Lecture 14. ENCODE

Michael Schatz

March 15, 2017

JHU 601.749: Applied Comparative Genomics

Project Proposal!
Due March 15

Project Proposal

Assignment Date: March 7, 2018
Due Date: Thursday, March 15, 2017 @ 11:59pm

Review the Project Ideas page

Work solo or form a team for your class project (no more than 3 people to a team).

The proposal should have the following components:

- Name of your team
- List of team members and email addresses
- Short title for your proposal
- 1 paragraph description of what you hope to do and how you will do it
- References to relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)

Submit the proposal as a single page PDF on blackboard. After submitting your proposal, we will schedule a time to discuss your proposal, especially to ensure you have access to the data that you need. The sooner that you submit your proposal, the sooner we can schedule the meeting. No late days can be used for the project.

Later, you will present your project in class during the last week of class. You will also submit a written report (5-7 pages) of your project, formatting as a Bioinformatics article (Intro, Methods, Results, Discussion, References). Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online

Please use Piazza to coordinate proposal plans!

HW6:
Due March 29

Assignment 6: RNA-seq and differential expression

Assignment Date: Thursday, March 15, 2018
Due Date: Thursday, March 29, 2018 @ 11:59pm

Assignment Overview

In this assignment, you will analyze gene expression data and learn how to make several kinds of plots in the environment of your choice. (We suggest Python or R.) **Make sure to show your work/code in your writeup!** As before, any questions about the assignment should be posted to Piazza.

Question 1. Time Series [10 pts]

This file contains pre-normalized expression values for 100 genes over 10 time points. Most genes have a stable background expression level, but some special genes show increased expression over the timecourse and some show decreased expression.

a. Cluster the genes using an algorithm of your choice. Which genes show increasing expression and which genes show decreasing expression, and how did you determine this? What is the background expression level (numerical value) and how did you determine this? [Hint: K-means and hierarchical clustering are common clustering algorithms you could try.]

b. Calculate the first two principal components of the expression matrix. Show the plot and color the points based on their cluster from part (a). Does the PC1 axis, PC2 axis, neither, or both correspond to the clustering?

c. Create a heatmap of the expression matrix. Order the genes by cluster, but keep the time points in numerical order.

Question 2. Sampling Simulation [10 pts]

A typical human cell has ~250,000 transcripts, and a typical bulk RNA-seq experiment may involve millions of cells. Consequently in an RNAseq experiment you may start with trillions of RNA molecules, although your sequencer will only give a few million to billions of reads. Therefore your RNAseq experiment will be a small sampling of the full composition. We hope the sequences will be a representative sample of the total population, but if your sample is very unlucky or biased it may not represent the true distribution. We will explore this concept by sampling a small subset of transcripts (1000 to 5000) out of a much larger set (1M) so that you can evaluate this bias.

In this file with 1,000,000 lines we provide an abstraction of RNA-seq data where normalization has been performed and the number of times a gene name occurs corresponds to the number of transcripts sequenced.

a. Randomly sample 1000 rows. Do this simulation 10 times and record the relative abundance of each of the 15 genes. Plot the mean vs. variance.

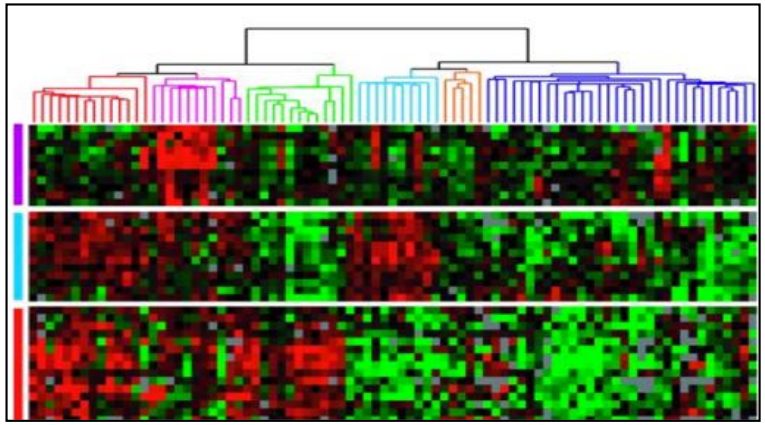
b. Do the same sampling experiment but sample 5000 rows each time. Again plot the mean vs. variance.

c. Is the variance greater in (a) or (b)?, and explain why. What is the relationship between abundance and variance?

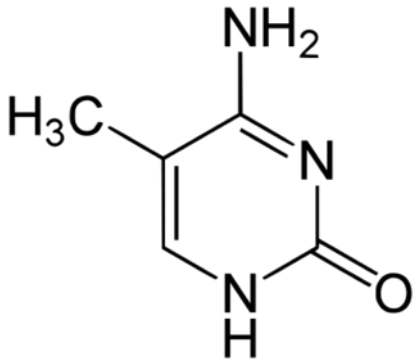
d. Suppose you had received data where the number of times a gene name occurs corresponds to the number of reads mapped to that gene. In a few sentences explain how would you normalize the data, and what additional information would you need? [Hint: why is read count not enough?]

*-seq in 4 short vignettes

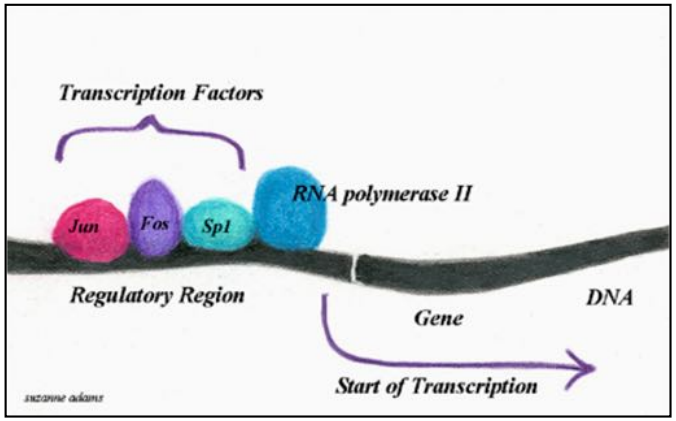
RNA-seq



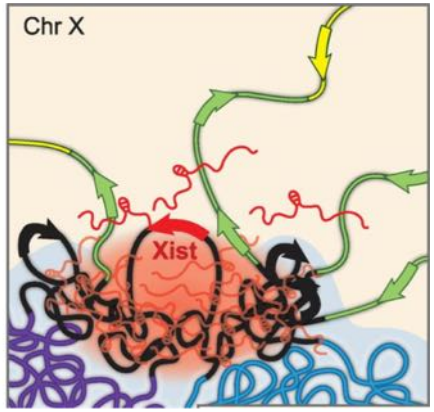
Methyl-seq



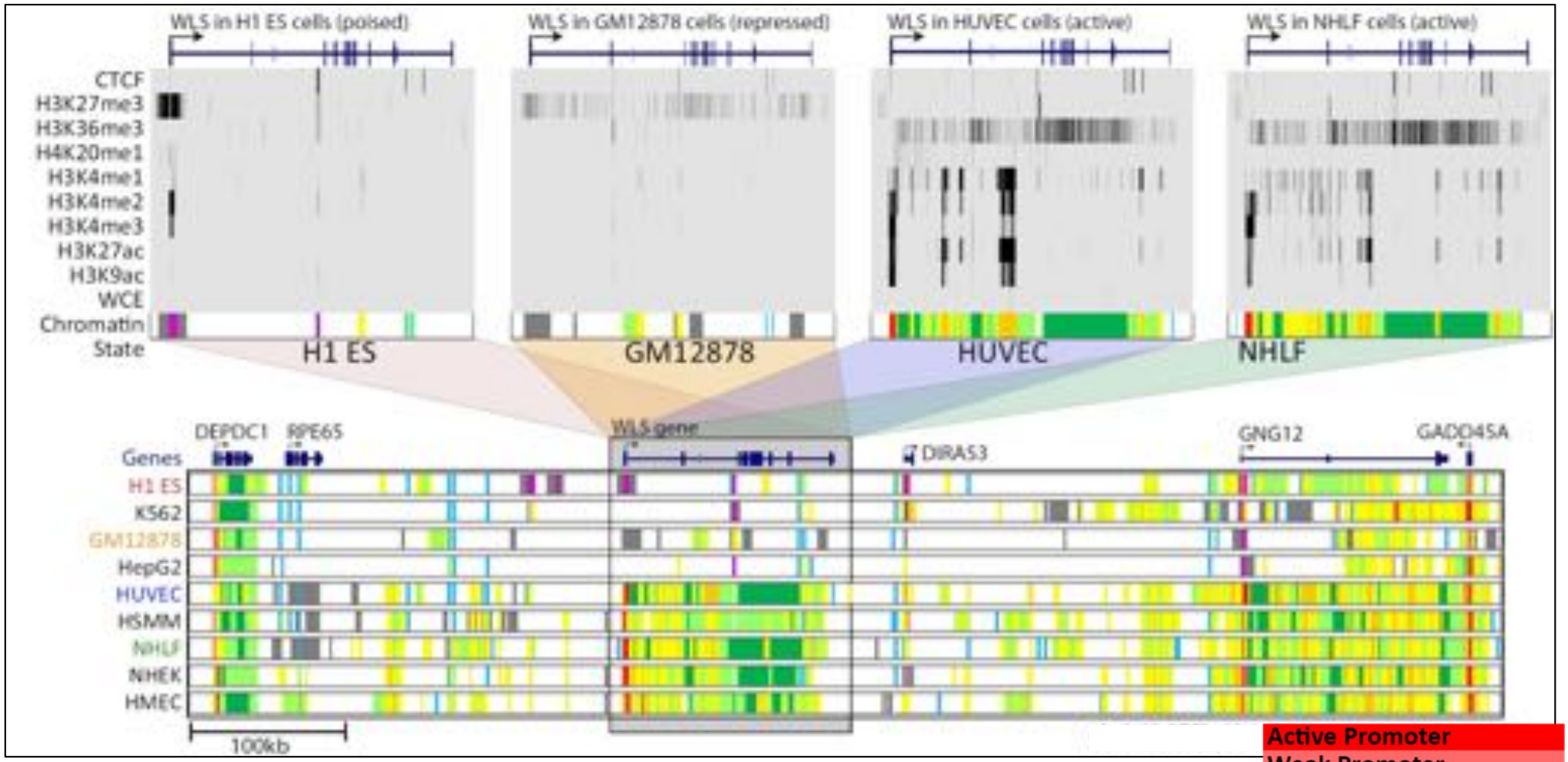
ChIP-seq



Hi-C



Chromatin states dynamics across nine ENCODE cell types



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

Ernst et al, Nature 2011

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

doi:10.1038/nature11247

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

ARTICLE

doi:10.1038/nature11232

The accessible chromatin landscape of the human genome

ARTICLE

doi:10.1038/nature11212

An expansive human regulatory lexicon encoded in transcription factor footprints

ARTICLE

doi:10.1038/nature11245

Architecture of the human regulatory network derived from ENCODE data

LETTER

doi:10.1038/nature11279

The long-range interaction landscape of gene promoters

ARTICLE

doi:10.1038/nature11233

Landscape of transcription in human cells

An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

ARTICLE

The accessible chromatin landscape of the human genome

Research

Long noncoding RNAs are rarely translated in two human cell lines

Research

Discovery of hundreds of mirtrons in mouse and human small RNA data

Resource

GENCODE: The reference human genome annotation for The ENCODE Project

Research

Personal and population genomics of human regulatory variation

Research

Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs

Method

Combining RT-PCR-seq and RNA-seq to catalog all genic elements encoded in the human genome

ARTICLE

Architecture of the human regulatory network derived from ENCODE data

LETTER

The long-range interaction landscape of gene promoters

Method

Predicting cell-type–specific gene expression from regions of open chromatin

Resource

ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

Resource

Annotation of functional variation in personal genomes using RegulomeDB

Method

Linking disease associations with regulatory information in the human genome

RESEARCH Open Access

Modeling gene expression using chromatin features in various cellular contexts

ARTICLE

Landscape of transcription in human cells

ARTICLE

An expansive human regulatory lexicon encoded in transcription factor footprints

RESEARCH Open Access

Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3

RESEARCH Open Access

Functional analysis of transcription factor binding sites in human promoters

RESEARCH Open Access

Analysis of variation at transcription factor binding sites in *Drosophila* and humans

RESEARCH Open Access

Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors


nature | ENCODE
Search
Search ENCODE
Go
Home | Research | Threads | Additional Research | News and Comment | About | Sponsor
nature
ENCODE explorer
THREADS
PAPERS
PRODUCED WITH SUPPORT FROM
illumina
THREAD OVERVIEW:
Non-coding RNA characterization
Many novel and previously known non-coding RNA species are characterized in ENCODE
Read Thread
What is ENCODE?
Threads: a new approach
Guide to the ENCODE explorer

Production Groups

- **A** **Broad Institute**
- **B** **Cold Spring Harbor**; Centre for Genomic Regulation (CRG);
- **C** **University of Connecticut Health Center**; UCSD
- **D** **HudsonAlpha;** Pennsylvania State; UC Irvine; Duke; Caltech
- **E** **UCSD**; Salk Institute ; Joint Genome Institute; Lawrence Berkeley National Laboratory; UCSD
- **F** **Stanford**; University of Chicago; Yale
- **G** **University of Washington**; Fred Hutchinson Cancer Research Center; University of Massachusetts Medical School

Data Coordination Center

- **H** **Stanford**; UCSC

Data Analysis Center

- **I** **University of Massachusetts Medical School**; Yale; MIT; Stanford; Harvard; University of Washington

Technology Development Groups

- **J** **MIT**
- **K** **Washington University, St. Louis**
- **L** **USC**; Ohio State University; UC, Davis
- **M** **University of Washington**
- **N** **Sloan-Kettering**; Weill Cornell Medical College
- **O** **Princeton;** Weizmann
- **P** **University of Michigan**
- **Q** **Broad Institute**
- **R** **University of Washington**; UCSF
- **S** **Advanced RNA Technologies, LLC**
- **T** **Harvard**

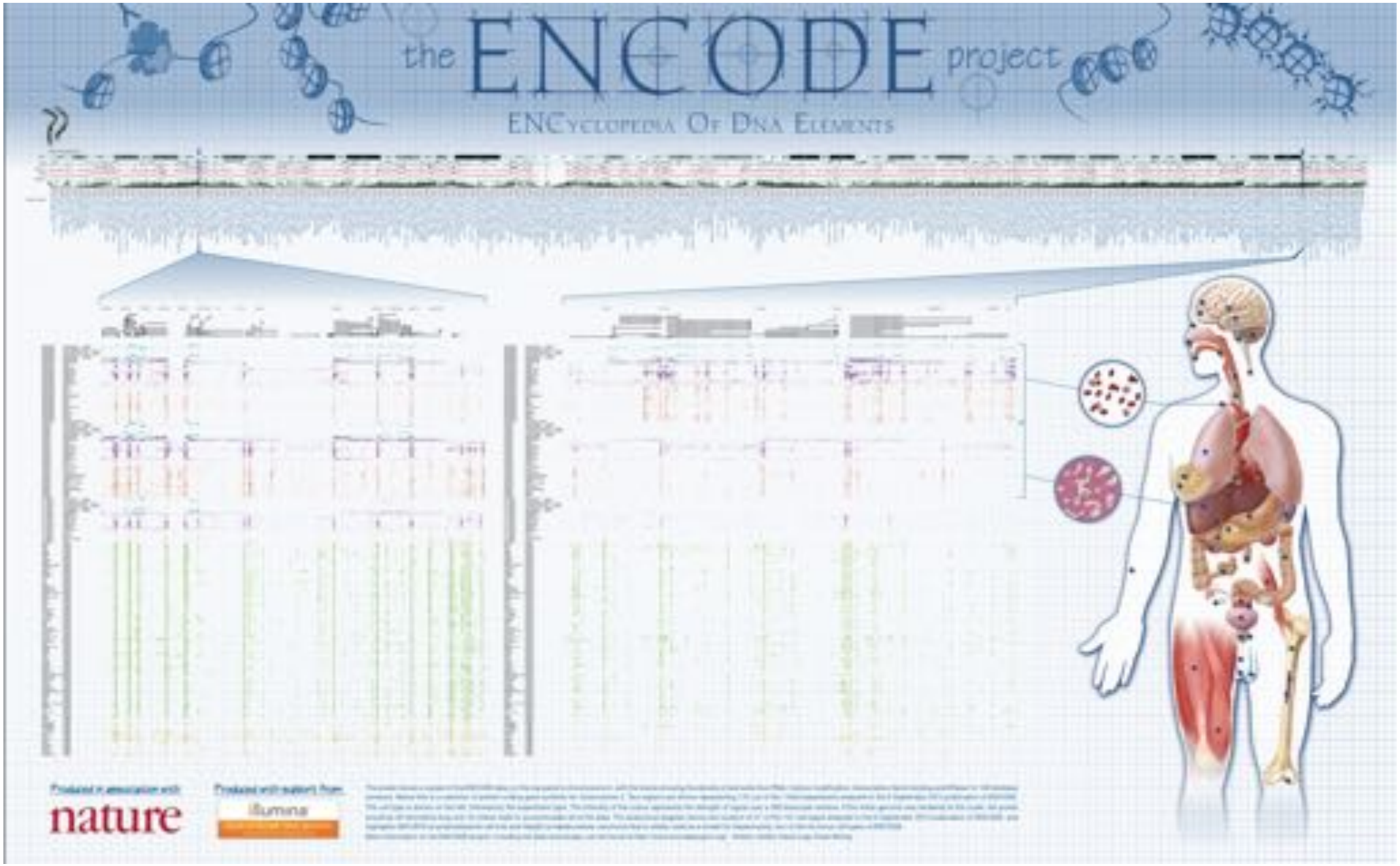
Computational Analysis Groups

- **U** **Berkeley**; Wayne State University
- **V** **MIT**
- **W** **University of Wisconsin**
- **X** **Sloan-Kettering**; Broad Institute
- **Y** **Stanford**
- **Z** **UCLA**

Affiliated Groups

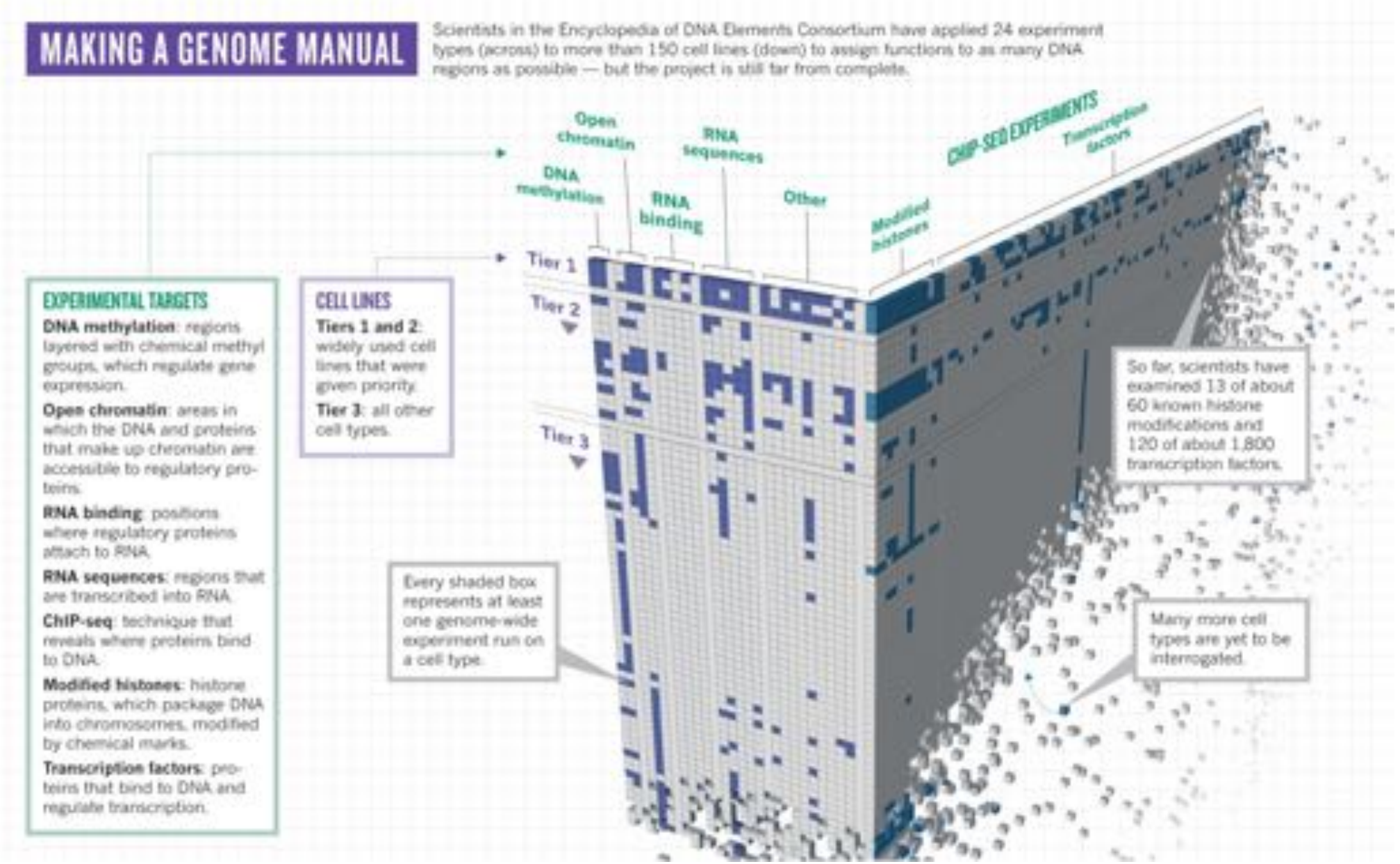
- **1** **Wellcome Trust Sanger Institute**
- **2** **Florida State University**

ENCODE Data Sets



1,640 data sets total over 147 different cell types

ENCODE Data Sets



1,640 data sets total over 147 different cell types

Cell Types

Tier 1 (3 samples, most complete analysis)

- ***GM12878 (NA12878)***: a lymphoblastoid cell line produced from the blood of a female donor with northern and western European ancestry by EBV transformation. It was one of the original HapMap cell lines and has been deeply sequenced using the Solexa/Illumina platform.
- ***K562:*** an immortalized cell line produced from a female patient with chronic myelogenous leukemia (CML). It is a widely used model for cell biology, biochemistry, and erythropoiesis. It grows well, is transfectable, and represents the mesoderm linage.
- ***H1-hESC***: H1-human embryonic stem cells

Tier 2 (9 samples, intermediate analysis)

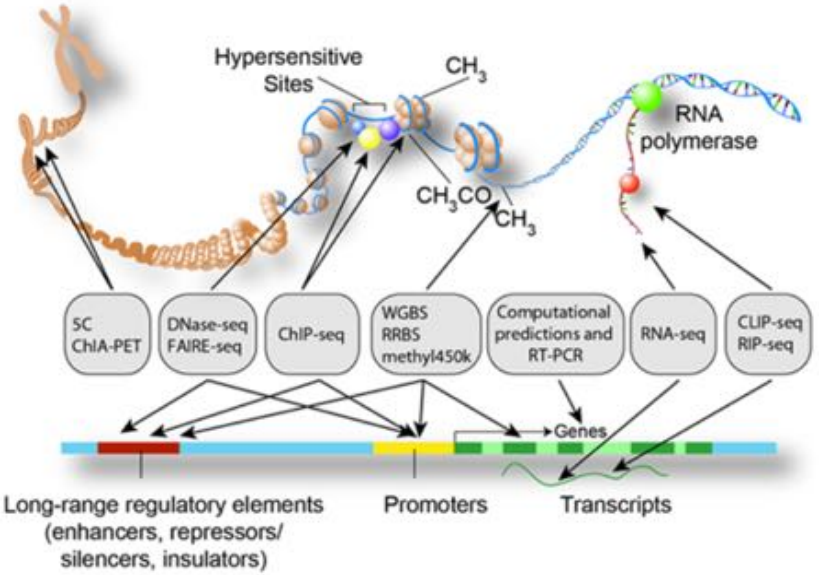
- ***HeLa-S3:*** cervical carcinoma cells
- ***HepG2:*** hepatoblastoma cells & model system for metabolism disorders
- ***HUVECs:*** Primary (non-transformed) human umbilical vein endothelial cells
- Several other major cell lines from cancer and normal tissues

Tier 3 (135 samples, partial analysis)

- Everything else: many major cell lines and body organs

Assays

1. ***RNA transcribed regions***
   - RNA-seq: General sequencing of RNA
   - CAGE: Identify transcription start sites
   - RNA-PET: full length RNA analysis and manual annotation



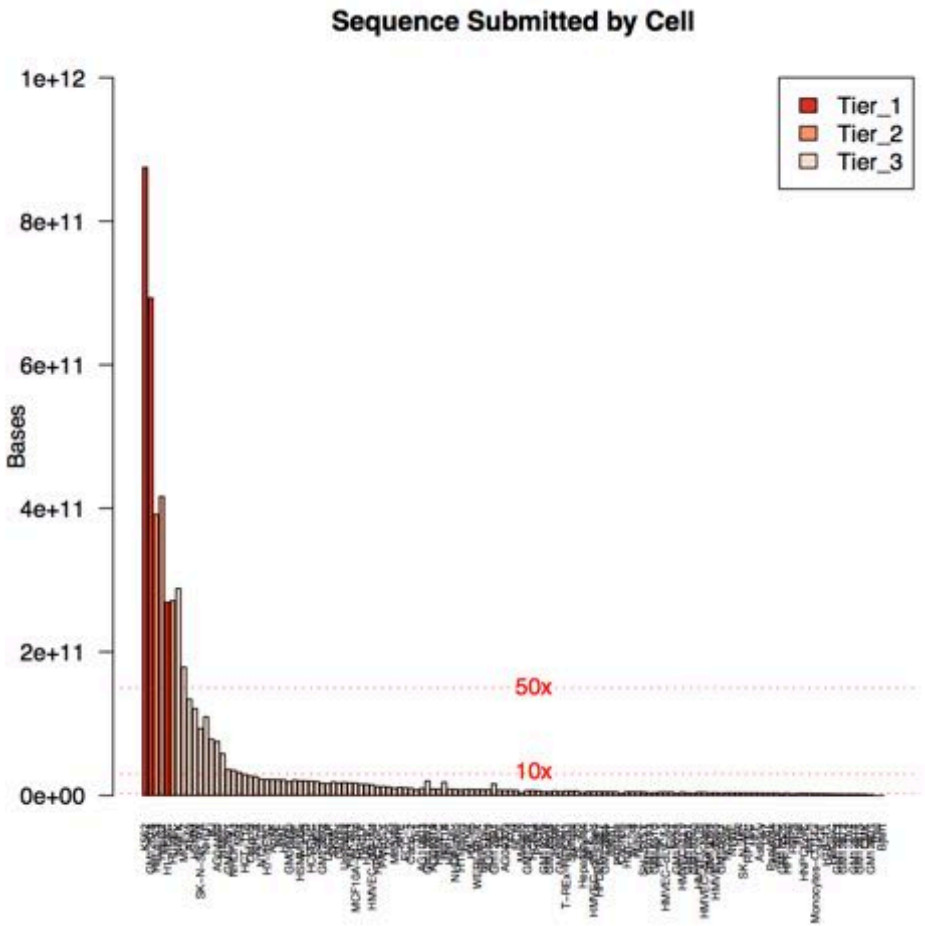
2. ***Protein-coding regions***
   - Mass Spectrometry: Sequencing of proteins

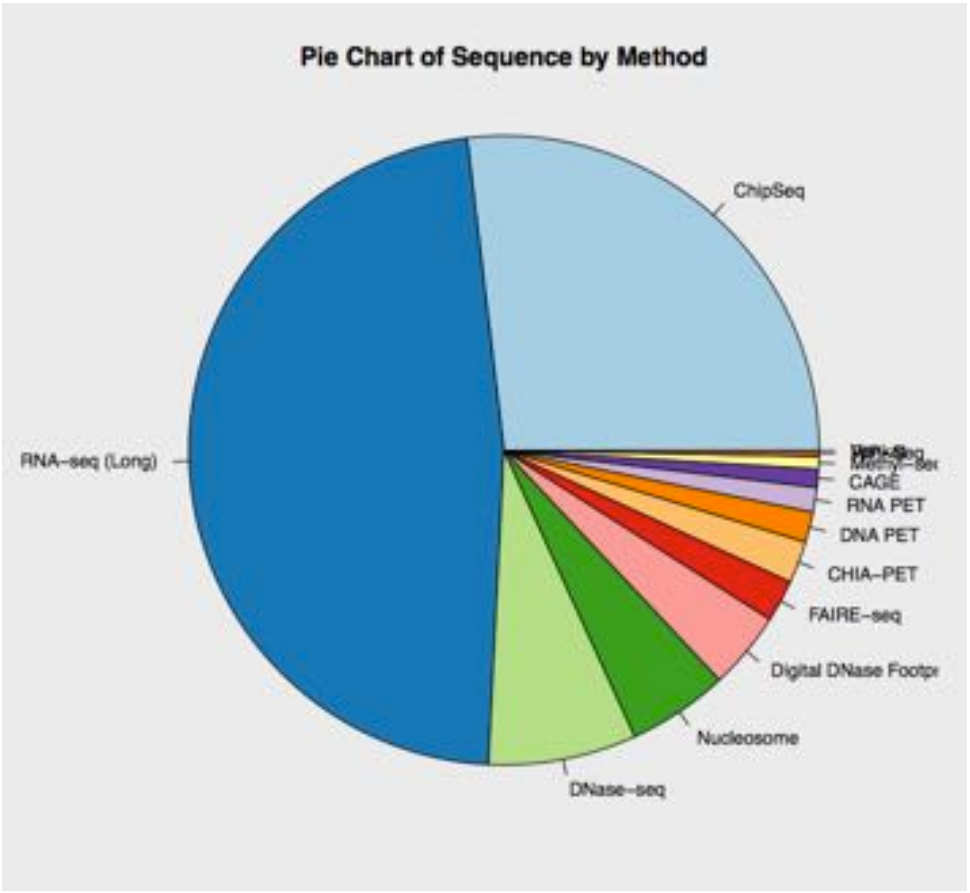
3. ***Transcription-factor-binding sites***
   - ChIP-seq: 119 of 1,800 known transcription factors
   - DNase-seq: open chromatin accessible to Dnase I cutting, "hallmark of regulatory regions"

4. ***Chromatin structure***
   - DNase-seq: 13 of more than 60 currently known histone or DNA modifications
   - FAIRE-seq: nucleosome-depleted regions
   - Histone ChIP-seq: histone proteins pull down and sequencing
   - MNase-seq: nucleosome identification

5. ***DNA methylation sites***
   - RRBS assay: Methyl-seq at targeted sites near restriction binding sites

Data Summary

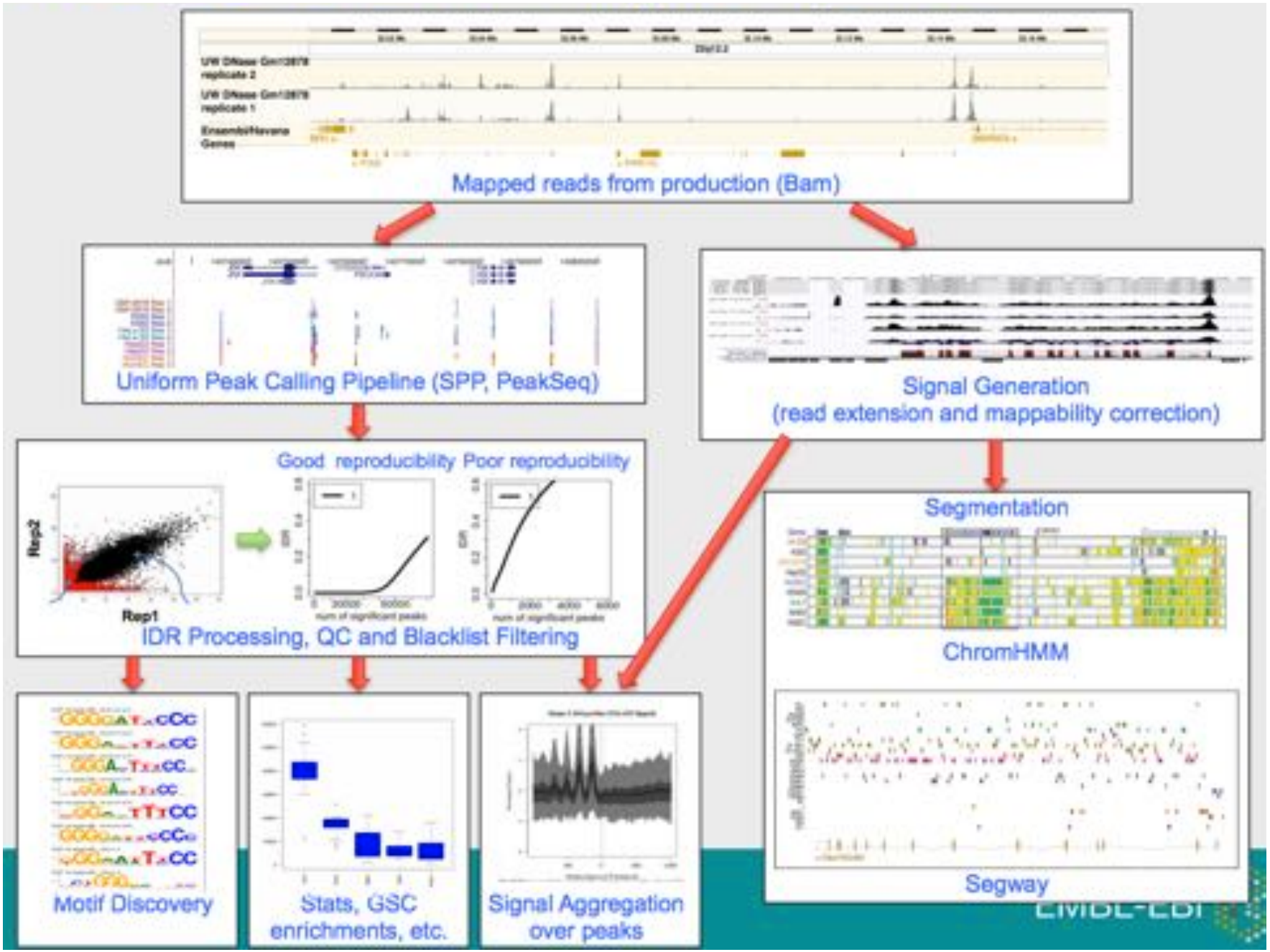




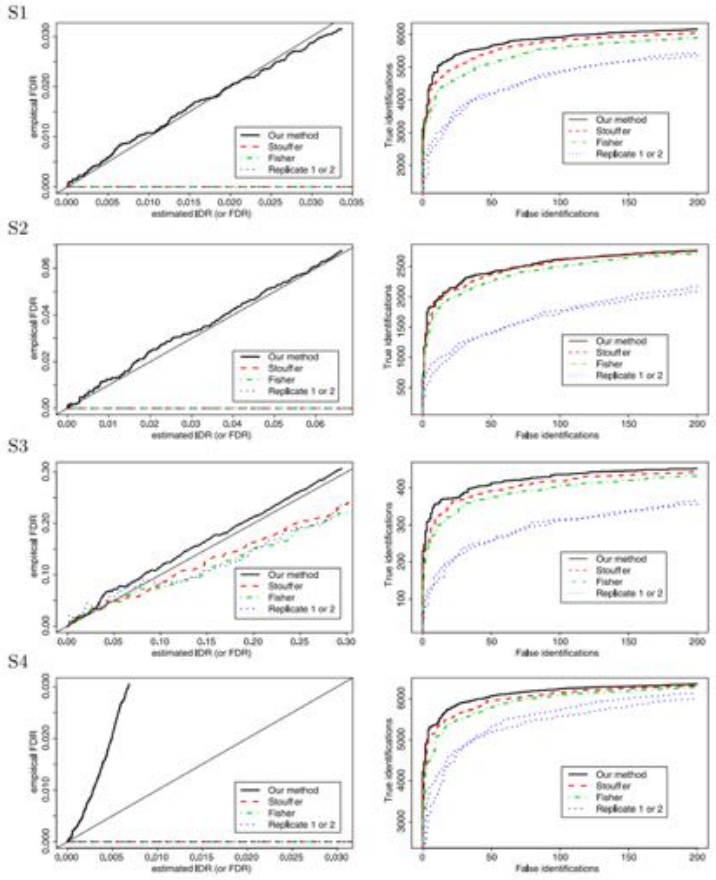
16031 files
1847 Experiments

>5 TeraBases
1716x of the Human Genome

Data Analysis Overview



Irreproducible Discovery Rate (IDR)



Resource

ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

Stephen G. Landt,[1,26] Georgi K. Marinov,[2,26] Anshul Kundaje,[3,26] Pouya Kheradpour,[4] Florencia Pauli,[5] Serafim Batzoglou,[3] Bradley E. Bernstein,[6] Peter Bickel,[7] James B. Brown,[7] Philip Cayting,[1] Yiwen Chen,[8] Gilberto DeSalvo,[2] Charles Epstein,[6] Katherine I. Fisher-Aylor,[2] Ghia Euskirchen,[1] Mark Gerstein,[9] Jason Gertz,[5] Alexander J. Hartemink,[10] Michael M. Hoffman,[11] Vishwanath R. Iyer,[12] Youngsook L. Jung,[13,14] Subhradip Karmakar,[15] Manolis Kellis,[4] Peter V. Kharchenko,[12] Qunhua Li,[16] Tao Liu,[8] X. Shirley Liu,[8] Lijia Ma,[15] Aleksandar Milosavljevic,[17] Richard M. Myers,[5] Peter J. Park,[13,14] Michael J. Pazin,[18] Marc D. Perry,[19] Debasish Raha,[20] Timothy E. Reddy,[5,27] Joel Rozowsky,[9] Noam Shoresh,[6] Arend Sidow,[1,21] Matthew Slattery,[15] John A. Stamatoyannopoulos,[11,22] Michael Y. Tolstorukov,[13,14] Kevin P. White,[15] Simon Xi,[23] Peggy J. Farnham,[24,28] Jason D. Lieb,[25,28] Barbara J. Wold,[2,28] and Michael Snyder[1,28]

[1–25] *[Author affiliations appear at the end of the paper.]*

Chromatin immunoprecipitation (ChIP) followed by high-throughput DNA sequencing (ChIP-seq) has become a valuable and widely used approach for mapping the genomic location of transcription-factor binding and histone modifications in living cells. Despite its widespread use, there are considerable differences in how these experiments are conducted, how the results are scored and evaluated for quality, and how the data and metadata are archived for public use. These practices affect the quality and utility of any global ChIP experiment. Through our experience in performing ChIP-seq experiments, the ENCODE and modENCODE consortia have developed a set of working standards and guidelines for ChIP experiments that are updated routinely. The current guidelines address antibody validation, experimental replication, sequencing depth, data and metadata reporting, and data quality assessment. We discuss how ChIP quality, assessed in these ways, affects different uses of ChIP-seq data. All data sets used in the analysis have been deposited for public viewing and downloading at the ENCODE (http://encodeproject.org/ENCODE/) and modENCODE (http://www.modencode.org/) portals.

[Supplemental material is available for this article.]

ARTICLE

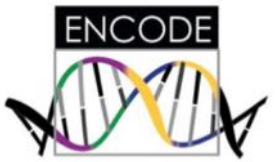
doi:10.1038/nature11247

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.

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.*

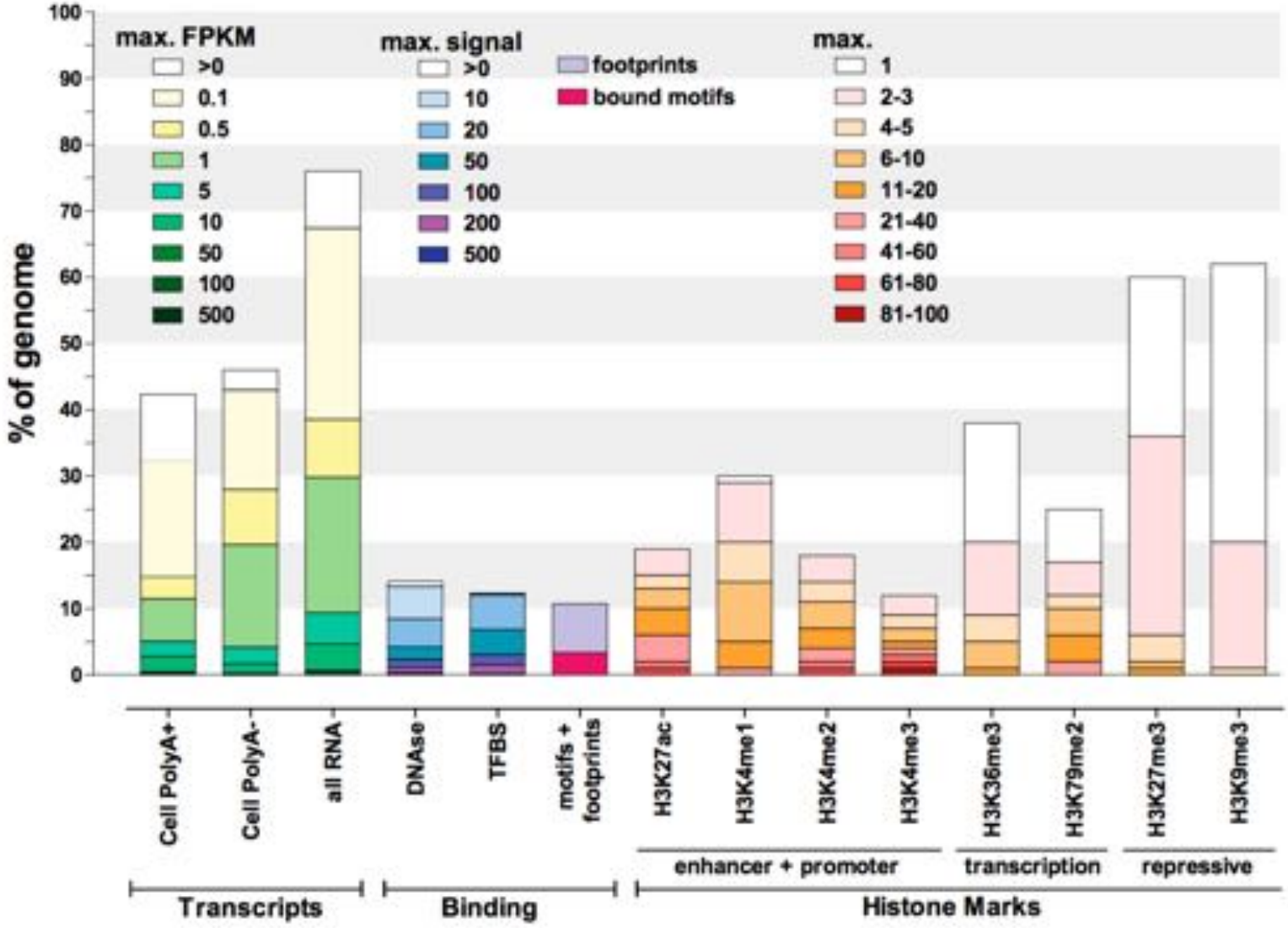
Summary of ENCODE elements

"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

Redefining the concept of a gene

As a consequence of both the expansion of genic regions by the discovery of new isoforms and the identification of novel intergenic transcripts, there has been a marked increase in the number of intergenic regions (from 32,481 to 60,250) due to their fragmentation and a decrease in their lengths (from 14,170 bp to 3,949 bp median length; Fig. 6). Concordantly, we observed an increased overlap of genic regions. As the determination of genic regions is currently defined by the cumulative lengths of the isoforms and their genetic association to phenotypic characteristics, the likely continued reduction in the lengths of intergenic regions will steadily lead to the overlap of most genes previously assumed to be distinct genetic loci. This supports and is consistent with earlier observations of a highly interleaved transcribed genome[12], but more importantly, prompts the reconsideration of the definition of a gene. As this is a consistent characteristic of annotated genomes, we would propose that the transcript be considered as the basic atomic unit of inheritance. Concomitantly, the term gene would then denote a higher-order concept intended to capture all those transcripts (eventually divorced from their genomic locations) that contribute to a given phenotypic trait. Co-published ENCODE-related papers can be explored online via the *Nature* ENCODE explorer (http://www.nature.com/ENCODE), a specially designed visualization tool that allows users to access the linked papers and investigate topics that are discussed in multiple papers via thematically organized threads.

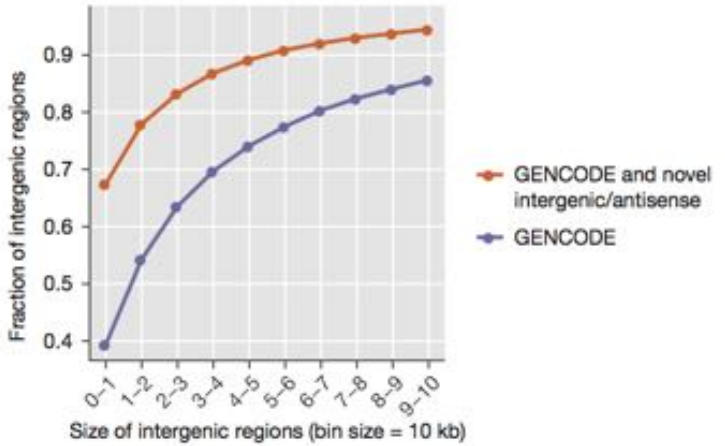


Figure 6 | Size distribution of intergenic regions. Novel genes increase the proportion of small intergenic regions.

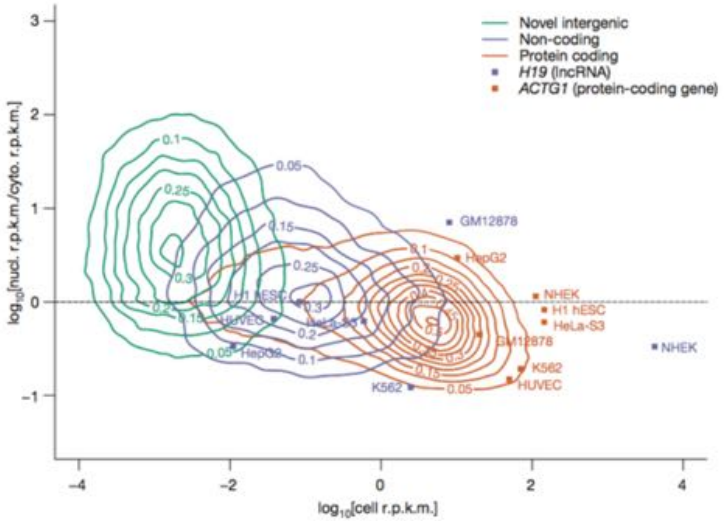
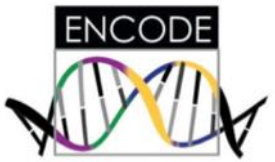


Figure 3 | Abundance of gene types in cellular compartments. Two-dimensional kernel density plots of nuclear over cytosolic enrichment (*y* axis) versus overall gene expression in the whole cell extract (*x* axis), for protein coding, long non-coding and novel genes over all cell lines. Only genes present

Landscape of transcription in human cells
Djebali et al. (2012) *Nature*. doi:10.1038/nature11233

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.*

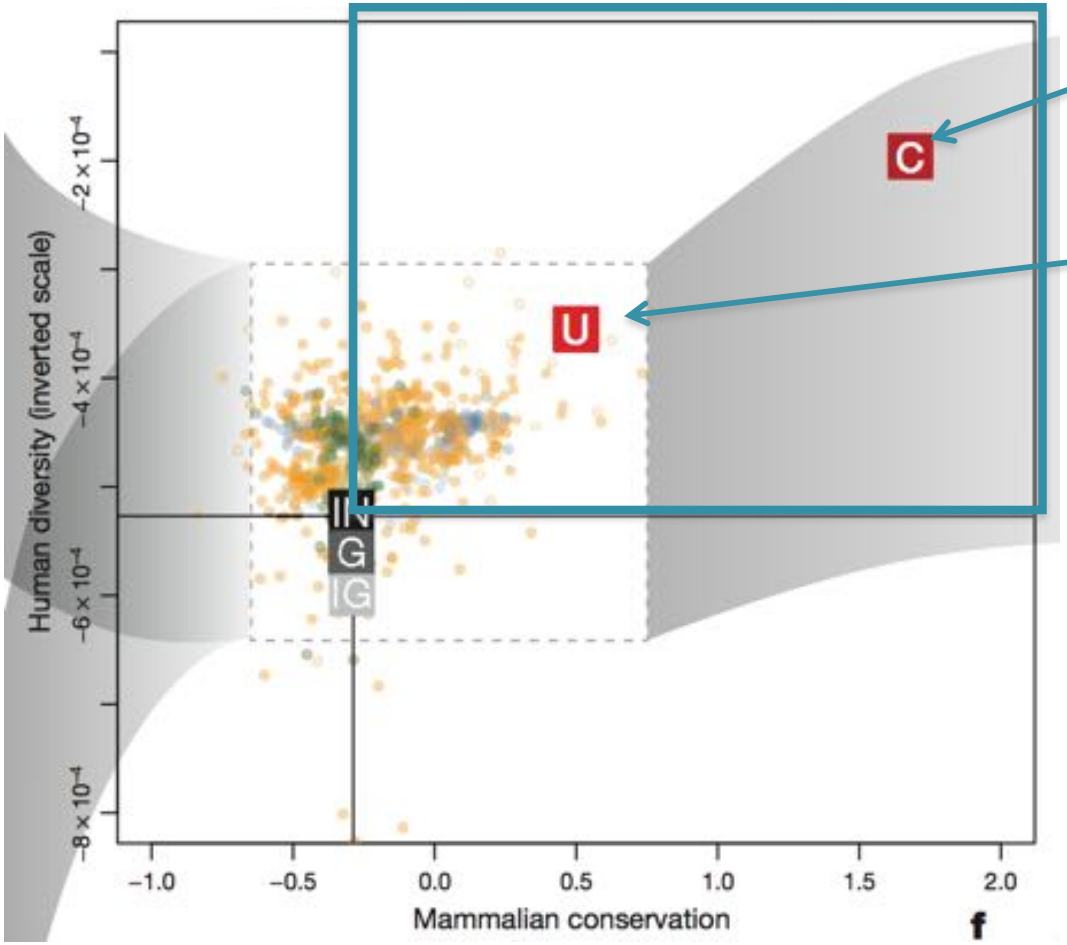
Impact and Evidence of Selection

Most constrained
=> Most likely functional

Average values across
Protein coding sequences

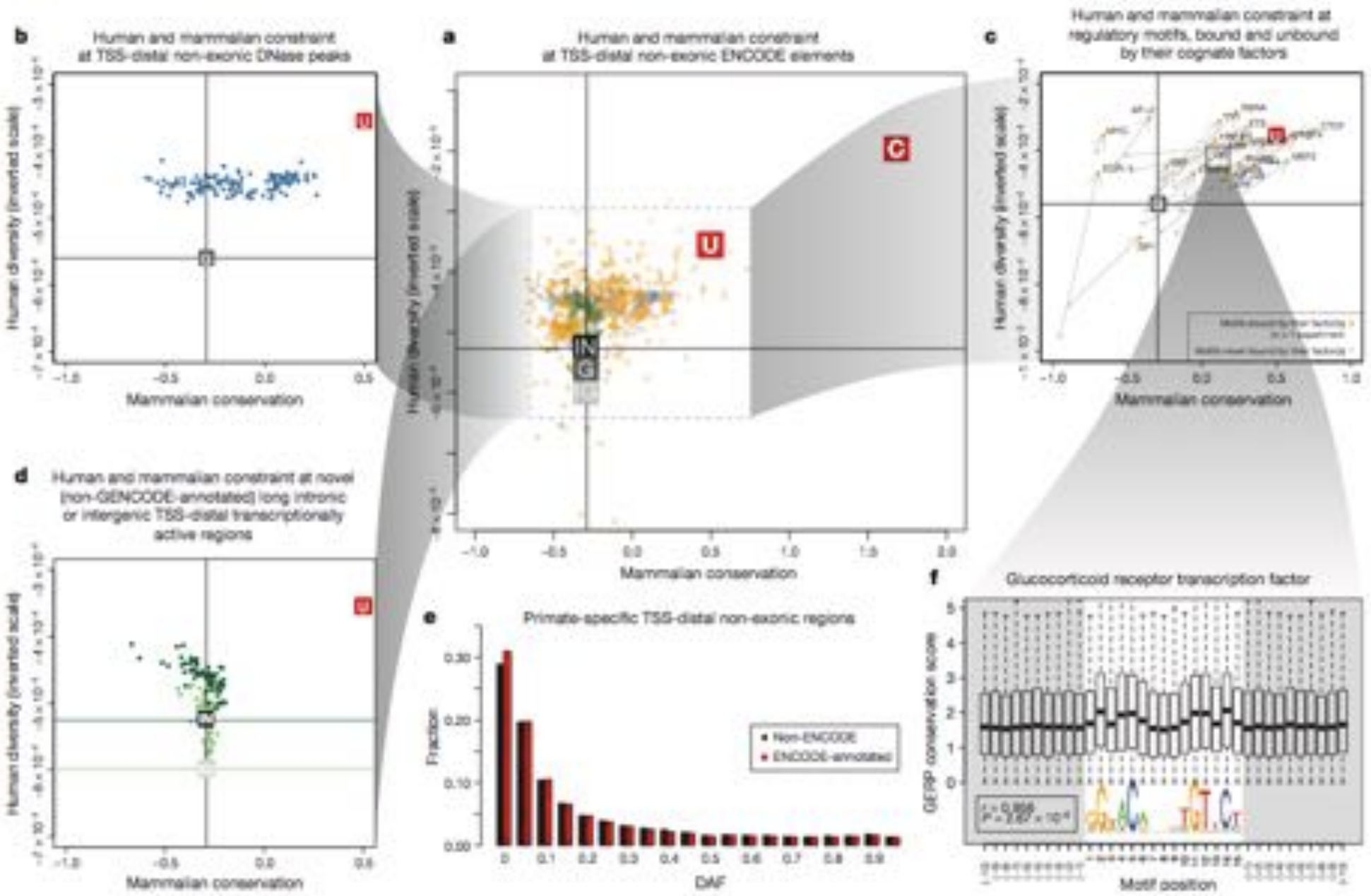
Average values across
UTR sequences

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

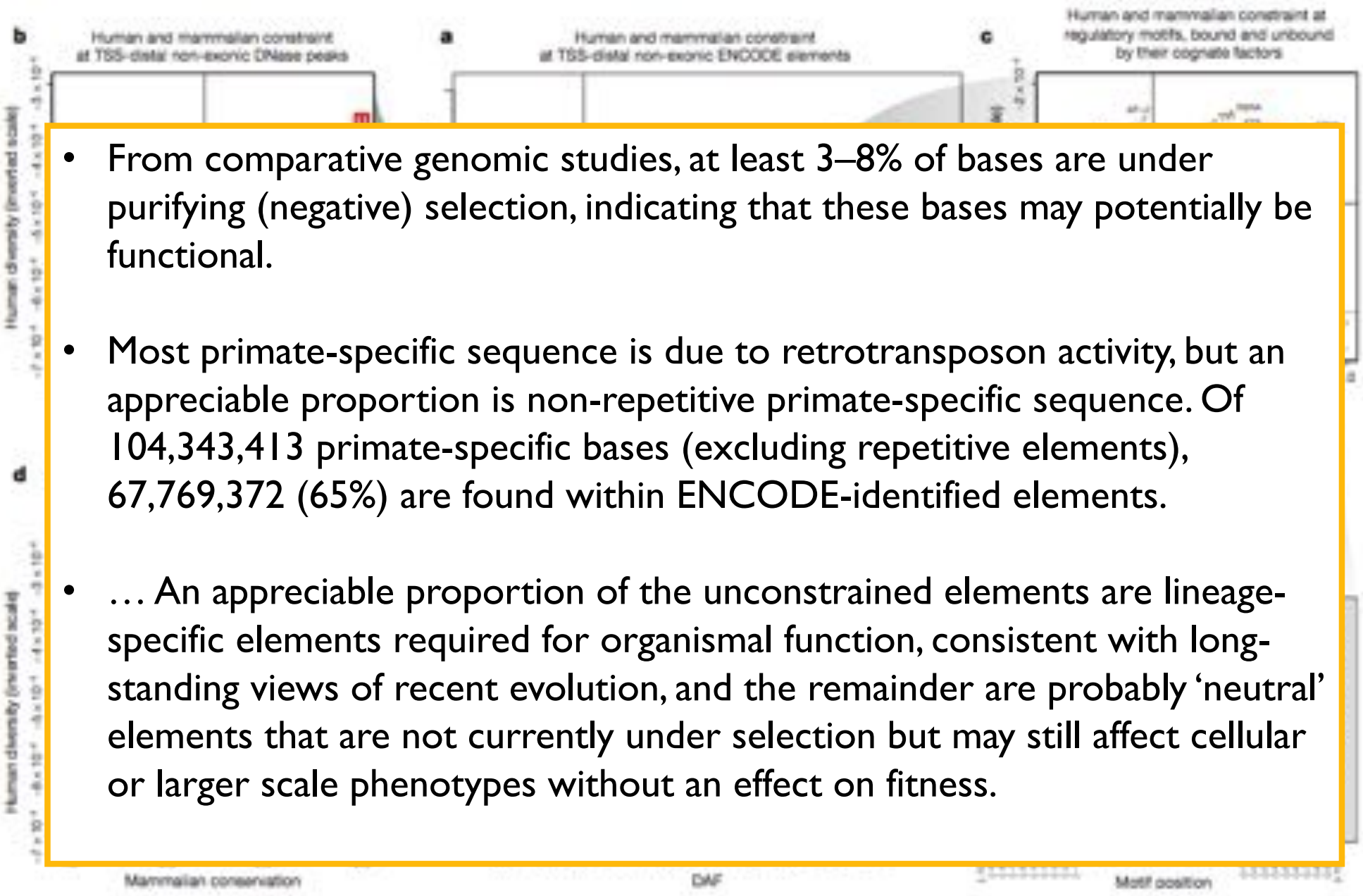


For a given ENCODE region, how much conservation do we see across 24 sequenced mammalian genomes?

Impact and Evidence of Selection

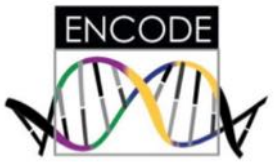


Impact and Evidence of Selection



- 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

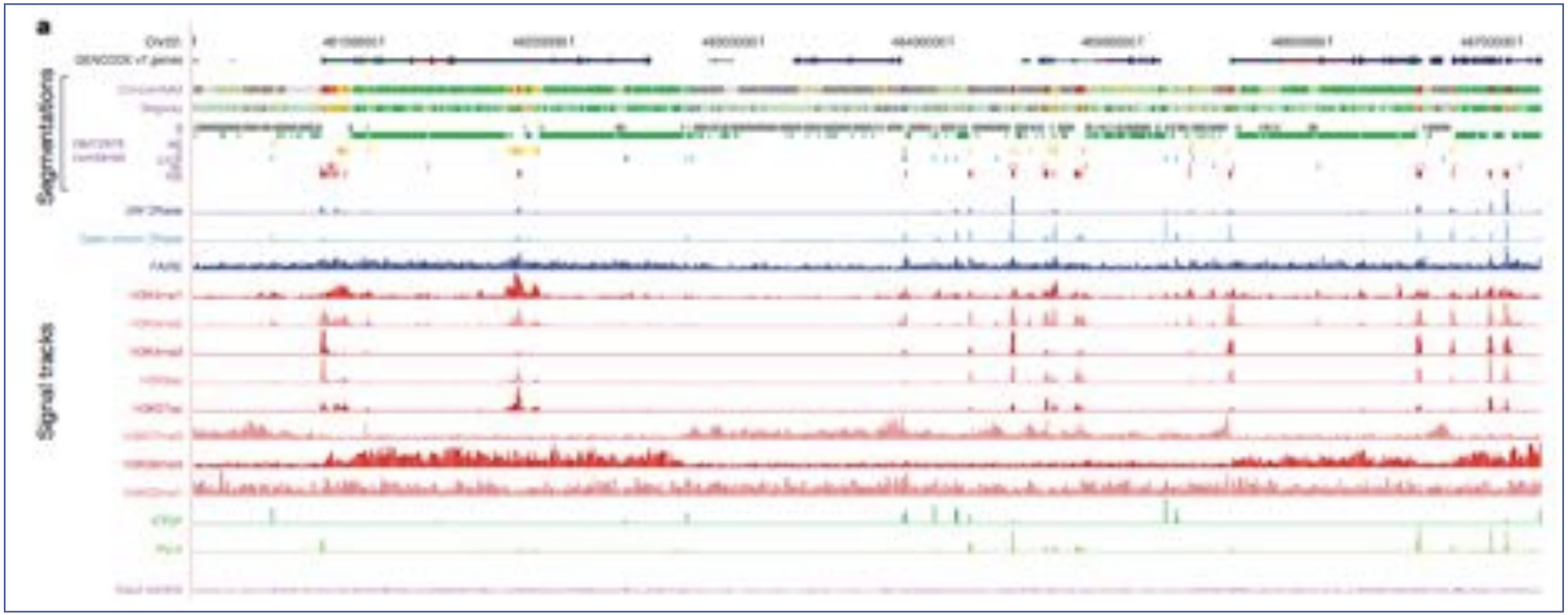


Table 3 | Summary of the combined state types

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

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.*

Histone Modifications

- Histones are the proteins around which DNA is wound into nucleosomes and at a higher level chromatin
- Histone modifications have been previously reported to indicate repressive/activating functional state
- Use ChIP-seq techniques to locate where they are in the genome
  - Cannot be predicted from sequence composition alone, highly dependent on cell type and cell state)

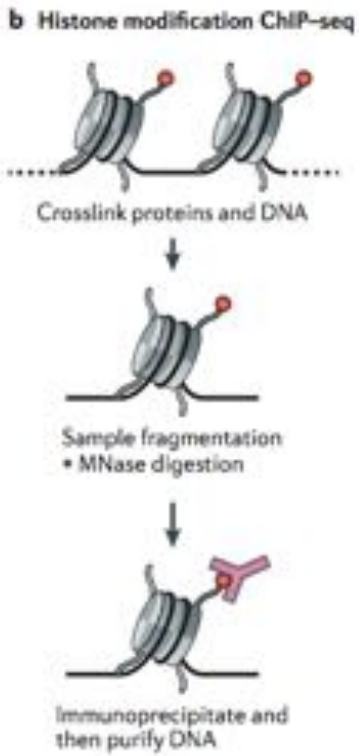
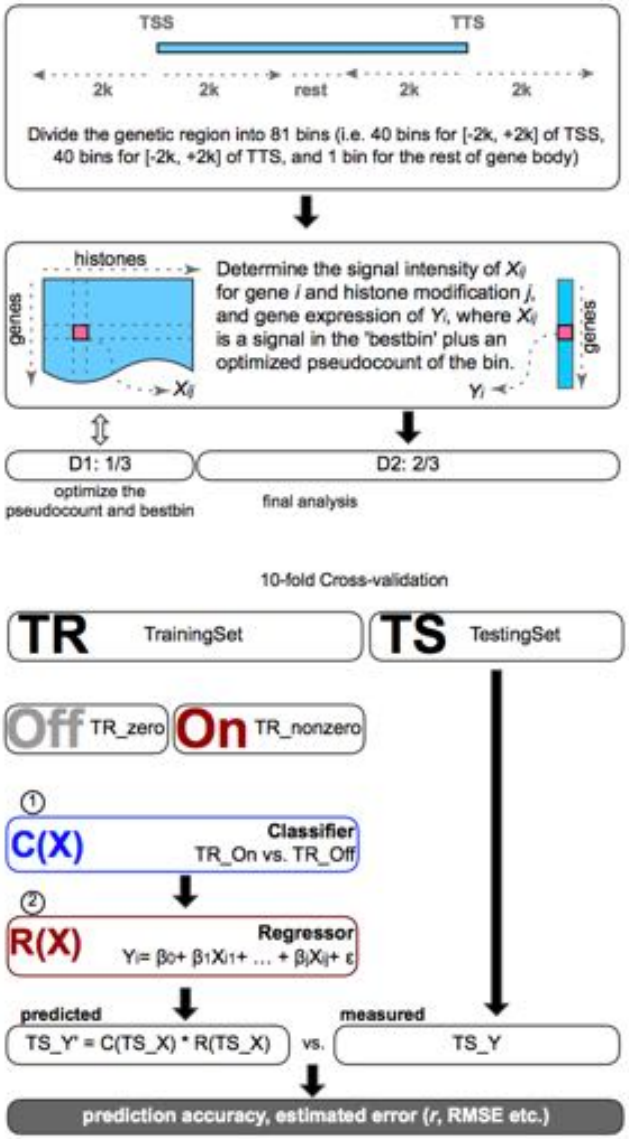
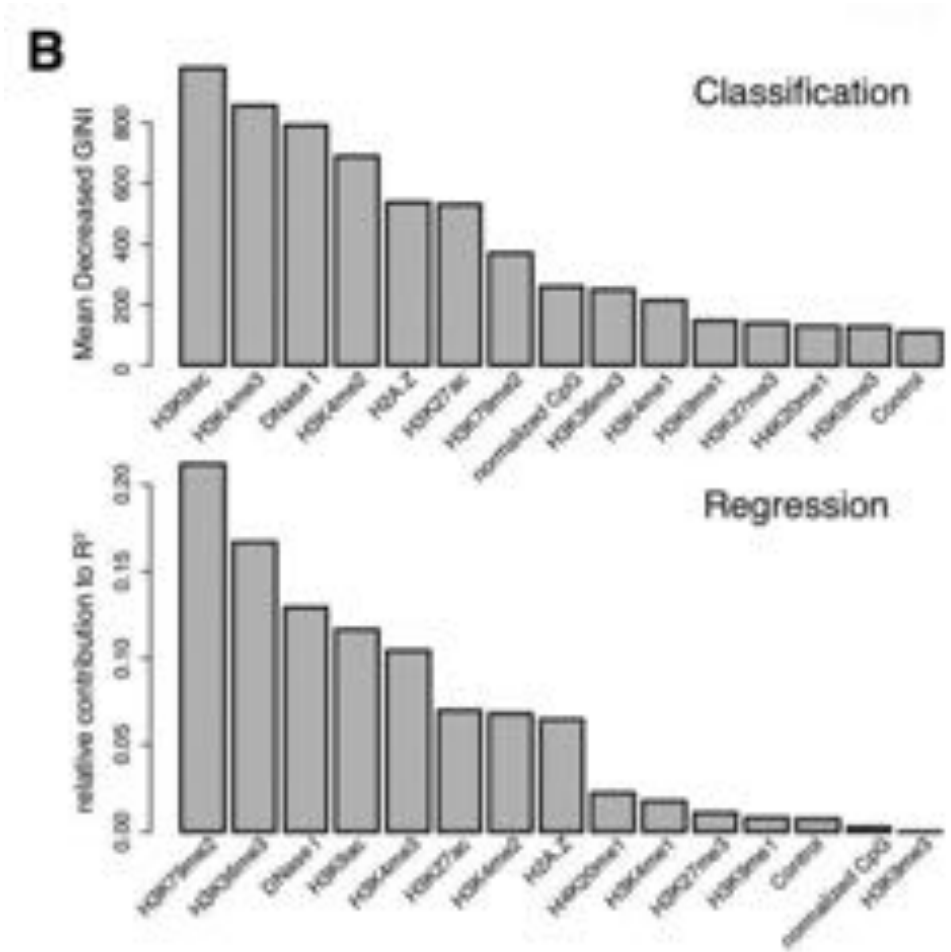


Table 2 | Summary of ENCODE histone modifications and variants

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

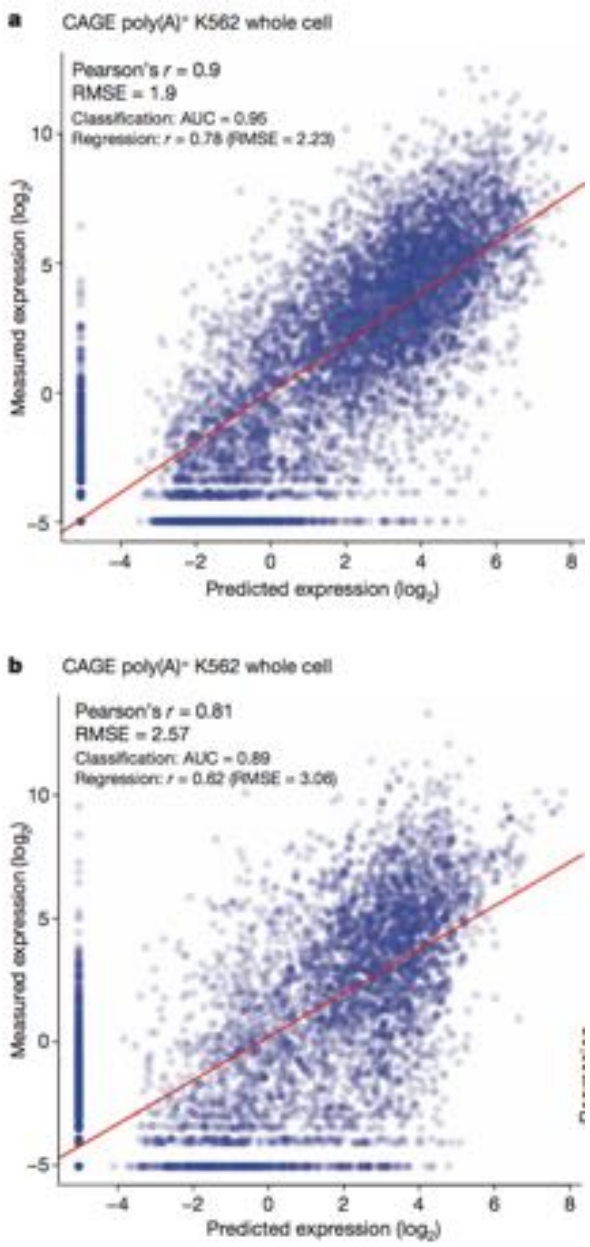
Expression Modeling





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

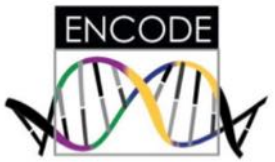
Expression Modeling



- 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

Figure 2 | Modelling transcription levels from histone modification and transcription-factor-binding patterns. **a**, **b**, Correlative models between either histone modifications or transcription factors, respectively, and RNA production as measured by CAGE tag density at TSSs in K562 cells. In each case the scatter plot shows the output of the correlation models (*x* axis) compared to observed values (*y* axis). The bar graphs show the most important histone modifications (**a**) or transcription factors (**b**) in both the initial classification phase (top bar graph) or the quantitative regression phase (bottom bar graph), with larger values indicating increasing importance of the variable in the model. Further analysis of other cell lines and RNA measurement types is reported elsewhere[59,79]. AUC, area under curve; Gini, Gini coefficient; RMSE, root mean square error.

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

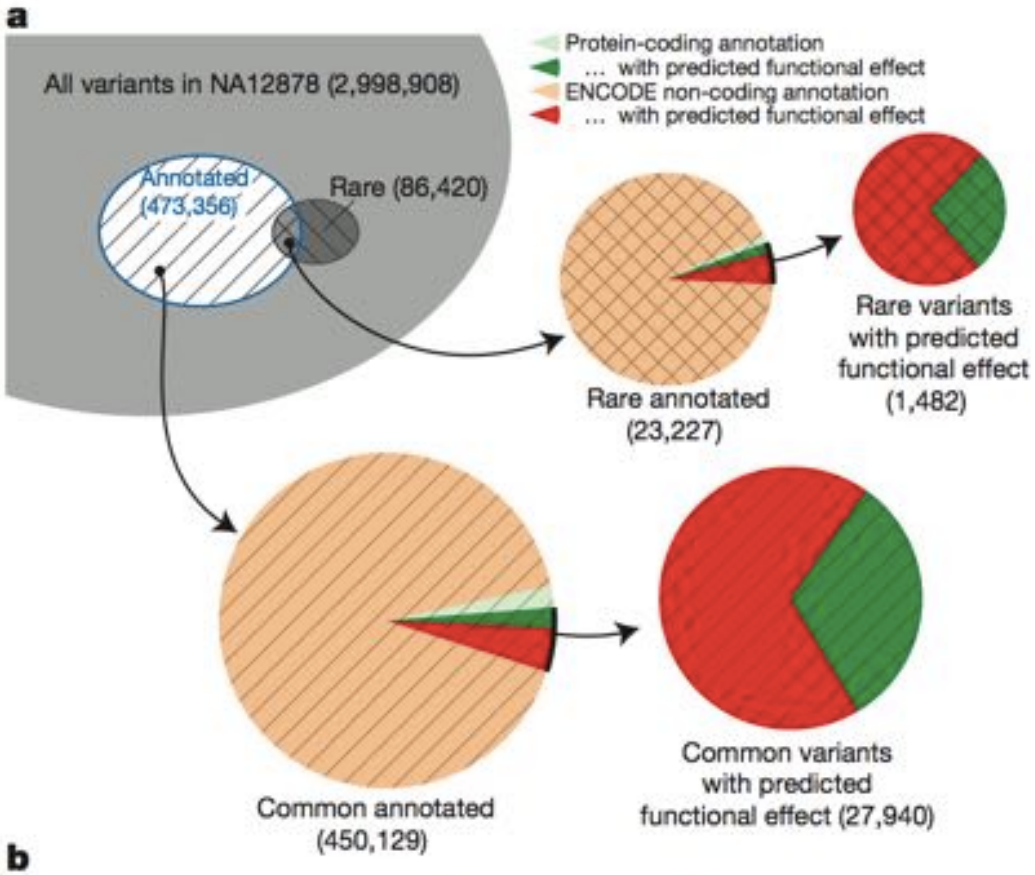


Figure 9 | Examining ENCODE elements on a per individual basis in the normal and cancer genome. **a**, Breakdown of variants in a single genome (NA12878) by both frequency (common or rare (that is, variants not present in the low-coverage sequencing of 179 individuals in the pilot 1 European panel of 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

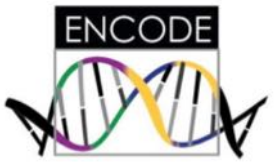
- Common or Rare (that is, variants not present in the low-coverage sequencing of 179 individuals in the pilot 1 European panel of the 1000 Genomes project)
- ENCODE annotation, including protein-coding gene and non-coding elements

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

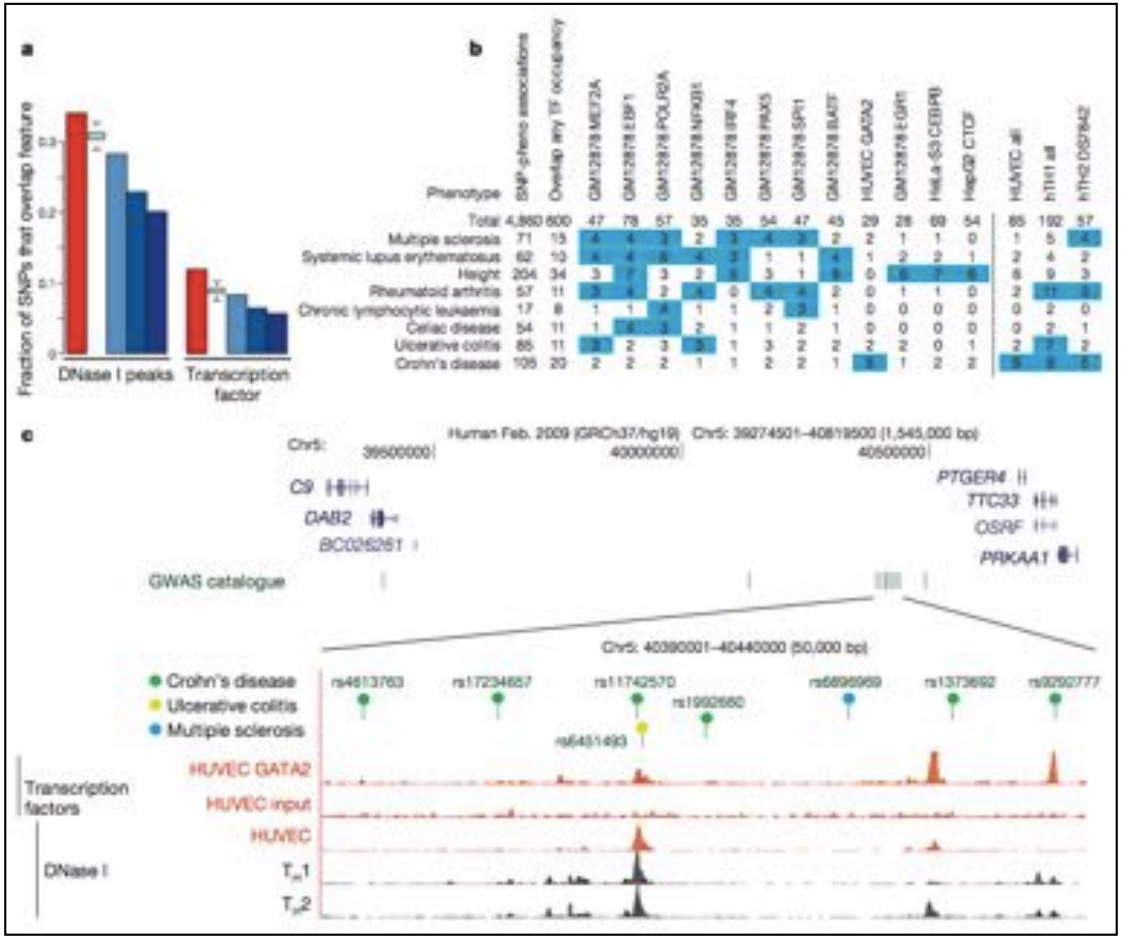
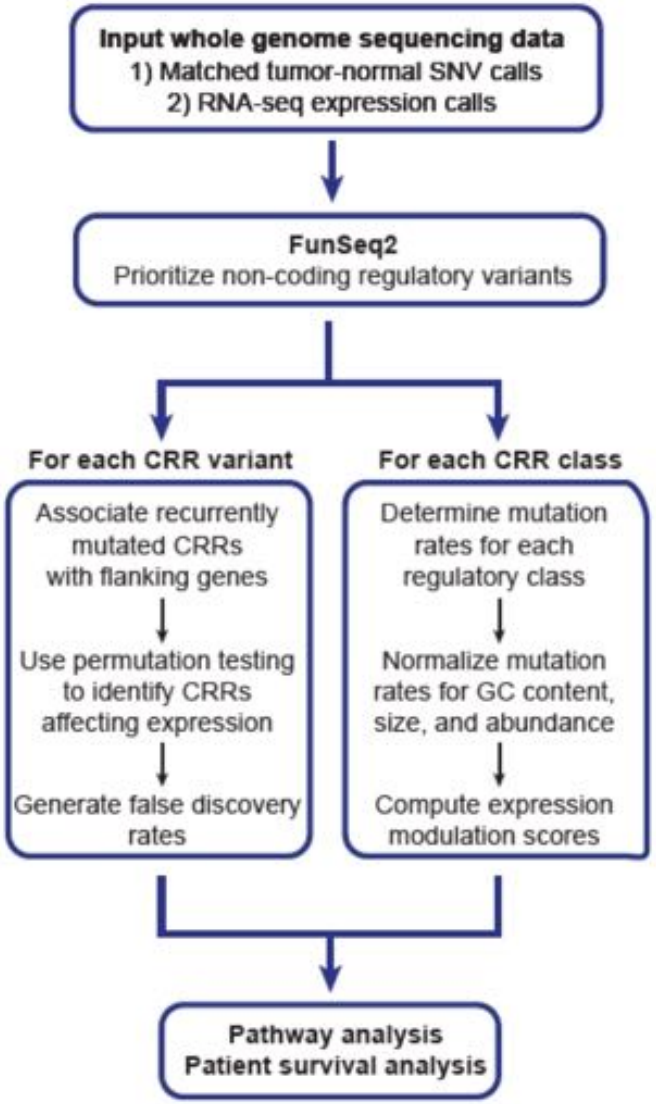


Figure 10 | Comparison of genome-wide-association-study-identified loci with ENCODE data. **a**, Overlap of lead SNPs in the NHGRI GWAS SNP catalogue (June 2011) with DHSs (left) or transcription-factor-binding sites (right) as red bars compared with various control SNP sets in blue. The control SNP 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 and Cancer



Coding alterations of PDAC are now fairly well established but non-coding mutations (NCMs) largely unexplored

- Developed GECCO to analyze the thousands of somatic mutations observed from hundreds of tumors to find potential drivers of gene expression and pathogenesis

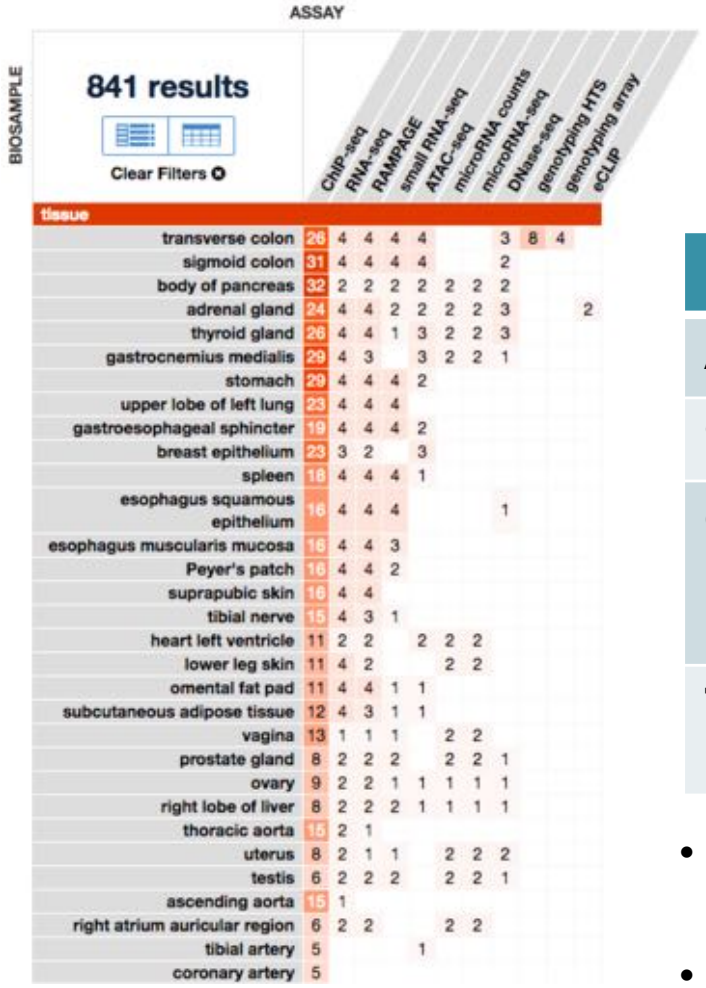
- NCMs are enriched in known and novel pathways
- NCMs correlate with changes in gene expression
- NCMs can demonstrably modulate gene expression
- NCMs correlate with novel clinical outcomes

NCMs are an important mechanism for tumor genome evolution

Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma
Feigin, M, Garvin, T et al. (2017) Nature Genetics. doi:10.1038/ng.3861

ENCODE Studies

>6000 Citations for main paper; >>10k for all papers



http://encodeproject.org

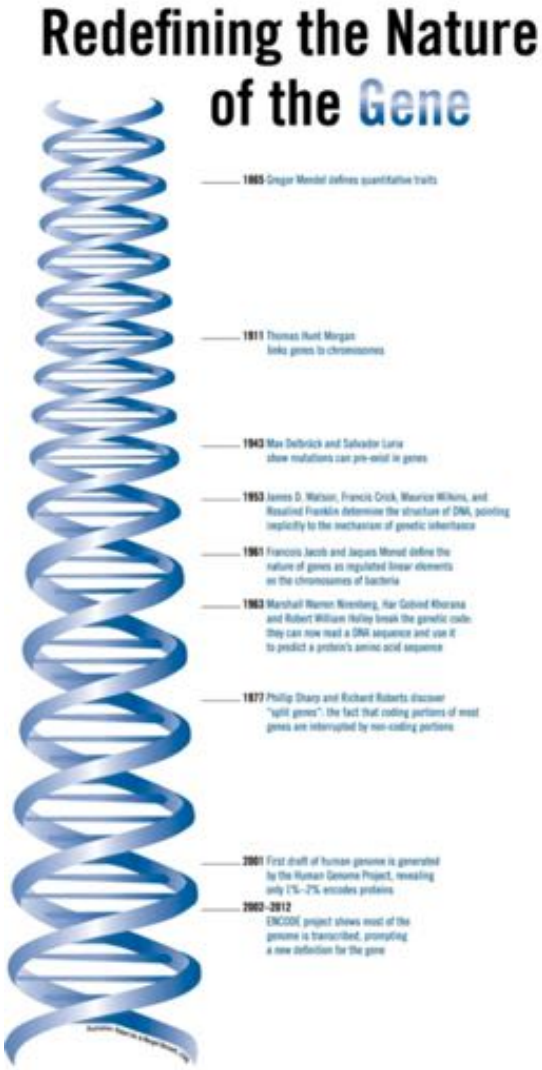
EN-TEx: Expression & Regulation Analysis of Personalized Genomes

	ENC-001	ENC-002	ENC-003	ENC-004
Age	37	54	53	51
Sex	Male	Male	Female	Female
Cause of Death	Anoxia	Anoxia	Cerebral Vascular Accident	Cerebral Vascular Accident
Total Libraries	319	299	488	299

- Sequenced the genomes for 2 male and 2 female samples using transverse colon tissue
- Large number of ChIP-seq, RNA-seq, ATAC-seq, DNase-seq, and other functional datasets available in dozens of tissues

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?

Comment on "Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions"

Phil Green* and Brent Ewing

Ward and Kellis (Reports, 28 September 2012, p. 1675; published online 5 September 2012) found altered patterns of human polymorphism in biochemically active but non–mammalian-conserved genomic regions relative to control regions and interpreted this as due to lineage-specific purifying selection. We find on closer inspection of their data that the polymorphism trends are primarily attributable to mutational variation and technical artifacts rather than selection.

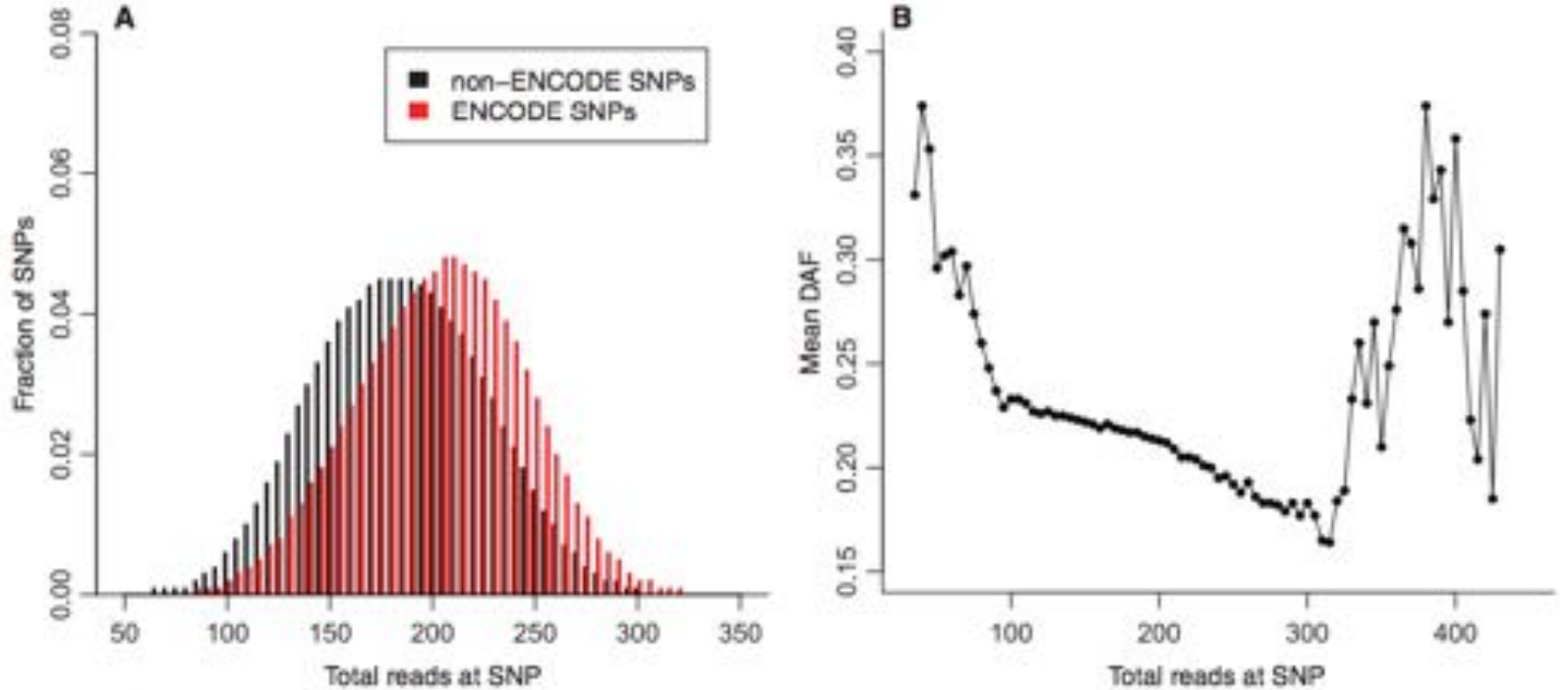


Fig. 1. Variation in 1000 Genomes read depth (totaled over 59 Yoruban individuals) and its impact on DAF. (**A**) Read-depth distribution for SNPs in neutral control (non-ENCODE) and ENCODE target regions. (**B**) DAF as a function of read depth, for non-ENCODE SNPs. DAF decreases with increasing depth, due to increasing sensitivity to detect rare variants; the reverse trend at depths above 300 likely reflects the presence of spurious "paralogue-collapse" SNPs .

Genome Biology and Evolution

SMBE

GBE

On the Immortality of Television Sets: "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE

Dan Graur[1,*], Yichen Zheng[1], Nicholas Price[1], Ricardo B.R. Azevedo[1], Rebecca A. Zufall[1], and Eran Elhaik[2]

[1]Department of Biology and Biochemistry, University of Houston

[2]Department of Mental Health, Johns Hopkins University Bloomberg School of Public Health

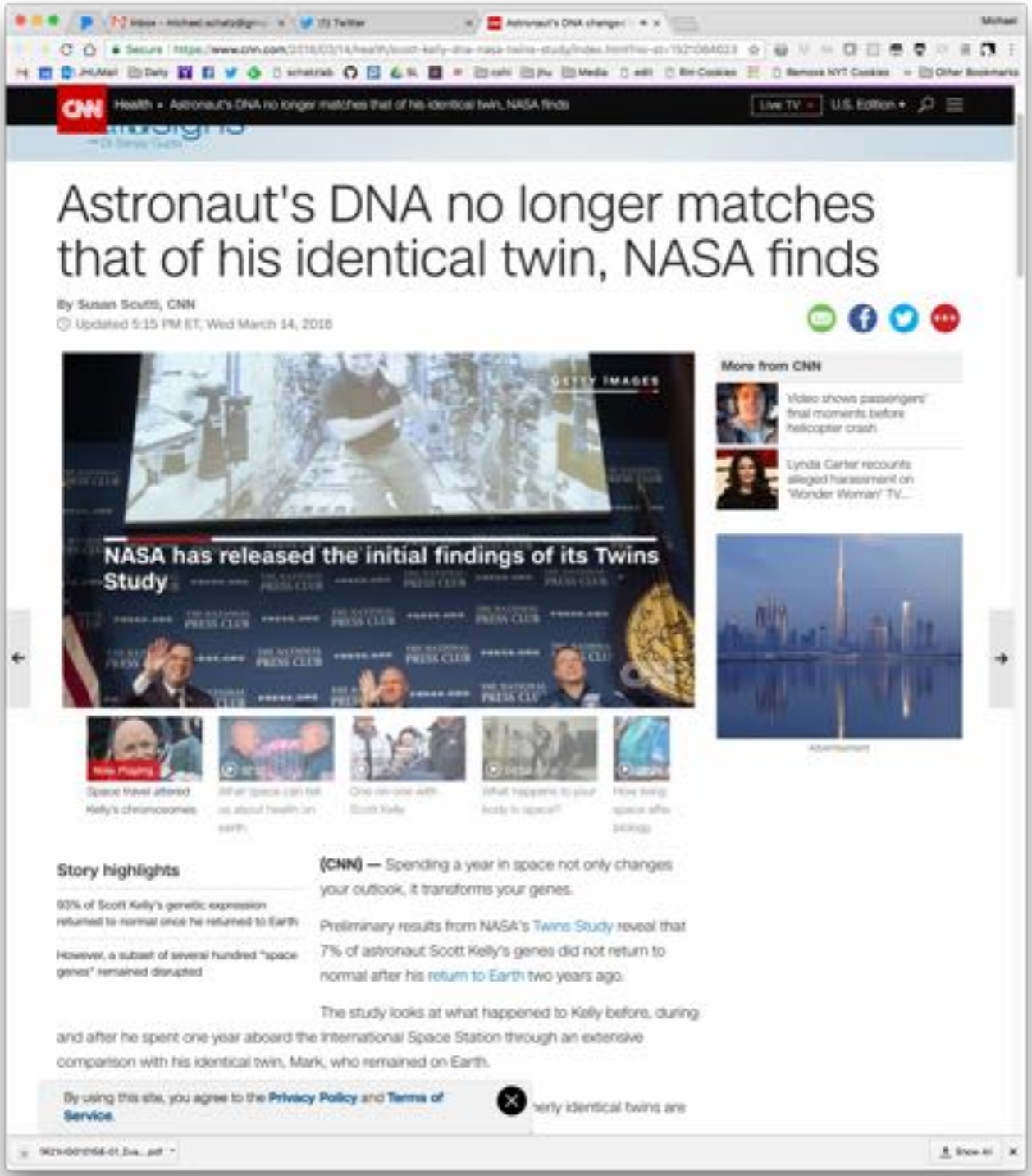
*Corresponding author: E-mail: dgraur@uh.edu.

Accepted: February 16, 2013

Abstract

A recent slew of ENCyclopedia Of DNA Elements (ENCODE) Consortium publications, specifically the article signed by all Consortium members, put forward the idea that more than 80% of the human genome is functional. This claim flies in the face of current estimates according to which the fraction of the genome that is evolutionarily conserved through purifying selection is less than 10%. Thus, according to the ENCODE Consortium, a biological function can be maintained indefinitely without selection, which implies that at least 80 – 10 = 70% of the genome is perfectly invulnerable to deleterious mutations, either because no mutation can ever occur in these "functional" regions or because no mutation in these regions can ever be deleterious. This absurd conclusion was reached through various means, chiefly by employing the seldom used "causal role" definition of biological function and then applying it inconsistently to different biochemical properties, by committing a logical fallacy known as "affirming the consequent," by failing to appreciate the crucial difference between "junk DNA" and "garbage DNA," by using analytical methods that yield biased errors and inflate estimates of functionality, by favoring statistical sensitivity over specificity, and by emphasizing statistical significance rather than the magnitude of the effect. Here, we detail the many logical and methodological transgressions involved in assigning functionality to almost every nucleotide in the human genome. The ENCODE results were predicted by one of its authors to necessitate the rewriting of textbooks. We agree, many textbooks dealing with marketing, mass-media hype, and public relations may well have to be rewritten.

Key words: junk DNA, genome functionality, selection, ENCODE project.


CNN
Health + Astronaut's DNA no longer matches that of his identical twin, NASA finds
Live TV
U.S. Edition
Astronaut's DNA no longer matches that of his identical twin, NASA finds
By Susan Scutti, CNN
Updated 5:15 PM ET, Wed March 14, 2018
GETTY IMAGES
NASA has released the initial findings of its Twins Study
More from CNN
Video shows passengers' final moments before helicopter crash
Lynda Carter recounts alleged harassment on 'Wonder Woman' TV...
Space travel altered Kelly's chromosomes
Story highlights
93% of Scott Kelly's genetic expression returned to normal once he returned to Earth
However, a subset of several hundred "space genes" remained disrupted
(CNN) — Spending a year in space not only changes your outlook, it transforms your genes.
Preliminary results from NASA's Twins Study reveal that 7% of astronaut Scott Kelly's genes did not return to normal after his return to Earth two years ago.
The study looks at what happened to Kelly before, during and after he spent one year aboard the International Space Station through an extensive comparison with his identical twin, Mark, who remained on Earth.
By using this site, you agree to the Privacy Policy and Terms of Service.

Essay

Current Biology Vol 23 No 7
R260

The ENCODE project: Missteps overshadowing a success

Two clichés of science journalism have now played out around the ENCODE project. ENCODE's publicity first presented a misleading "all the textbooks

*"To clarify what noise means, I propose the **Random Genome Project**. Suppose we put a few million bases of entirely random synthetic DNA into a human cell, and do an ENCODE project on it. Will it be reproducibly transcribed into mRNA-like transcripts, reproducibly bound by DNA-binding proteins, and reproducibly wrapped around histones marked by specific chromatin modifications? I think yes.*

A striking feature of genetic regulation is that regulatory factors (proteins or RNAs) generally recognize and bind to small sites, small enough that any given factor will find specific binding sites even in random DNA. Promoters, enhancers, splice sites, poly-A addition sites, and other functional features in the genome all have substantial random occurrence frequencies. These sites are not nonspecific in a random genome. They are specific sequences, albeit randomly occurring and not under selection for any function.

Would biochemical activities in the random genome be regulated under different conditions? For example, would they be cell type-specific? Surely yes, because the regulatory factors themselves (such as transcription factors) are regulated and expressed in specific cell types and conditions."

Essay

Current Biology Vol 23 No 7
R260

The ENCODE project: Missteps overshadowing a success

***"There are three categories of big science: the big experiment, the map, and the leading wedge**. A big experiment is driven by a single question or hypothesis test, but requires a large scale community investment. […] A map is a data resource — comprehensive, complete, closed ended — to be used by multiple groups, over a long time, for multiple purposes. The decision to build a map is a cost/benefit calculation, weighed against individual labs who are already making piecemeal maps in an ill coordinated fashion, especially when small groups lack technical expertise to make the map well. A failure mode with a map is to miscalculate the cost/benefit analysis and make a map that too few individual labs will use.*

ENCODE and some of its critics have fallen into similar traps. In trying to make the result sound important, ENCODE's publicity spun it retrospectively as a hypothesis test, but ENCODE was not designed to test anything. ENCODE is a map: it should have been published and defended as such. And while its critics argue over an interpretation that wasn't in ENCODE's mission to begin with, ENCODE's planners should also recognize that as ENCODE now moves into a new funding phase, it may be headed for a failure mode in its actual mission. The cost/benefit calculation is rapidly changing. ENCODE's technologies (all based on high throughput sequencing) are now widely and inexpensively available in individual labs.

Bruce Alberts is Editor-in-Chief of *Science*.

EDITORIAL

The End of "Small Science"?

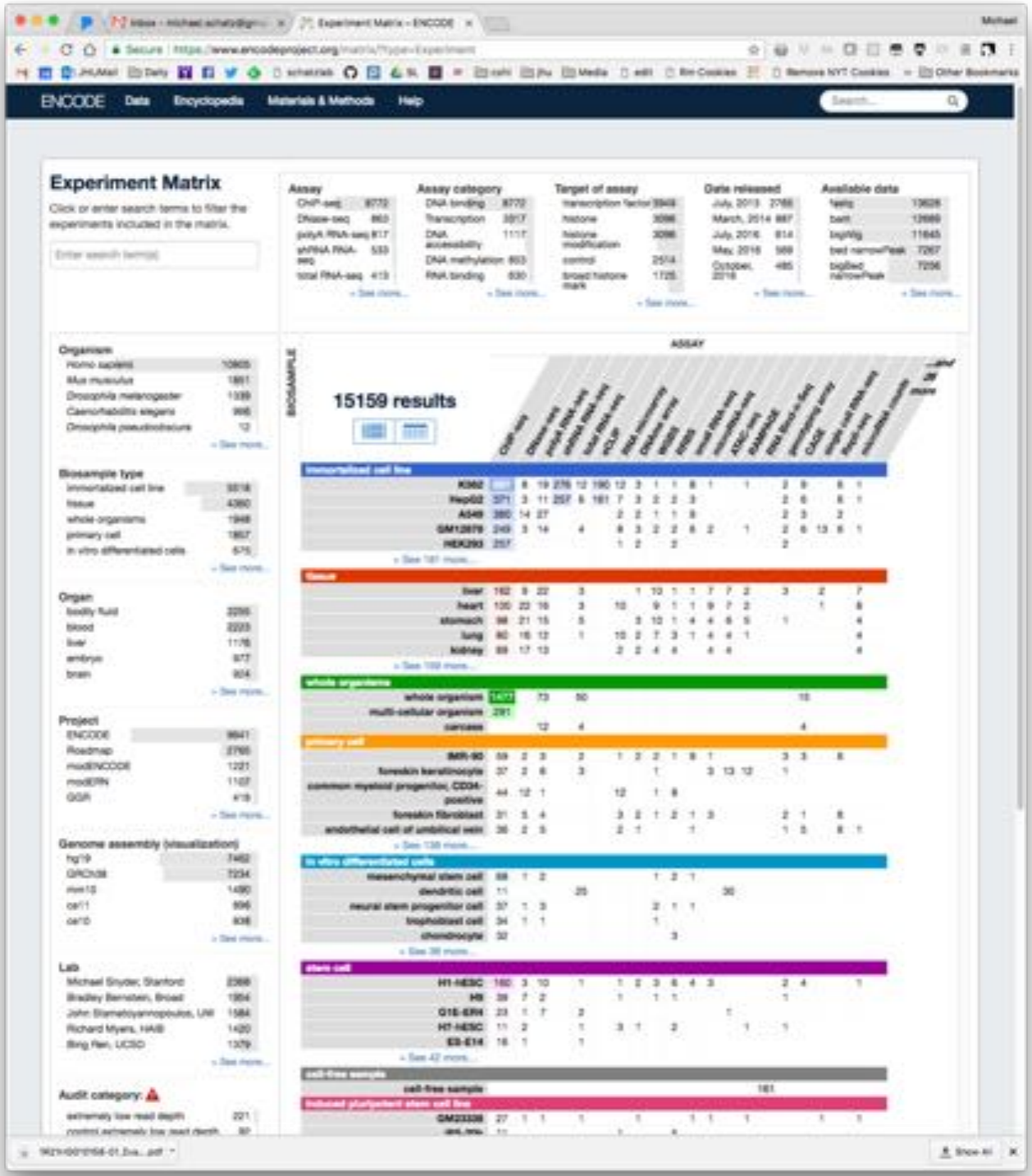
I AM PROMPTED TO WRITE THIS EDITORIAL BY THE RELEASE OF 30 PAPERS THIS MONTH FROM THE ENCODE Project Consortium. This decade-long project involved an international team of 442 scientists who have compiled what is being called an "encyclopedia of DNA elements," a comprehensive list of functional elements in the human genome. The detailed overview is expected to spur further research on the fundamentals of life, health, and disease. ENCODE exemplifies a "big-science" style of research that continues to sweep the headlines, and the increased efficiency of data production by such projects is impressive. Does this mean that the highly successful "small-science" era of biological research will soon be over? Will government funding increasingly favor big-science projects? I certainly hope that the answer is no.

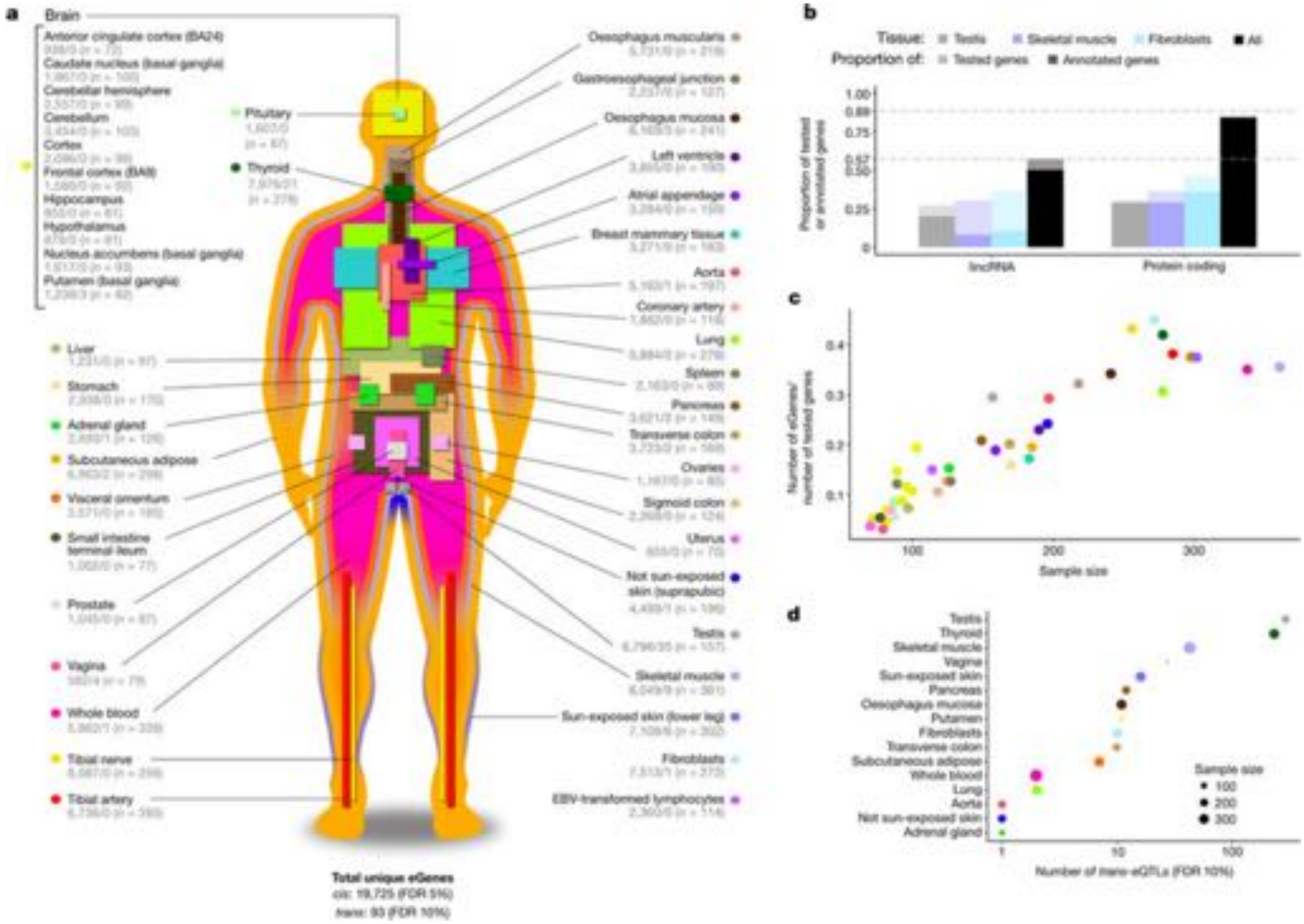
...

Each year, the amount of factual information that scientists acquire about cells increases and, stimulated by -omics projects, the compilations of data expand at a tremendous rate. But the grand challenges that remain in attaining a deep understanding of the chemistry of life will require going beyond detailed catalogs. Ensuring a successful future for the biological sciences will require restraint in the growth of large centers and -omics–like projects, so as to provide more financial support for the critical work of innovative small laboratories striving to understand the wonderful complexity of living systems.

– Bruce Alberts

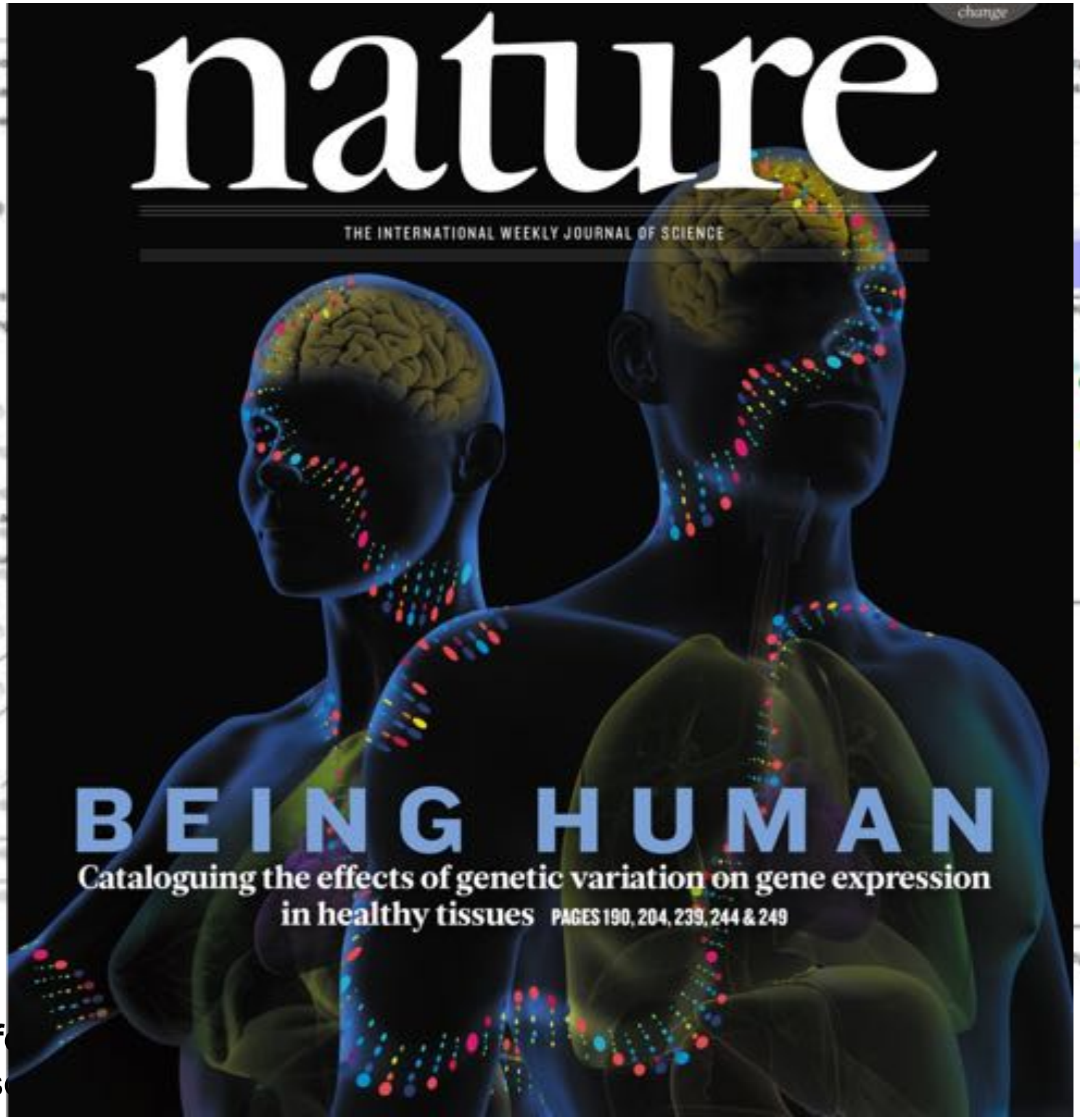
10.1126/science.1230529


ENCODE
Data
Encyclopedia
Materials & Methods
Help
Experiment Matrix
Click or enter search terms to filter the experiments included in the matrix.
Assay
Assay category
Target of assay
Data released
Available data
Organism
Biosample type
Organism
Project
Genome assembly (visualization)
Lab
Audit category:
15159 results
ASSAY
BIOSAMPLE
immortalized cell line
tissue
whole organisms
primary cell
in vitro differentiated cells
stem cell
cell-free sample
induced pluripotent stem cell line



Genetic effects on gene expression across human tissues

GTEx Consortium (2017) Nature. doi:10.1038/nature24277



Genetic eff
GTEx Cons