# Gene Regulation & Review

Michael Schatz

November 1, 2023

Lecture 19. Applied Comparative Genomics

# Upcoming events

~~Mon Oct 30: Regular class | Prelim report assigned~~
Wed Nov 1: Review class

Mon Nov 6: Midterm exam (1 page of notes allowed)
Wed Nov 8: Regular class | Review exam

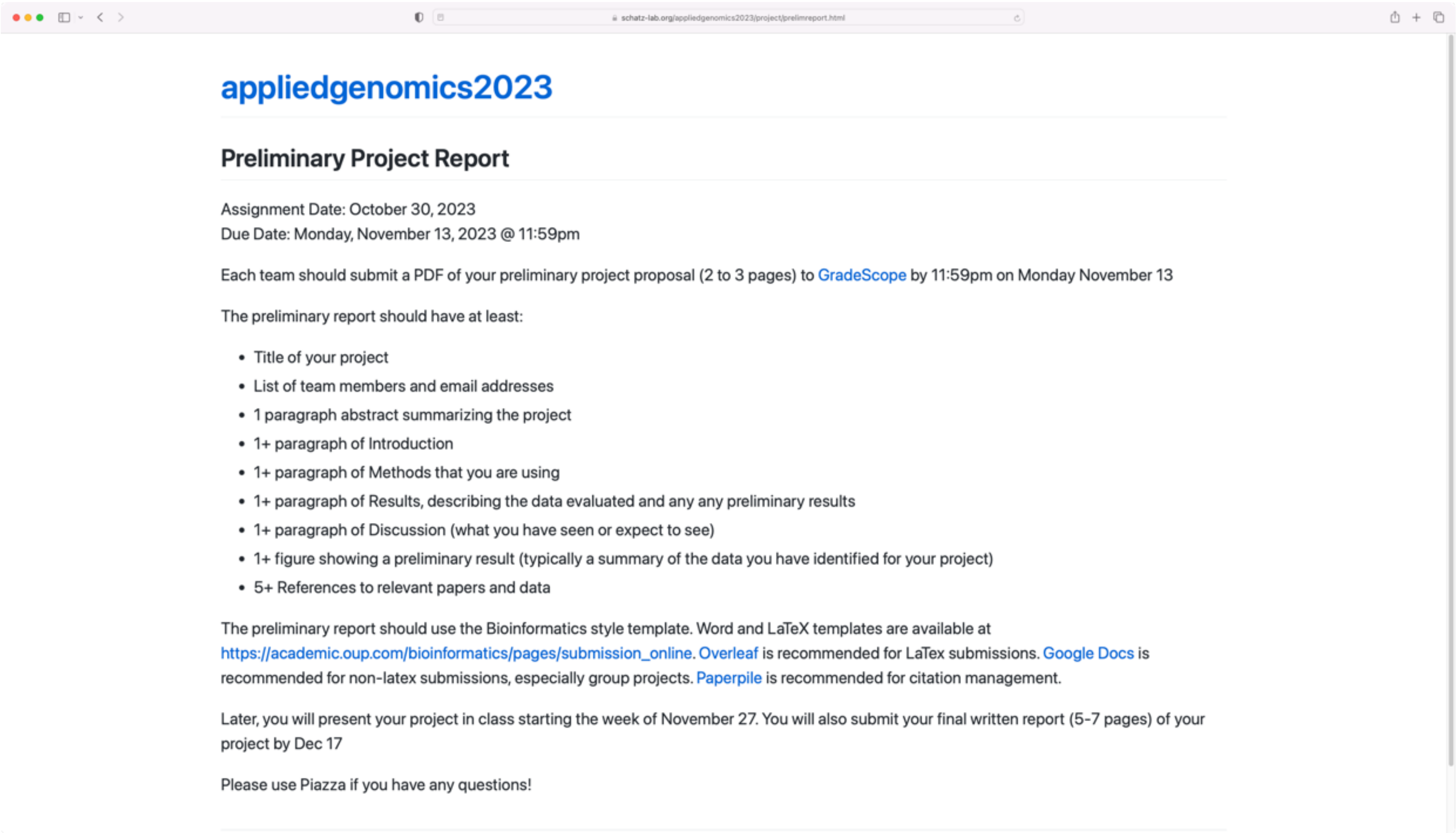
Mon Nov 13: Regular class | Prelim report due
Wed Nov 15: Regular class

Mon Nov 20: Thanksgiving break
Wed Nov 22: Thanksgiving break

Mon Nov 27: In class presentation (random order)
Wed Nov 29: In class presentation (random order)
Mon Dec 4: In class presentation (random order)
Wed Dec 6: No class!

Sun Dec 17: Final report due!!!!

# Preliminary Report
# Due Monday November 13



schatz-lab.org/appliedgenomics2023/project/prelimreport.html

## appliedgenomics2023

### Preliminary Project Report

Assignment Date: October 30, 2023
Due Date: Monday, November 13, 2023 @ 11:59pm

Each team should submit a PDF of your preliminary project proposal (2 to 3 pages) to GradeScope by 11:59pm on Monday November 13

The preliminary report should have at least:

- Title of your project
- List of team members and email addresses
- 1 paragraph abstract summarizing the project
- 1+ paragraph of Introduction
- 1+ paragraph of Methods that you are using
- 1+ paragraph of Results, describing the data evaluated and any any preliminary results
- 1+ paragraph of Discussion (what you have seen or expect to see)
- 1+ figure showing a preliminary result (typically a summary of the data you have identified for your project)
- 5+ References to relevant papers and data

The preliminary report should use the Bioinformatics style template. Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online. Overleaf is recommended for LaTex submissions. Google Docs is recommended for non-latex submissions, especially group projects. Paperpile is recommended for citation management.

Later, you will present your project in class starting the week of November 27. You will also submit your final written report (5-7 pages) of your project by Dec 17

Please use Piazza if you have any questions!

# Exam Topics

**Genomics**

- Genomics Technologies
  - Illumina, PacBio, Nanopore
- Kmer profiling
- Genome Assembly
- Whole Genome Alignment
- Read mapping
- Variant Identification
- Gene Finding
- BLAST
- RNA-seq
- Methyl-seq, Chip-Seq, Hi-C
- Genome Annotation

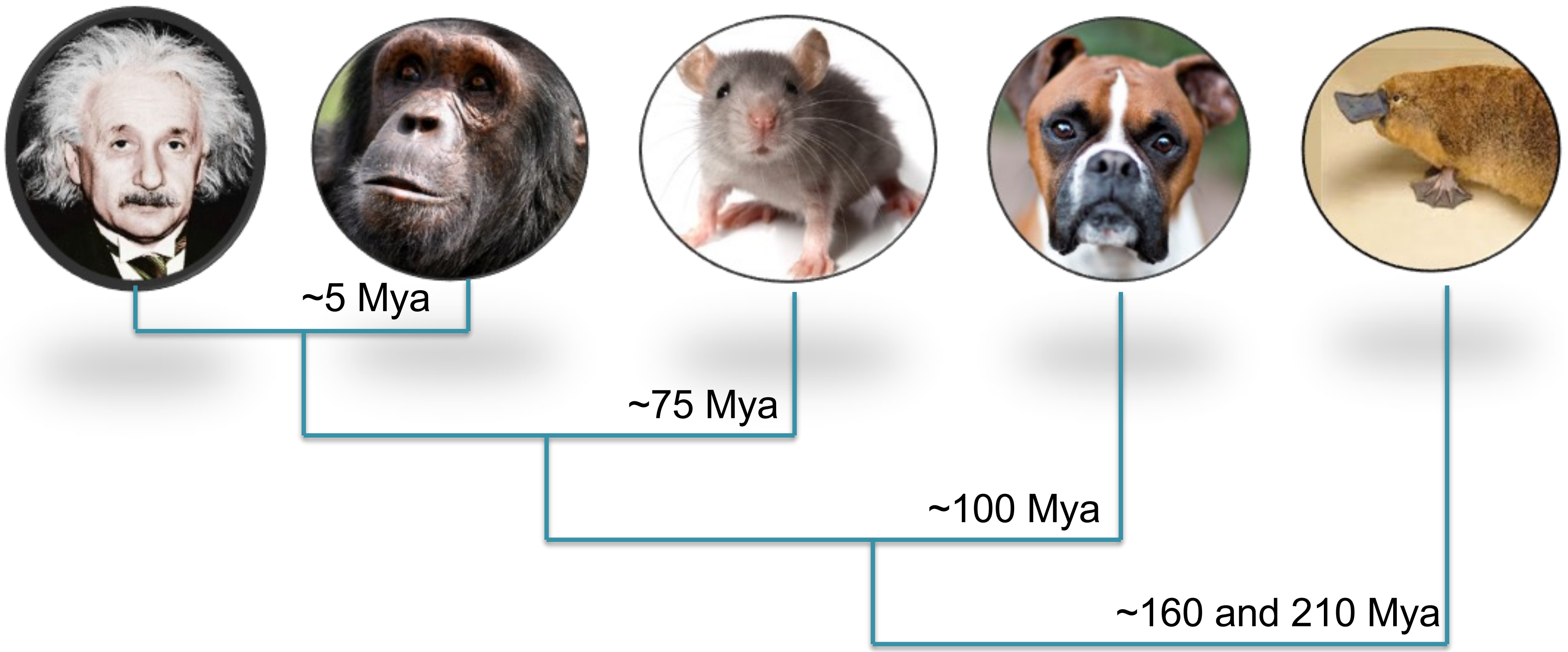
**Quantitative Techniques**

- Normal, Poisson, Binomial, P-value
- de Bruijn and overlap graphs
- Minimizers
- Dot plots
- Quality Values (Phred Scale)
- Full text indexing & BWT
- Seed & Extend
- Hidden Markov Models
- PCA / t-SNE / UMAP
- Convolutional Neural Networks
- Differential Expression
- Expectation Maximization

**What is the goal? What is the approach? What are the key challenges?**

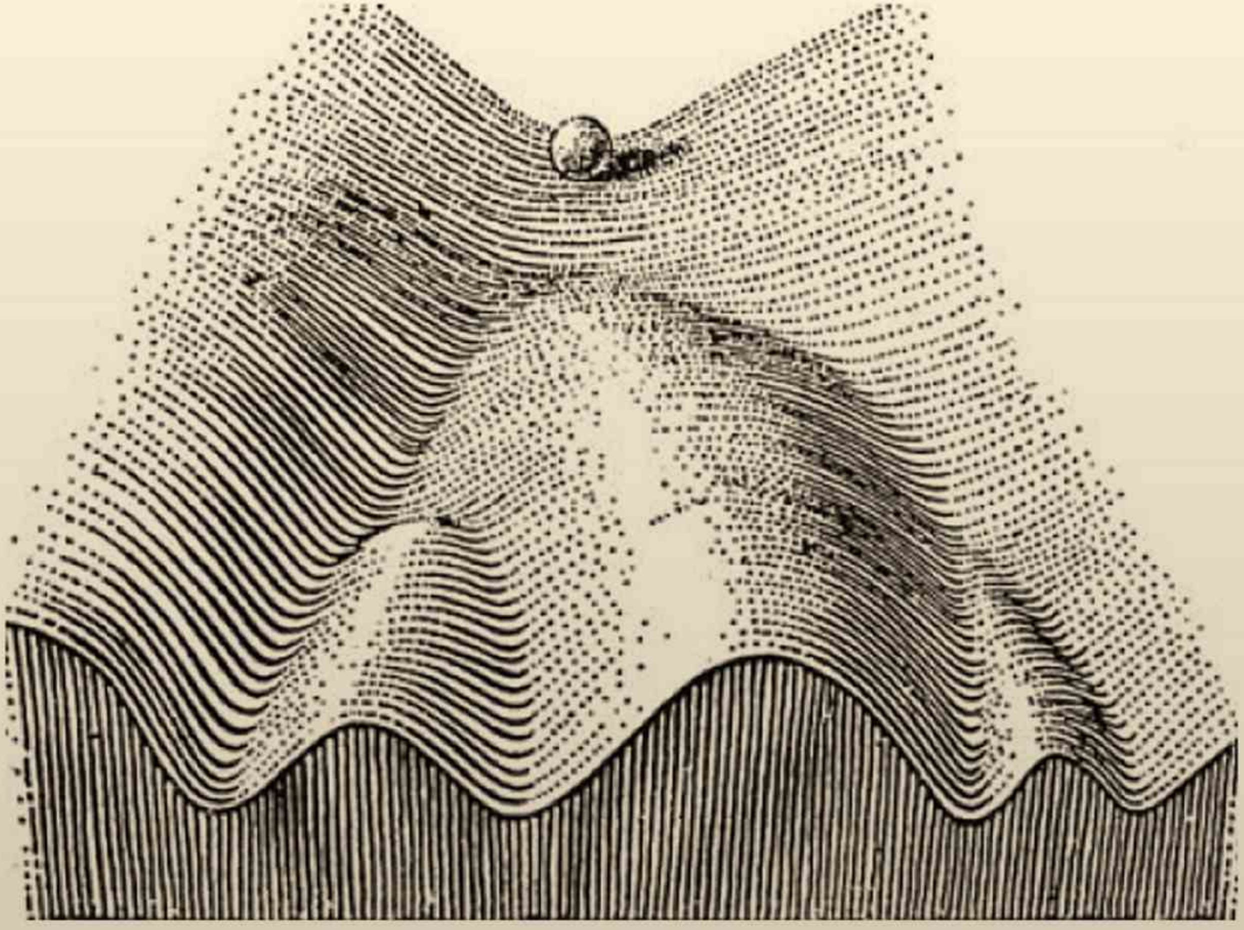
**How did we explore these topics in the homeworks and lectures?**

# Human Evolution



***As expected, the majority of platypus genes (82%; 15,312 out of 18,596) have orthologues in these five other amniotes*** (Supplementary Table 5). The remaining 'orphan' genes are expected to primarily reflect rapidly evolving genes, for which no other homologues are discernible, erroneous predictions, and true lineage-specific genes that have been lost in each of the other five species under consideration.

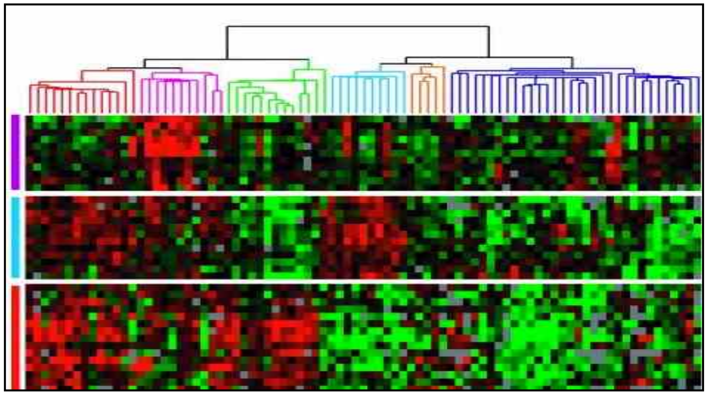
***Genome analysis of the platypus reveals unique signatures of evolution***
(2008) *Nature.* 453, 175-183 doi:10.1038/nature06936



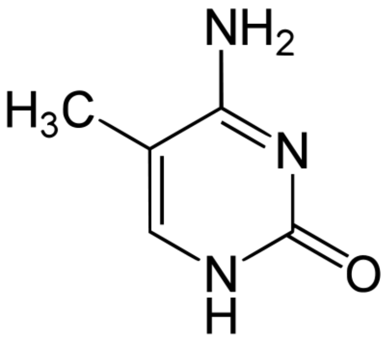
**The Strategy of the Genes**
CH Waddington (1957)

# *-seq in 4 short vignettes

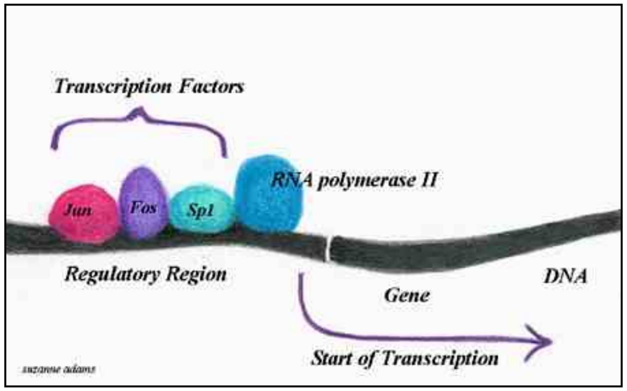
## RNA-seq



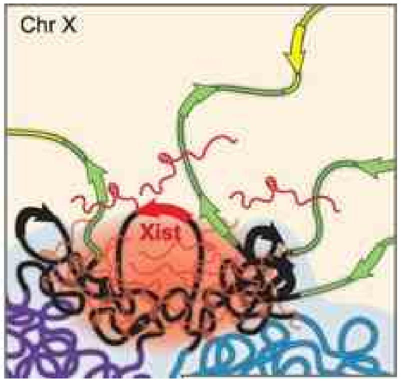
## Methyl-seq



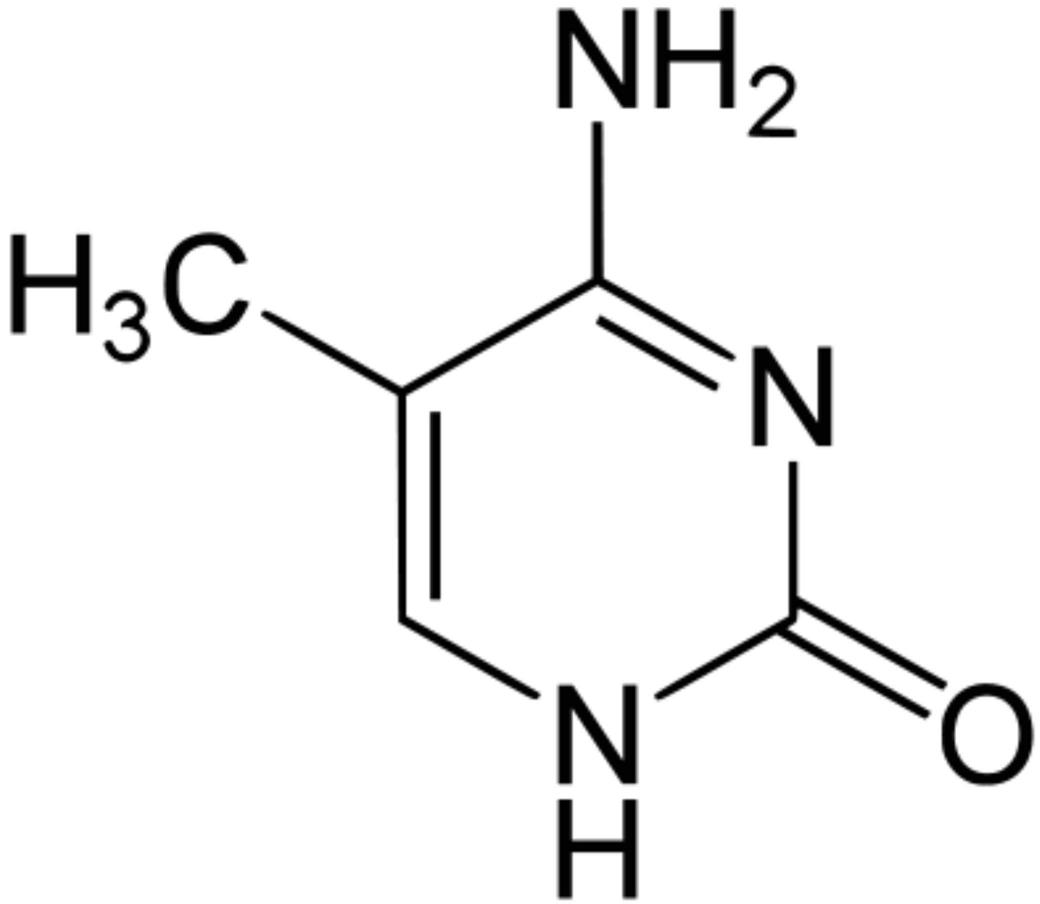
## ChIP-seq



## Hi-C



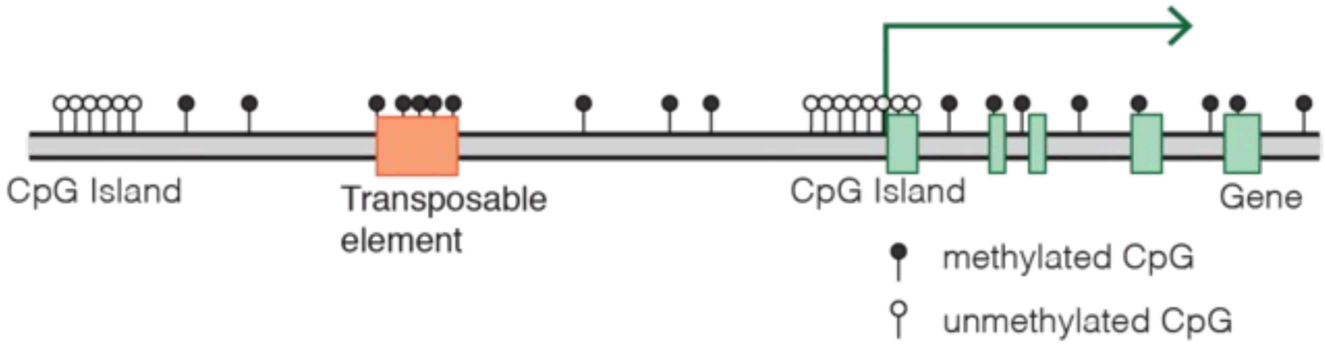
# Methyl-seq



**Finding the fifth base: Genome-wide sequencing of cytosine methylation**
Lister and Ecker (2009) *Genome Research.* 19: 959-966

# Methylation of CpG Islands

Typical mammalian DNA methylation landscape



***CpG islands are (usually) defined as regions with***

1) a length greater than 200bp,
2) a G+C content greater than 50%,
3) a ratio of observed to expected CpG greater than 0.6

***Methylation in promoter regions correlates negatively with gene expression.***

- CpG-dense promoters of actively transcribed genes are never methylated
- In mouse and human, around 60-70% of genes have a CpG island in their promoter region and most of these CpG islands remain unmethylated independently of the transcriptional activity of the gene
- Methylation of DNA itself may physically impede the binding of transcriptional proteins to the gene
- Methylated DNA may be bound by proteins known as methyl-CpG-binding domain proteins (MBDs) that can modify histones, thereby forming compact, inactive chromatin, termed heterochromatin.

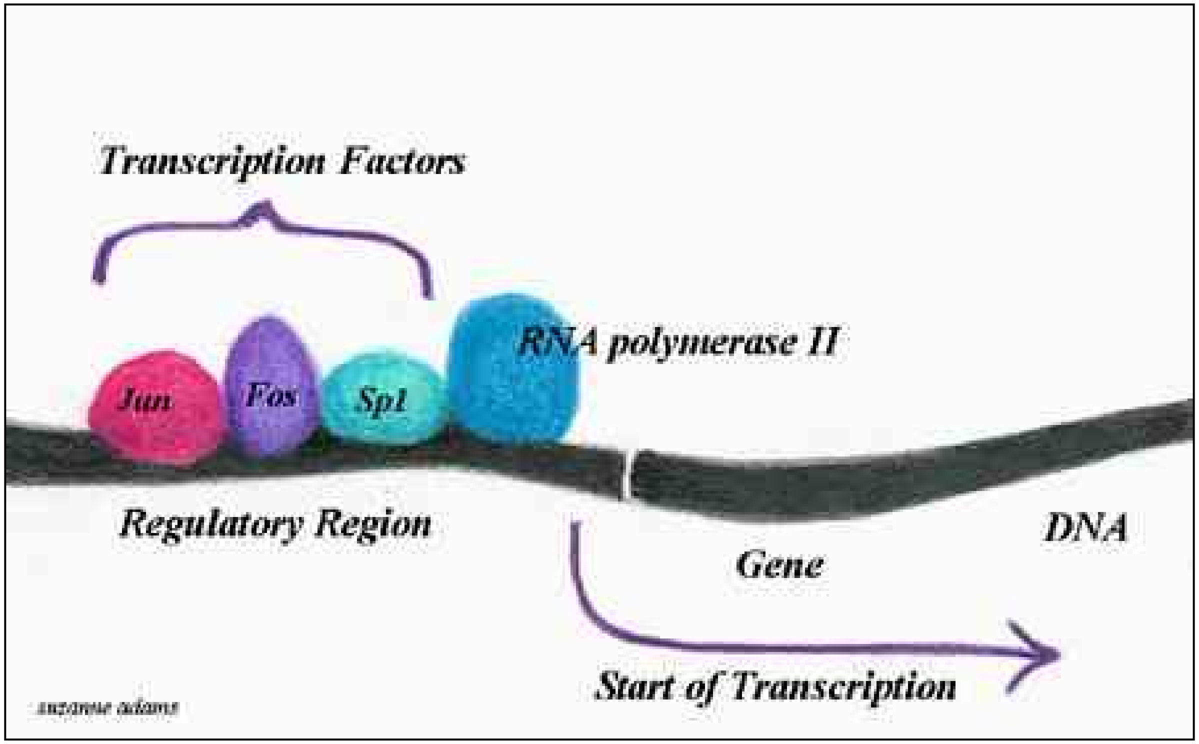
# Bisulfite Conversion

**Tr**
**w**

REMOVE CAP AND BREAK FOIL SEAL BEFORE USING
INGREDIENTS: Chili, Sugar, Salt, Garlic, Distilled Vinegar, Potassium Sorbate and Sodium Bisulfite as preservatives, Xanthan Gum.
CONTAINS SULPHITE (SODIUM BISULFITE)
INGREDIENTES: Chiles, Azúcar, Sal, Ajo, Vinagre, Sorbato de Potasio y Bisulfito de Sodio como conservadores, Goma de Xanthan.
CONTIENE SULFITO (BISULFITO DE SODIO)
STORE IN COOL DRY PLACE
MADE IN U.S.A.

**Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications**
Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

# ChIP-seq

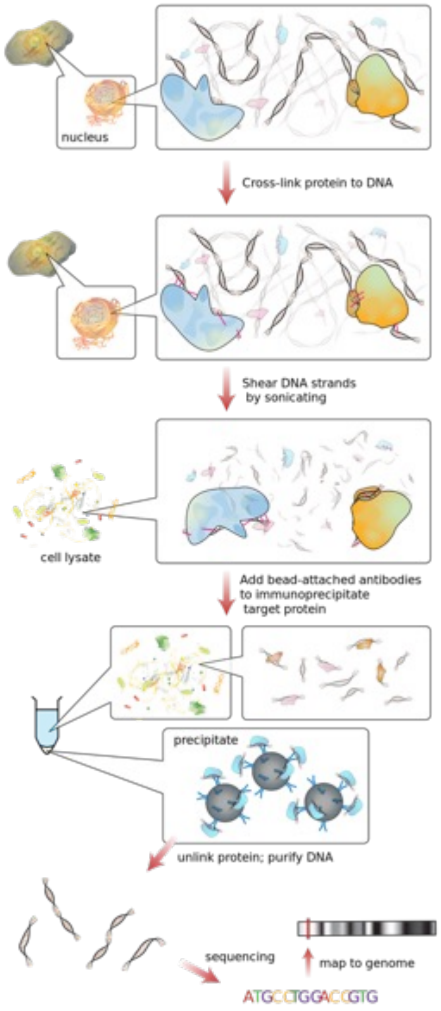


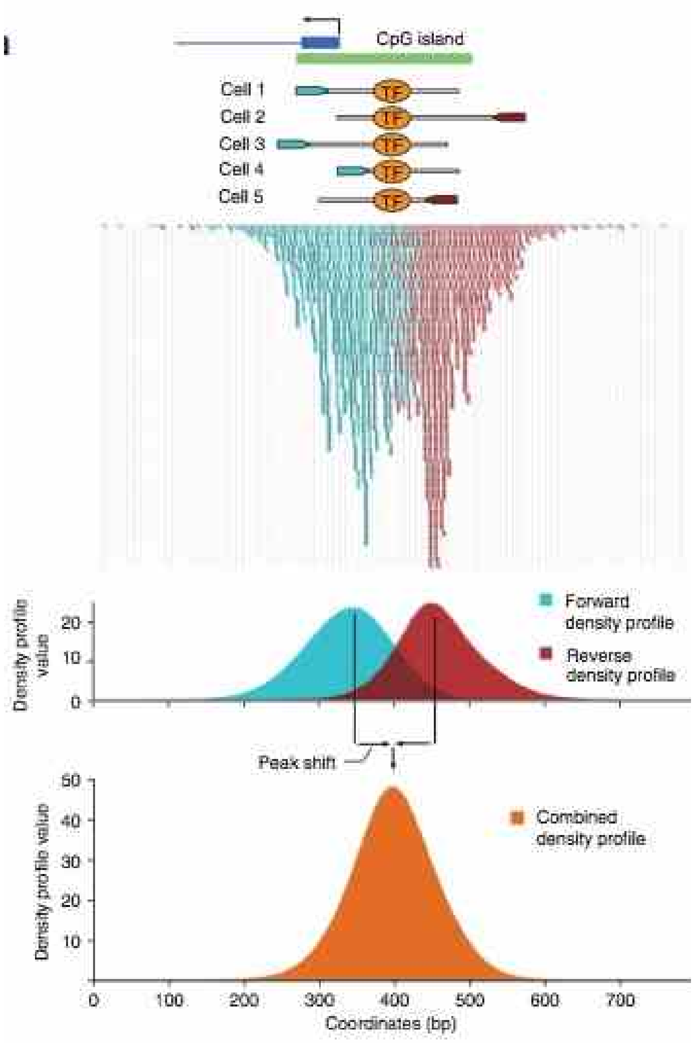
**Genome-wide mapping of in vivo protein-DNA interactions.**
Johnson *et al* (2007) *Science*. 316(5830):1497-502

# ChIP-seq: TF Binding

**Goals:**

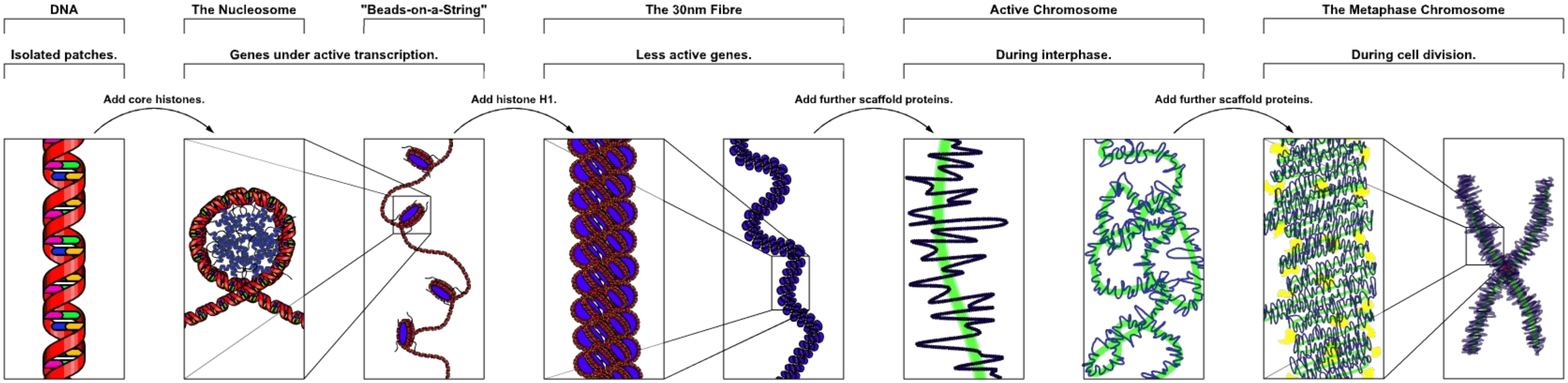
- Where are transcription factors and other proteins binding to the DNA?
- How strongly are they binding?
- Do the protein binding patterns change over developmental stages or when the cells are stressed?


nucleus
Cross-link protein to DNA
Shear DNA strands by sonicating
cell lysate
Add bead-attached antibodies to immunoprecipitate target protein
precipitate
unlink protein; purify DNA
sequencing
map to genome
ATGCCTGGACCGTG



**Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data**
Valouev et al (2008) *Nature Methods.* 5, 829 - 834

# Chromatin compaction model



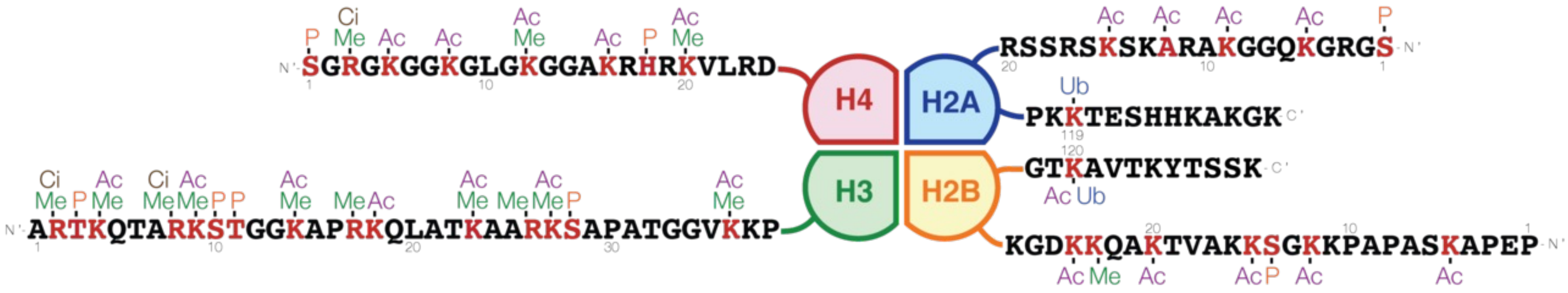
***Nucleosome is a basic unit of DNA packaging in eukaryotes***

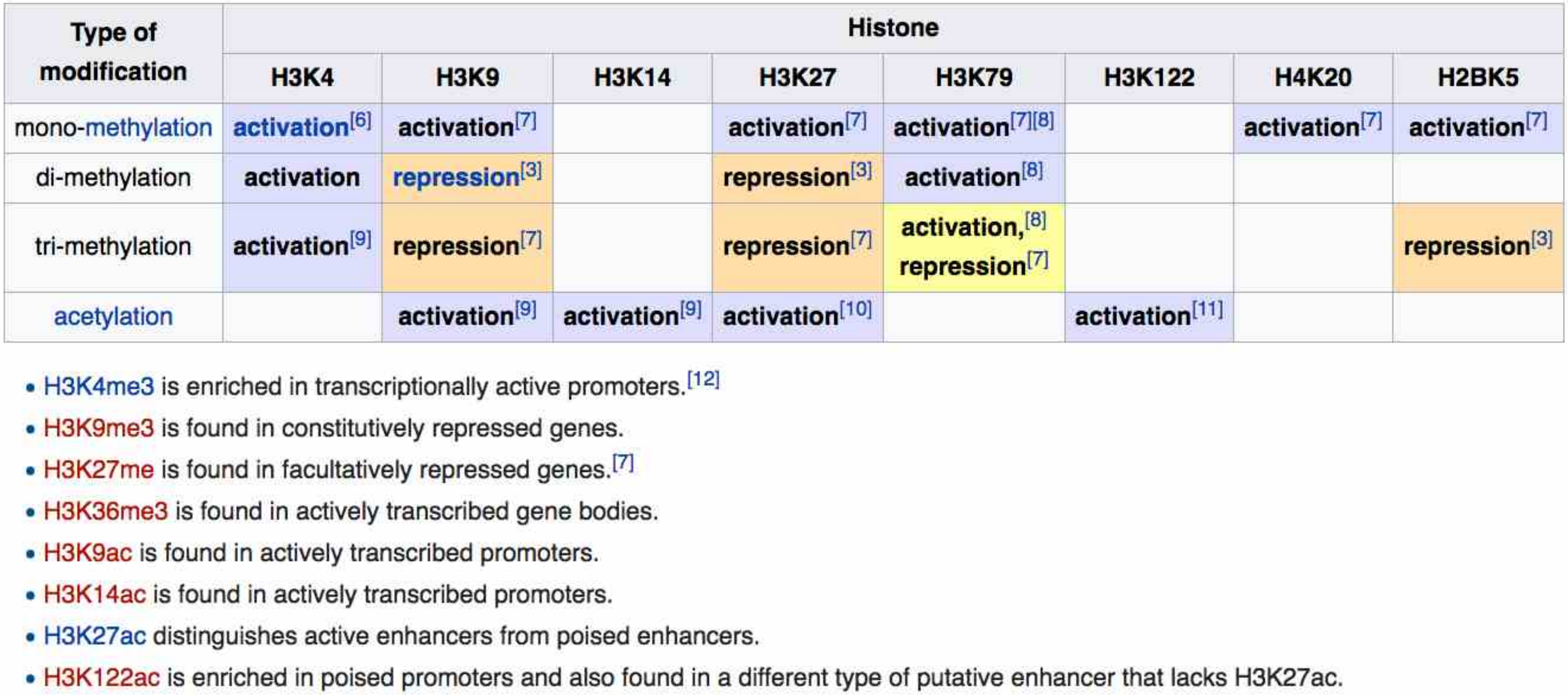
- Consists of a segment of 146bp DNA wound in sequence around eight histone protein cores (thread wrapped around a spool) followed by a ~38bp linker
- Under active transcription, nucleosomes appear as “beads-on-a-string”, but are more densely packed for less active genes

***Nucleosomes form the fundamental repeating units of eukaryotic chromatin***

- Used to pack the large eukaryotic genomes into the nucleus while still ensuring appropriate access to it (in mammalian cells approximately 2 m of linear DNA have to be packed into a nucleus of roughly 10 μm diameter).

# ChIP-seq: Histone Modifications

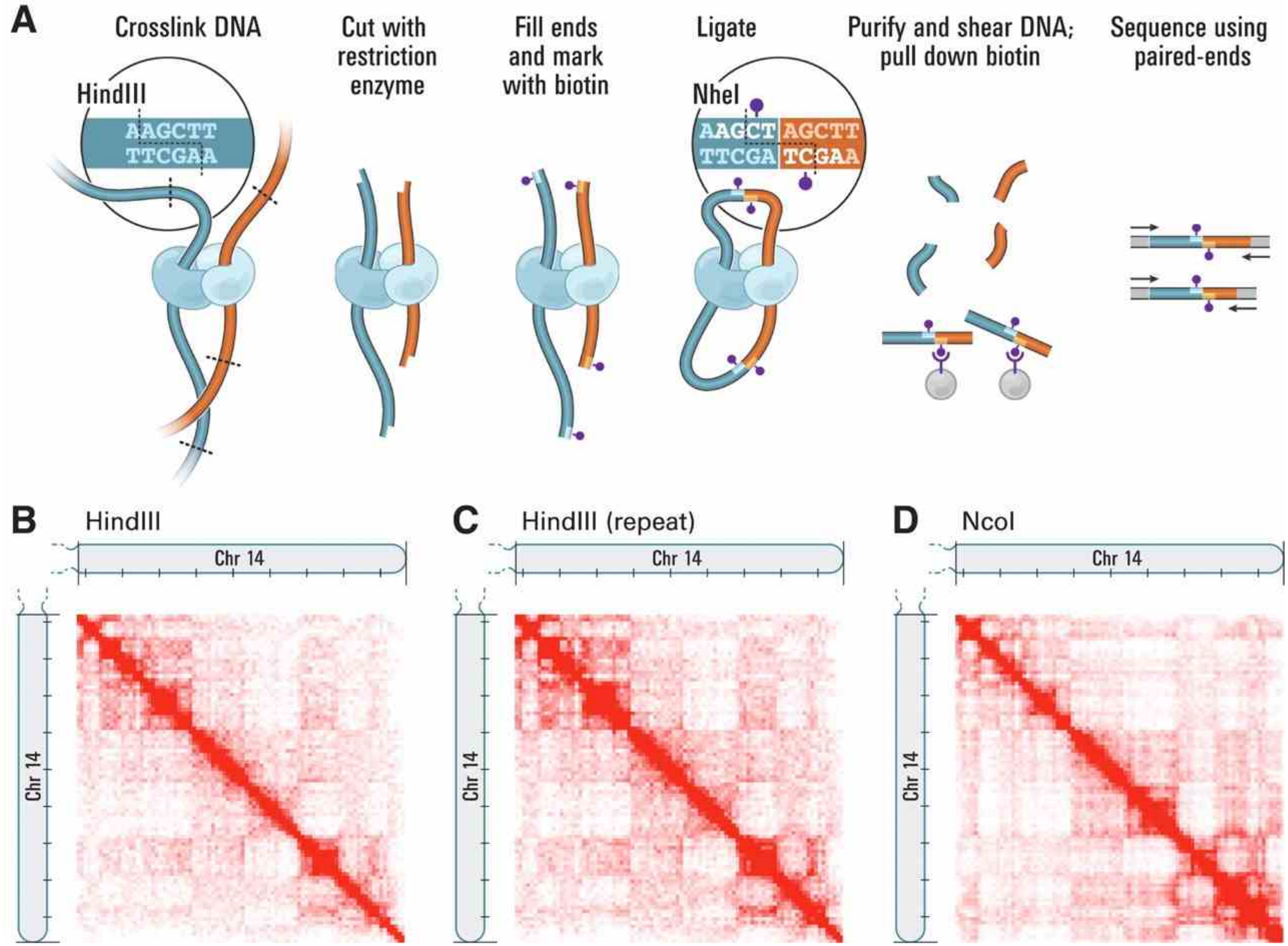




| Type of modification | Histone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | H3K4 | H3K9 | H3K14 | H3K27 | H3K79 | H3K122 | H4K20 | H2BK5 |
| mono-methylation | activation[6] | activation[7] | | activation[7] | activation[7][8] | | activation[7] | activation[7] |
| di-methylation | activation | repression[3] | | repression[3] | activation[8] | | | |
| tri-methylation | activation[9] | repression[7] | | repression[7] | activation,[8]<br>repression[7] | | | repression[3] |
| acetylation | | activation[9] | activation[9] | activation[10] | | activation[11] | | |

- H3K4me3 is enriched in transcriptionally active promoters.[12]
- H3K9me3 is found in constitutively repressed genes.
- H3K27me is found in facultatively repressed genes.[7]
- H3K36me3 is found in actively transcribed gene bodies.
- H3K9ac is found in actively transcribed promoters.
- H3K14ac is found in actively transcribed promoters.
- H3K27ac distinguishes active enhancers from poised enhancers.
- H3K122ac is enriched in poised promoters and also found in a different type of putative enhancer that lacks H3K27ac.

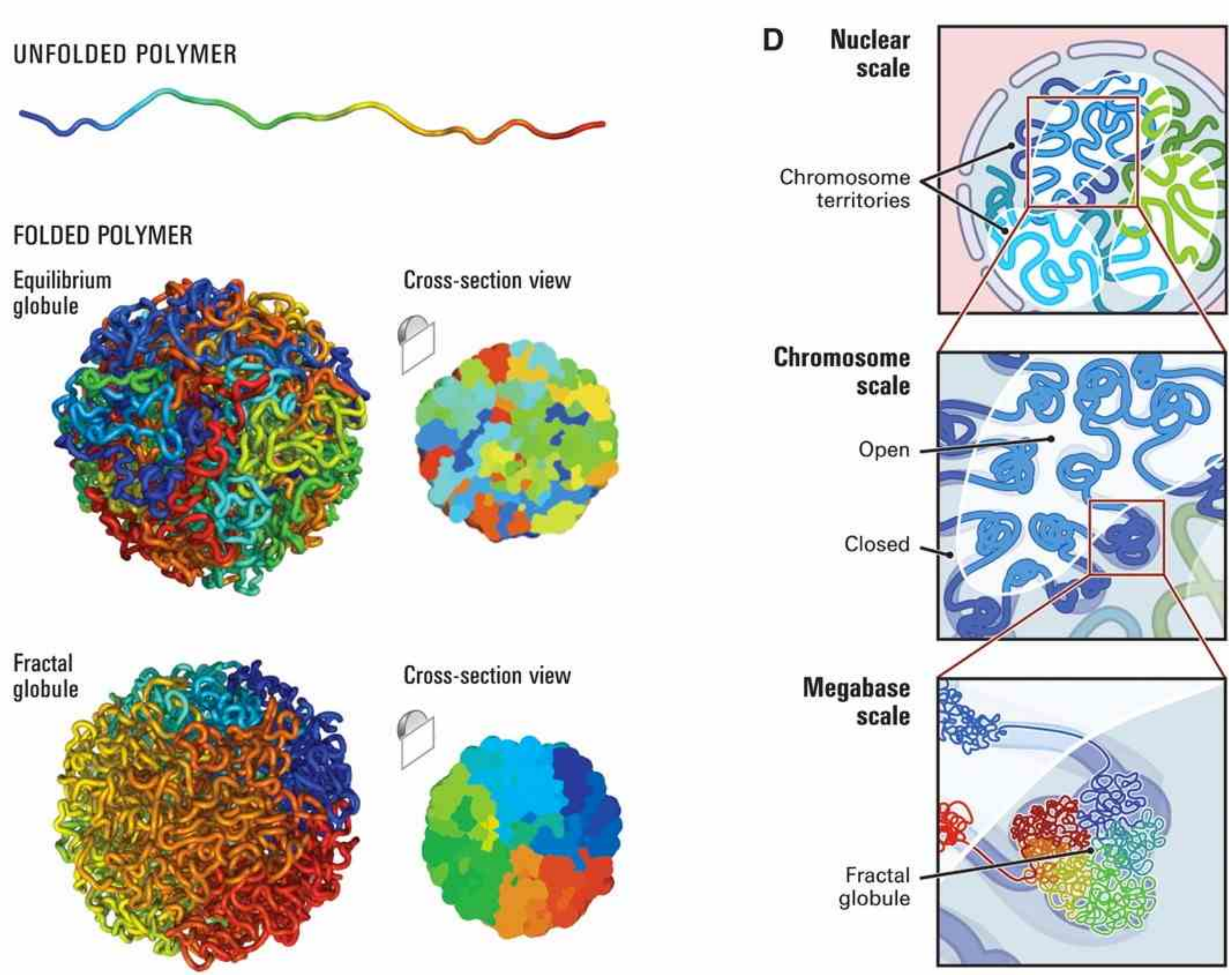
# HI-C: Mapping the folding of DNA



**Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome**

Liberman-Aiden et al. (2009) *Science*. 326 (5950): 289-293

# HI-C: Mapping the folding of DNA



**Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome**
Liberman-Aiden et al. (2009) *Science*. 326 (5950): 289-293

# Gene Regulation in 3-dimensions

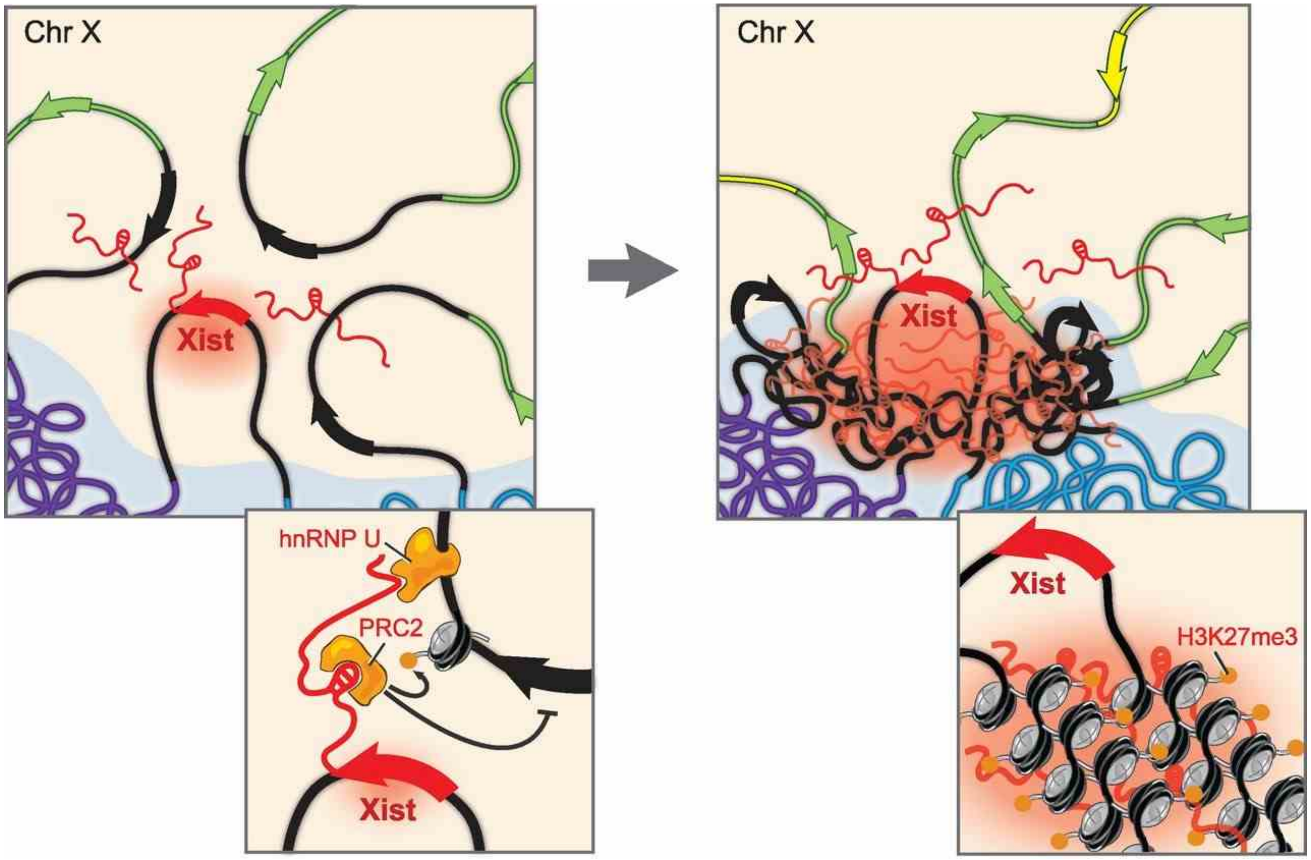
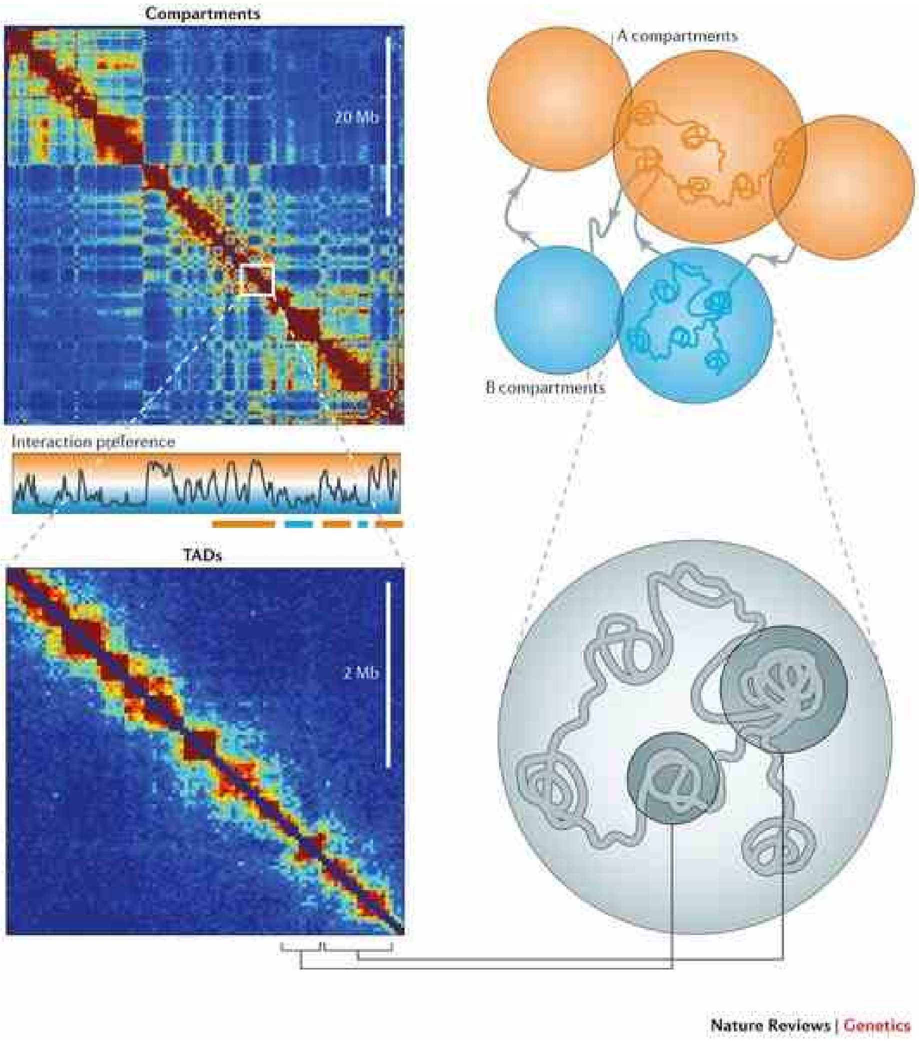


**Fig 6. A model for how Xist exploits and alters three-dimensional genome architecture to spread across the X chromosome.**

**The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome**
Engreitz et al. (2013) *Science*. 341 (6147)

# Genome compartments & TADs



***Mammalian genomes have a pattern of interactions that can be approximated by two compartments called A and B***

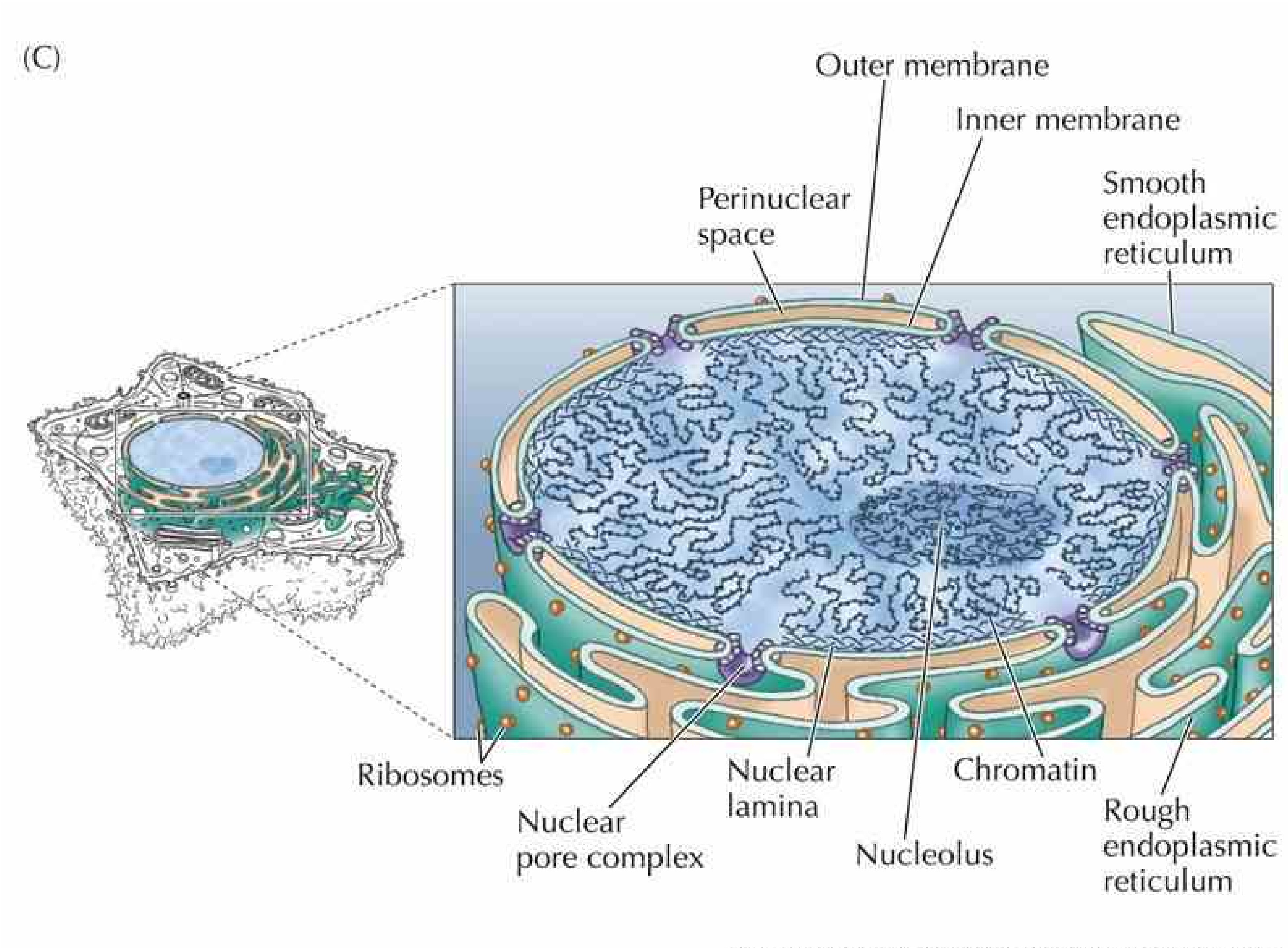
- alternate along chromosomes and have a characteristic size of ~5 Mb each.
- A compartments (orange) preferentially interact with other A compartments; B compartments (blue) associate with other B compartments.
- A compartments are largely euchromatic, transcriptionally active regions.

***Topologically associating domains (TADs)***

- TADs are smaller (~400–500 kb)
- Can be active or inactive, and adjacent TADs are not necessarily of opposite chromatin status.
- TADs are hard-wired features of chromosomes, and groups of adjacent TADs can organize in A and B compartments

**Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data**
Dekker et al. (2013) *Nature Reviews Genetics 14, 390–403*

# “Lamina-Associated Domains are the B compartment”

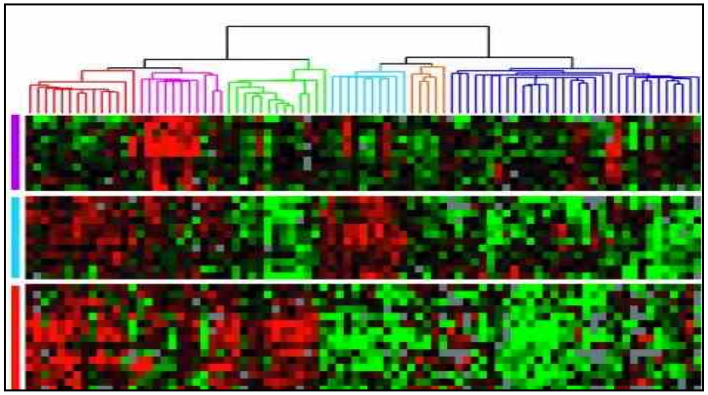


THE CELL, Fourth Edition, Figure 9.1 (Part 3) © 2006 ASM Press and Sinauer Associates, Inc.

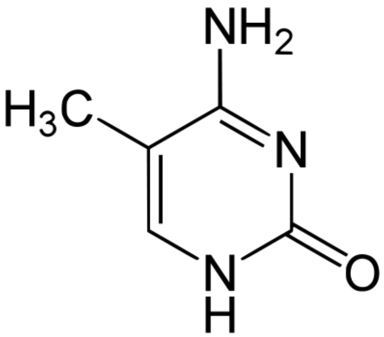
**Chromosome Conformation Paints Reveal the Role of Lamina Association in Genome Organization and Regulation**
Luperchio et al. (2017) bioRxiv. *doi: https://doi.org/10.1101/122226*

# Putting it all together!

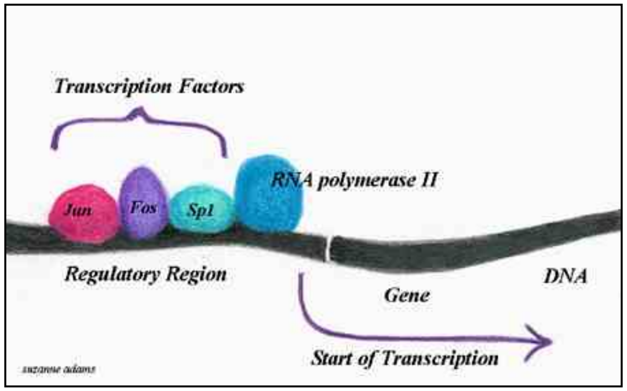
## RNA-seq



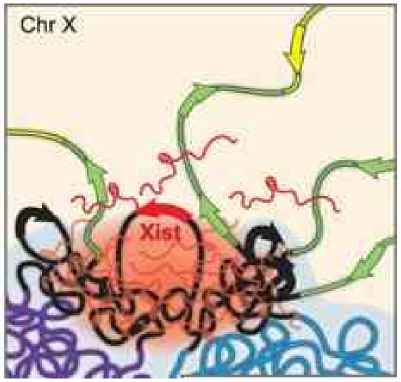
## Methyl-seq

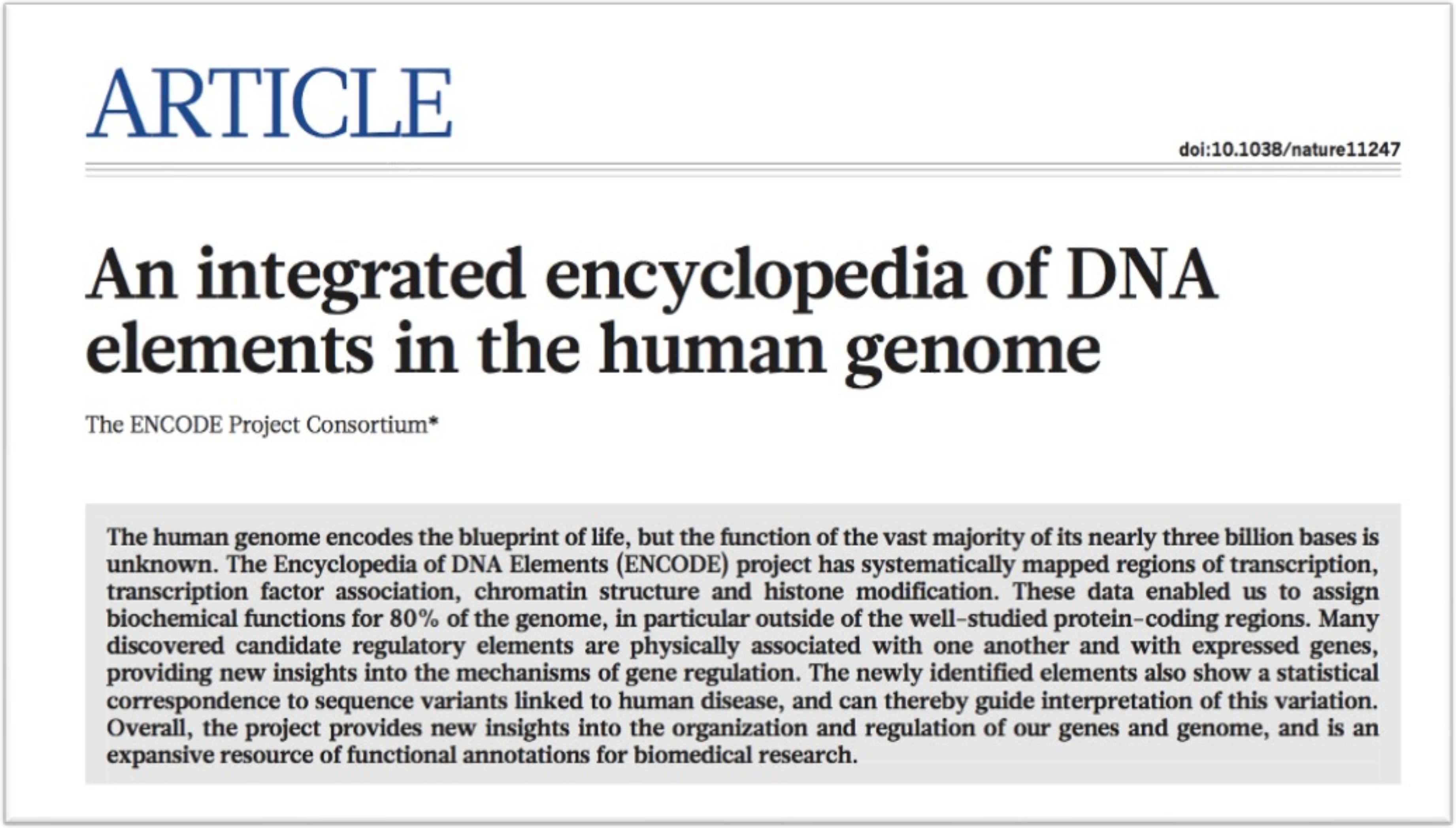


## ChIP-seq



## Hi-C





ARTICLE

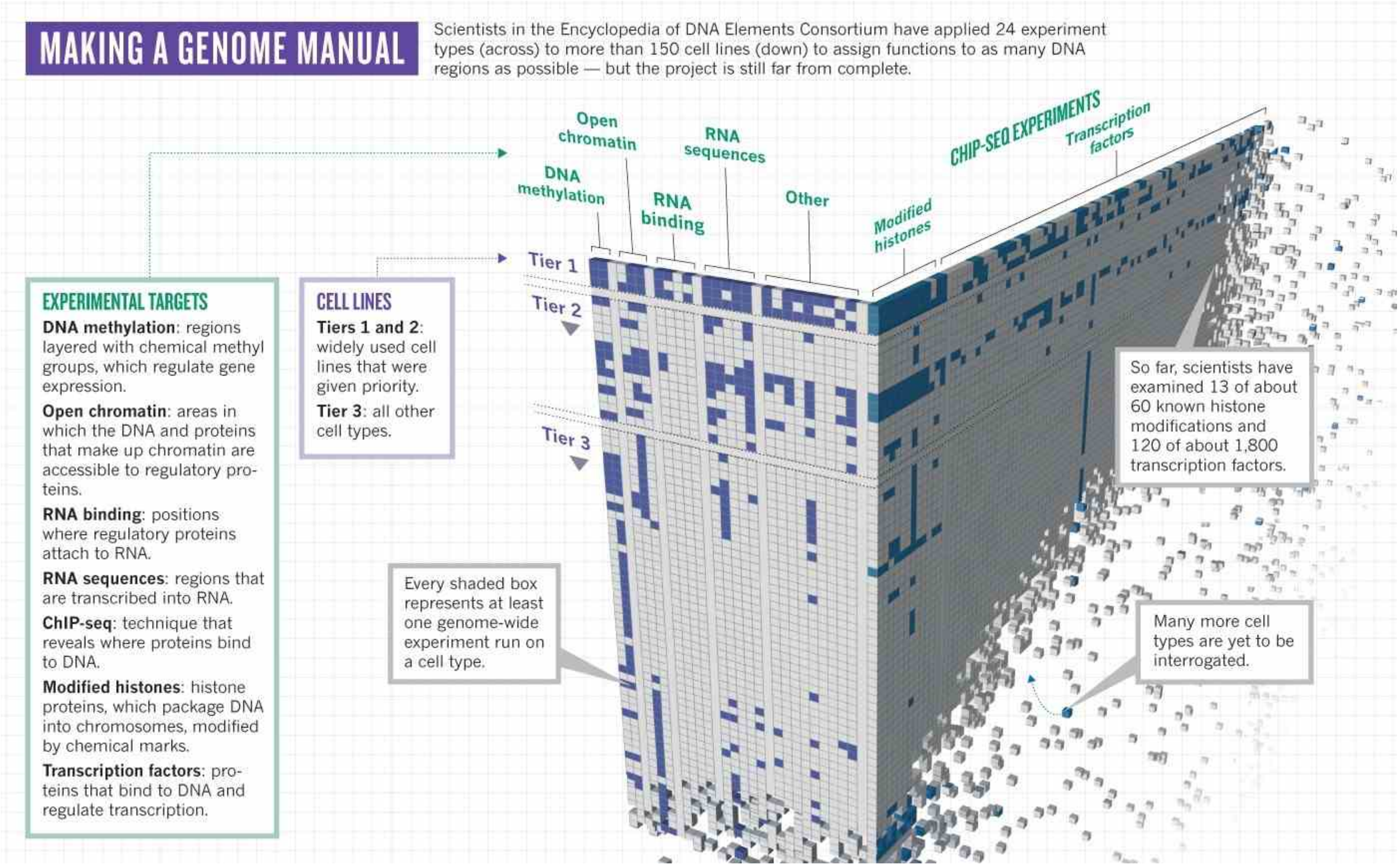
doi:10.1038/nature11247

# 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.

# ENCODE Data Sets



## MAKING A GENOME MANUAL

Scientists in the Encyclopedia of DNA Elements Consortium have applied 24 experiment types (across) to more than 150 cell lines (down) to assign functions to as many DNA regions as possible — but the project is still far from complete.

### EXPERIMENTAL TARGETS

**DNA methylation**: regions layered with chemical methyl groups, which regulate gene expression.

**Open chromatin**: areas in which the DNA and proteins that make up chromatin are accessible to regulatory proteins.

**RNA binding**: positions where regulatory proteins attach to RNA.

**RNA sequences**: regions that are transcribed into RNA.

**ChIP-seq**: technique that reveals where proteins bind to DNA.

**Modified histones**: histone proteins, which package DNA into chromosomes, modified by chemical marks.

**Transcription factors**: proteins that bind to DNA and regulate transcription.

### CELL LINES

**Tiers 1 and 2**: widely used cell lines that were given priority.

**Tier 3**: all other cell types.

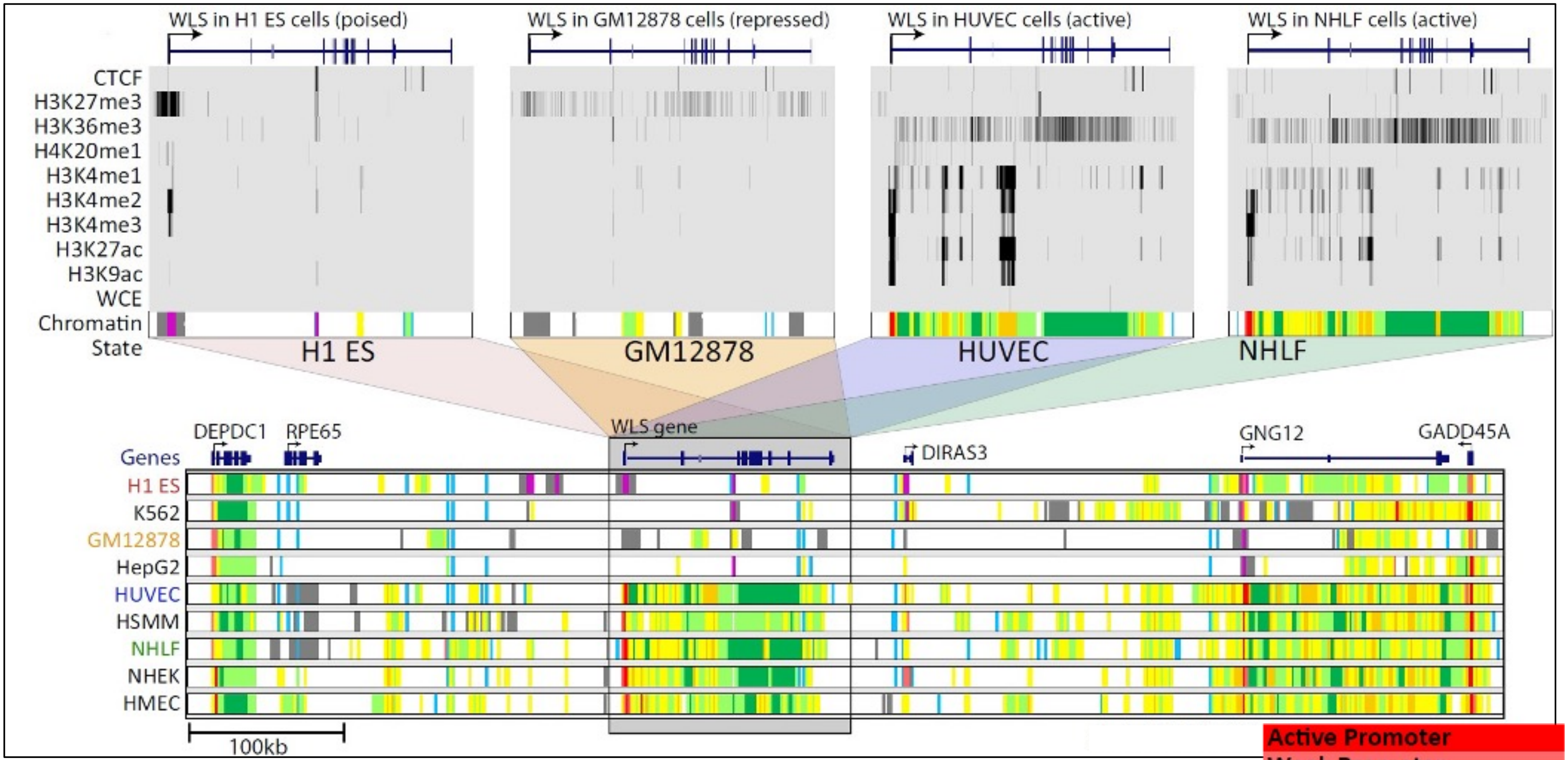
Every shaded box represents at least one genome-wide experiment run on a cell type.

So far, scientists have examined 13 of about 60 known histone modifications and 120 of about 1,800 transcription factors.

Many more cell types are yet to be interrogated.

***1,640 data sets total over 147 different cell types***

# Chromatin states dynamics across nine cell types



- Single annotation track for each cell type
- Summarize cell-type activity at a glance
- Can study 9-cell activity pattern across ↓

🟥 Active Promoter
🟥 Weak Promoter
🟪 Inactive/poised Promoter
🟧 Strong enhancer
🟧 Strong enhancer
🟨 Weak/poised enhancer
🟨 Weak/poised enhancer
🟦 Insulator
🟩 Transcriptional transition
🟩 Transcriptional elongation
🟩 Weak transcribed
⬛ Polycomb-repressed
⬜ Heterochrom; low signal

Ernst et al, Nature 2011

# Major Findings

1. *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.*

2. *Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.*

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

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

5. *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.*

6. *Single nucleotide polymorphisms (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 transcription factor.*

# Major Findings

1. ***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.***

2. *Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.*

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

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

5. *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.*

6. *Single nucleotide polymorphisms (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 transcription factor.*

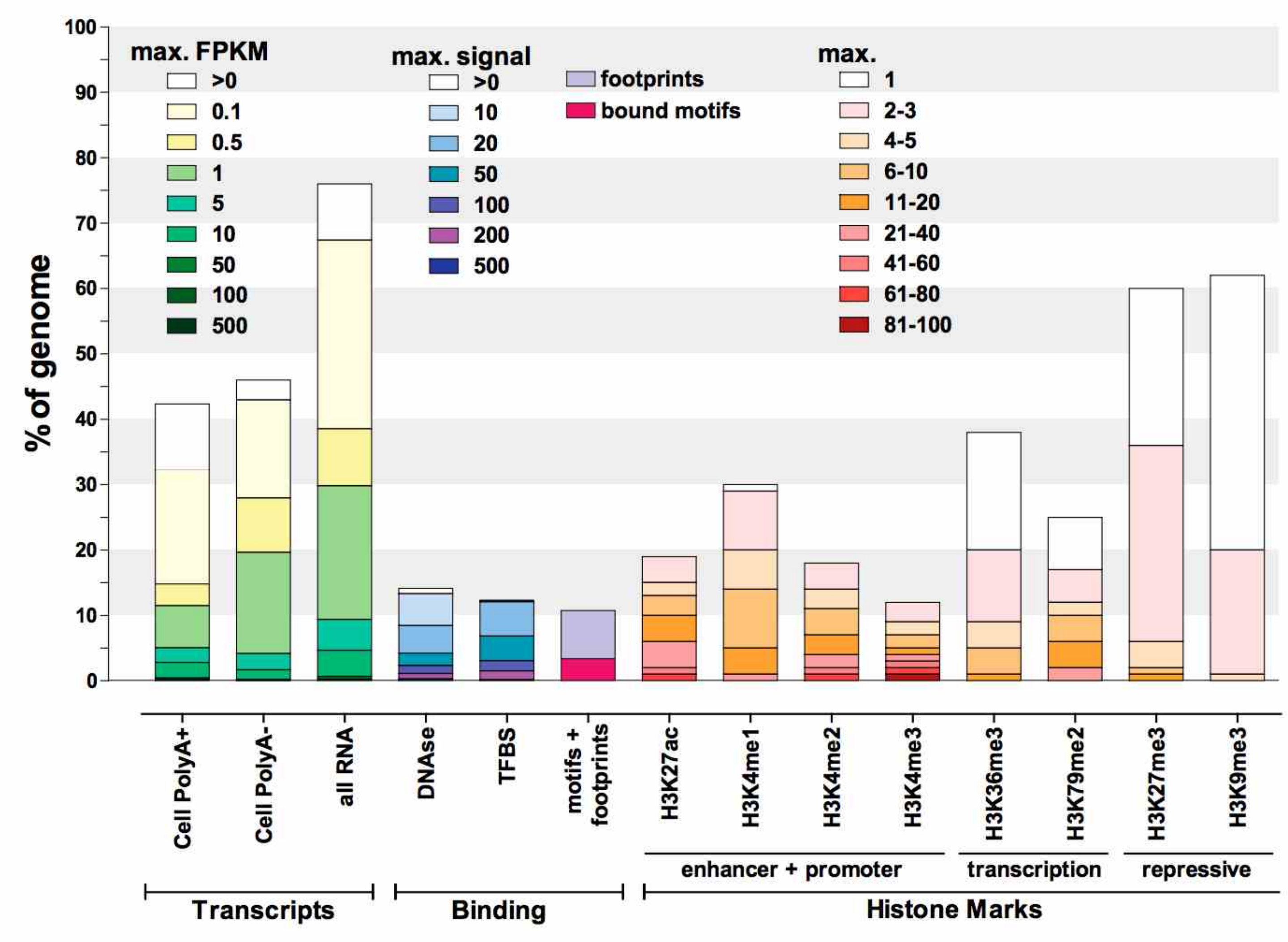
# Summary of ENCODE elements

*"Accounting for all these elements, a surprisingly large amount of the human genome, 80.4%, is covered by at least one ENCODE-identified element"*

- 62% transcribed
- 56% enriched for histone marks
- 15% open chromatin
- 8% TF binding
- 19% At least one DHS or TF Chip-seq peak
- 4% TF binding site motif
- (Note protein coding genes comprise ~2.94% of the genome)

*"Given that the ENCODE project did not assay all cell types, or all transcription factors, and in particular has sampled few specialized or developmentally restricted cell lineages,* ***these proportions must be underestimates of the total amount of functional bases."***

# Pervasive Transcription and Regulation



***Defining functional DNA elements in the human genome***
Kellis et al (2014). *PNAS* 6131–6138, doi: 10.1073/pnas.1318948111

# Major Findings

1. *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.*

2. ***Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.***

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

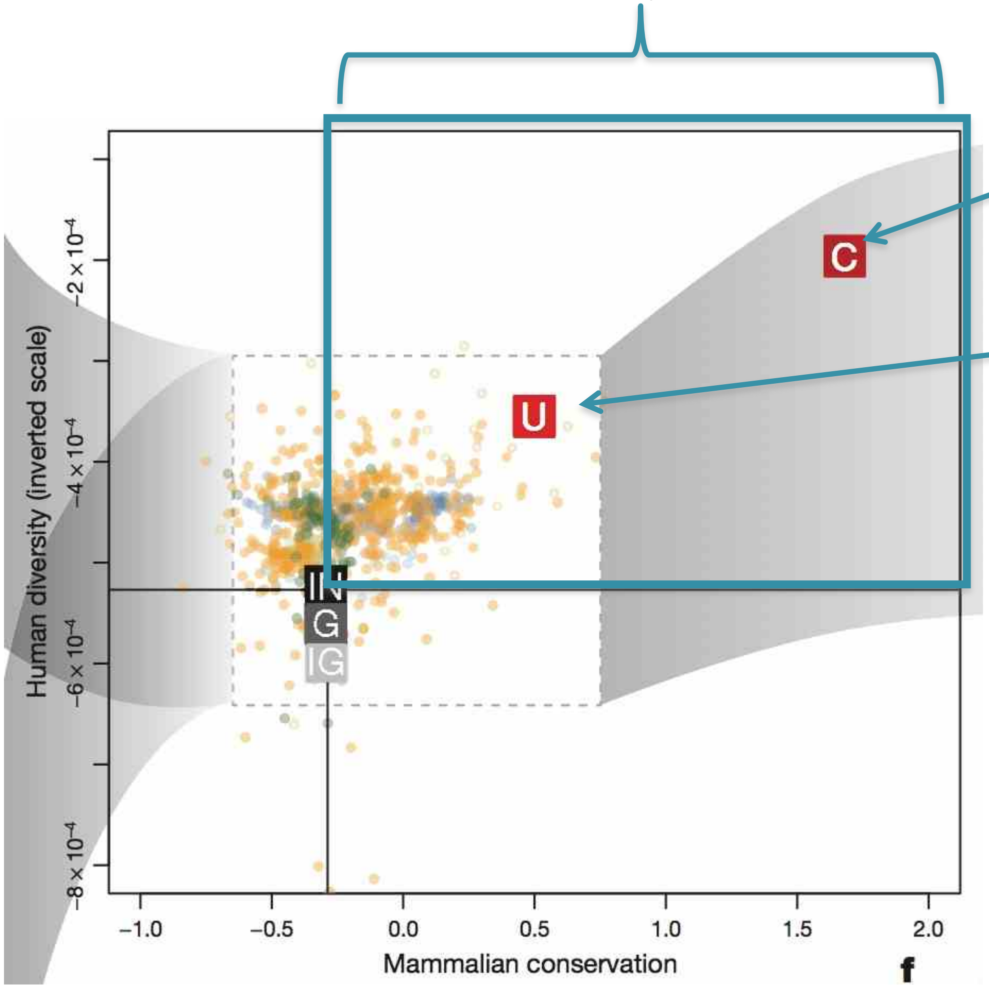
4. *It is possible to correlate quantitatively RNA sequence production and processing with both chromatin marks and transcription factor binding at promoters, indicating that promoter functionality can explain most of the variation in RNA expression.*

5. *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.*

6. *Single nucleotide polymorphisms (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 transcription factor.*

# Impact and Evidence of Selection

Most constrained
=> Most likely functional



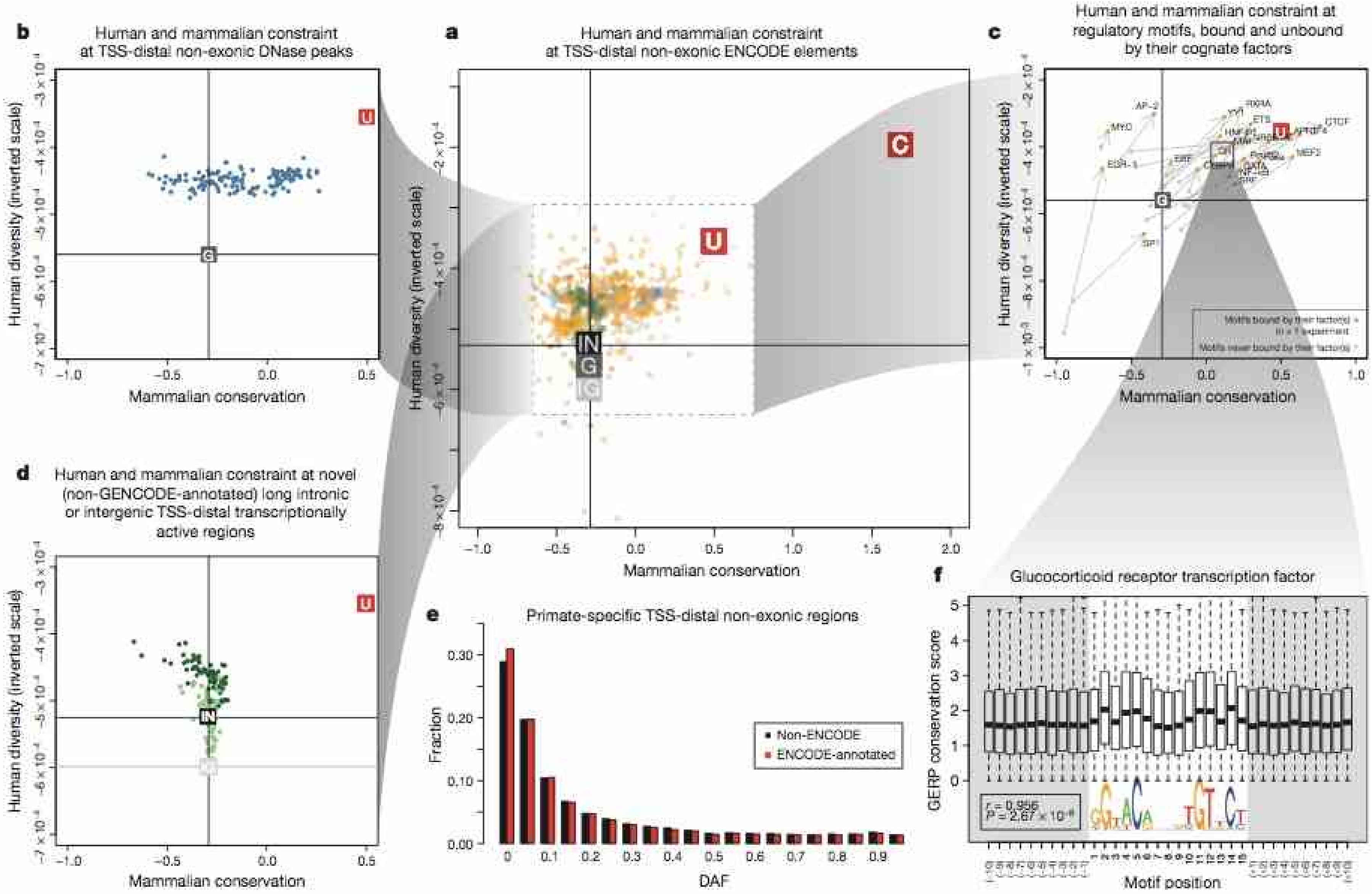
Average values across
Protein coding sequences

Average values across
UTR sequences

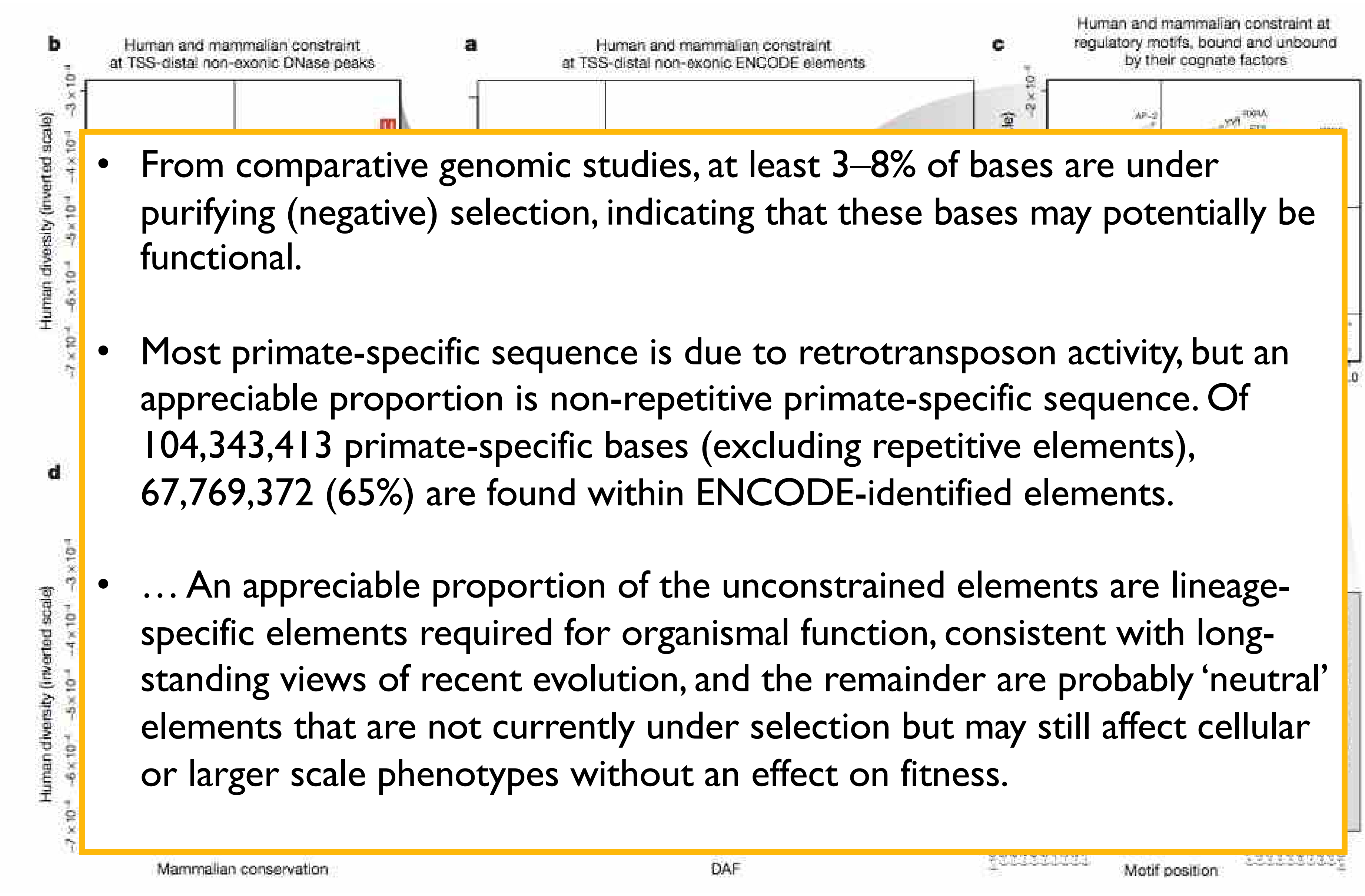
For a given ENCODE region, how much conservation do we see across modern humans
(1000 genomes project)

For a given ENCODE region, how much conservation do we see across 24 sequenced mammalian genomes?

# Impact and Evidence of Selection



# Impact and Evidence of Selection



- From comparative genomic studies, at least 3–8% of bases are under purifying (negative) selection, indicating that these bases may potentially be functional.

- Most primate-specific sequence is due to retrotransposon activity, but an appreciable proportion is non-repetitive primate-specific sequence. Of 104,343,413 primate-specific bases (excluding repetitive elements), 67,769,372 (65%) are found within ENCODE-identified elements.

- … An appreciable proportion of the unconstrained elements are lineage-specific elements required for organismal function, consistent with long-standing views of recent evolution, and the remainder are probably ‘neutral’ elements that are not currently under selection but may still affect cellular or larger scale phenotypes without an effect on fitness.

# Major Findings

1. *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.*

2. *Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.*

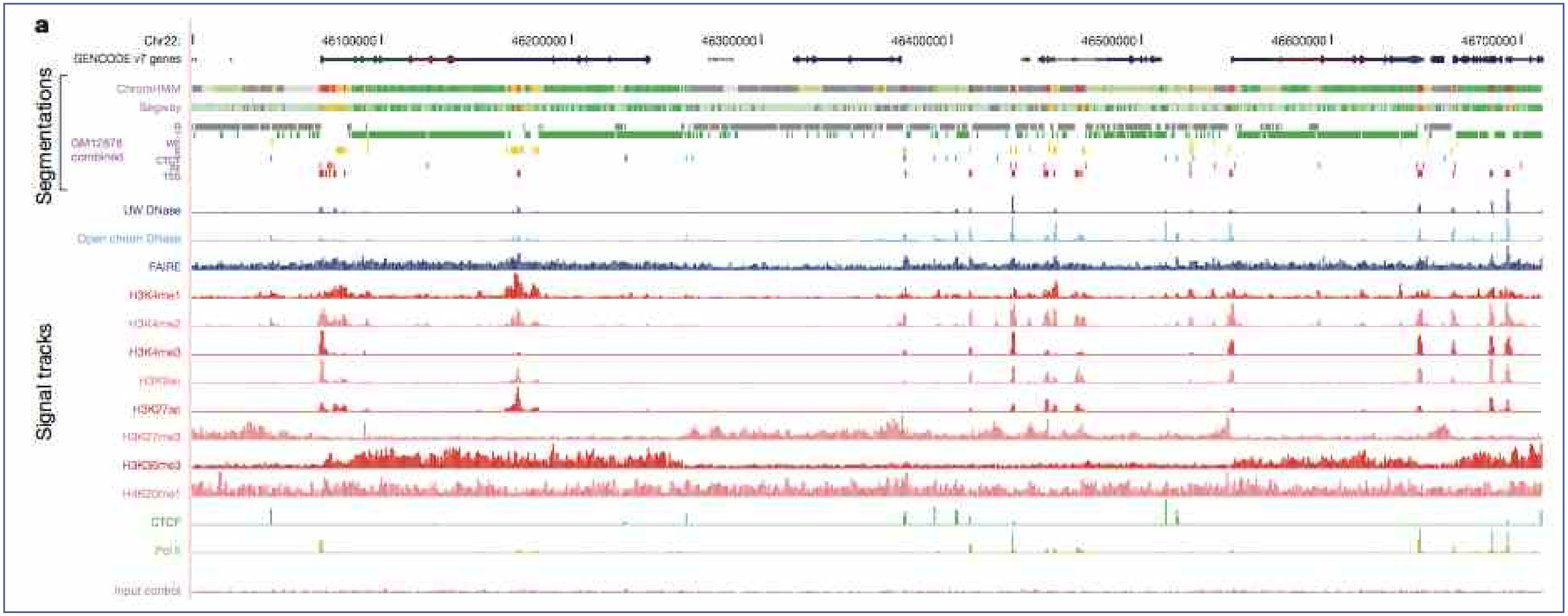
3. ***Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands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

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

5. *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.*

6. *Single nucleotide polymorphisms (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 transcription factor.*

# Signal Integration



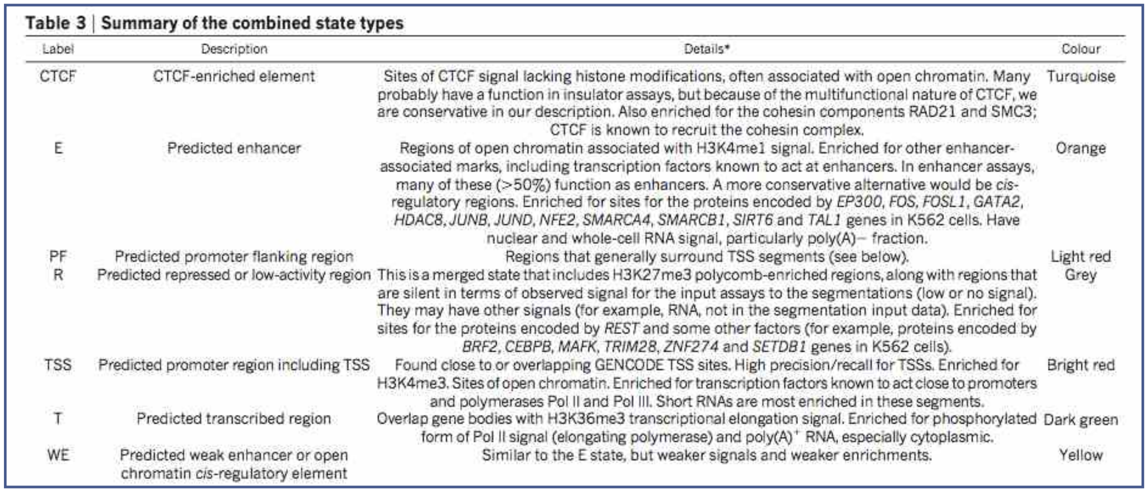


**Table 3 | Summary of the combined state types**

| Label | Description | Details* | Colour |
|---|---|---|---|
| CTCF | CTCF-enriched element | Sites of CTCF signal lacking histone modifications, often associated with open chromatin. Many probably have a function in insulator assays, but because of the multifunctional nature of CTCF, we are conservative in our description. Also enriched for the cohesin components RAD21 and SMC3; CTCF is known to recruit the cohesin complex. | Turquoise |
| E | Predicted enhancer | Regions of open chromatin associated with H3K4me1 signal. Enriched for other enhancer-associated marks, including transcription factors known to act at enhancers. In enhancer assays, many of these (>50%) function as enhancers. A more conservative alternative would be *cis*-regulatory regions. Enriched for sites for the proteins encoded by *EP300*, *FOS*, *FOSL1*, *GATA2*, *HDAC8*, *JUNB*, *JUND*, *NFE2*, *SMARCA4*, *SMARCB1*, *SIRT6* and *TAL1* genes in K562 cells. Have nuclear and whole-cell RNA signal, particularly poly(A)− fraction. | Orange |
| PF | Predicted promoter flanking region | Regions that generally surround TSS segments (see below). | Light red |
| R | Predicted repressed or low-activity region | This is a merged state that includes H3K27me3 polycomb-enriched regions, along with regions that are silent in terms of observed signal for the input assays to the segmentations (low or no signal). They may have other signals (for example, RNA, not in the segmentation input data). Enriched for sites for the proteins encoded by *REST* and some other factors (for example, proteins encoded by *BRF2*, *CEBPB*, *MAFK*, *TRIM28*, *ZNF274* and *SETDB1* genes in K562 cells). | Grey |
| TSS | Predicted promoter region including TSS | Found close to or overlapping GENCODE TSS sites. High precision/recall for TSSs. Enriched for H3K4me3. Sites of open chromatin. Enriched for transcription factors known to act close to promoters and polymerases Pol II and Pol III. Short RNAs are most enriched in these segments. | Bright red |
| T | Predicted transcribed region | Overlap gene bodies with H3K36me3 transcriptional elongation signal. Enriched for phosphorylated form of Pol II signal (elongating polymerase) and poly(A)$^{+}$ RNA, especially cytoplasmic. | Dark green |
| WE | Predicted weak enhancer or open chromatin *cis*-regulatory element | Similar to the E state, but weaker signals and weaker enrichments. | Yellow |

- Use ChromHMM and Segway to Summarize the individual assays into 7 functional/regulatory states

# Major Findings

ENCODE

1. *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.*

2. *Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.*

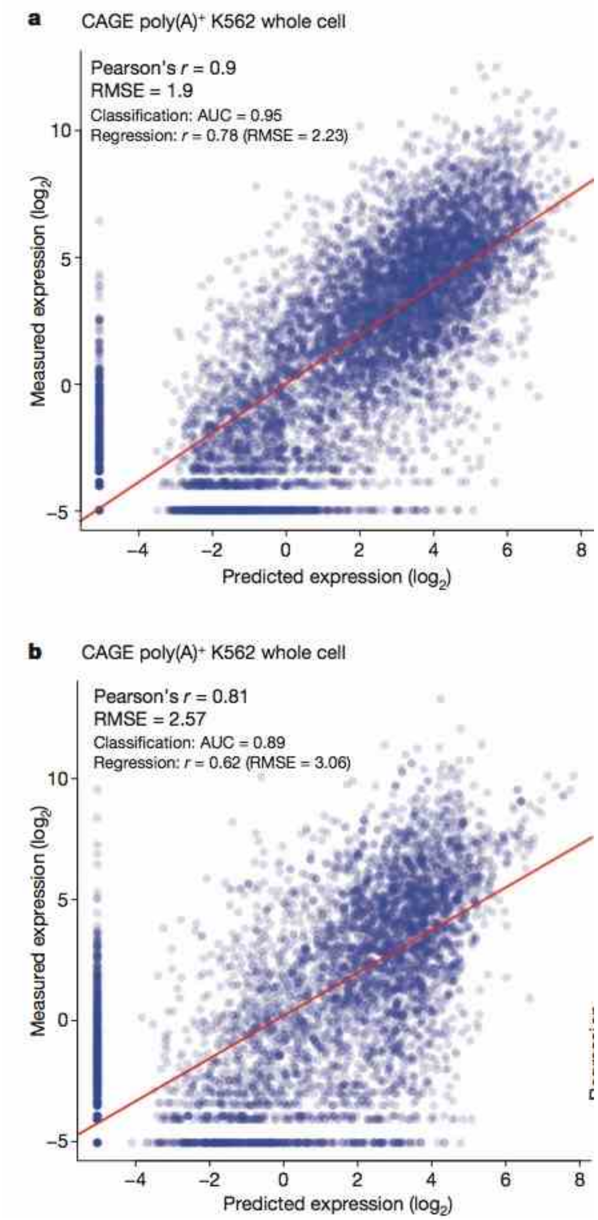
3. *Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of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

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

5. *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.*

6. *Single nucleotide polymorphisms (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 transcription factor.*

# Expression Modeling



- Developed predictive models to explore the interaction between histone modifications and transcription factor binding towards level of transcription

- The best models had two components: an initial classification component (on/off) and a second quantitative model component

- Together, these correlation models indicate both that a limited set of chromatin marks are sufficient to 'explain' transcription and that a variety of transcription factors might have broad roles in general transcription levels across many genes

**Modeling gene expression using chromatin features in various cellular context**
Dong et al. (2012) *Genome Biology.* 12:R53

# Major Findings

1. *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.*

2. *Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.*

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

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

5. ***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.***

6. *Single nucleotide polymorphisms (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 transcription factor.*

# Many variants in ENCODE-regions

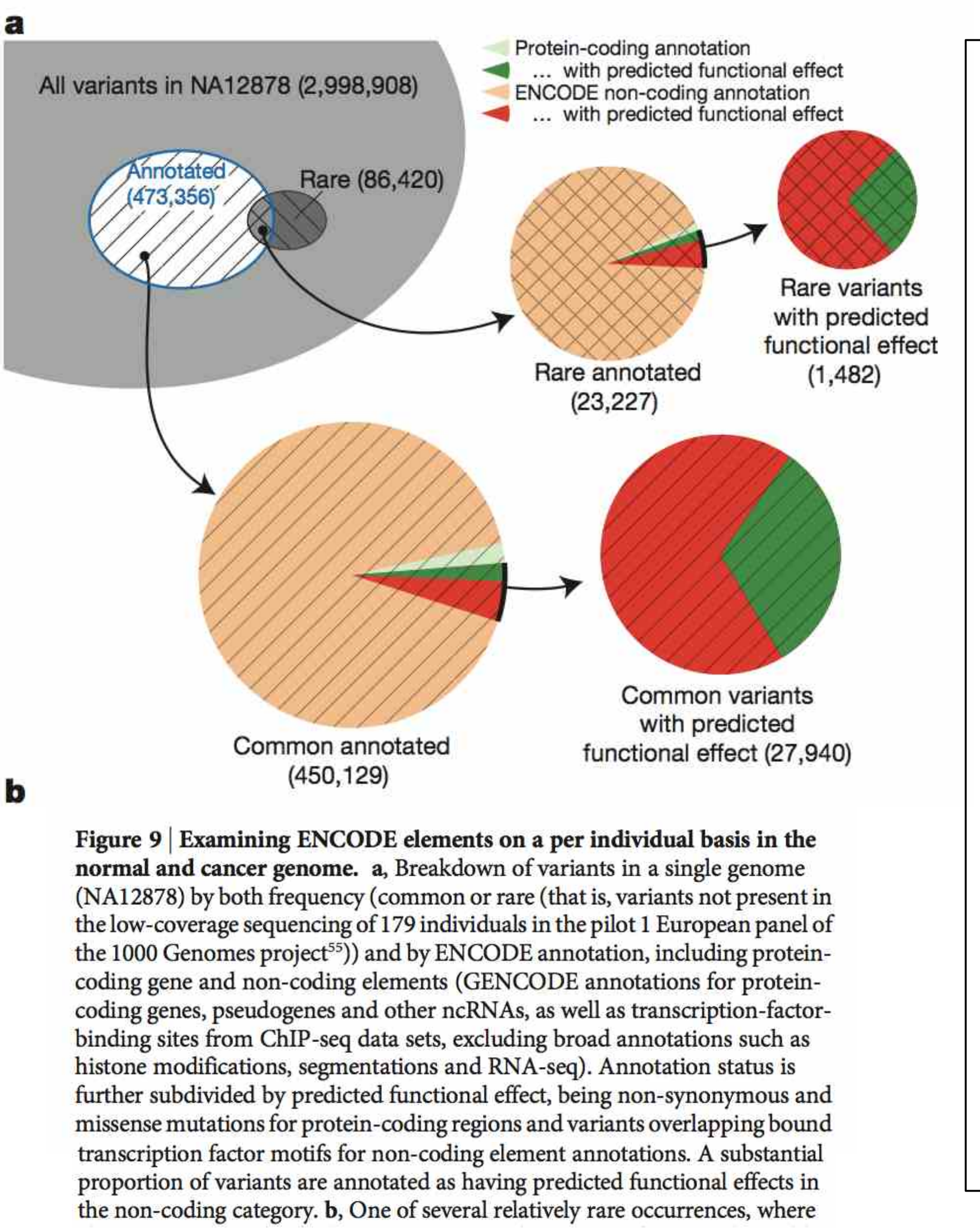


**Figure 9 | Examining ENCODE elements on a per individual basis in the normal and cancer genome.** **a**, Breakdown of variants in a single genome (NA12878) by both frequency (common or rare (that is, variants not present in the low-coverage sequencing of 179 individuals in the pilot 1 European panel of the 1000 Genomes project[55])) and by ENCODE annotation, including protein-coding gene and non-coding elements (GENCODE annotations for protein-coding genes, pseudogenes and other ncRNAs, as well as transcription-factor-binding sites from ChIP-seq data sets, excluding broad annotations such as histone modifications, segmentations and RNA-seq). Annotation status is further subdivided by predicted functional effect, being non-synonymous and missense mutations for protein-coding regions and variants overlapping bound transcription factor motifs for non-coding element annotations. A substantial proportion of variants are annotated as having predicted functional effects in the non-coding category. **b**, One of several relatively rare occurrences, where

Breakdown of variants by frequency

- Common or Rare (that is, variants not present in the low-coverage sequencing of 179 individuals in the pilot 1 European panel of the 1000 Genomes project)
- ENCODE annotation, including protein-coding gene and non-coding elements

Annotation status is further subdivided by predicted functional effect

- non-synonymous and missense mutations for protein-coding regions and variants overlapping bound transcription factor motifs for non-coding element annotations.

***A substantial proportion of variants are annotated as having predicted functional effects in the non-coding category.***

# Major Findings

1. *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.*

2. *Primate-specific elements as well as elements without detectable mammalian constraint show, in aggregate, evidence of negative selection; thus, some of them are expected to be functional.*

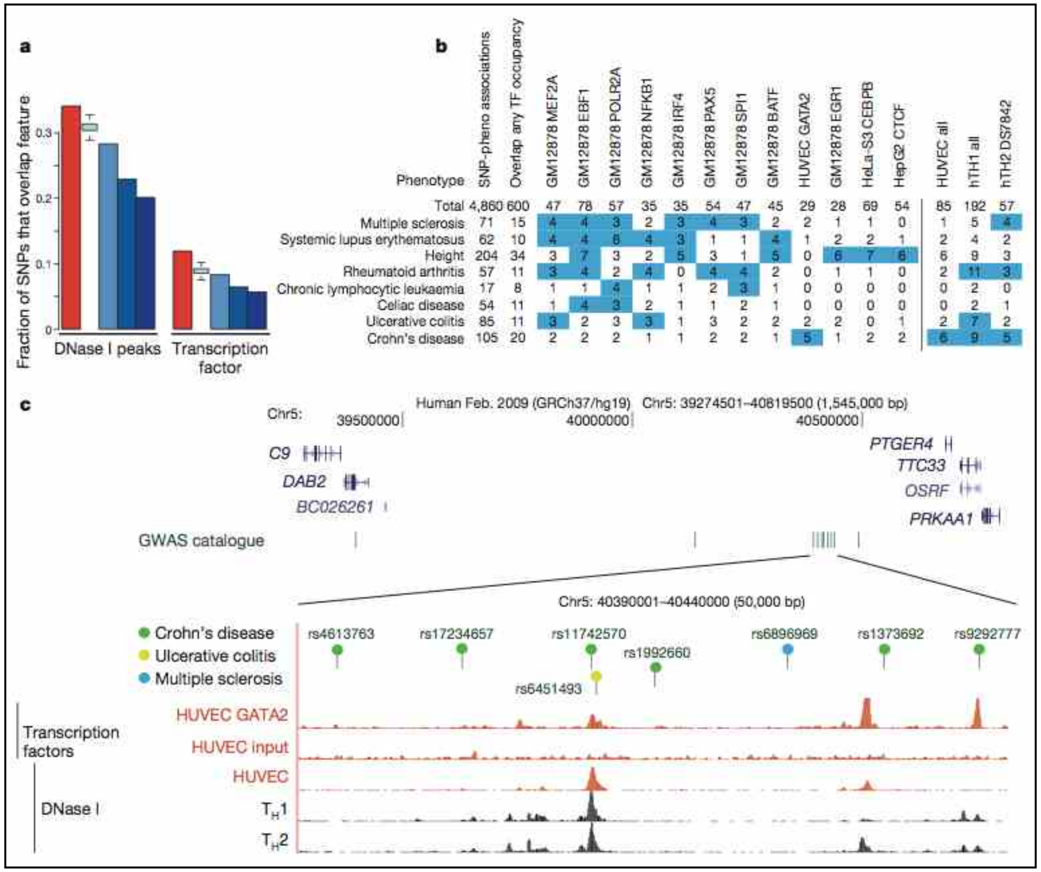
3. *Classifying the genome into seven chromatin states indicates an initial set of 399,124 regions with enhancer-like features and 70,292 regions with promoter-like features, as well as hundreds of thousands of quiescent regions. High-resolution analyses further subdivide the genome into thousands of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

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

5. *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.*

6. ***Single nucleotide polymorphisms (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 transcription factor.***

# ENCODE and Disease

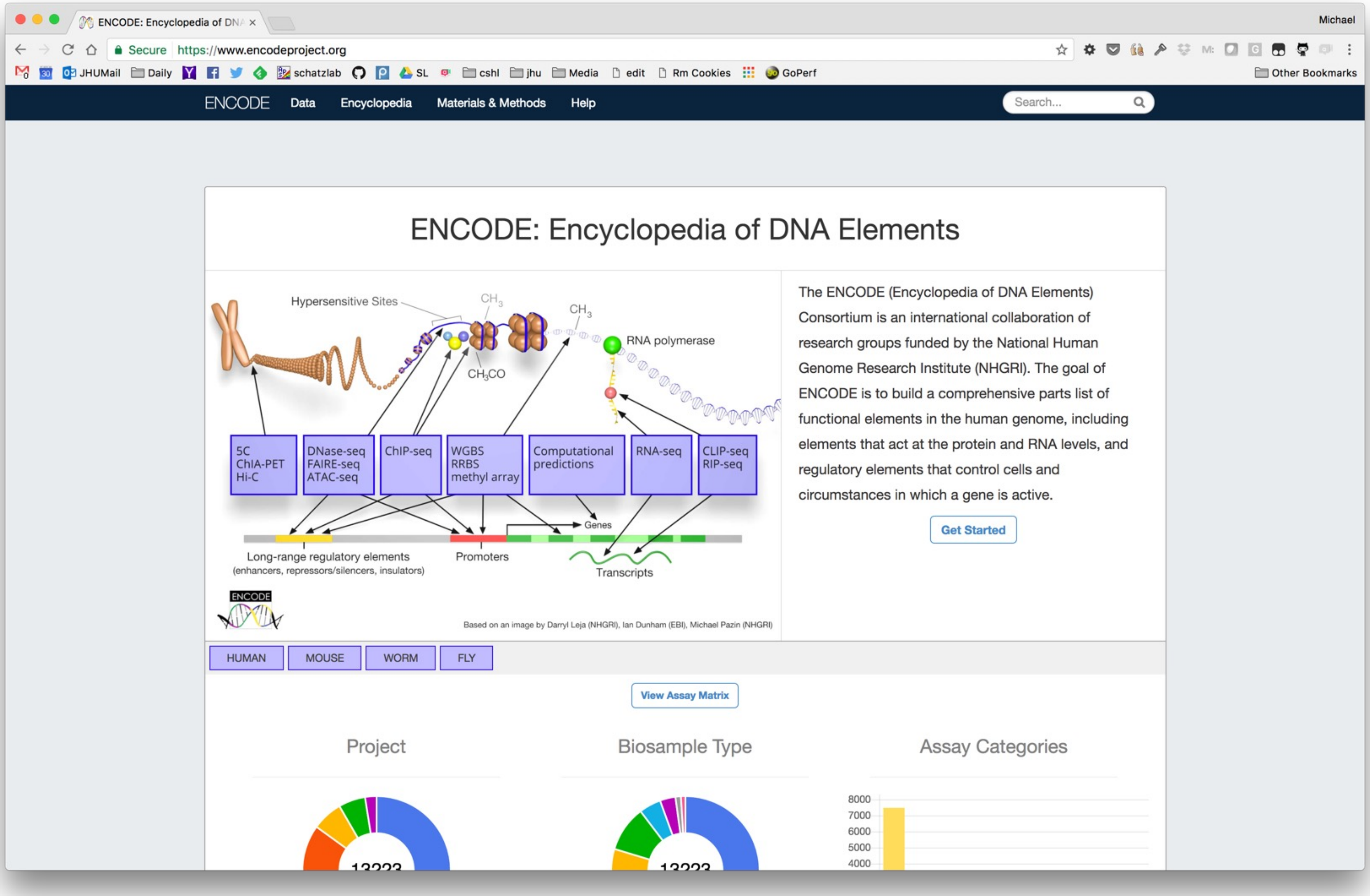


| Phenotype | SNP-pheno associations | Overlap any TF occupancy | GM12878 MEF2A | GM12878 EBF1 | GM12878 POLR2A | GM12878 NFKB1 | GM12878 IRF4 | GM12878 PAX5 | GM12878 SPI1 | GM12878 BATF | HUVEC GATA2 | GM12878 EGR1 | HeLa-S3 CEBPB | HepG2 CTCF | HUVEC all | hTH1 all | hTH2 DS7842 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total | 4,860 | 600 | 47 | 78 | 57 | 35 | 35 | 54 | 47 | 45 | 29 | 28 | 69 | 54 | 85 | 192 | 57 |
| Multiple sclerosis | 71 | 15 | 4 | 4 | 3 | 2 | 3 | 4 | 3 | 2 | 2 | 1 | 1 | 0 | 1 | 5 | 4 |
| Systemic lupus erythematosus | 62 | 10 | 4 | 4 | 6 | 4 | 3 | 1 | 1 | 4 | 1 | 2 | 2 | 1 | 2 | 4 | 2 |
| Height | 204 | 34 | 3 | 7 | 3 | 2 | 5 | 3 | 1 | 5 | 0 | 6 | 7 | 6 | 6 | 9 | 3 |
| Rheumatoid arthritis | 57 | 11 | 3 | 4 | 2 | 4 | 0 | 4 | 4 | 2 | 0 | 1 | 1 | 0 | 2 | 11 | 3 |
| Chronic lymphocytic leukaemia | 17 | 8 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 0 |
| Celiac disease | 54 | 11 | 1 | 4 | 3 | 2 | 1 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 1 |
| Ulcerative colitis | 85 | 11 | 3 | 2 | 3 | 3 | 1 | 3 | 2 | 2 | 2 | 2 | 0 | 1 | 2 | 7 | 2 |
| Crohn's disease | 105 | 20 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 5 | 1 | 2 | 2 | 6 | 9 | 5 |

**Figure 10 | Comparison of genome-wide-association-study-identified loci with ENCODE data. a,** Overlap of lead SNPs in the NHGRI GWAS SNP catalogue (June 2011) with DHSs (left) or transcription-factor-binding sites (right) as red bars compared with various control SNP sets in blue. The control SNP sets are (from left to right): SNPs on the Illumina 2.5M chip as an example of a widely used GWAS SNP typing panel; SNPs from the 1000 Genomes project; SNPs extracted from 24 personal genomes (see personal genome variants track at http://main.genome-browser.bx.psu.edu (ref. 80)), all shown as blue bars. In addition, a further control used 1,000 randomizations from the genotyping SNP panel, matching the SNPs with each NHGRI catalogue SNP for allele frequency and distance to the nearest TSS (light blue bars with bounds at 1.5 times the interquartile range). For both DHSs and transcription-factor-binding regions, a larger proportion of overlaps with GWAS-implicated SNPs is found compared to any of the controls sets. **b,** Aggregate overlap of phenotypes to selected transcription-factor-binding sites (left matrix) or DHSs in selected cell lines (right matrix), with a count of overlaps between the phenotype and the cell line/factor. Values in blue squares pass an empirical *P*-value threshold ≤0.01 (based on the same analysis of overlaps between randomly chosen, GWAS-matched SNPs and these epigenetic features) and have at least a count of three overlaps. The *P* value for the total number of phenotype–transcription factor associations is <0.001. **c,** Several SNPs associated with Crohn's disease and other inflammatory diseases that reside in a large gene desert on chromosome 5, along with some epigenetic features indicative of function. The SNP (rs11742570) strongly associated to Crohn's disease overlaps a GATA2 transcription-factor-binding signal determined in HUVECs. This region is also DNase I hypersensitive in HUVECs and T-helper $T_H1$ and $T_H2$ cells. An interactive version of this figure is available in the online version of the paper.

- 88% of GWAS SNPs are intronic or intergenic of unknown function
- We found that 12% of these GWAS-SNPs overlap transcription-factor-occupied regions whereas 34% overlap DHSs
- GWAS SNPs are particularly enriched in the segmentation classes associated with enhancers and TSSs across several cell types

# ENCODE Studies



## ENCODE: Encyclopedia of DNA Elements

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

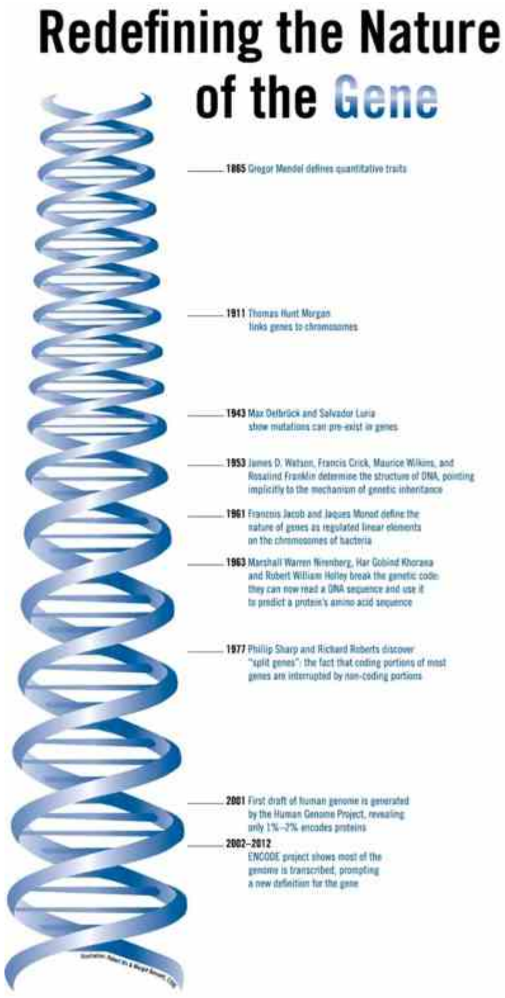
The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at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

>5000 Citations for main paper; >>10k for all papers

# Summary & Critique

- ***Summary***
  - *The unprecedented number of functional elements identified in this study provides a valuable resource to the scientific community as well as significantly enhances our understanding of the human genome.*

- ***Critique***
  - Was it correct?
  - What is functional?
  - What is conservation?
  - What was the control?
  - What are the tradeoffs of organizing so much funding ($288M!) around a single project; will other groups successfully use these data?


Redefining the Nature of the Gene
1865 Gregor Mendel defines quantitative traits
1911 Thomas Hunt Morgan links genes to chromosomes
1943 Max Delbrück and Salvador Luria show mutations can pre-exist in genes
1953 James D. Watson, Francis Crick, Maurice Wilkins, and Rosalind Franklin determine the structure of DNA, pointing implicitly to the mechanism of genetic inheritance
1961 Francois Jacob and Jaques Monod define the nature of genes as regulated linear elements on the chromosomes of bacteria
1963 Marshall Warren Nirenberg, Har Gobind Khorana and Robert William Holley break the genetic code: they can now read a DNA sequence and use it to predict a protein's amino acid sequence
1977 Phillip Sharp and Richard Roberts discover "split genes": the fact that coding portions of most genes are interrupted by non-coding portions
2001 First draft of human genome is generated by the Human Genome Project, revealing only 1%–2% encodes proteins
2002–2012 ENCODE project shows most of the genome is transcribed, prompting a new definition for the gene

# Exam Topics

**Genomics**

- Genomics Technologies
  - Illumina, PacBio, Nanopore
- Kmer profiling
- Genome Assembly
- Whole Genome Alignment
- Read mapping
- Variant Identification
- Gene Finding
- BLAST
- RNA-seq
- Methyl-seq, Chip-Seq, Hi-C
- Genome Annotation

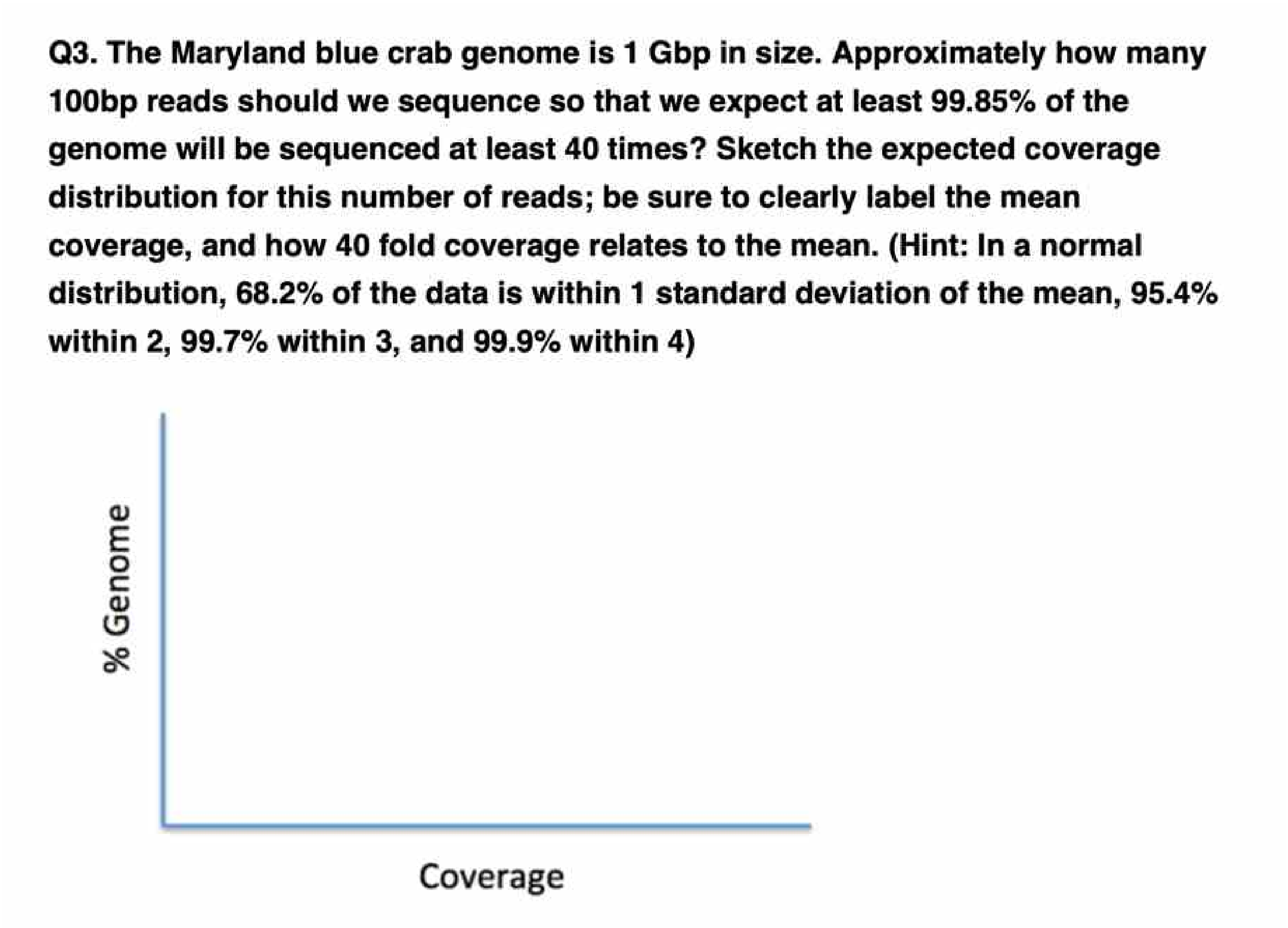
**Quantitative Techniques**

- Normal, Poisson, Binomial, P-value
- de Bruijn and overlap graphs
- Minimizers
- Dot plots
- Quality Values (Phred Scale)
- Full text indexing & BWT
- Seed & Extend
- Hidden Markov Models
- PCA / t-SNE / UMAP
- Convolutional Neural Networks
- Differential Expression
- Expectation Maximization

**What is the goal? What is the approach? What are the key challenges?**

**How did we explore these topics in the homeworks and lectures?**

# Sample Question



Q3. The Maryland blue crab genome is 1 Gbp in size. Approximately how many 100bp reads should we sequence so that we expect at least 99.85% of the genome will be sequenced at least 40 times? Sketch the expected coverage distribution for this number of reads; be sure to clearly label the mean coverage, and how 40 fold coverage relates to the mean. (Hint: In a normal distribution, 68.2% of the data is within 1 standard deviation of the mean, 95.4% within 2, 99.7% within 3, and 99.9% within 4)