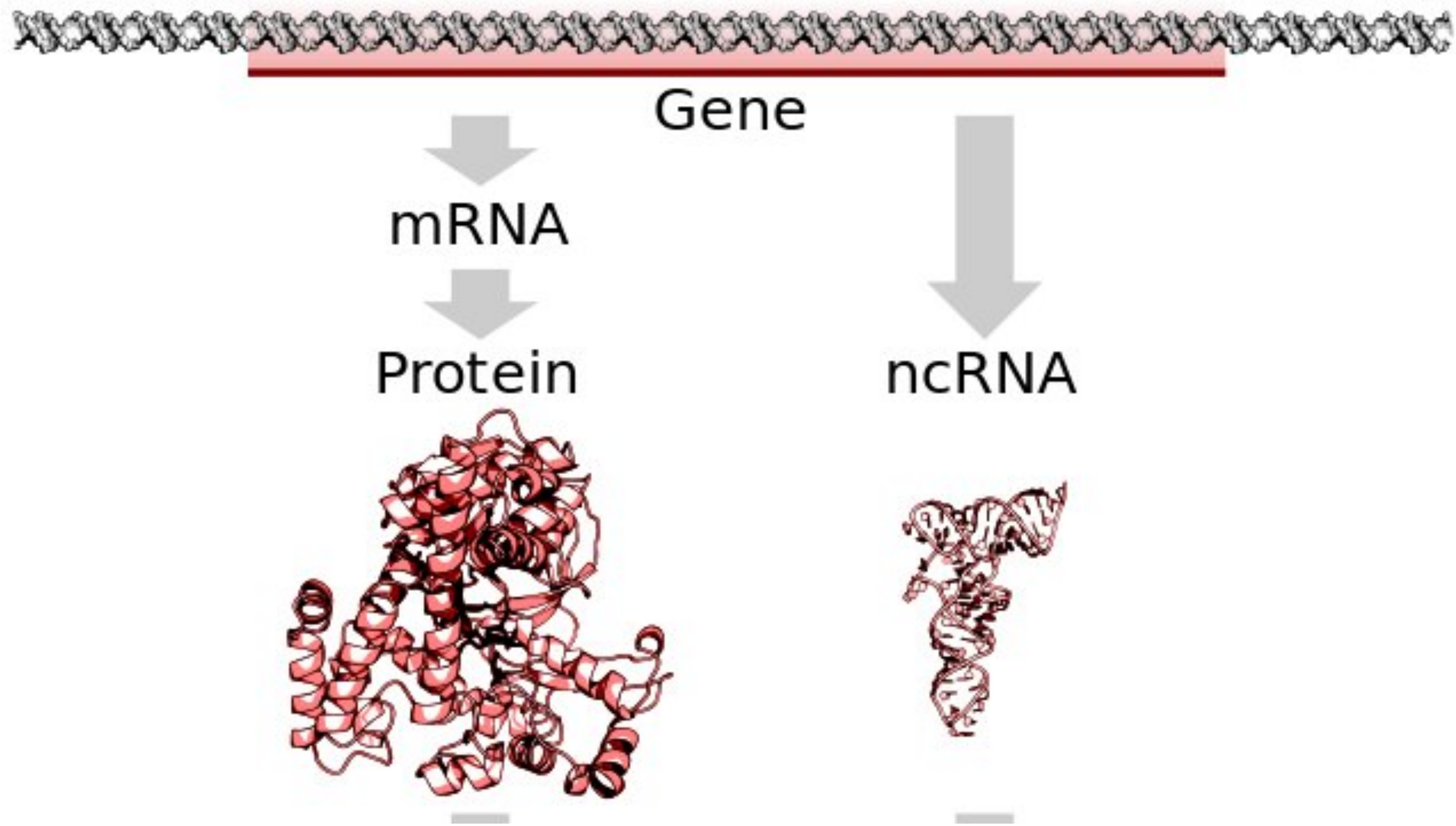
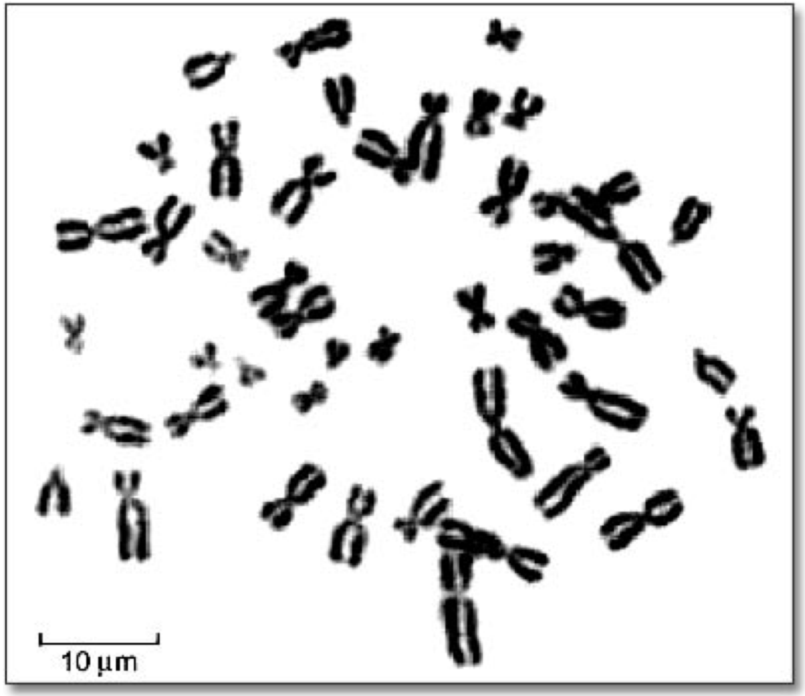
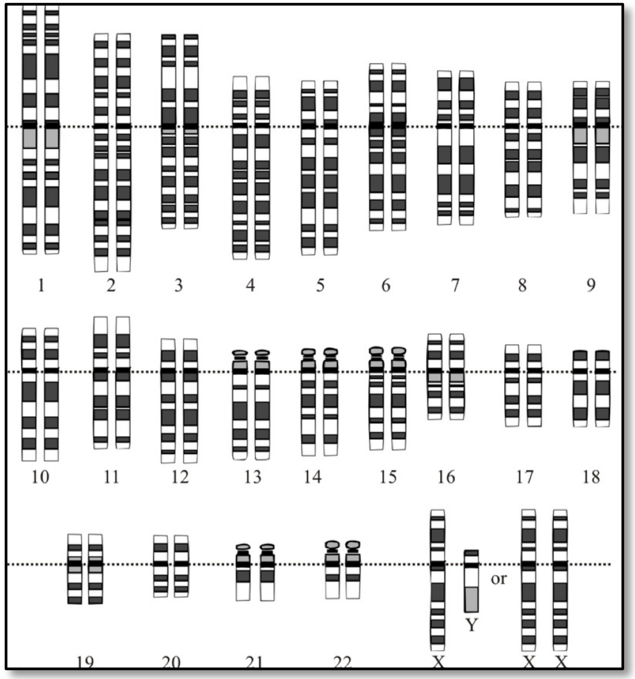
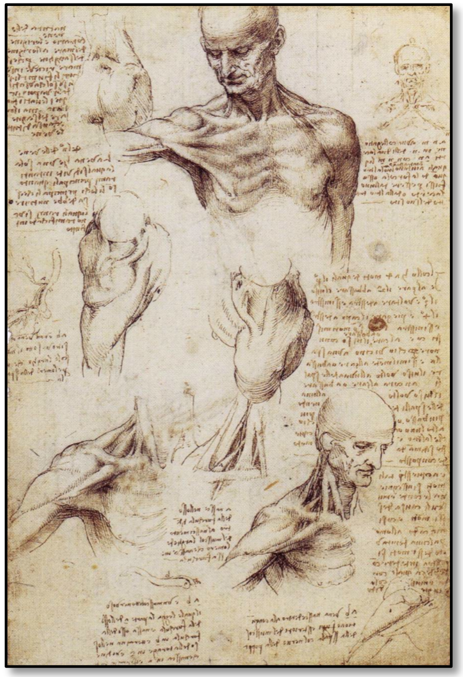
# Non coding RNA Biology
## AA 2024/2025



# The human genome is highly structured







The human genome:

22 autosome pairs

2 Sex chromosome pairs (XX o XY)

Total haploid genome $3x10^9$

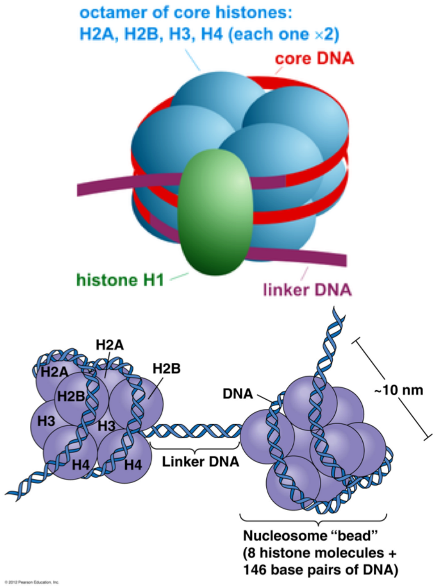
# The human genome is highly structured

Chromatin: DNA + protein in nucleus
Organisation of genetic information

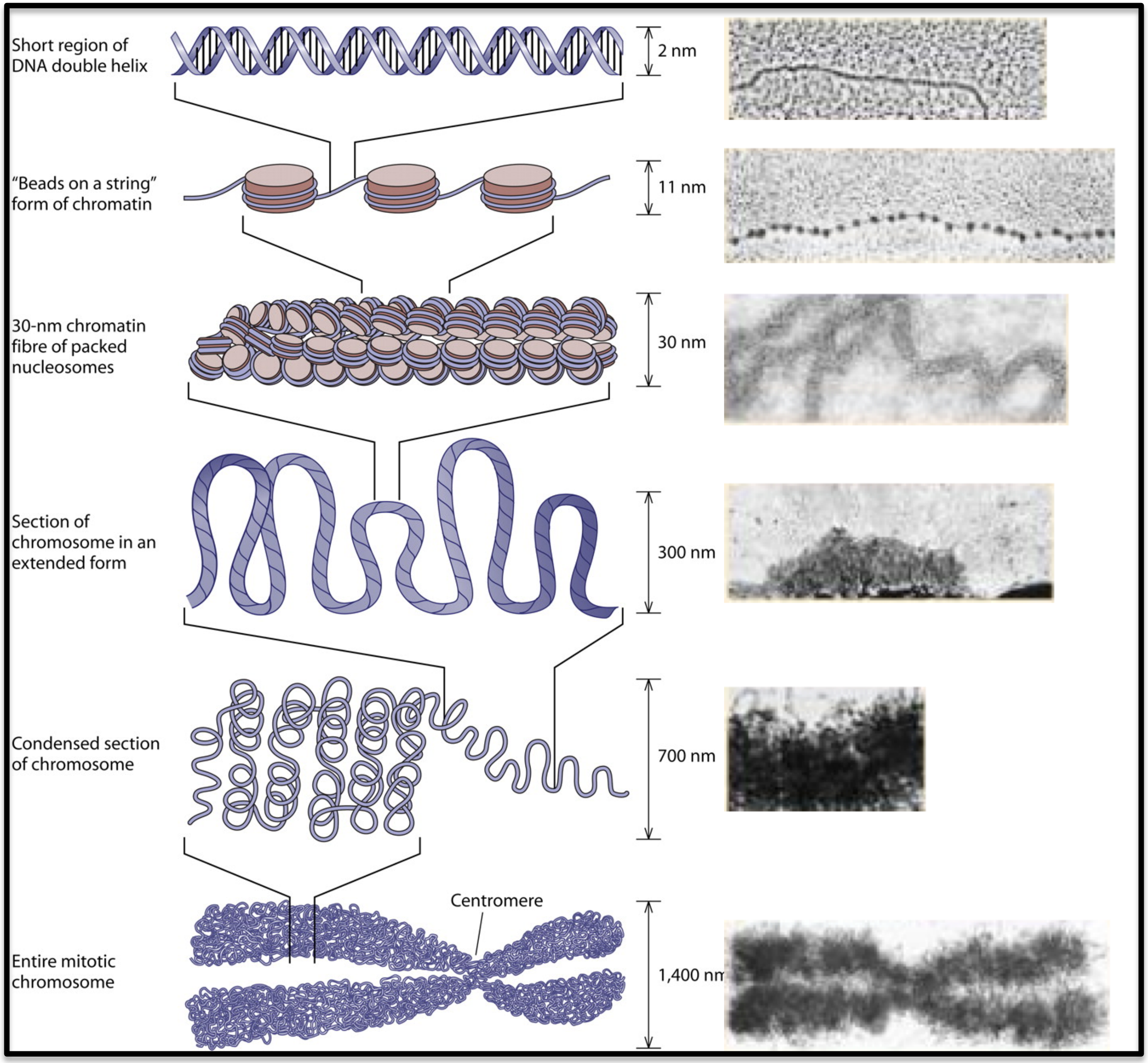
**Function:**

Packaging of DNA
Compaction of DNA
Definition of reagions of gene
Expression (euchromatin) or repression (heterochromatin)

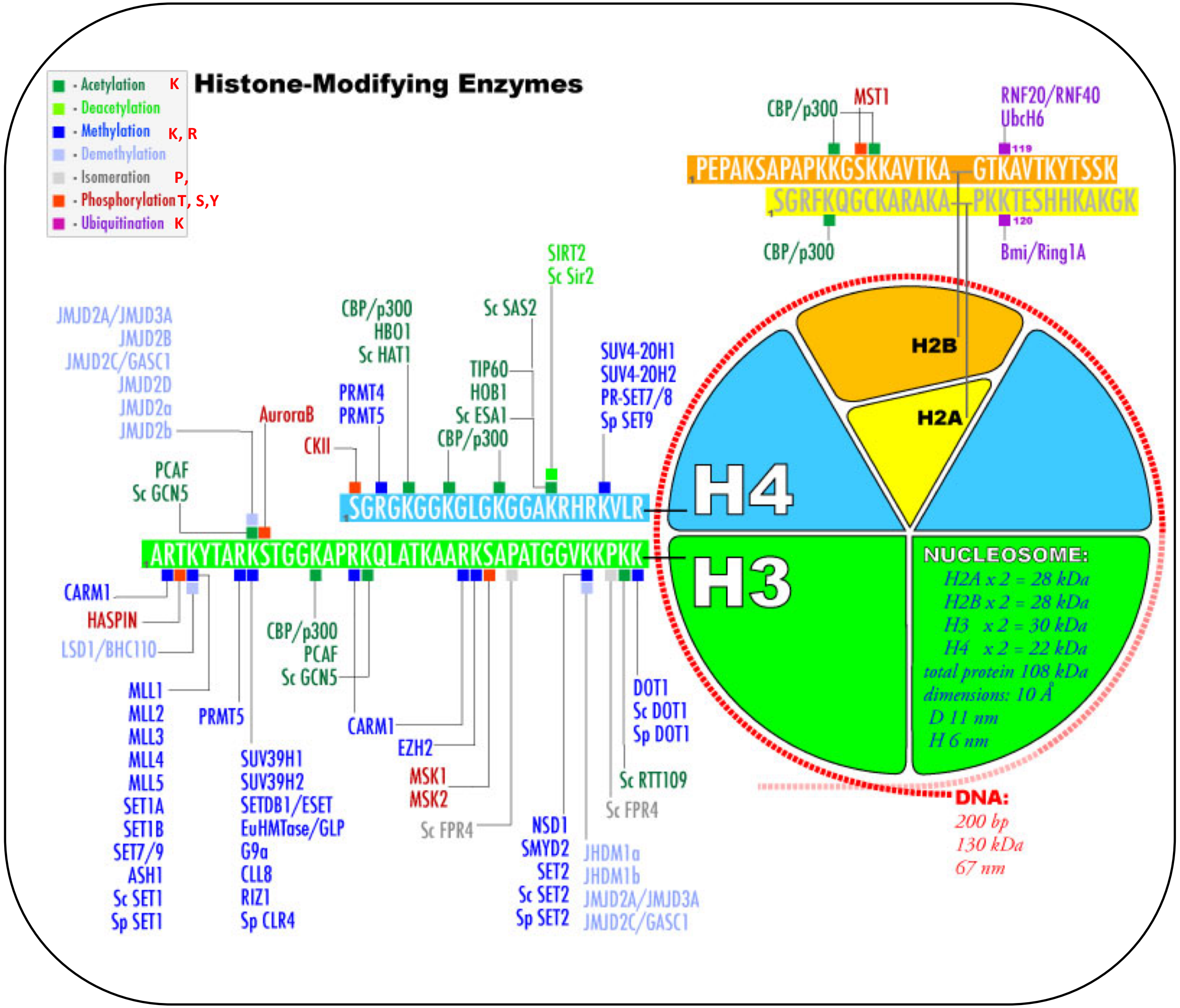
-Increasing stability of DNA
-Prevention of damage
-Control of replication, gene expression
-Cell cycle



**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

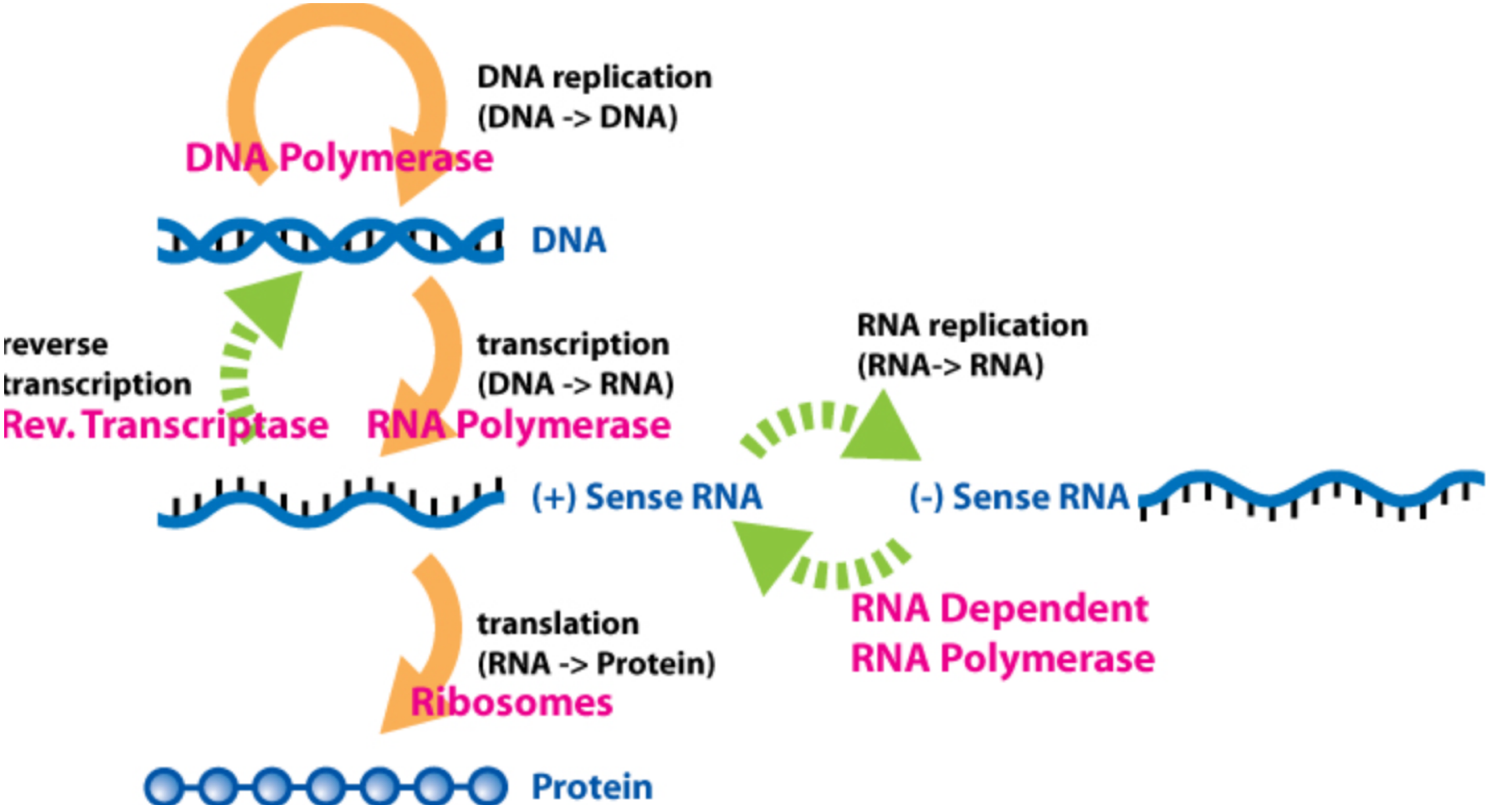
# The central dogma of molecular biology
# …and a protein centred point of view

DNA

RNA

Protein

Biological function

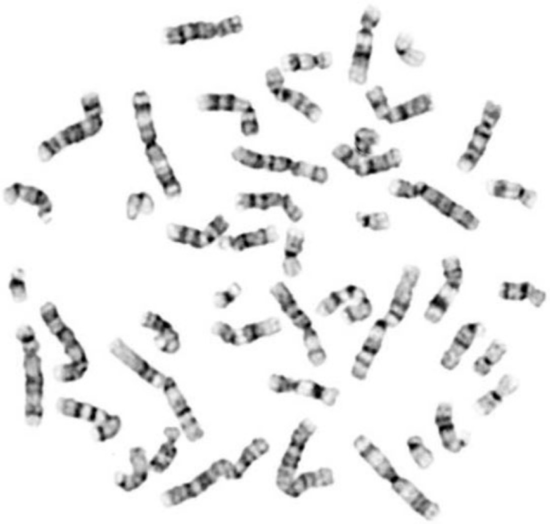


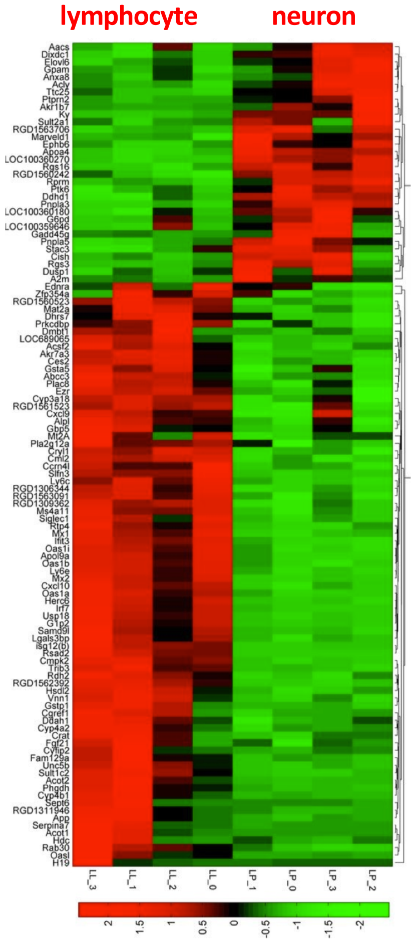
# The human genome encodes information that underlies cell specification in multi-cellular organisms

GENOME

Specific gene expression programs

Cell function





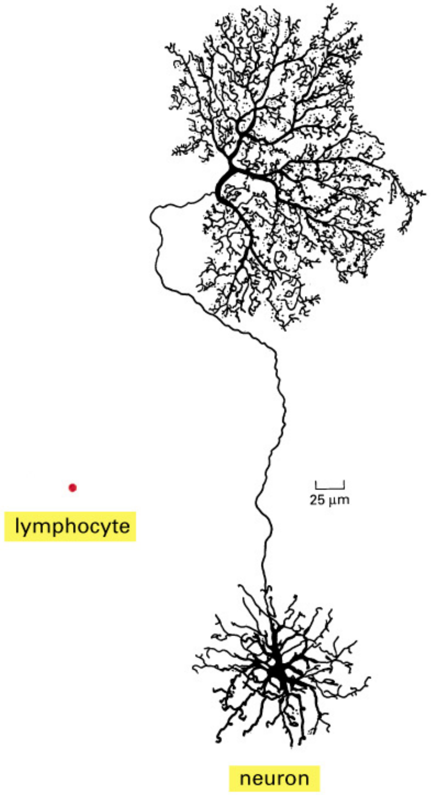
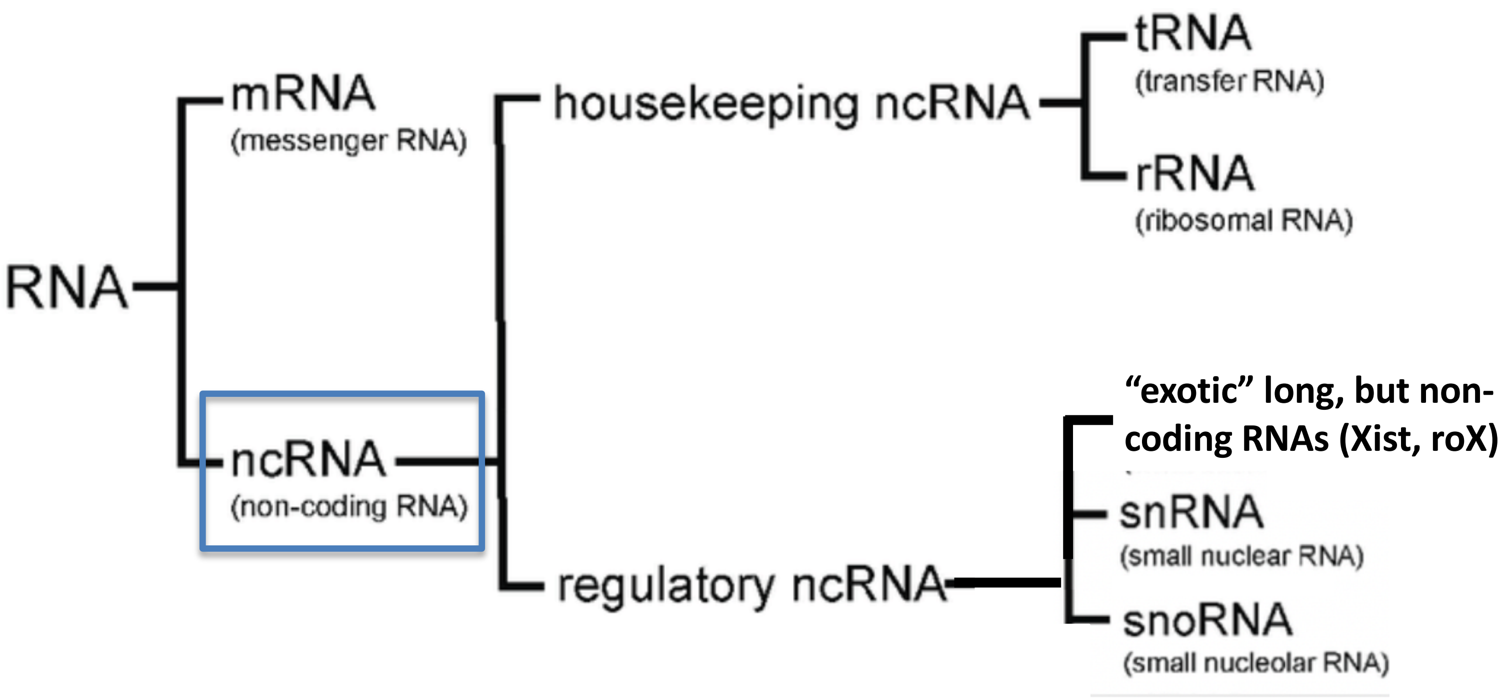


Figure 7–1. Molecular Biology of the Cell, 4th Edition.

*Genetic information must be highly organized*

# A classic view on eukaryotic non-coding RNAs

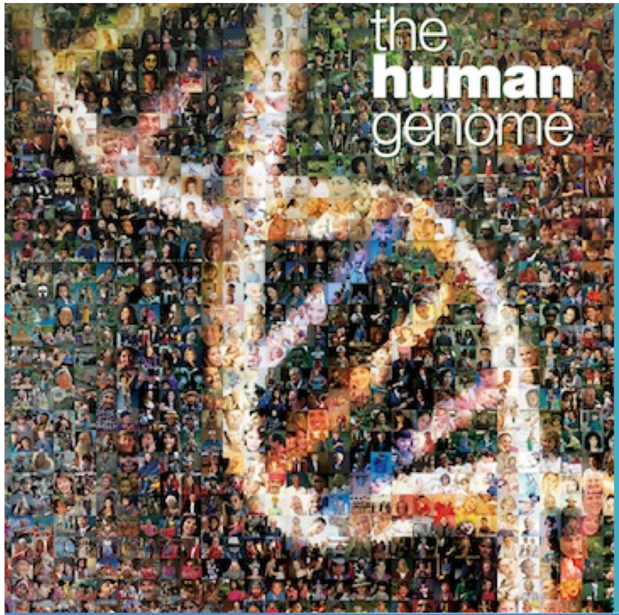
until late 1990ies



1. mRNAs → protein coding → development, differentiation, disease

2. **ncRNAs → defined biochemical activity to ensure mRNA processing and protein expression**

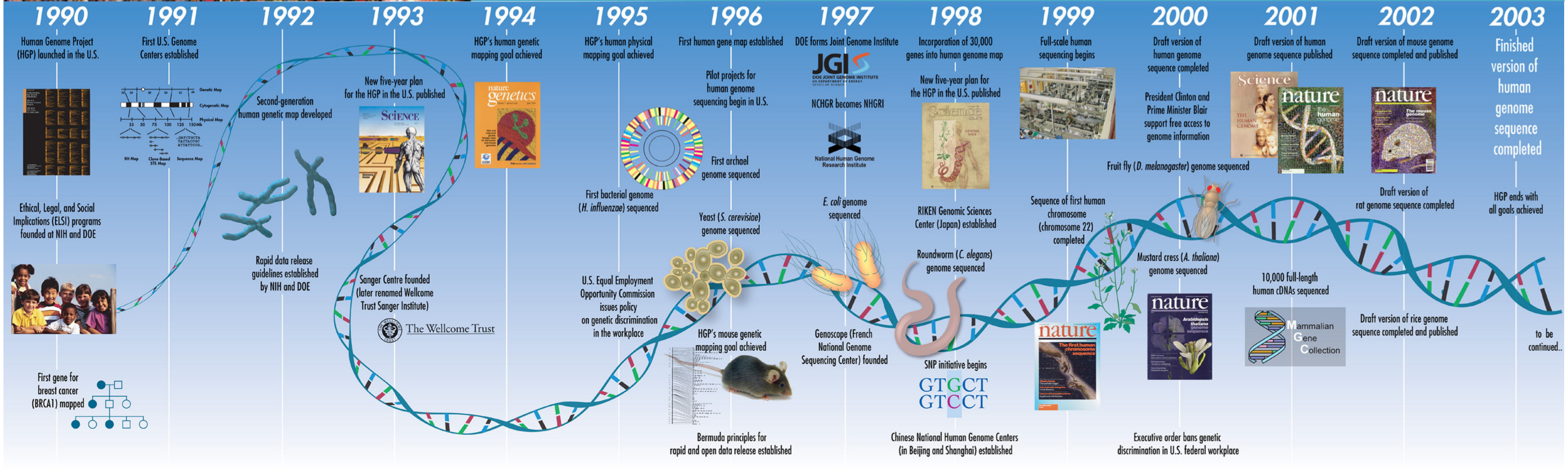
**3. A few "exotic RNAs" (such as Xist) → function identified due to genetic experiments, no idea on biochemical activity**



# WHOLE GENOME SEQEUNCING TO GRASP THE COMPLEXITY OF GENETIC INFORMATION

# THE HUMAN GENOME PROJECT

# SEQEUNCING OF 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

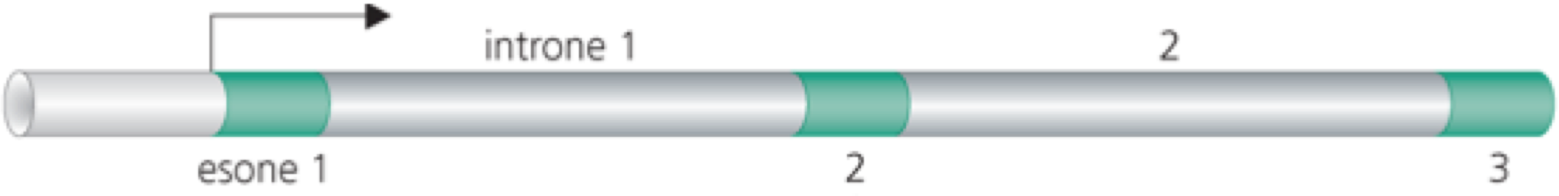
2000
Draft version of human genome sequence completed
President Clinton and Prime Minister Blair support free access to genome information
Fruit fly (*D. melanogaster*) genome sequenced
Mustard cress (*A. thaliana*) genome sequenced
Executive order bans genetic discrimination in U.S. federal workplace

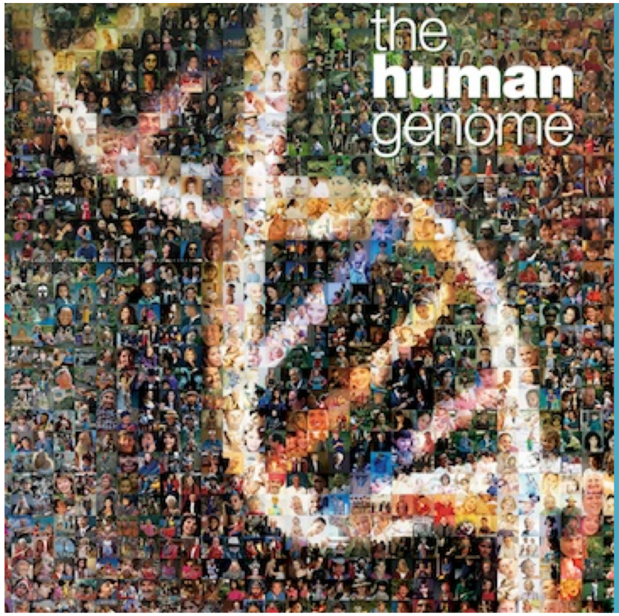
2001
Draft version of human genome sequence published
10,000 full-length human cDNAs sequenced

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!





# WHOLE GENOME SEQEUNCING TO GRASP THE COMPLEXITY OF GENETIC INFORMATION

# DNA SEQUENCING OF MULTIPLE SPECIES GAVE SURPRISES

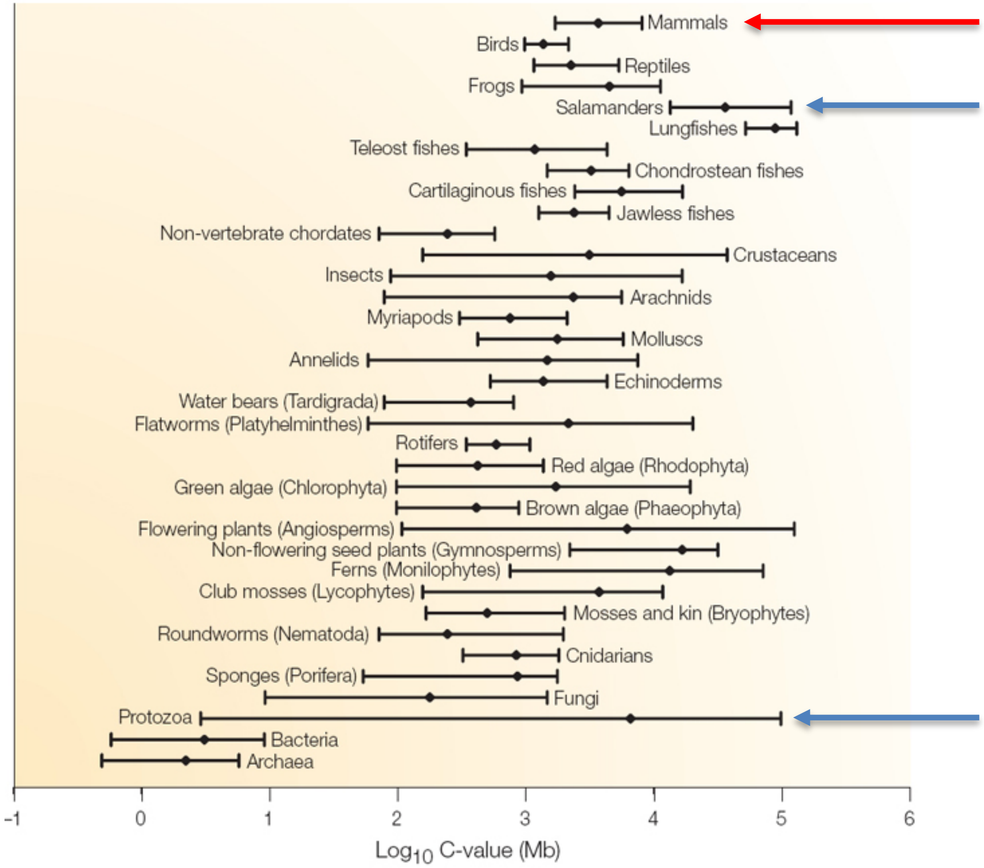
## Confirmation of the C-value paradox:

The amount of DNA in a haploid genome (the 1C value) does correspond strongly to the complexity of an organism. 1C values can be extremely variable.

## Vertebrates:

Only 1-2% of the genome is composed of exons that encode protein

## What DNA sequences are present in «junk» regions of genomic DNA?

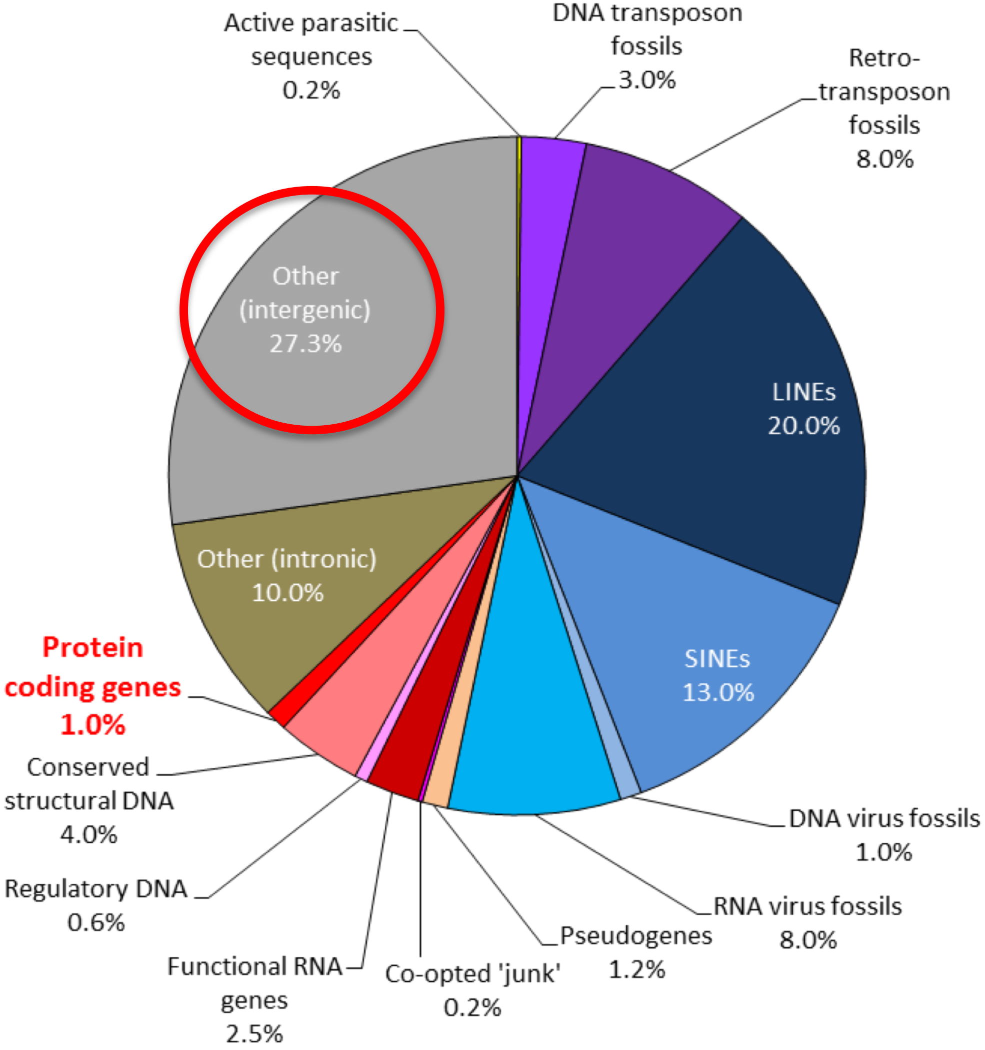


# 98% OF THE HUMAN GENOMIC DNA DOES NOT ENCODE FOR PROTEINS

**ca 50% transposable elements**

**1-2% protein coding genes**

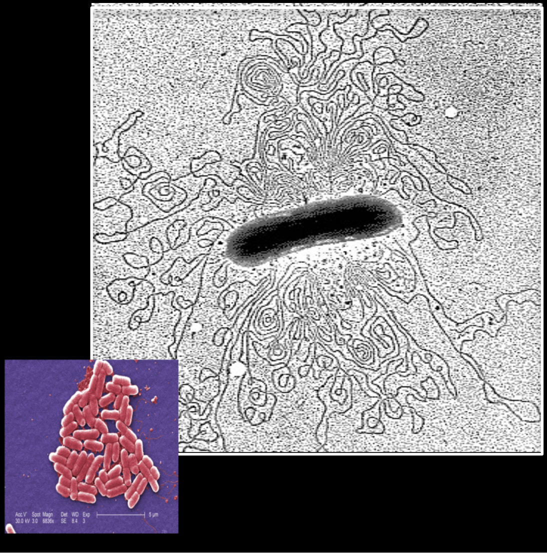
**0.5-1% pseudogenes**



**→ Vast genome sequnences without biological functions?**

# The C-value and G-value paradox

*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%** |

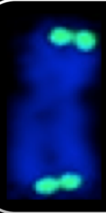
Disconnection of biological complexity and genome size:

- G-value paradox: number of genes does not correlate with genome size
- C-Value paradox: the amount of DNA in a haploid genome (the 1C value) does correspond strongly to the complexity of an organism.

***ncDNA CONTENT INCREASES WITH ORGANISMAL COMPLEXITY***

# Early examples of "exotic" functional ncRNAs

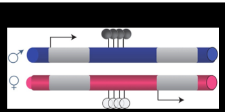
AGING-CANCER



Telomere function

hTR lncRNA

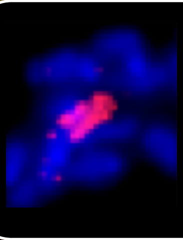
DEVELOPMENTAL DEFECTS



mono-allelic Gene expression

Airn lncRNA

Xist lncRNA



X inactivation Silencing 1000 genes on the Xi

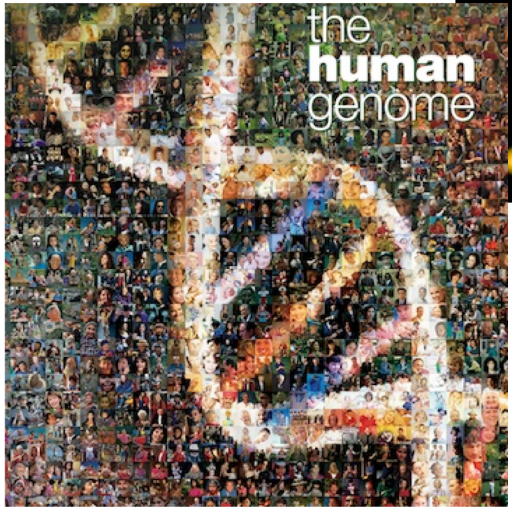
LETHAL

Relevant for development and disease

….others?

# Imagine you live in 2000…..

and you ask yourself a question:



….what can be done to identify new classes of RNAs that origin from non-coding regions and carry biological function

…what techniques to you apply

…are there techniques/instruments that can help in this quest?

Form 5 groups - 10 minutes discussion; 2 persons present ideas

# Classic automated sanger sequencing approaches are not sufficinet to capture transcriptome complexity

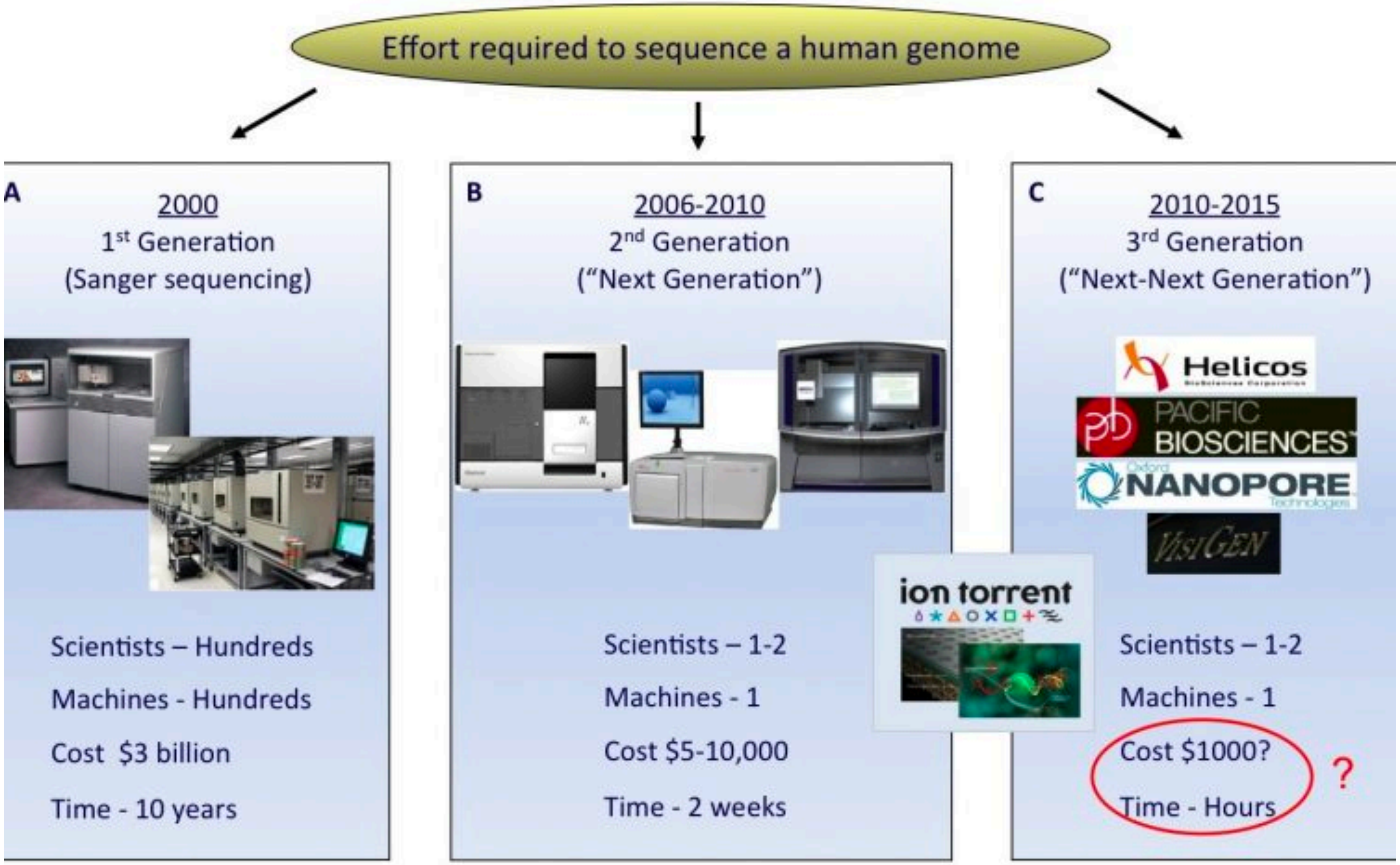
**What are the problems in the use of classic DNA sequencing techniques in the discovery of new functional elements (RNAs) in eukaryotes????????**

- Slow
- Cost intensive
- Labor intensive
- Biased towards highly expressed transcripts
- Non productive for sequencing of larger genomes/transciptomes
- Sequencing sample preparation for genome studies is labour intensive
- Combined, multiple analyses of particular sample is almost impossible (transcriptome and epigenome)

**What type of methods do we need to get a better resolution of the eukaryotic genome????**

- Fast method
- Cheap methods
- Efficient and reproducible methods
- Excellent detection of low abundance targets (low expressed RNAs)
- Multiple coverage of target with sequence reads
- Sequencing sample preparation also from limited source (small cell populations)
- Combined, multiple analyses of particular sample (transcriptome and epigenome) to link biological information

# A short wrap-up on sequencing techniques



Effort required to sequence a human genome

A

2000

1st Generation

(Sanger sequencing)

Scientists – Hundreds

Machines - Hundreds

Cost $3 billion

Time - 10 years

B

2006-2010

2nd Generation

("Next Generation")

Scientists – 1-2

Machines - 1

Cost $5-10,000

Time - 2 weeks

C

2010-2015

3rd Generation

("Next-Next Generation")

Scientists – 1-2

Machines - 1

Cost $1000?

Time - Hours

?

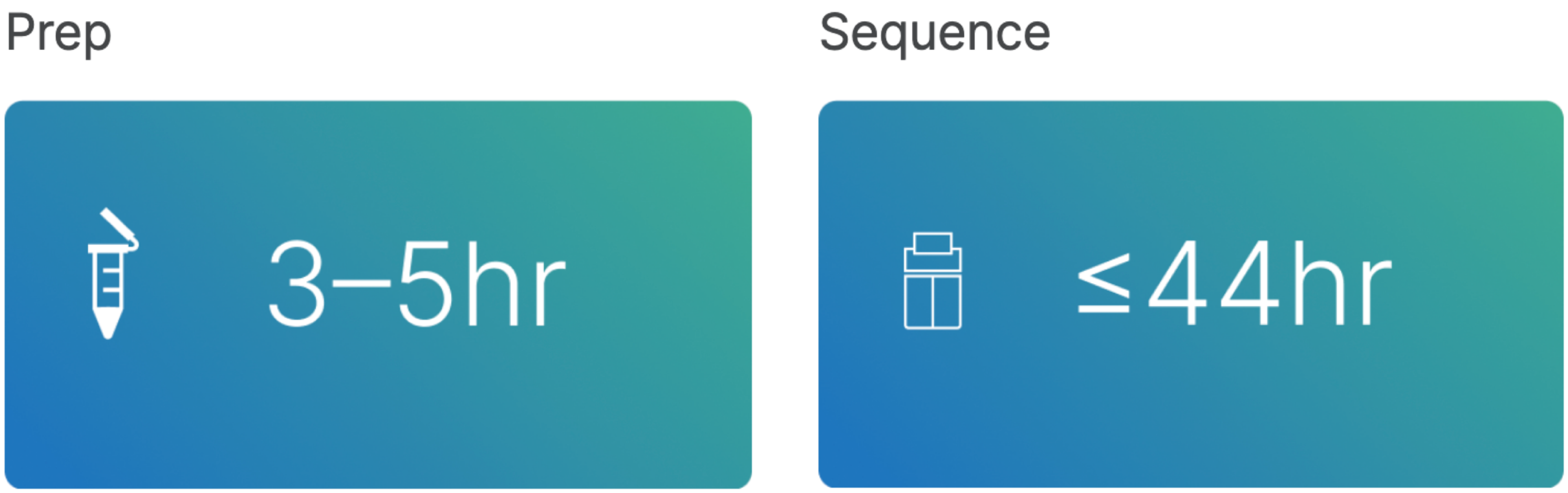
*2022: RNA seq: €400*
*2022: ChIP seq: €500*
*2023: RNA seq: €300*
*2023: ChIP seq: €350*

*…one vial of restriction enzyme: € 250*

# 3. Massive Parallel Sequencing – a revolution in genome sequencing

Output:

up to 6 Tb and 20 billion sequence reads in < 2 days.
typically 30-100 fold coverage
(each nucleotide in the sequnced DNA is represented by 30 – 100 seqeunce reads)



Prep

3–5hr

Sequence

≤44hr

**Data analysis:**

**Limiting factor: trained personnel, equipment and biological interpretation of the seqeuncing data**

**Experimental conditions and models systems need to be chosen carefully**

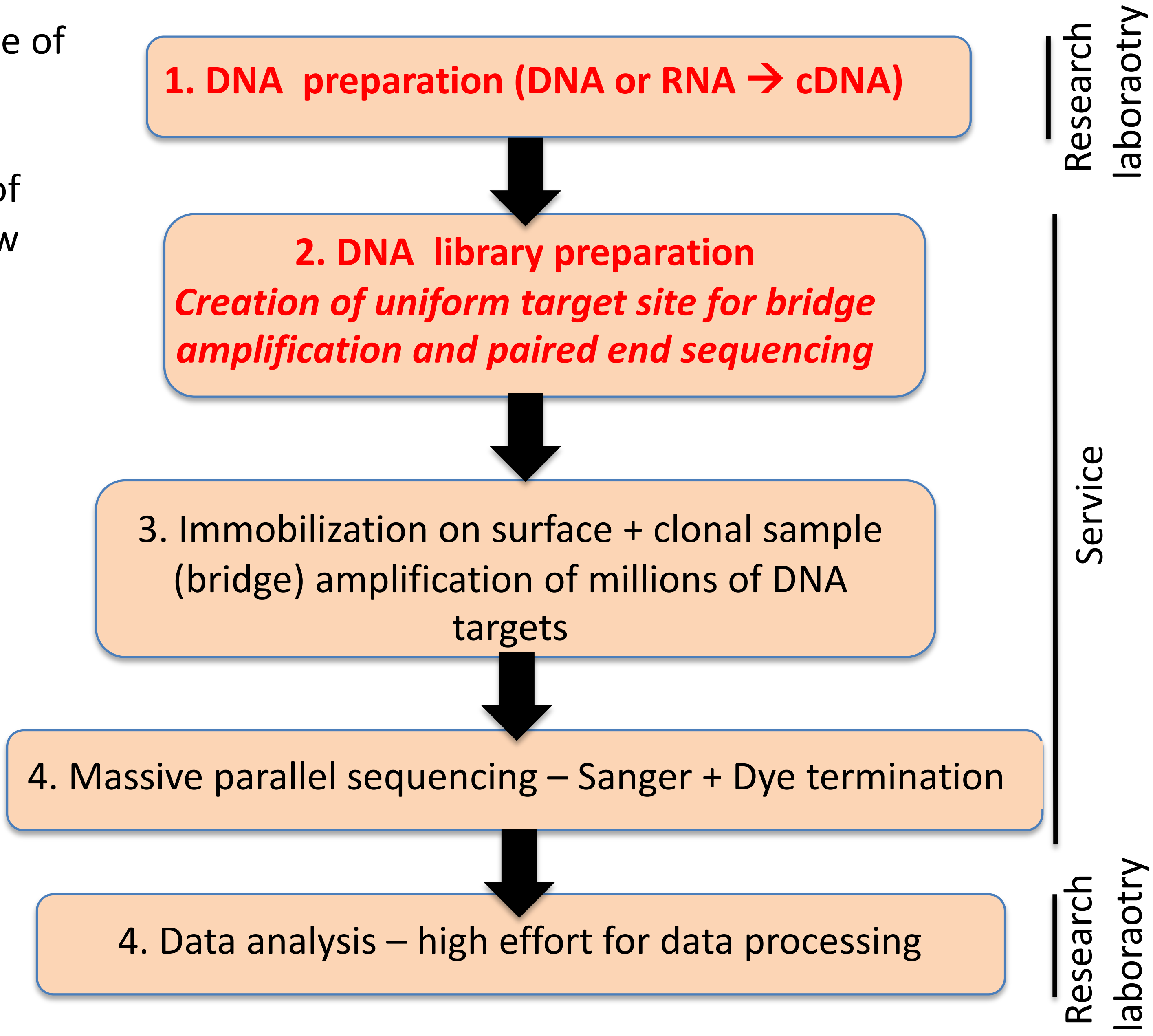
# 3. Massive Parallel Sequencing – how does it work?

**DNA seq** – genome sequence of many organisms

**RNA seq** – all RNAs (cDNA) of many organisms – also at low abundance

**ChiP seq**

**ATAC A**ssay for **T**ransposase-**A**ccessible **C**hromatin using **seq**uencing**…..**



1. DNA preparation (DNA or RNA → cDNA)

2. DNA library preparation
*Creation of uniform target site for bridge amplification and paired end sequencing*

3. Immobilization on surface + clonal sample (bridge) amplification of millions of DNA targets

4. Massive parallel sequencing – Sanger + Dye termination

4. Data analysis – high effort for data processing

# 3b Illumina: massive parallel sequencing: RNA seq

## RNA seq allows the sequencing of different classes of RNA

1. Selection of cell type/tissue/organism

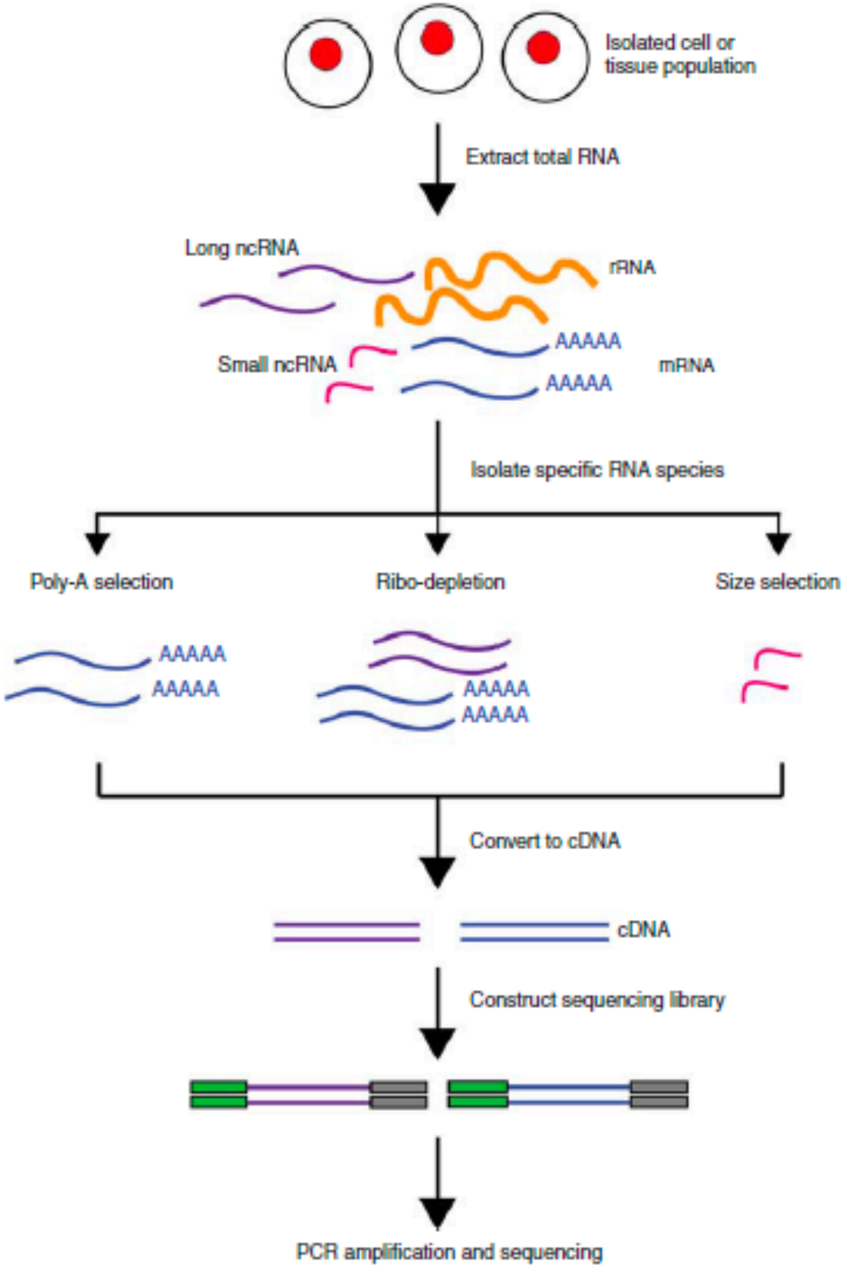
2. Fractionation of RNA types

3. cDNA synthesis and end polishing

4. Adapter ligation
(sequencing primer target site)

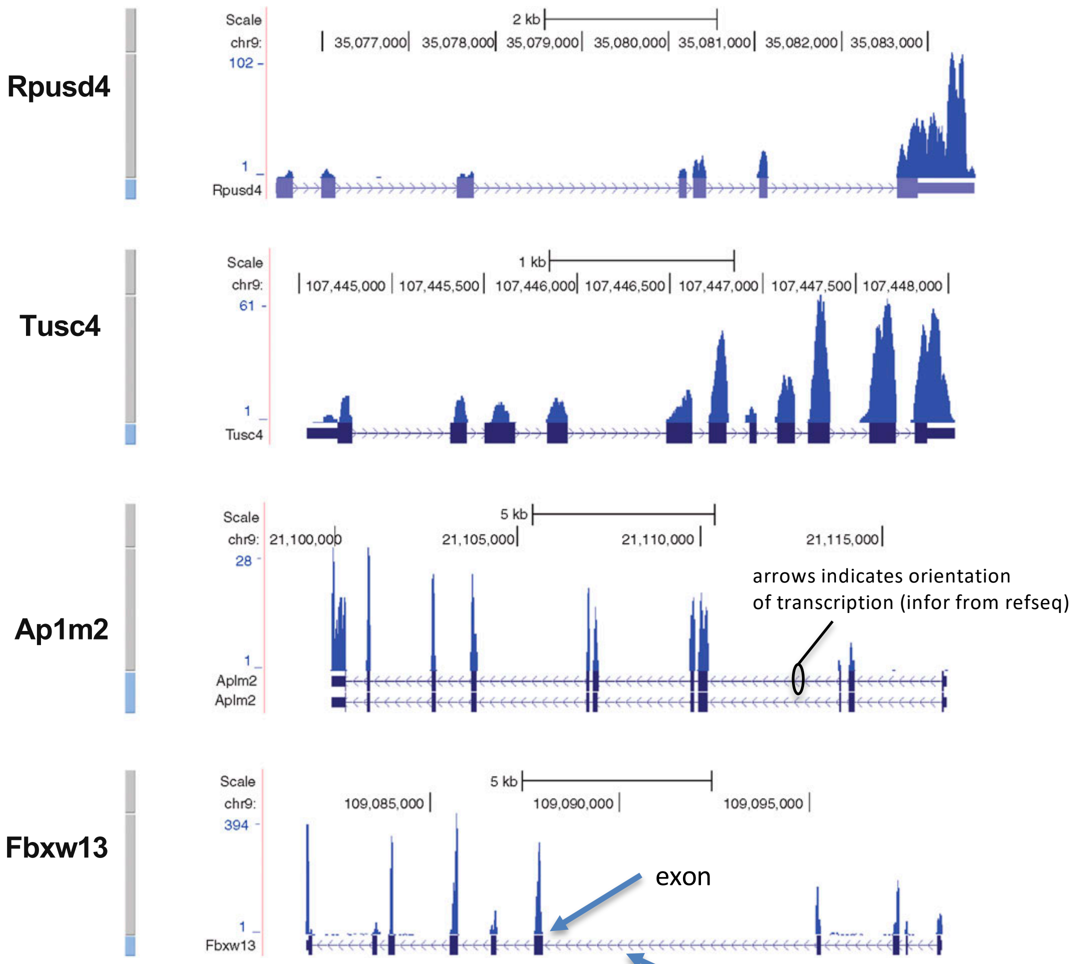
5. PCR amplification

6. Cluster amplification and sequencing in flow cell



# Coverage blot of typical RNA seq experiment

Zoom into defined region in the genome of protein coding genes



arrows indicates orientation of transcription (infor from refseq)

exon

intron

Coverage plots of RNA-Seq reads from a single wild-type mature oocyte. Analysis was performed using the UCSC genome browser. Depicted here are the base coverage files for Rpusd4, Tusc4, Ap1m2 and Fbxw13 on chromosome 9.

**Sequence reads build up to peak**

**Peaks build up on exonic sequences in reference genome**

**No peaks on intronic sequences (degraded RNA)**

## Qualitative Information

**The start and end of a peak allows to identify a defined (mature) RNA unit with respect to the reference sequence (i.e. exon: start and end point) (i.e. miRNA: start and end of processed miRNA)**

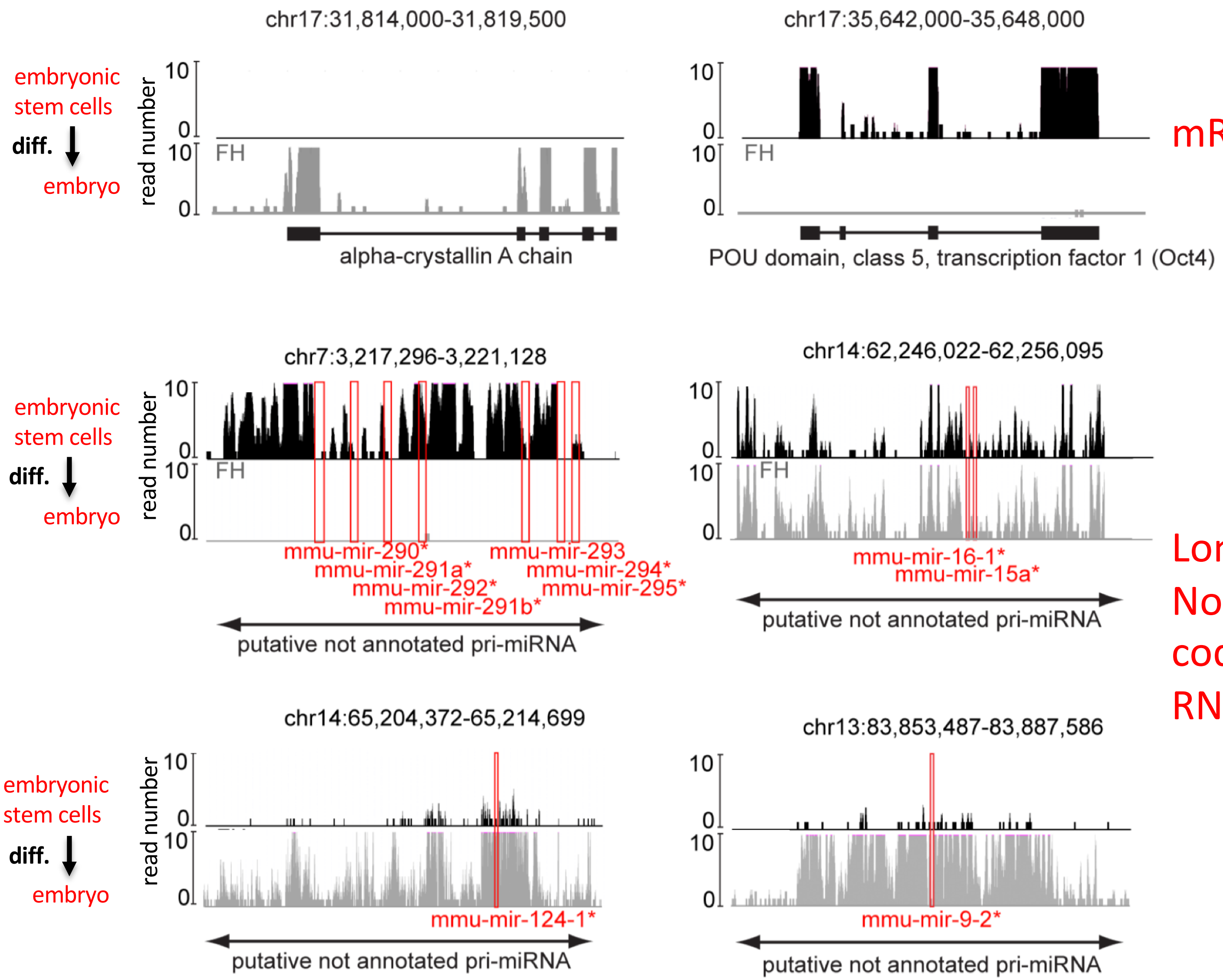
## Quantitative information

**Computational analysis of sequencing data allows to Correlate coverage blot of individual transcripts to gene expression levels**

## Tissue specific expression

**RNAseq in different tissues, cell types or differentiation (differential processing, expression)**
**RNAseq in different species (conservation)**
**Single cell information**

# Non-coding RNAs are complex in their expression and processing into mature ncRNA



mRNA

**Alpha crystallin A chain:**

expressed in embryo not ES cells,
clear exon intron structure
no other significant reads

**Oct4:**

expressed in ES cells,
clear exon intron structure
no other significant reads

Long Non-coding RNA

**Non-coding pri-miRNAs**

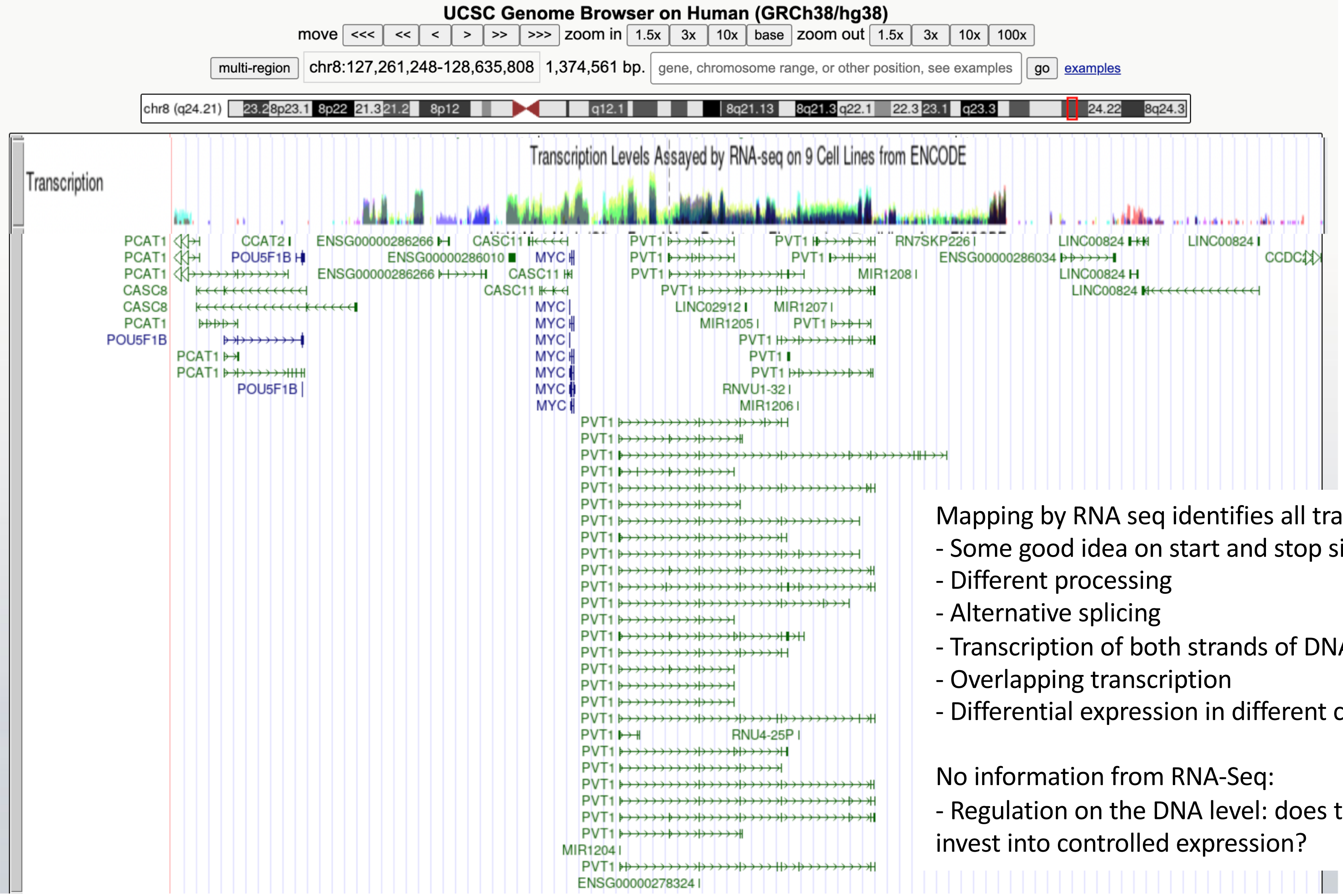
Reads map along a longer stretch on Chr7/14

Chr7 lncRNA: differential regulation of expression

Chr14 lncRNA: common expression

Annotation with miRNAs (processing known)

# RNA Seq identifies new type of RNA elements – coding and non-coding


UCSC Genome Browser on Human (GRCh38/hg38)
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region chr8:127,261,248-128,635,808 1,374,561 bp. gene, chromosome range, or other position, see examples go examples
chr8 (q24.21) 23.2 8p23.1 8p22 21.3 21.2 8p12 q12.1 8q21.13 8q21.3 q22.1 22.3 23.1 q23.3 24.22 8q24.3
Transcription
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
PCAT1
PCAT1
PCAT1
CASC8
CASC8
PCAT1
POU5F1B
CCAT2
POU5F1B
PCAT1
PCAT1
POU5F1B
ENSG00000286266
ENSG00000286010
ENSG00000286266
CASC11
CASC11
CASC11
MYC
MYC
MYC
MYC
MYC
MYC
MYC
MYC
MYC
MYC
PVT1
PVT1
PVT1
PVT1
LINC02912
MIR1205
PVT1
PVT1
PVT1
RNVU1-32
MIR1206
PVT1
PVT1
MIR1207
PVT1
RN7SKP226
MIR1208
ENSG00000286034
LINC00824
LINC00824
LINC00824
LINC00824
CCDC
PVT1
RNU4-25P
MIR1204
ENSG00000278324

Mapping by RNA seq identifies all transcipts.
- Some good idea on start and stop site
- Different processing
- Alternative splicing
- Transcription of both strands of DNA
- Overlapping transcription
- Differential expression in different cell types

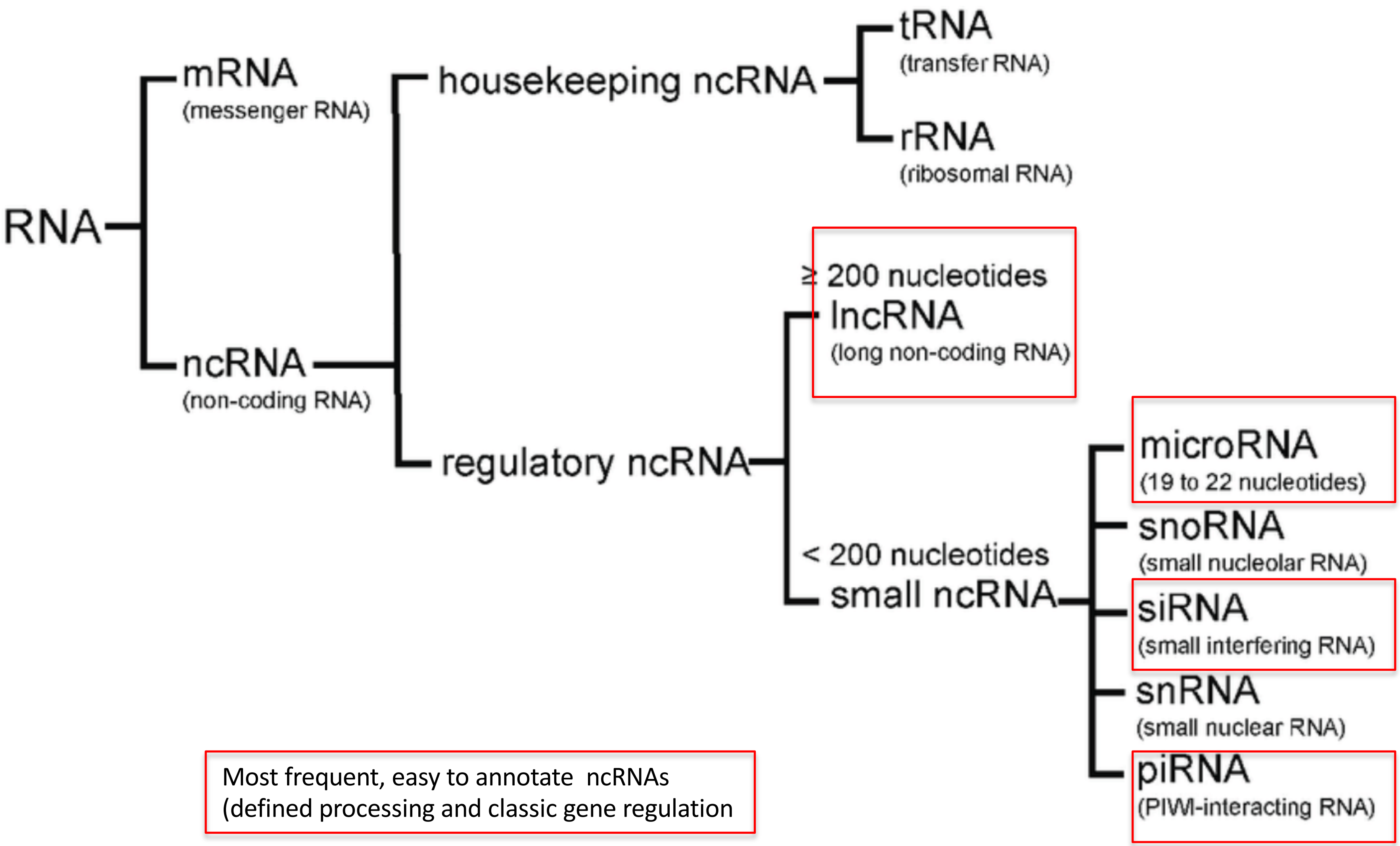
No information from RNA-Seq:
- Regulation on the DNA level: does the cell invest into controlled expression?

Blue: protein coding transcript

Green: non-coding transcripts of different types (miRNA, lncRNA)

Note: all variants identified in different cell types are shown

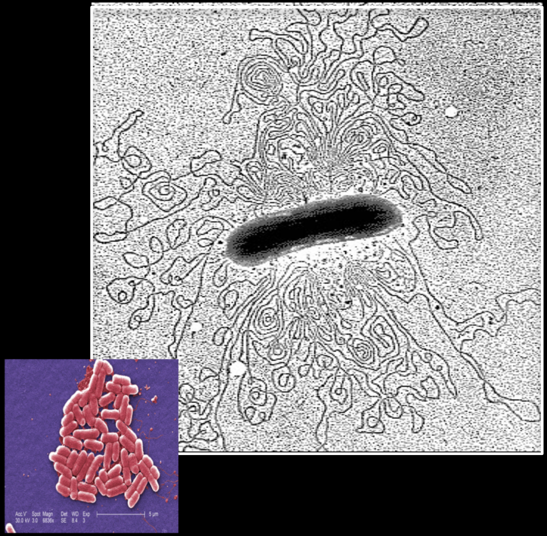
# RNA Seq identifies new type of RNA elements



Most frequent, easy to annotate ncRNAs
(defined processing and classic gene regulation

# The non-coding genome (r)evolution

*E.coli*

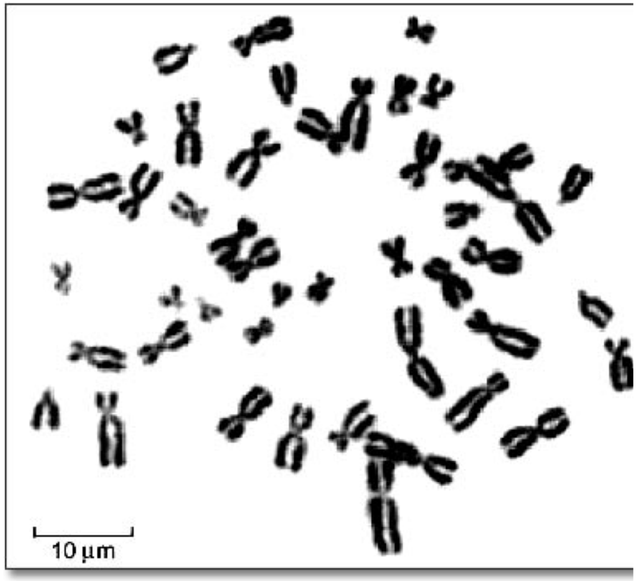


*C. elegans*

*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 |

# Almost all regions in the genome are subject to regulation and transcription – what about non-coding gene regulation?



**The vast majority (80,4%) of the human genome in at least one biochemical RNA event in at least one cell type**

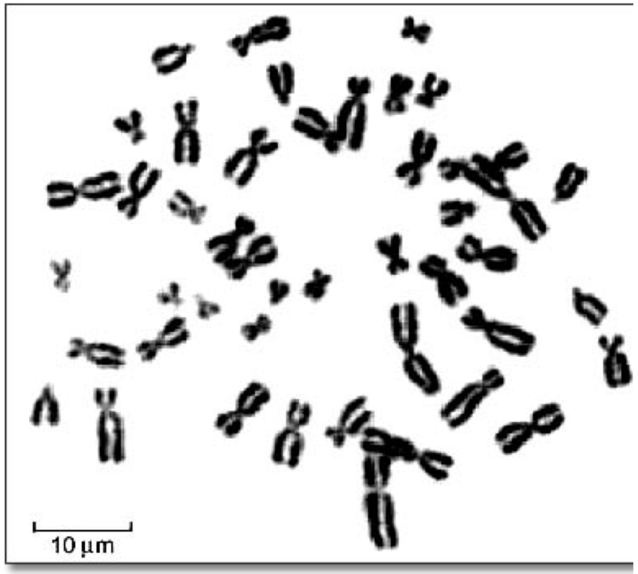
**Coding transcripts are expressed at relatively high levels, Non-coding RNAs tend to be expressed at low levels**

**How would you define a potential functionally relevant transcripts?**

**Can we use genomics data to propose functionally relevant transcripts? How?**

Think about 10 minutes and propose an idea!

# 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.

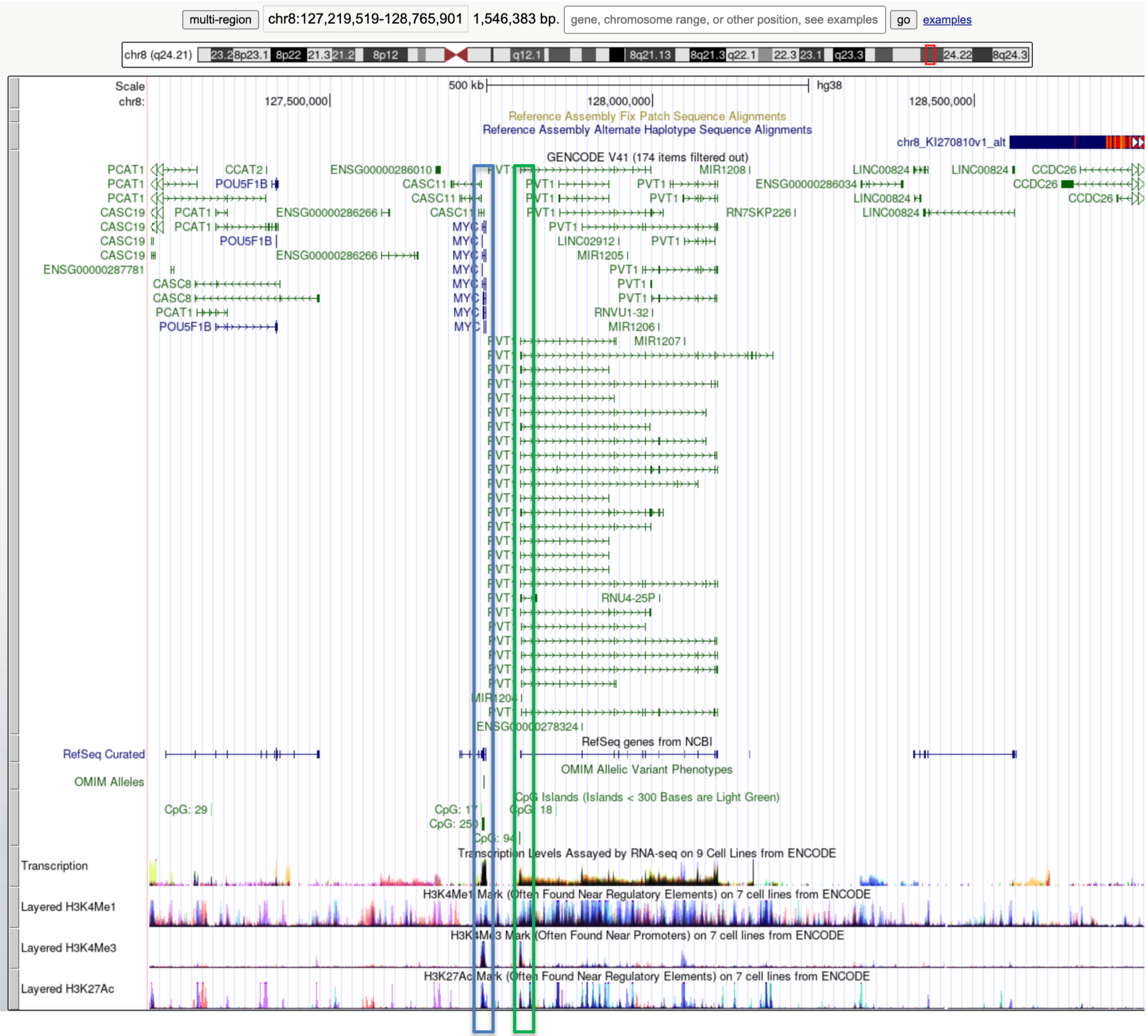
***THIS FINDING SHOWS THAT COMPLEXITY GOES WIDE BEYOND PROTEIN CODING GENES***

***THOUSANDS OF NEW GENES THAT HAVE NEVER BEEN STUDIED BEFORE***

# *Combine RNAseq data with known chromatin features at gene promoters*

🟩 Coding gene

🟦 Non-coding gene



multi-region
chr8:127,219,519-128,765,901
1,546,383 bp.
gene, chromosome range, or other position, see examples
go
examples
chr8 (q24.21)
Scale
chr8:
500 kb
hg38
127,500,000
128,000,000
128,500,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
chr8_KI270810v1_alt
GENCODE V41 (174 items filtered out)
PCAT1
CASC19
CCAT2
POU5F1B
CASC8
ENSG00000287781
ENSG00000286010
ENSG00000286266
CASC11
MYC
PVT1
MIR1208
MIR1205
MIR1206
MIR1207
MIR1204
LINC02912
RNVU1-32
RNU4-25P
RN7SKP226
ENSG00000286034
ENSG00000278324
LINC00824
CCDC26
RefSeq Curated
RefSeq genes from NCBI
OMIM Alleles
OMIM Allelic Variant Phenotypes
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 29
CpG: 17
CpG: 250
CpG: 94
CpG: 18
Transcription
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
Layered H3K4Me1
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me3
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE

**Chromatin features**

histones RNA CpG

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

## → Mapping transcription units

## → Mapping regulatory units

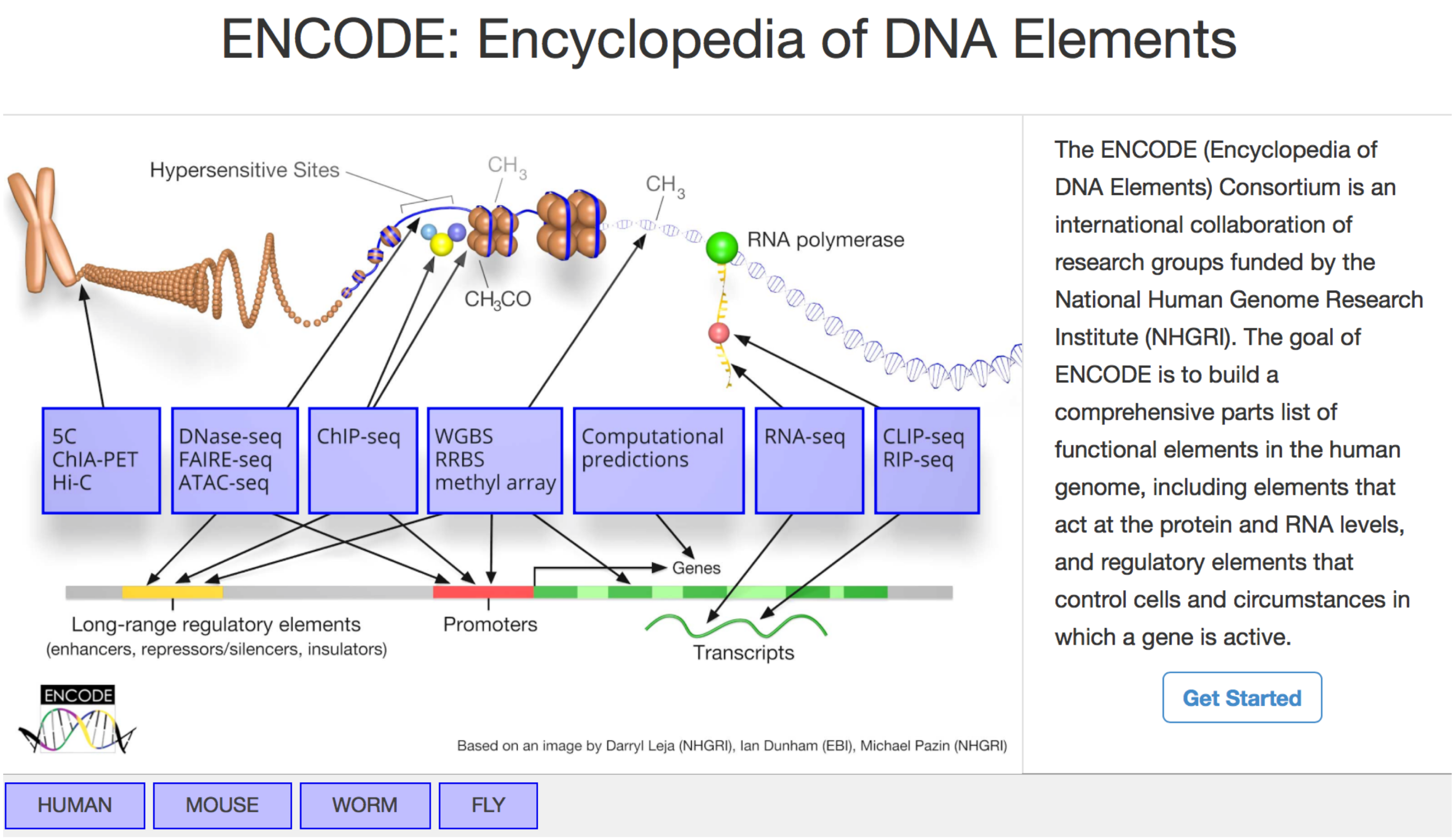
The Encyclopedia of DNA Elements (ENCODE) is a public research project launched by the US National Human Genome Research Institute (NHGRI) in September 2003.

**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**

# ENCODE MASSIVE EXPERIMENTAL INPUT

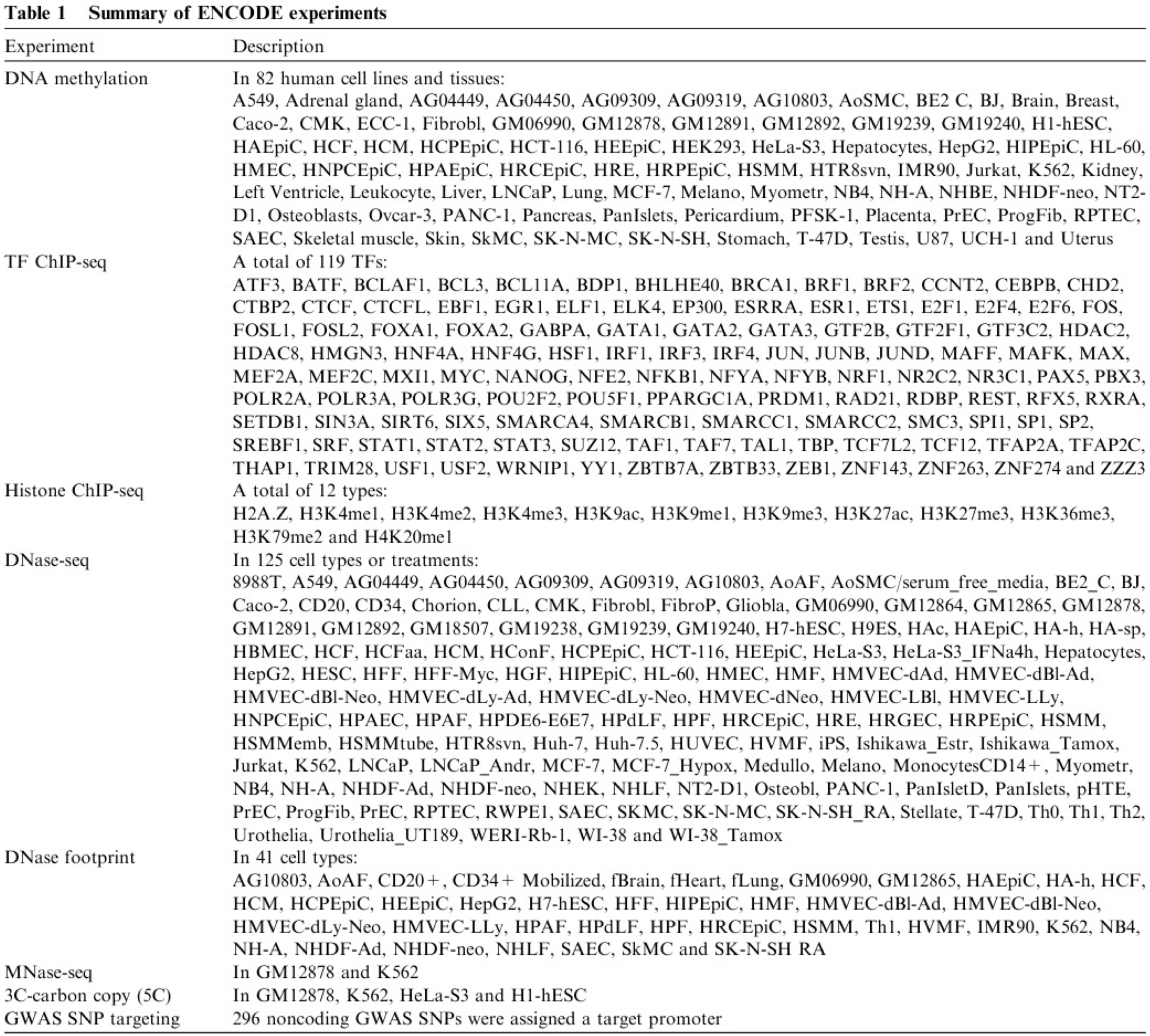


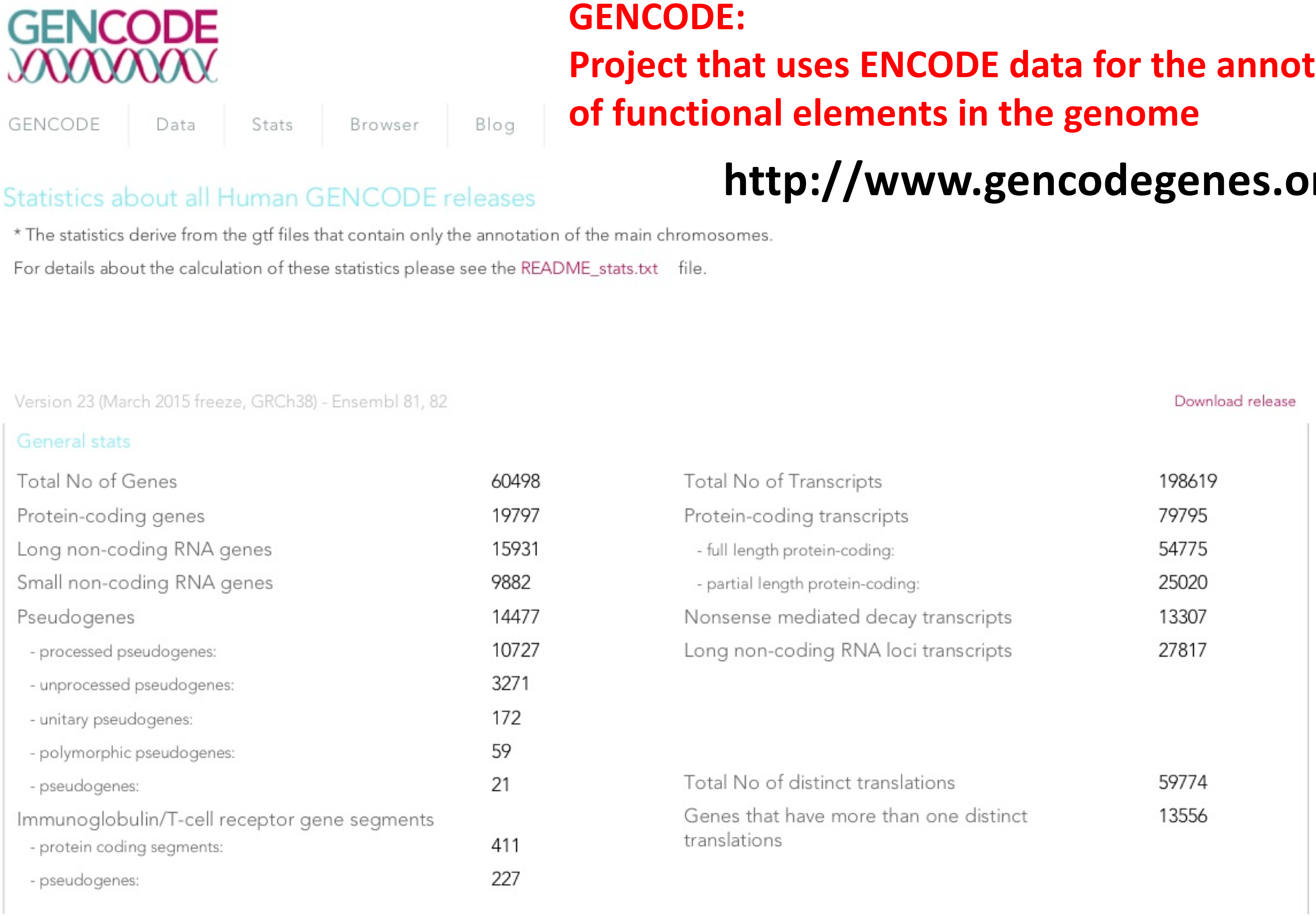
**Table 1 Summary of ENCODE experiments**

| Experiment | Description |
|---|---|
| DNA methylation | In 82 human cell lines and tissues:<br>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:<br>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:<br>H2A.Z, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me1, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K79me2 and H4K20me1 |
| DNase-seq | In 125 cell types or treatments:<br>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:<br>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

General stats

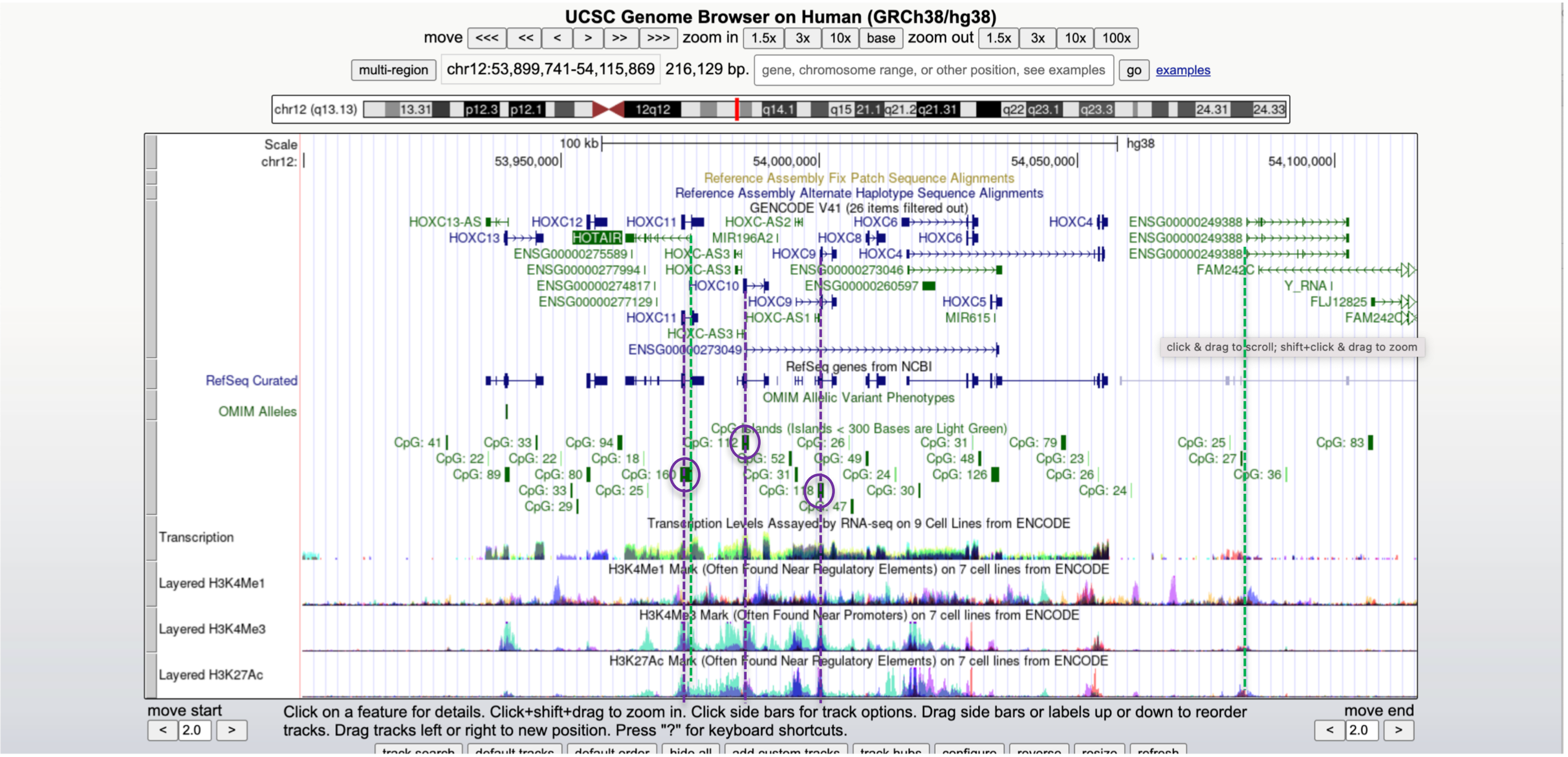
| | | | |
|---|---|---|---|
| Total No of Genes | 60498 | Total No of Transcripts | 198619 |
| Protein-coding genes | 19797 | Protein-coding transcripts | 79795 |
| Long non-coding RNA genes | 15931 | - full length protein-coding: | 54775 |
| Small non-coding RNA genes | 9882 | - partial length protein-coding: | 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 | | |

**HOME WORK:**

Use the **UCSC genome browser** ( https://genome.ucsc.edu/index.html) to check coding and ncRNA expression in a Hox gene cluster

- Select human genome,
- check for HOXC9 gene region (use zoom in and our function to visualize a genome region of ca. 100kb up and 100 kb downstream of HOXC9
- 1. Get an imagination on how many coding and non coding genes are in this region
- Go to tool selection: select CpG Island function; select ENCODE regulation
- Click on gray bar next to ChIP peaks; click integrated regulation from encode tracks; select ”transcription, layered H3K4Me1, layered H3K4Me3, layered H3K27”, select “submit”
- 2. Try to individuate peaks in ChIP data that have a particular pattern with respect to transcripts

- Be ready to show next lecture!!

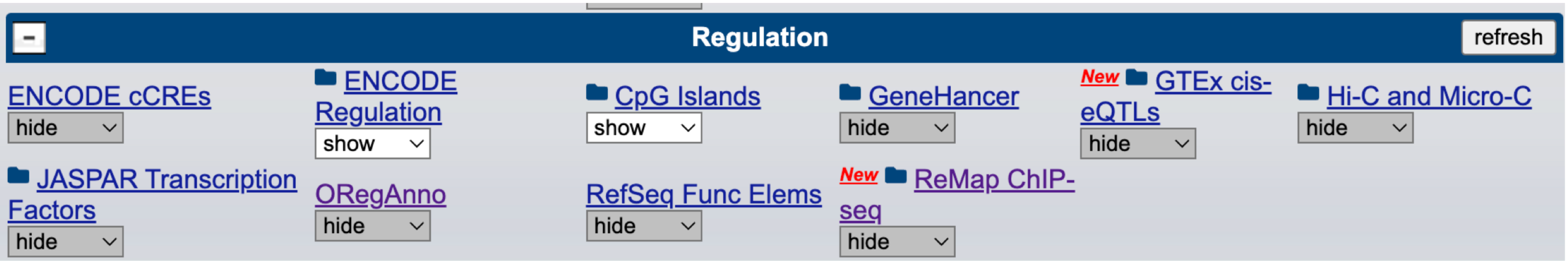


UCSC Genome Browser on Human (GRCh38/hg38)

chr12:53,899,741-54,115,869 216,129 bp.

Click on a feature for details. Click+shift+drag to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region
gene, chromosome range, or other position, see examples
go
examples
chr12 (q13.13)
13.31 p12.3 p12.1 12q12 q14.1 q15 21.1 q21.2 q21.31 q22 q23.1 q23.3 24.31 24.33
Scale
chr12:
100 kb
hg38
53,950,000
54,000,000
54,050,000
54,100,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V41 (26 items filtered out)
HOXC13-AS
HOXC12
HOXC11
HOXC-AS2
HOXC6
HOXC4
ENSG00000249388
HOXC13
HOTAIR
MIR196A2
HOXC8
HOXC6
ENSG00000249388
ENSG00000275589
HOXC-AS3
HOXC9
HOXC4
ENSG00000249388
ENSG00000277994
HOXC-AS3
ENSG00000273046
FAM242C
ENSG00000274817
HOXC10
ENSG00000260597
Y_RNA
ENSG00000277129
HOXC9
HOXC5
FLJ12825
HOXC11
HOXC-AS1
MIR615
FAM242C
HOXC-AS3
ENSG00000273049
click & drag to scroll; shift+click & drag to zoom
RefSeq genes from NCBI
RefSeq Curated
OMIM Allelic Variant Phenotypes
OMIM Alleles
CpG Islands (Islands < 300 Bases are Light Green)
CpG: 41
CpG: 33
CpG: 94
CpG: 112
CpG: 26
CpG: 31
CpG: 79
CpG: 25
CpG: 83
CpG: 22
CpG: 22
CpG: 18
CpG: 52
CpG: 49
CpG: 48
CpG: 23
CpG: 27
CpG: 89
CpG: 80
CpG: 160
CpG: 31
CpG: 24
CpG: 126
CpG: 26
CpG: 36
CpG: 33
CpG: 25
CpG: 118
CpG: 30
CpG: 24
CpG: 29
CpG: 47
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
Transcription
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me1
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
Layered H3K4Me3
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
move start
< 2.0 >
move end
< 2.0 >
track search default tracks default order hide all add custom tracks track hubs configure reverse resize refresh



ENCODE cCREs: hide

ENCODE Regulation: show

CpG Islands: show

GeneHancer: hide

*New* GTEx cis-eQTLs: hide

Hi-C and Micro-C: hide

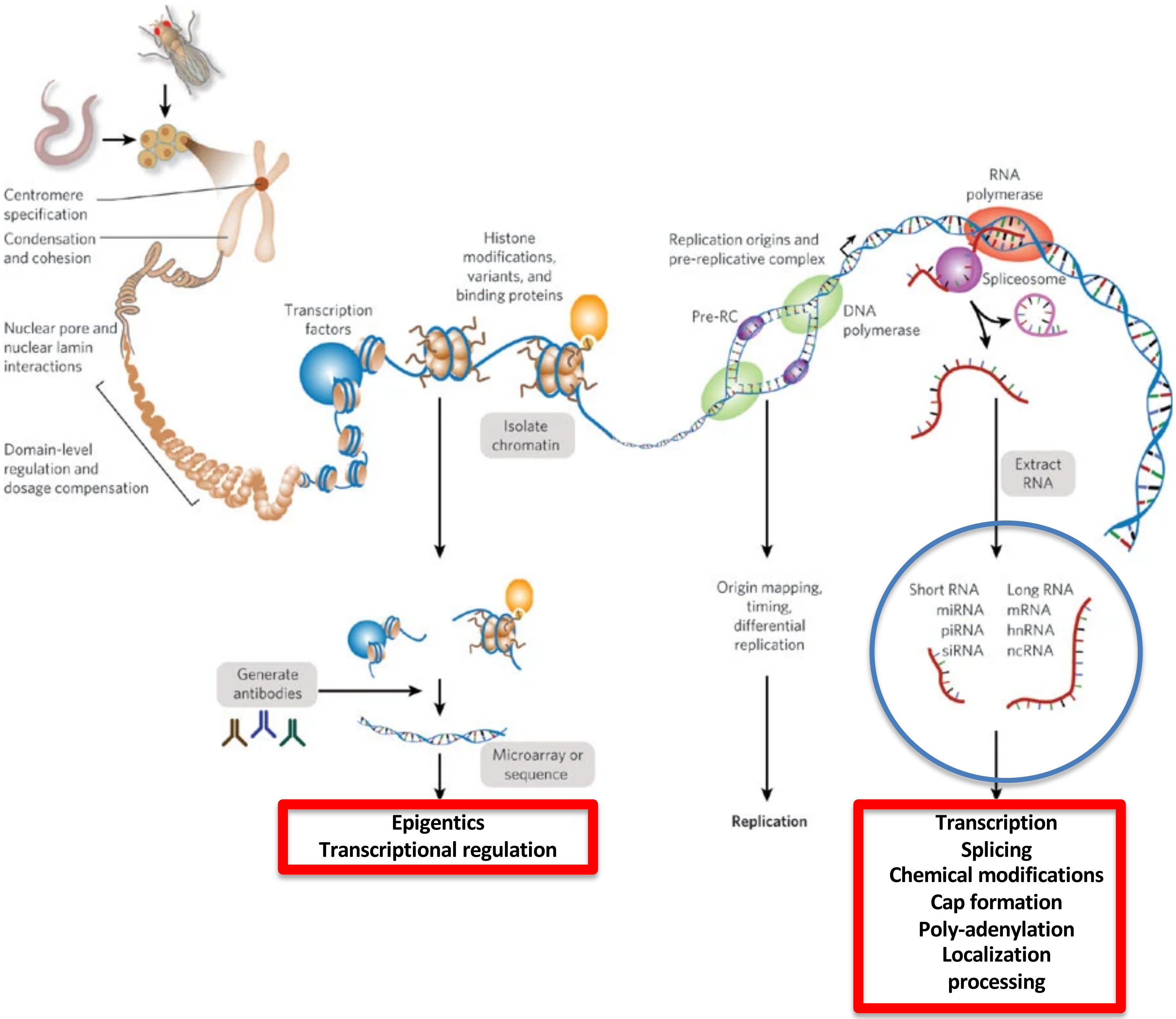
JASPAR Transcription Factors: hide

ORegAnno: hide

RefSeq Func Elems: hide

*New* ReMap ChIP-seq: hide

# How to unlock the genome and identify new functional elements ??


Centromere specification
Condensation and cohesion
Nuclear pore and nuclear lamin interactions
Domain-level regulation and dosage compensation
Transcription factors
Histone modifications, variants, and binding proteins
Isolate chromatin
Replication origins and pre-replicative complex
Pre-RC
DNA polymerase
RNA polymerase
Spliceosome
Extract RNA
Generate antibodies
Microarray or sequence
Origin mapping, timing, differential replication
Short RNA
miRNA
piRNA
siRNA
Long RNA
mRNA
hnRNA
ncRNA
Replication

**Epigentics**
**Transcriptional regulation**

**Transcription**
**Splicing**
**Chemical modifications**
**Cap formation**
**Poly-adenylation**
**Localization**
**processing**

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

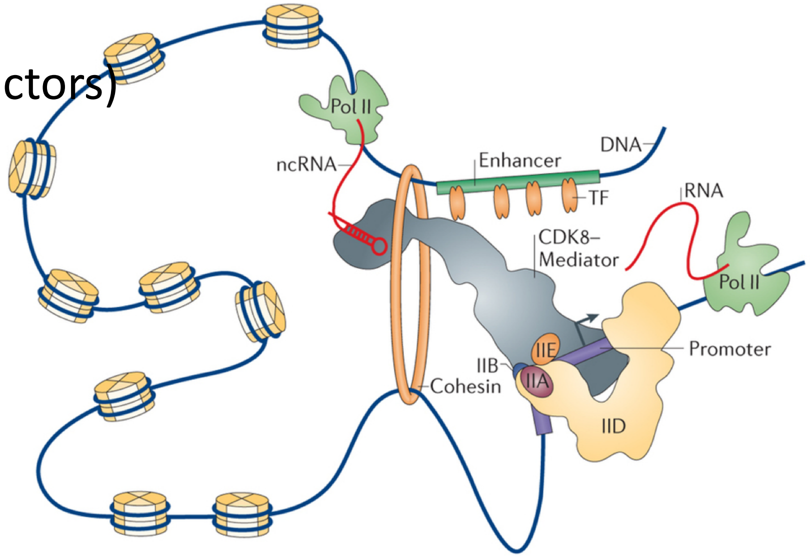
1. DNA Seqeuncing (Human genome project, DNA-Seq) → ALREADY DONE

**2. Landscape of transcription: Sequencing of RNA (total RNA, small/large RNA, CAGE)**
- **Determine the transcriptome of a give cell or cell type (bulk RNA sequencing)**
- **Determine the intiation site of transcription → identify the expected position of regulatory elements (promoter, CpG islands, modified histones): 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

***THE ACQUISITION OF MECHANISMS OF GENE REGUALTION IS A STRONG INDICATOR FOR FUNCTIONAL RELEVANCE OF lncRNAs***



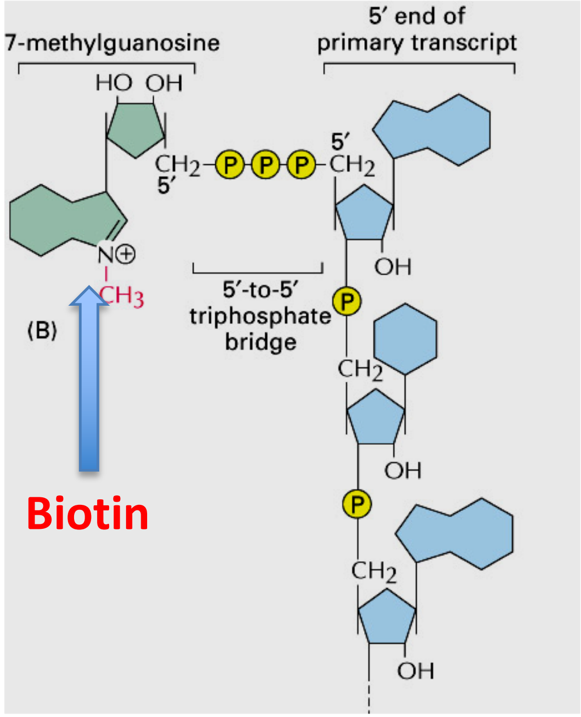
Nature Reviews | Molecular Cell Biology

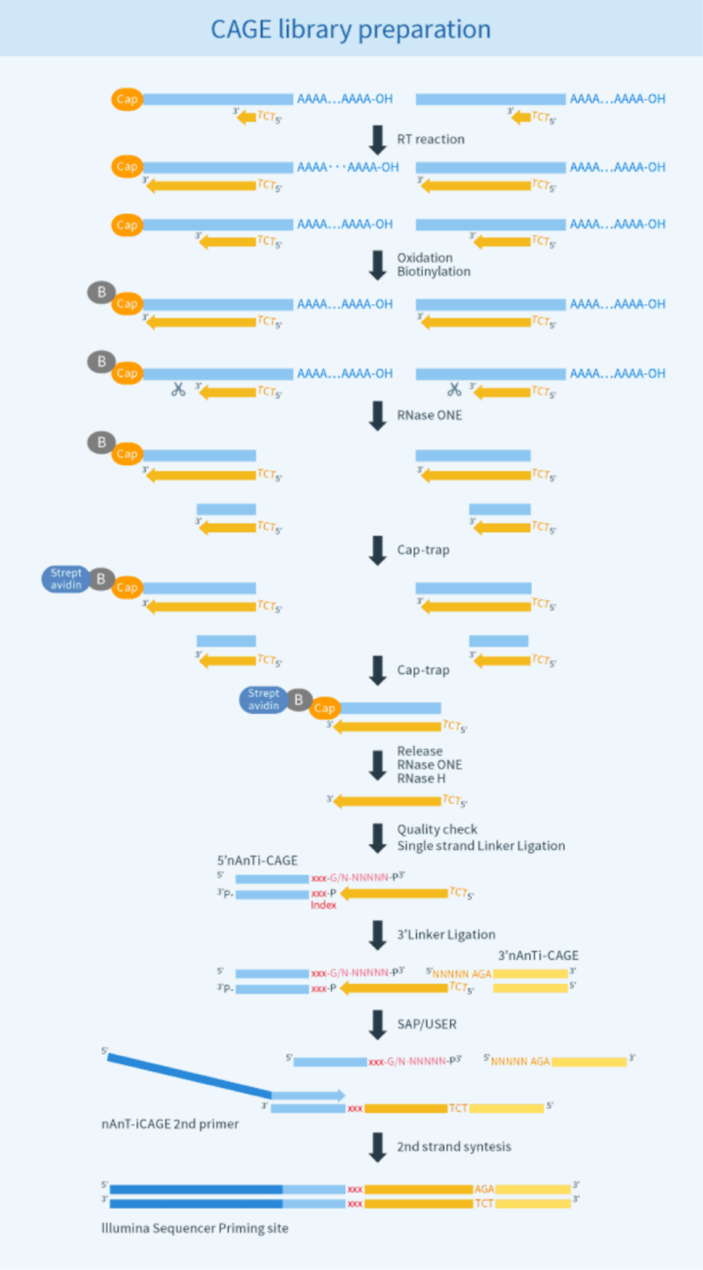
# 2. Landscape of transcription: CAGE (Cap Analysis of Gene Expression)

## Identifying transcriptional start sites

**Mapping 5' end of transcripts**
**Getting information on localization of regulatory region**
**Limited to RNA polymerase II transripts**

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.





Reverse transcription
- some RT products reach 5'end of RNA

Chemical modification of CAP with Biotin

Digestion of ssRNA
- remains only cDNA/RNA hybrid

Mix of cDNA/RNAhybrod with or without biotin-CAP

Trap the biotin-CAP with streptavidine beads

RNaseH digests RNA paired with DNA
RNaseONE digests ssRNA

ssDNA strand (cDNA) remains

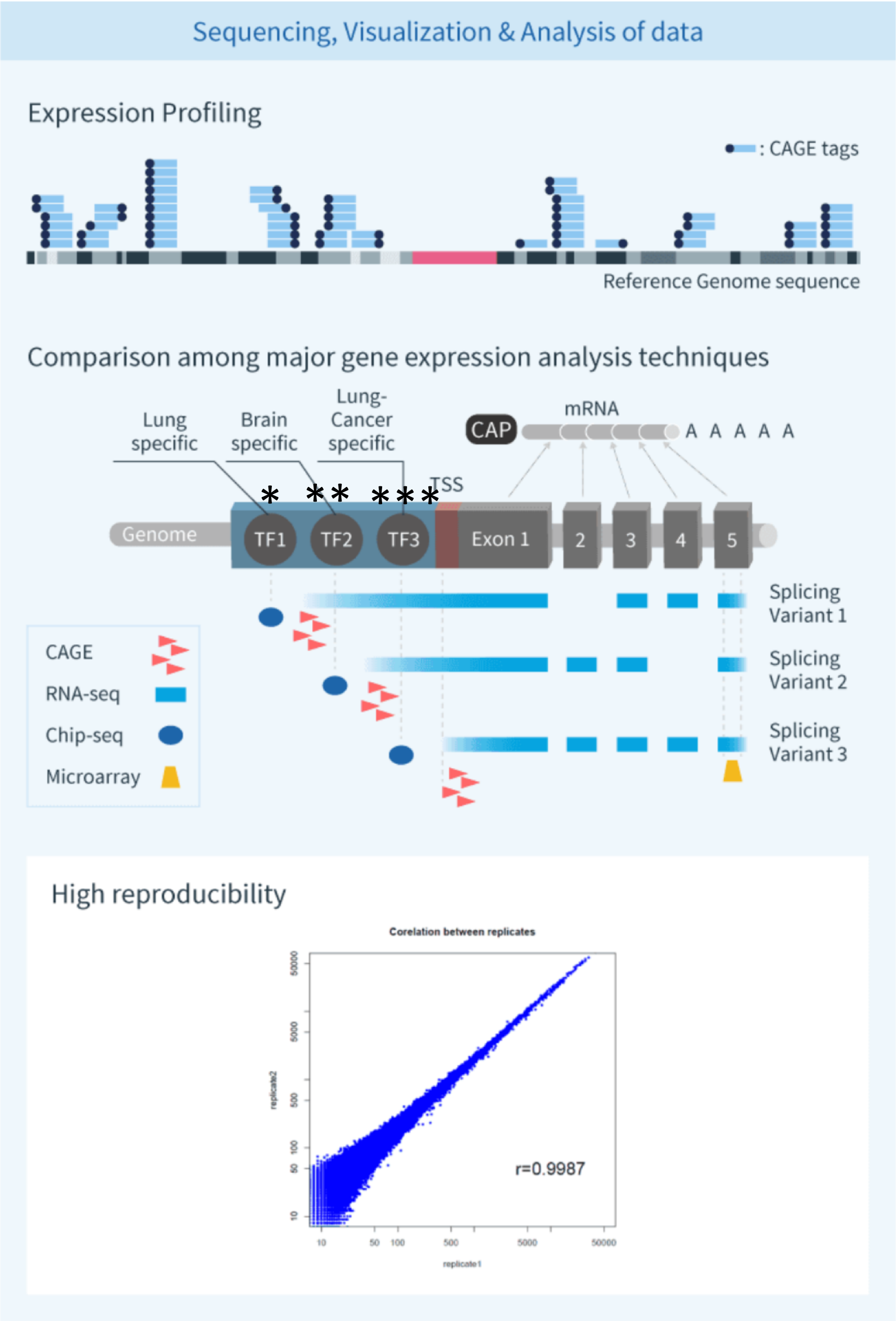
Tagging of 3'end of DNA with linker

Tagging of 5' end of DNA with linker

Second strand synthesis and Library preparation for massive parallel sequencing

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

# 2. Landscape of transcription: CAGE (Cap Analysis of Gene Expression)



***RNAseq is «focussed» on 5' end of transcripts → only type of RNAs represented in library***

| | |
|---|---|
| * | Tissue specific promoter 1 |
| ** | Tissue specific promoter 2 |
| *** | Tissue specific promoter 3 |

CAGE seq identifies 5' end of RNA and promoter regions

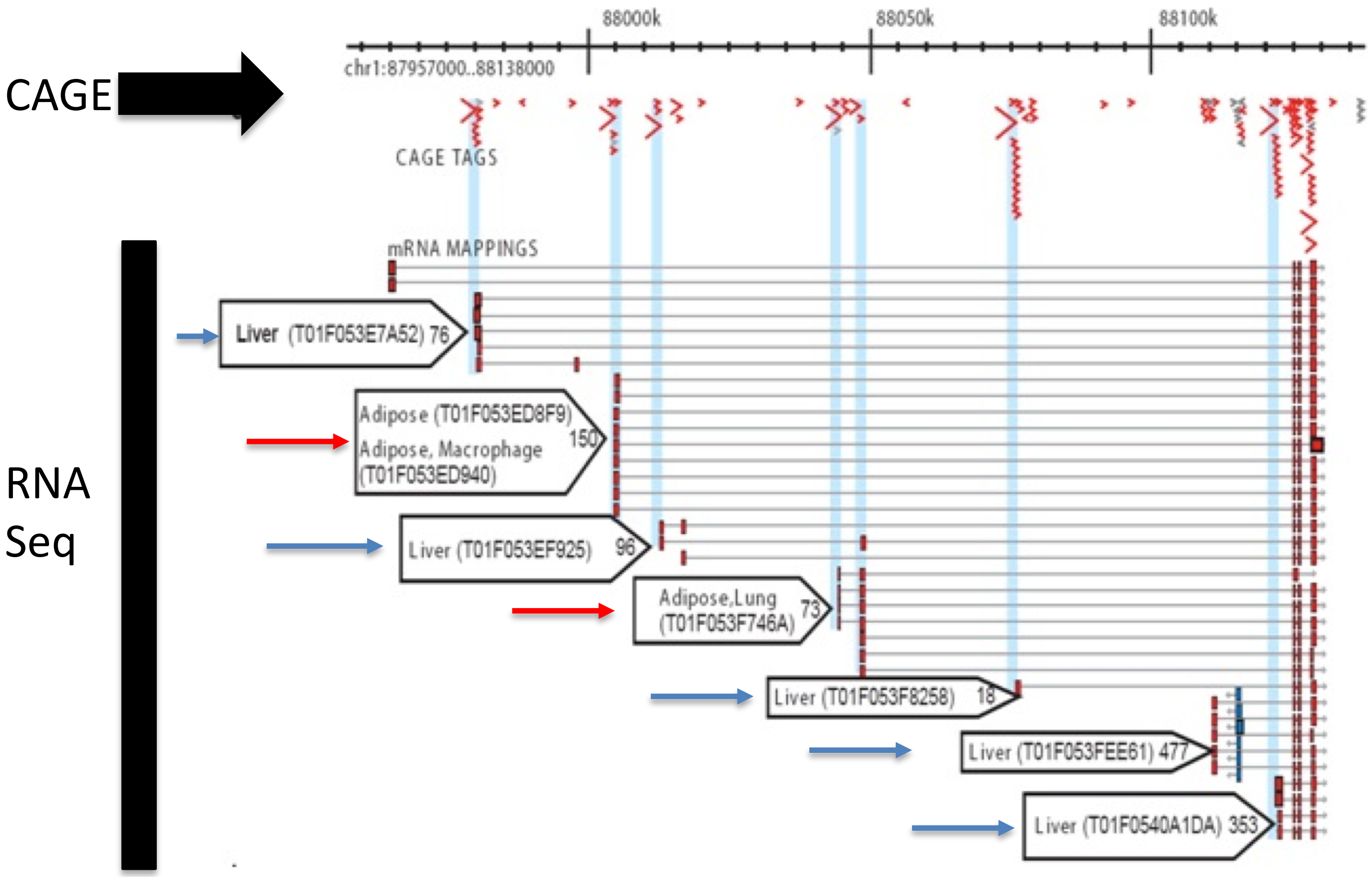
Identification of transcipt variants of gene in a single cell type with different promoters

Information on tissue specific gene regulation

Note: classic RNA seq provides information on "body" of transcript (length, alterntive splicing, expression levels)

# 2. Landscape of transcription: CAGE (Cap Analysis of Gene Expression)

## An example:



**Reads from CAGE and RNAseq experiments**

***RNA seq: can only detect aligned transcripts without detailed information on TSS.***

***CAGE: Excellent tool to identify Transcriptional start sites***

***Liver: same mRNA encoding gene, transcript variants with different start sites***

***In particular for non-coding genes That do not provide addional information from RNA sequence (i.e. triplette code for translation)***

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

***Identification of 5' end is essential: gives information on the position of a putative promoter***

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

1. DNA Seqeuncing (Human genome project, DNA-Seq)

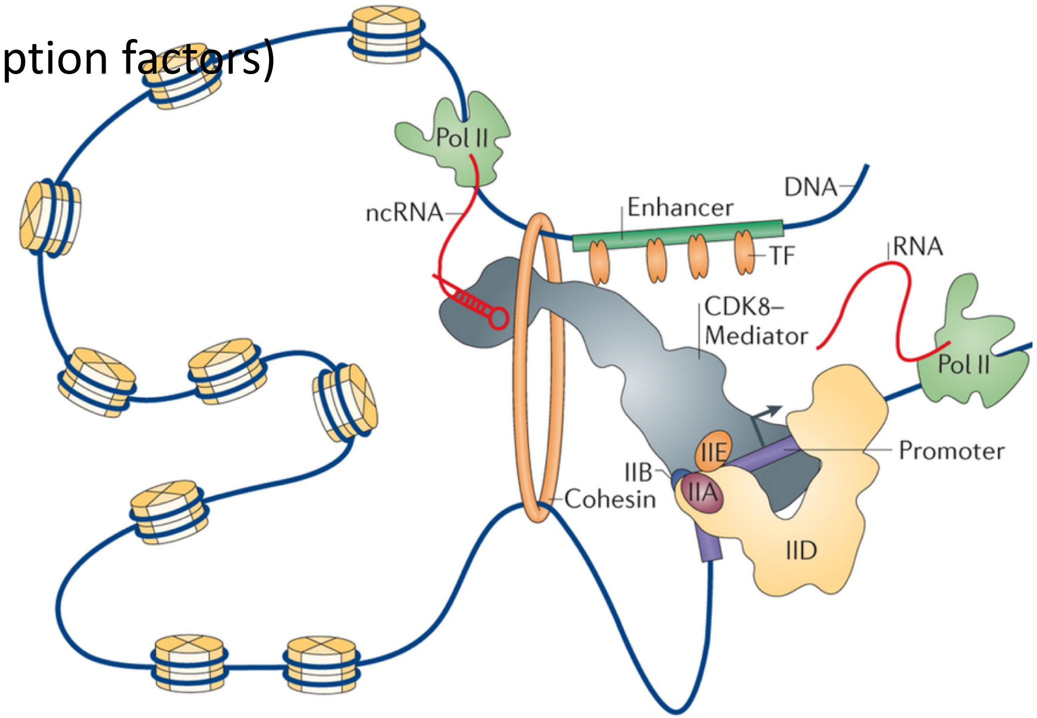
2. Landscape of transcription: Seqeuncing of RNA (total RNA, small/large RNA, CAGE)

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

***THE ACQUISITION OF MECHANISMS OF GENE REGUALTION IS A STRONG INDICATOR FOR FUNCTIONAL RELEVANCE OF lncRNAs***



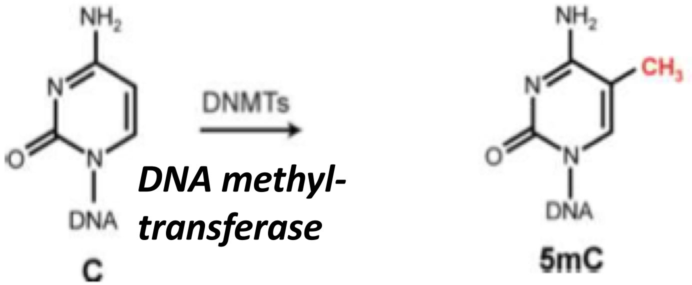
Nature Reviews | Molecular Cell Biolo

# 3. DNA methylation: educed representation bisulfite sequencing (RRBS)

## Identifying CpG islands

## - Identifying information that control gene expression

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

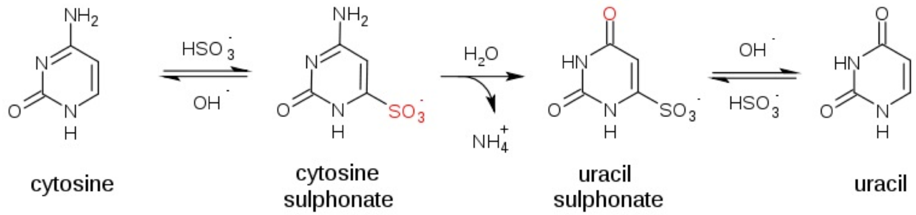


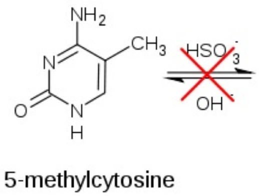
***DNA methyl-transferase***

active gene

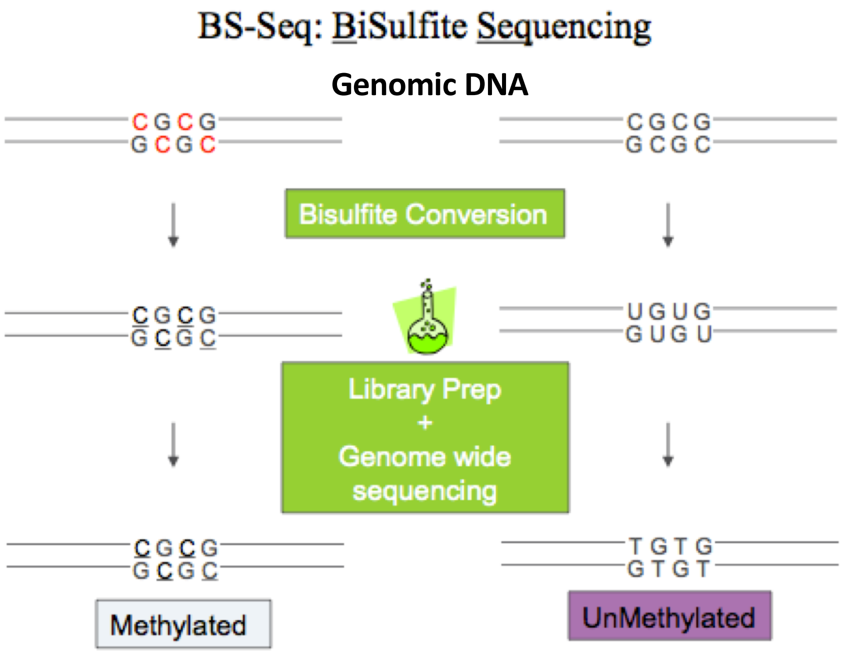
Silenced gene

- **Bi-sulfite conversion: C→U conversion by sodium bisulfite treatment**



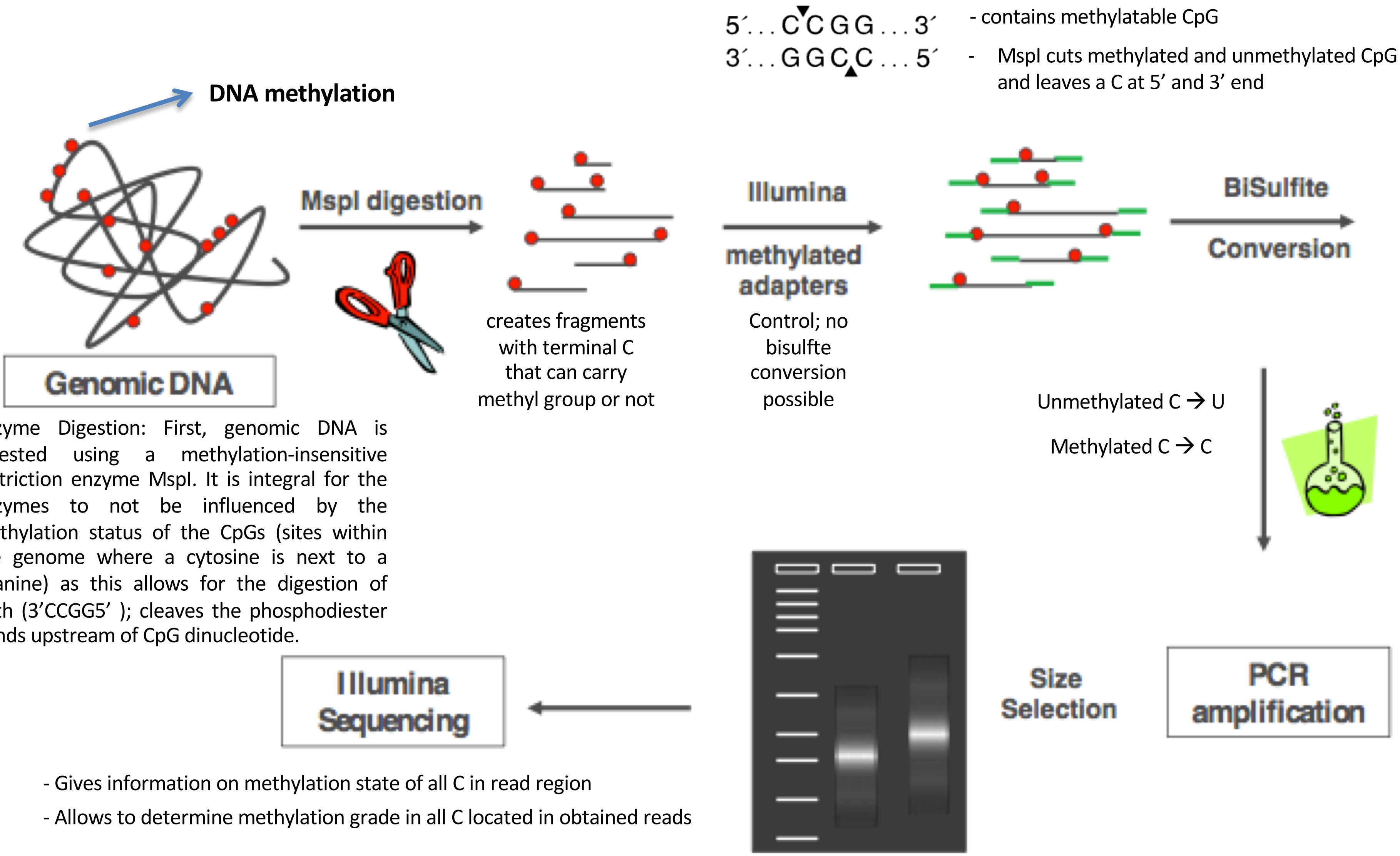


**methylated C cannot be converted!!**



# 3. 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.



5´...CCGG...3´
3´...GGCC...5´

- contains methylatable CpG
- MspI cuts methylated and unmethylated CpG and leaves a C at 5’ and 3’ end

DNA methylation

Genomic DNA

MspI digestion

creates fragments with terminal C that can carry methyl group or not

Illumina methylated adapters

Control; no bisulfte conversion possible

BiSulfite Conversion

Unmethylated C → U

Methylated C → C

PCR amplification

Size Selection

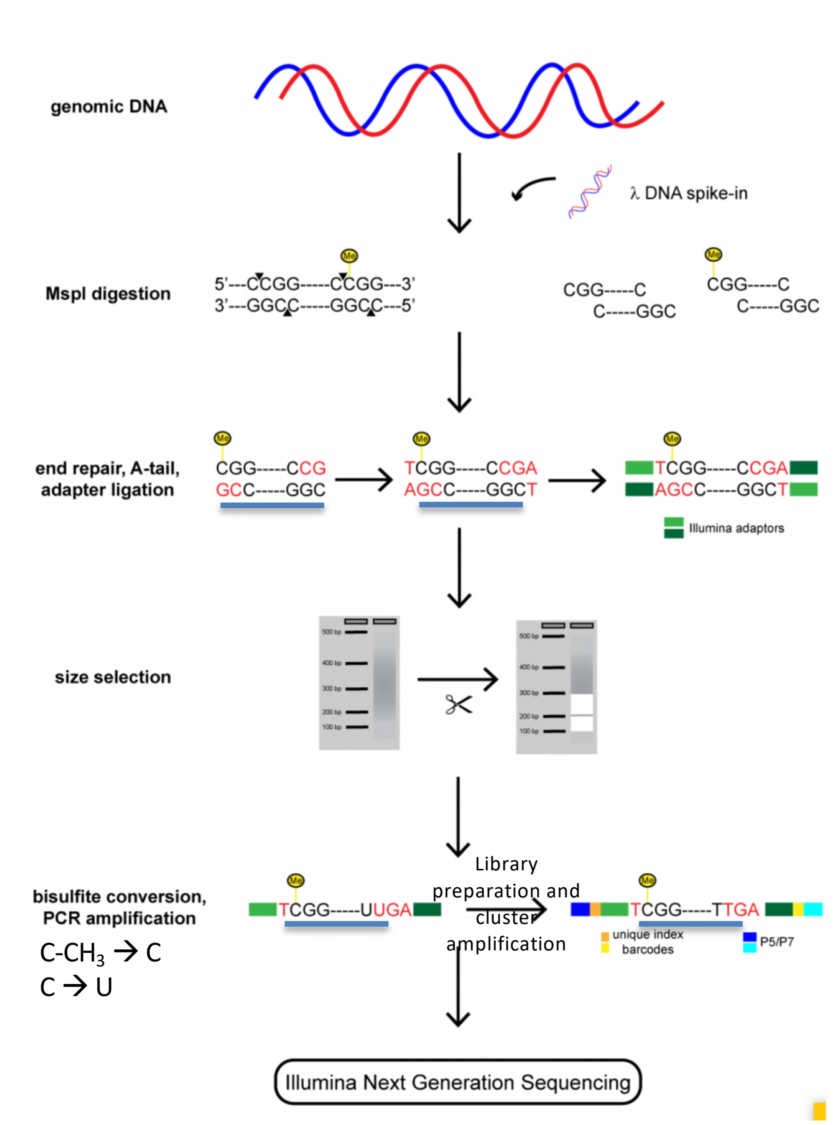
Illumina Sequencing

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.

- Gives information on methylation state of all C in read region
- Allows to determine methylation grade in all C located in obtained reads

# 3. DNA methylation: Reduced representation bisulfite sequencing (RRBS)

**Ideally 2 experimental conditions or 2 different biological samples**
(for example: WT and knock-out;
ES cells and differentiated ES cells)



Mapping of reads against reference genome

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.

5´. . . C▼CGG . . . 3´
3´. . . GGC▲C . . . 5´

- MspI cuts methylated and unmethylated CpG and leaves a C at 5’ and 3’ end with or without - $CH_3$

λ: Lambda DNA: prepared from methylation deficient bacteria
(no $CH_3$ → all C will be converted to U)

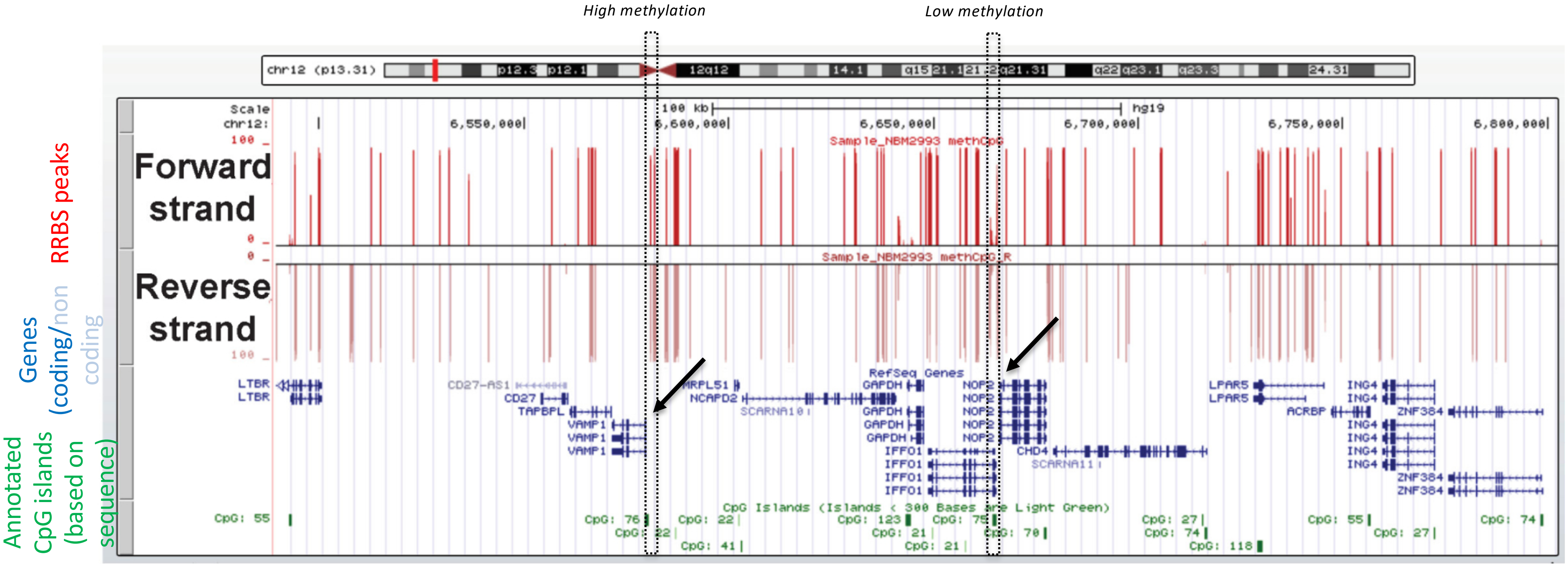
- Gives information on methylation state of all C in read region

- Allows to allocate methylated sequences (read build up diagram against refgenome)

- Allows to determine methylation grade in all C located in in differentially methylated regions

# 3. DNA methylation: Reduced representation bisulfite sequencing (RRBS)

- Aligment of RRBS data with classic CpG islands annotation (annotated based on sequence content)
- Clustering of RRBS reads in zones with CpG islands
- Analysis of methylation grade from RRBS data



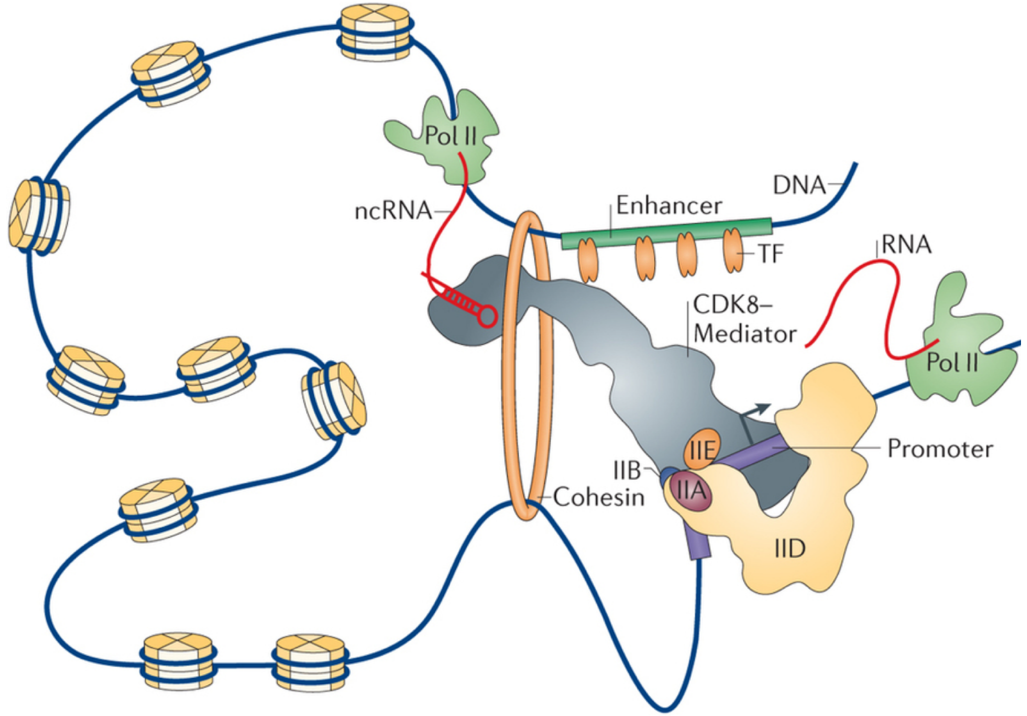
University of California, Santa Cruz (UCSC) genome browser[43] image of representative data from an RRBS sequencing lane. The y-axis scale bar represents 0-100% methylation at each cytosine covered with a minimum of 10x. The top custom track represents the forward strand and the lower custom track represents the reverse strand. Shown is chr12:6,489,523-6,802,422 (hg19) inclusive of refseq genes and CpG islands within this genomic region.

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

1. DNA Sequencing (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

***THE ACQUISITION OF MECHANISMS OF GENE REGUALTION IS A STRONG INDICATOR FOR FUNCTIONAL RELEVANCE OF lncRNAs***



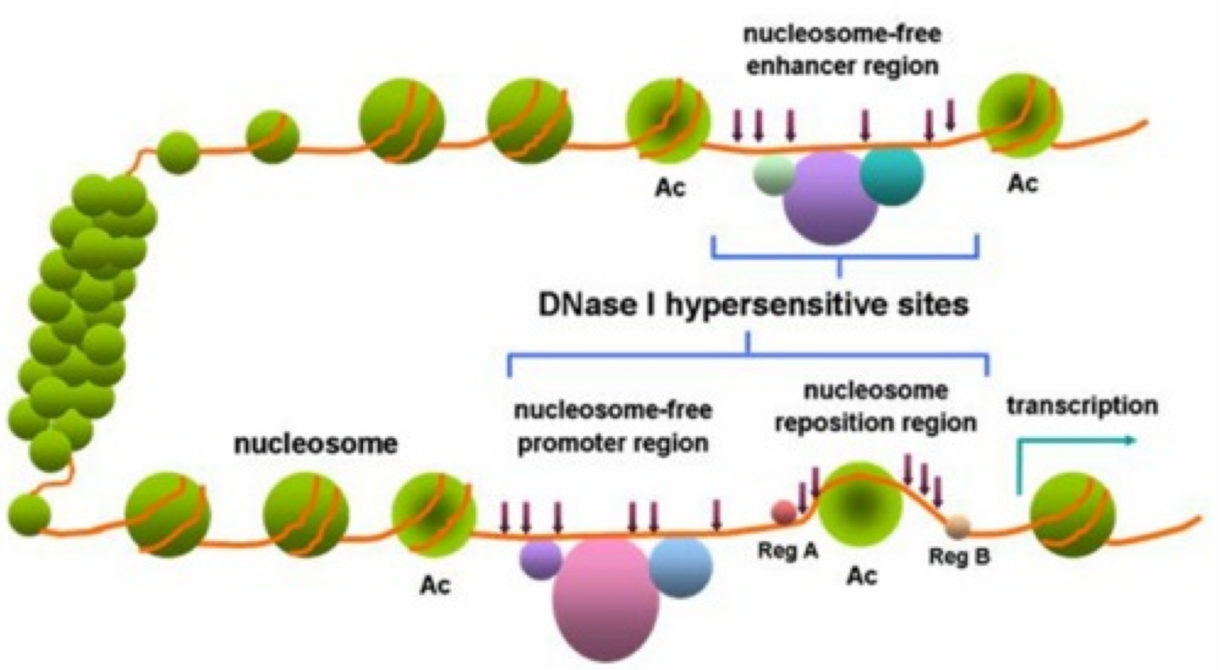
Nature Reviews | Molecular Cell Biolo

# 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

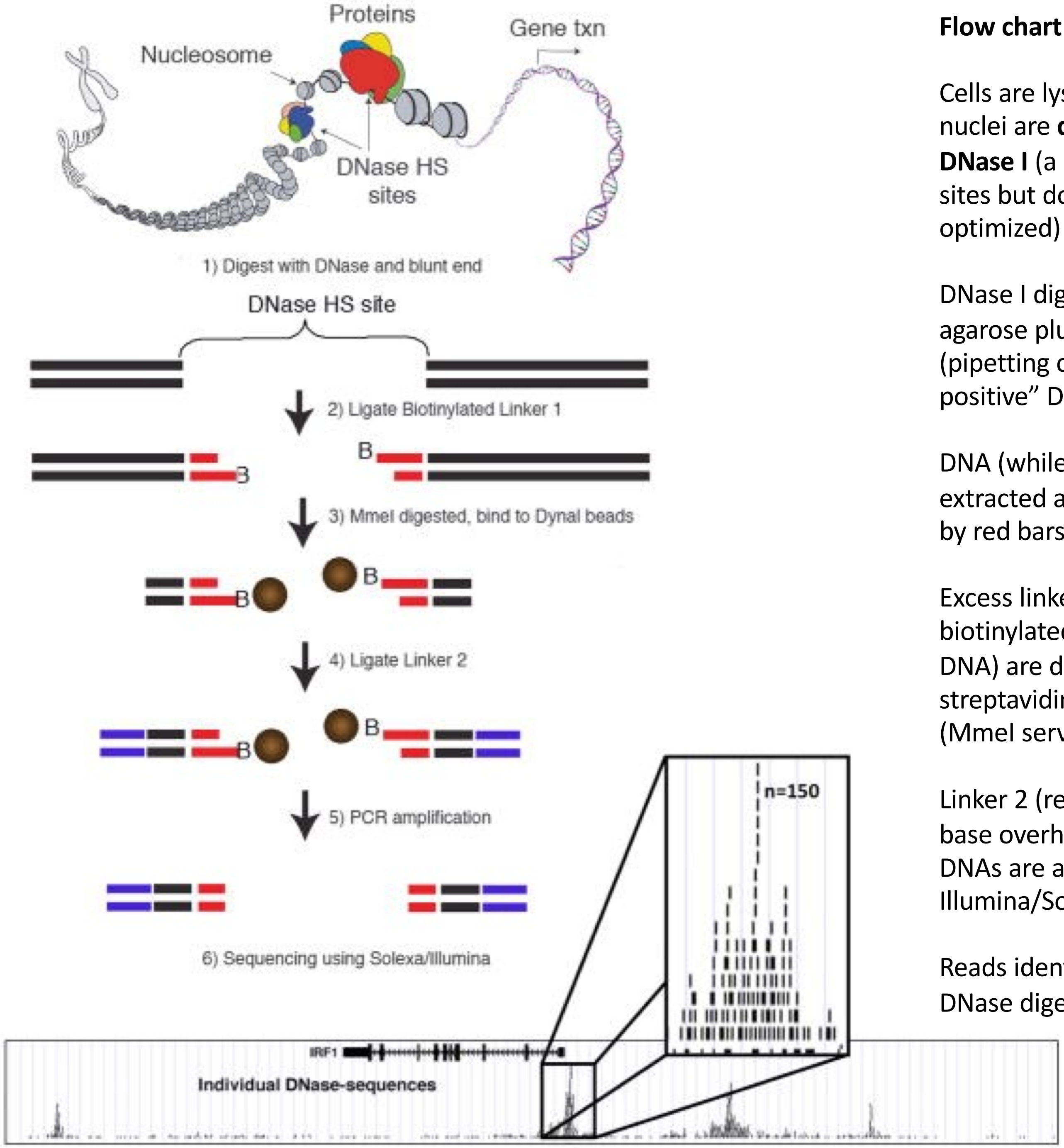
## DNase hypersensitive sites mark sequences involved in gene regulation

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.



nucleosome-free enhancer region
Ac
Ac
DNase I hypersensitive sites
nucleosome
nucleosome-free promoter region
nucleosome reposition region
transcription
Ac
Reg A
Ac
Reg B

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


Proteins
Nucleosome
Gene txn
DNase HS sites
1) Digest with DNase and blunt end
DNase HS site
2) Ligate Biotinylated Linker 1
B
B
3) MmeI digested, bind to Dynal beads
B
B
4) Ligate Linker 2
B
B
5) PCR amplification
6) Sequencing using Solexa/Illumina
n=150
IRF1
Individual DNase-sequences

**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** (a concentration that allows digestion of sensitive sites but does not cleave all linker regions → need to be optimized)

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).

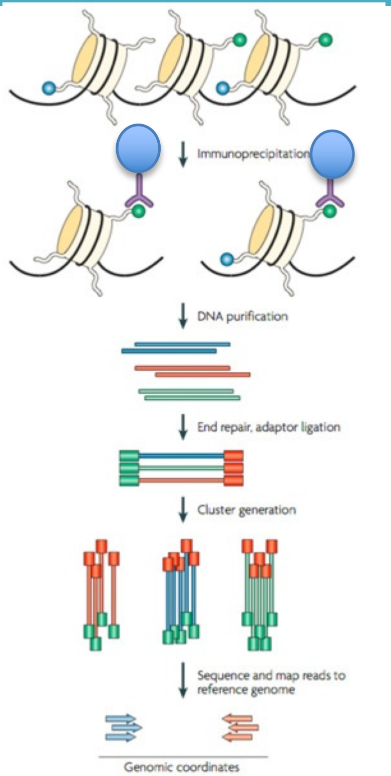
DNA (while still in the plugs) are then blunt-ended, extracted and ligated to biotinylated linker 1 (represented by red bars in the figure).

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). (MmeI serves to fragment DNA)

Linker 2 (represented by the blue bars) is ligated to the 2 base overhang generated by MmeI, and the tagged 20 bp DNAs are amplified by PCR and sequenced by Illumina/Solexa.

Reads identify the borders of the gap created by the DNase digest

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



H3K4me3 (active chromatin mark)

H3K27me3 (repressive chromatin mark)

H3K27Ac (regulatory elements)

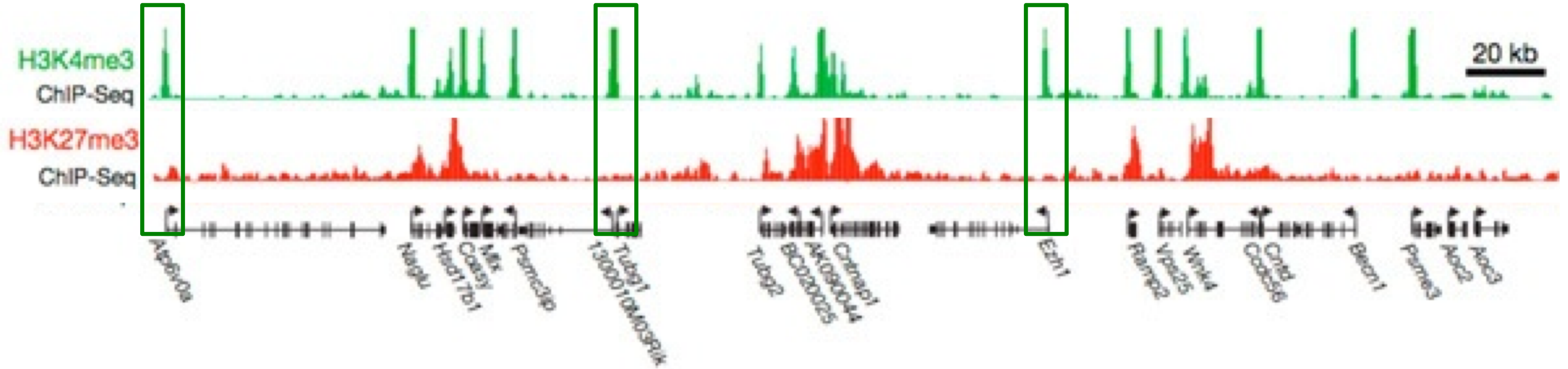
● magnetic beads covered with specific antibody

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

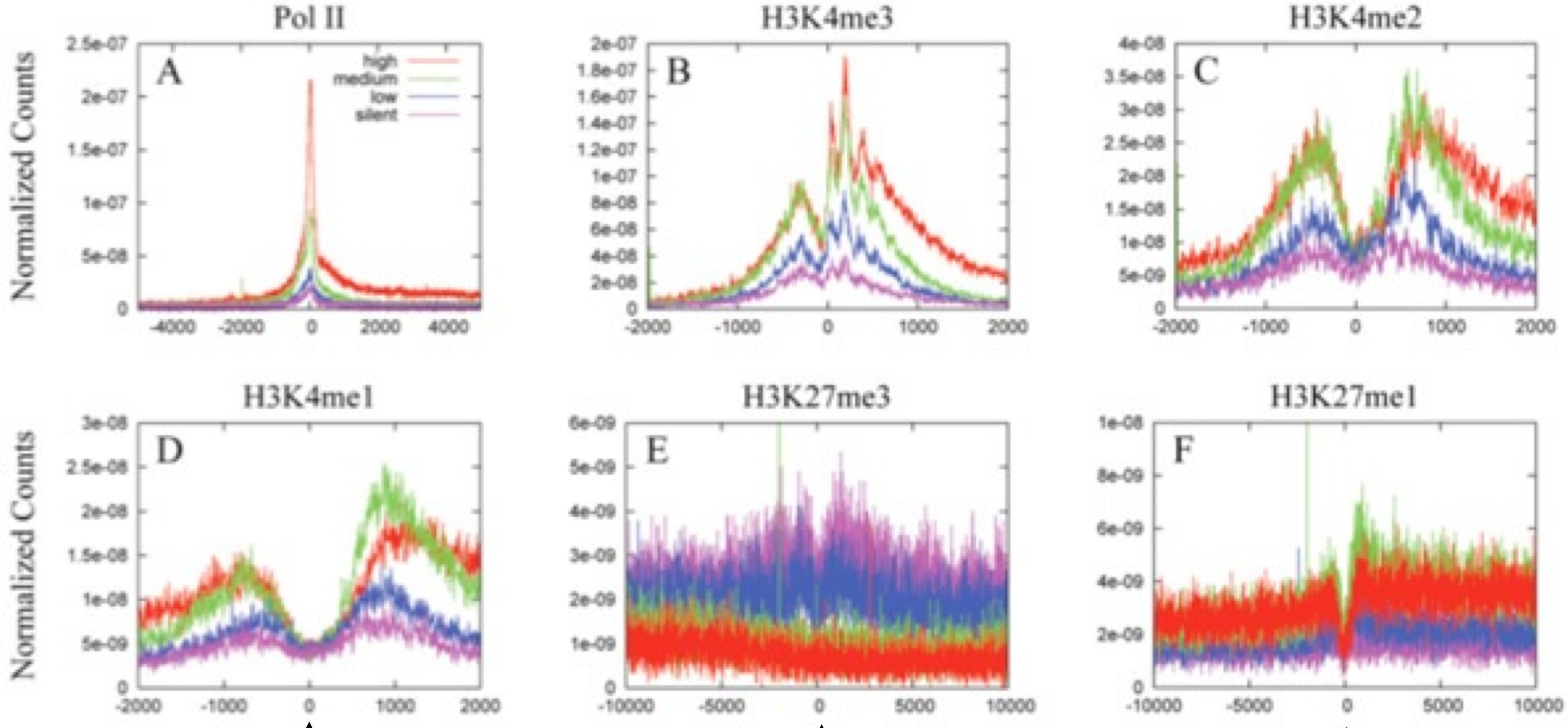
Enrichment for the histone modification or transcription factor

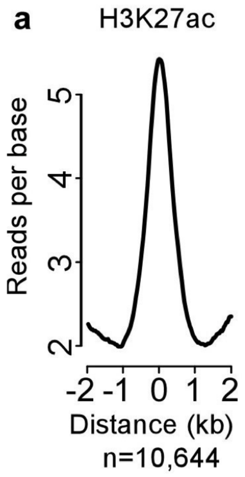


The results indicate that some modifications (H3K4me) are correlated with increased gene expression, while others (H3K27me3) correlate with decreases gene expression. The peaks observed in the H3K4me3 for genes at high expression levels occur at +50, +210, and +360 based which correlates well with the 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 RNA Pol II target genes***





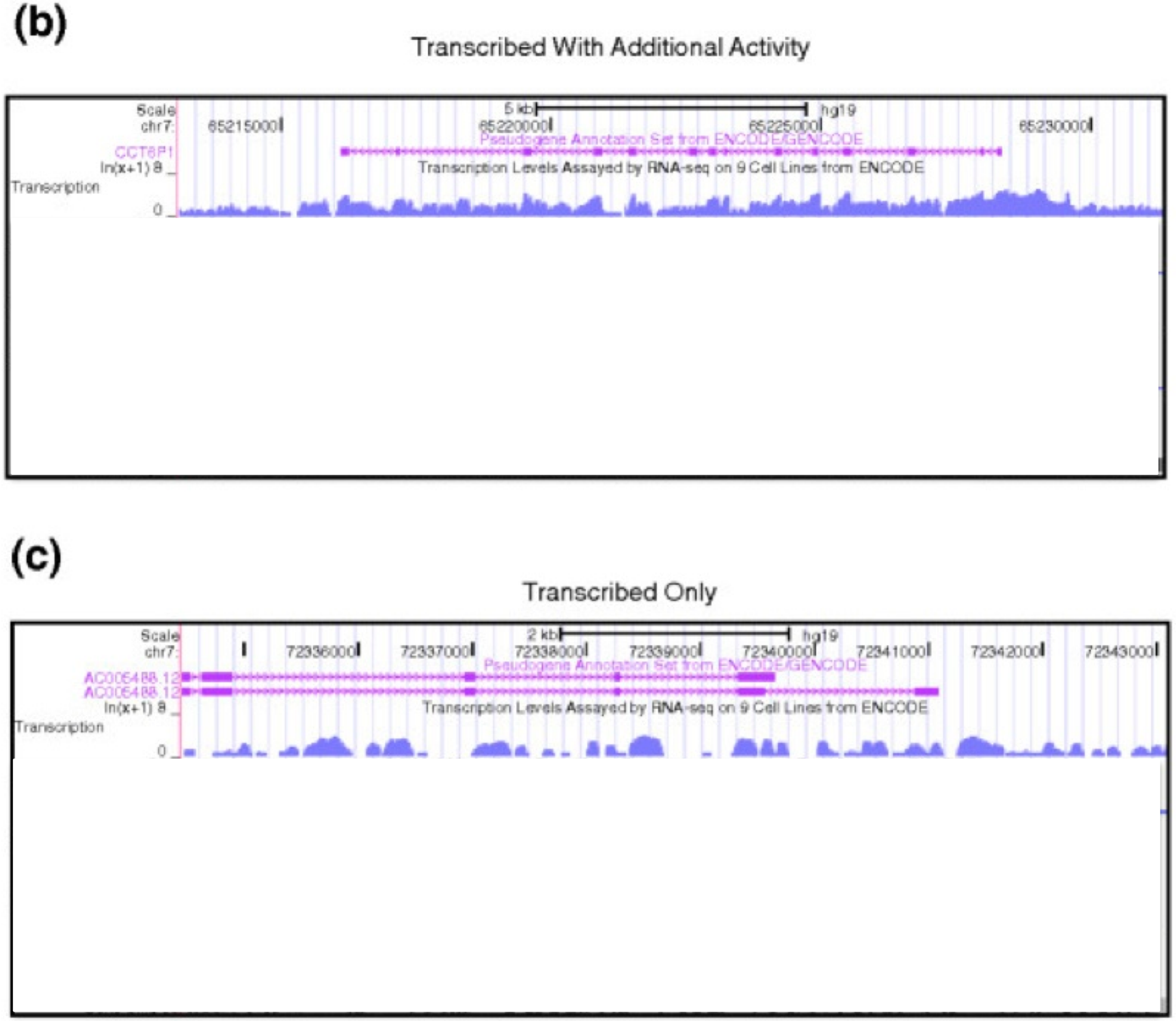
transcriptional start site = position 0

Regulatory elements

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

**Same method can be used to localize transcription factors**

# AN EXAMPLE:ORGANISATION OF A FUNCTIONAL ELEMENT: PSEUDOGENES


(b)

Transcribed With Additional Activity

Scale
chr7:
5 kb
hg19
65215000
65220000
65225000
65230000
Pseudogene Annotation Set from ENCODE/GENCODE
CCT6P1
ln(x+1) 8
Transcription
0
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE

(c)

Transcribed Only

Scale
chr7:
2 kb
hg19
72336000
72337000
72338000
72339000
72340000
72341000
72342000
72343000
Pseudogene Annotation Set from ENCODE/GENCODE
AC006488.12
AC006488.12
ln(x+1) 8
Transcription
0
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE

lncRNA 1 (Pseudogene CCT6P1)

lncRNA 2 (Pseudogene AC0064BB12)

Question:

Which of the lncRNA is a "real" functional RNA?

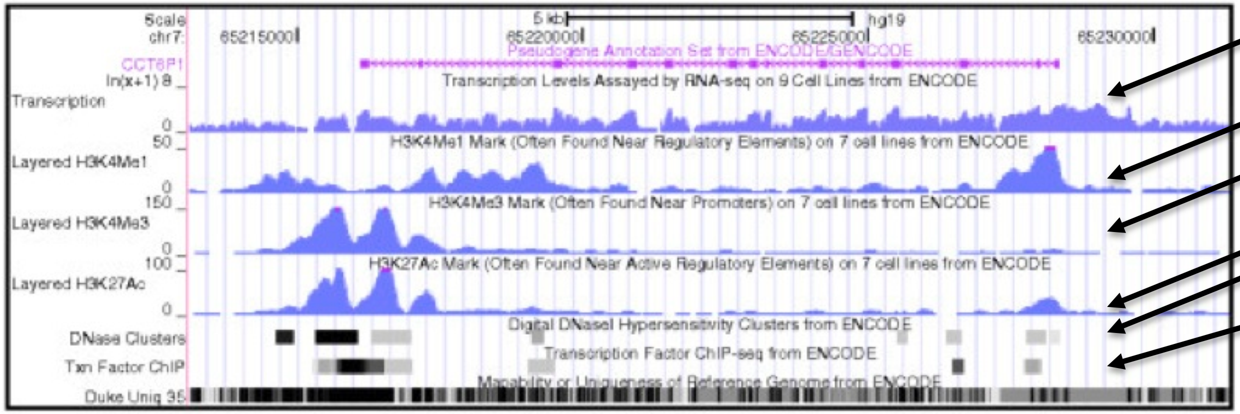
Which of the lncRNA is a result of transcriptional noise?

**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

# AN EXAMPLE: ORGANISATION OF A FUNCTIONAL ELEMENT: PSEUDOGENES

Transcribed With Additional Activity


Scale
chr7:
5 kb
hg19
65215000
65220000
65225000
65230000
Pseudogene Annotation Set from ENCODE/GENCODE
CCT6P1
ln(x+1) 8
Transcription
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
Layered H3K4Me1
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me3
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
Duke Uniq 35
Mapability or Uniqueness of Reference Genome from ENCODE

## lncRNA 1 (Pseudogene CCT6P1)

RNA expression: PRESENT

RNA Polymerase II: not shown

H4K4me1: near regulatory elements

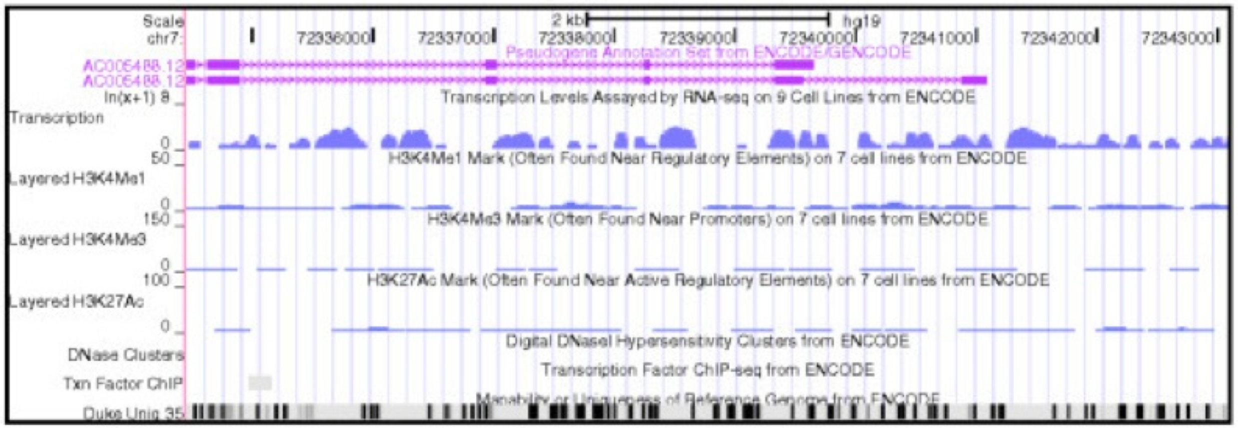
H3K4me3: near promoters

H3K27Ac: near regulatory elements

DNAse hypersensitive sites: at regulatory elements

Transcription factor (TF) binding: Near promoter

Transcribed Only


Scale
chr7:
2 kb
hg19
72336000
72337000
72338000
72339000
72340000
72341000
72342000
72343000
Pseudogene Annotation Set from ENCODE/GENCODE
AC006488.12
AC006488.12
ln(x+1) 8
Transcription
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
Layered H3K4Me1
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me3
H3K4Me3 Mark (Often Found Near Promoters) on 7 cell lines from ENCODE
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
Duke Uniq 35
Mapability or Uniqueness of Reference Genome from ENCODE

## lncRNA 2 (Pseudogene AC0064BB12)

RNA expression: PRESENT

Chromatin shows low active marks

Poor definition

Presumably non-functional RNA transcripts

**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)

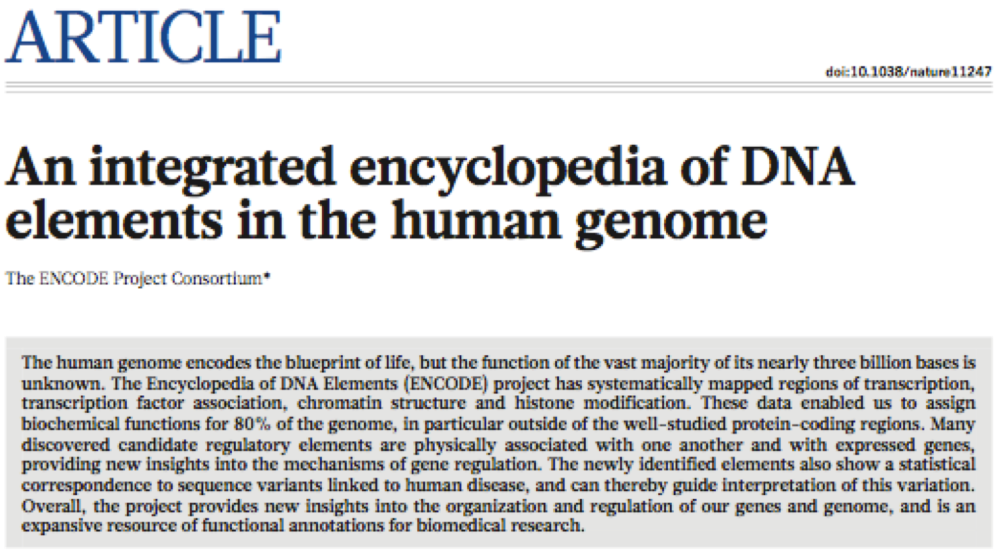
WORK STILL IN PRGRESS

*http://www.genome.gov/encode/*

Aim: a catalog of manually curated list of genes/transcripts (GENCODE)

*http://www.gencodegenes.org/*

Release ENCODE V4 (2020)



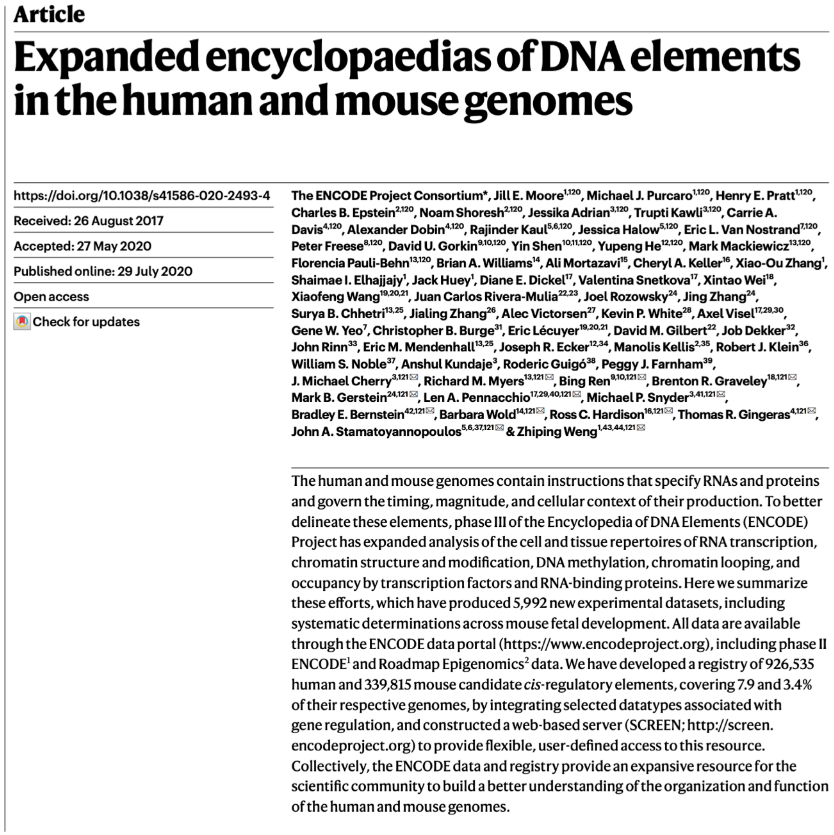
ARTICLE

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.



Article

# Expanded encyclopaedias of DNA elements in the human and mouse genomes

https://doi.org/10.1038/s41586-020-2493-4

Received: 26 August 2017

Accepted: 27 May 2020

Published online: 29 July 2020

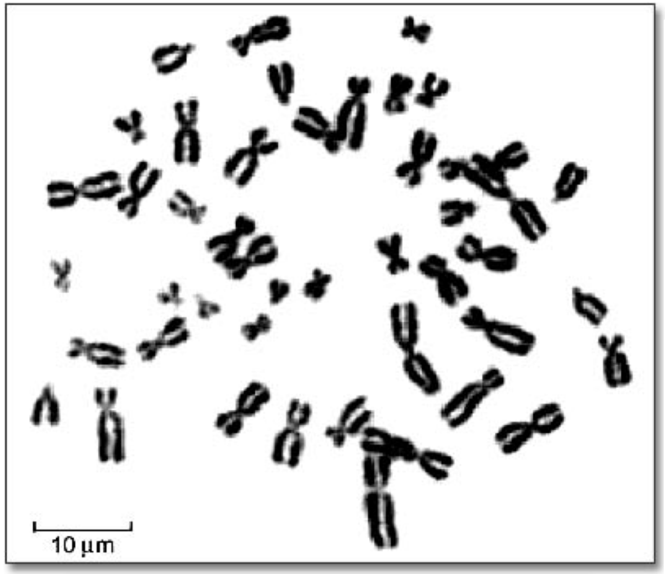
Open access

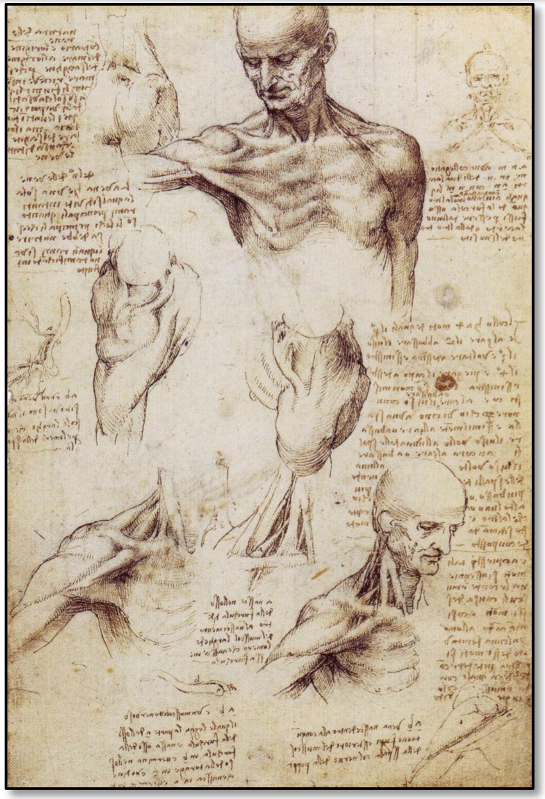
Check for updates

The ENCODE Project Consortium*, Jill E. Moore[1,120], Michael J. Purcaro[1,120], Henry E. Pratt[1,120], Charles B. Epstein[2,120], Noam Shoresh[2,120], Jessika Adrian[3,120], Trupti Kawli[3,120], Carrie A. Davis[4,120], Alexander Dobin[4,120], Rajinder Kaul[5,6,120], Jessica Halow[5,120], Eric L. Van Nostrand[7,120], Peter Freese[8,120], David U. Gorkin[9,10,120], Yin Shen[10,11,120], Yupeng He[12,120], Mark Mackiewicz[13,120], Florencia Pauli-Behn[13,120], Brian A. Williams[14], Ali Mortazavi[15], Cheryl A. Keller[16], Xiao-Ou Zhang[1], Shaimae I. Elhajjajy[1], Jack Huey[1], Diane E. Dickel[17], Valentina Snetkova[17], Xintao Wei[18], Xiaofeng Wang[19,20,21], Juan Carlos Rivera-Mulia[22,23], Joel Rozowsky[24], Jing Zhang[24], Surya B. Chhetri[13,25], Jialing Zhang[26], Alec Victorsen[27], Kevin P. White[28], Axel Visel[17,29,30], Gene W. Yeo[7], Christopher B. Burge[31], Eric Lécuyer[19,20,21], David M. Gilbert[22], Job Dekker[32], John Rinn[33], Eric M. Mendenhall[13,25], Joseph R. Ecker[12,34], Manolis Kellis[2,35], Robert J. Klein[36], William S. Noble[37], Anshul Kundaje[3], Roderic Guigó[38], Peggy J. Farnham[39], J. Michael Cherry[3,121] ✉, Richard M. Myers[13,121] ✉, Bing Ren[9,10,121] ✉, Brenton R. Graveley[18,121] ✉, Mark B. Gerstein[24,121] ✉, Len A. Pennacchio[17,29,40,121] ✉, Michael P. Snyder[3,41,121] ✉, Bradley E. Bernstein[42,121] ✉, Barbara Wold[14,121] ✉, Ross C. Hardison[16,121] ✉, Thomas R. Gingeras[4,121] ✉, John A. Stamatoyannopoulos[5,6,37,121] ✉ & Zhiping Weng[1,43,44,121] ✉

The human and mouse genomes contain instructions that specify RNAs and proteins and govern the timing, magnitude, and cellular context of their production. To better delineate these elements, phase III of the Encyclopedia of DNA Elements (ENCODE) Project has expanded analysis of the cell and tissue repertoires of RNA transcription, chromatin structure and modification, DNA methylation, chromatin looping, and occupancy by transcription factors and RNA-binding proteins. Here we summarize these efforts, which have produced 5,992 new experimental datasets, including systematic determinations across mouse fetal development. All data are available through the ENCODE data portal (https://www.encodeproject.org), including phase II ENCODE[1] and Roadmap Epigenomics[2] data. We have developed a registry of 926,535 human and 339,815 mouse candidate *cis*-regulatory elements, covering 7.9 and 3.4% of their respective genomes, by integrating selected datatypes associated with gene regulation, and constructed a web-based server (SCREEN; http://screen.encodeproject.org) to provide flexible, user-defined access to this resource. Collectively, the ENCODE data and registry provide an expansive resource for the scientific community to build a better understanding of the organization and function of the human and mouse genomes.

# 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.

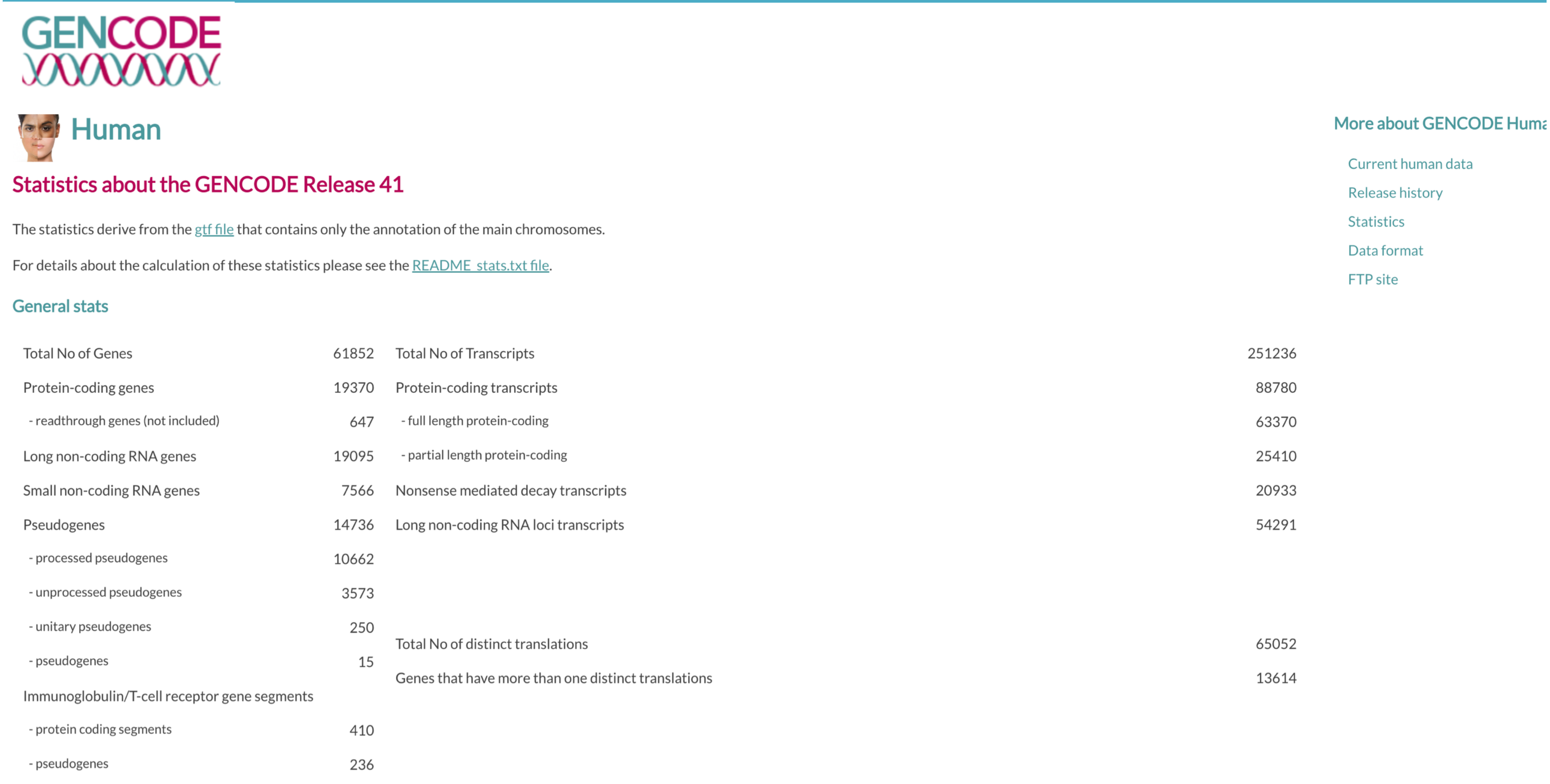
It is possible to quantitatively correlate RNA sequence production and processing with both chromatin marks and transcription factor (TF) binding at promoters, indicating that promoter functionality can explain the majority of RNA expression variation.

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 27.09.2022:
# Project that uses ENCODE for the annotation of functional elements in the genome

**http://www.gencodegenes.org/**



GENCODE

## Human

### Statistics about the GENCODE Release 41

The statistics derive from the gtf file that contains only the annotation of the main chromosomes.

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

#### General stats

| | | | |
|---|---|---|---|
| Total No of Genes | 61852 | Total No of Transcripts | 251236 |
| Protein-coding genes | 19370 | Protein-coding transcripts | 88780 |
| - readthrough genes (not included) | 647 | - full length protein-coding | 63370 |
| Long non-coding RNA genes | 19095 | - partial length protein-coding | 25410 |
| Small non-coding RNA genes | 7566 | Nonsense mediated decay transcripts | 20933 |
| Pseudogenes | 14736 | Long non-coding RNA loci transcripts | 54291 |
| - processed pseudogenes | 10662 | | |
| - unprocessed pseudogenes | 3573 | | |
| - unitary pseudogenes | 250 | Total No of distinct translations | 65052 |
| - pseudogenes | 15 | Genes that have more than one distinct translations | 13614 |
| Immunoglobulin/T-cell receptor gene segments | | | |
| - protein coding segments | 410 | | |
| - pseudogenes | 236 | | |

More about GENCODE Huma

Current human data

Release history

Statistics

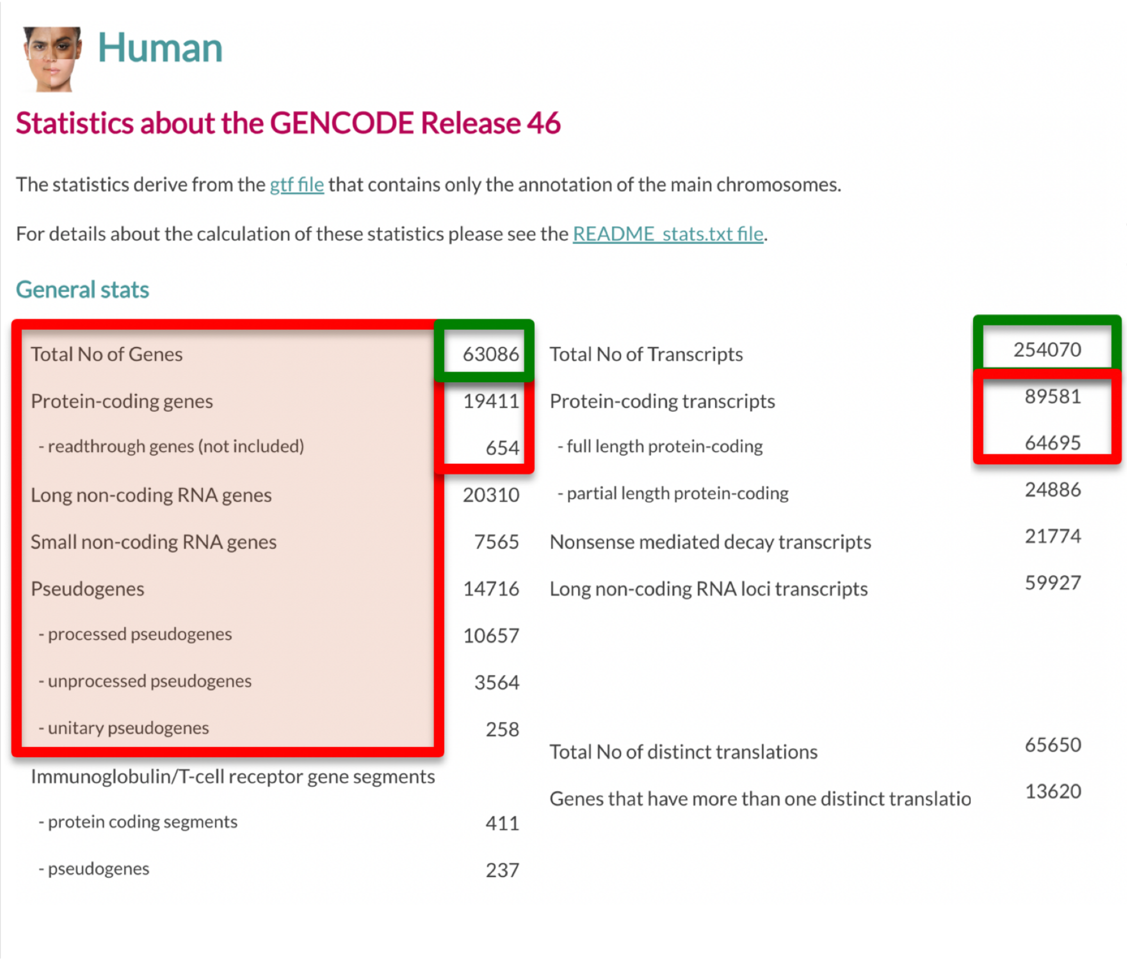
Data format

FTP site

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

# GENCODE – STATUS 23.09.2024:
# Project that uses ENCODE for the annotation of functional elements in the genome

http://www.gencodegenes.org/



## Human

### Statistics about the GENCODE Release 46

The statistics derive from the gtf file that contains only the annotation of the main chromosomes.

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

#### General stats

| | |
|---|---|
| Total No of Genes | 63086 |
| Protein-coding genes | 19411 |
| - readthrough genes (not included) | 654 |
| Long non-coding RNA genes | 20310 |
| Small non-coding RNA genes | 7565 |
| Pseudogenes | 14716 |
| - processed pseudogenes | 10657 |
| - unprocessed pseudogenes | 3564 |
| - unitary pseudogenes | 258 |
| Immunoglobulin/T-cell receptor gene segments | |
| - protein coding segments | 411 |
| - pseudogenes | 237 |

| | |
|---|---|
| Total No of Transcripts | 254070 |
| Protein-coding transcripts | 89581 |
| - full length protein-coding | 64695 |
| - partial length protein-coding | 24886 |
| Nonsense mediated decay transcripts | 21774 |
| Long non-coding RNA loci transcripts | 59927 |
| Total No of distinct translations | 65650 |
| Genes that have more than one distinct translatio | 13620 |

# release
41→46

+800

+800

+5000

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