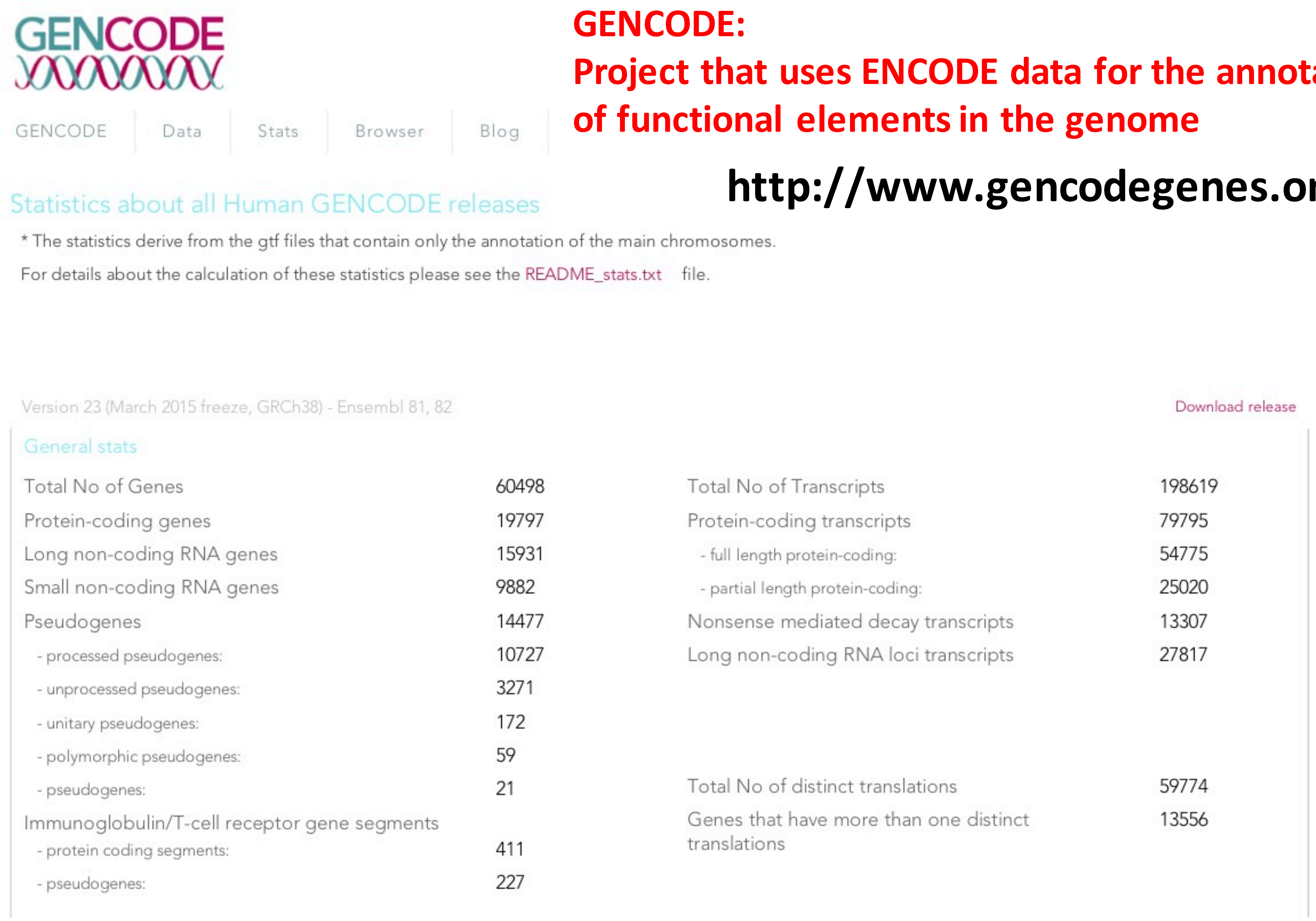
Table 1 Summary of ENCODE experiments

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

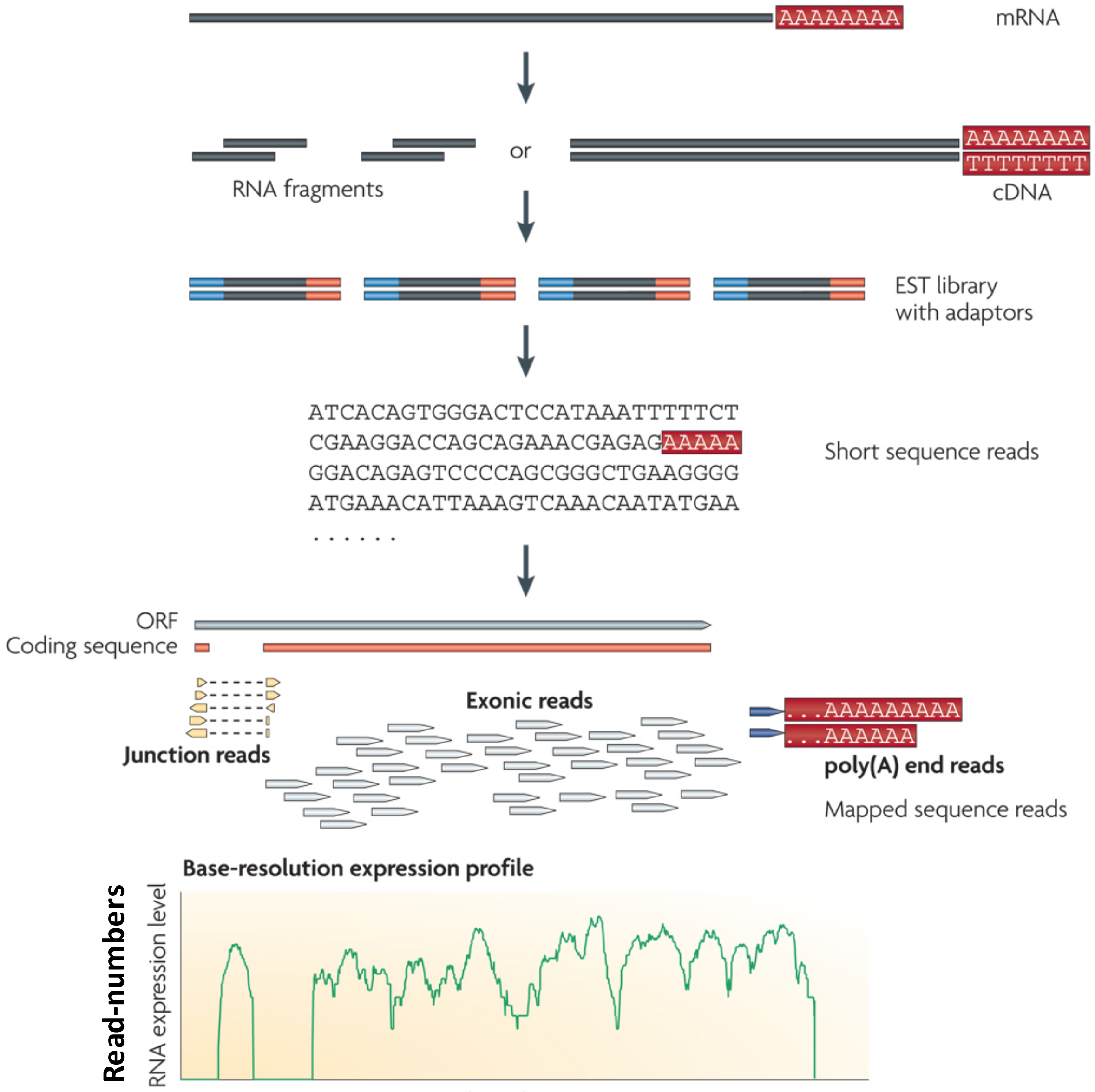
Total No of Genes	60498
Protein-coding genes	19797
Long non-coding RNA genes	15931
Small non-coding RNA genes	9882
Pseudogenes	14477
- processed pseudogenes:	10727
- unprocessed pseudogenes:	3271
- unitary pseudogenes:	172
- polymorphic pseudogenes:	59
- pseudogenes:	21
Immunoglobulin/T-cell receptor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

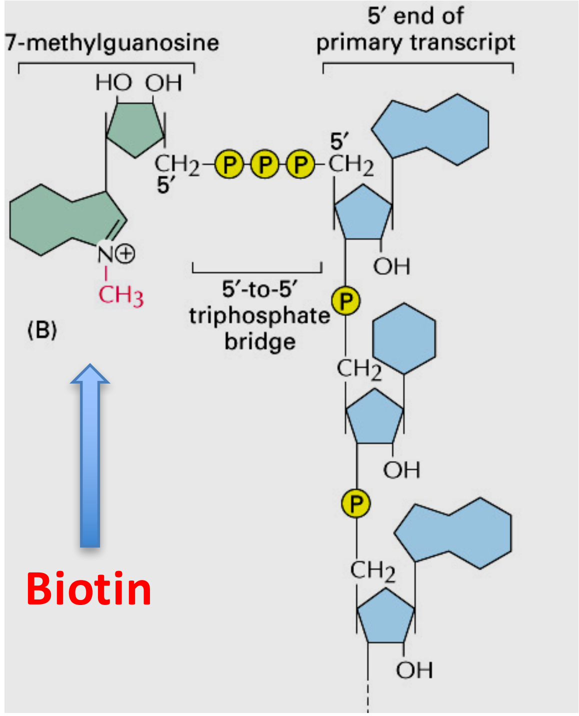


DNA sequence corresponding With identified RNAs

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.



Commonly starting from 50µg total RNA.

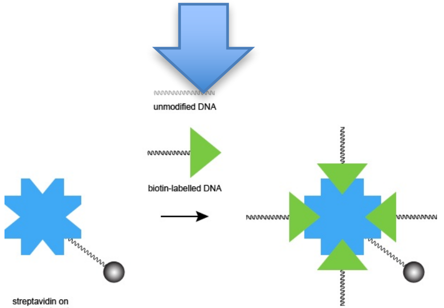
1st Strand cDNA Synthesis

(Covering poly(A-) mRNA and long mRNA.)

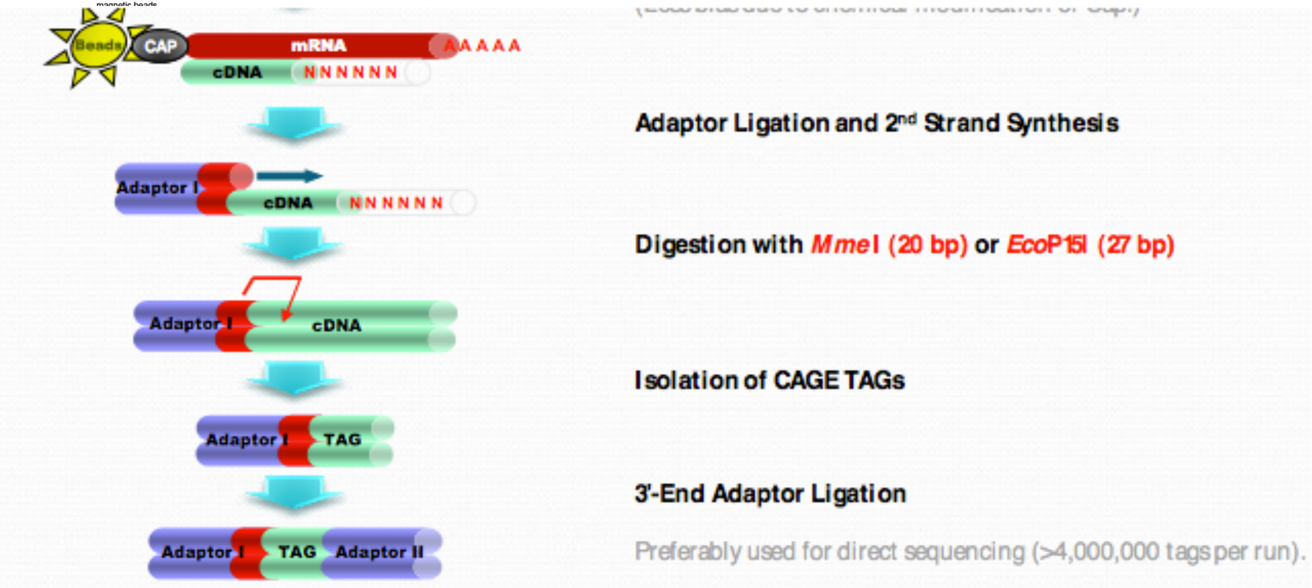
5′-End Selection on Beads by Cap Trapper

(Less bias due to chemical modification of Cap.)

Chemical reaction: Biotin is added to 5′CAP



Concentration of Biotinylated CAPs = concentration of 5′ ends



Adaptor Ligation and 2nd Strand Synthesis

Digestion with *MmeI* (20 bp) or *EcoP15I* (27 bp)

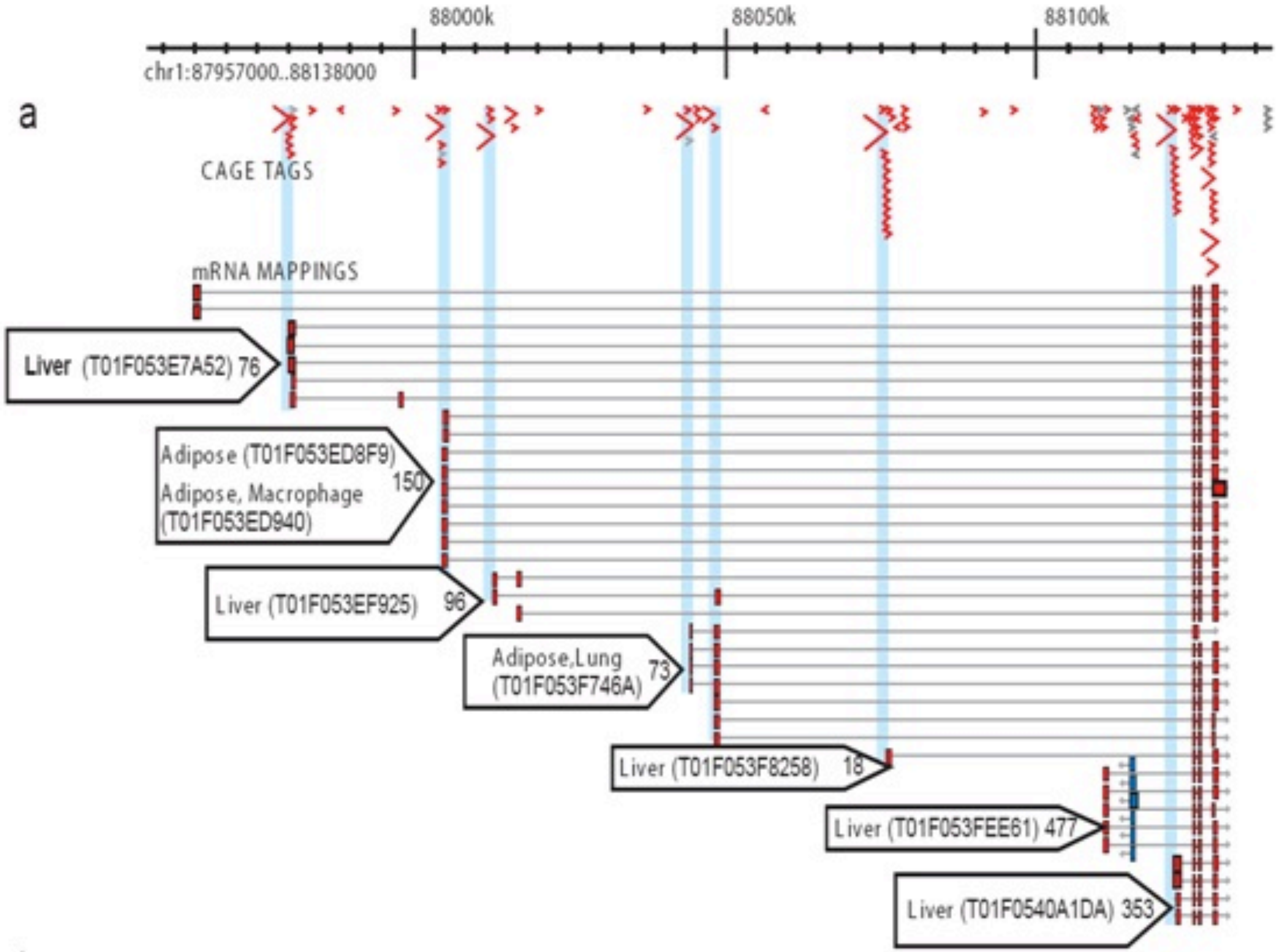
Isolation of CAGE TAGs

3′-End Adaptor Ligation

Preferably used for direct sequencing (>4,000,000 tags per run).

Massive parallel sequencing

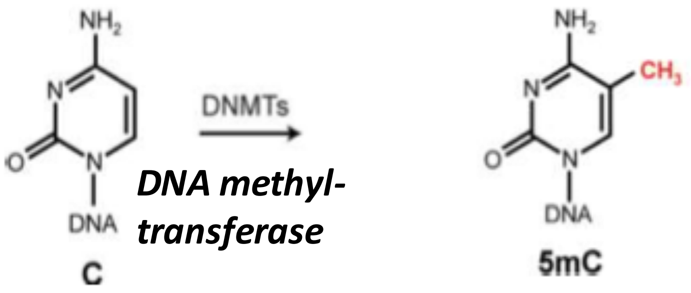
2. RNA Seq variant technology: CAGE (Cap Analysis of Gene Expression)



Excellent tool
To identify
transcriptional
start sites

2. DNA methylation: educed representation bisulfite sequencing (RRBS)

Methylation of cytosine at CpG dinucleotides is an important epigenetic regulatory modification in many eukaryotic genomes.

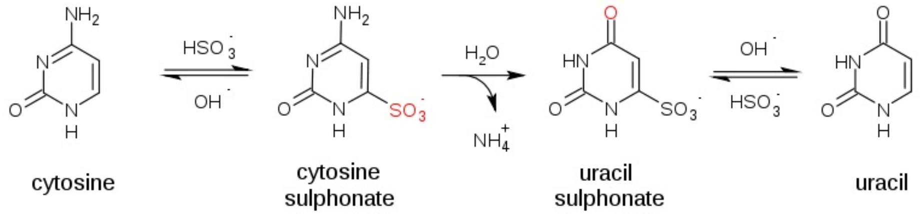


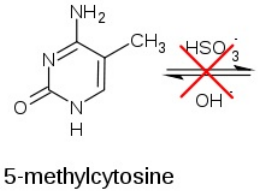
DNA methyl-transferase

active gene

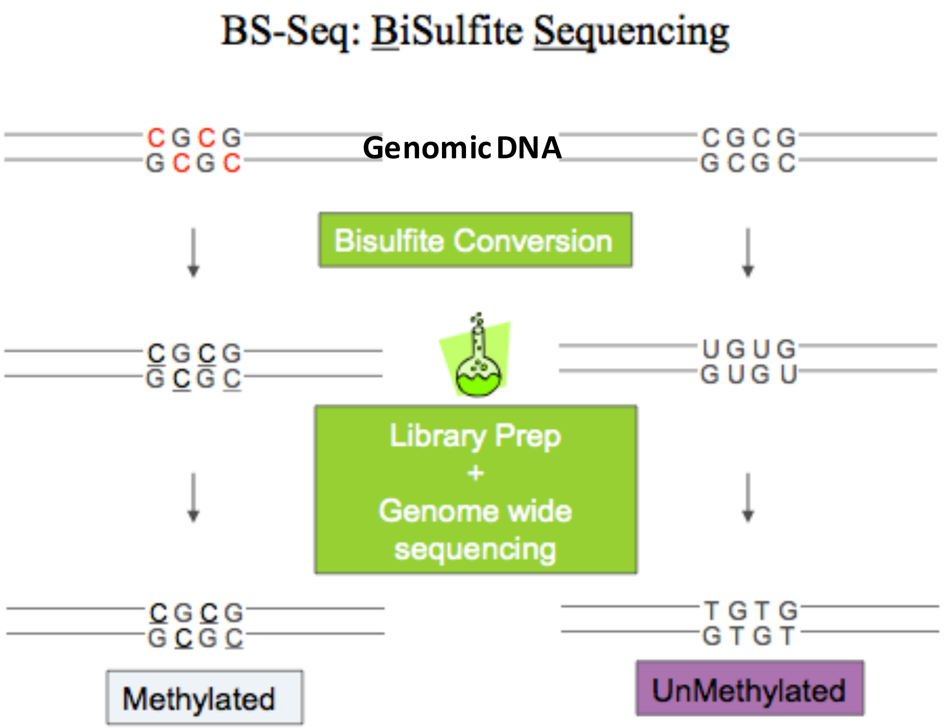
Silenced gene

Bi-sulfite conversion: C→U conversion



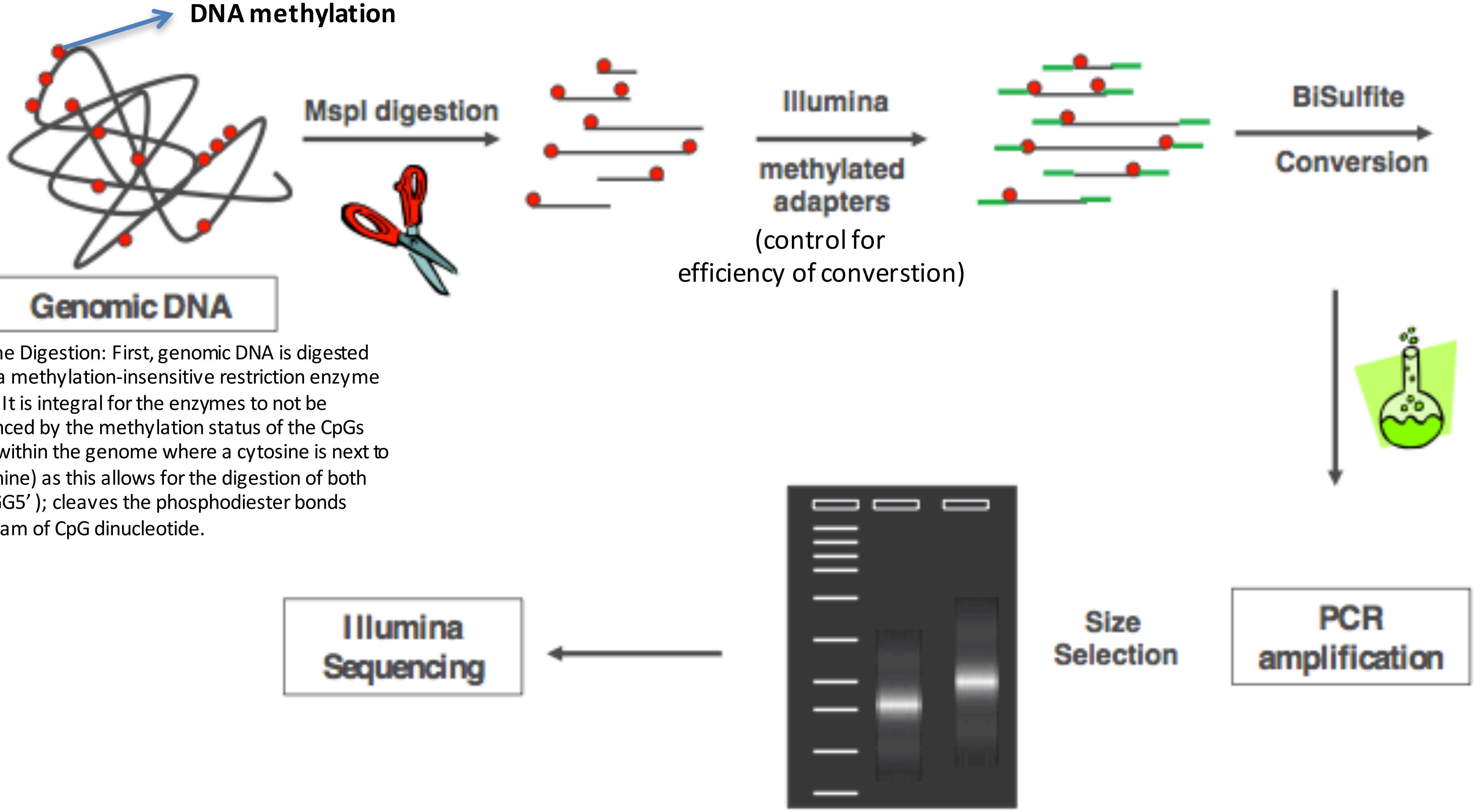


methylated C cannot be converted!!



2. DNA methylation: Reduced representation bisulfite sequencing (RRBS)

Reduced representation bisulfite sequencing (RRBS) is an efficient and high-throughput technique used to analyze the genome-wide methylation profiles on a single nucleotide level. This technique combines restriction enzymes and bisulfite sequencing in order to enrich for the areas of the genome that have a high CpG content. Due to the high cost and depth of sequencing needed to analyze methylation status in the entire genome. The fragments that comprise the reduced genome still include the majority of promoters, as well as regions such as repeated sequences that are difficult to profile using conventional bisulfite sequencing approaches.



(control for efficiency of converstion)

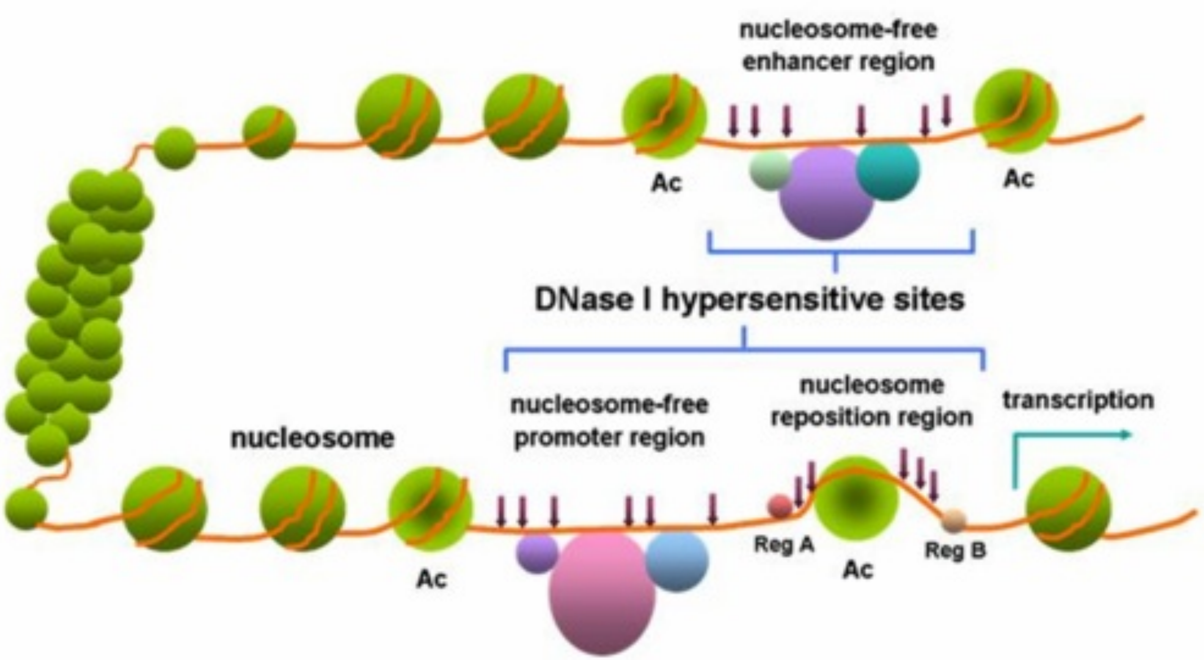
Enzyme Digestion: First, genomic DNA is digested using a methylation-insensitive restriction enzyme MspI. It is integral for the enzymes to not be influenced by the methylation status of the CpGs (sites within the genome where a cytosine is next to a guanine) as this allows for the digestion of both (3’CCGG5’); cleaves the phosphodiester bonds upstream of CpG dinucleotide.

4. Local chromatin structure: determination of DNAse I hypersensitivity (DNase Seq)

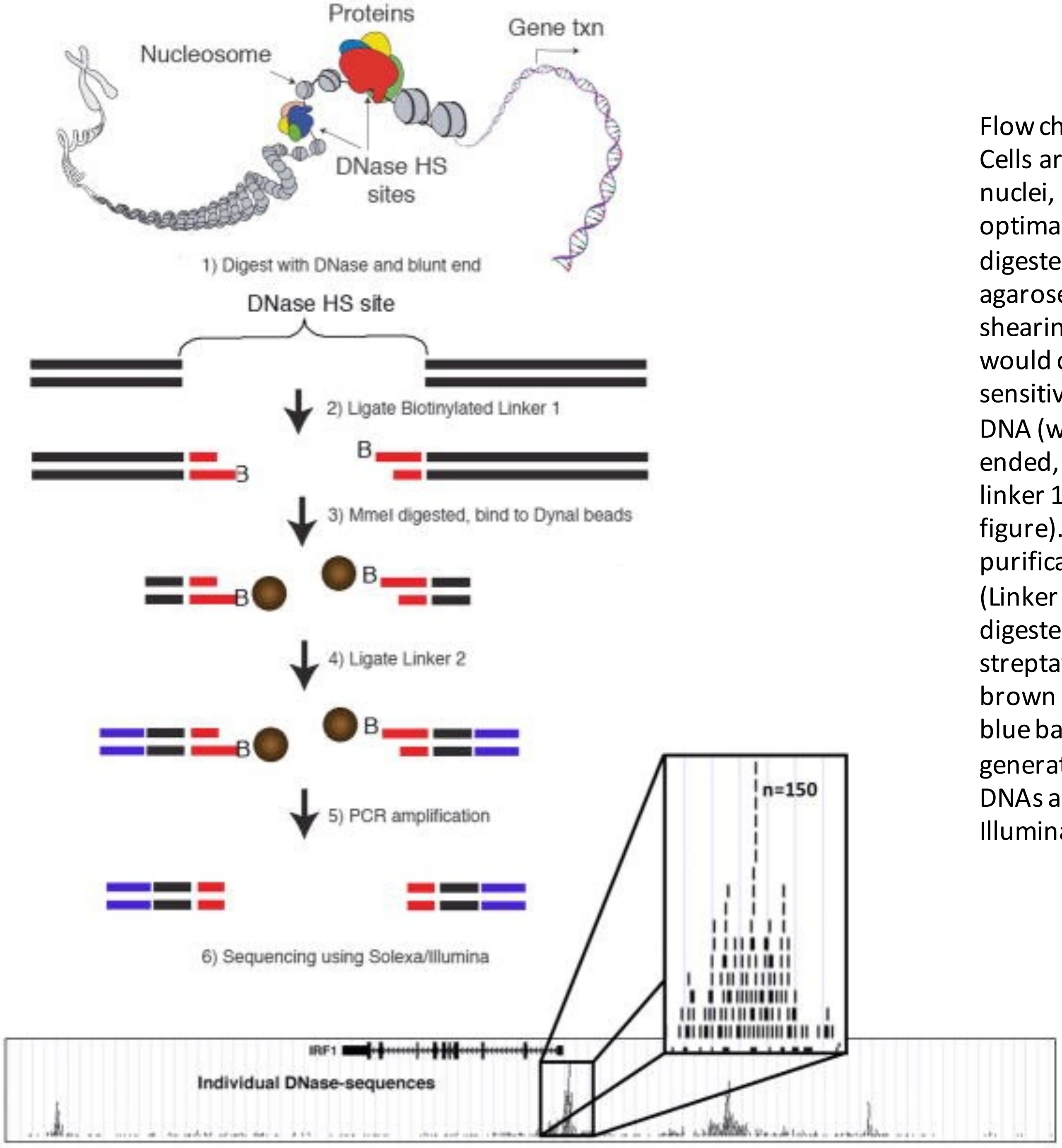
- determination of DNAse I hypersensitivity (DNase Seq)
- Nucleosome occupancy (MNase-seq)
- ChIP-seq (chromatin modifications, transcription factors)
- 3 Dimensional space interaction

DNase hypersensitive sites mark sequences involved in gene regulation

DNase I hypersensitive sites (DHSs) are regions of chromatin that are sensitive to cleavage by the DNase I enzyme. In these specific regions of the genome, chromatin has lost its condensed structure, exposing the DNA and making it accessible. This raises the availability of DNA to degradation by enzymes, such as DNase I. These accessible chromatin zones are functionally related to transcriptional activity, since this remodeled state is necessary for the binding of proteins such as transcription factors.

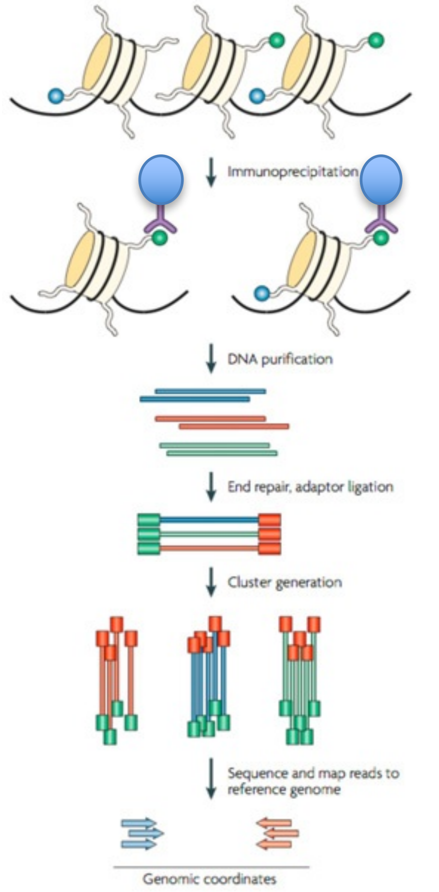


4. Local chromatin structure: determination of DNAse I hypersensitivity (DNase Seq)



Flow chart of DNase-seq protocol.
Cells are lysed with detergent to release nuclei, and the nuclei are digested with optimal concentrations of DNase I. DNase I digested DNA is immobilized in low-melt gel agarose plugs to reduce additional random shearing. (pipetting can cause breaks that would cause "false positive" DNase hyper sensitive sites).
DNA (while still in the plugs) are then blunt-ended, extracted and ligated to biotinylated linker 1 (represented by red bars in the figure). Excess linker is removed by gel purification, and biotinylated fragments (Linker 1 plus 20 bases of genomic DNA) are digested with MmeI, and captured by streptavidin-coated beads (represented by brown balls). Linker 2 (represented by the blue bars) is ligated to the 2 base overhang generated by MmeI, and the ditagged 20 bp DNAs are amplified by PCR and sequenced by Illumina/Solexa.

4. Local chromatin structure:
Chromatin immunoprecipitation sequencing (ChIP-seq)



H3K4me3
(active chromatin mark)

H3K27me3
(repressive chromatin mark)

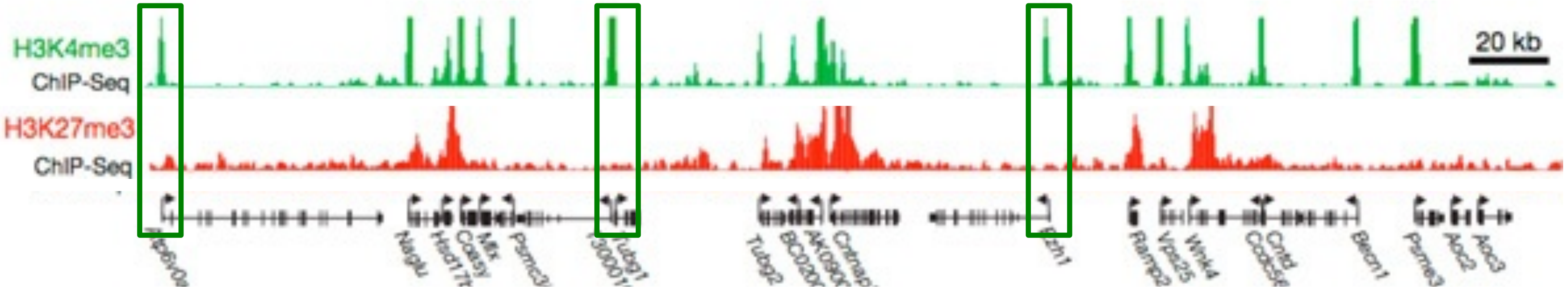
● magnetic beads
covered with specific antibody

1. Cell fixation-proteins and DNA are crosslinked
2. Sonication of DNA (fragmentation)
3. Immunoprecipitation of chromatin using

Specific antibodies: histone modifications or transcription Factors

4. Purify beads (magnet), washing of beads + elution of immunoprecipitated material
5. Library construction
6. Massive parallel sequencing
7. Align sequencing results to genomic sequence
8. Increase in read-number for a particular sequence indicates

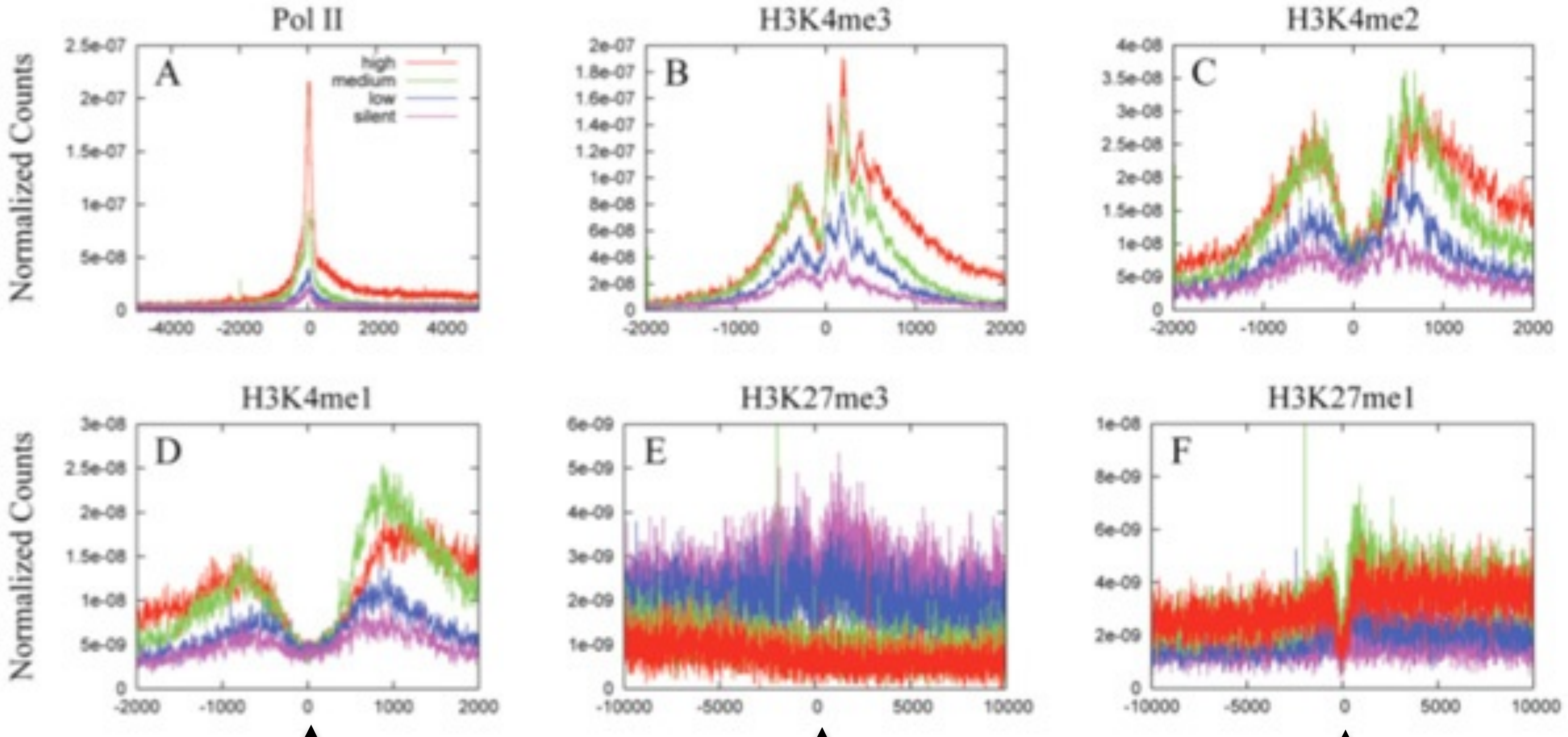
Enrichment for the histone modification or transcription factor

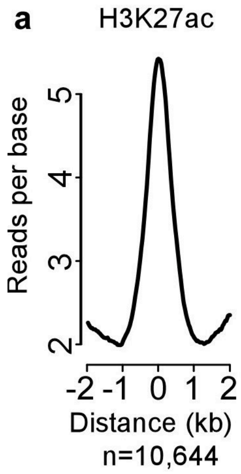


The results indicate that some modifications (H3K4me) are correlated with increased gene expression, while others (H3K27me3) correlate with decreases gene expression. The peaks observed in the H3K4me3 for genes at high expression levels occur at +50, +210, and +360 based which correlates well with the known spacing interval for nucleosome positioning. Furthermore, the dip in abundance at the transcriptional start site is consistent with local nucleosome depletion of actively expressed genes.

4. Local chromatin structure:
Chromatin immunoprecipitation sequencing (ChIP-seq)

A special chromatin code marks the transcriptional start site of Pol II target genes





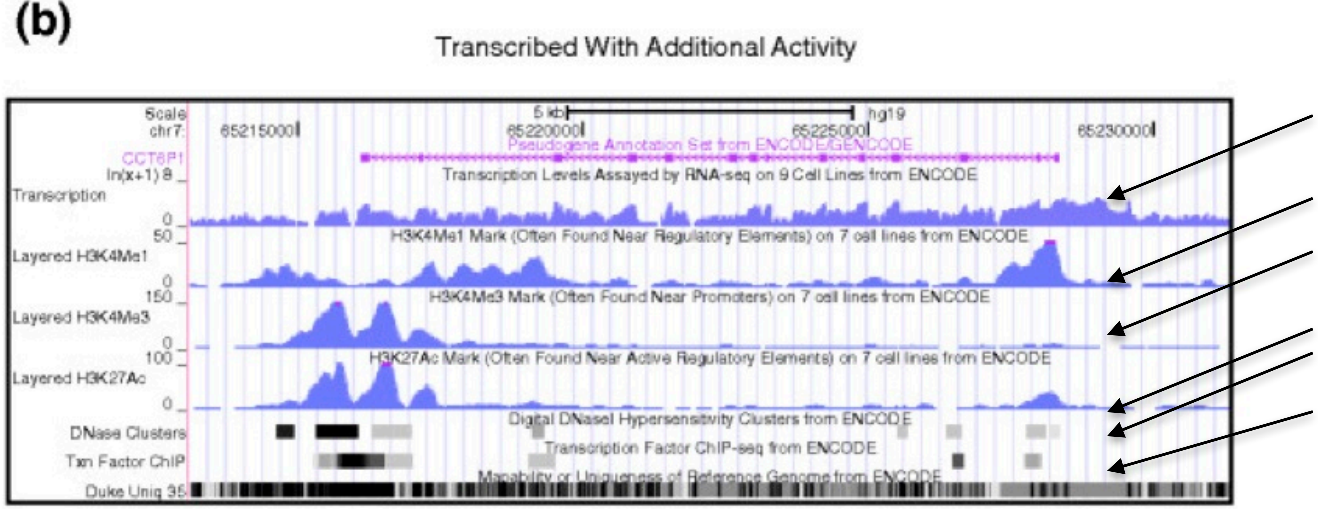
Position 0:
RNA Polymerase II: peak
H4K4me3: peak
H3K4me2: drop
H3K4me1: drop
H3K27me3: low
H3K27me1: drop

transcriptional start site = position 0

Regulatory elements

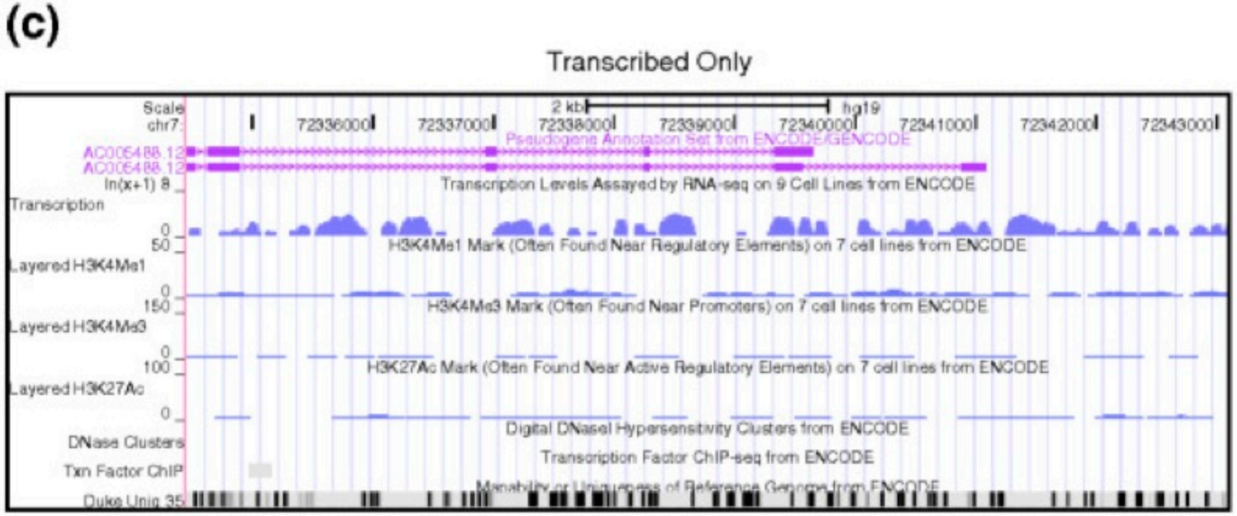
Same method can be used to localize transcription factors

AN EXAMPLE:ORGANISATION OF A FUNCTIONAL ELEMENT: PSEUDOGENES



Pseudogene CCT6P1

RNA expression: PRESENT
RNA Polymerase II: not shown
H4K4me1: near regulatory elements
H3K4me3: near promoters
H3K27Ac: near regulatory elements
DNAse hypersensitive sites: at regulatory elements
Transcription factor (TF) binding: Near promoter



Pseudogene AC0064BB12

RNA expression: PRESENT
Chromatin shows actve marks
Poor definition

Summary of pseudogene annotation and case studies. **(a)** A heatmap showing the annotation for transcribed pseudogenes including active chromatin segmentation, DNaseI hypersensitivity, active promoter, active Pol2, and conserved sequences. Raw data were from the K562 cell line. **(b)** A transcribed duplicated pseudogene (Ensembl gene ID: ENST00000434500.1; genomic location, chr7: 65216129-65228323) showing consistent active chromatin accessibility, histone marks, and TFBSs in its upstream sequences. **(c)** A transcribed processed pseudogene (Ensembl gene ID: ENST00000355920.3; genomic location, chr7: 72333321-72339656) with no active chromatin features or conserved sequences. **(d)** A non-transcribed duplicated pseudogene showing partial activity patterns (Ensembl gene ID: ENST00000429752.2; genomic location, chr1: 109646053-109647388). **(e)** Examples of partially active pseudogenes. E1 and E2 are examples of duplicated pseudogenes. E1 shows *UGT1A2P* (Ensembl gene ID: ENST00000454886), indicated by the green arrowhead. *UTG1A2P* is a non-transcribed pseudogene with active chromatin and it is under negative selection. Coding exons of protein-coding paralogous loci are represented by dark green boxes and UTR exons by filled red boxes. E2 shows *FAM86EP* (Ensembl gene ID: ENST00000510506) as open green boxes, which is a transcribed pseudogene with active chromatin and upstream TFBSs and Pol2 binding sites. The transcript models associated with the locus are displayed as filled red boxes. Black arrowheads indicate features novel to the pseudogene locus. E3 and E4 show two unitary pseudogenes. E3 shows *DOC2GP* (Ensembl gene ID: ENST00000514950) as open green boxes, and transcript models associated with the locus are shown as filled red boxes. E4 shows *SLC22A20* (Ensembl gene ID: ENST00000530038). Again, the pseudogene model is represented as open green boxes, transcript models associated with the locus as filled red boxes, and black arrowheads indicate features novel to the pseudogene locus. E5 and E6 show two processed pseudogenes. E5 shows pseudogene *EGLN1* (Ensembl gene ID: ENST00000531623) inserted into duplicated pseudogene *SCAND2* (Ensembl gene ID: ENST00000541103), which is a transcribed pseudogene showing active chromatin but no upstream regulatory regions as seen in the parent gene. The pseudogene models are represented as open green boxes, transcript models associated with the locus are displayed as filled red boxes, and black arrowheads indicate features novel to the pseudogene locus. E6 shows a processed pseudogene *RP11-409K20* (Ensembl gene ID: ENST00000417984; filled green box), which has been inserted into a CpG island, indicated by an orange arrowhead. sRNA, small RNA.

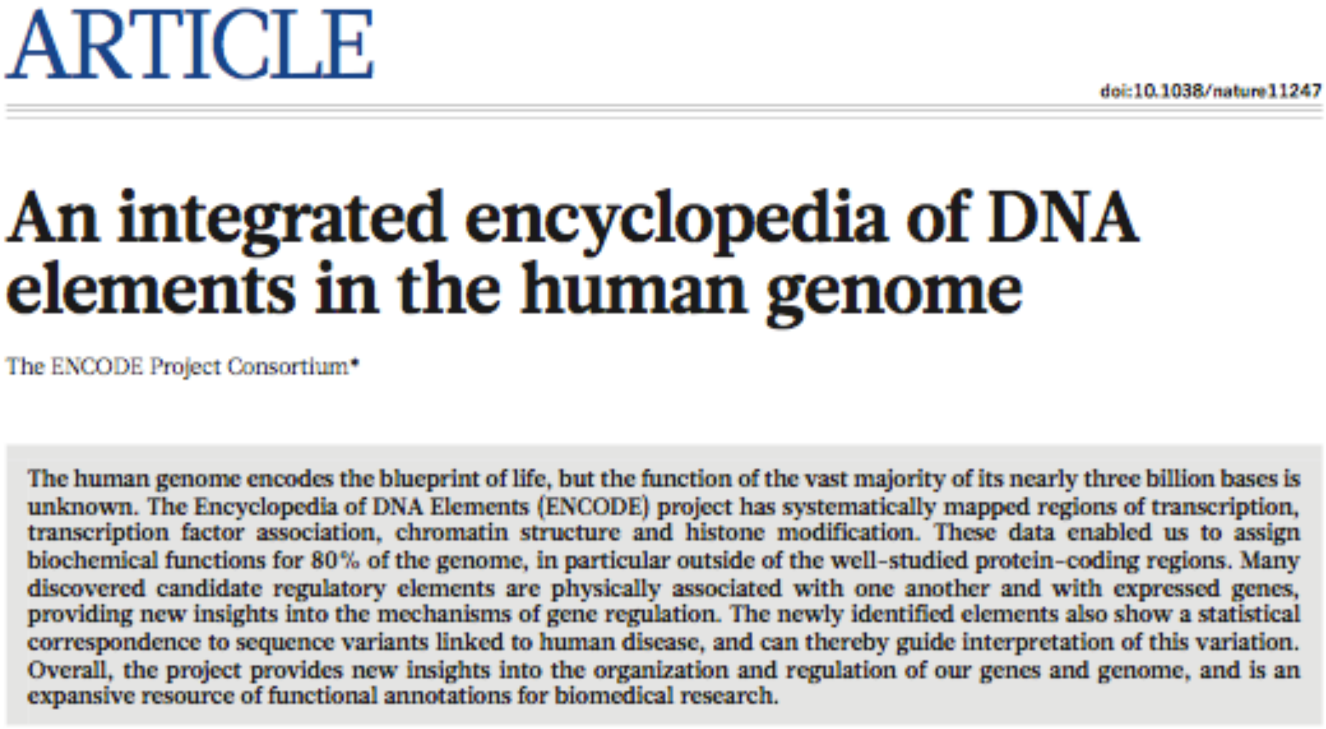
Pei *et al.* *Genome Biology* 2012 **13**:R51 doi:10.1186/gb-2012-13-9-r51

Aim: Identify functional elements of the genome (ENCODE)

WORK STILL IN PRGRESS ***http://www.genome.gov/encode/***

Aim: a catalog of <u>manually curated</u> list of genes/transcripts (GENCODE) ***http://www.gencodegenes.org/***

Release ENCODE7 (2012); new release expected 12/2015)



ARTICLE

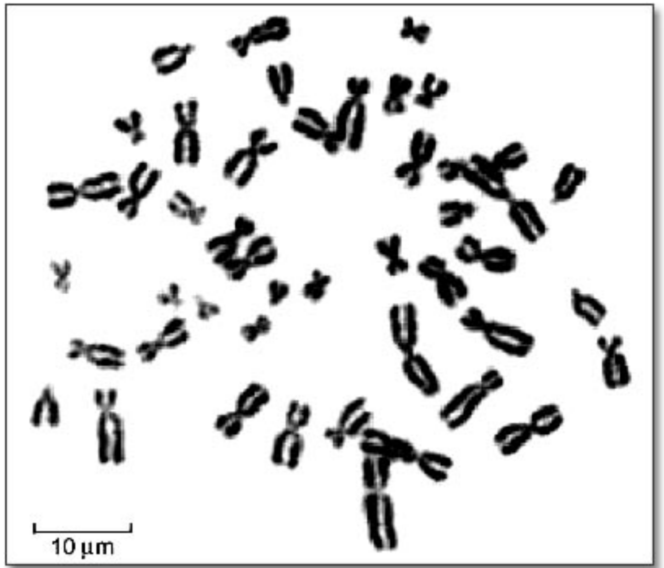
doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

Almost all regions in the genome are subjecte to regualtion and transcription



The vast majority (80.4%) of the human genome participates in at least one biochemical RNA and/or chromatin associated event in at least one cell type. Much of the genome lies close to a regulatory event: 95% of the genome lies within 8kb of a DNA-protein interaction (as assayed by bound ChIP-seq motifs or DNaseI footprints), and 99% is within 1.7kb of at least one of the biochemical events measured by ENCODE.

Classifying the genome into seven chromatin states suggests an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of narrow states with distinct functional properties.

It is possible to quantitatively correlate RNA sequence production and processing with both chromatin marks and transcription factor (TF) binding at promoters, indicating that promoter functionality can explain the majority of RNA expression variation.

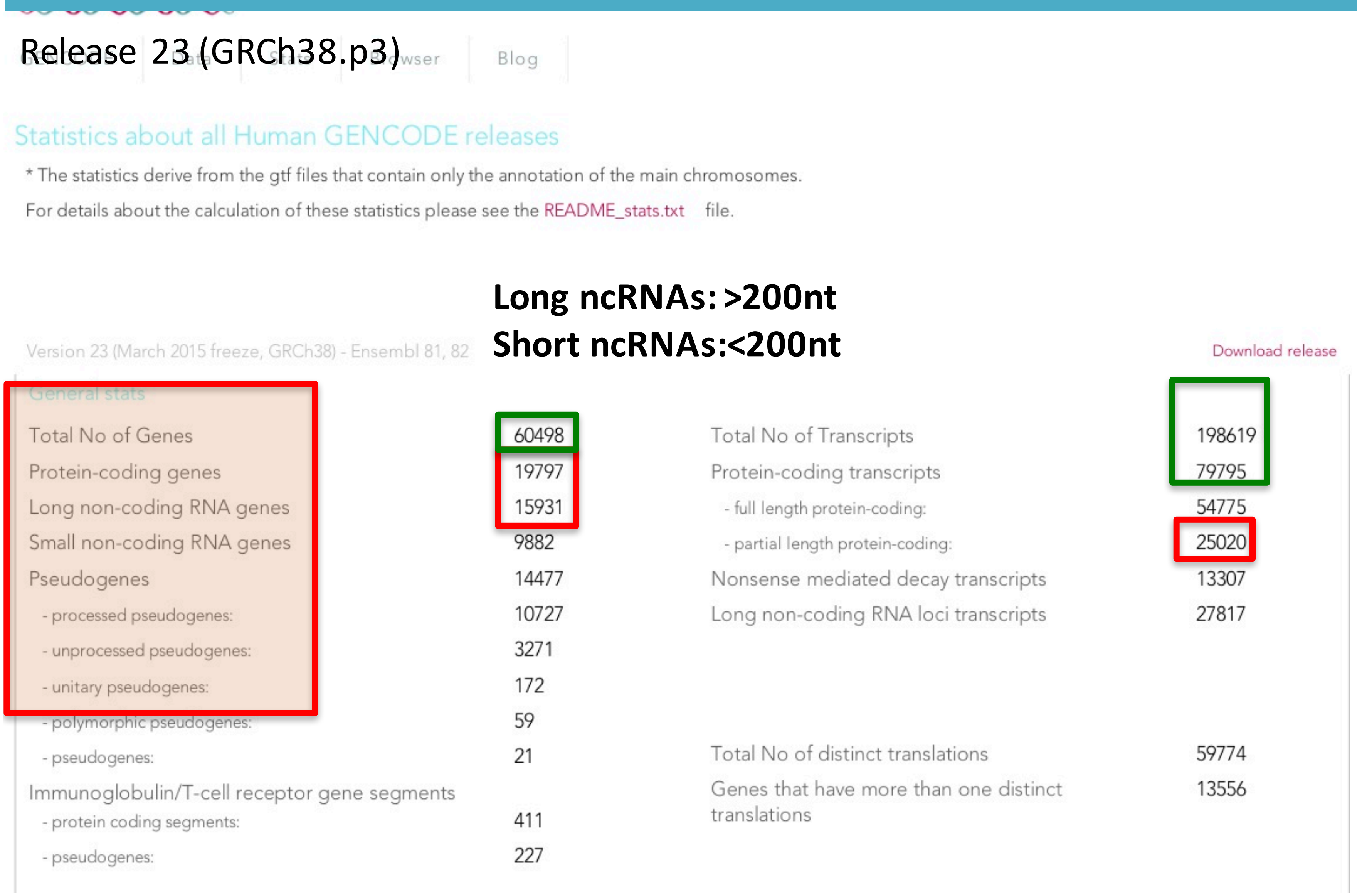
Many non-coding variants in individual genome sequences lie in ENCODE-annotated functional regions; this number is at least as large as those that lie in protein coding genes.

SNPs associated with disease by GWAS are enriched within non-coding functional elements, with a majority residing in or near ENCODE-defined regions that are outside of protein coding genes. In many cases, the disease phenotypes can be associated with a specific cell type or TF.

GENCODE – STATUS 09.11.2015:
Project that uses ENCODE for the annotation of functional elements in the genome

http://www.gencodegenes.org/

http://www.gencodegenes.org/



Release 23 (GRCh38.p3)

Blog

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.

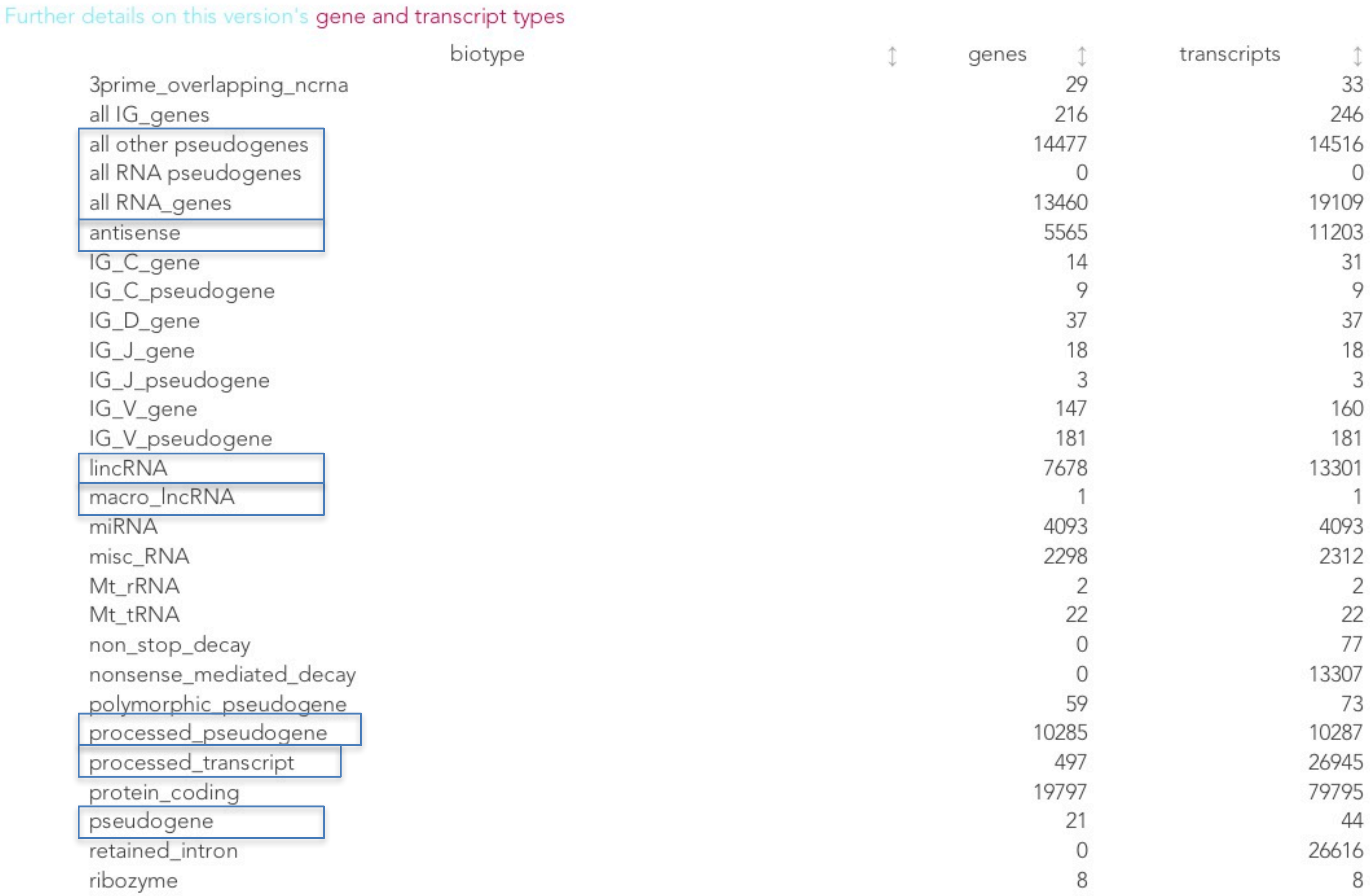
Long ncRNAs: >200nt
Short ncRNAs:<200nt

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

Download release

General stats			
Total No of Genes	60498	Total No of Transcripts	198619
Protein-coding genes	19797	Protein-coding transcripts	79795
Long non-coding RNA genes	15931	- full length protein-coding:	54775
Small non-coding RNA genes	9882	- partial length protein-coding:	25020
Pseudogenes	14477	Nonsense mediated decay transcripts	13307
- processed pseudogenes:	10727	Long non-coding RNA loci transcripts	27817
- unprocessed pseudogenes:	3271		
- unitary pseudogenes:	172		
- polymorphic pseudogenes:	59		
- pseudogenes:	21	Total No of distinct translations	59774
Immunoglobulin/T-cell receptor gene segments		Genes that have more than one distinct translations	13556
- protein coding segments:	411		
- pseudogenes:	227		

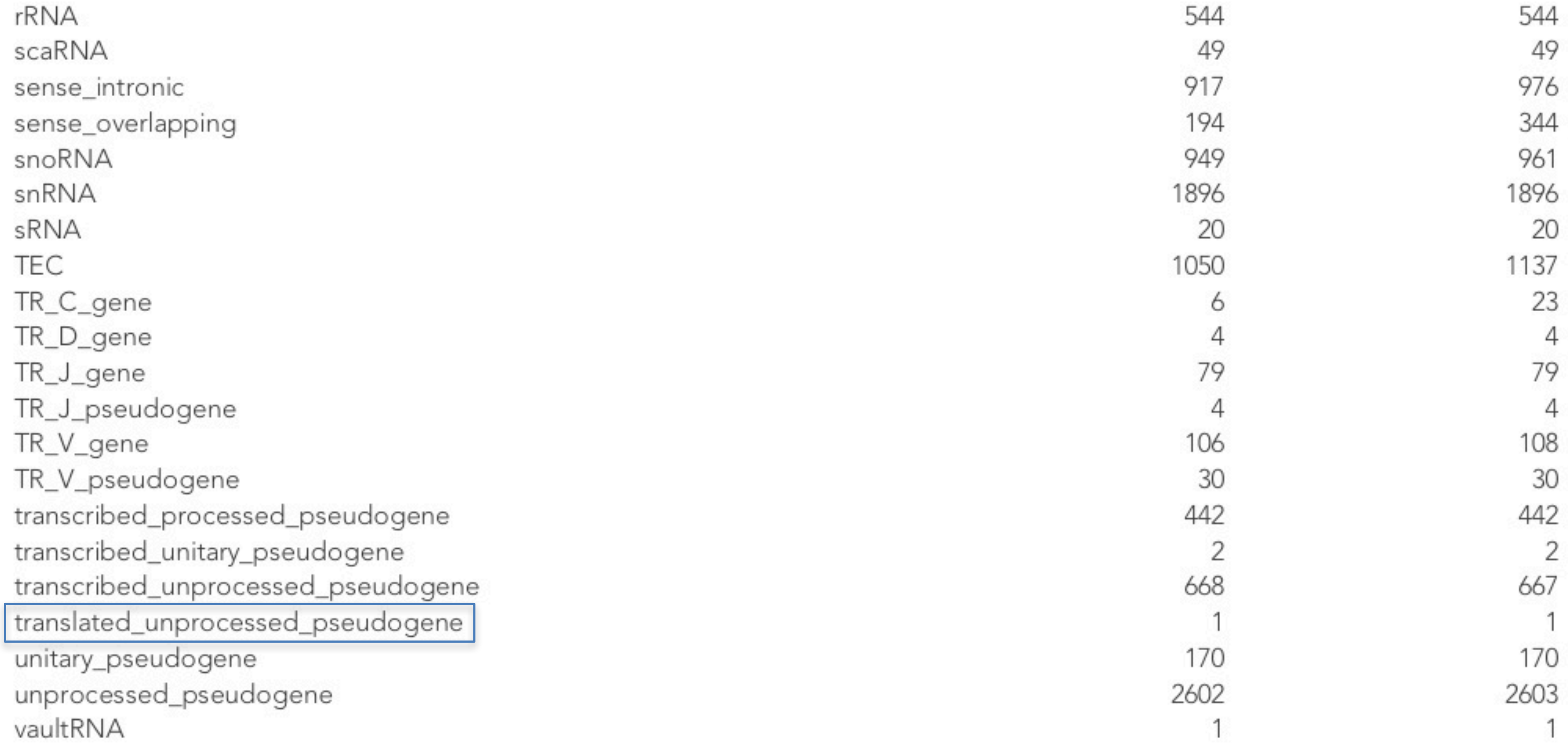
ANNOTATED TRANSCRIPT TYPES (ENCODE ; 11/2015)



Further details on this version's gene and transcript types

biotype	genes	transcripts
3prime_overlapping_ncrna	29	33
all IG_genes	216	246
all other pseudogenes	14477	14516
all RNA pseudogenes	0	0
all RNA_genes	13460	19109
antisense	5565	11203
IG_C_gene	14	31
IG_C_pseudogene	9	9
IG_D_gene	37	37
IG_J_gene	18	18
IG_J_pseudogene	3	3
IG_V_gene	147	160
IG_V_pseudogene	181	181
lincRNA	7678	13301
macro_lncRNA	1	1
miRNA	4093	4093
misc_RNA	2298	2312
Mt_rRNA	2	2
Mt_tRNA	22	22
non_stop_decay	0	77
nonsense_mediated_decay	0	13307
polymorphic_pseudogene	59	73
processed_pseudogene	10285	10287
processed_transcript	497	26945
protein_coding	19797	79795
pseudogene	21	44
retained_intron	0	26616
ribozyme	8	8

ANNOTATED TRANSCRIPT TYPES (ENCODE ; 11/2015)



rRNA	544	544
scaRNA	49	49
sense_intronic	917	976
sense_overlapping	194	344
snoRNA	949	961
snRNA	1896	1896
sRNA	20	20
TEC	1050	1137
TR_C_gene	6	23
TR_D_gene	4	4
TR_J_gene	79	79
TR_J_pseudogene	4	4
TR_V_gene	106	108
TR_V_pseudogene	30	30
transcribed_processed_pseudogene	442	442
transcribed_unitary_pseudogene	2	2
transcribed_unprocessed_pseudogene	668	667
translated_unprocessed_pseudogene	1	1
unitary_pseudogene	170	170
unprocessed_pseudogene	2602	2603
vaultRNA	1	1

NOTE: These are annotated ncRNA transcripts/gene: they are subjected to gene Regulatory mechanisms.

NOTE: ncRNAs can also be generated outside of defined transcription units!!!

Example: DNA damage repair RNAs (DDRNA)