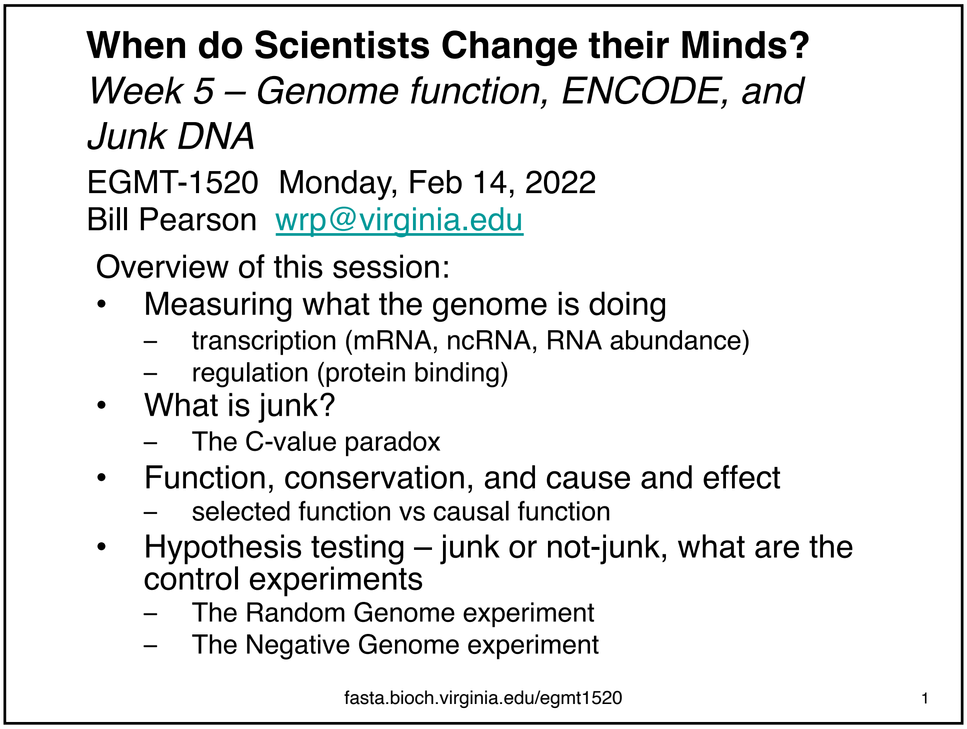


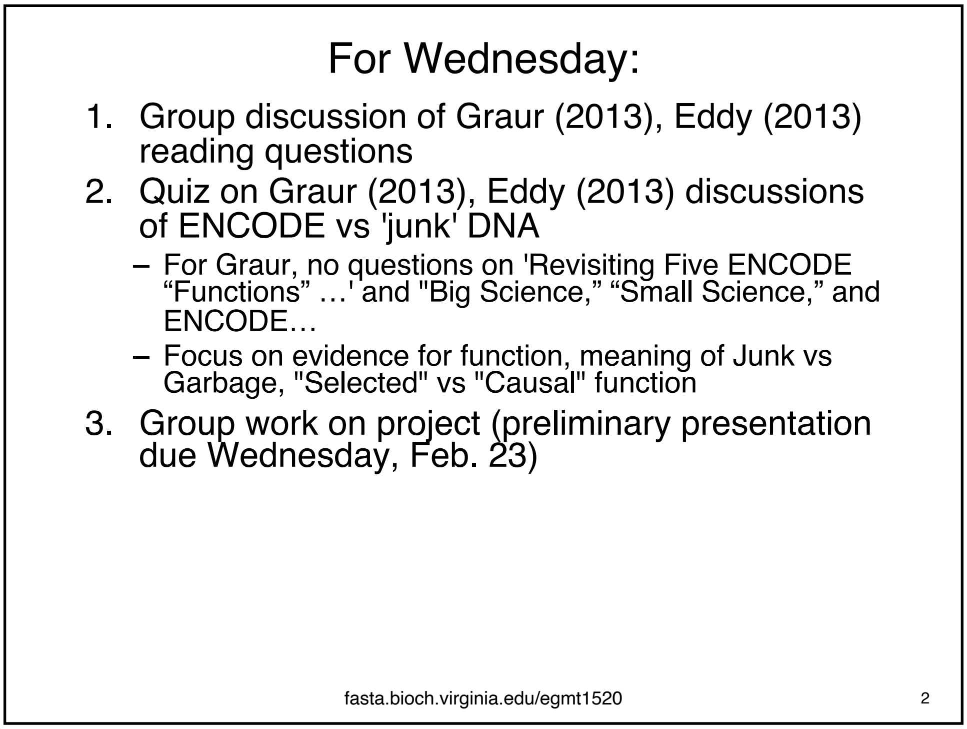
When do Scientists Change their Minds?

Week 5 – Genome function, ENCODE, and Junk DNA

EGMT-1520 Monday, Feb 14, 2022
Bill Pearson wrp@virginia.edu

Overview of this session:

- Measuring what the genome is doing
  - transcription (mRNA, ncRNA, RNA abundance)
  - regulation (protein binding)
- What is junk?
  - The C-value paradox
- Function, conservation, and cause and effect
  - selected function vs causal function
- Hypothesis testing – junk or not-junk, what are the control experiments
  - The Random Genome experiment
  - The Negative Genome experiment

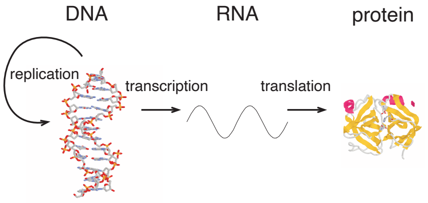


For Wednesday:

1. Group discussion of Graur (2013), Eddy (2013) reading questions
2. Quiz on Graur (2013), Eddy (2013) discussions of ENCODE vs 'junk' DNA
   - For Graur, no questions on 'Revisiting Five ENCODE “Functions” …' and "Big Science,” “Small Science,” and ENCODE…
   - Focus on evidence for function, meaning of Junk vs Garbage, "Selected" vs "Causal" function
3. Group work on project (preliminary presentation due Wednesday, Feb. 23)

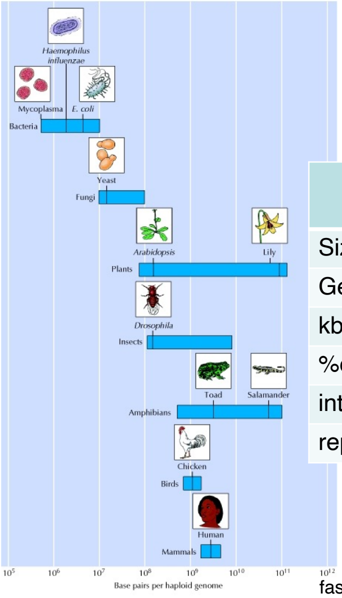
What does a (eukaryotic) gene look like?

- Central dogma:



- Parts of a gene:
  - start of transcription (beginning of the mRNA)
  - start of translation (beginning of the protein)
  - end of translation (end of the protein)
  - poly-A addition site (end of processed mRNA)
  - end of transcription

What is in a genome?



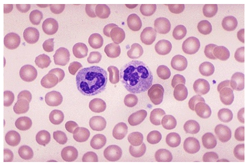
Cooper, GM (2000) The Cell: A Molecular Approach. 2nd edition. Fig 4.1

	E. col	Plas.	Yeast	Plant (ARATH)	Homo
Size(Mb)	4.64	22.8	12.5	115	3289
Genes	4288	5268	5770	25.5K	~25K
kb/Gene	0.95	4.34	2.09	4.53	27
%coding	87.8	52.6	70.5	28.8	1.3
introns	0	7406	272	107K	53K
repeat%	<1	<1	2.4	15	46

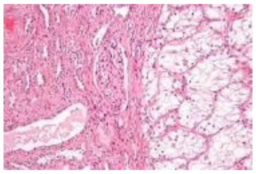
Pevsner, Table 16-1

All cells have the same genome (DNA)
Cells in different tissues are different

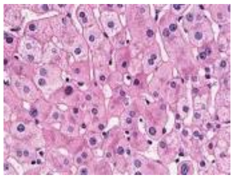
red blood cell



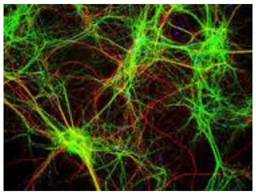
kidney



liver

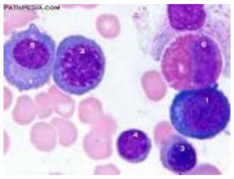


brain



muscle

lymphocyte

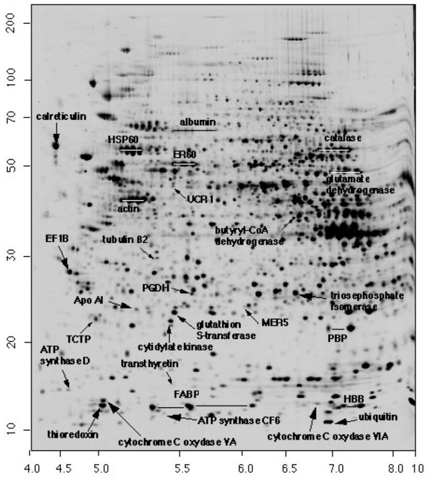


because they express different proteins from different mRNAs

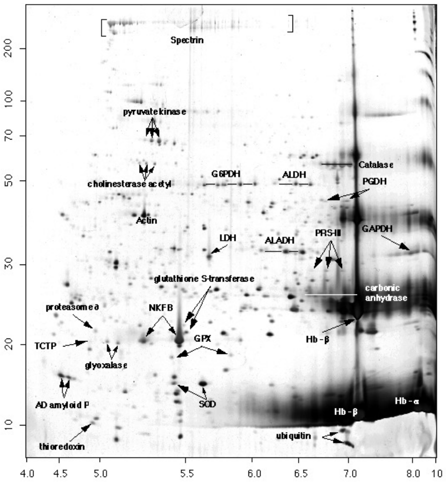
5

Cells in different tissues are different

human HEPG2 (liver cell line)



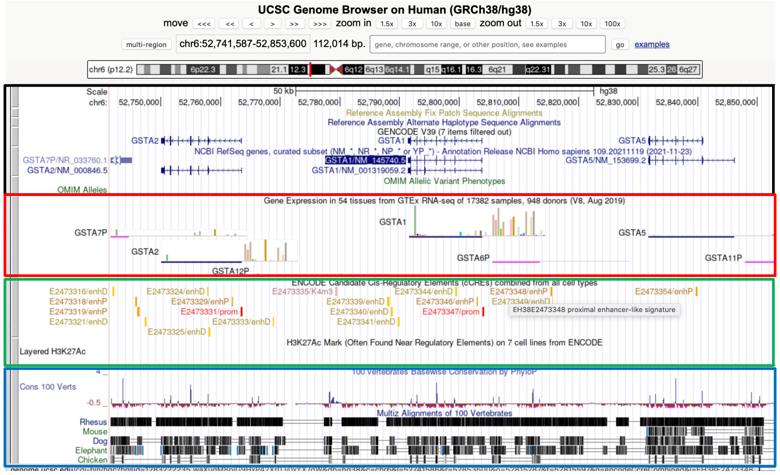
human red blood cell



because they express different proteins from different mRNAs

6

All cells have the same DNA
But that DNA is used differently in different cells


UCSC Genome Browser on Human (GRCh38/hg38)

chr6:52,741,587-52,853,600 112,014 bp.

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
multi-region
gene, chromosome range, or other position, see examples
go
examples
chr6 (p12.2) 6p22.3 21.1 12.3 6q12 6q13 6q14.1 q15 q16.1 16.3 6q21 q22.31 25.3 26 6q27
Scale 50 kb hg38
chr6: 52,750,000 52,760,000 52,770,000 52,780,000 52,790,000 52,800,000 52,810,000 52,820,000 52,830,000 52,840,000 52,850,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V39 (7 items filtered out)
GSTA2
GSTA1
GSTA5
NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release NCBI Homo sapiens 109.20211119 (2021-11-23)
GSTA7P/NR_033760.1
GSTA2/NM_000846.5
GSTA1/NM_145740.5
GSTA1/NM_001319059.2
GSTA5/NM_153699.2
OMIM Allelic Variant Phenotypes
OMIM Alleles
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)
GSTA7P
GSTA2
GSTA12P
GSTA1
GSTA6P
GSTA5
GSTA11P
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
E2473316/enhD
E2473318/enhP
E2473319/enhP
E2473321/enhD
E2473324/enhD
E2473329/enhP
E2473331/prom
E2473325/enhD
E2473333/enhD
E2473335/K4m3
E2473339/enhD
E2473340/enhD
E2473341/enhD
E2473344/enhD
E2473346/enhP
E2473347/prom
E2473348/enhP
E2473349/enhD
EH38E2473348 proximal enhancer-like signature
E2473354/enhP
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
100 Vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
4
-0.5
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken

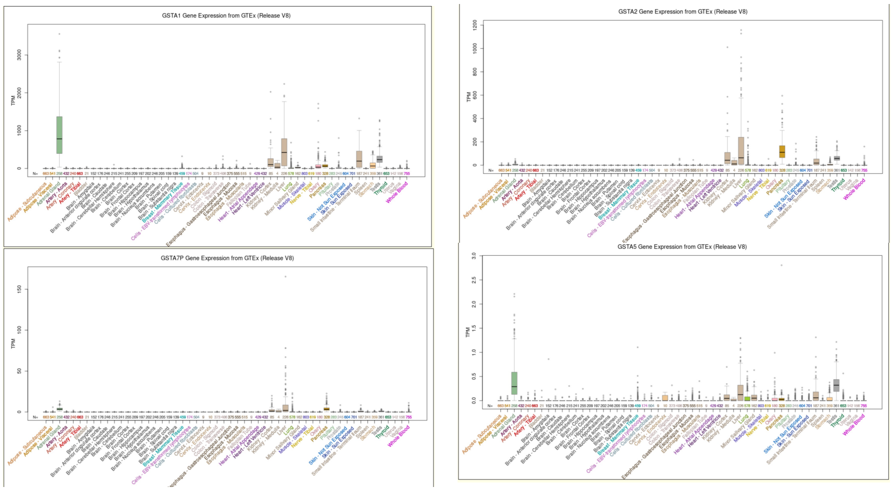
Structure

Expression

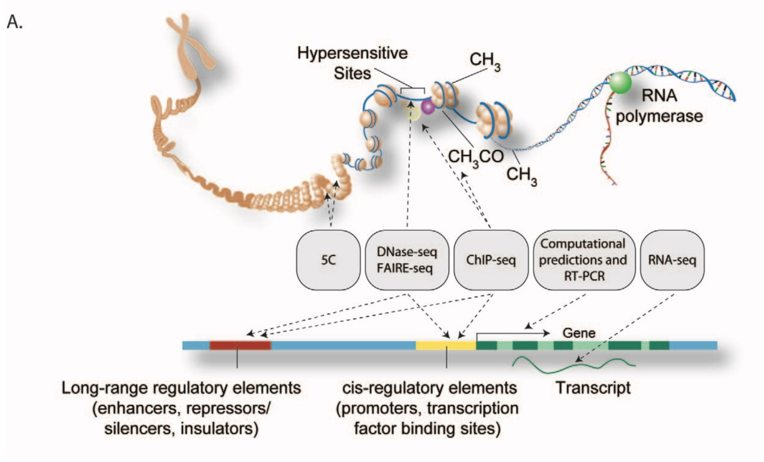
Binding

Conservation

Different gene expression in different cells



The Encyclopedia of DNA Elements (ENCODE)


A.
Hypersensitive Sites
CH_3
RNA polymerase
CH_3CO
CH_3
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

PLoS Biol (2011 **9:**e1001046). ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9,** e1001046 (2011).

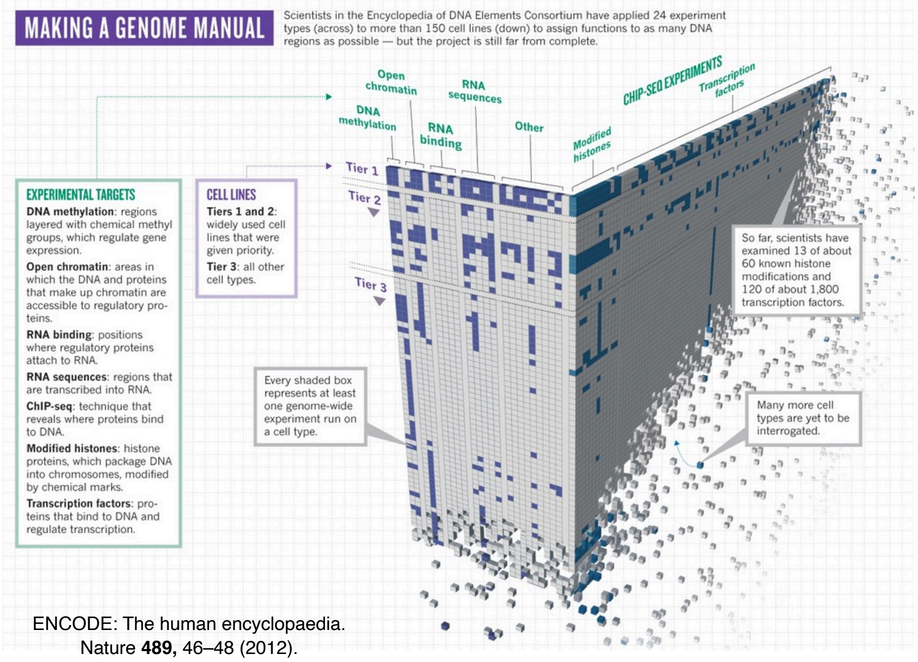
ENCODE measures of function

Measurements (assays):

- RNA Expression (RNAseq, CAGE, RNA-PET)
- Chromatin accessibility
  - DNase hypersensitivity
  - FAIRE
- Transcription factory binding (ChIP-seq)
- CpG methylation
- Histone modification

Data sources: Cell lines (immortalized cells grown in culture)

- Tier1: erythroleukaemia cells, B-lymphoblastoid, human embryonic stem cells
- Tier 2: HeLa, HepG2, HUVECs



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.

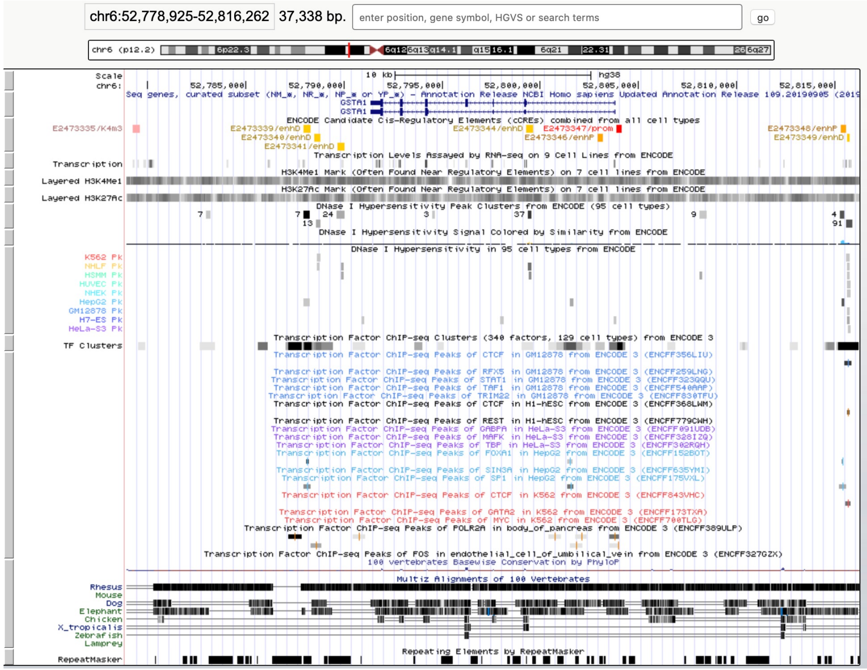
Many more cell types are yet to be interrogated.

ENCODE: The human encyclopaedia.
Nature **489,** 46–48 (2012).

ENCODE measures of function

Measurements : 80.4% of genome "functional"

- RNA Expression:
  - 62% of genome "expressed" as RNAseq or Gencode exons
  - 5.5% of total RNA in Gencode exons
  - 31% outside of annotated genes (mostly introns)
- Transcription factor binding (ChIP-seq)
  - 119 DNA binding proteins, 72 cell types
  - 231 Mb (8.1%) of genome bound to binding proteins in all cell types

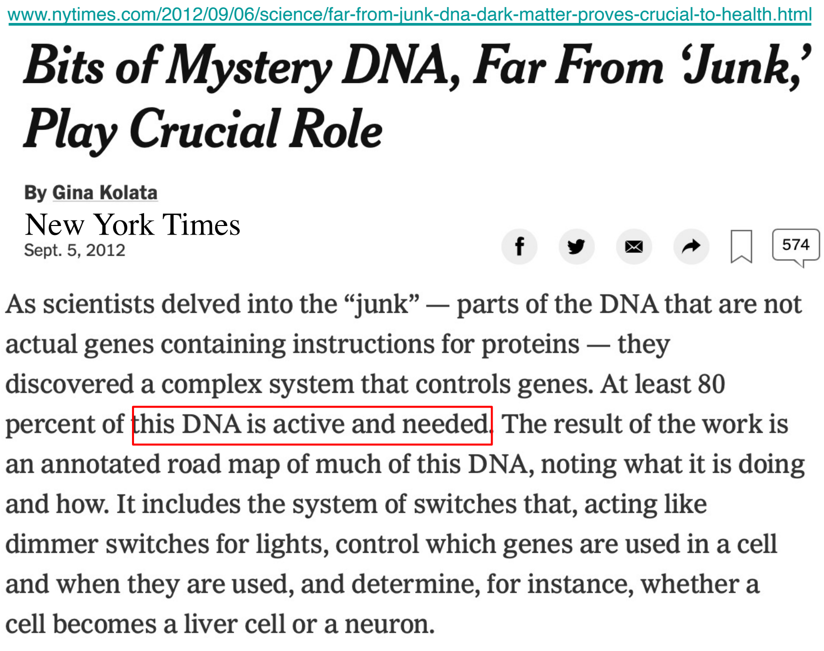


chr6:52,778,925-52,816,262 37,338 bp.

enter position, gene symbol, HGVS or search terms
go
chr6 (p12.2)
6p22.3
6q12
6q13
q14.1
q15
16.1
6q21
22.31
26
6q27
Scale
chr6:
10 kb
hg38
52,785,000|
52,790,000|
52,795,000|
52,800,000|
52,805,000|
52,810,000|
52,815,000|
Seq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release NCBI Homo sapiens Updated Annotation Release 109.20190905 (2019
GSTA1
GSTA1
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
E2473335/K4m3
E2473339/enhD
E2473340/enhD
E2473341/enhD
E2473344/enhD
E2473347/prom
E2473346/enhP
E2473348/enhP
E2473349/enhD
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
Transcription
H3K4Me1 Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K4Me1
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
7
7
24
13
3
37
9
4
91
DNase I Hypersensitivity Signal Colored by Similarity from ENCODE
DNase I Hypersensitivity in 95 cell types from ENCODE
K562 Pk
NHLF Pk
HSMM Pk
HUVEC Pk
NHEK Pk
HepG2 Pk
GM12878 Pk
H7-ES Pk
HeLa-S3 Pk
Transcription Factor ChIP-seq Clusters (340 factors, 129 cell types) from ENCODE 3
TF Clusters
Transcription Factor ChIP-seq Peaks of CTCF in GM12878 from ENCODE 3 (ENCFF356LIU)
Transcription Factor ChIP-seq Peaks of RFX5 in GM12878 from ENCODE 3 (ENCFF259LNG)
Transcription Factor ChIP-seq Peaks of STAT1 in GM12878 from ENCODE 3 (ENCFF323QQU)
Transcription Factor ChIP-seq Peaks of TAF1 in GM12878 from ENCODE 3 (ENCFF540AAP)
Transcription Factor ChIP-seq Peaks of TRIM22 in GM12878 from ENCODE 3 (ENCFF830TFU)
Transcription Factor ChIP-seq Peaks of CTCF in H1-hESC from ENCODE 3 (ENCFF368LWM)
Transcription Factor ChIP-seq Peaks of REST in H1-hESC from ENCODE 3 (ENCFF779CWH)
Transcription Factor ChIP-seq Peaks of GABPA in HeLa-S3 from ENCODE 3 (ENCFF091UDB)
Transcription Factor ChIP-seq Peaks of MAFK in HeLa-S3 from ENCODE 3 (ENCFF328IZQ)
Transcription Factor ChIP-seq Peaks of TBP in HeLa-S3 from ENCODE 3 (ENCFF302RQH)
Transcription Factor ChIP-seq Peaks of FOXA1 in HepG2 from ENCODE 3 (ENCFF152BOT)
Transcription Factor ChIP-seq Peaks of SIN3A in HepG2 from ENCODE 3 (ENCFF635YMI)
Transcription Factor ChIP-seq Peaks of SP1 in HepG2 from ENCODE 3 (ENCFF175VXL)
Transcription Factor ChIP-seq Peaks of CTCF in K562 from ENCODE 3 (ENCFF843VHC)
Transcription Factor ChIP-seq Peaks of GATA2 in K562 from ENCODE 3 (ENCFF173TXA)
Transcription Factor ChIP-seq Peaks of MYC in K562 from ENCODE 3 (ENCFF700TLG)
Transcription Factor ChIP-seq Peaks of POLR2A in body_of_pancreas from ENCODE 3 (ENCFF389ULP)
Transcription Factor ChIP-seq Peaks of FOS in endothelial_cell_of_umbilical_vein from ENCODE 3 (ENCFF327GZX)
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Repeating Elements by RepeatMasker
RepeatMasker

fasta.bioch.virginia.edu/egmt1520 13

13



www.nytimes.com/2012/09/06/science/far-from-junk-dna-dark-matter-proves-crucial-to-health.html

Bits of Mystery DNA, Far From 'Junk,' Play Crucial Role

By Gina Kolata

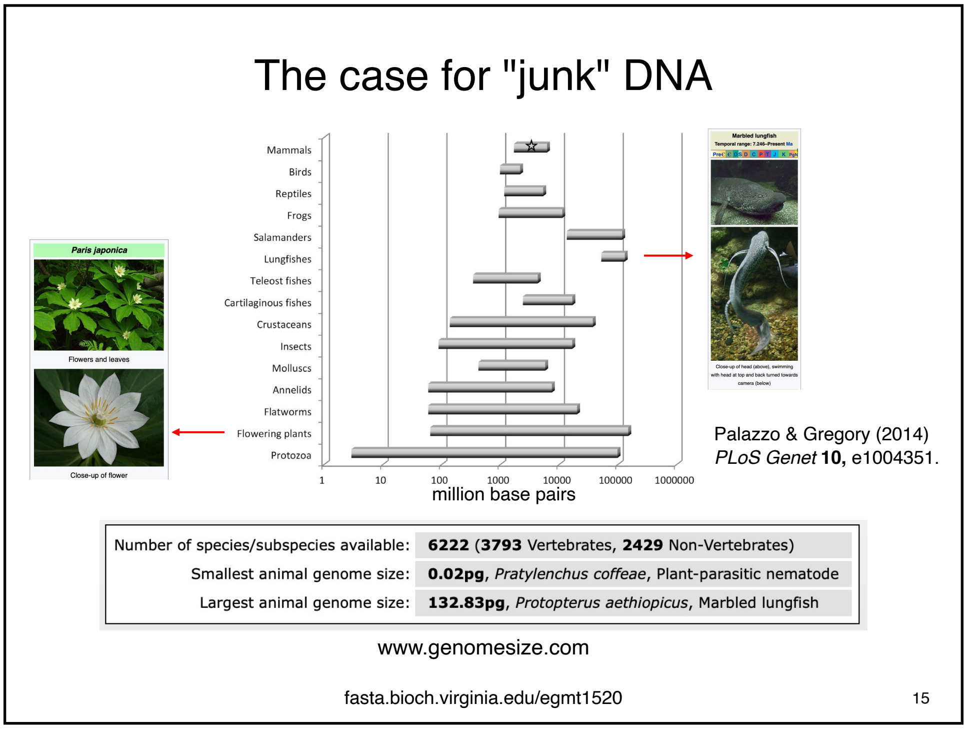
New York Times

Sept. 5, 2012

As scientists delved into the "junk" — parts of the DNA that are not actual genes containing instructions for proteins — they discovered a complex system that controls genes. At least 80 percent of this DNA is active and needed. The result of the work is an annotated road map of much of this DNA, noting what it is doing and how. It includes the system of switches that, acting like dimmer switches for lights, control which genes are used in a cell and when they are used, and determine, for instance, whether a cell becomes a liver cell or a neuron.

fasta.bioch.virginia.edu/egmt1520 14

14



The case for "junk" DNA

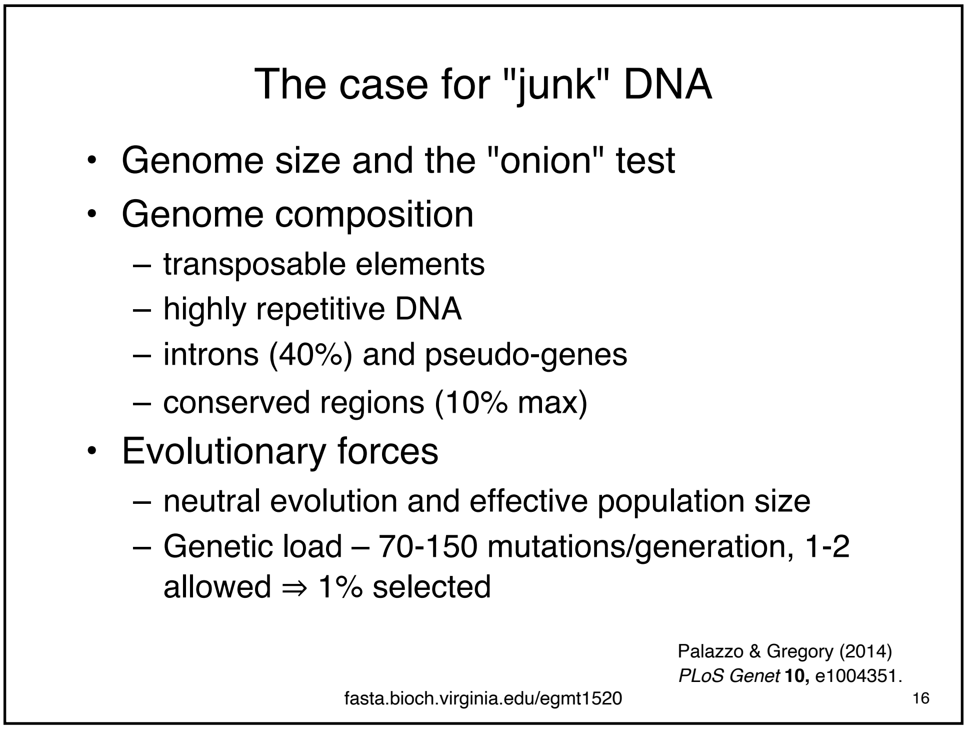
Palazzo & Gregory (2014) *PLoS Genet* **10,** e1004351.

Number of species/subspecies available:	**6222** (**3793** Vertebrates, **2429** Non-Vertebrates)
Smallest animal genome size:	**0.02pg**, *Pratylenchus coffeae*, Plant-parasitic nematode
Largest animal genome size:	**132.83pg**, *Protopterus aethiopicus*, Marbled lungfish

www.genomesize.com

fasta.bioch.virginia.edu/egmt1520 15

15



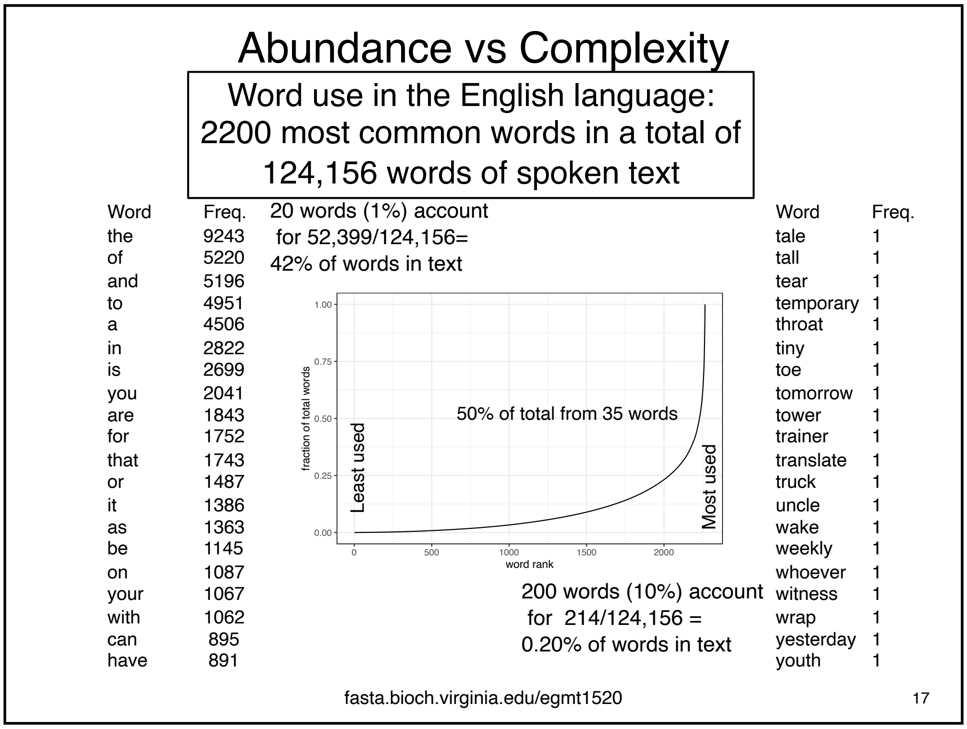
The case for "junk" DNA

- Genome size and the "onion" test
- Genome composition
  - transposable elements
  - highly repetitive DNA
  - introns (40%) and pseudo-genes
  - conserved regions (10% max)
- Evolutionary forces
  - neutral evolution and effective population size
  - Genetic load – 70-150 mutations/generation, 1-2 allowed ⇒ 1% selected

Palazzo & Gregory (2014) *PLoS Genet* **10,** e1004351.

fasta.bioch.virginia.edu/egmt1520 16

16



Abundance vs Complexity

Word use in the English language:
2200 most common words in a total of
124,156 words of spoken text

Word	Freq.
the	9243
of	5220
and	5196
to	4951
a	4506
in	2822
is	2699
you	2041
are	1843
for	1752
that	1743
or	1487
it	1386
as	1363
be	1145
on	1087
your	1067
with	1062
can	895
have	891

20 words (1%) account
for 52,399/124,156=
42% of words in text

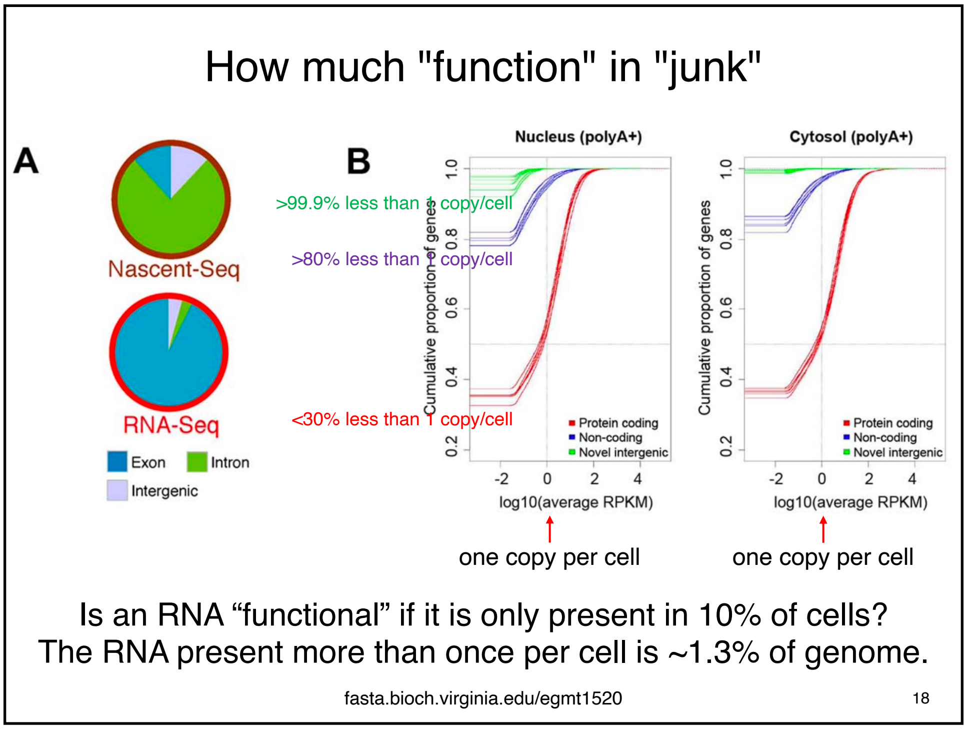
50% of total from 35 words

200 words (10%) account
for 214/124,156 =
0.20% of words in text

Word	Freq.
tale	1
tall	1
tear	1
temporary	1
throat	1
tiny	1
toe	1
tomorrow	1
tower	1
trainer	1
translate	1
truck	1
uncle	1
wake	1
weekly	1
whoever	1
witness	1
wrap	1
yesterday	1
youth	1

fasta.bioch.virginia.edu/egmt1520 17

17



How much "function" in "junk"

>99.9% less than 1 copy/cell

>80% less than 1 copy/cell

<30% less than 1 copy/cell

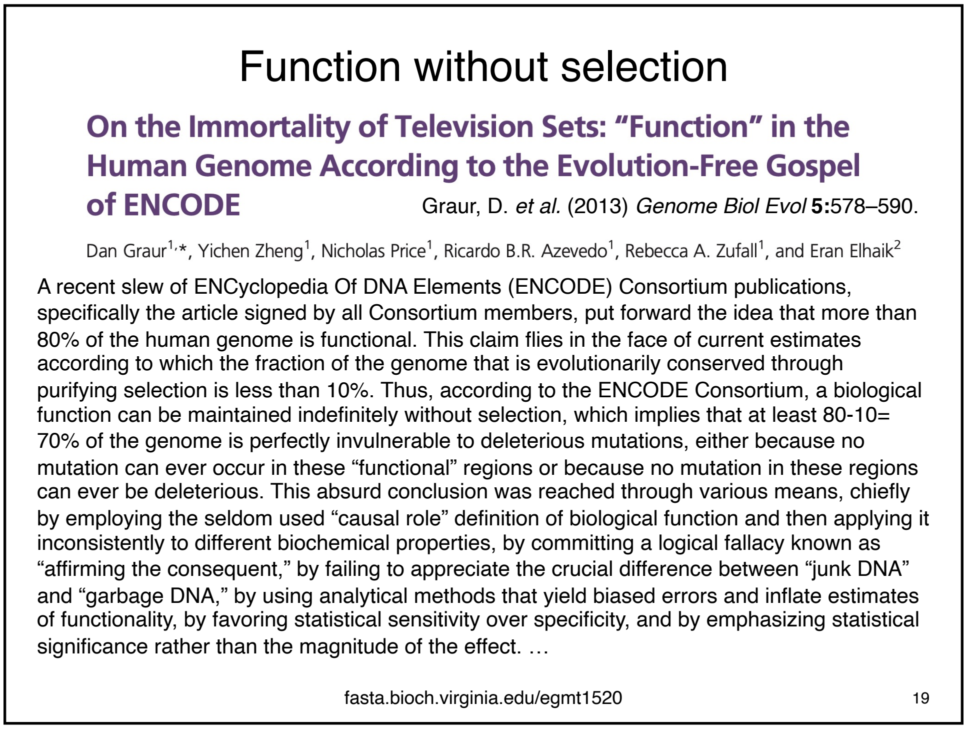
one copy per cell

one copy per cell

Is an RNA “functional” if it is only present in 10% of cells?
The RNA present more than once per cell is ~1.3% of genome.

fasta.bioch.virginia.edu/egmt1520 18

18



Function without selection

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

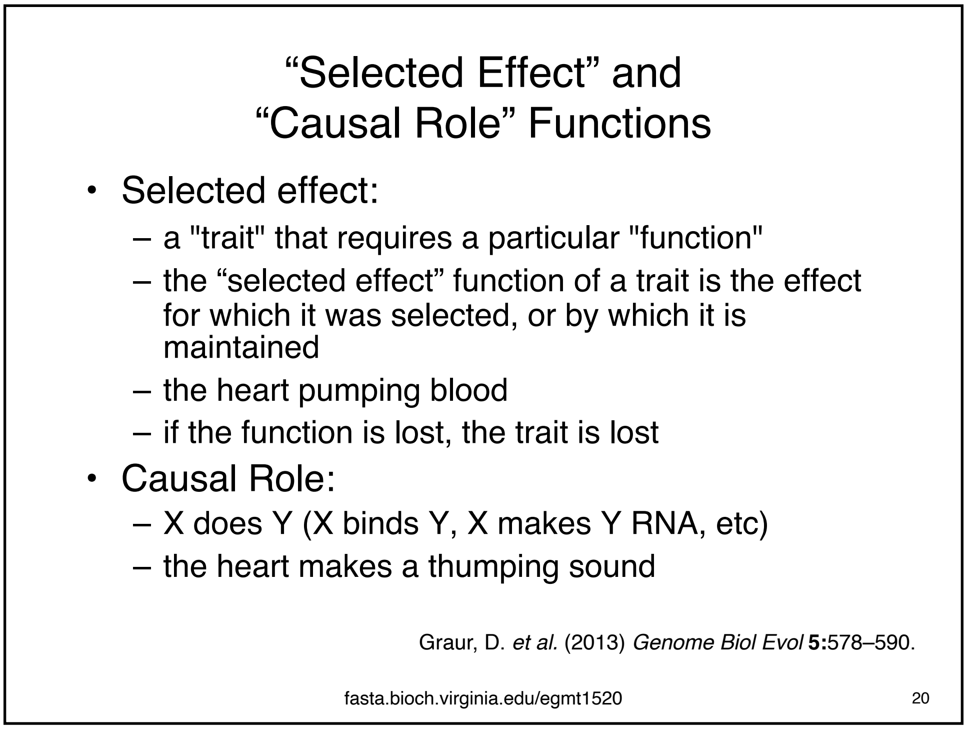
Graur, D. *et al.* (2013) *Genome Biol Evol* **5:**578–590.

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

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. …

fasta.bioch.virginia.edu/egmt1520 19

19



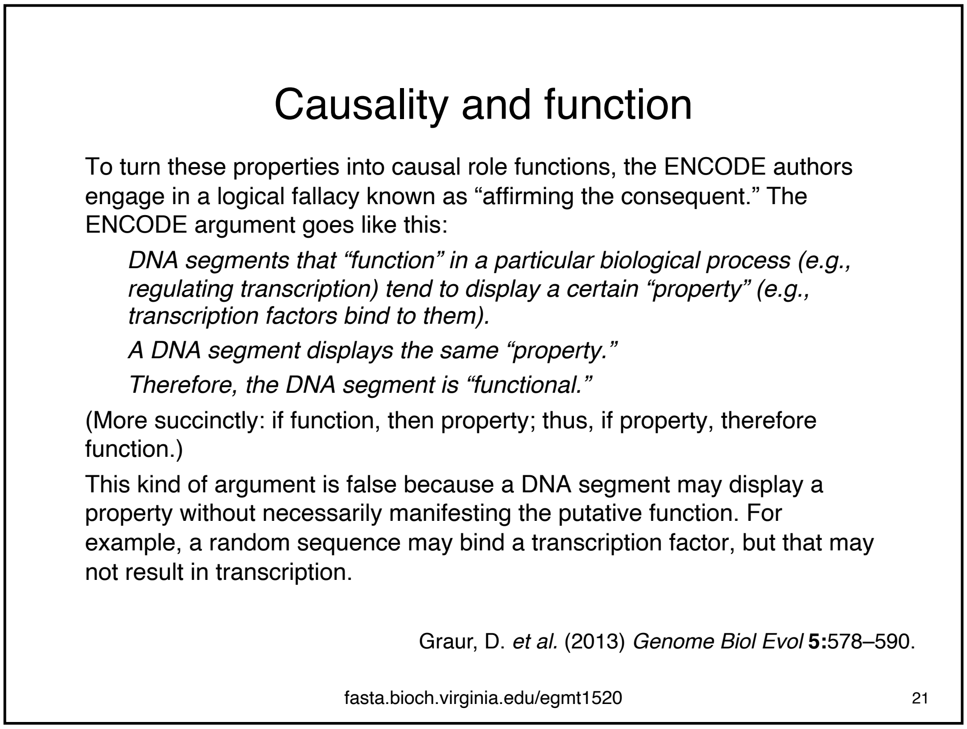
"Selected Effect" and "Causal Role" Functions

- Selected effect:
  - a "trait" that requires a particular "function"
  - the "selected effect" function of a trait is the effect for which it was selected, or by which it is maintained
  - the heart pumping blood
  - if the function is lost, the trait is lost
- Causal Role:
  - X does Y (X binds Y, X makes Y RNA, etc)
  - the heart makes a thumping sound

Graur, D. *et al.* (2013) *Genome Biol Evol* **5:**578–590.

fasta.bioch.virginia.edu/egmt1520 20

20



Causality and function

To turn these properties into causal role functions, the ENCODE authors engage in a logical fallacy known as “affirming the consequent.” The ENCODE argument goes like this:

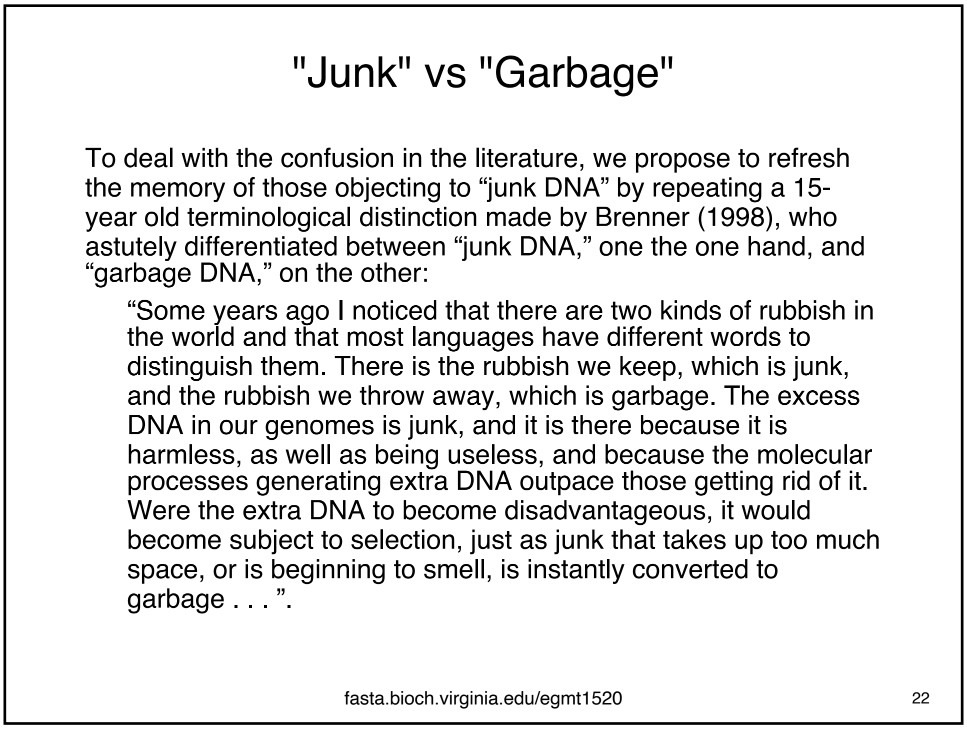
> *DNA segments that “function” in a particular biological process (e.g., regulating transcription) tend to display a certain “property” (e.g., transcription factors bind to them).*
>
> *A DNA segment displays the same “property.”*
>
> *Therefore, the DNA segment is “functional.”*

(More succinctly: if function, then property; thus, if property, therefore function.)

This kind of argument is false because a DNA segment may display a property without necessarily manifesting the putative function. For example, a random sequence may bind a transcription factor, but that may not result in transcription.

Graur, D. *et al.* (2013) *Genome Biol Evol* **5:**578–590.

21



"Junk" vs "Garbage"

To deal with the confusion in the literature, we propose to refresh the memory of those objecting to “junk DNA” by repeating a 15-year old terminological distinction made by Brenner (1998), who astutely differentiated between “junk DNA,” one the one hand, and “garbage DNA,” on the other:

> “Some years ago I noticed that there are two kinds of rubbish in the world and that most languages have different words to distinguish them. There is the rubbish we keep, which is junk, and the rubbish we throw away, which is garbage. The excess DNA in our genomes is junk, and it is there because it is harmless, as well as being useless, and because the molecular processes generating extra DNA outpace those getting rid of it. Were the extra DNA to become disadvantageous, it would become subject to selection, just as junk that takes up too much space, or is beginning to smell, is instantly converted to garbage . . . ”.

22

Meanings of (biological) function

Trying to conceptualize the forces that act on genome evolution is not just a matter of semantics. We can envision the human genome as a perfectly honed machine, or we can think of it as a wild landscape littered and layered with successions of decomposing molecular replicators, like dead weeds decaying into fertile soil. How much DNA does it take to *design* a human? How much DNA does it take to *evolve* a human?

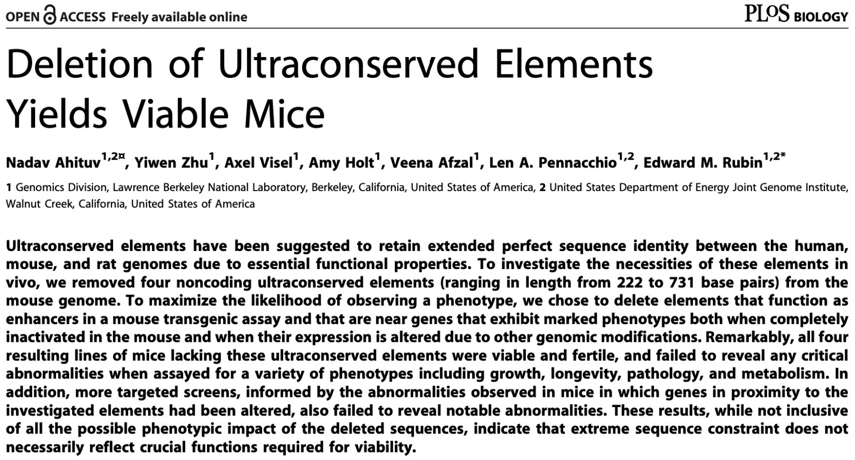
They are not the same question, and the gap between them is where we seek an understanding of genome evolution.

Eddy, S. R. (2013) *Curr Biol* **23,** R259–61

Different meanings of (biological) function

- The Random Genome Project (the negative control) Eddy, S. R. (2013) *Curr Biol* **23,** R259–61
  - synthesize random DNA sequence
  - add to a human (cell-line) genome
  - measure transcription, chromatin accessibility, transcription-factor (TF) binding
  - look for reproducible cell-type specific differences
- The negative genome project
  - look for regions of the genome that are highly conserved Ahituv, N. *et al.* (2007) *PLoS Biol* 5**:**e234.
  - delete regions (in mice) and look for effect

The negative genome project



OPEN ACCESS Freely available online PLoS BIOLOGY

Deletion of Ultraconserved Elements Yields Viable Mice

Nadav Ahituv[1,2¤], Yiwen Zhu[1], Axel Visel[1], Amy Holt[1], Veena Afzal[1], Len A. Pennacchio[1,2], Edward M. Rubin[1,2*]

[1] Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, [2] United States Department of Energy Joint Genome Institute, Walnut Creek, California, United States of America

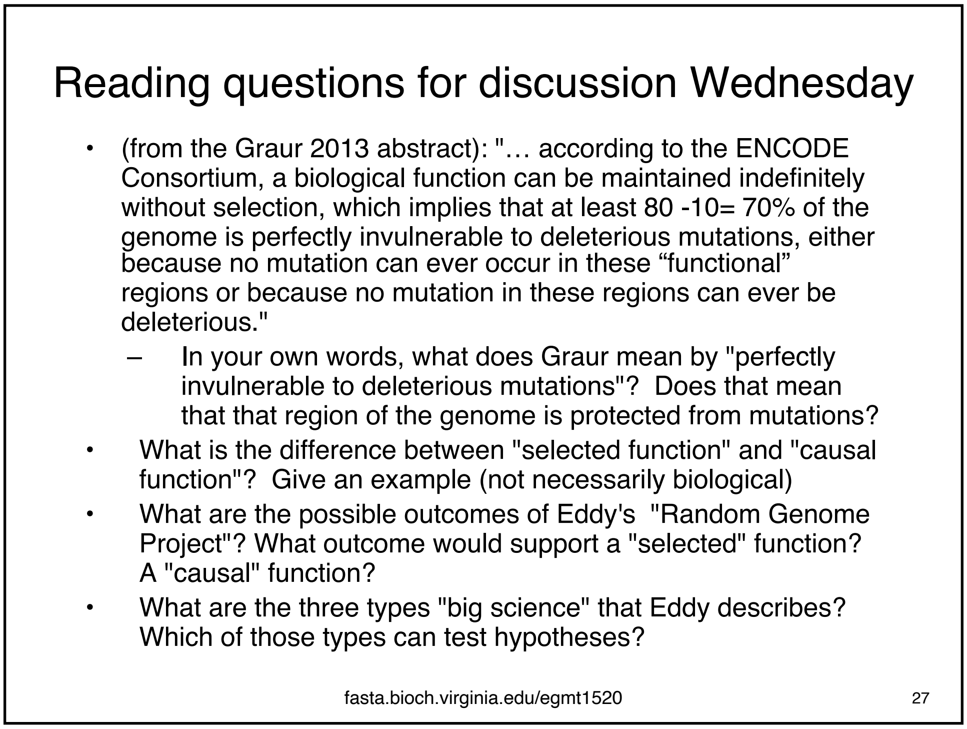
Ultraconserved elements have been suggested to retain extended perfect sequence identity between the human, mouse, and rat genomes due to essential functional properties. To investigate the necessities of these elements in vivo, we removed four noncoding ultraconserved elements (ranging in length from 222 to 731 base pairs) from the mouse genome. To maximize the likelihood of observing a phenotype, we chose to delete elements that function as enhancers in a mouse transgenic assay and that are near genes that exhibit marked phenotypes both when completely inactivated in the mouse and when their expression is altered due to other genomic modifications. Remarkably, all four resulting lines of mice lacking these ultraconserved elements were viable and fertile, and failed to reveal any critical abnormalities when assayed for a variety of phenotypes including growth, longevity, pathology, and metabolism. In addition, more targeted screens, informed by the abnormalities observed in mice in which genes in proximity to the investigated elements had been altered, also failed to reveal notable abnormalities. These results, while not inclusive of all the possible phenotypic impact of the deleted sequences, indicate that extreme sequence constraint does not necessarily reflect crucial functions required for viability.

80% Functional DNA? Pro and Con

- Pros:
  - A large fraction of the human genome is transcribed
  - There are millions of sites where transcription factors bind reproducibly in different cell types
  - There are tens of thousands of different cell types, whatever number we measure must be an underestimate
  - evolutionary selection can easily be missed if the sequences are short (TF binding motifs are conserved, but not detectable)
- Cons:
  - Genome size is not constrained among closely related species (e.g. higher plants, fish) – the C-value paradox
  - Most transcription is either of introns (which can vary dramatically in size and number for similar organisms) or very low abundance (< 1 copy per cell, transcriptional noise)
  - reproducible binding does not guarantee function

Selection vs Neutrality

Is there a "reason" ("function") for all that stuff?



Reading questions for discussion Wednesday

- (from the Graur 2013 abstract): "… 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."
  - In your own words, what does Graur mean by "perfectly invulnerable to deleterious mutations"? Does that mean that that region of the genome is protected from mutations?
- What is the difference between "selected function" and "causal function"? Give an example (not necessarily biological)
- What are the possible outcomes of Eddy's "Random Genome Project"? What outcome would support a "selected" function? A "causal" function?
- What are the three types "big science" that Eddy describes? Which of those types can test hypotheses?