umcg

Department of genetics

Lude Franke > ENCODE project

Department of Genetics, UMC Groningen

Nature, 6 september 2012

ARTICLE

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

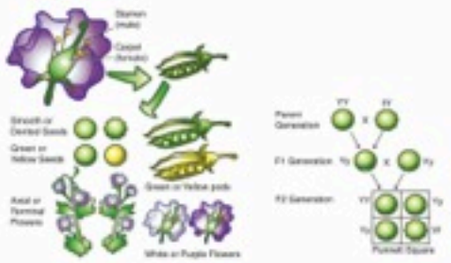
The ENCODE Project Consortium*

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 project

GENOME RESEARCH

Gregor Mendel



The **laws of inheritance** were described.

1865

1869

The **nucleic acids** were isolated and studied by Friedrich Miescher.

1900

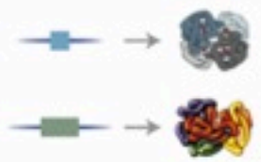
The rediscovery of Mendel's work by Carl Correns, Erich von Tschermak-Seysenegg, and Hugo De Vries prompted the foundation of **genetics**.

Thomas Morgan

1910

Studies in *Drosophila melanogaster* suggest a **linear model of genes** on chromosomes, like 'beads on a string.'

1927



One gene, one enzyme; Then **one gene, one protein**.

Artificial **transmutation of the gene** by X-ray was reported by Hermann Müller.

1941

Francis Crick Jame

The DNase experiment by Avery, MacLeod, and McCarty suggested **transformation is induced by DNA**.

1944

Gene as a discrete heredity unit *Gene as a distinct locus* *Gene as a physical molecule* *Gene as a protein b*

ZHANG MMVII

A term invented almost a century ago, 'gene,' with its beguilingly simple orthography, has become a central concept in biology. Given a specific meaning at its coinage, this word has evolved into something complex and elusive over the years, reflecting our ever-expanding knowledge in genetics and in life sciences at large. The stunning discoveries made in the ENCODE Project—like many before that significantly enriched the meaning of this term—are harbingers of another tide of change in our understanding of what a gene is.

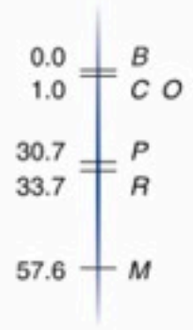
1909

The first appearance of the word '**gene**,' derived from the Greek *genesis* or *genos*.

Wilhelm Johannsen

1913

Alfred Sturtevant constructed the **first genetic map**.



1928

Griffith's experiment demonstrating type-switching in *pneumococcus* suggested a **transforming principle**.

Hershey and Chase determined that **DNA is the genetic material**.

Alfred Hershey Martha Chase

The **op** by Fra Jacques strated

Pro

ENCODE project

GENOME RESEARCH

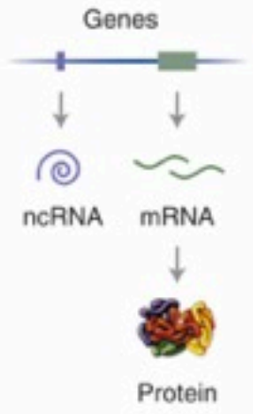
s Crick James Watson

experiment
cLeod, and
suggested
rmation is
d by DNA.

1944

The **DNA double helix** structure was solved.

1953



The '**Central Dogma**' of molecular biology was proposed by Francis Crick.

1958

The **first sequence** of a gene, *COAT_BPMS2*, was determined.

1972

The first large-scale **gene function** analysis using gene expression in yeast

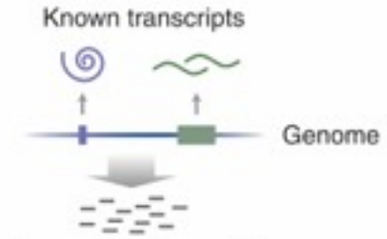
1994

GENSCAN, a computer program for **gene structure prediction**, became available.

1997

The drafts of the **human genome sequence** were published.

2001



The ENCODE Project highlighted the **complexity** of gene transcription and regulation.

2007

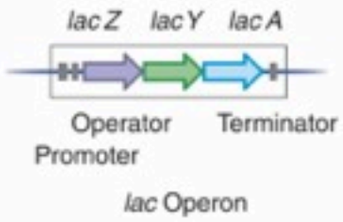
Gene as a physical molecule
Gene as a protein blueprint
Gene as transcribed code
Gene as ORF sequence pattern
Gene as annotated genomic entity
Gene as ...

1952

and
ined
the
rial.

tha Chase

The **operon**, described by François Jacob and Jacques Monod, demonstrated **transcriptional control**.



lac Operon

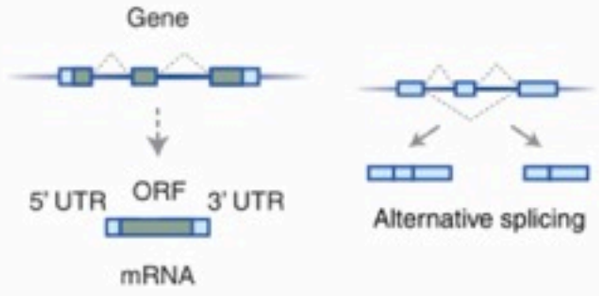
1961

1965

The **genetic code** was deciphered by Marshall Nirenberg, Har Gobind Khorana, and others.

1977

Introns and the mechanism of **RNA splicing** were discovered by Phillip Sharp and Richard Roberts demonstrating 'split gene structure.'

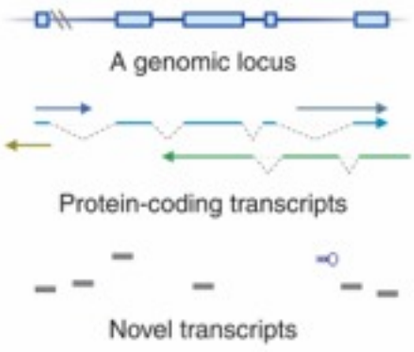


2003

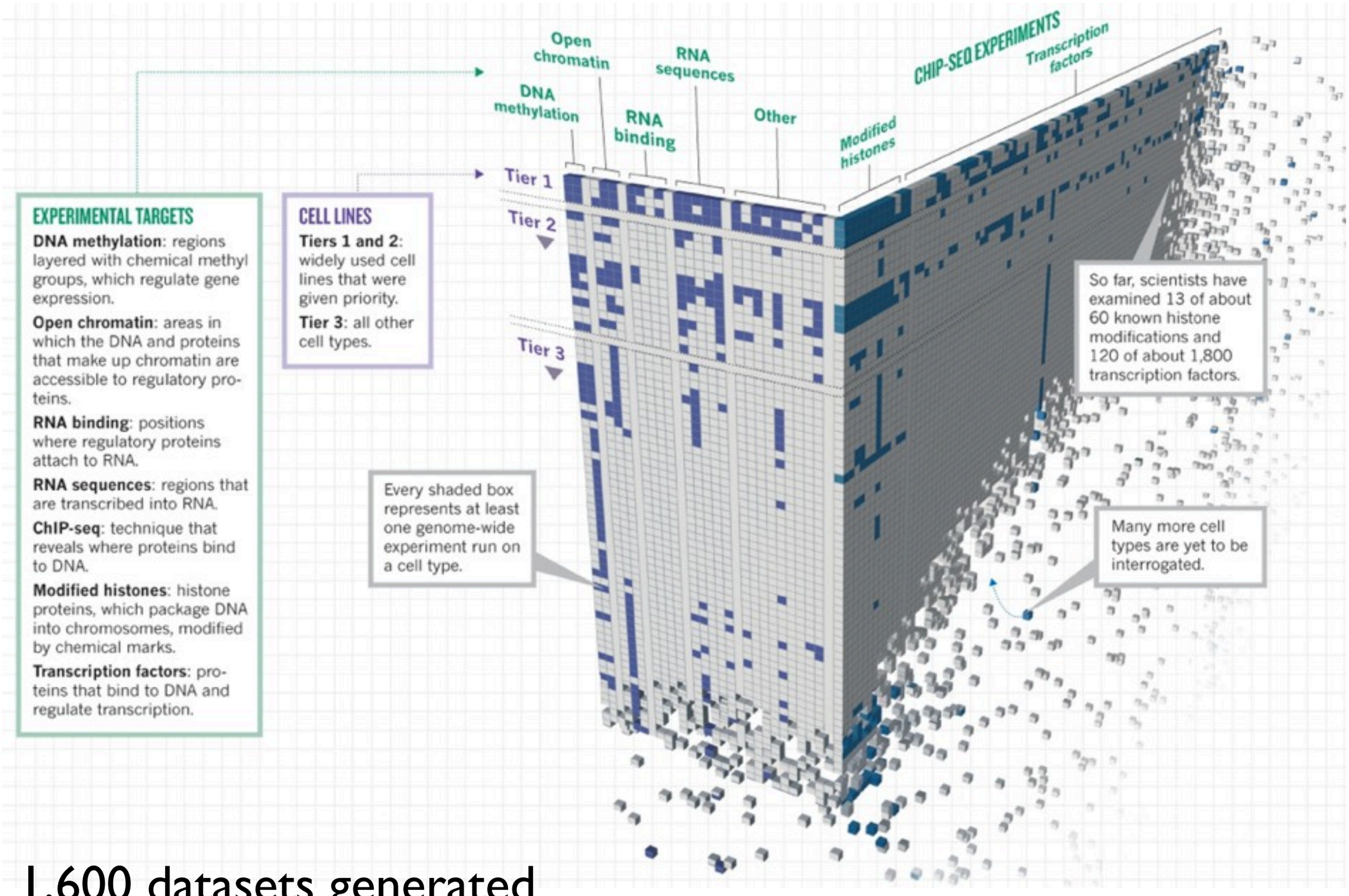
The **ENCODE Project** was launched.

2007

The pilot phase of the ENCODE Project was finished. New gene models are proposed.

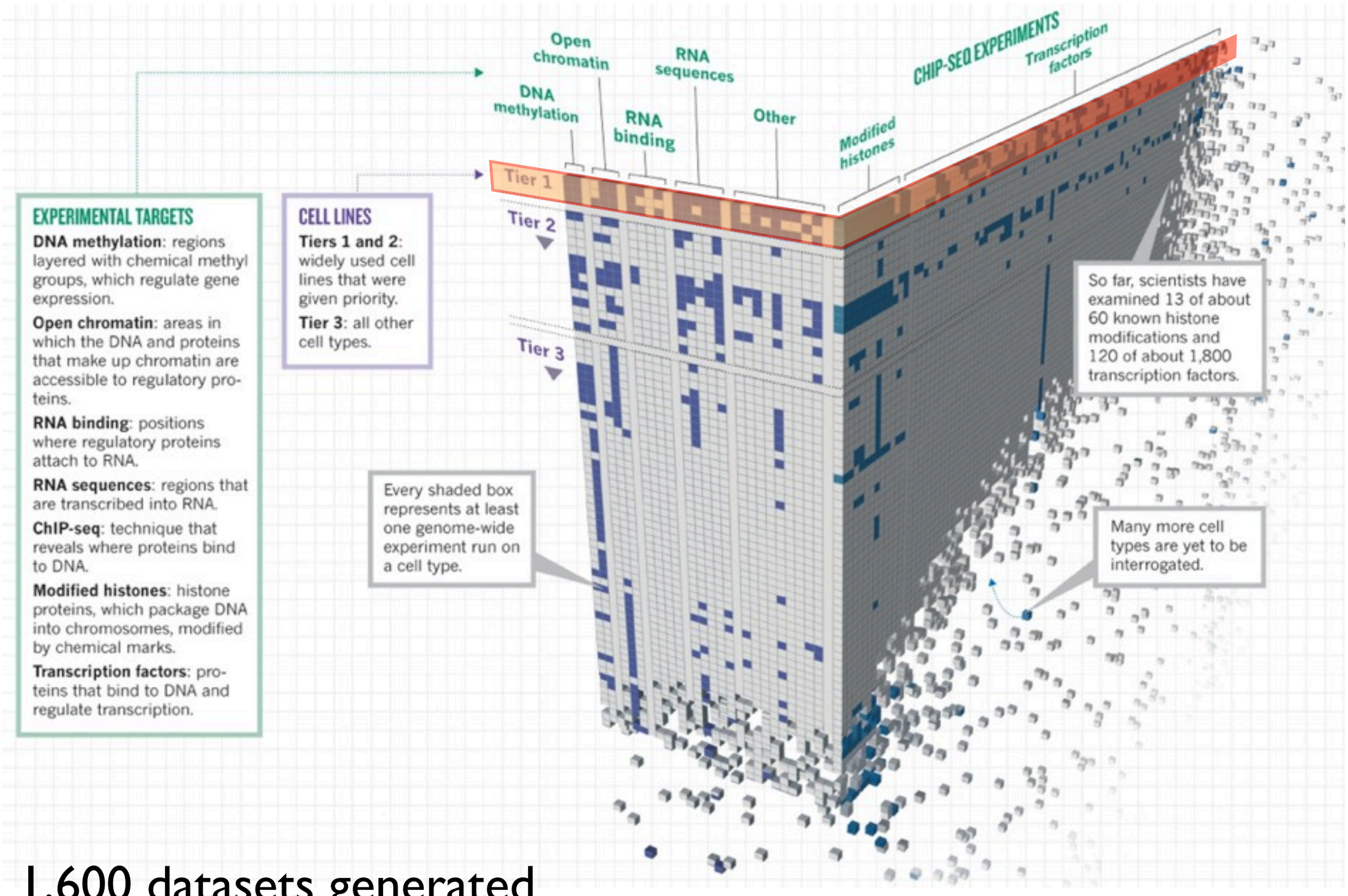


What is ENCODE?



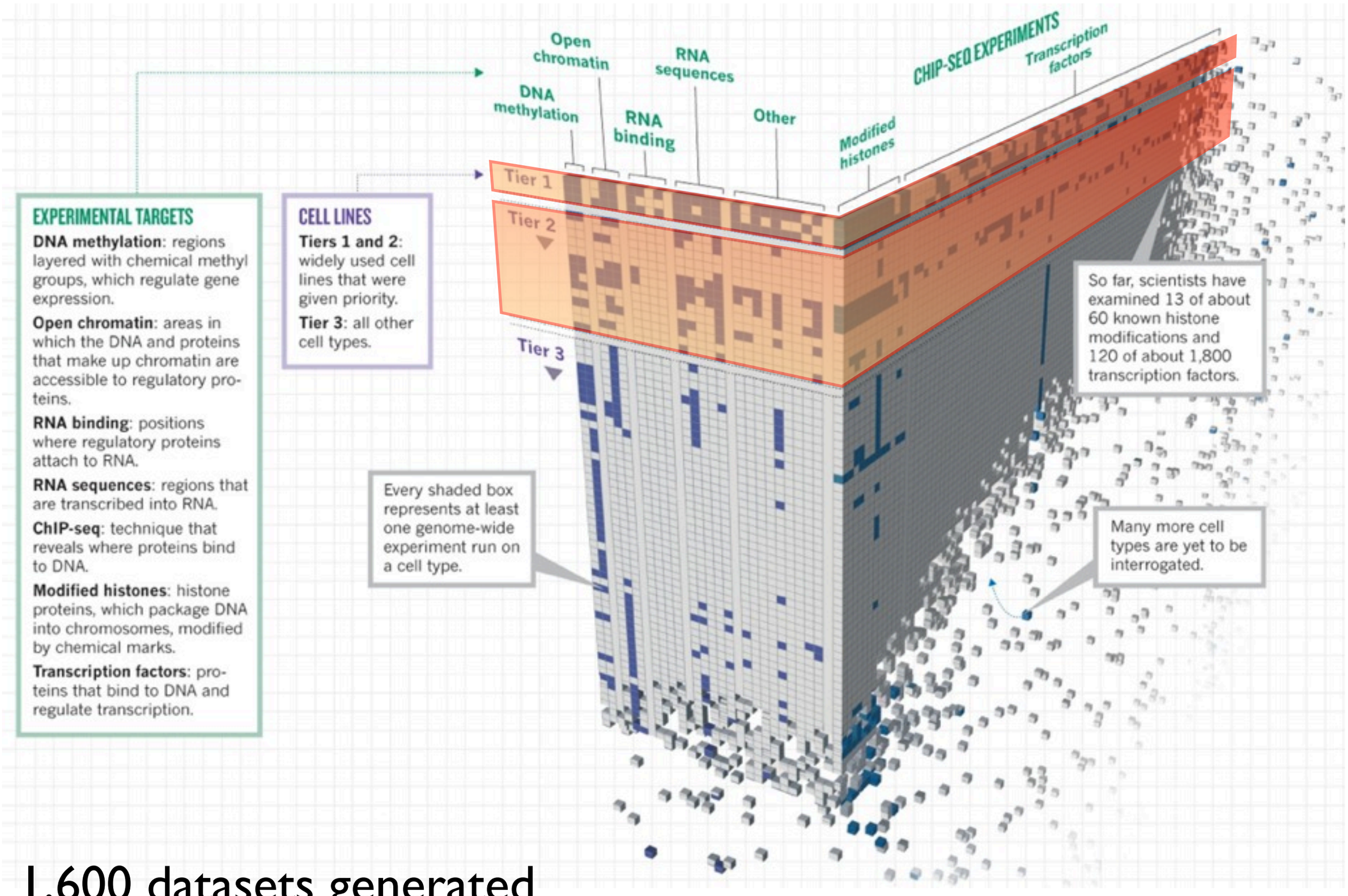
1,600 datasets generated

What is ENCODE?



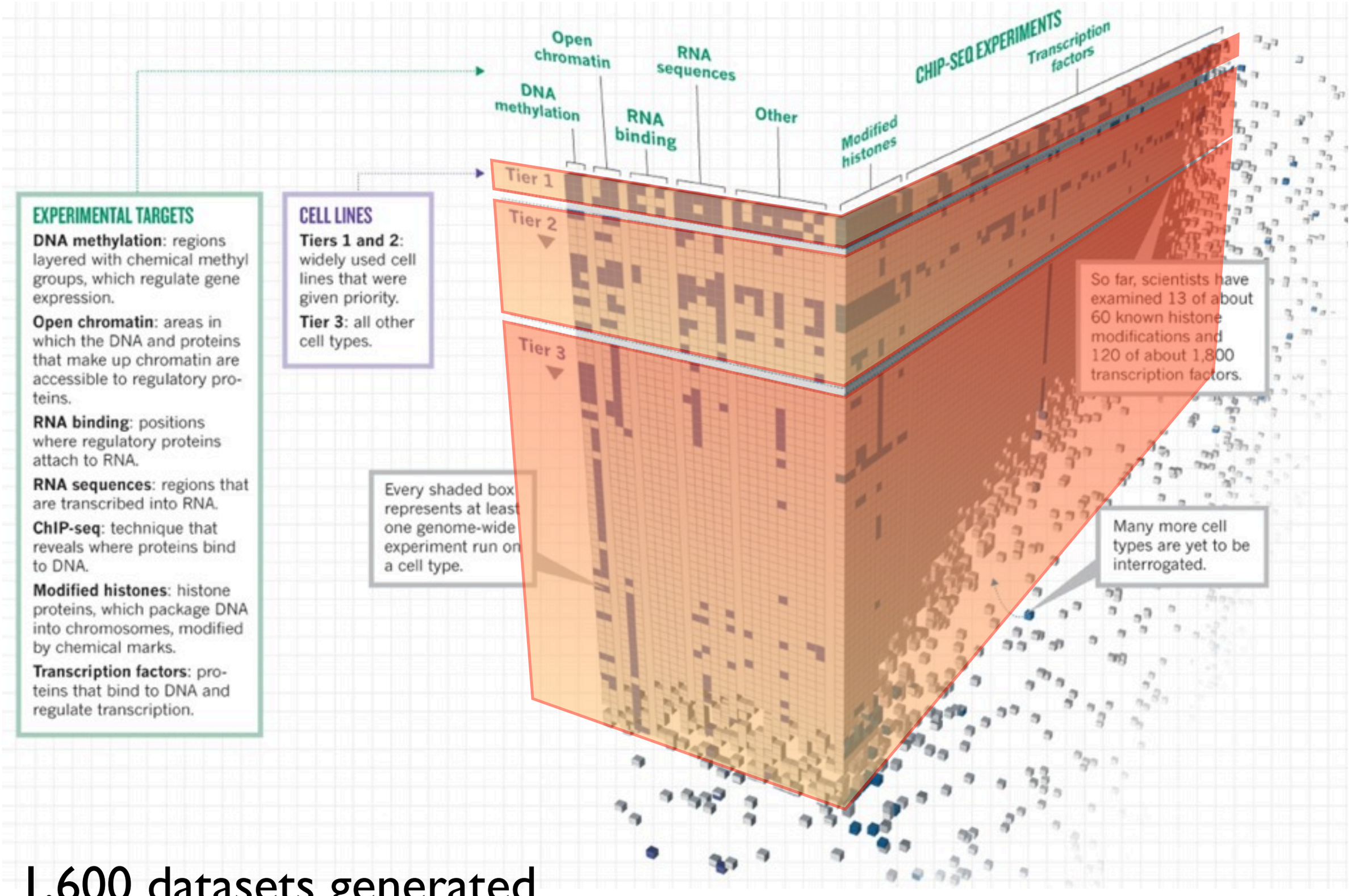
1,600 datasets generated

What is ENCODE?



1,600 datasets generated

What is ENCODE?



1,600 datasets generated

Types of experiments

RNA expression

RNA-seq. Isolation of RNA sequences, often with different purification techniques to isolate different fractions of RNA followed by high-throughput sequencing.

CAGE. Capture of the methylated cap at the 5′ end of RNA, followed by high-throughput sequencing of a small tag adjacent to the 5′ methylated caps. 5′ methylated caps are formed at the initiation of transcription, although other mechanisms also methylate 5′ ends of RNA.

RNA-PET. Simultaneous capture of RNAs with both a 5′ methyl cap and a poly(A) tail, which is indicative of a full-length RNA. This is then followed by sequencing a short tag from each end by high-throughput sequencing.

Protein binding to DNA

ChIP-seq. Chromatin immunoprecipitation followed by sequencing. Specific regions of crosslinked chromatin, which is genomic DNA in complex with its bound proteins, are selected by using an antibody to a specific epitope. The enriched sample is then subjected to high-throughput sequencing to determine the regions in the genome most often bound by the protein to which the antibody was directed. Most often used are antibodies to any chromatin-associated epitope, including transcription factors, chromatin binding proteins and specific chemical modifications on histone proteins.

Open chromatin

DNase-seq. Adaption of established regulatory sequence assay to modern techniques. The DNase I enzyme will preferentially cut live chromatin preparations at sites where nearby there are specific (non-histone) proteins. The resulting cut points are then sequenced using high-throughput sequencing to determine those sites 'hypersensitive' to DNase I, corresponding to open chromatin.

Nucleosome Regulatory factor binding

FAIRE-seq. Formaldehyde assisted isolation of regulatory elements. FAIRE isolates nucleosome-depleted genomic regions by exploiting the difference in crosslinking efficiency between nucleosomes (high) and sequence-specific regulatory factors (low). FAIRE consists of crosslinking, phenol extraction, and sequencing the DNA fragments in the aqueous phase.

Methy-lation

RRBS. Reduced representation bisulphite sequencing. Bisulphite treatment of DNA sequence converts unmethylated cytosines to uracil. To focus the assay and save costs, specific restriction enzymes that cut around CpG dinucleotides can reduce the genome to a portion specifically enriched in CpGs. This enriched sample is then sequenced to determine the methylation status of individual cytosines quantitatively.

Which cell-types have been investigated most?

Tier 1. Tier 1 cell types were the highest-priority set and comprised three widely studied cell lines: K562 erythroleukaemia cells; GM12878, a B-lymphoblastoid cell line that is also part of the 1000 Genomes project (http://1000genomes.org)[55]; and the H1 embryonic stem cell (H1 hESC) line.

Tier 2. The second-priority set of cell types in the ENCODE project which included HeLa-S3 cervical carcinoma cells, HepG2 hepatoblastoma cells and primary (non-transformed) human umbilical vein endothelial cells (HUVECs).

Tier 3. Any other ENCODE cell types not in tier 1 or tier 2.

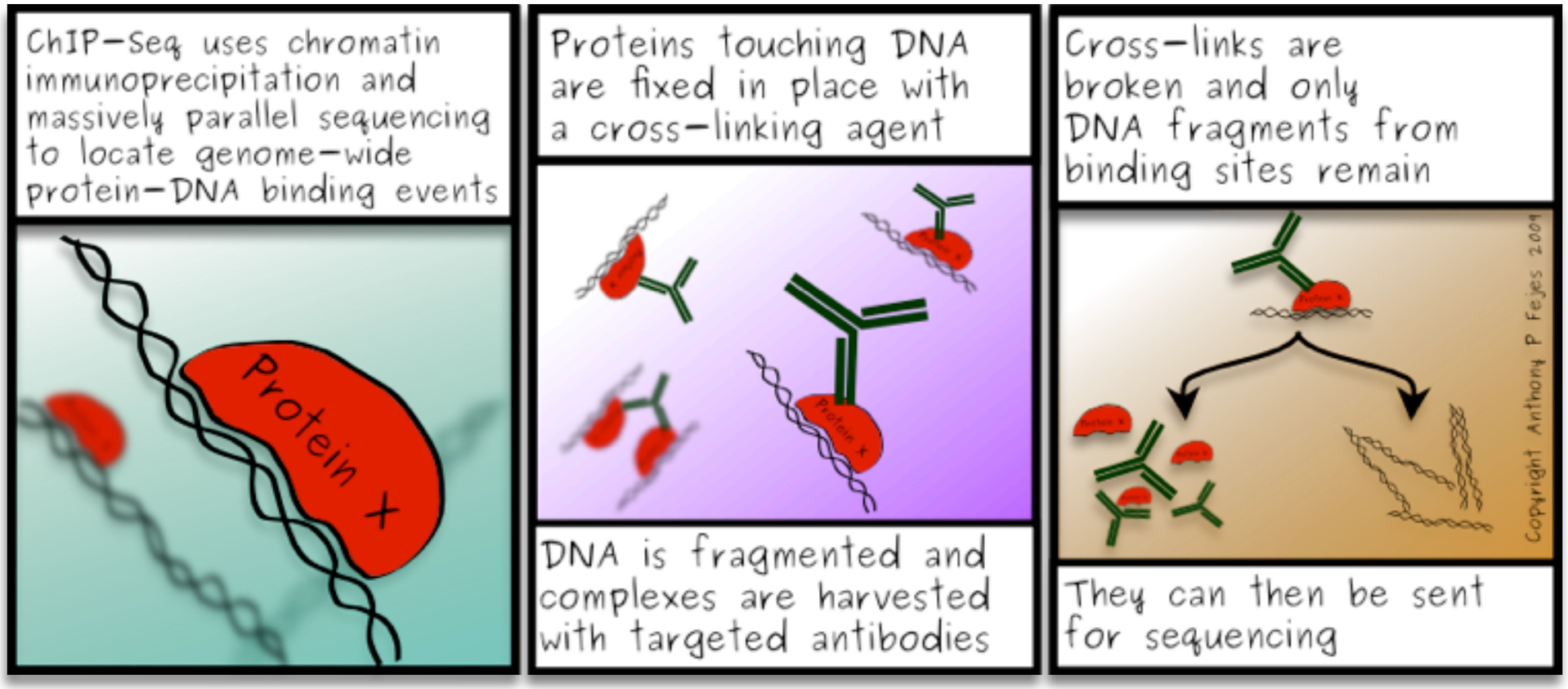
RNA analyses:

RNA-seq:
- 62% of genomic bases are expressed (long reads > 200 bp)
- Only 5.5% of these map within known exons

CAGE-seq (5' cap-targeted RNA isolation and sequencing)
- 62,043 transcription start sites (TSSs)
- 44% of these map within 5' of known transcript (thus 56% not!)

Many reads shorter than 200 bp observed:
- Reflect transferRNA, microRNA, small nuclear RNA, small nucleolar RNA

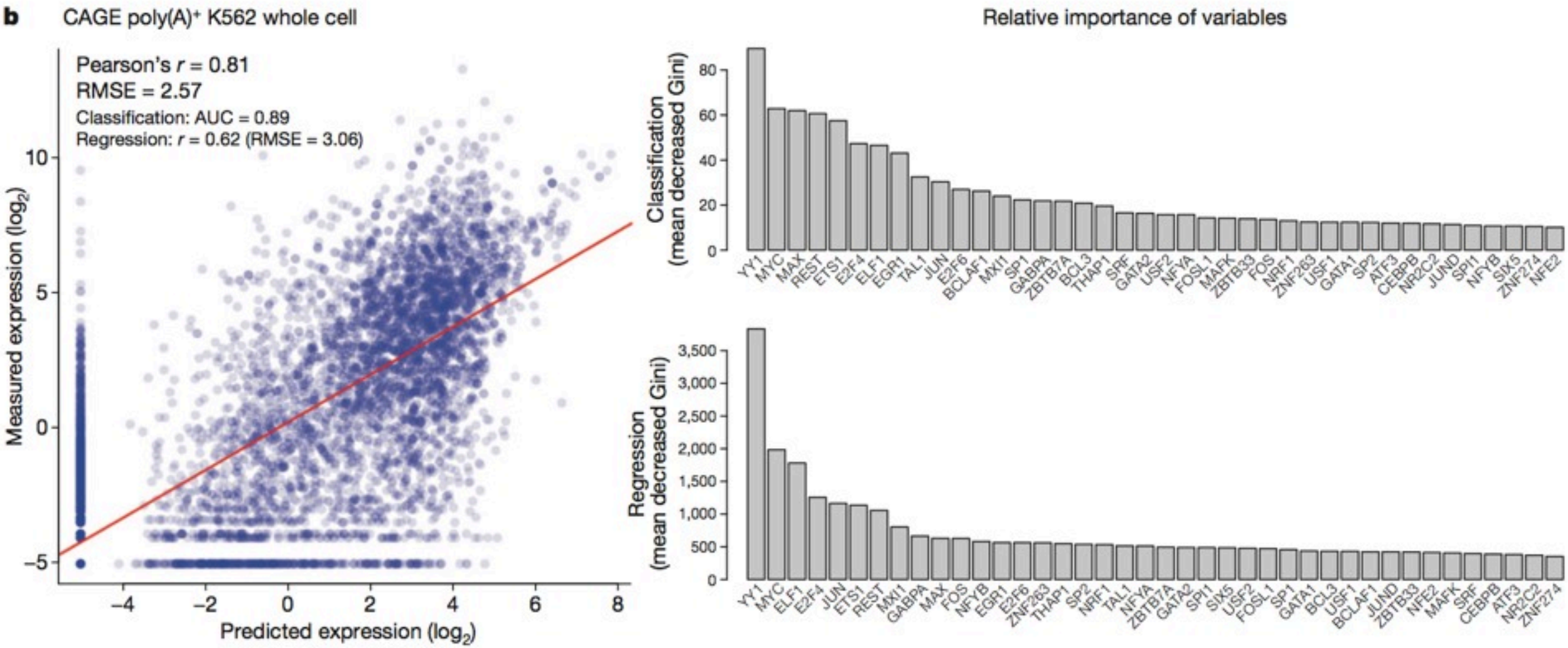
Protein binding to DNA: ChIP-seq



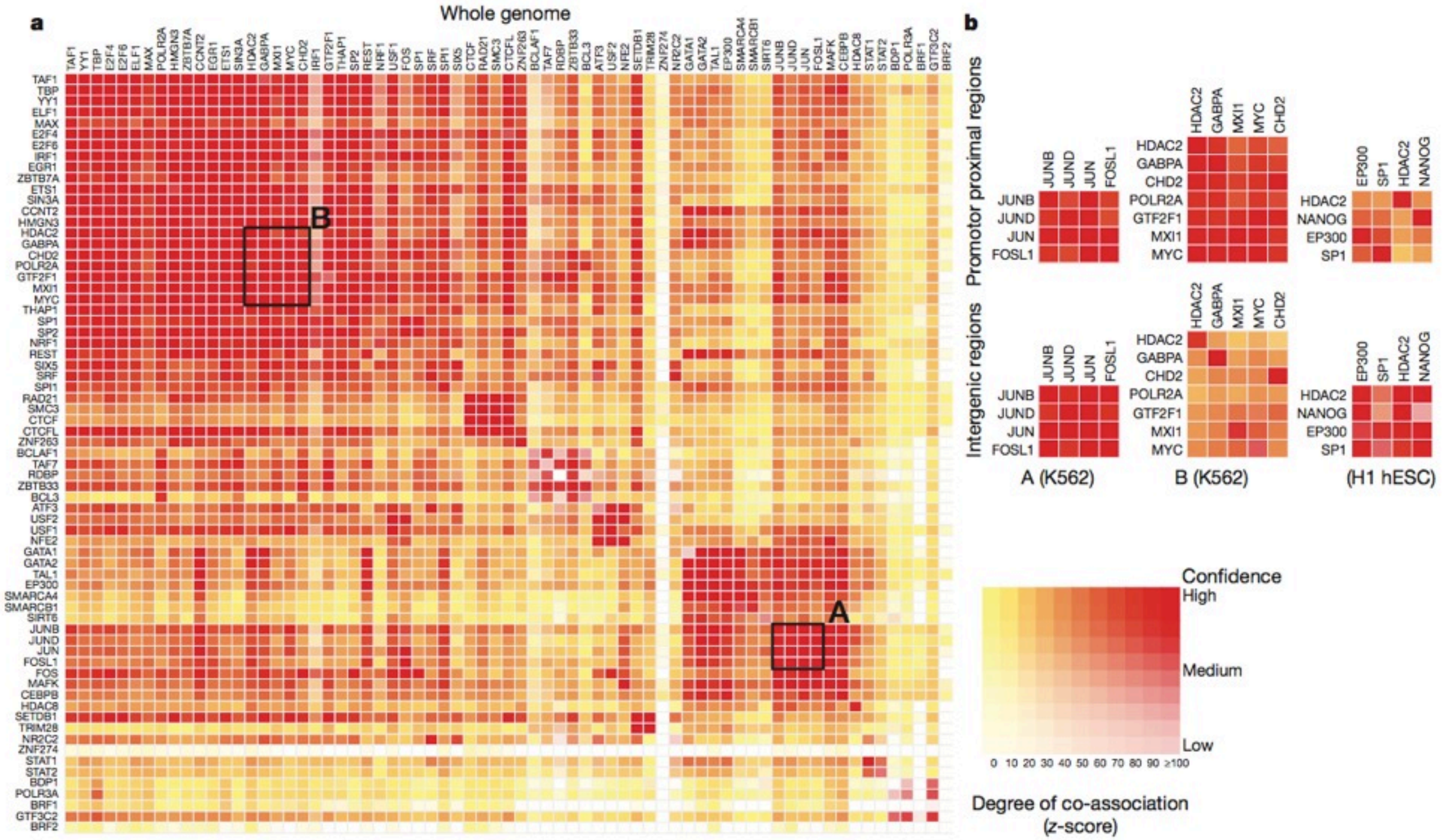
Which parts of the DNA are physically interacting with proteins?

- 199 different DNA-binding proteins assayed in 72 cell-types using ChIP-seq
- 636,336 binding regions identified, covering 231 megabases (8.1% of genome)
- 86% of these regions contain strong DNA-binding motif (in most cases the expected motif is enriched)
- All this information is available at public resource FactorBook (www.factorbook.org)

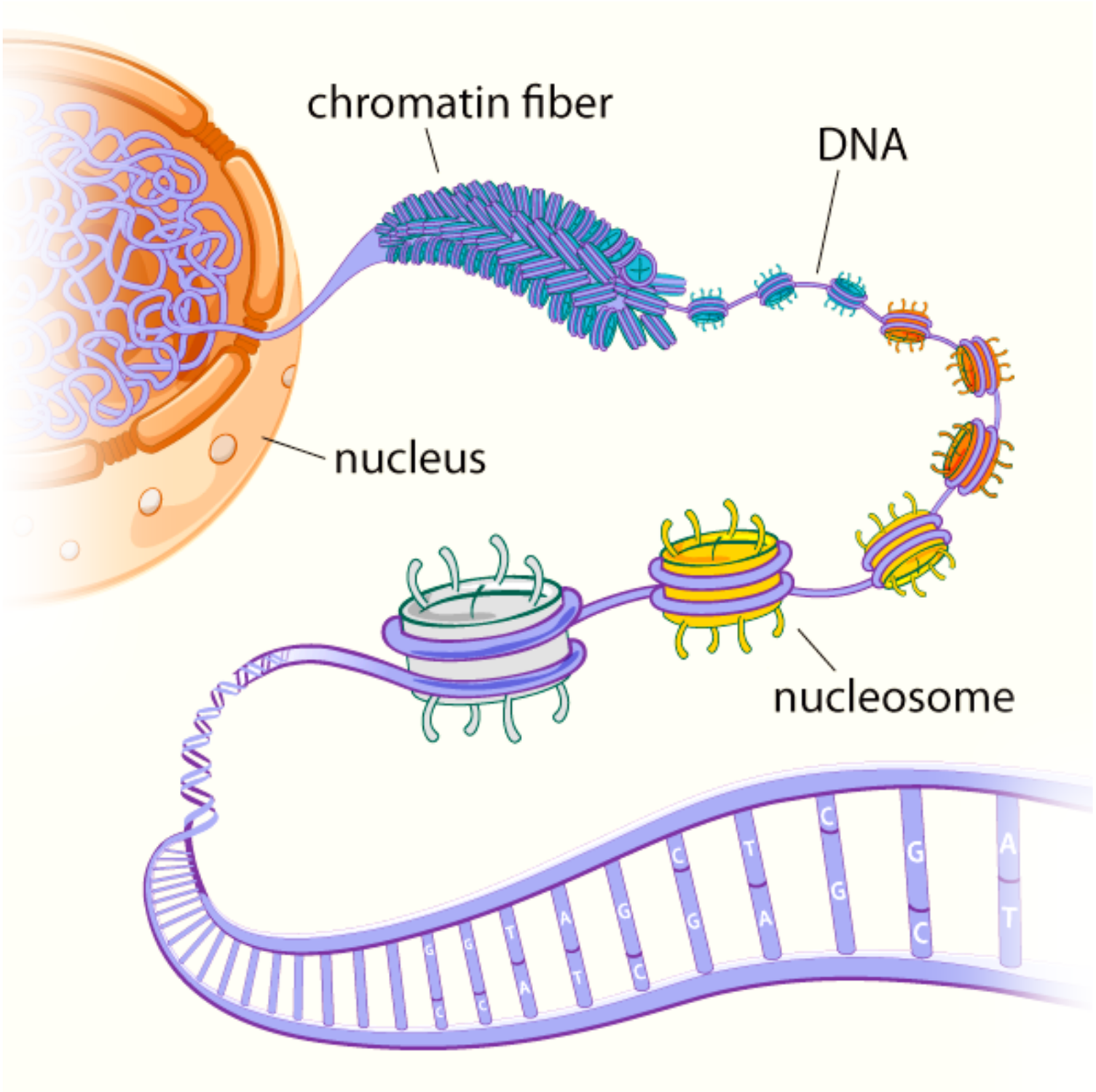
Predict expression levels based on bound TFs



Co-association between transcription factors

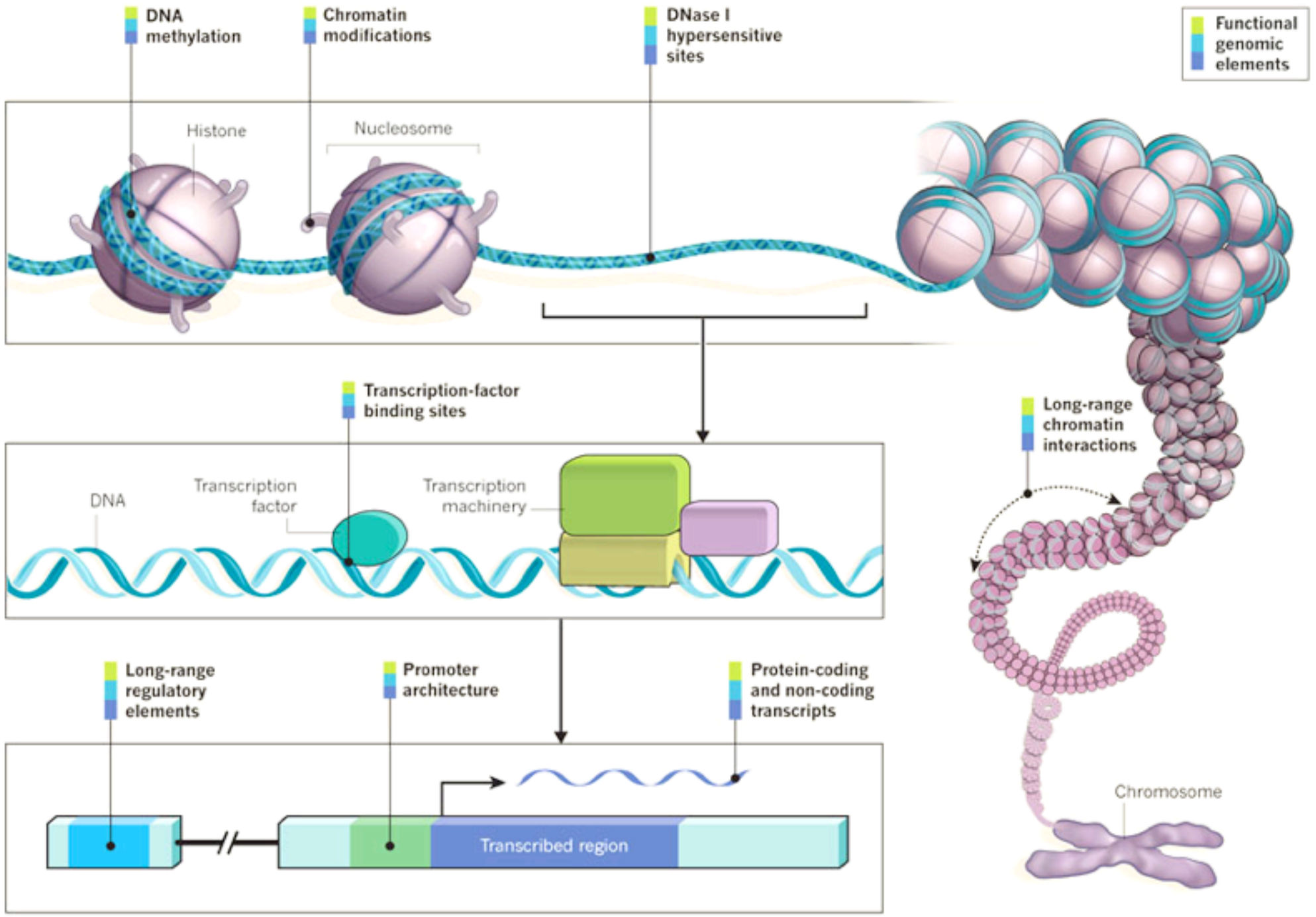


Epigenetics

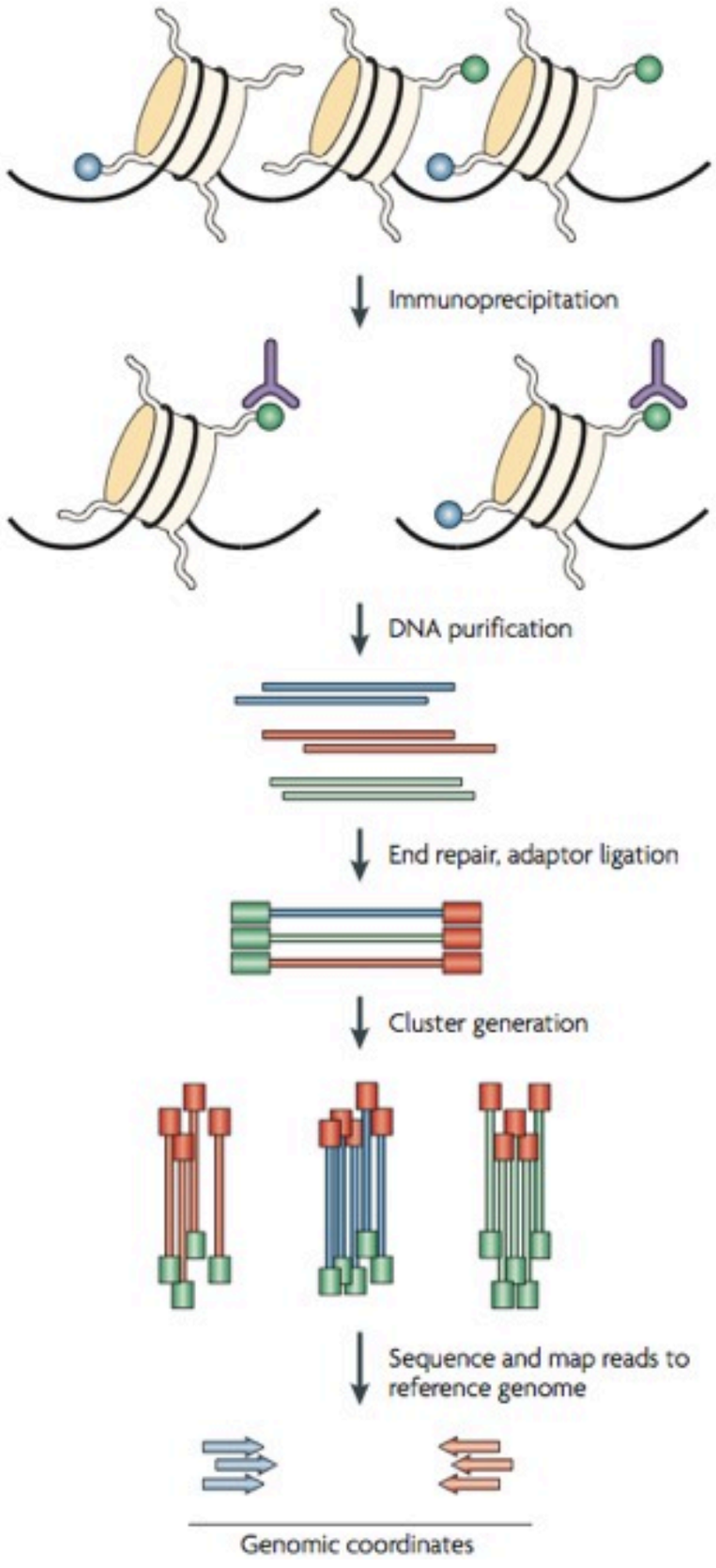


DNase I hypersensitive sites and footprints

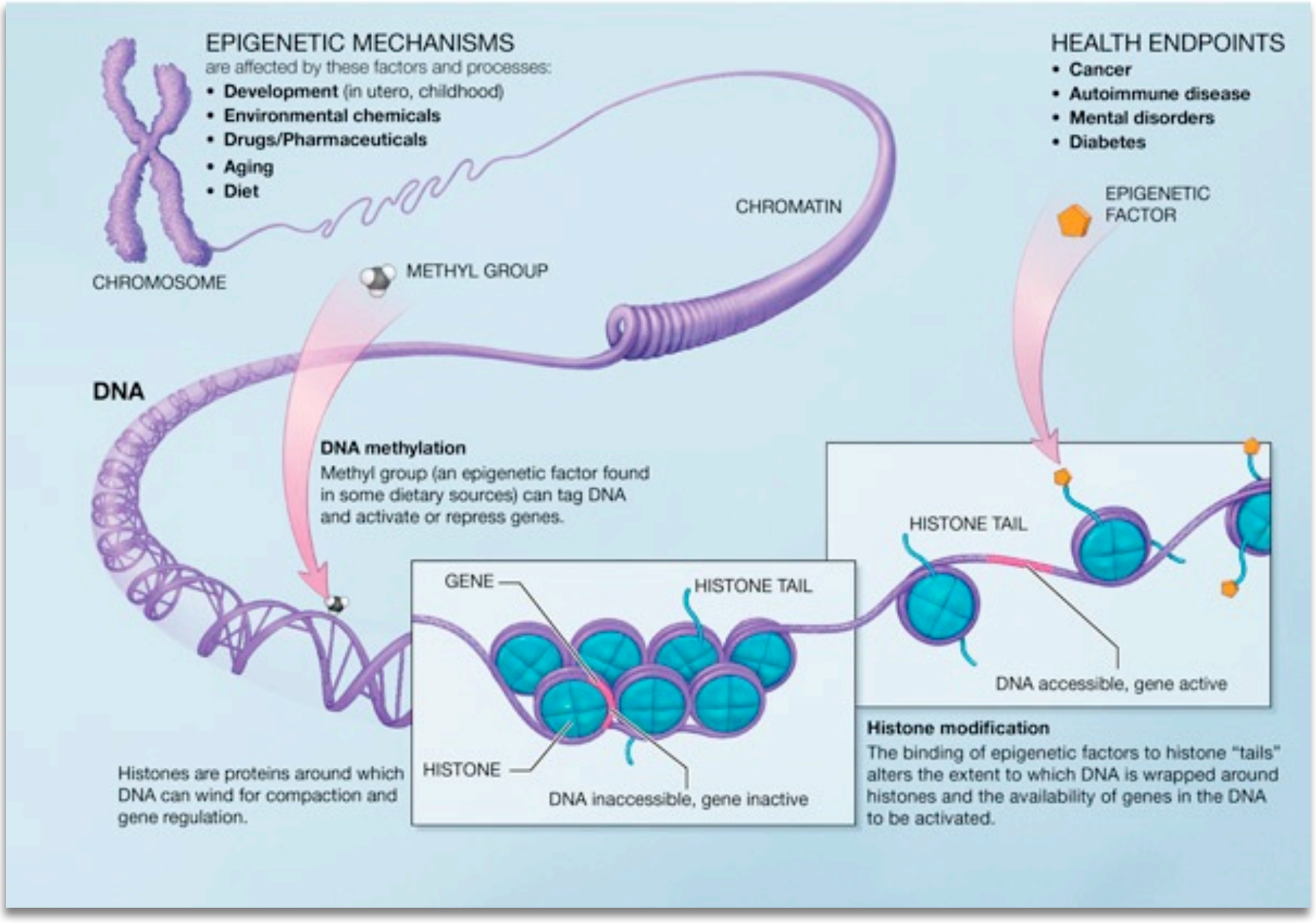
- Accessible chromatin (identified by DNase I hypersensitivity): hallmark of regulatory DNA regions
- 2.89 million non-overlapping hypersensitive sites found in 125 cell types.
- Far majority map distal to transcription start sites (TSSs)
- 98.5% of the transcription factor binding lie within accessible chromatin



Histone modifications



Histone modifications



Histone modifications

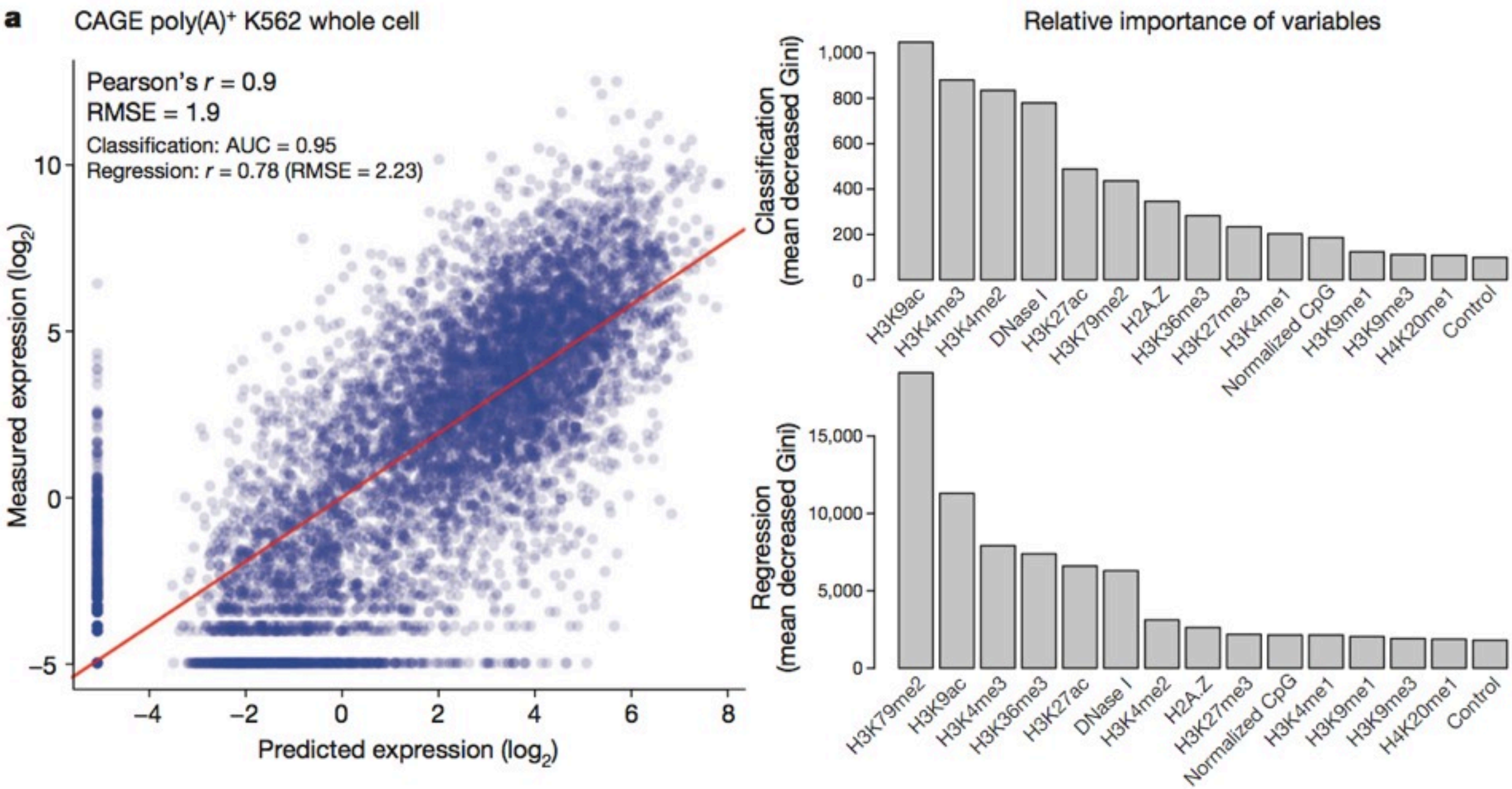
- 12 histone modifications assayed

Table 2 | Summary of ENCODE histone modifications and variants

Histone modification or variant	Signal characteristics	Putative functions
H2A.Z	Peak	Histone protein variant (H2A.Z) associated with regulatory elements with dynamic chromatin
H3K4me1	Peak/region	Mark of regulatory elements associated with enhancers and other distal elements, but also enriched downstream of transcription starts
H3K4me2	Peak	Mark of regulatory elements associated with promoters and enhancers
H3K4me3	Peak	Mark of regulatory elements primarily associated with promoters/transcription starts
H3K9ac	Peak	Mark of active regulatory elements with preference for promoters
H3K9me1	Region	Preference for the 5′ end of genes
H3K9me3	Peak/region	Repressive mark associated with constitutive heterochromatin and repetitive elements
H3K27ac	Peak	Mark of active regulatory elements; may distinguish active enhancers and promoters from their inactive counterparts
H3K27me3	Region	Repressive mark established by polycomb complex activity associated with repressive domains and silent developmental genes
H3K36me3	Region	Elongation mark associated with transcribed portions of genes, with preference for 3′ regions after intron 1
H3K79me2	Region	Transcription-associated mark, with preference for 5′ end of genes
H4K20me1	Region	Preference for 5′ end of genes

- Global patterns of histone modifications are highly variable across cell-types
- Can be used to predict functional attributes!

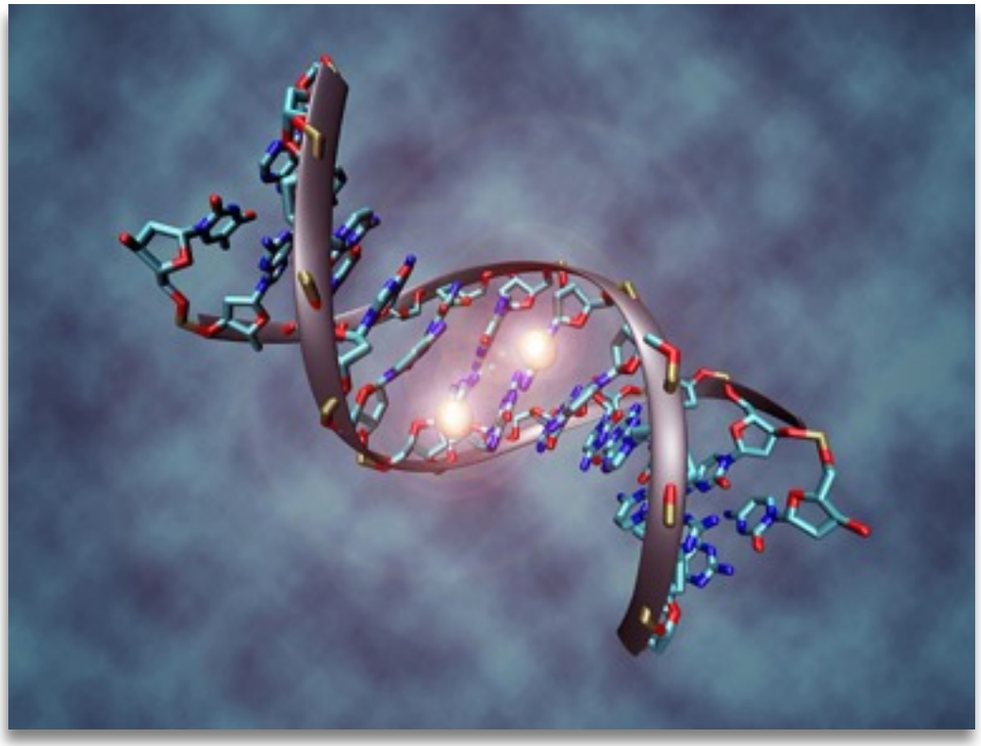
Histone modifications can predict expression levels



DNA methylation

DNA can be methylated: Addition of methyl group

- Bisulphite sequencing (RRBS)
- 82 cell-lines and tissues assayed
- 96% of CpGs show differential methylation in at least one cell type or tissue
- Most variable methylated CpGs found in gene bodies and intergenic regions



Conclusions main ENCODE paper

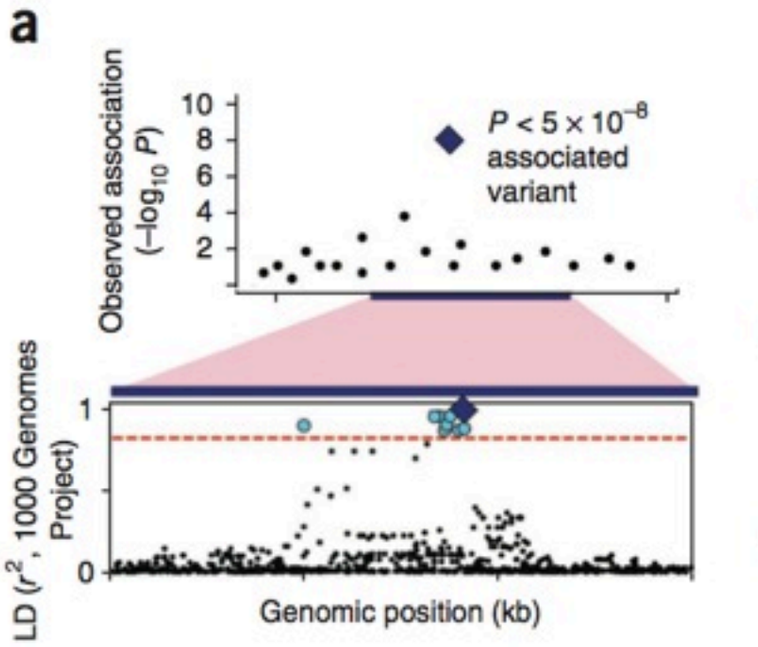
Many different analyses conducted, resulting in 1,640 datasets:

- Vast majority (80.4%) of genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type
- 95% of genome lies within 8 kb of a DNA-protein interaction (ChIP-seq, DNase I)
- 99% of genome lies within 1.7kb of at least one measured biochemical event
- Primate-specific DNA elements show evidence of negative selection: thus some are functional
- Seven chromatin states ascertained: 400,000 regions show enhancer-like features, 70,000 regions show promotor-like features
- It is possible to predict RNA expression levels by using chromatin state information
- Many non-coding genetic variants (SNPs) lie in functional regions: They are expected to have a functional consequence!
- SNPs associated with disease are enriched within non-coding functional regions, often the disease phenotypes can be associated with a specific cell type or transcription factor

Chromatin marks identify critical cell types for fine mapping complex trait variants

nature genetics

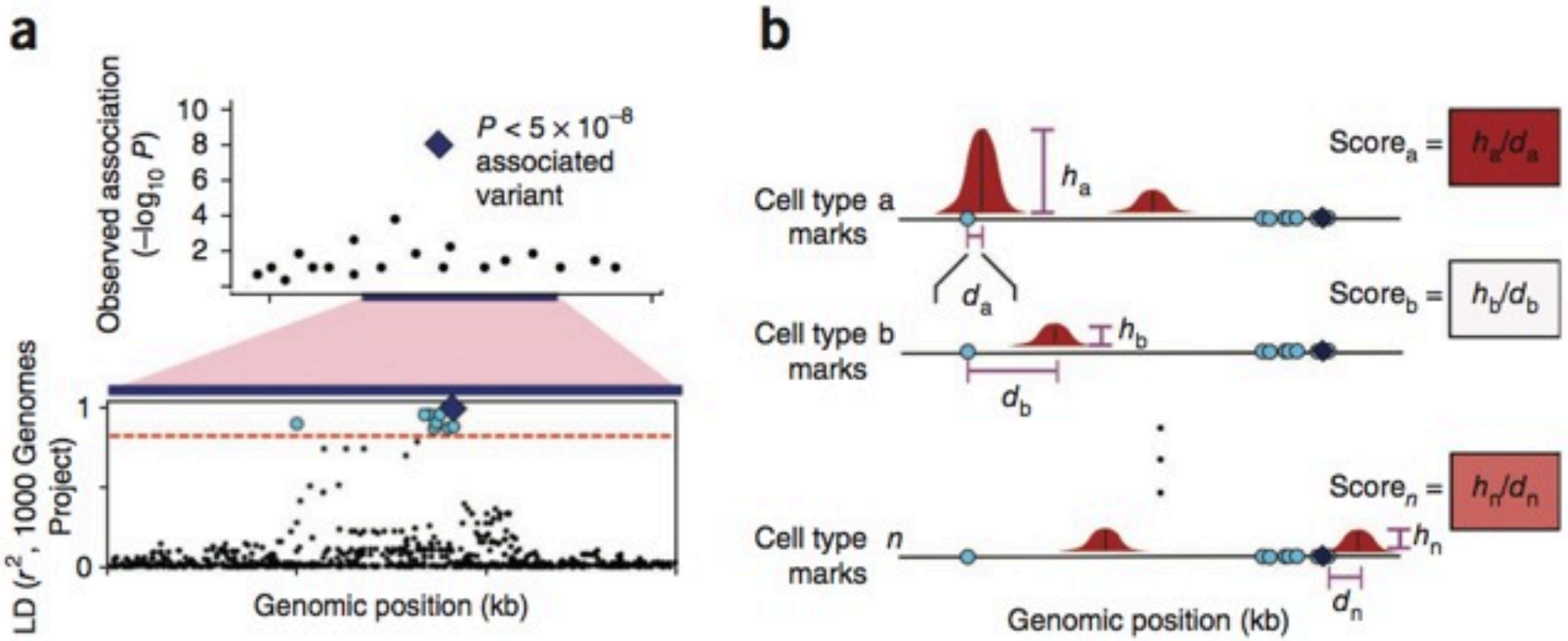
Gosia Trynka[1–4,8], Cynthia Sandor[1–4,8], Buhm Han[1–4], Han Xu[5], Barbara E Stranger[1,4,7], X Shirley Liu[5] & Soumya Raychaudhuri[1–4,6]



Chromatin marks identify critical cell types for fine mapping complex trait variants

nature genetics

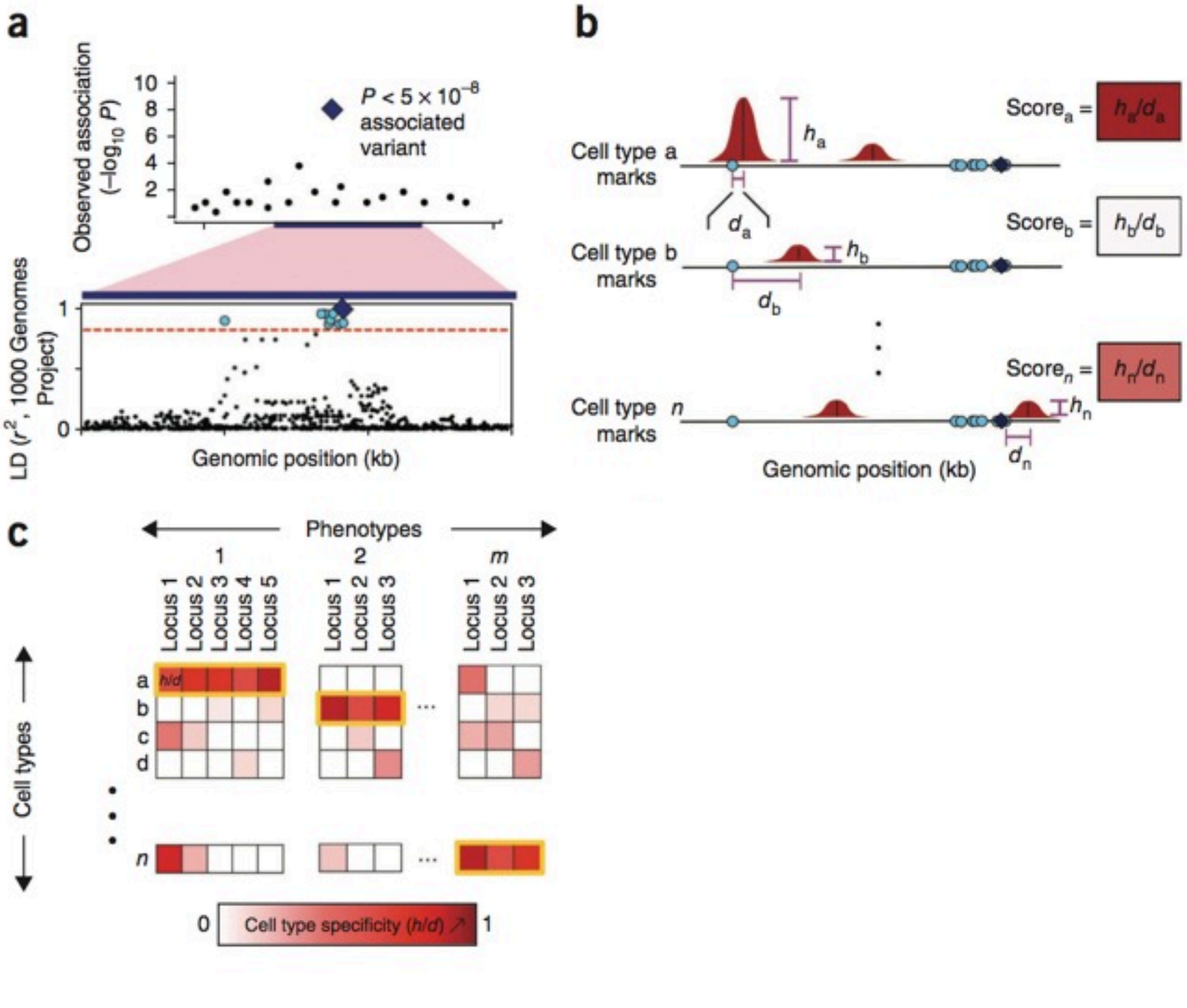
Gosia Trynka[1–4,8], Cynthia Sandor[1–4,8], Buhm Han[1–4], Han Xu[5], Barbara E Stranger[1,4,7], X Shirley Liu[5] & Soumya Raychaudhuri[1–4,6]



Chromatin marks identify critical cell types for fine mapping complex trait variants

nature genetics

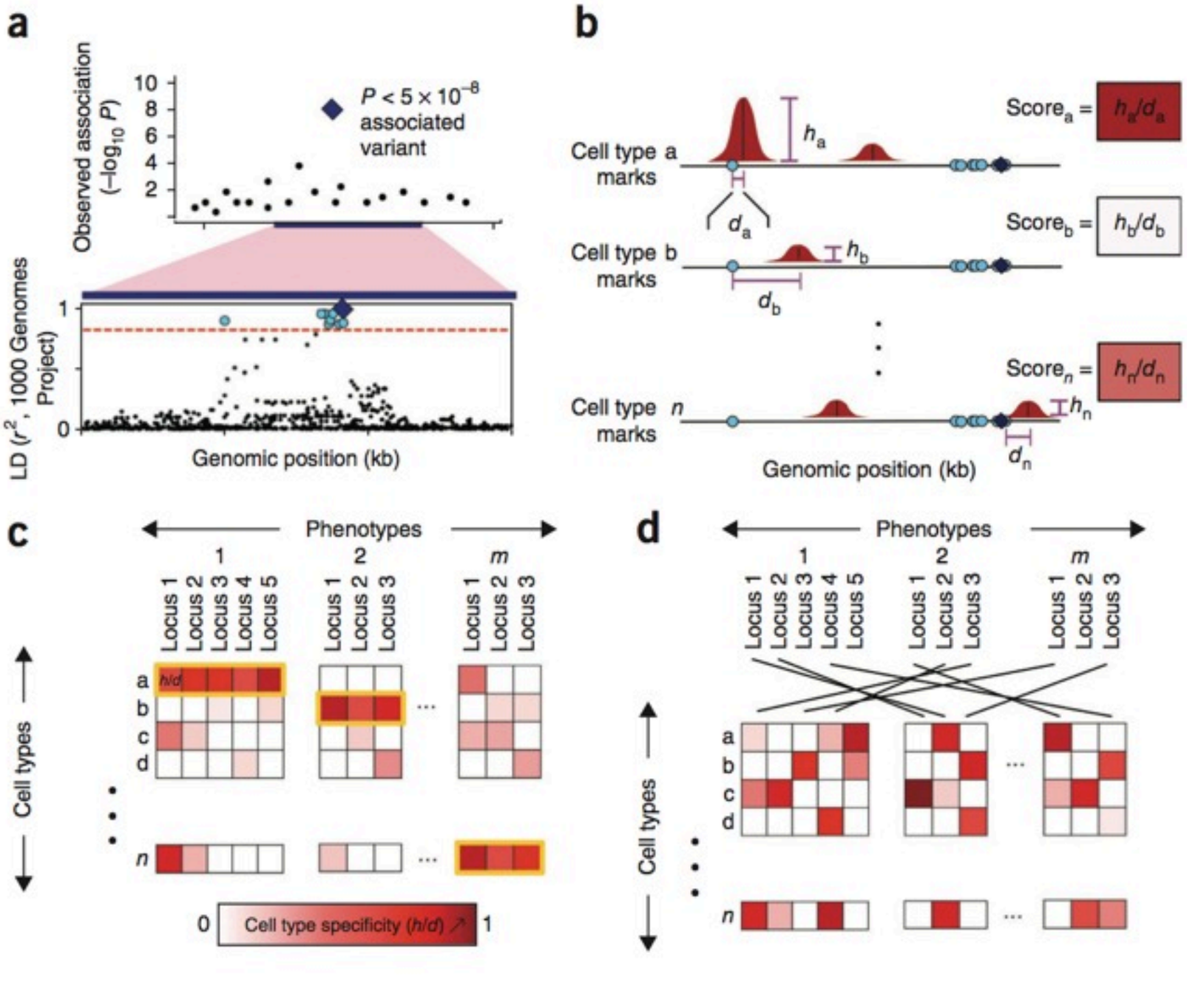
Gosia Trynka[1–4,8], Cynthia Sandor[1–4,8], Buhm Han[1–4], Han Xu[5], Barbara E Stranger[1,4,7], X Shirley Liu[5] & Soumya Raychaudhuri[1–4,6]



Chromatin marks identify critical cell types for fine mapping complex trait variants

Gosia Trynka[1–4,8], Cynthia Sandor[1–4,8], Buhm Han[1–4], Han Xu[5], Barbara E Stranger[1,4,7], X Shirley Liu[5] & Soumya Raychaudhuri[1–4,6]

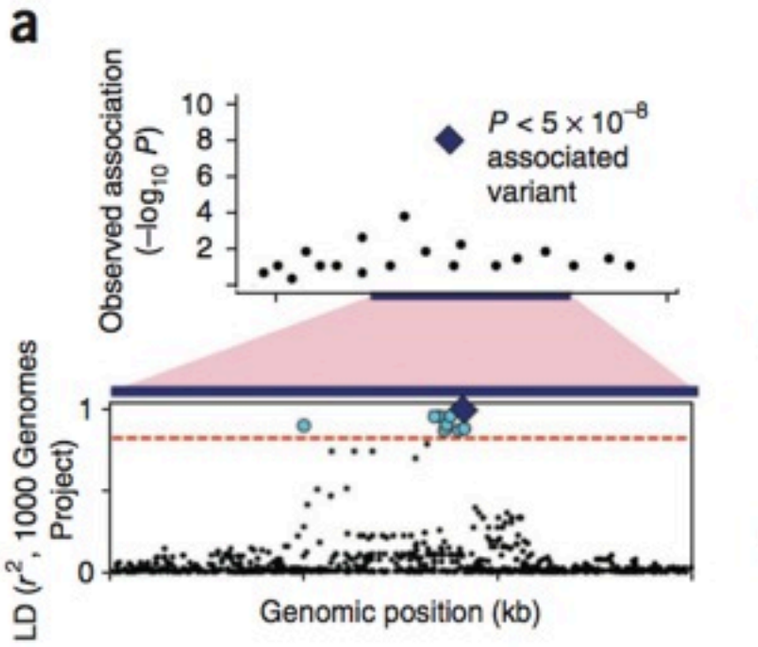
nature genetics



Chromatin marks identify critical cell types for fine mapping complex trait variants

nature genetics

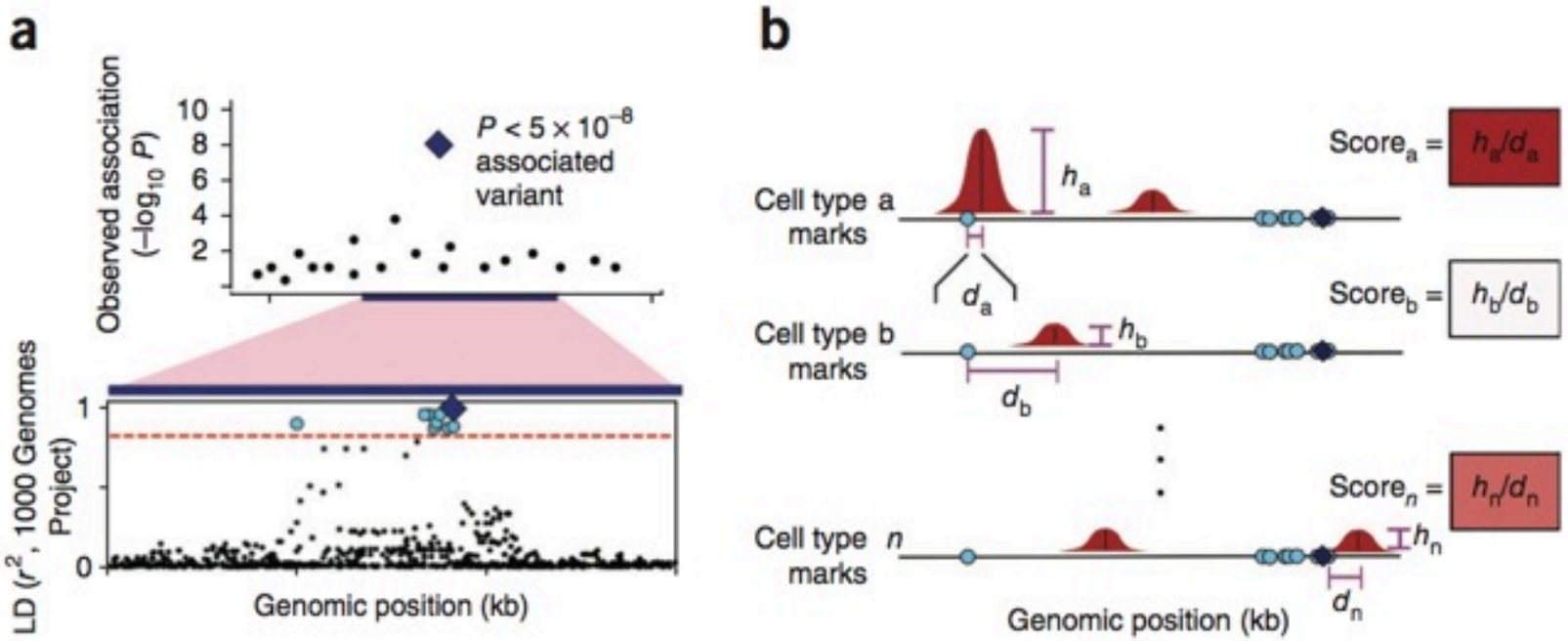
Gosia Trynka[1–4,8], Cynthia Sandor[1–4,8], Buhm Han[1–4], Han Xu[5], Barbara E Stranger[1,4,7], X Shirley Liu[5] & Soumya Raychaudhuri[1–4,6]



Chromatin marks identify critical cell types for fine mapping complex trait variants

nature genetics

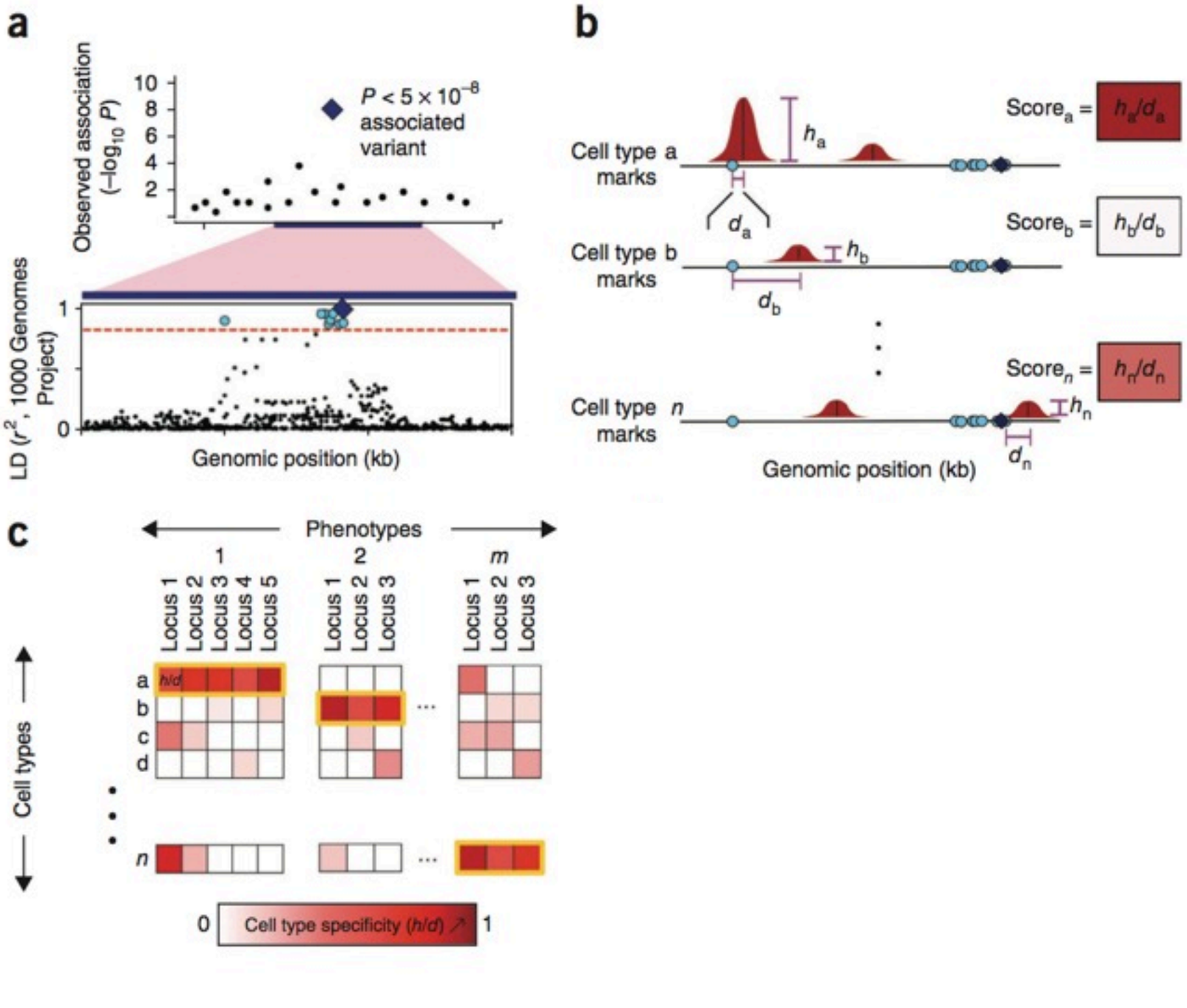
Gosia Trynka[1–4,8], Cynthia Sandor[1–4,8], Buhm Han[1–4], Han Xu[5], Barbara E Stranger[1,4,7], X Shirley Liu[5] & Soumya Raychaudhuri[1–4,6]



Chromatin marks identify critical cell types for fine mapping complex trait variants

nature genetics

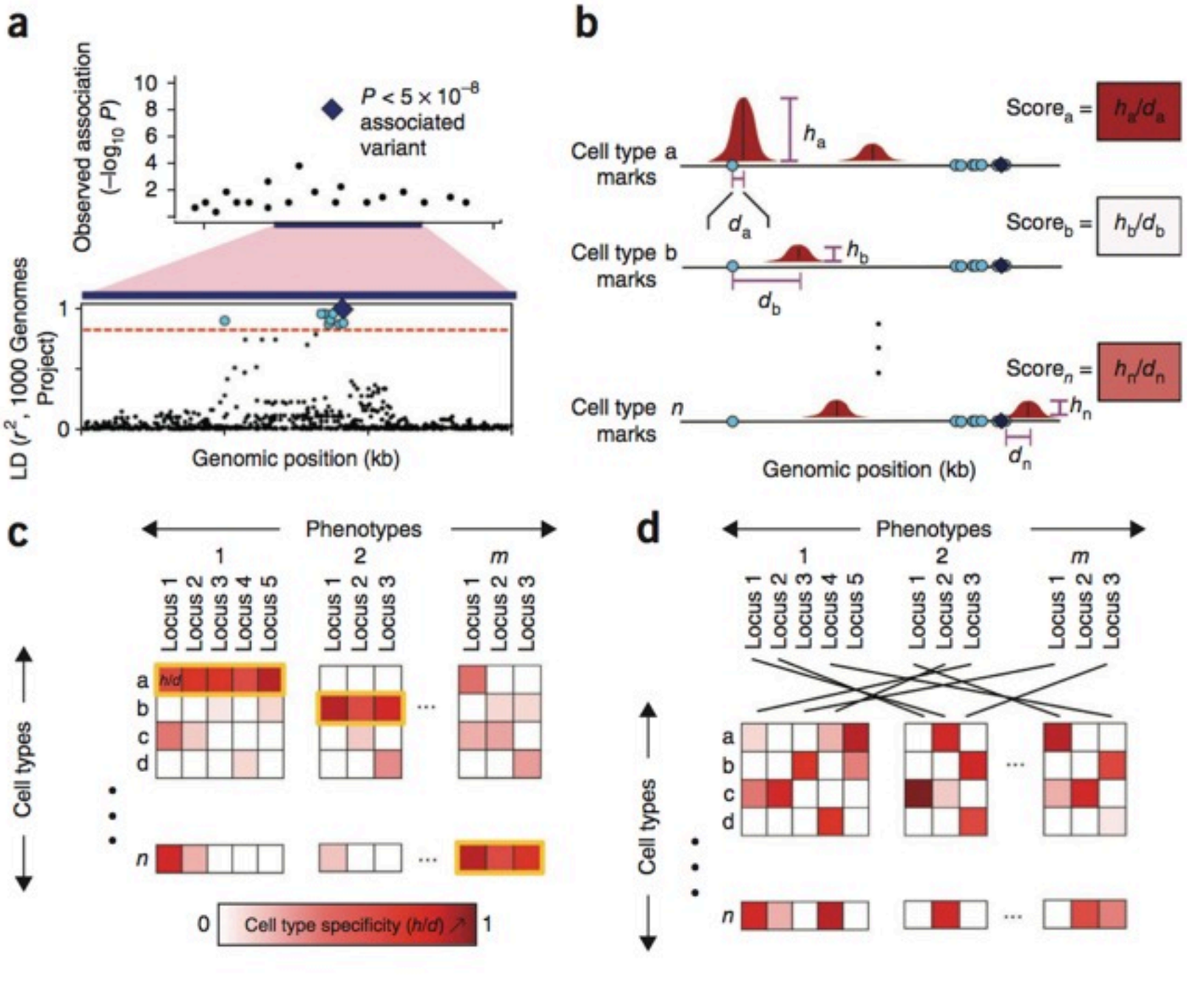
Gosia Trynka[1–4,8], Cynthia Sandor[1–4,8], Buhm Han[1–4], Han Xu[5], Barbara E Stranger[1,4,7], X Shirley Liu[5] & Soumya Raychaudhuri[1–4,6]

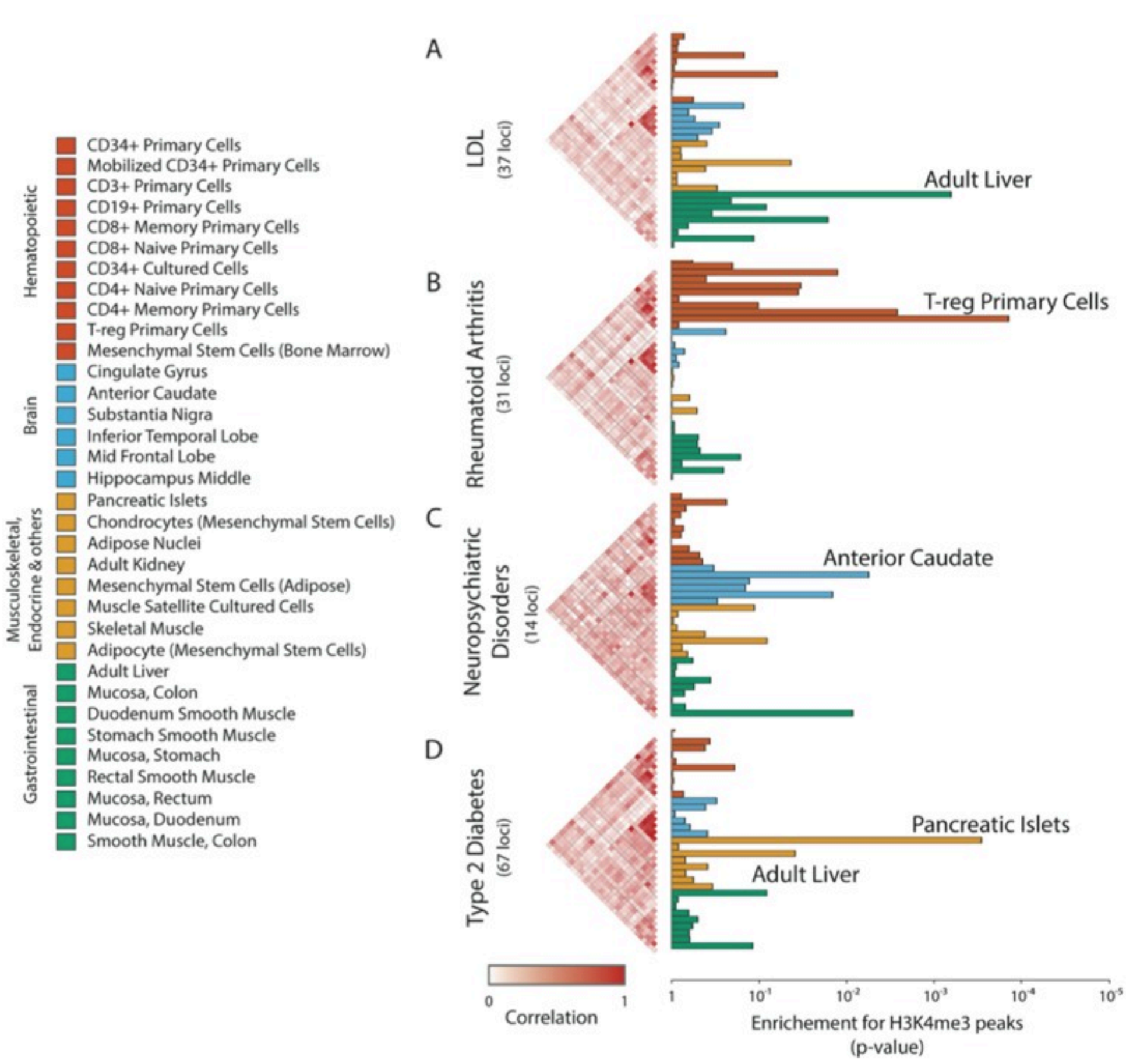


Chromatin marks identify critical cell types for fine mapping complex trait variants

Gosia Trynka[1–4,8], Cynthia Sandor[1–4,8], Buhm Han[1–4], Han Xu[5], Barbara E Stranger[1,4,7], X Shirley Liu[5] & Soumya Raychaudhuri[1–4,6]

nature genetics





Trynka *et al*, Nature Genetics 2012