Comparisons of Mammalian Genomes

Chris Ponting

MRC Functional Genomics Unit

University of Oxford

Comparisons inevitably involve alignments

```
CGACATTAA--ATAGGCATAGCAGGACCAGATACCAGATCAAAGGCTTCAGGCGCA
CGACGTTAACGATTGGC---GCAGTATCAGATACCCGATCAAAG----CAGACGCA
```

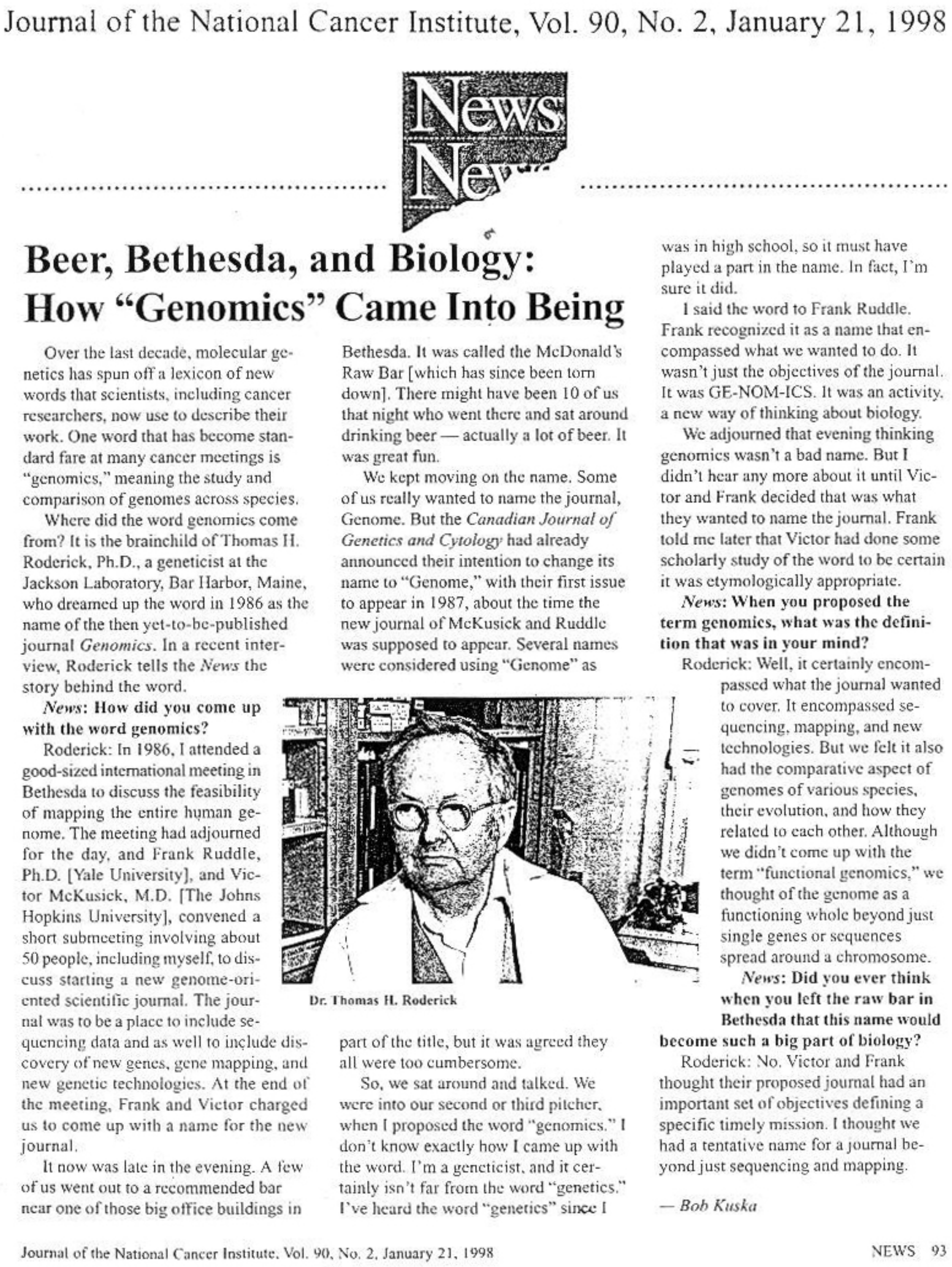
“GE-NOM-ICS...

It was an activity, a new way of thinking about biology.

It encompassed sequencing, mapping, and new technologies.

It also had the comparative aspect of genomes of various species, their evolution, and how they are related to each other.”

Thomas Roderick, who coined the term.



Journal of the National Cancer Institute, Vol. 90, No. 2, January 21, 1998

Beer, Bethesda, and Biology: How "Genomics" Came Into Being

Over the last decade, molecular genetics has spun off a lexicon of new words that scientists, including cancer researchers, now use to describe their work. One word that has become standard fare at many cancer meetings is "genomics," meaning the study and comparison of genomes across species.

Where did the word genomics come from? It is the brainchild of Thomas H. Roderick, Ph.D., a geneticist at the Jackson Laboratory, Bar Harbor, Maine, who dreamed up the word in 1986 as the name of the then yet-to-be-published journal *Genomics*. In a recent interview, Roderick tells the *News* the story behind the word.

***News*: How did you come up with the word genomics?**

Roderick: In 1986, I attended a good-sized international meeting in Bethesda to discuss the feasibility of mapping the entire human genome. The meeting had adjourned for the day, and Frank Ruddle, Ph.D. [Yale University], and Victor McKusick, M.D. [The Johns Hopkins University], convened a short submeeting involving about 50 people, including myself, to discuss starting a new genome-oriented scientific journal. The journal was to be a place to include sequencing data and as well to include discovery of new genes, gene mapping, and new genetic technologies. At the end of the meeting, Frank and Victor charged us to come up with a name for the new journal.

It now was late in the evening. A few of us went out to a recommended bar near one of those big office buildings in Bethesda. It was called the McDonald's Raw Bar [which has since been torn down]. There might have been 10 of us that night who went there and sat around drinking beer — actually a lot of beer. It was great fun.

We kept moving on the name. Some of us really wanted to name the journal, Genome. But the *Canadian Journal of Genetics and Cytology* had already announced their intention to change its name to "Genome," with their first issue to appear in 1987, about the time the new journal of McKusick and Ruddle was supposed to appear. Several names were considered using "Genome" as part of the title, but it was agreed they all were too cumbersome.

So, we sat around and talked. We were into our second or third pitcher, when I proposed the word "genomics." I don't know exactly how I came up with the word. I'm a geneticist, and it certainly isn't far from the word "genetics." I've heard the word "genetics" since I was in high school, so it must have played a part in the name. In fact, I'm sure it did.

Dr. Thomas H. Roderick

I said the word to Frank Ruddle. Frank recognized it as a name that encompassed what we wanted to do. It wasn't just the objectives of the journal. It was GE-NOM-ICS. It was an activity, a new way of thinking about biology.

We adjourned that evening thinking genomics wasn't a bad name. But I didn't hear any more about it until Victor and Frank decided that was what they wanted to name the journal. Frank told me later that Victor had done some scholarly study of the word to be certain it was etymologically appropriate.

***News*: When you proposed the term genomics, what was the definition that was in your mind?**

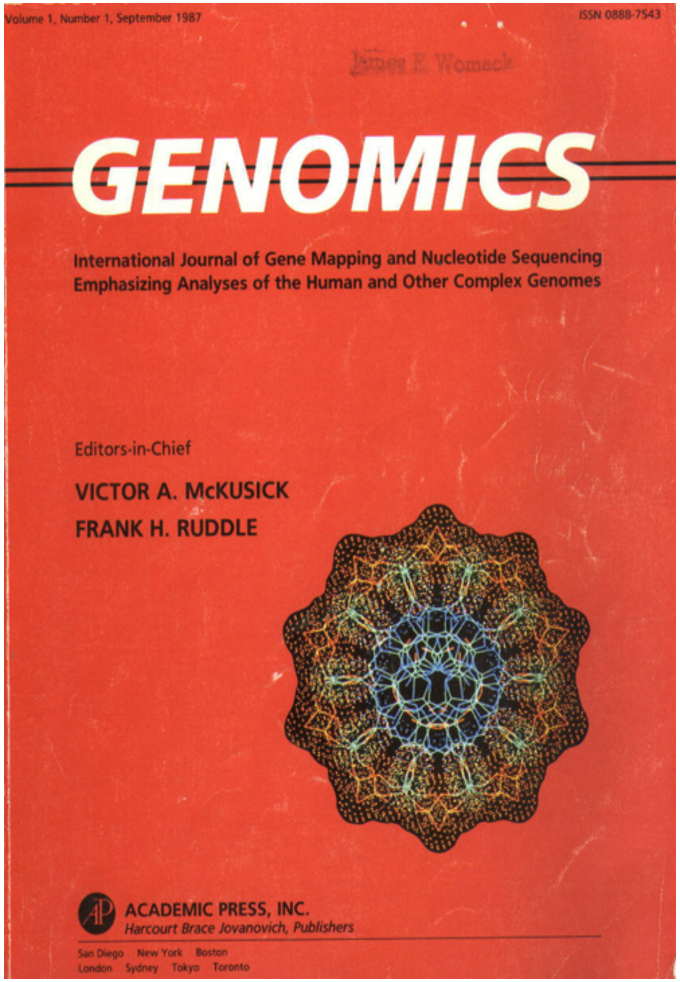
Roderick: Well, it certainly encompassed what the journal wanted to cover. It encompassed sequencing, mapping, and new technologies. But we felt it also had the comparative aspect of genomes of various species, their evolution, and how they related to each other. Although we didn't come up with the term "functional genomics," we thought of the genome as a functioning whole beyond just single genes or sequences spread around a chromosome.

***News*: Did you ever think when you left the raw bar in Bethesda that this name would become such a big part of biology?**

Roderick: No. Victor and Frank thought their proposed journal had an important set of objectives defining a specific timely mission. I thought we had a tentative name for a journal beyond just sequencing and mapping.

— *Bob Kuska*

1987



Volume 1, Number 1, September 1987 ISSN 0888-7543

James E. Womack

GENOMICS

International Journal of Gene Mapping and Nucleotide Sequencing
Emphasizing Analyses of the Human and Other Complex Genomes

Editors-in-Chief

VICTOR A. McKUSICK
FRANK H. RUDDLE

ACADEMIC PRESS, INC.
Harcourt Brace Jovanovich, Publishers
San Diego New York Boston
London Sydney Tokyo Toronto

nature
The mouse genome
15 February 2001
nature
the human genome
Finished
nature
The chicken
Cracking the code
nature
genome
mammalian evolution
nature
QUANTUM MEMORY
Controlling single photons
THE GENETICS OF NONSENSE
A cellular balancing act
IN PURSUIT OF PLEASURE
Dopamine's role revisited
THE DOG
NATUREJOBS
nature
THE OPOSSUM GENOME
Marsupial sequence illuminates mammalian evolution
HOT JUPITERS
HUMAN GENOMICS
A roadmap for structural variation
SORTING OUT THE CLIMATE
Do we have the technology?

15 February 2001
nature
£5.45 €8.29 FFr54 DM16 Lire16000
www.nature.com
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special

nature
BIODIVERSITY
What makes a species?
SATURN'S ROTATION
TROPOSPHERIC OZONE
Down from the heights
GENO
A PHY
Twelve Drosop
400 million ye
nature
CARBON DIOXIDE CAPTURE
Cagey ZIFs for emissions control
'FLIPPING' ENZYMES
Action mechanism of HIV's Achilles' heel
SCIENCE AND MUSIC
First in a new series
NATUREJOBS
Up Canada way
nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
FAMILY TIES

Inter-Species Sequence Comparisons

Era of Genomes and Evolution

1999 2000 2001 2002 2003 2004 2005

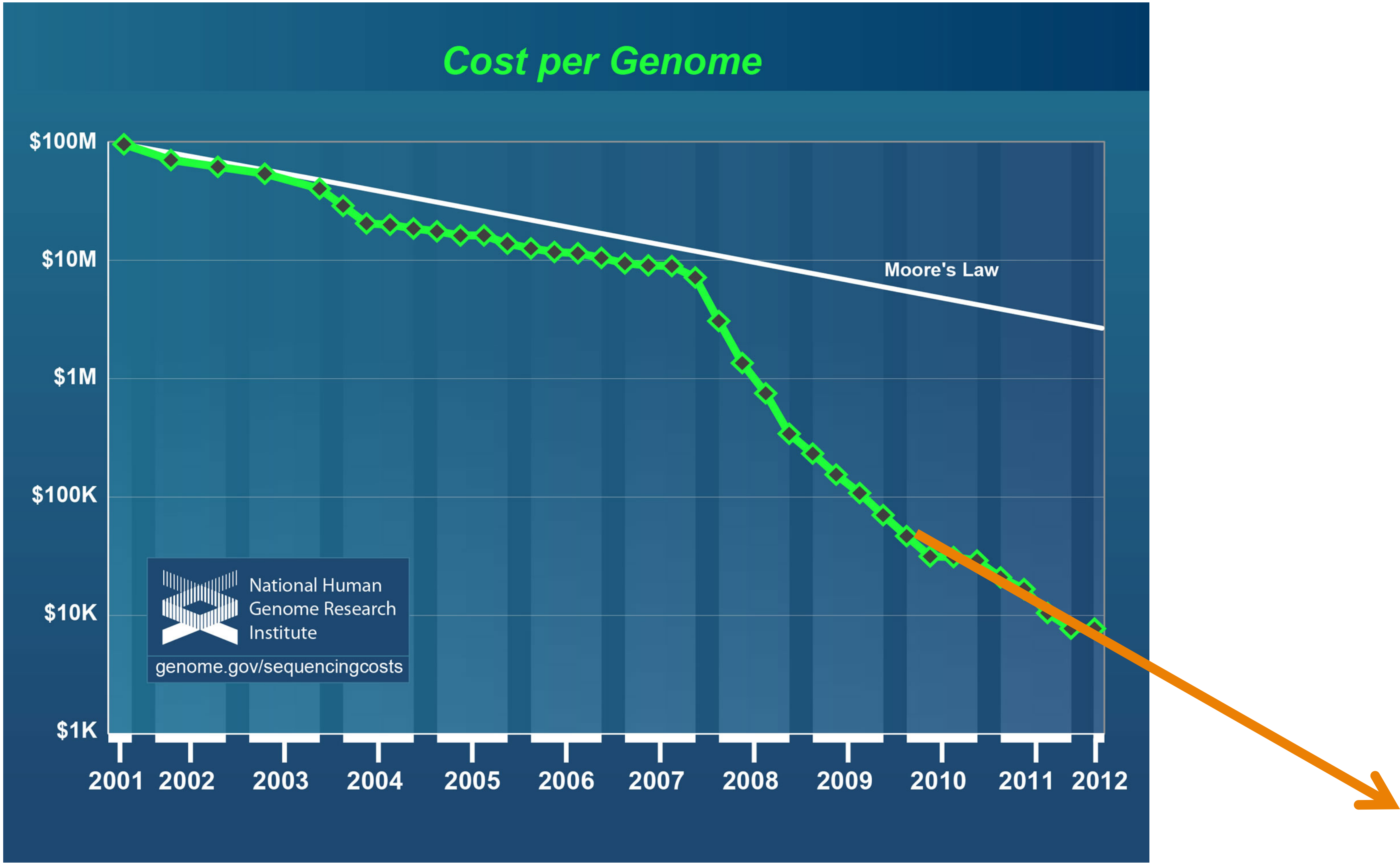
2006 2007 2008 2009 2010

2011 2012 2013 2014 2015 2016 2017 2018 2019 2020

Era of Disease Genomics

Intra-Species Sequence Comparisons

the first human genome cost ~ $3b



http://www.genome.gov/images/content/cost_per_genome.jpg

Summary

Part 1. 50 slides

- Human Genome Project
- Mouse Genome & Comparisons with Human

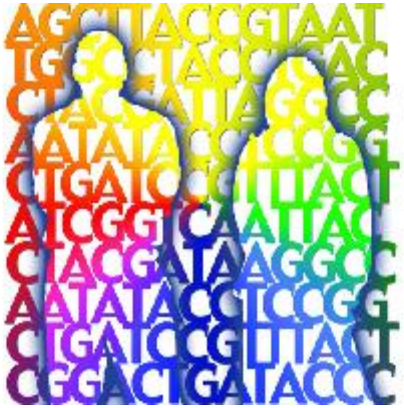
Part 2. 50 slides

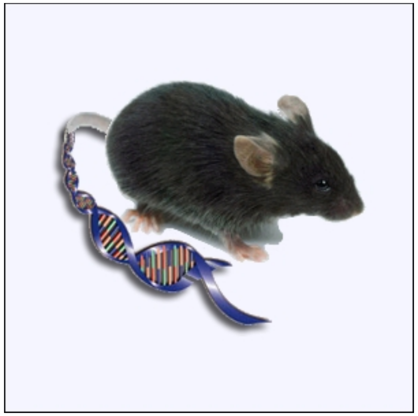
- The functional portion of the genome
- The ENCODE project
- Transcript maps

Part 3. 10 slides

- The Future

Part 1: Human and Mouse Genomes

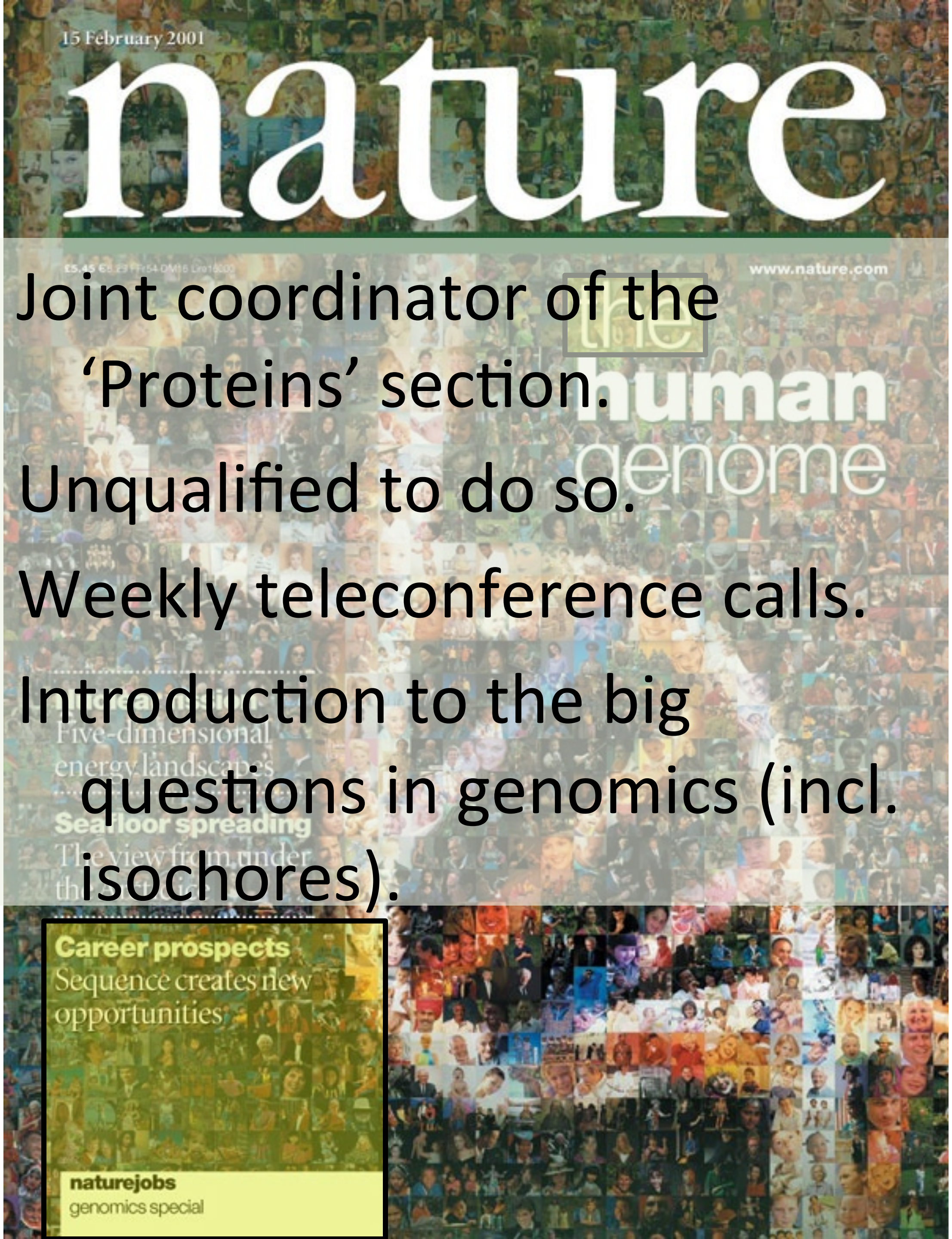




Pre-Genome Sequences

- The genome, and its genes, were not circumscribed.
- Studies could never be comprehensive (‘*genomic*’).
- Relatively little understanding about the extent and layers of transcriptional regulation.
- Genetics focused more on gene discovery, than on gene evolution or mechanism.
- Comparative Genomics was a pipe dream.

15 Feb 2001

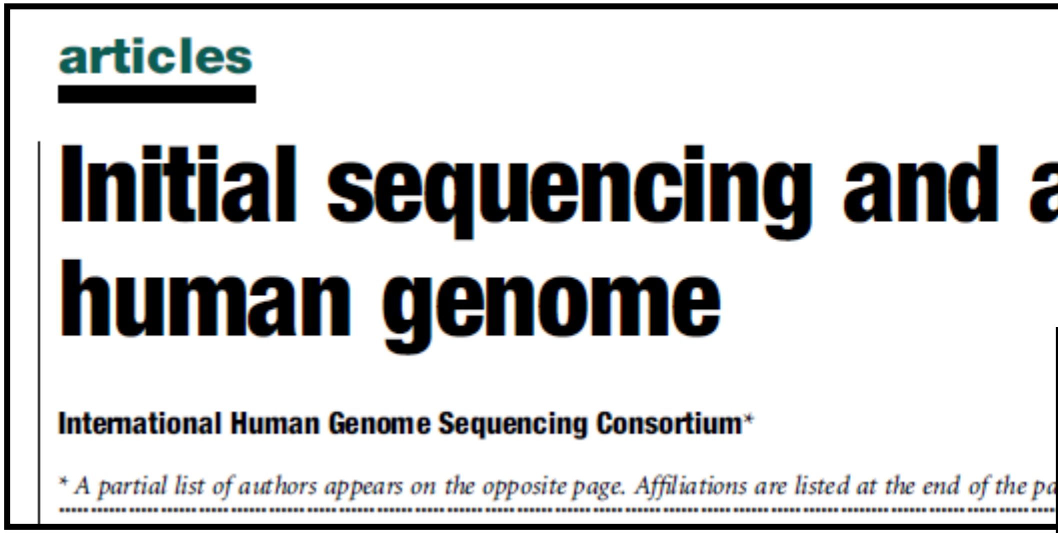


Joint coordinator of the ‘Proteins’ section.

Unqualified to do so.

Weekly teleconference calls.

Introduction to the big questions in genomics (incl. isochores).

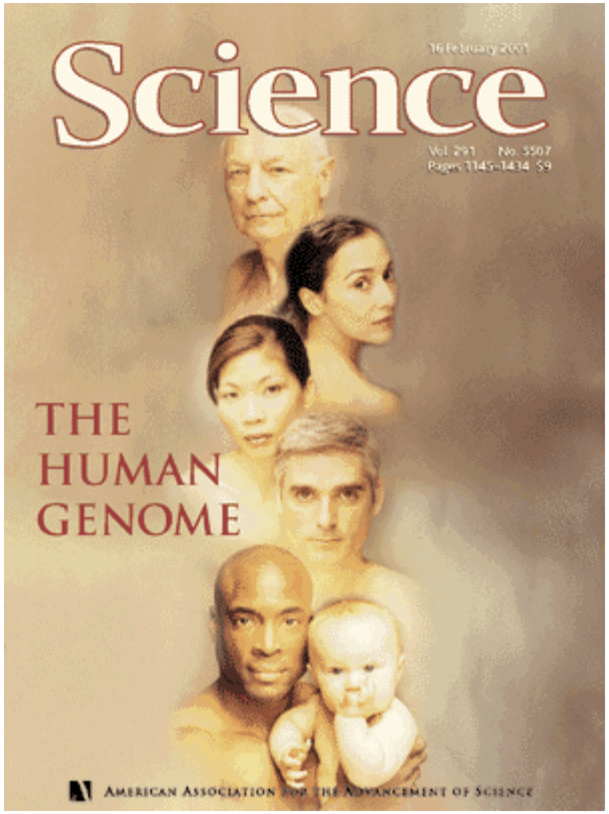


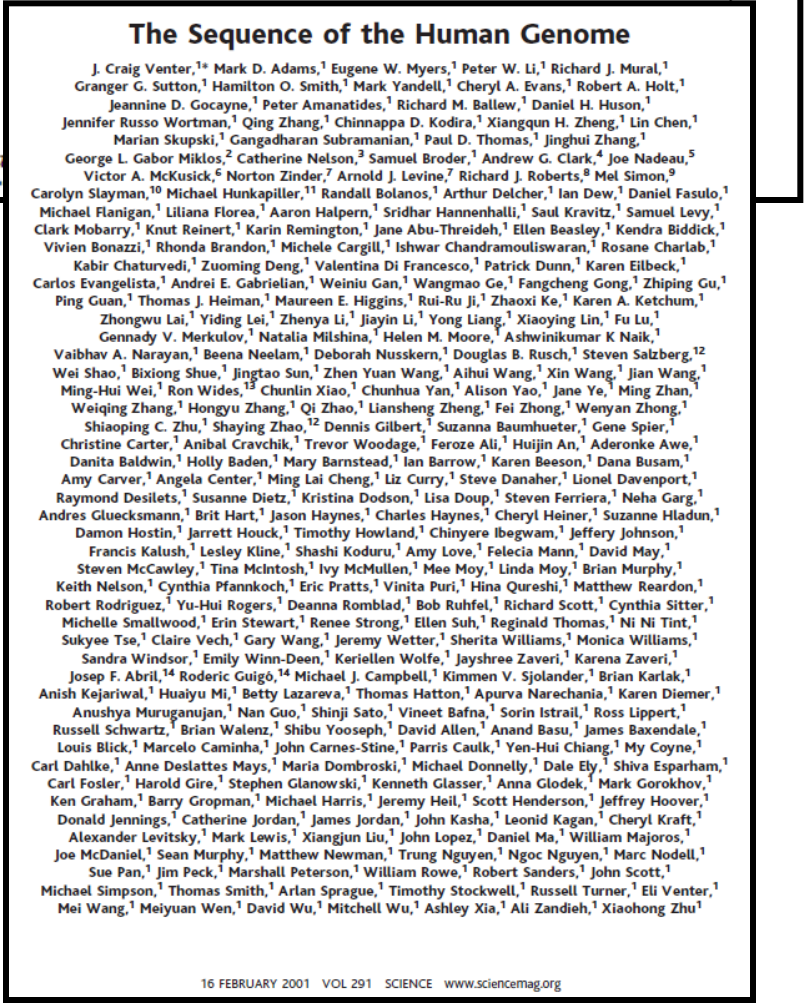
articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* A partial list of authors appears on the opposite page. Affiliations are listed at the end of the pa



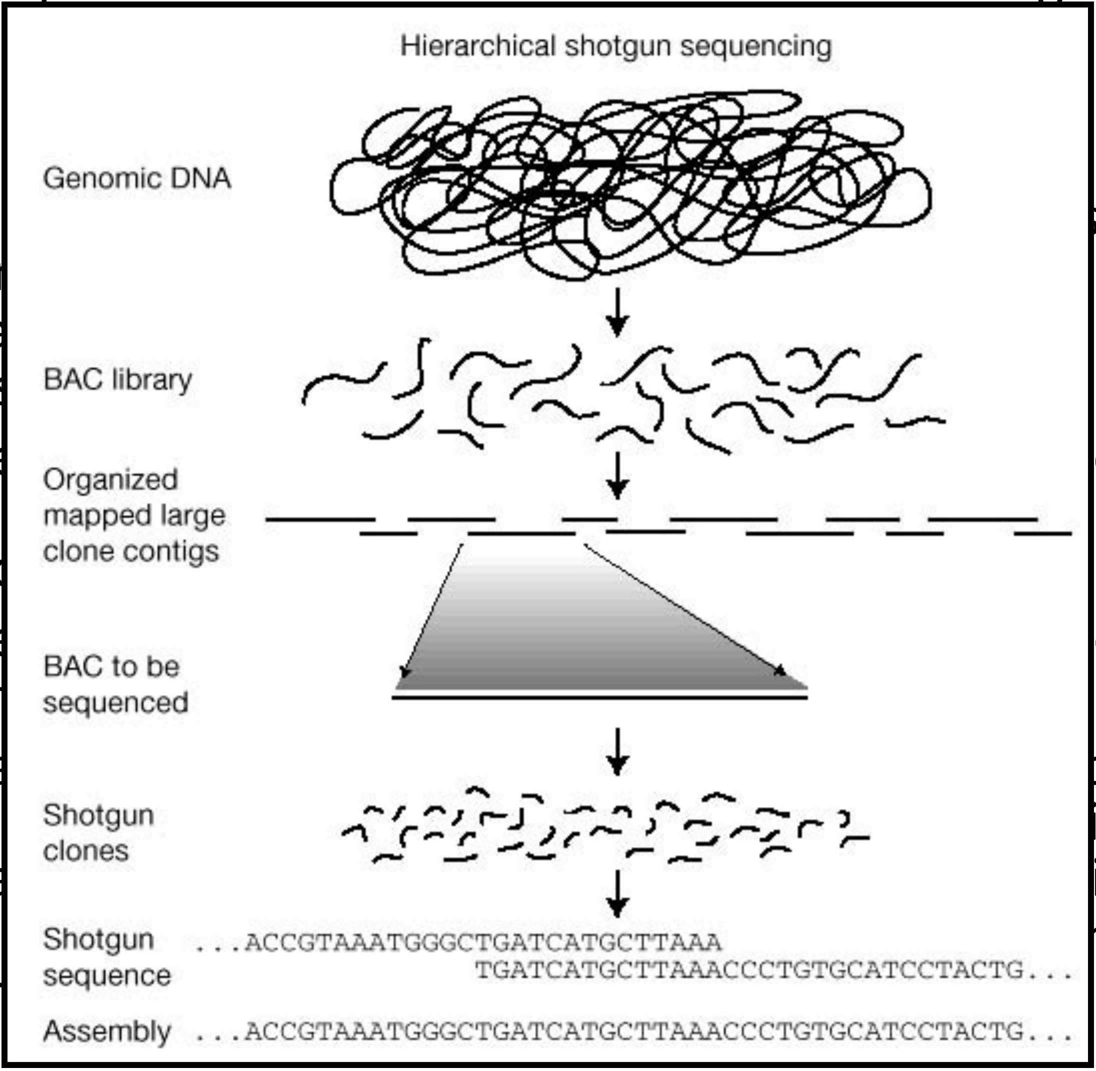


The Sequence of the Human Genome

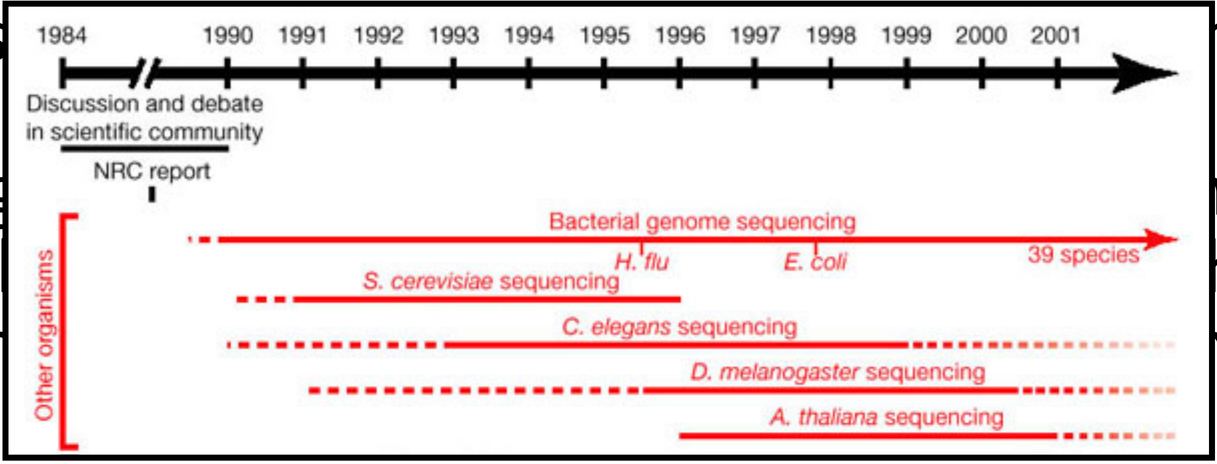
J. Craig Venter,[1*] Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1] Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1] Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1] Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1] Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1] George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5] Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9] Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1] Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1] Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1] Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1] Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1] Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1] Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1] Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1] Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1] Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12] Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1] Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1] Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1] Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1] Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1] Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1] Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1] Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1] Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1] Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1] Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1] Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1] Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1] Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1] Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1] Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1] Sandra Windsor,[1] Emily Winn-Deen,[1] Keriellen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1] Josep F. Abril,[14] Roderic Guigó,[14] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1] Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1] Anushya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1] Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1] Louis Blick,[1] Marcelo Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1] Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1] Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1] Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1] Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1] Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1] Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1] Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1] Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1] Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

16 FEBRUARY 2001 VOL 291 SCIENCE www.sciencemag.org

- Here we report the results of a collaboration involving 20 groups
from the nce,
Germany human
genome.
- The draft sical map
covering human
genome a ic
databases
- The seque od, with
coverage roughly
fifteen mo
- The seque estriction
and updat
- The task a osing all
gaps and illion bases
are in fina ity of the
sequence hould
proceed r

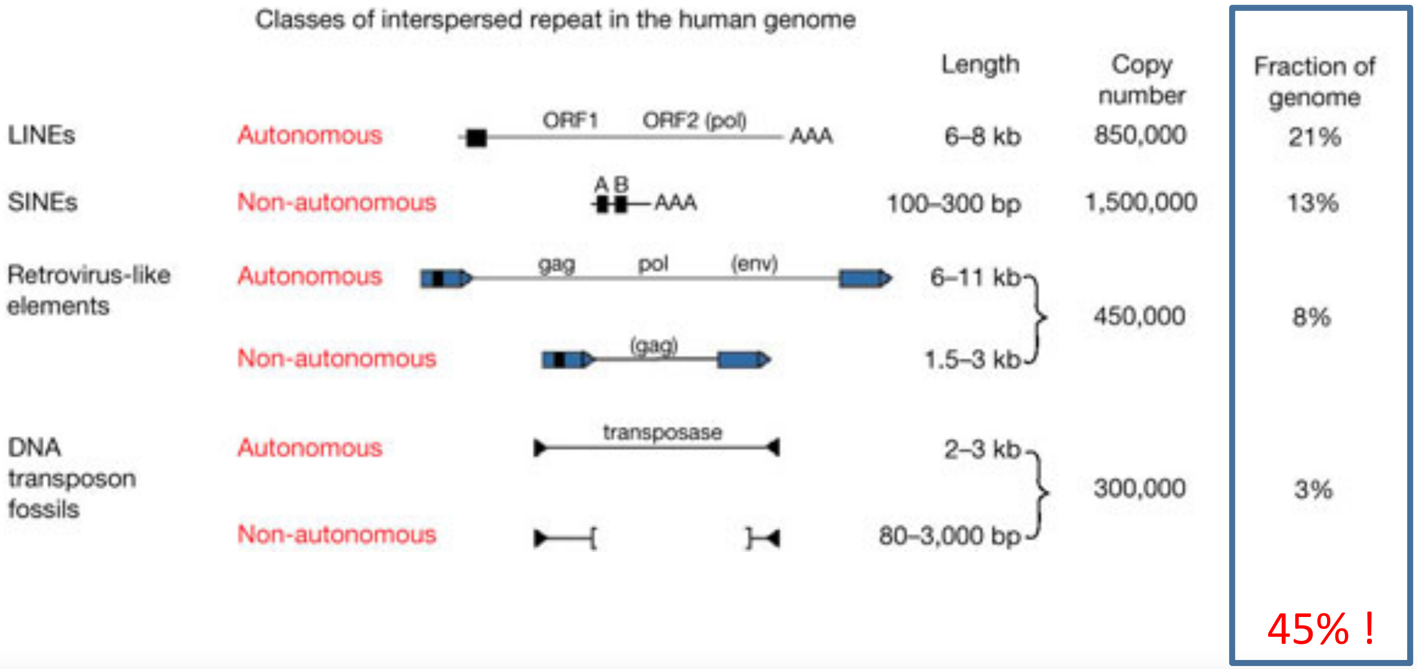

Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

- The sequence of the human genome is of interest in s me to be ext mes as larg ne and eig h gen ne to be ext is the genome of our own species.



- Much work remains to be done to produce a complete finished sequence, but the vast trove of information that has become available through this collaborative effort allows a global perspective on the human genome. Although the details will change as the sequence is finished, many points are already clear.

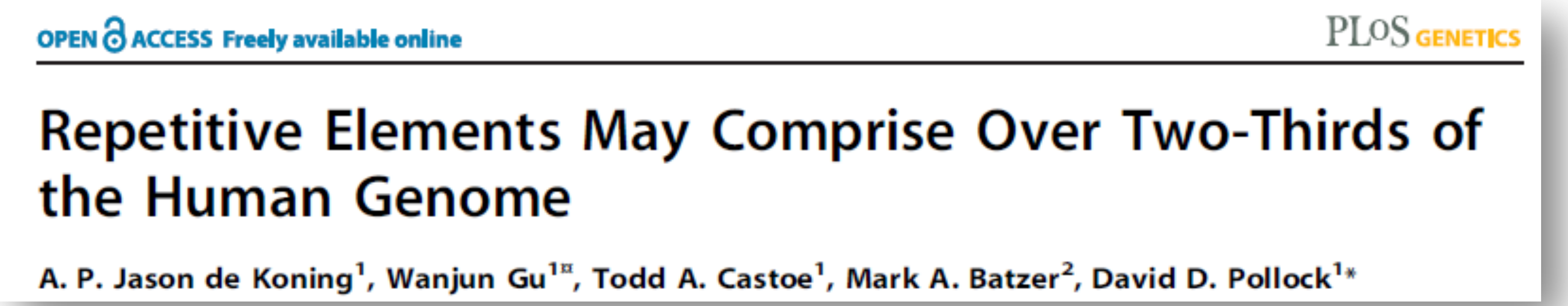
Transposable elements dominate the human genome



Classes of interspersed repeat in the human genome

		Length	Copy number	Fraction of genome
LINEs	Autonomous	6–8 kb	850,000	21%
SINEs	Non-autonomous	100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous	6–11 kb	450,000	8%
	Non-autonomous	1.5–3 kb		
DNA transposon fossils	Autonomous	2–3 kb	300,000	3%
	Non-autonomous	80–3,000 bp		

45% !

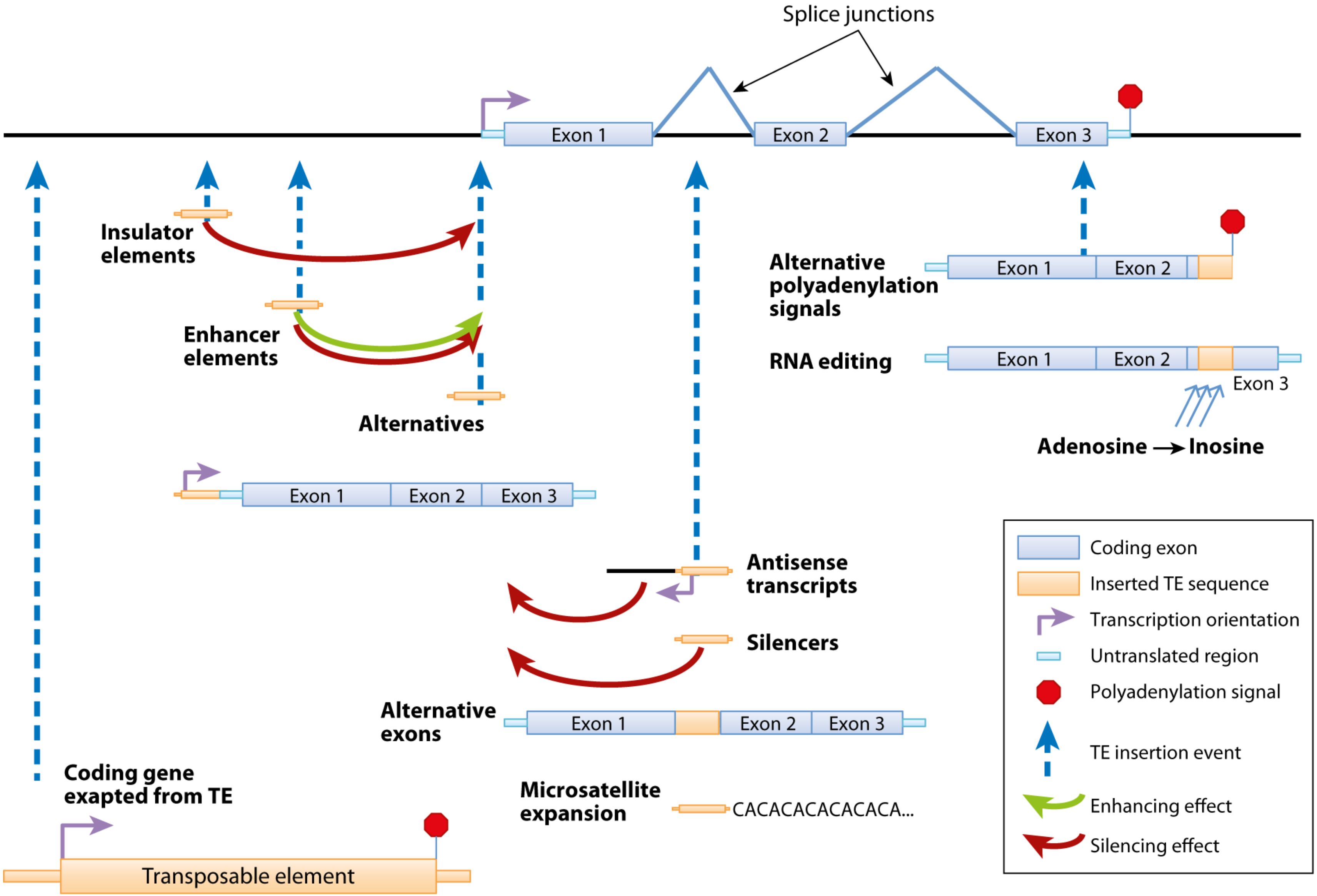


OPEN ACCESS Freely available online

PLoS GENETICS

Repetitive Elements May Comprise Over Two-Thirds of the Human Genome

A. P. Jason de Koning[1], Wanjun Gu[1¤], Todd A. Castoe[1], Mark A. Batzer[2], David D. Pollock[1*]



Ponting CP, et al. 2011.
Annu. Rev. Genomics Hum. Genet. 12:275–99

Isochores

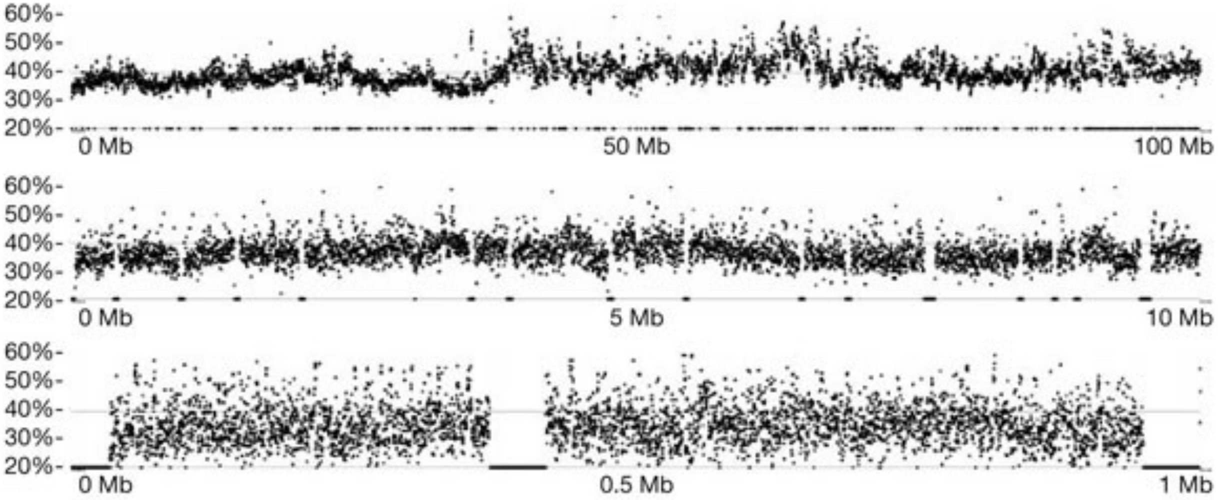
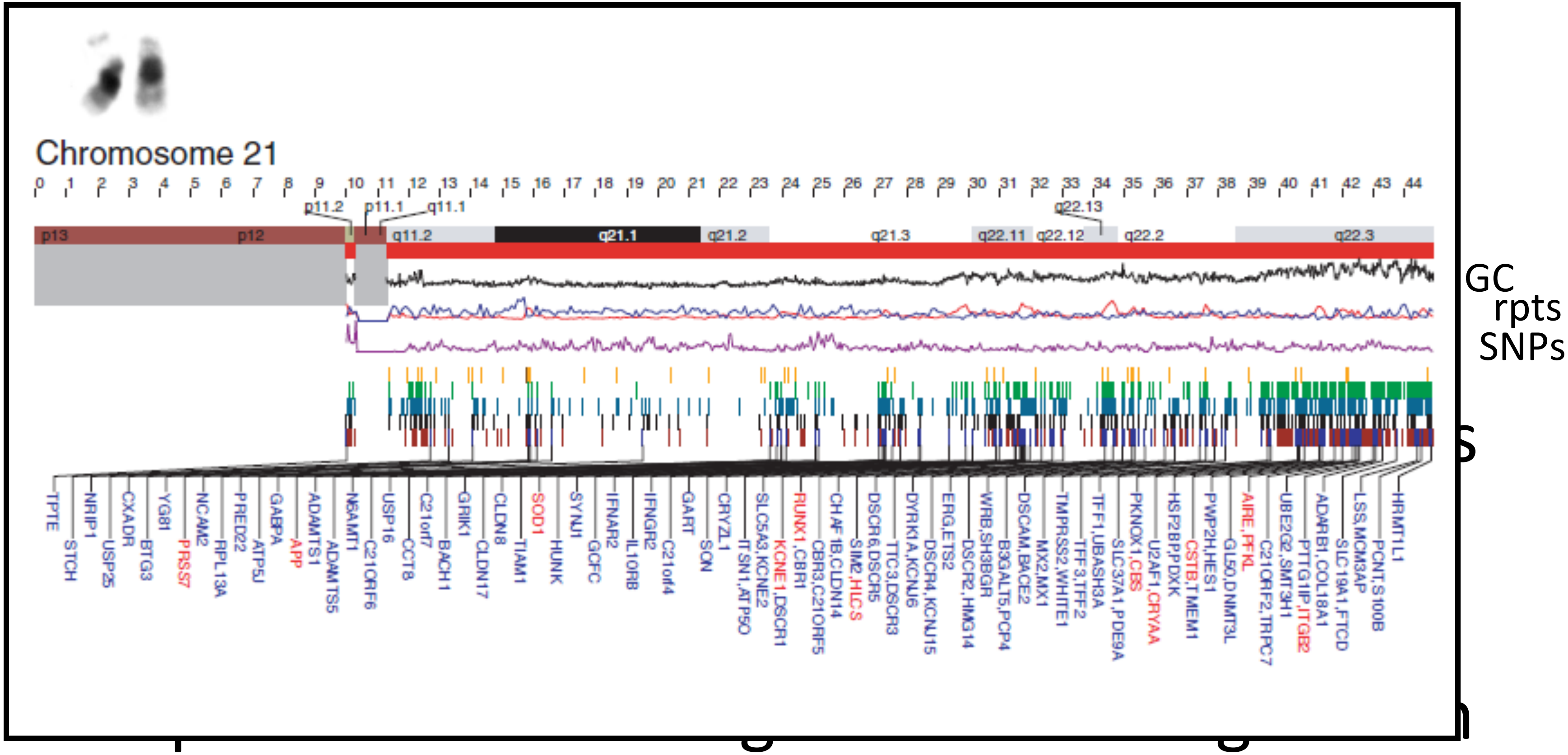


FIGURE 13. Variation in GC content at various scales

Bernardi et al. described the genome as being composed of a mosaic of compositionally homogeneous regions dubbed ‘isochores’.

- The genomic landscape shows marked variation in the distribution of a number of



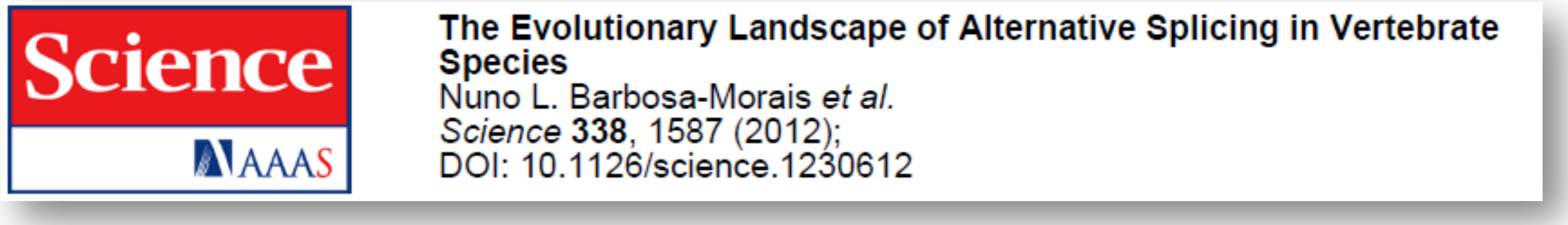
s

n

the clusters.

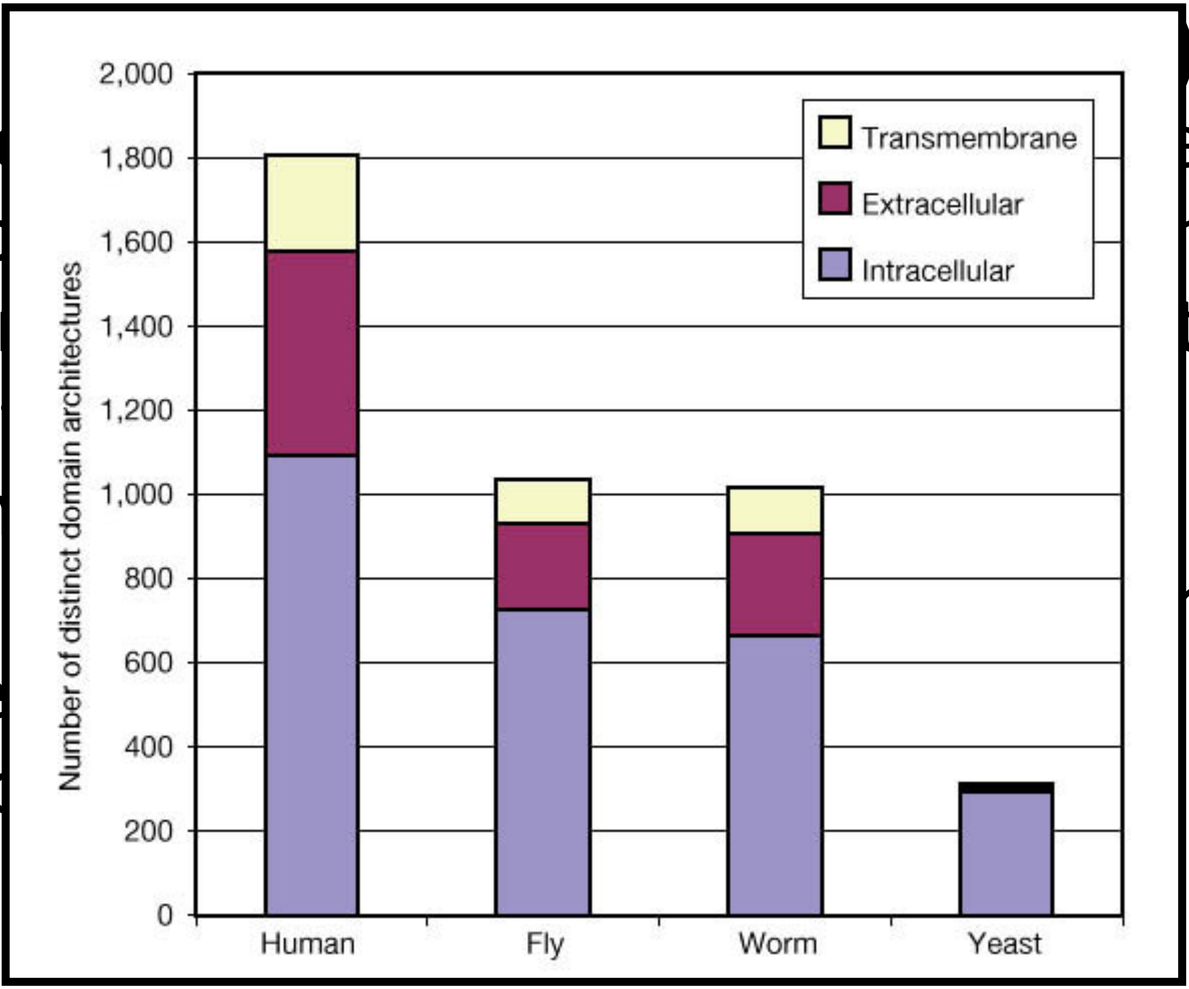
- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.

20,000 protein coding genes

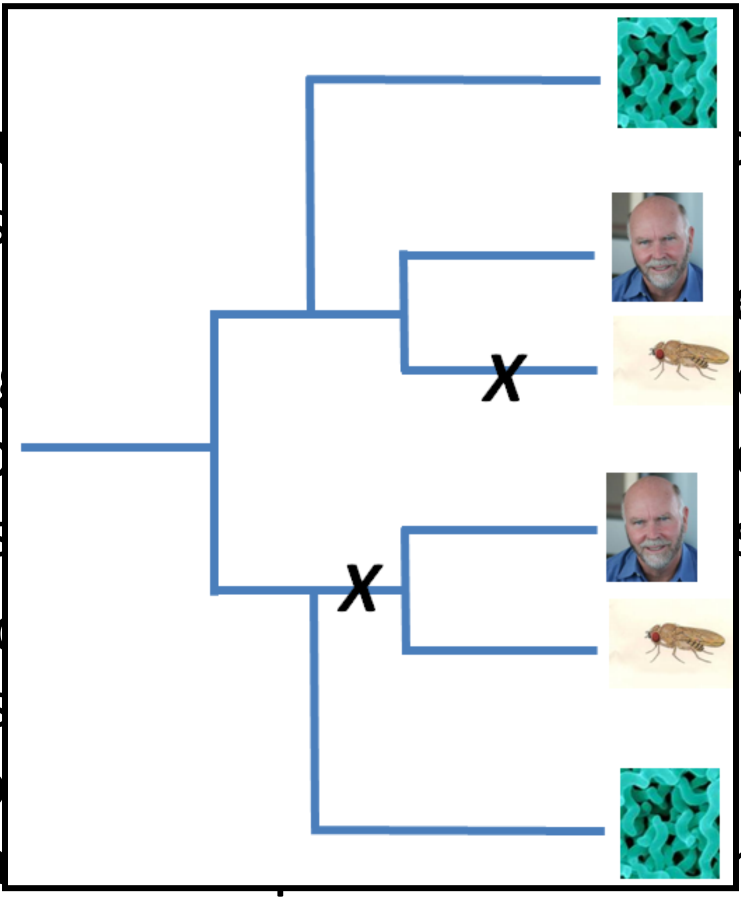


The Evolutionary Landscape of Alternative Splicing in Vertebrate Species
Nuno L. Barbosa-Morais *et al.*
Science **338**, 1587 (2012);
DOI: 10.1126/science.1230612

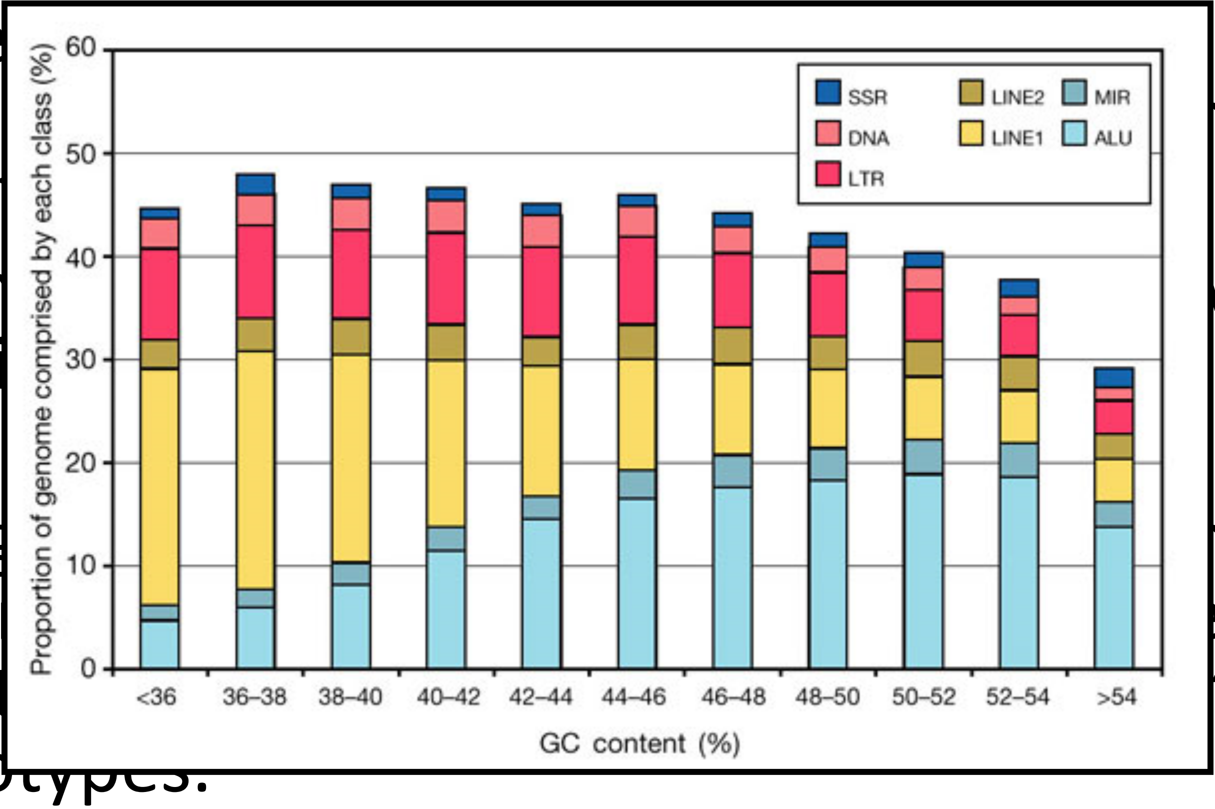
- The f)
encod e
comp his is
due i te-
speci
estim the
fact t ranged
pre-e
collec



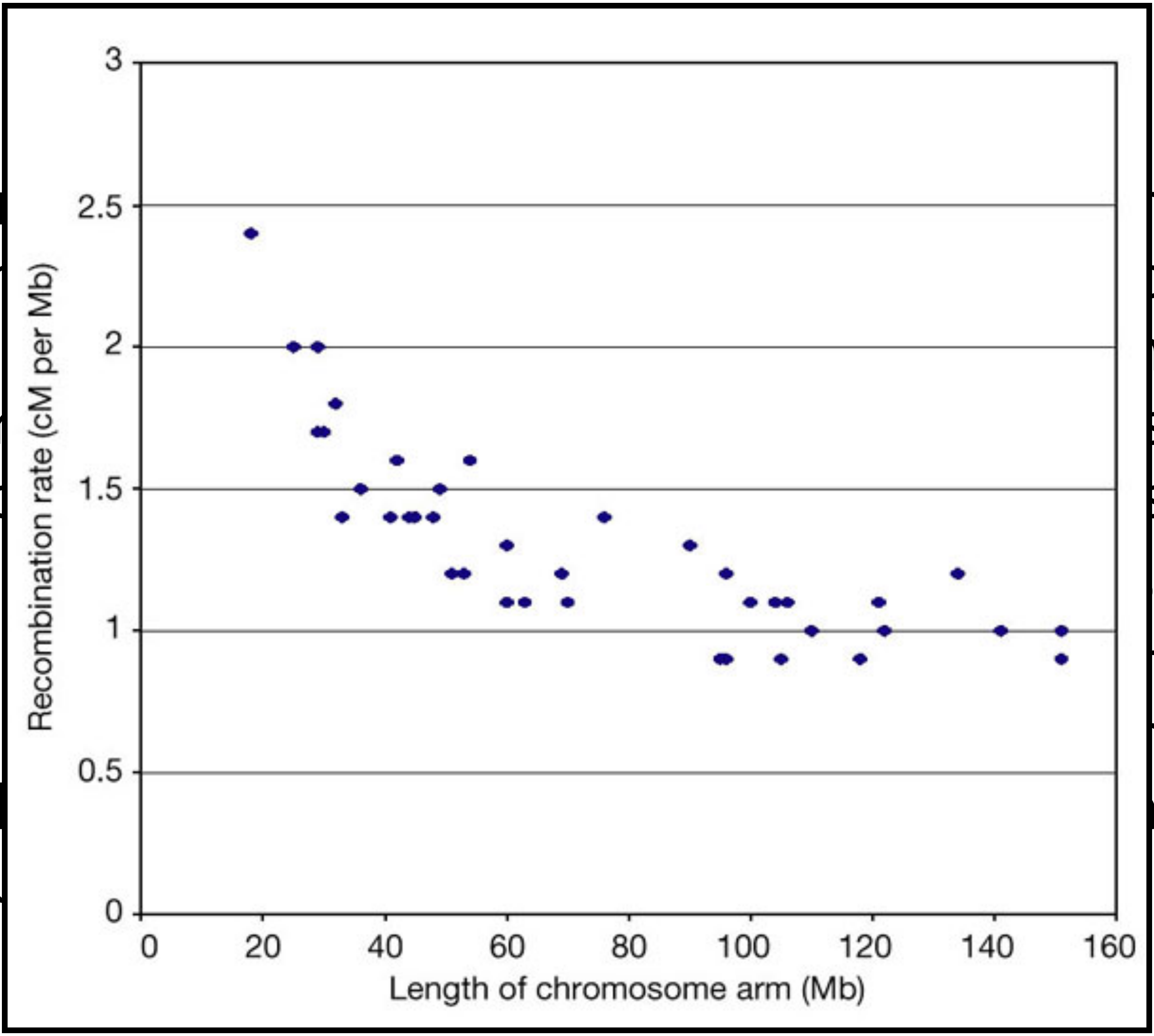
- Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been derived from transposable elements.
- Although abo ome derives
from transpos been a marked
decline in the ements in the
hominid linea ear to have
become comp erminal repeat
(LTR) retropos so.
- The pericentr regions of
chromosomes nt segmental
duplications o re in the
genome. Segn n more frequent
in humans than in yeast, fly or worm.



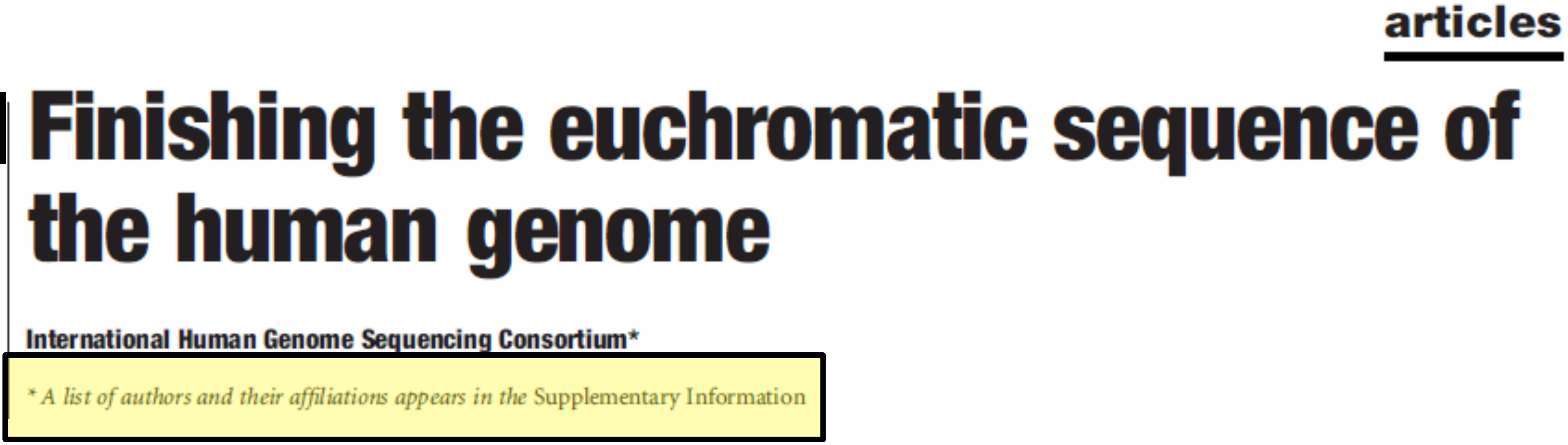
- Analysis of the organization of Alu elements explains the longstanding mystery of their surprising genomic distribution, and suggests that there may be strong selection in favour of prefe C-rich regio ay bene
- The r wice as high i that most
- Cytog nes confi egions are st ’ in karyotypes.



- Recor distal
region osomes
and o a
patter st one
crosso s.
- More rphisms
(SNPs ied.
This c ome-
wide es in
the hu



2004



articles

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

* *A list of authors and their affiliations appears in the* Supplementary Information

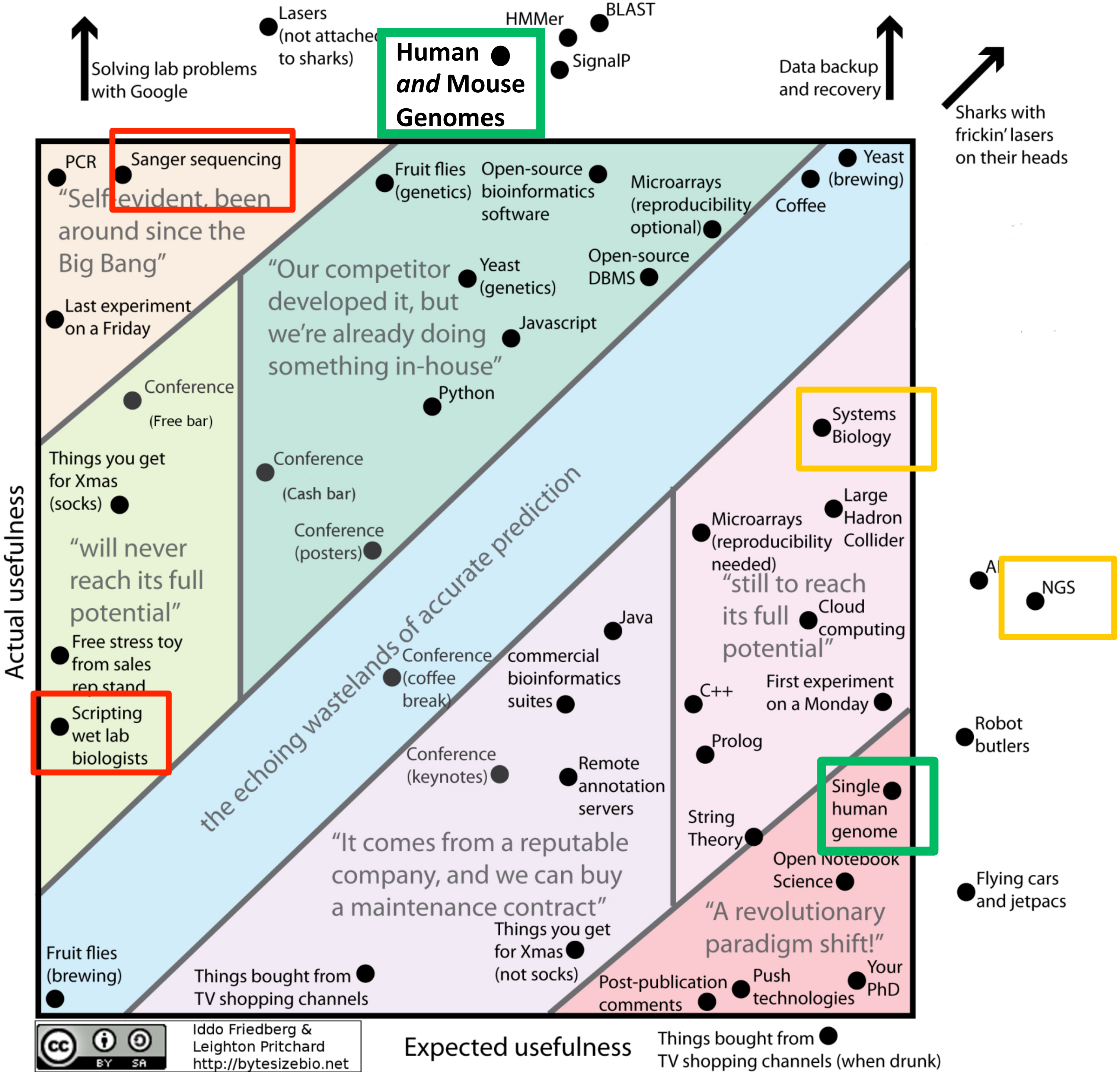
- "The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps.
- It covers 99% of the euchromatic genome and is accurate to an error rate of 1 event per 100,000 bases.
- Notably, the human genome seems to encode only 20,000–25,000 protein-coding genes."

GRCh37, the Genome Reference Consortium human genome (build 37) is derived from thirteen anonymous volunteers from Buffalo, New York

The human genome: More questions than answers

- How many genes? How much functional DNA?
- Are transposon-derived sequences functional?
- How did the heterogeneity in GC (‘isochores’) arise and how is it sustained?
- How is recombination controlled?
- How does the human genome, and its genes, differ from those of more closely-related genomes? Is it at all unusual?
- How is transcription regulated?

How Useful is a Human Genome?

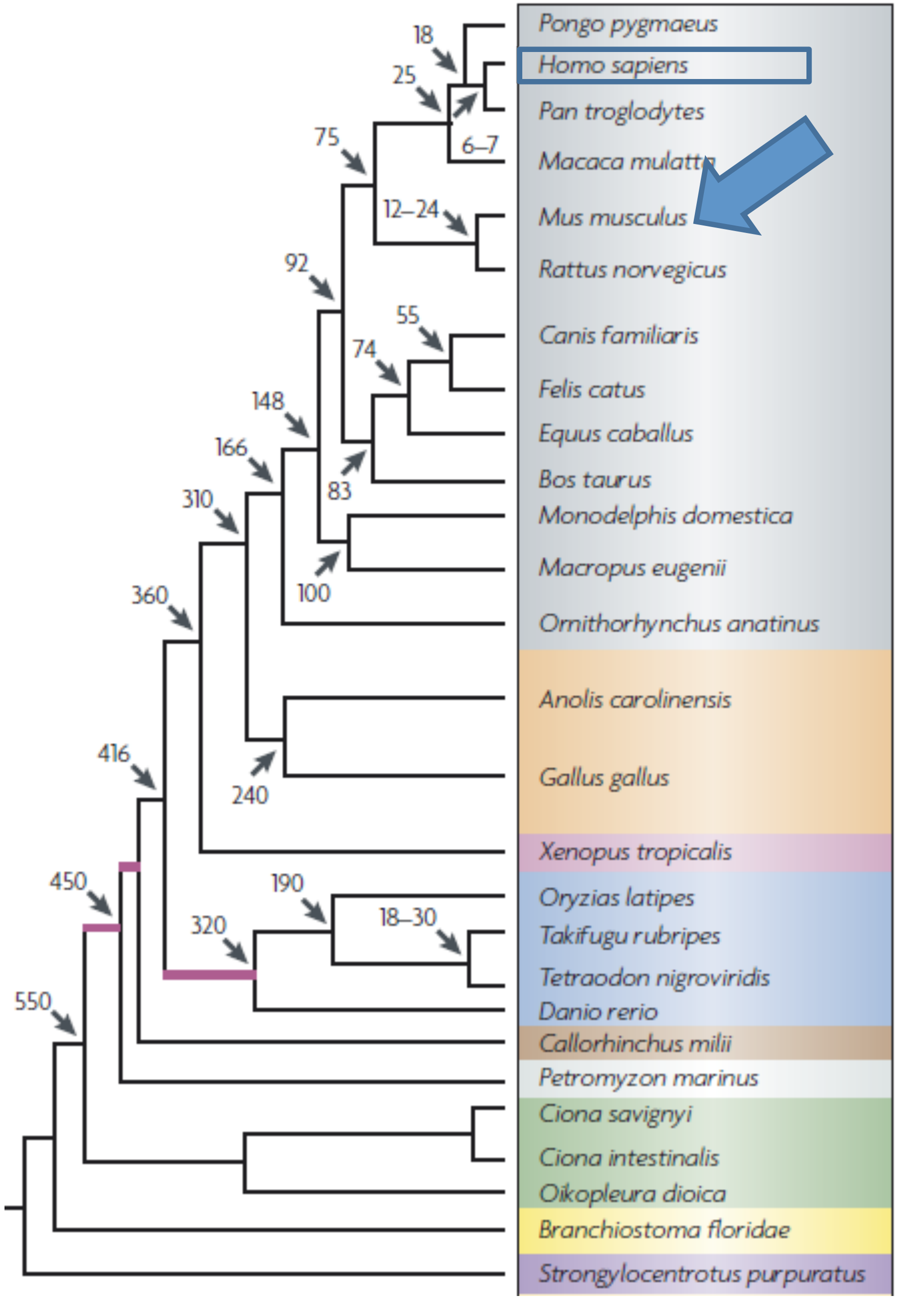

Solving lab problems with Google
Lasers (not attached to sharks)
Human and Mouse Genomes
HMMer
BLAST
SignalP
Data backup and recovery
Sharks with frickin' lasers on their heads
PCR
Sanger sequencing
"Self evident, been around since the Big Bang"
Last experiment on a Friday
Conference (Free bar)
Things you get for Xmas (socks)
"will never reach its full potential"
Free stress toy from sales rep stand
Scripting wet lab biologists
Fruit flies (genetics)
Open-source bioinformatics software
Microarrays (reproducibility optional)
Yeast (brewing)
Coffee
"Our competitor developed it, but we're already doing something in-house"
Yeast (genetics)
Open-source DBMS
Javascript
Python
Conference (Cash bar)
Conference (posters)
the echoing wastelands of accurate prediction
Conference (coffee break)
Conference (keynotes)
Java
commercial bioinformatics suites
Remote annotation servers
"It comes from a reputable company, and we can buy a maintenance contract"
Things you get for Xmas (not socks)
Things bought from TV shopping channels
Fruit flies (brewing)
Systems Biology
Large Hadron Collider
Microarrays (reproducibility needed)
"still to reach its full potential"
Cloud computing
First experiment on a Monday
C++
Prolog
String Theory
Single human genome
Open Notebook Science
"A revolutionary paradigm shift!"
Post-publication comments
Push technologies
Your PhD
AI
NGS
Robot butlers
Flying cars and jetpacs
Actual usefulness
Expected usefulness
Things bought from TV shopping channels (when drunk)

Iddo Friedberg &
Leighton Pritchard
http://bytesizebio.net

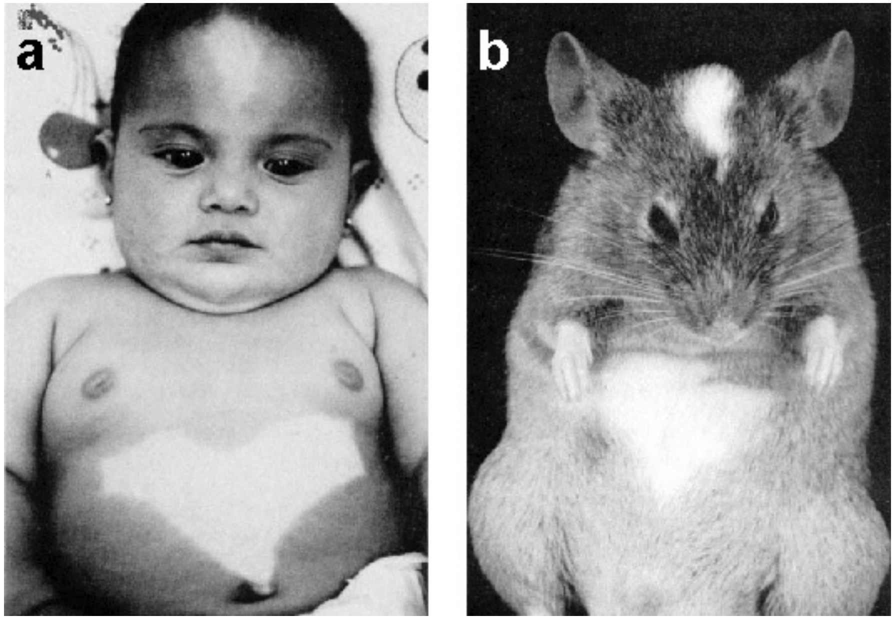
Evolutionary yardsticks

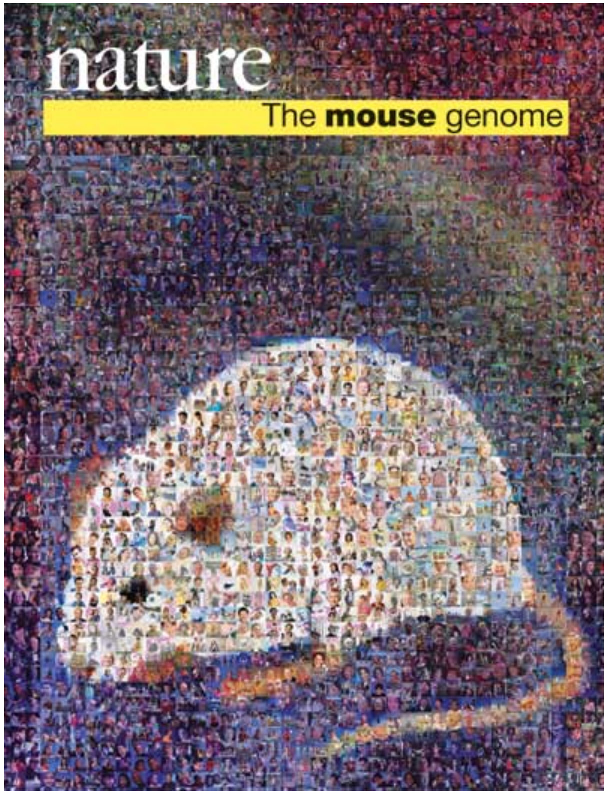
mammalian & invertebrate noncoding sequences do *not* align

c 2008

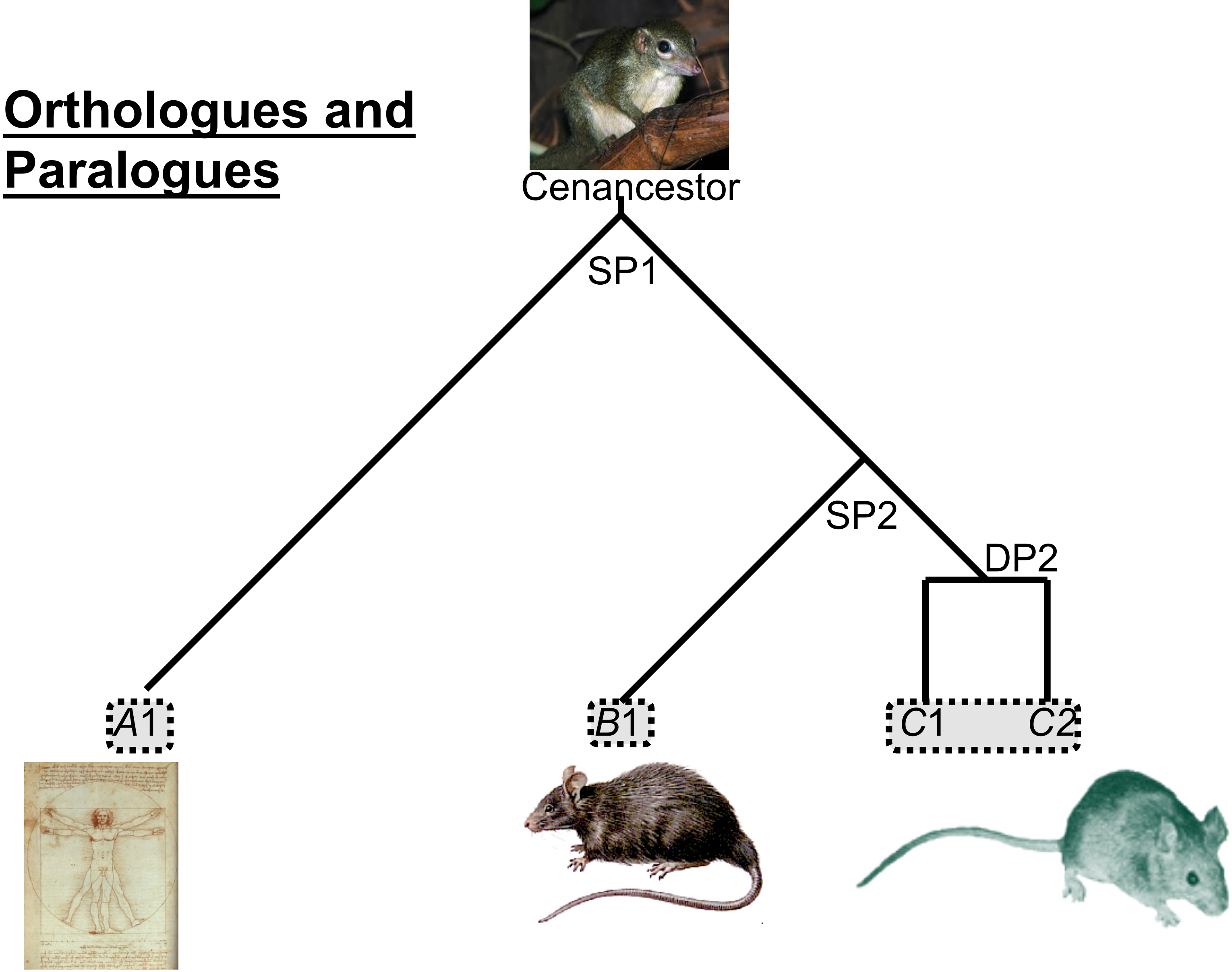


Mouse vs Human



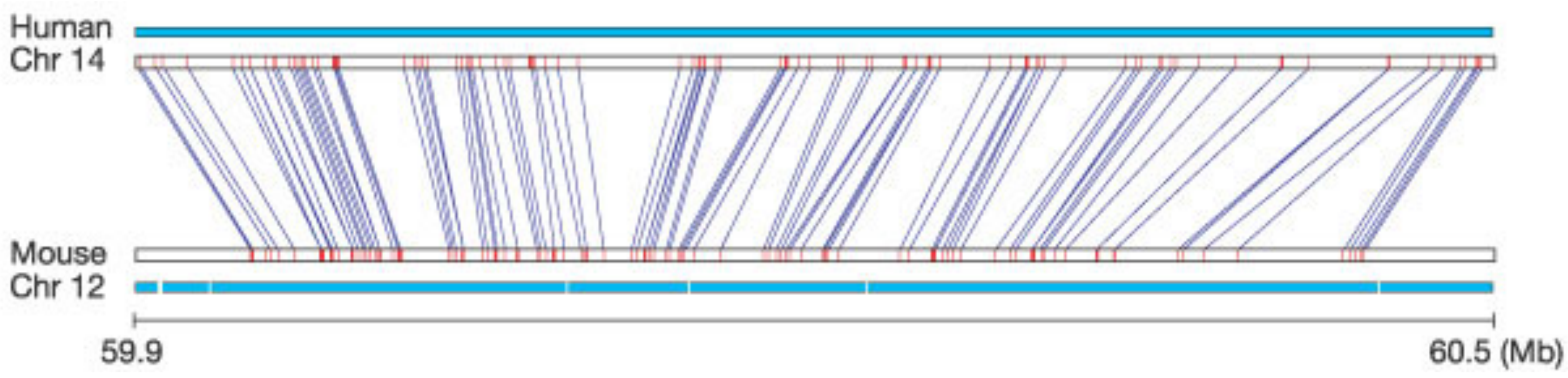


Orthologues and Paralogues



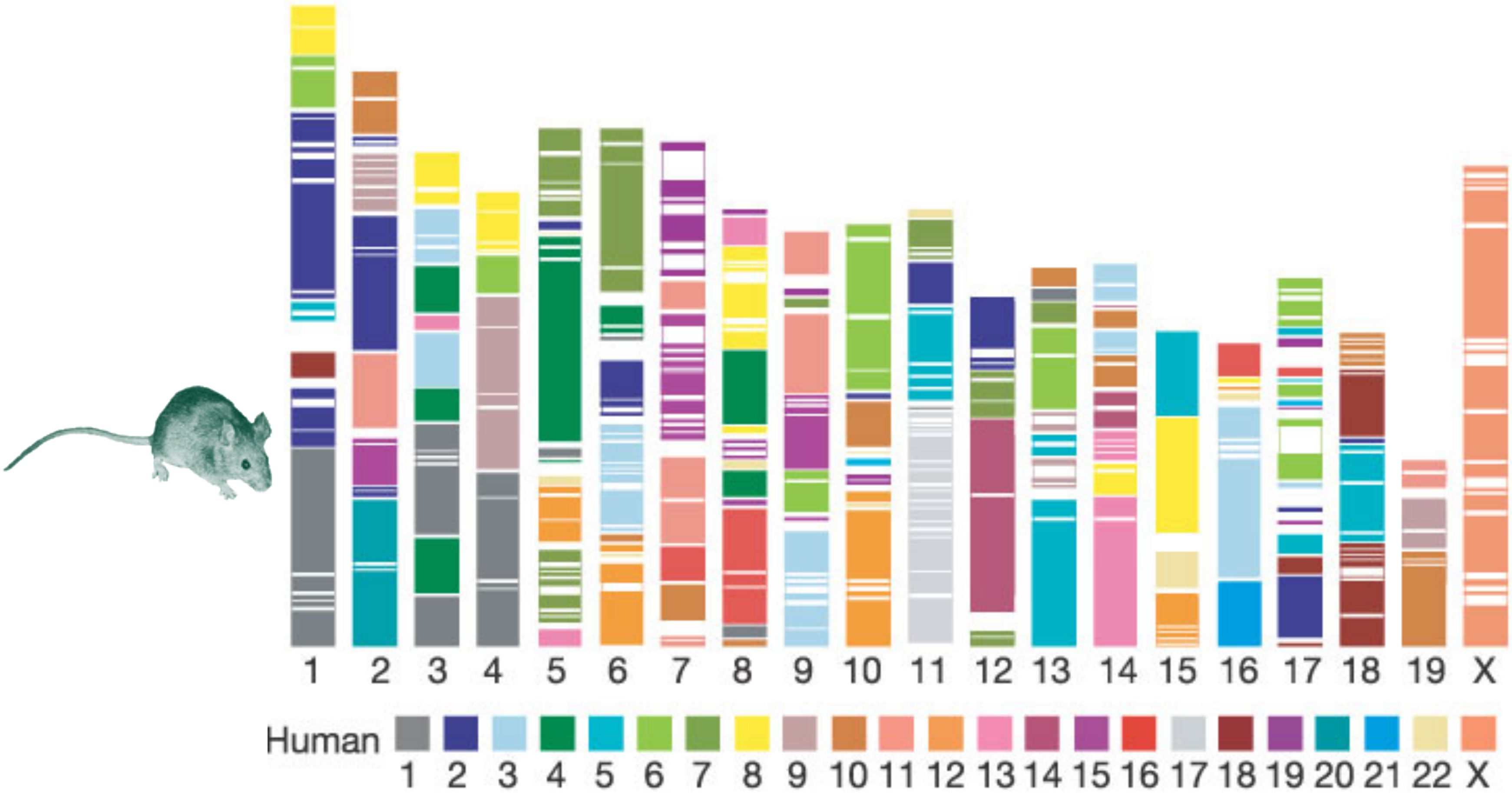
C1 and C2 are paralogues
A1 and B1 and (C1 and C2) are orthologues

Human and mouse
"local synteny"

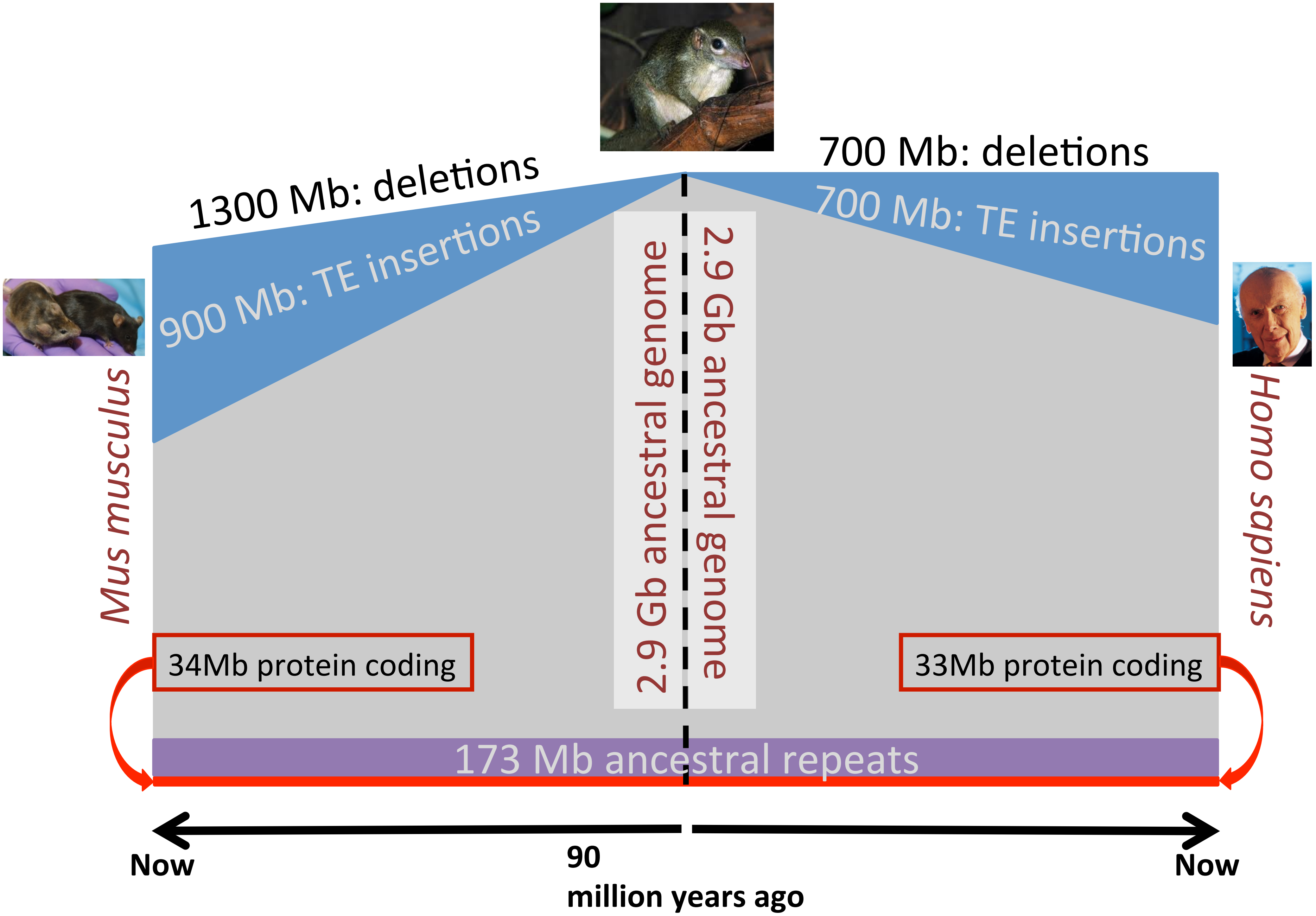


"Syntenic" regions contain orthologues!

Human and mouse chromosomes: global orthology



Only 40% of the human genome aligns to the mouse genome.



1300 Mb: deletions

900 Mb: TE insertions

700 Mb: deletions

700 Mb: TE insertions

Mus musculus

Homo sapiens

2.9 Gb ancestral genome

2.9 Gb ancestral genome

34Mb protein coding

33Mb protein coding

173 Mb ancestral repeats

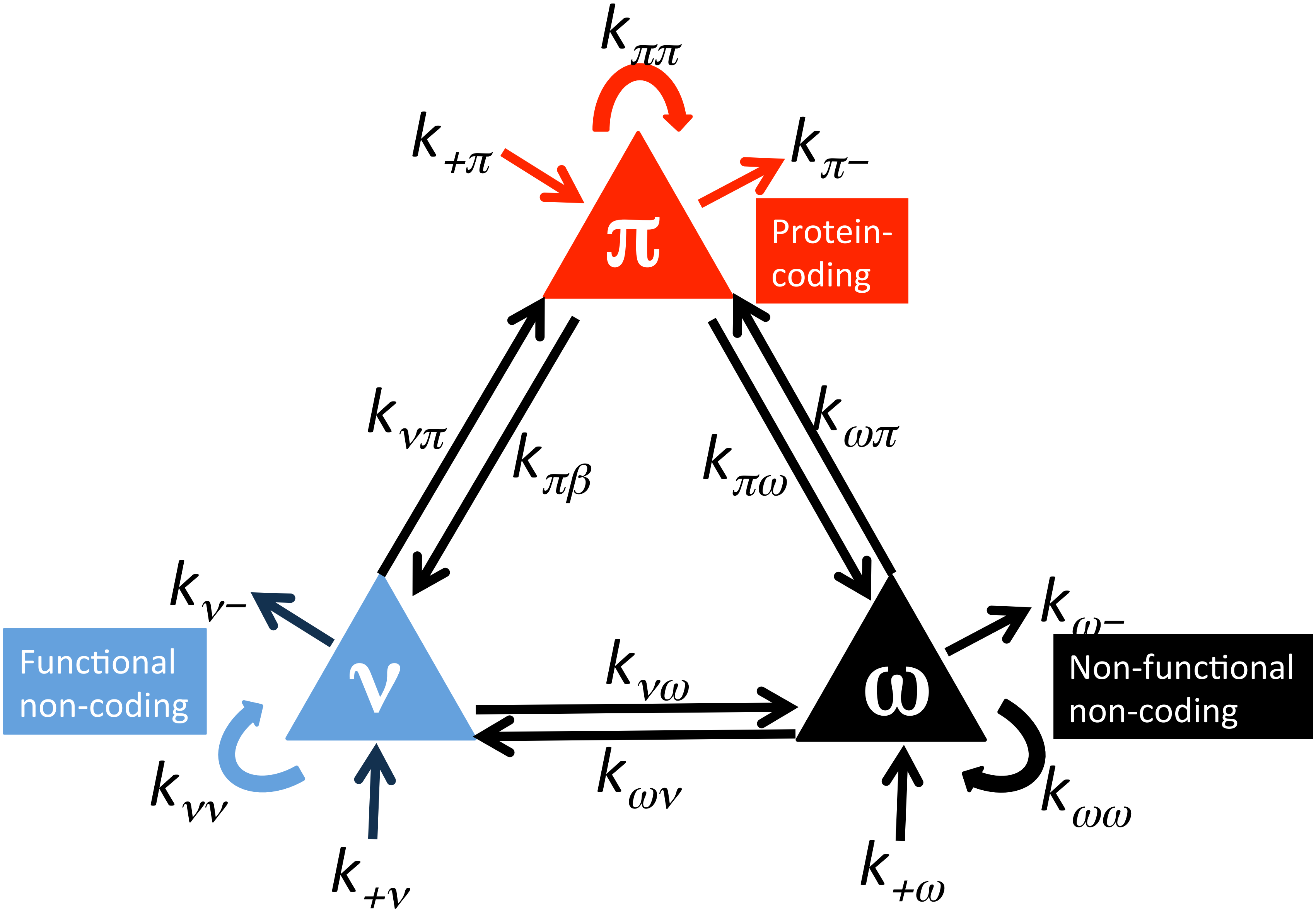
Now

90

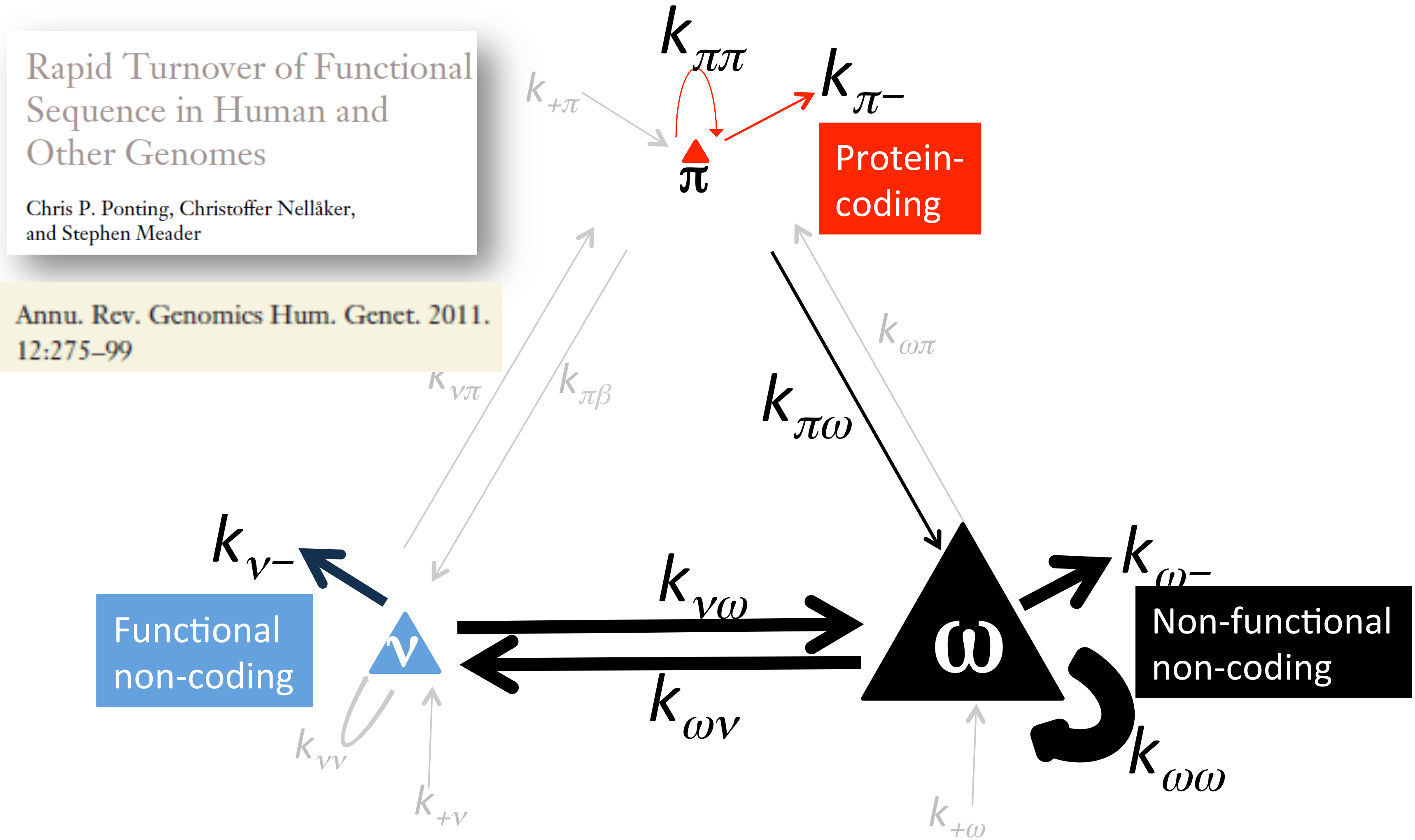
million years ago

Now

Three-state model of a mammalian genome



The dynamics of a mammalian genome are dominated by transposable elements



Rapid Turnover of Functional Sequence in Human and Other Genomes

Chris P. Ponting, Christoffer Nellåker, and Stephen Meader

Annu. Rev. Genomics Hum. Genet. 2011. 12:275–99

Conservation, Constraint and Function

1. Conserved sequence is not necessarily constrained: e.g. human-chimpanzee sequence
2. Constrained sequence is not necessarily conserved: e.g. lineage-specific function or high local mutation rates
3. Sequence evolving adaptively is functional but not constrained.
4. Positive selection does not necessarily imply adaptive evolution: e.g. clonal selection for germ-line cells

Gene sequence conservation

The exonic structures of essentially all human genes (major transcripts) are conserved in mouse.

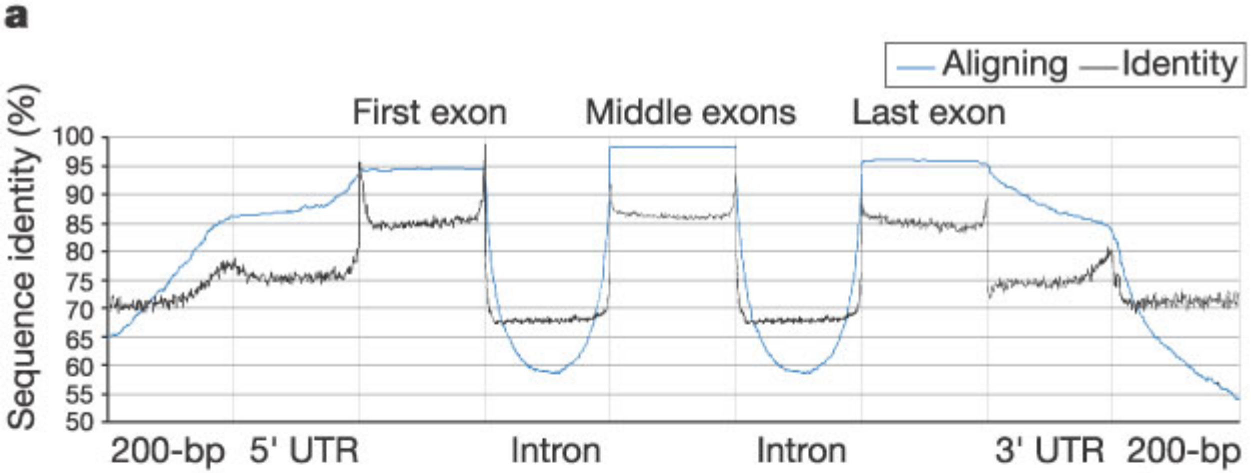
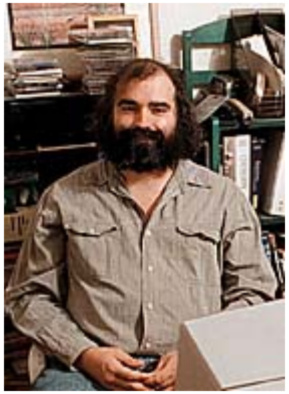
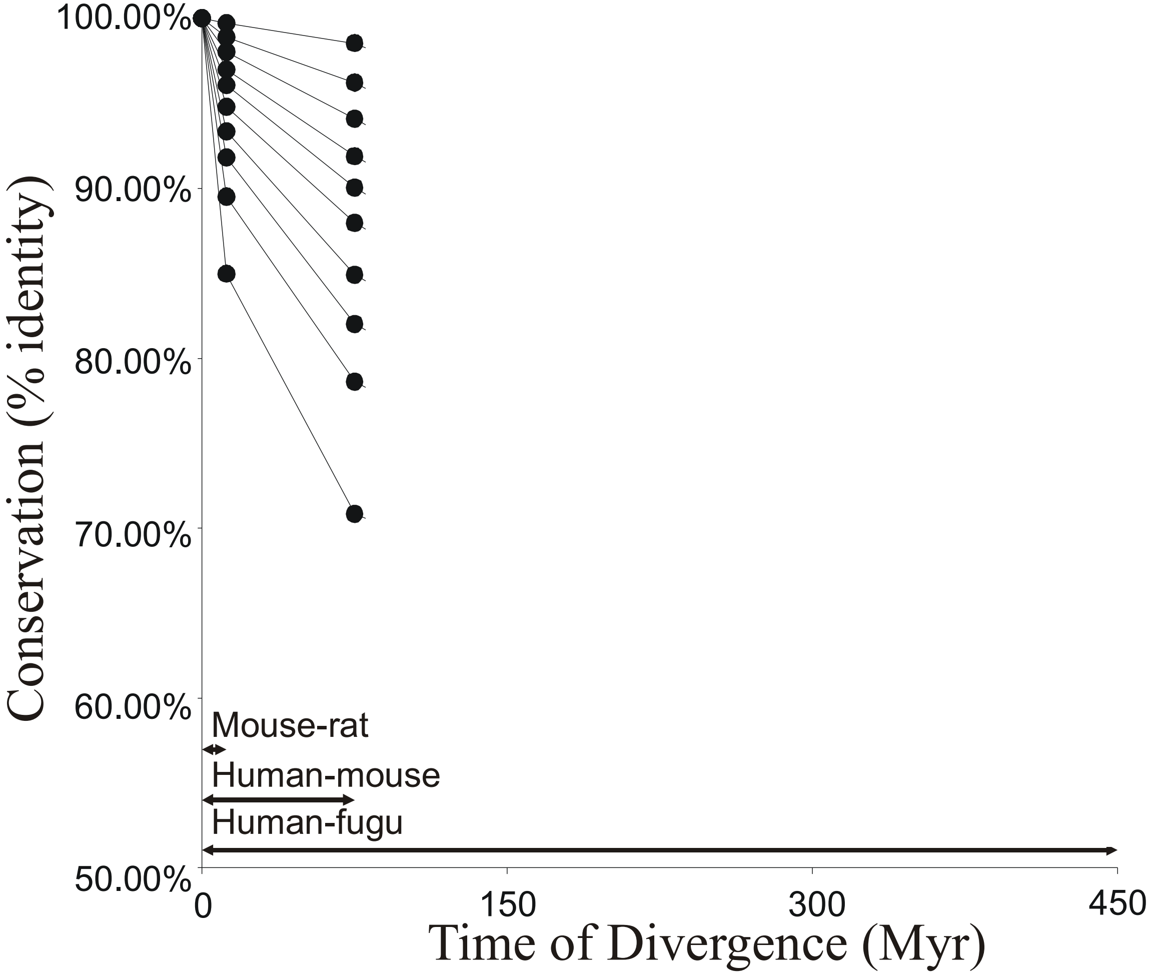


Figure 25. Sequence conservation between mouse and human genes
Mouse genome paper *Nature* 420, 520-562



Gene Sequence Conservation is Clock-like



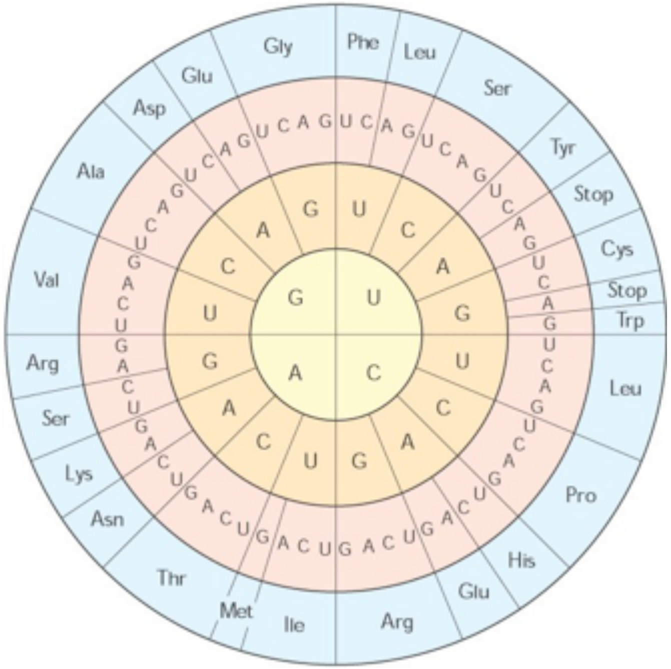
A model of neutral evolution

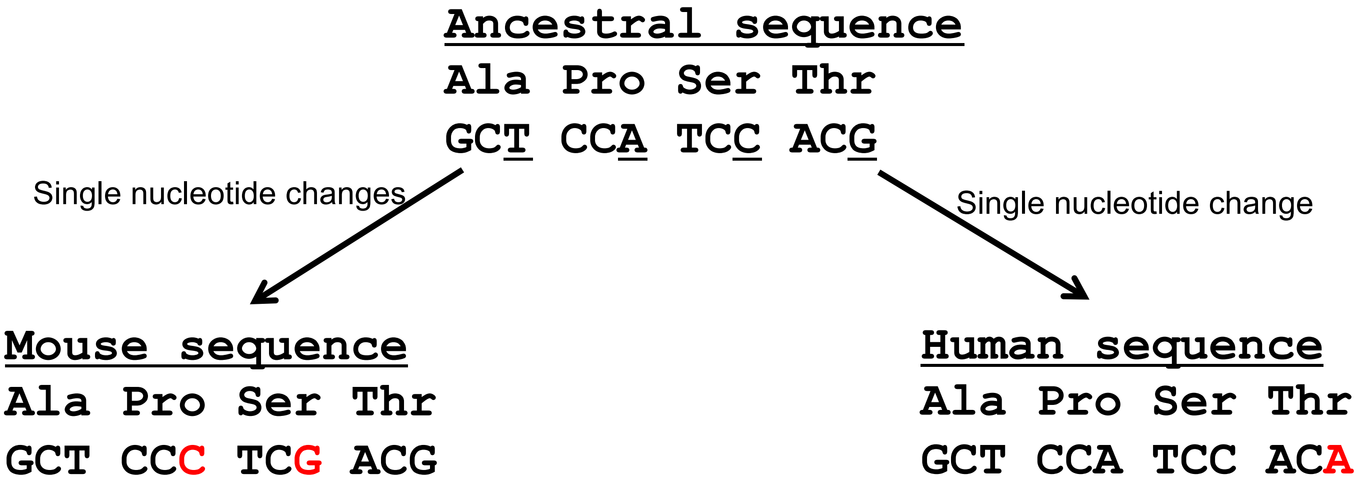
- d_S – the number of synonymous substitutions per synonymous site
- takes advantage of the redundant genetic code
- 4D sites GCx (ALA), CCx (PRO), TCx (SER), ACx (THR), CGx (ARG), GGx (GLY), CTx (LEU), GTx (VAL)
- "how much would a gene's sequence have changed if selection had <u>not</u> acted upon it?"

Hearing silence: non-neutral evolution at synonymous sites in mammals

J. V. Chamary*, Joanna L. Parmley‡ and Laurence D. Hurst‡

Nature Reviews Genetics 7, 98





Ancestral sequence

Ala Pro Ser Thr

GCT CCA TCC ACG

Single nucleotide changes

Single nucleotide change

Mouse sequence

Ala Pro Ser Thr

GCT CCC TCG ACG

Human sequence

Ala Pro Ser Thr

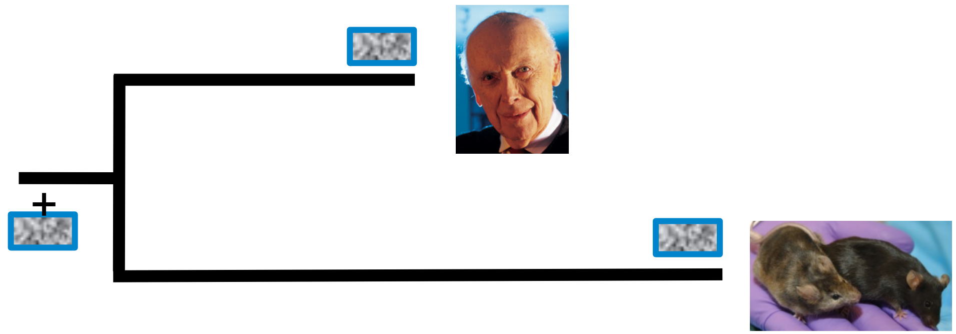
GCT CCA TCC ACA

3 nucleotide substitutions at 4 synonymous sites

d_S = ¾ or 0.75

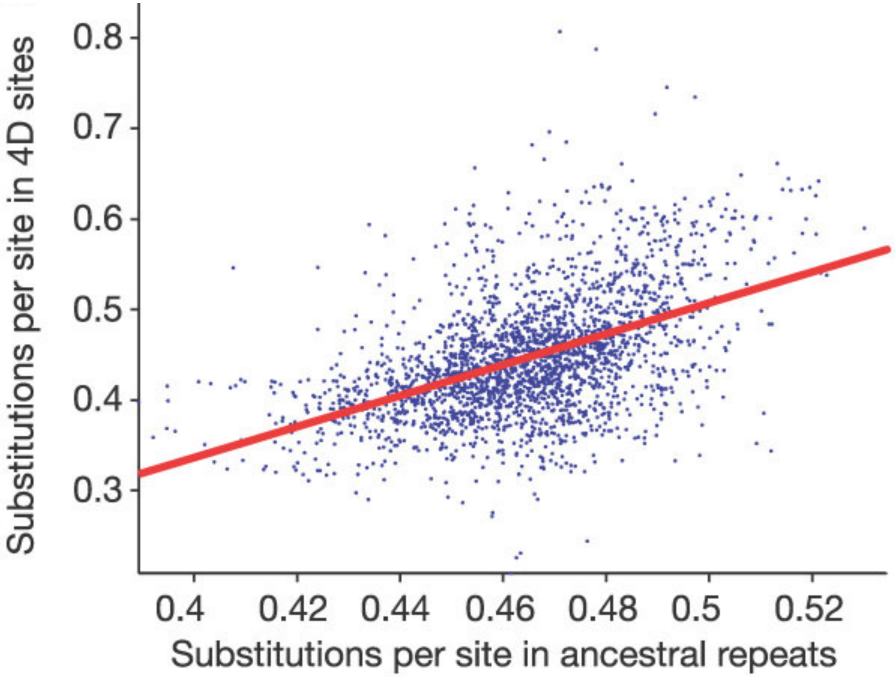
“Ancestral repeats”

- Transposable element-derived sequence that inserted prior to the last common ancestor of human and mouse.



- It is commonly assumed that evolution of Ancestral Repeats (ARs) has been neutral.

Neutral Rates Vary According to Location

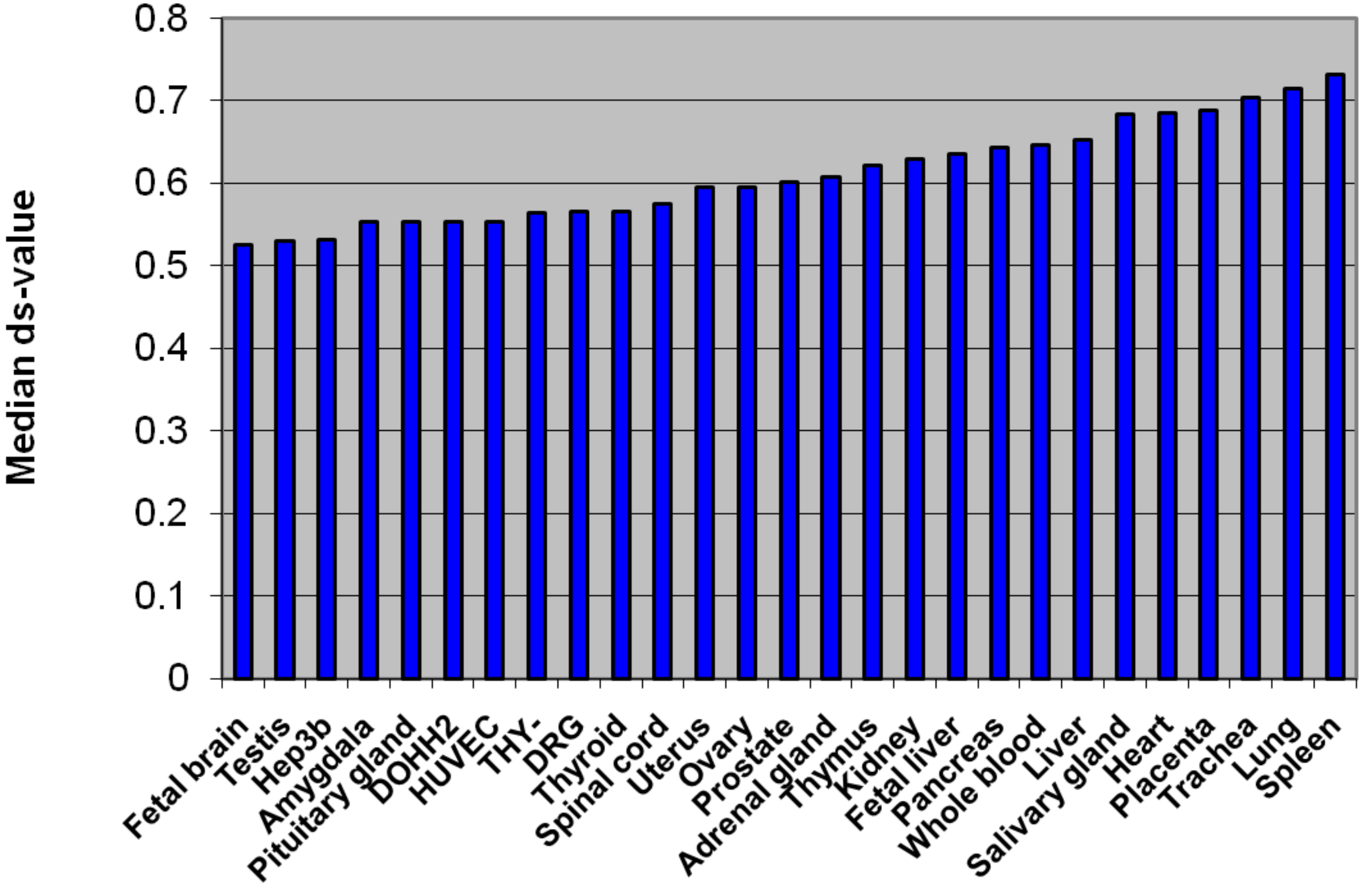


see also
Hardison et al.
Genome Res. 2003
13: 13-26.

Variation in rates of mutation and/or rates of repair?

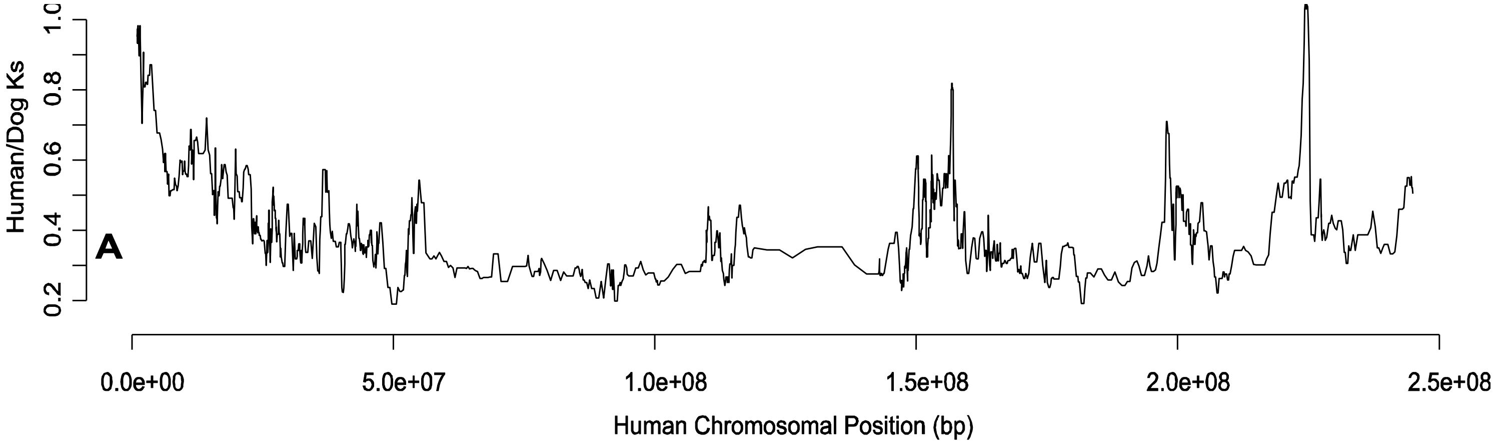
- Transcription-associated mutational strand asymmetry (Phil Green et al. Nature Genetics 33: 514-7)
- Associated with transcription-coupled repair processes (Majewski, Am J Human Genet 73, 688-692)
- Genes transcribed in the germline at high levels, when mutated, are repaired more readily, than those not transcribed in the germline.
- Majewski estimates that 71%-91% of genes are transcribed in the germline!

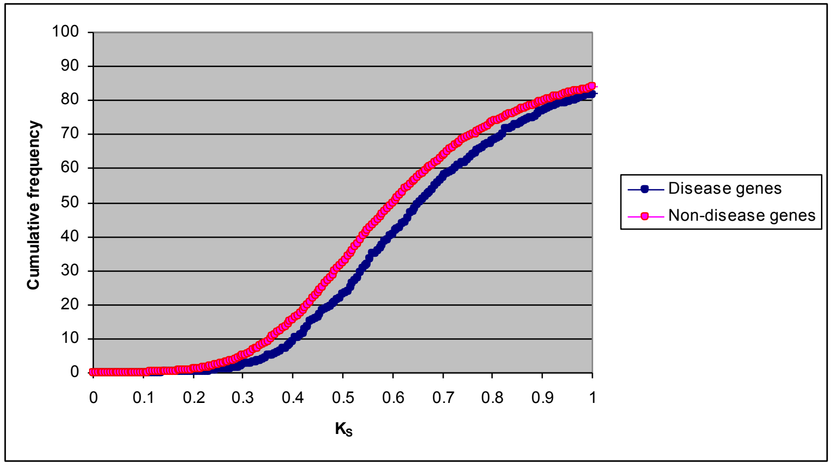
Tissue-specific genes' d_s

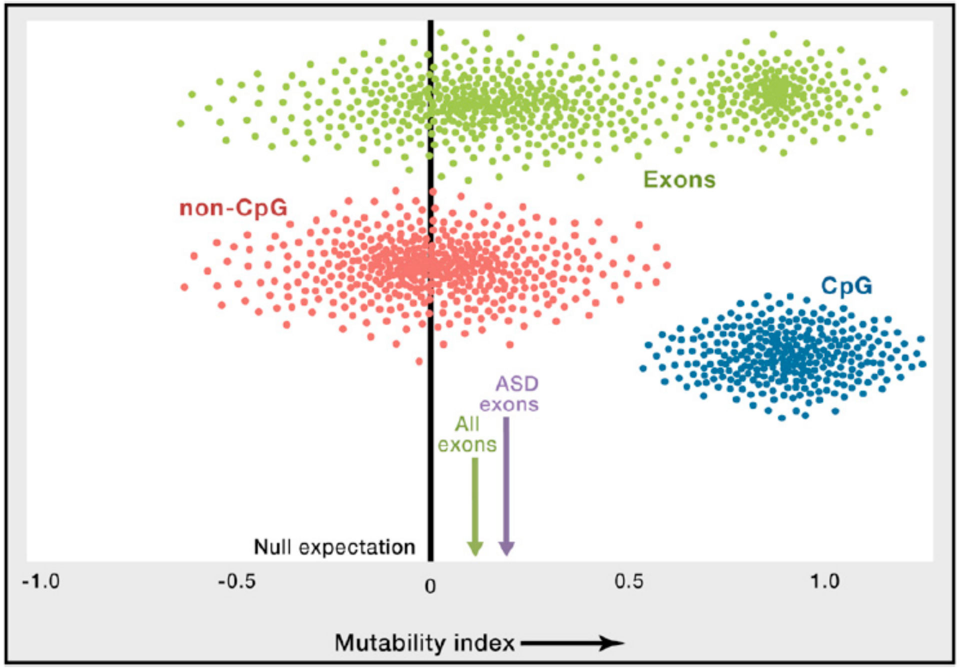


Winter et al. *Genome Research* 14:54-61, 2004

d_s variation







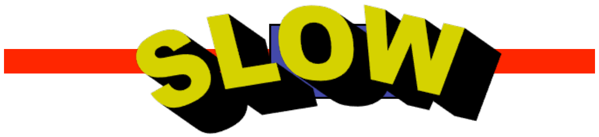
Leading Edge
Previews

Cell

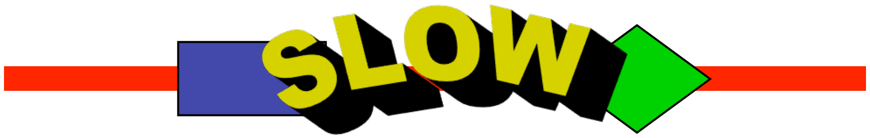
Loaded Dice for Human Genome Mutation

Modes of Protein Evolution

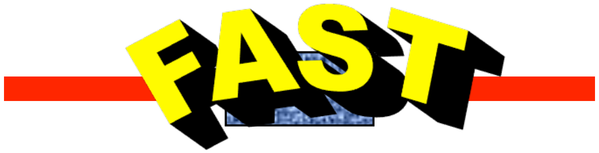
- *De novo* creation



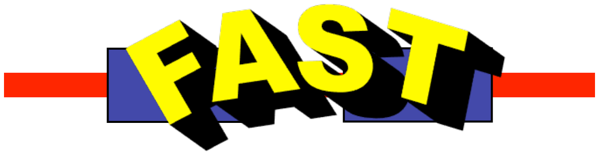
- Gene fusion / fission



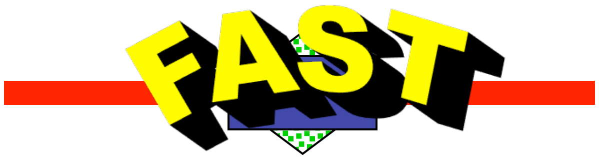
- Rapid sequence change



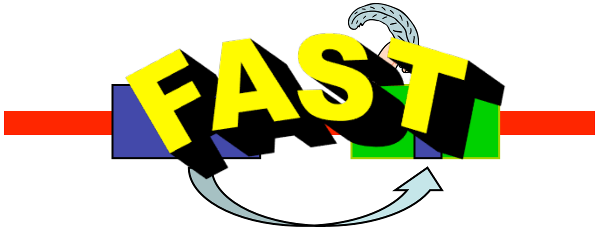
- Gene duplication



- Pseudogenisation



- Gene conversion

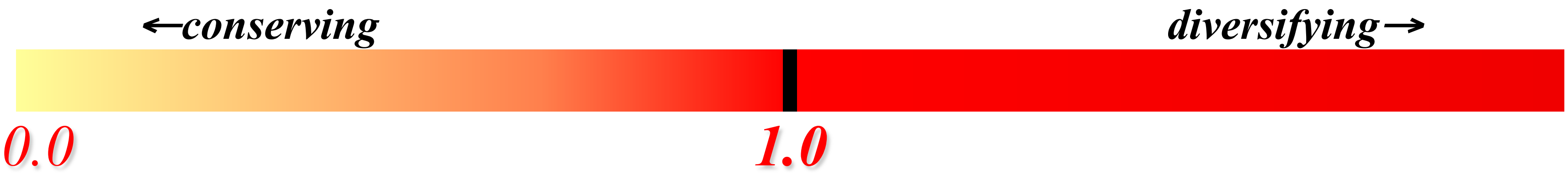


A model for non-neutral evolution

- d_N – the number of non-synonymous (amino acid changing) substitutions per non-synonymous site

- What proportion of possible amino acid-changing substitutions has occurred?

- dN/dS, ω —
  A model of selective pressure



Slowly & rapidly-evolving proteins

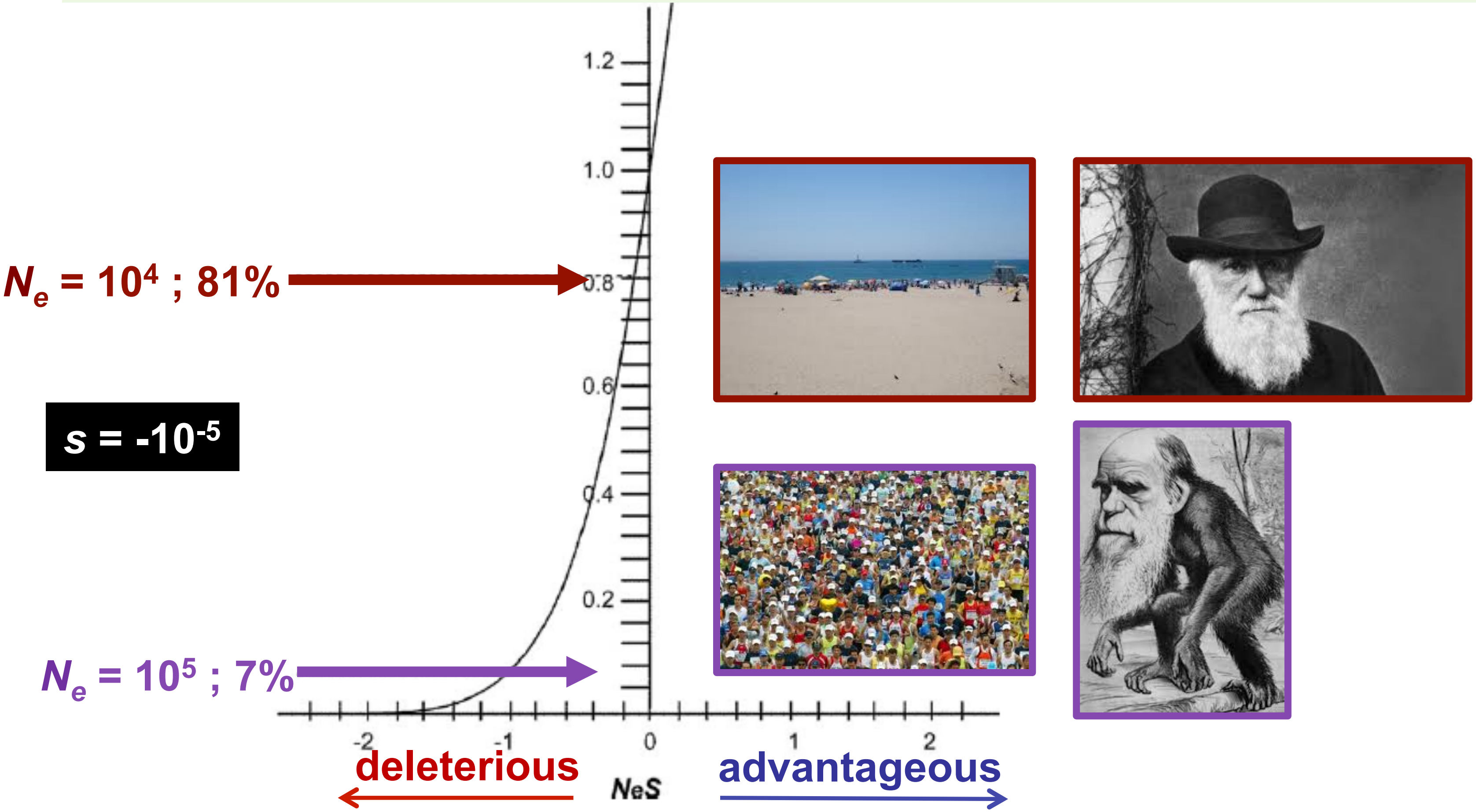
Slow (dN/dS is small ~0.1)

- Developmental genes
- Brain-expressed genes
- Big genes with many regulatory elements
- Genes that have escaped being duplicated over many tens of millions of years
- Domain structures
- Catalytic domains
- Intracellular proteins

Rapid (dN/dS is larger > 0.25)

- Environmental genes
- Testis-expressed genes
- Single exon genes
- Genes frequently duplicated or deleted over evolutionary time
- Unstructured regions
- Non-enzymes
- Extracellular proteins

Fixation probability of a deleterious allele: effect of N_e



$N_e = 10^4$; 81%

$s = -10^{-5}$

$N_e = 10^5$; 7%

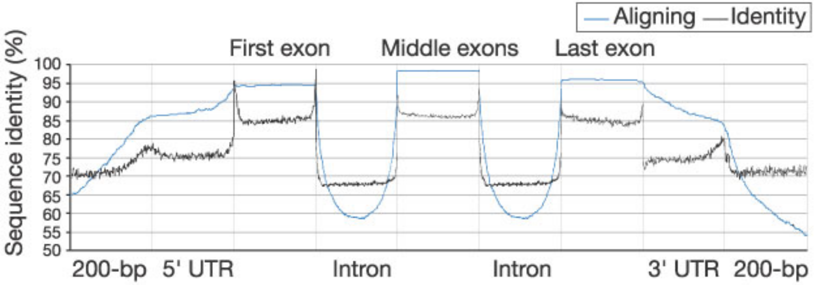
Fig. 1. The probability of fixation of a new variant with respect to the neutral expectation of $1/(2N)$ graphed as a function of the product of the effective population size and the selection coefficient of the variant ($N_e \times s$). The dashed lines represent cases in which $s = -10^{-5}$ but N_e takes on different values, either 10,000 (upper dashed line) or 100,000 (lower dashed line).

Nonadaptive Processes in Primate and Human Evolution

Eugene E. Harris*

YEARBOOK OF PHYSICAL ANTHROPOLOGY 53:13–45 (2010)

Positive Selection, dN/dS > 1



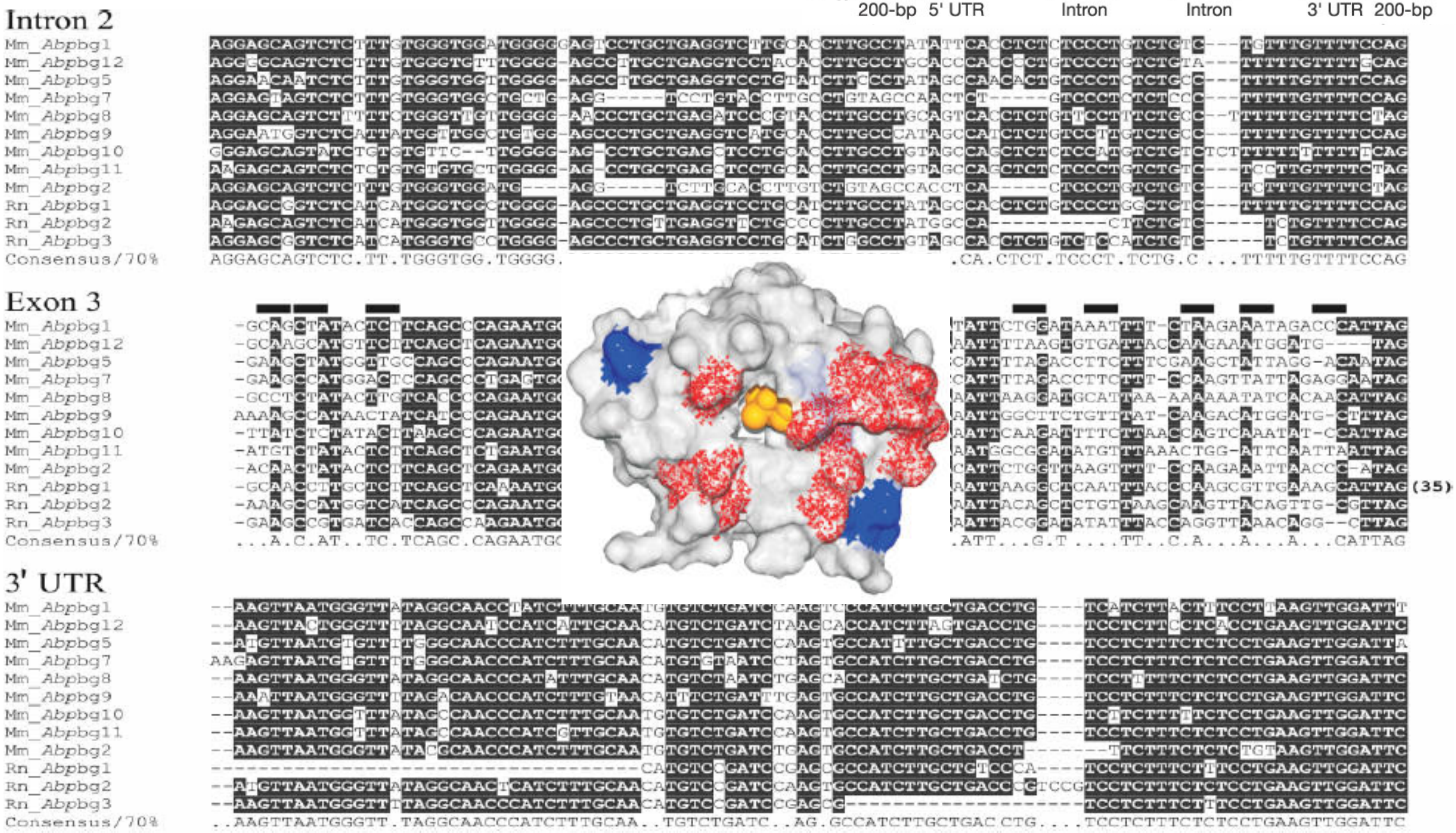
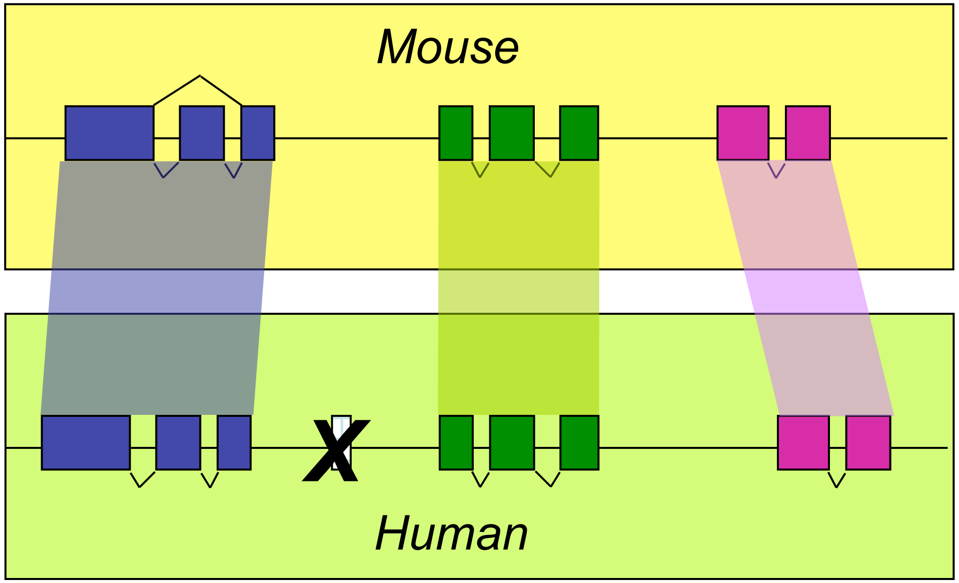


Figure 4 Multiple nucleotide sequence alignment of mouse and rat *Abpbg*-like exons 3 and surrounding genomic DNA. Genomic DNA corresponding to exon 3 (98 positions) and 100 nucleotide positions of both flanking intronic and 3′-UTR sequence was aligned with HMMER, and manually adjusted. We found that 81.3%, 50.5%, and 92.6% of the sites in the intron, exon, and 3′-UTR, respectively, exhibited ≥70% consensus. In these calculations, positions with fewer than 50% gaps were considered. The 14 codons of exon 3 corresponding to predicted ω⁺ sites are shown by horizontal bars.

Mouse & Human: Protein Coding Gene Census



PloS Biology♪
May 2009♪

- Mouse gene count = 20,210; Human gene count = 19,042.
- Captures only genes that have homologues in one or the other genome.
- Captures duplicates (that preserve exon structure).
- Misses fast evolvers.
- Doesn't consider copy number variable genes.

75% (80%) of Mouse (Human) Genes have a single orthologue in Human (Mouse)

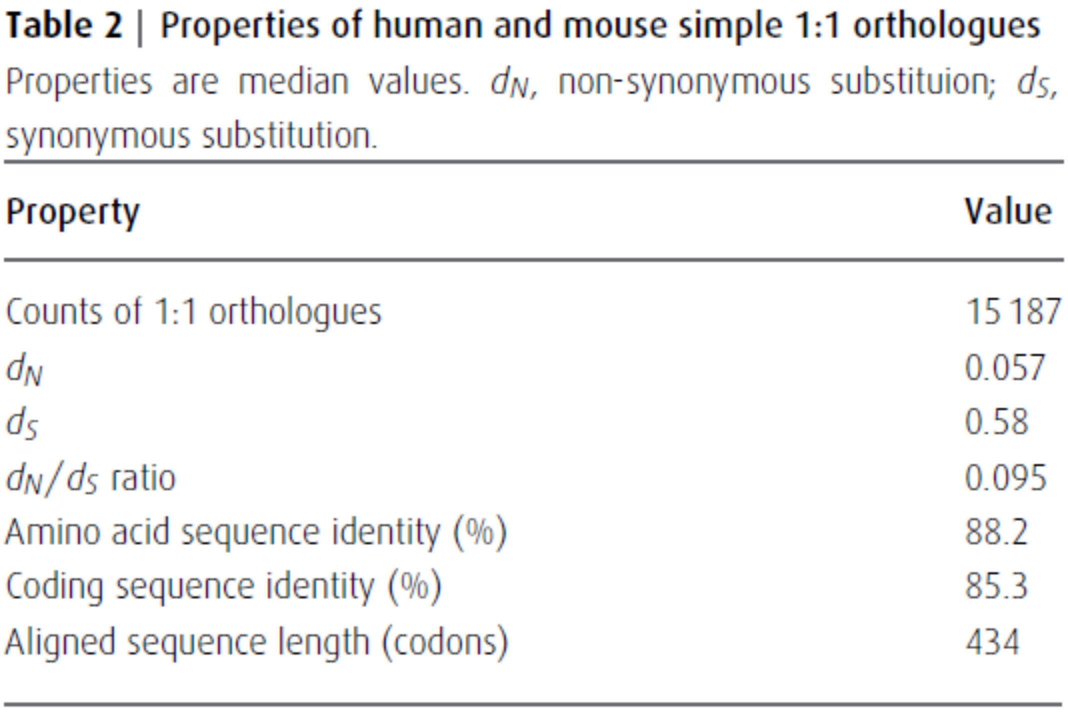
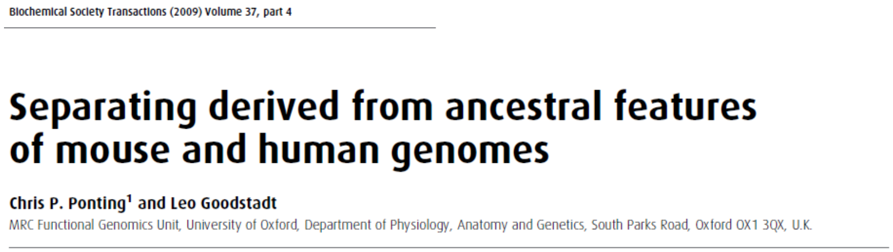


Table 2 | Properties of human and mouse simple 1:1 orthologues
Properties are median values. d_N, non-synonymous substituion; d_S, synonymous substitution.

Property	Value
Counts of 1:1 orthologues	15 187
d_N	0.057
d_S	0.58
d_N/d_S ratio	0.095
Amino acid sequence identity (%)	88.2
Coding sequence identity (%)	85.3
Aligned sequence length (codons)	434



Biochemical Society Transactions (2009) Volume 37, part 4

Separating derived from ancestral features of mouse and human genomes

Chris P. Ponting[1] and Leo Goodstadt
MRC Functional Genomics Unit, University of Oxford, Department of Physiology, Anatomy and Genetics, South Parks Road, Oxford OX1 3QX, U.K.

nature

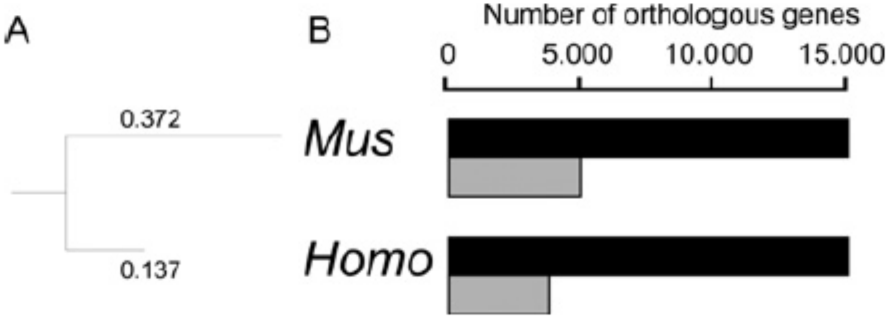
the mouse **genome**

The 2.5-Gb mouse genome sequence reported on page 520, from the C57BL/6J strain, reveals about 30,000 genes, with 99% having direct counterparts in humans.

Since the publication of the human genome, the scientific community has been eagerly awaiting the results of the mouse genome sequencing project. This week's issue contains a landmark publication from the Mouse Genome Sequencing Consortium that many say holds more promise for our future than even

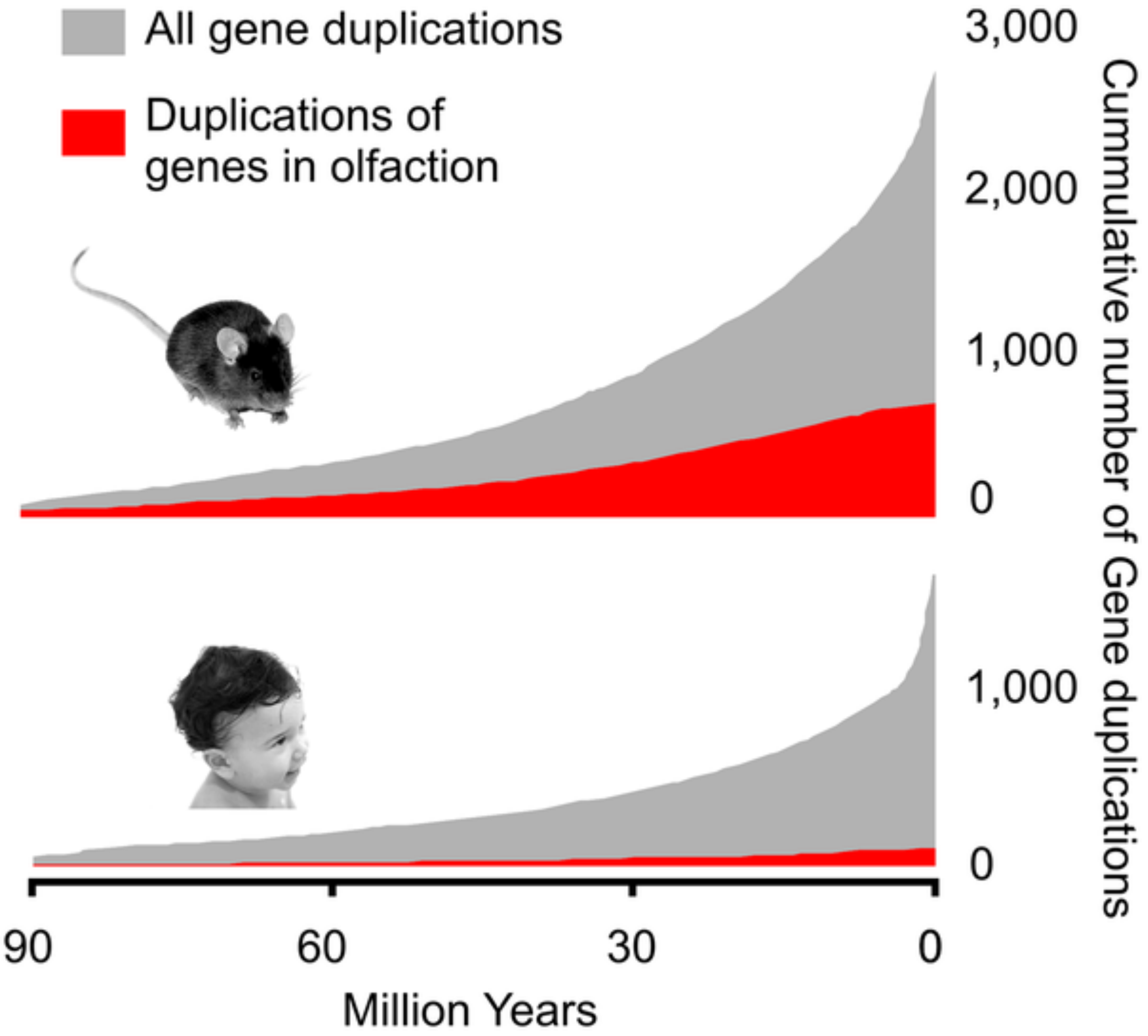
Figure 2 | Mouse genes have a higher synonymous nucleotide substitution rate (d_S) and have accumulated more lineage-specific duplicates than human genes

(**A**) Mouse and human phylogeny drawn to the d_S scale. (**B**) The number of 1:1 mouse and human orthologues (black) and the number of gene duplicates unique to each species (grey).

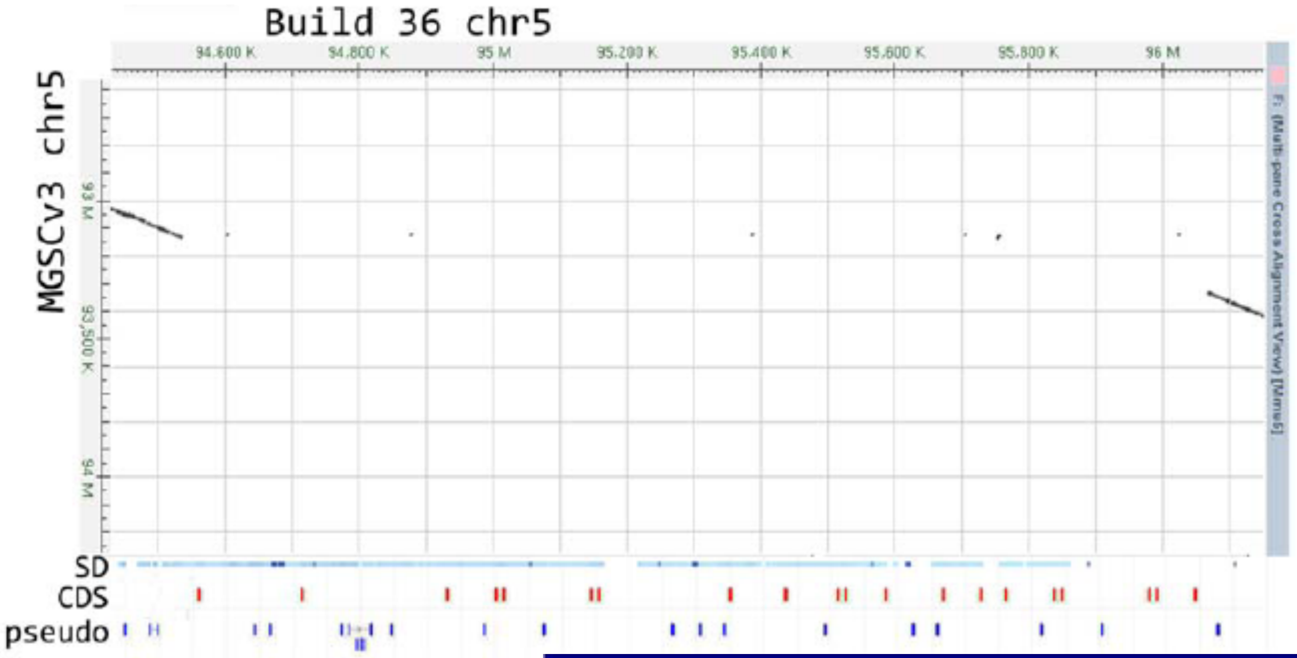


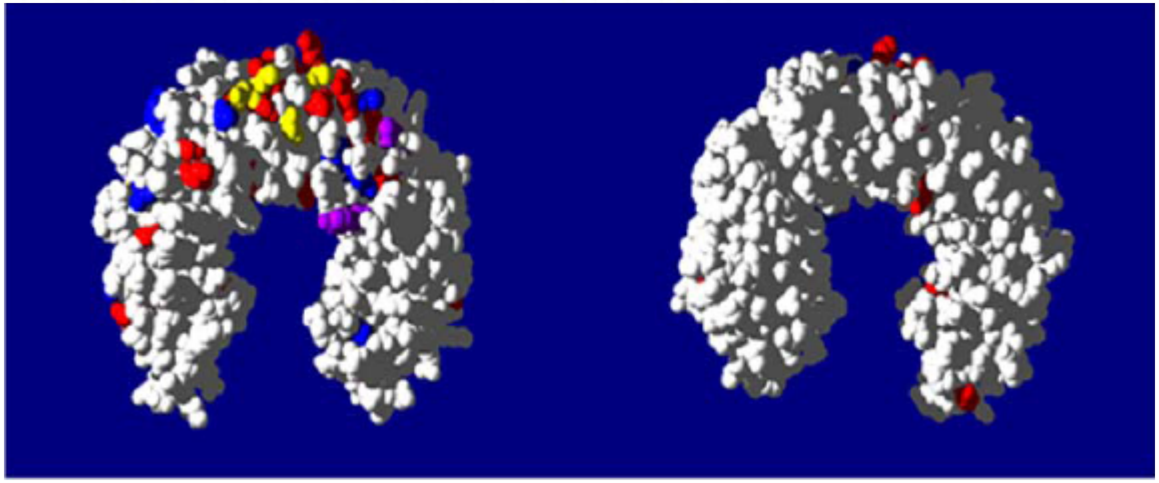
Ritu Dhand Chief Biology Editor

Lineage-specific paralogues



Consecutive highly-similar gene sequences hinder genome assembly





Mouse Segmental Duplications are mainly in *cis*

Segmental duplications: >1 kb fragments of genomic sequence with high sequence identity (>90%) that map to multiple locations

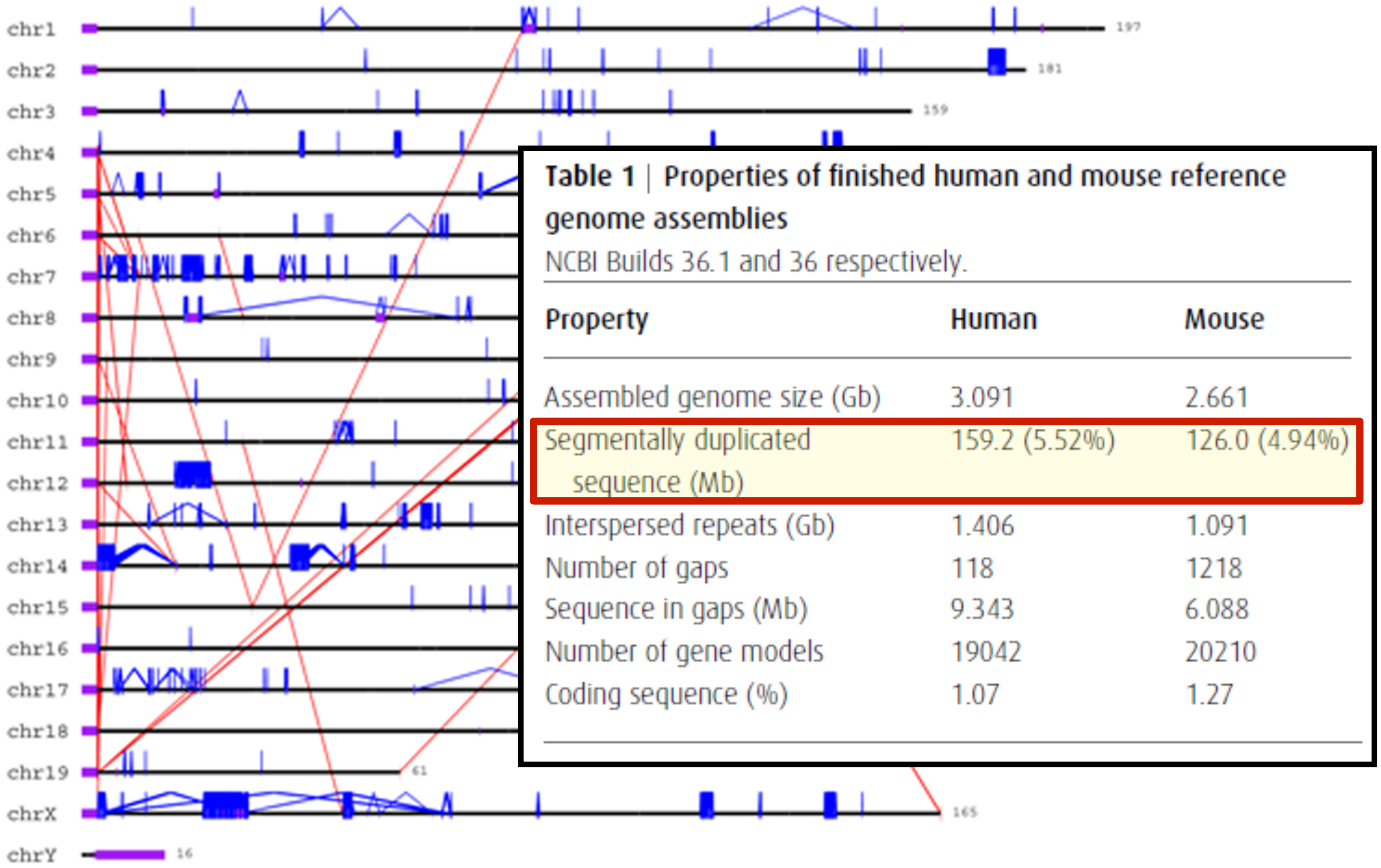


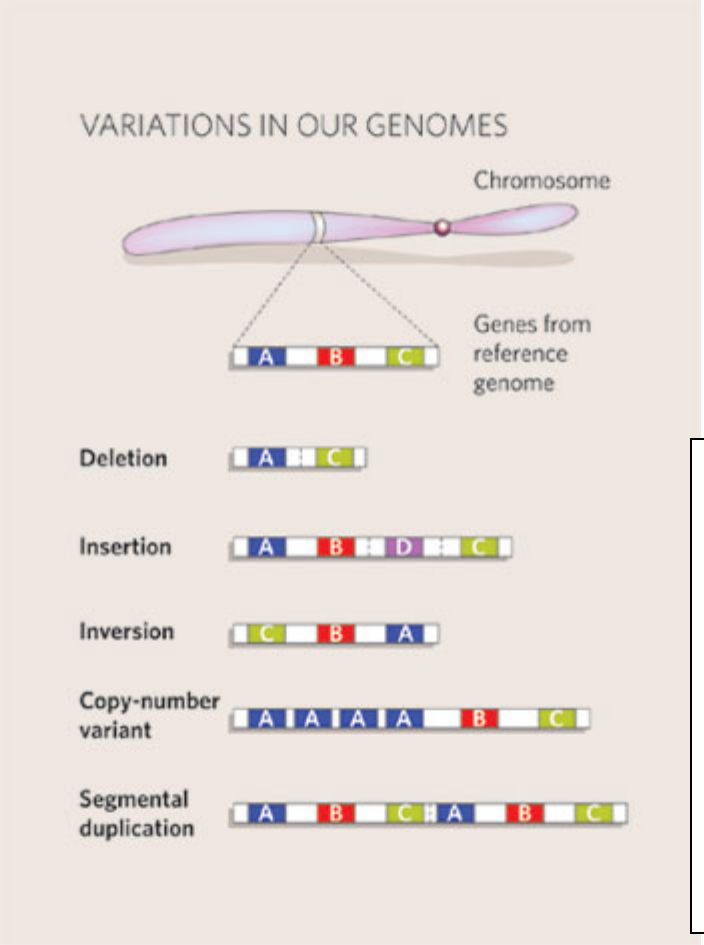
Table 1 | **Properties of finished human and mouse reference genome assemblies**

NCBI Builds 36.1 and 36 respectively.

Property	Human	Mouse
Assembled genome size (Gb)	3.091	2.661
Segmentally duplicated sequence (Mb)	159.2 (5.52%)	126.0 (4.94%)
Interspersed repeats (Gb)	1.406	1.091
Number of gaps	118	1218
Sequence in gaps (Mb)	9.343	6.088
Number of gene models	19042	20210
Coding sequence (%)	1.07	1.27

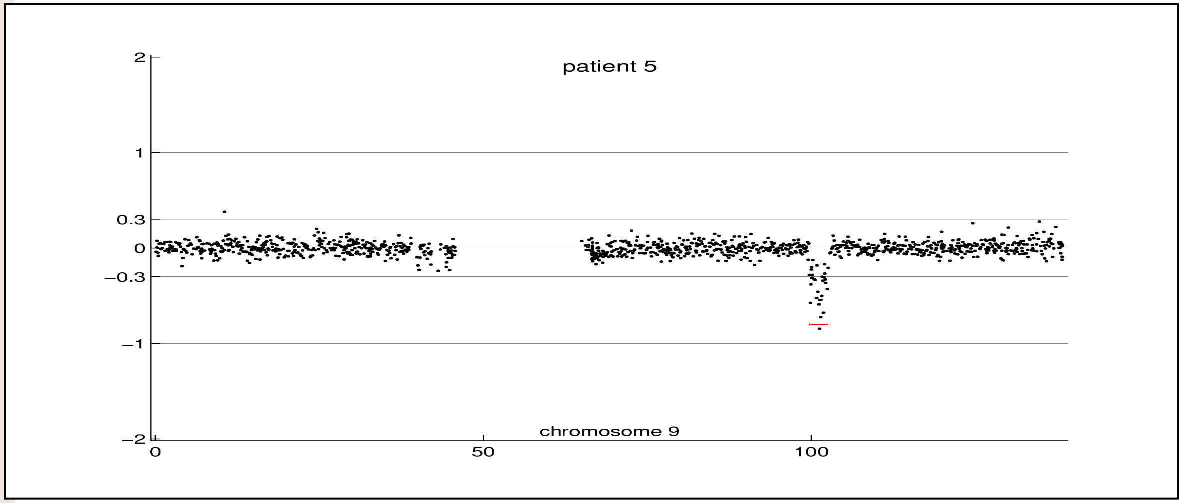
Build36

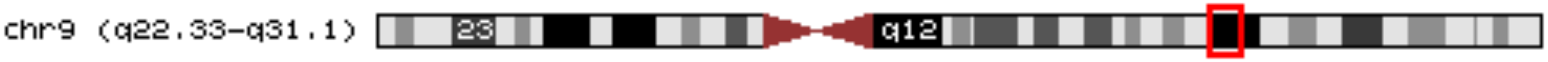
Copy Number Variants (CNVs)



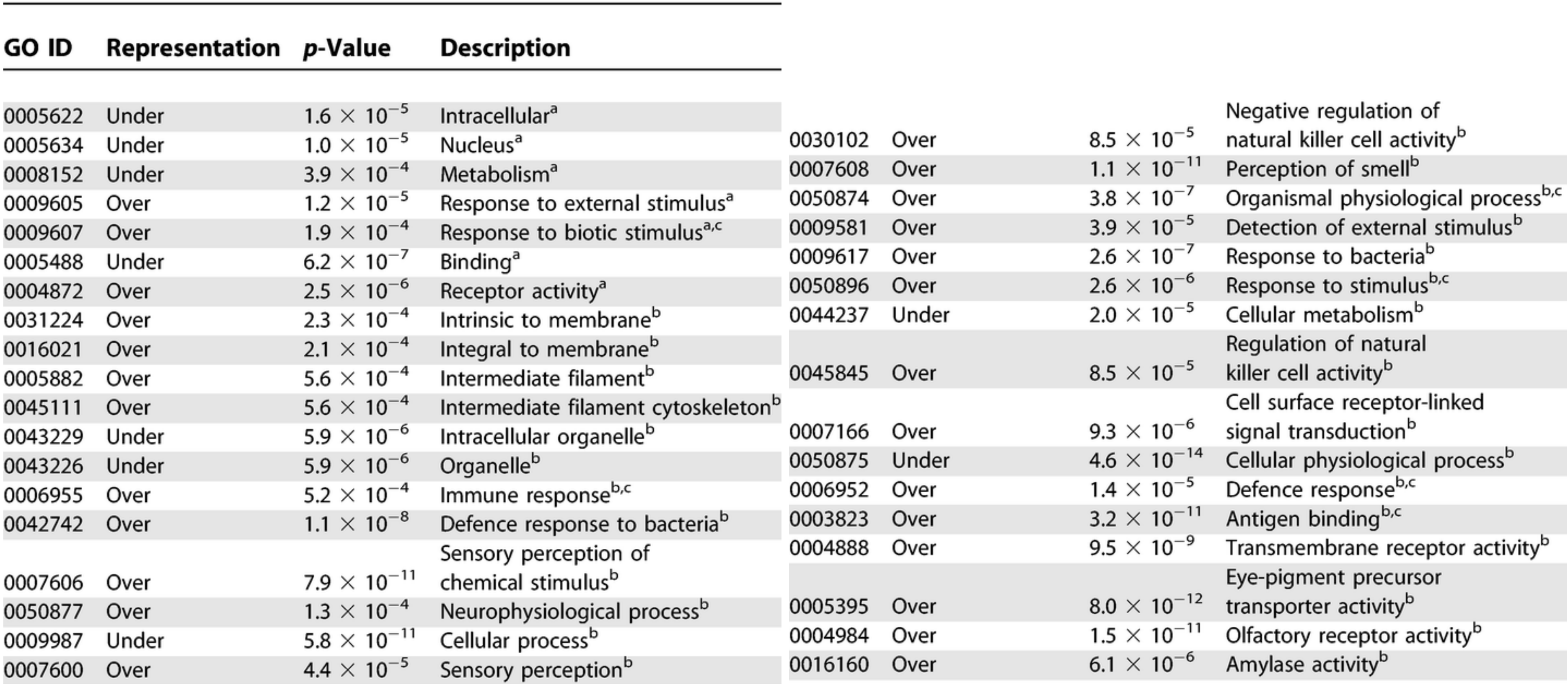
More bases differ in CNVs between individual genomes than they do in SNPs.

Segmental duplications and CNVs often coincide.





Immunity, defence, chemosensation genes

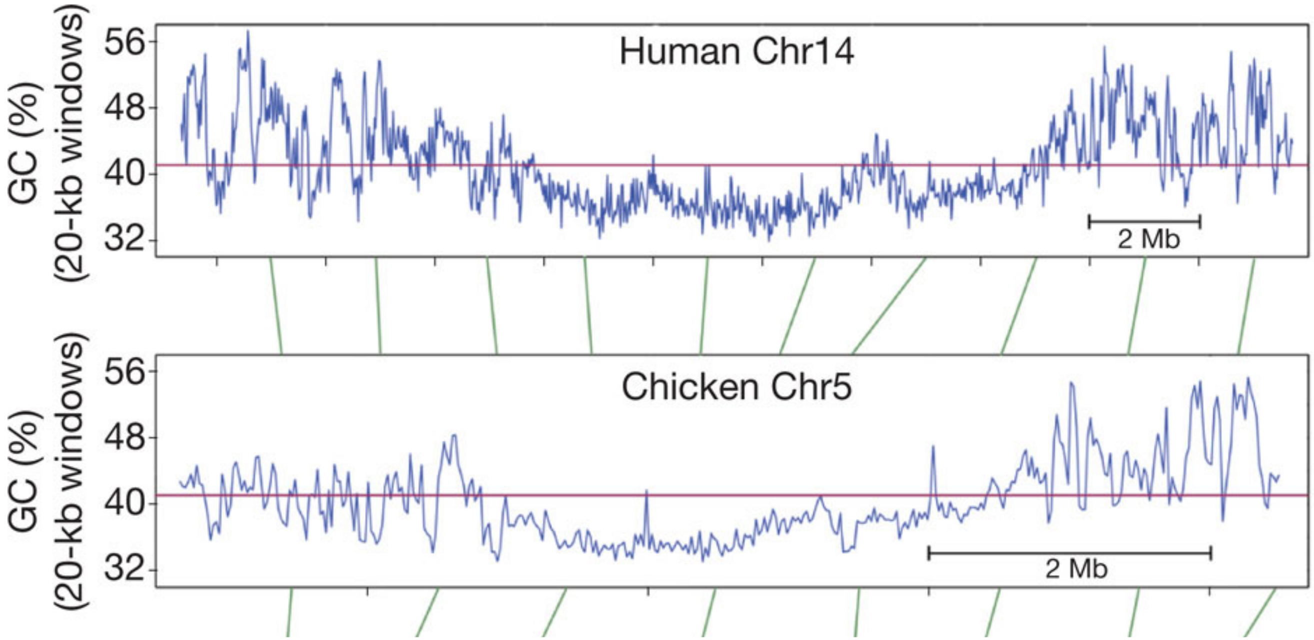


GO ID	Representation	*p*-Value	Description
0005622	Under	1.6×10^{-5}	Intracellular[a]
0005634	Under	1.0×10^{-5}	Nucleus[a]
0008152	Under	3.9×10^{-4}	Metabolism[a]
0009605	Over	1.2×10^{-5}	Response to external stimulus[a]
0009607	Over	1.9×10^{-4}	Response to biotic stimulus[a,c]
0005488	Under	6.2×10^{-7}	Binding[a]
0004872	Over	2.5×10^{-6}	Receptor activity[a]
0031224	Over	2.3×10^{-4}	Intrinsic to membrane[b]
0016021	Over	2.1×10^{-4}	Integral to membrane[b]
0005882	Over	5.6×10^{-4}	Intermediate filament[b]
0045111	Over	5.6×10^{-4}	Intermediate filament cytoskeleton[b]
0043229	Under	5.9×10^{-6}	Intracellular organelle[b]
0043226	Under	5.9×10^{-6}	Organelle[b]
0006955	Over	5.2×10^{-4}	Immune response[b,c]
0042742	Over	1.1×10^{-8}	Defence response to bacteria[b]
0007606	Over	7.9×10^{-11}	Sensory perception of chemical stimulus[b]
0050877	Over	1.3×10^{-4}	Neurophysiological process[b]
0009987	Under	5.8×10^{-11}	Cellular process[b]
0007600	Over	4.4×10^{-5}	Sensory perception[b]
0030102	Over	8.5×10^{-5}	Negative regulation of natural killer cell activity[b]
0007608	Over	1.1×10^{-11}	Perception of smell[b]
0050874	Over	3.8×10^{-7}	Organismal physiological process[b,c]
0009581	Over	3.9×10^{-5}	Detection of external stimulus[b]
0009617	Over	2.6×10^{-7}	Response to bacteria[b]
0050896	Over	2.6×10^{-6}	Response to stimulus[b,c]
0044237	Under	2.0×10^{-5}	Cellular metabolism[b]
0045845	Over	8.5×10^{-5}	Regulation of natural killer cell activity[b]
0007166	Over	9.3×10^{-6}	Cell surface receptor-linked signal transduction[b]
0050875	Under	4.6×10^{-14}	Cellular physiological process[b]
0006952	Over	1.4×10^{-5}	Defence response[b,c]
0003823	Over	3.2×10^{-11}	Antigen binding[b,c]
0004888	Over	9.5×10^{-9}	Transmembrane receptor activity[b]
0005395	Over	8.0×10^{-12}	Eye-pigment precursor transporter activity[b]
0004984	Over	1.5×10^{-11}	Olfactory receptor activity[b]
0016160	Over	6.1×10^{-6}	Amylase activity[b]

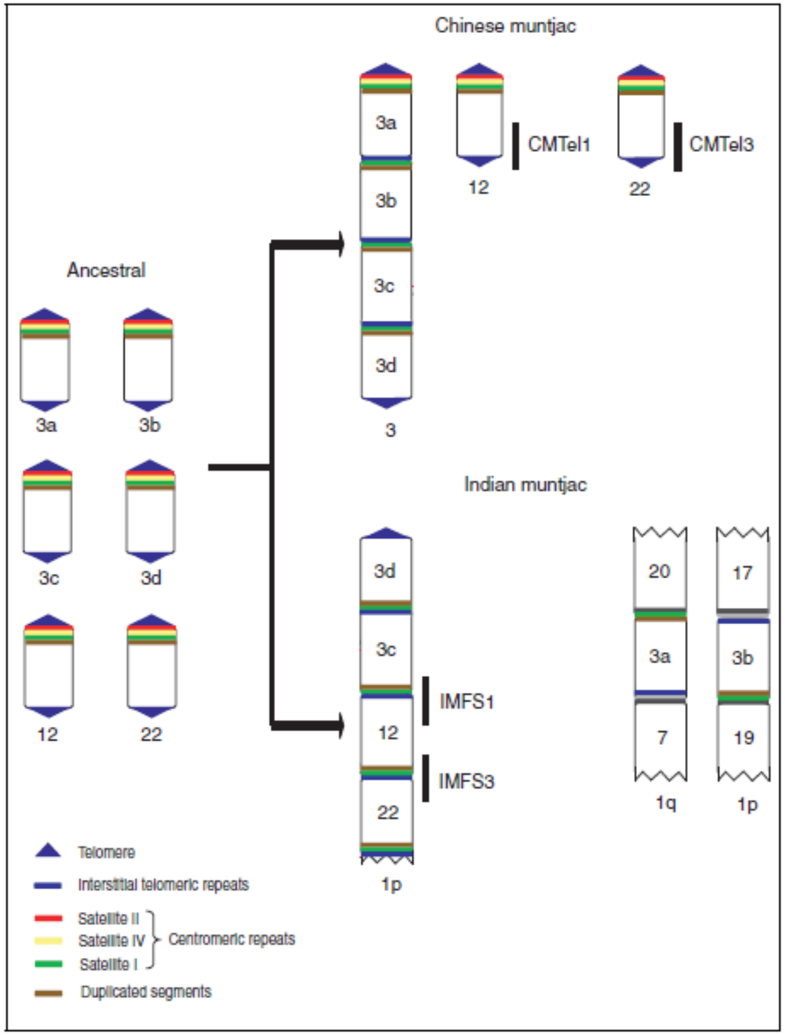
Little evidence that common CNVs are either adaptive or are associated with disease.

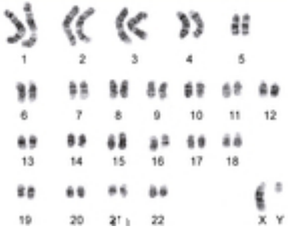
Evolutionary questions without adequate answers

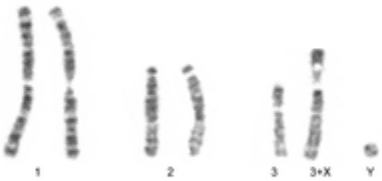
(Non-)Conservation of Isochores – Why?



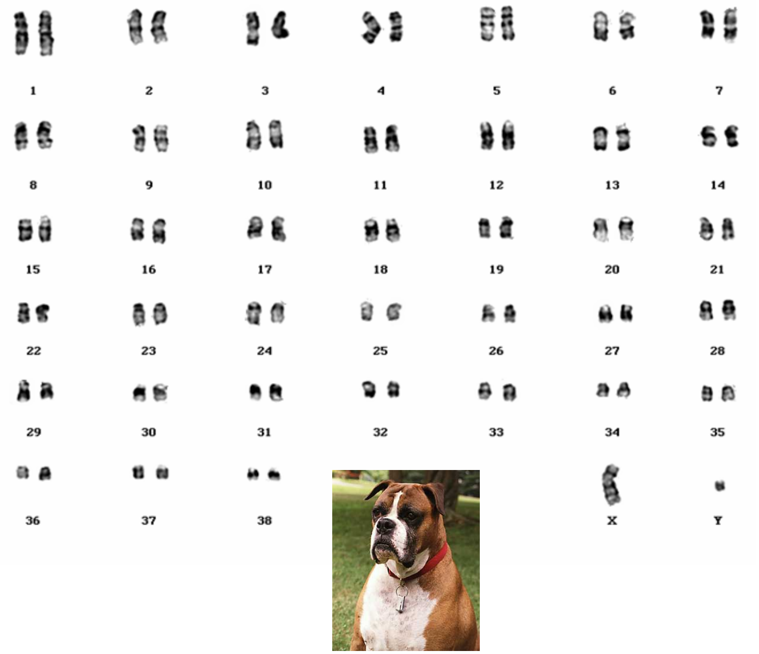
Why rapid variations in karyotypes?

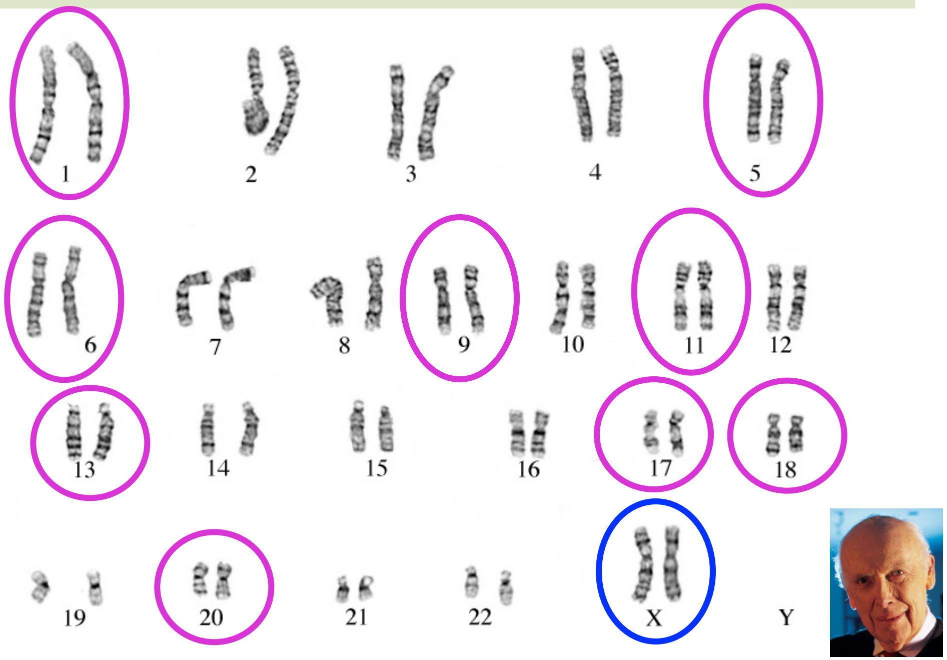






Mammalian Karyotypes





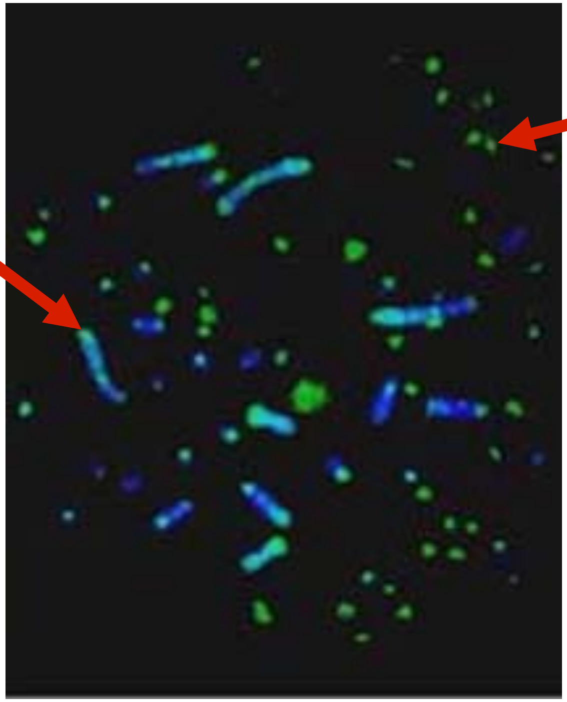
Ancestral chromosomes

Why do birds & lizards have small (micro) chromosomes?

Large (macro-) chromosomes:

more DNA but lower gene density;

lower mutation rate; lower G+C

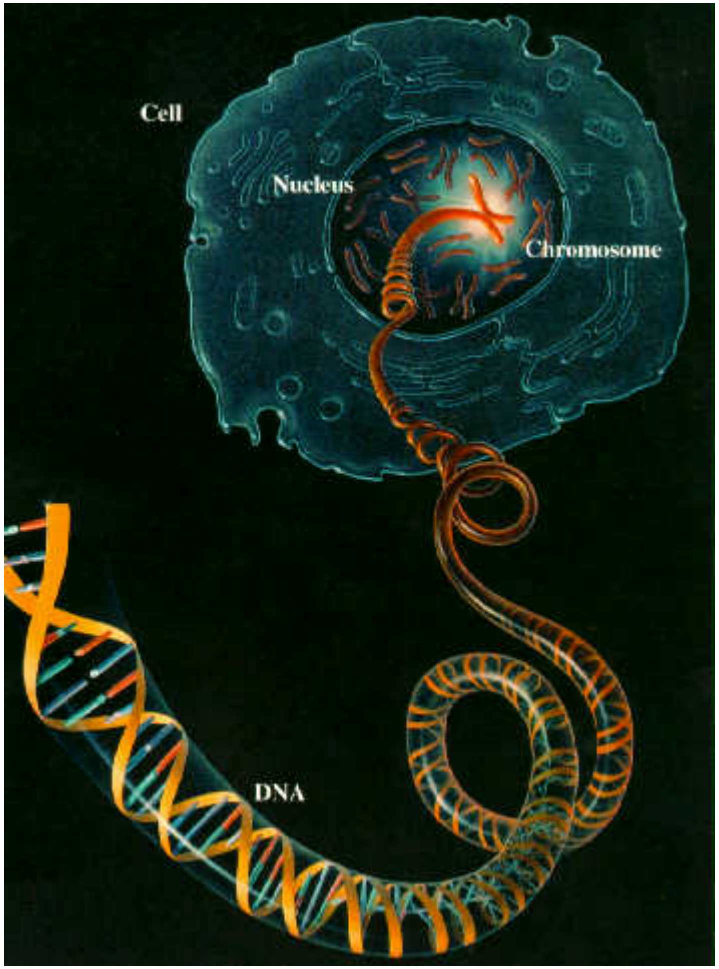


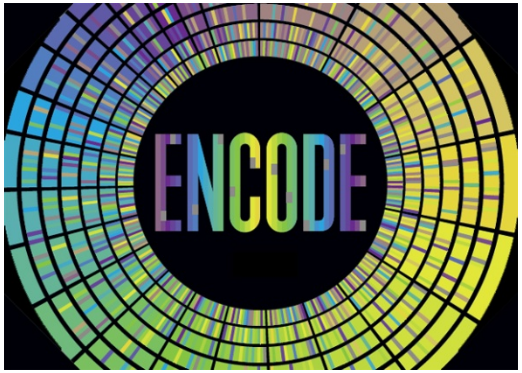
Small (micro-) chromosomes:

25% of the DNA but half the genes;

higher mutation rate; higher G+C

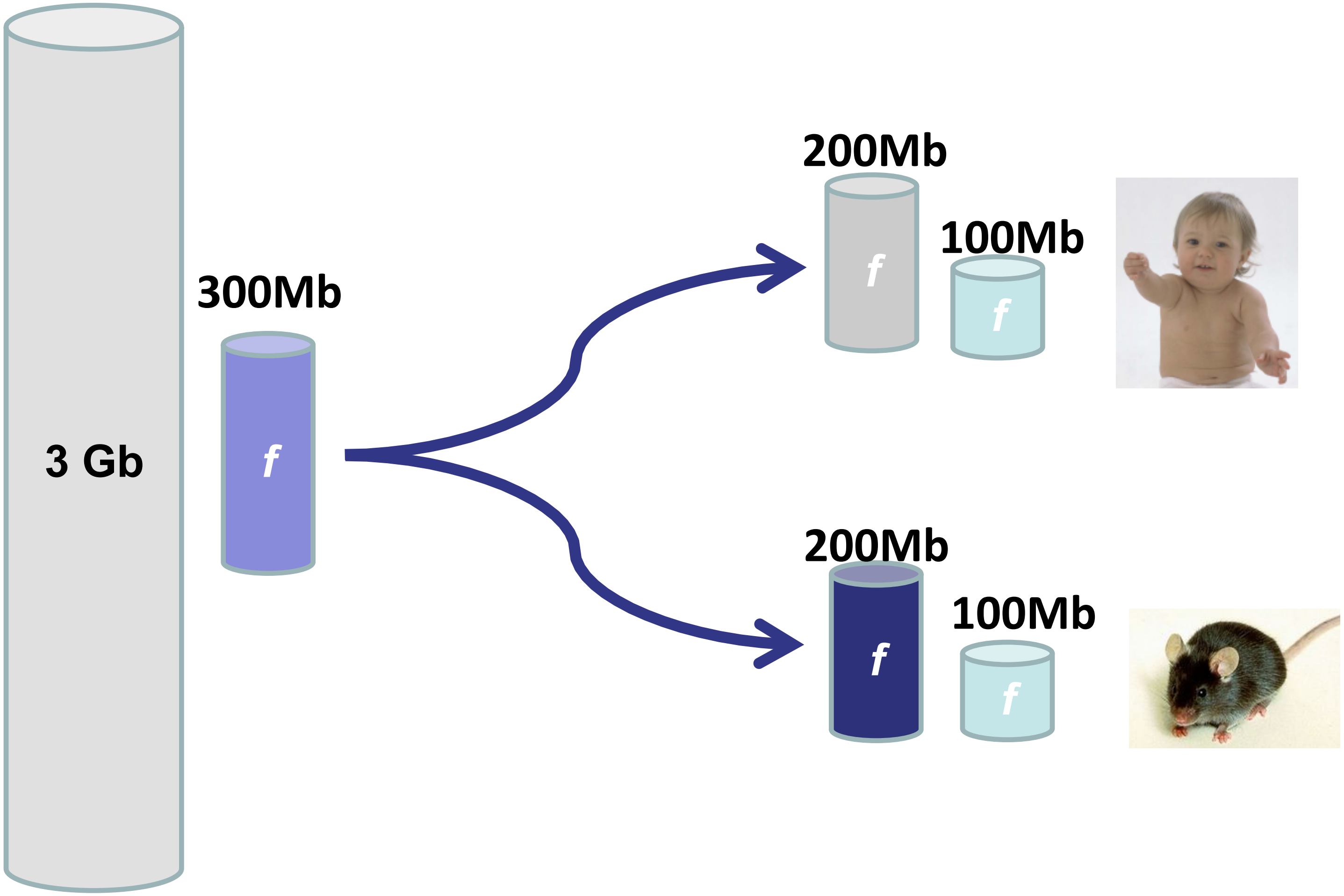
Part 2: functional DNA & transcript maps





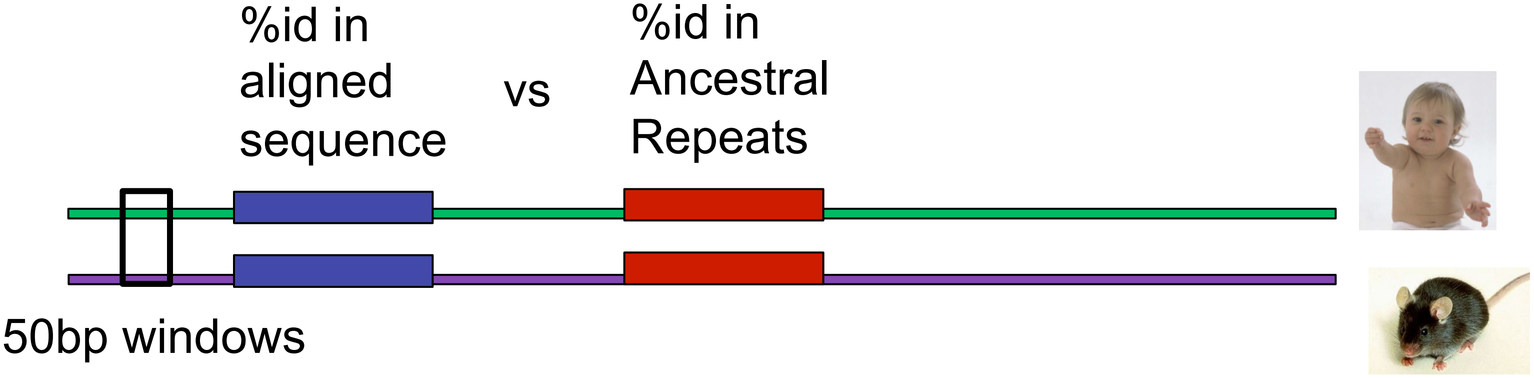
How much of our genome is biologically functional?

My view: 10% ...



Mouse Genome Paper 2002
Nucleotide Substitution Model

By comparing the extent of genome-wide sequence conservation to the neutral rate, the proportion of small (50–100 bp) segments in the mammalian genome that is under (purifying) selection can be estimated to be about 5%



%id in aligned sequence

vs

%id in Ancestral Repeats

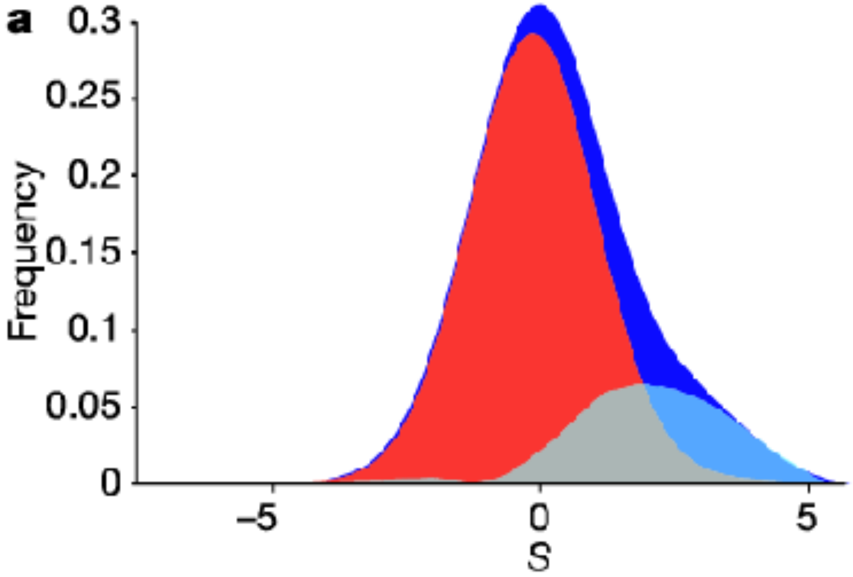
50bp windows

The Share of Human Genomic DNA under Selection Estimated from Human–Mouse Genomic Alignments

F. Chiaromonte,* R.J. Weber,† K.M. Roskin,† M. Diekhans,† W.J. Kent,†
and D. Haussler‡

*Department of Statistics and Department of Health Evaluation Sciences, Pennsylvania State University, University Park, Pennsylvania 16803; †Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064; ‡Howard Hughes Medical Institute, University of California, Santa Cruz, California 95064

Cold Spring Harbor Symposia on Quantitative Biology, Volume LXVIII. © 2003 Cold Spring Harbor Laboratory Press



Nature **420**, 520-562 (5 December 2002)

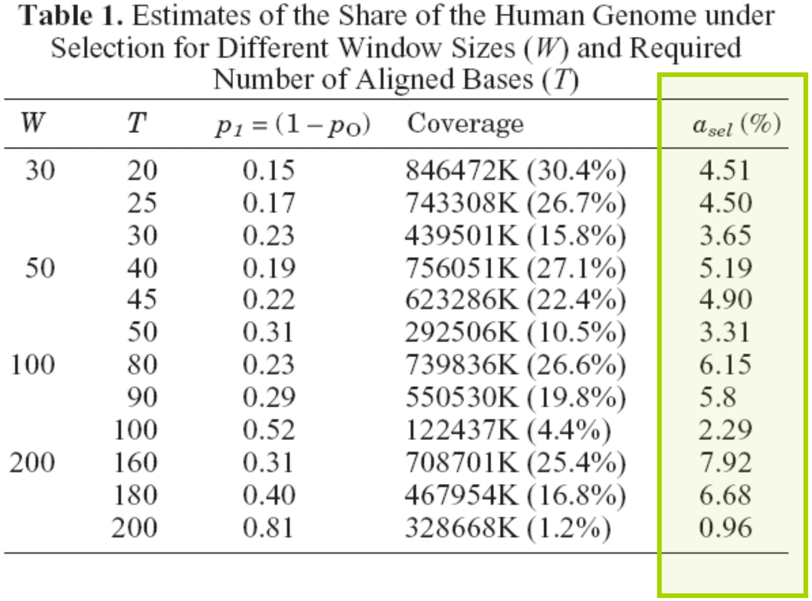
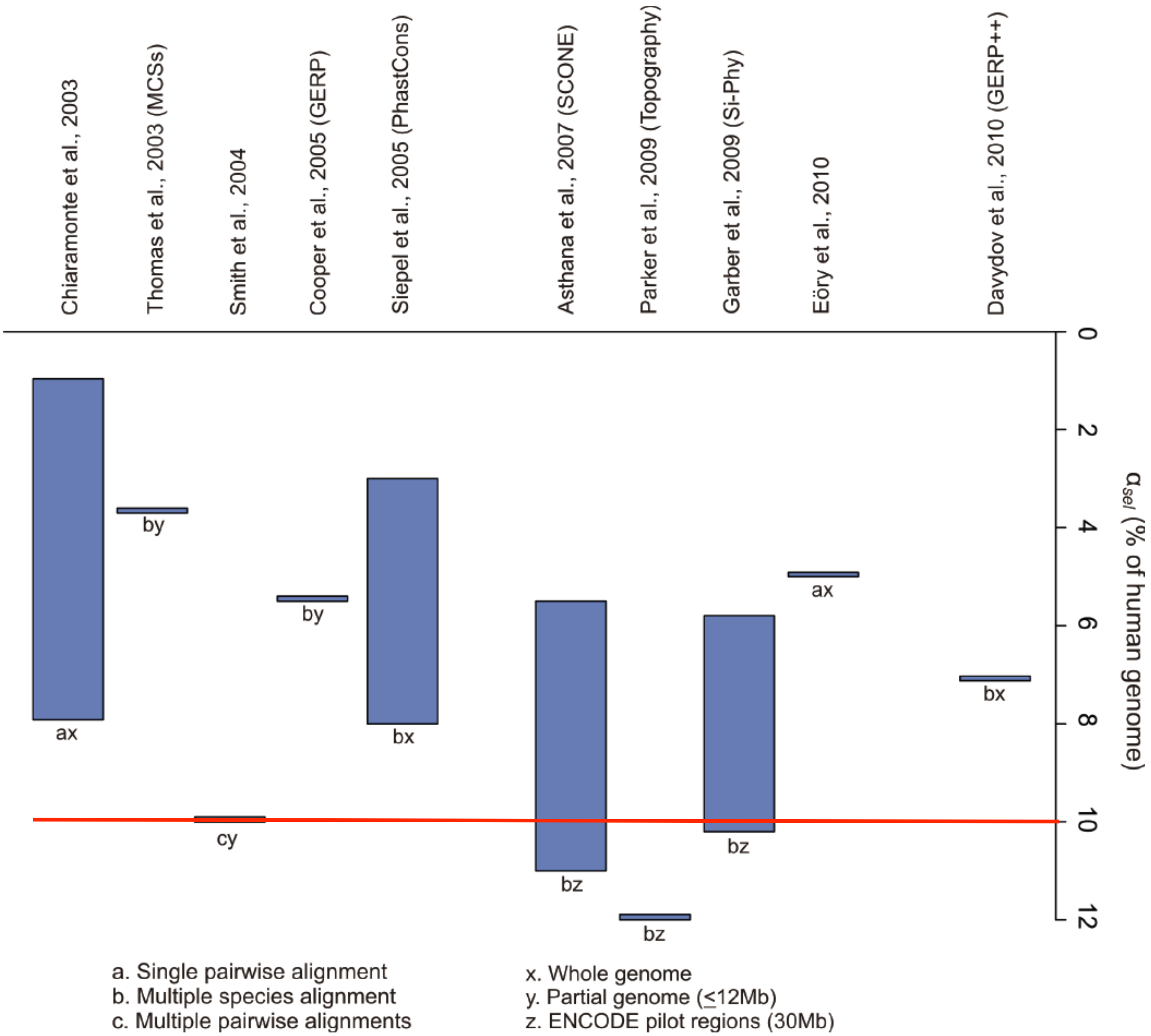
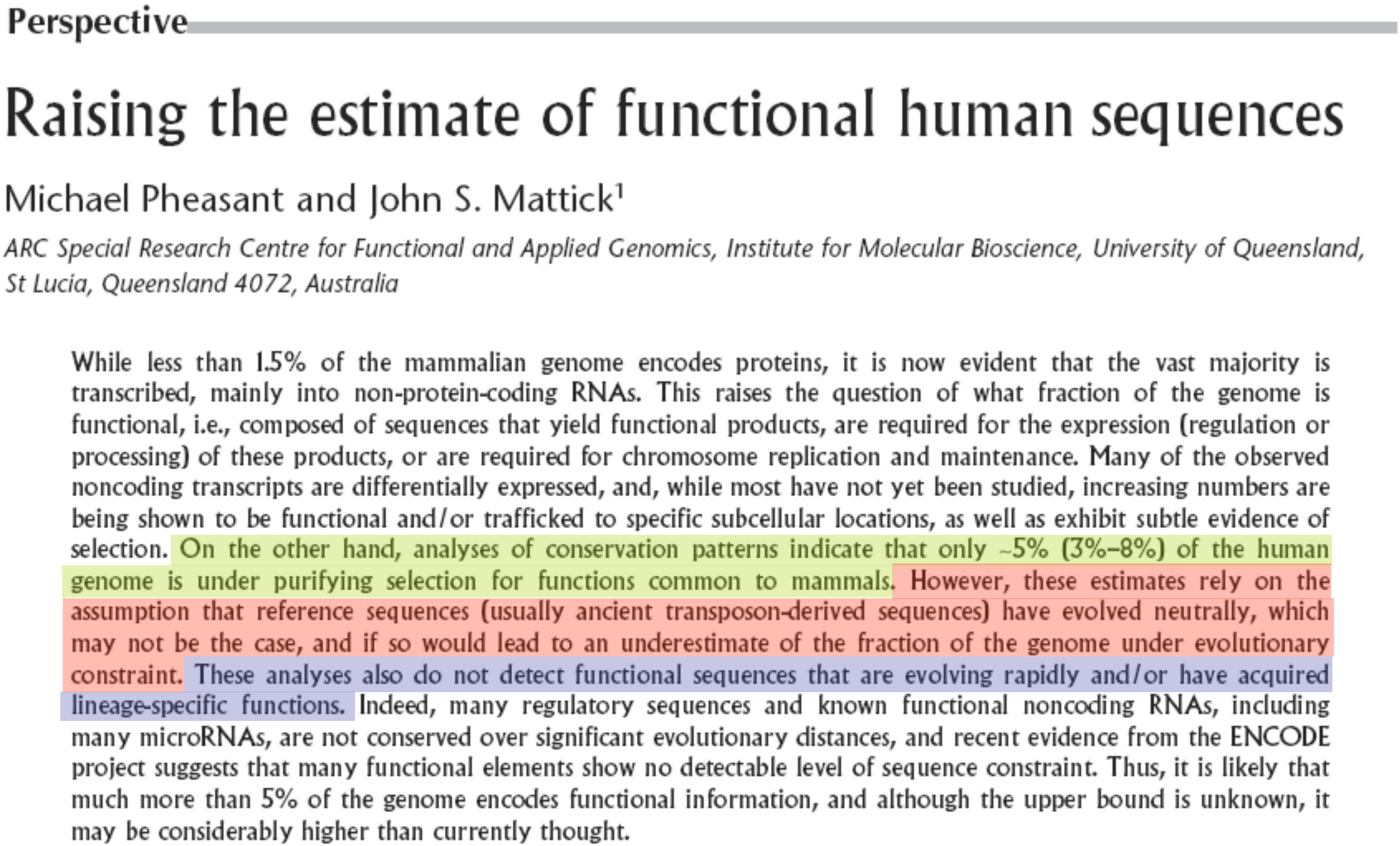


Table 1. Estimates of the Share of the Human Genome under Selection for Different Window Sizes (W) and Required Number of Aligned Bases (T)

W	T	$p_1 = (1 - p_O)$	Coverage	a_{sel} (%)
30	20	0.15	846472K (30.4%)	4.51
	25	0.17	743308K (26.7%)	4.50
	30	0.23	439501K (15.8%)	3.65
50	40	0.19	756051K (27.1%)	5.19
	45	0.22	623286K (22.4%)	4.90
	50	0.31	292506K (10.5%)	3.31
100	80	0.23	739836K (26.6%)	6.15
	90	0.29	550530K (19.8%)	5.8
	100	0.52	122437K (4.4%)	2.29
200	160	0.31	708701K (25.4%)	7.92
	180	0.40	467954K (16.8%)	6.68
	200	0.81	328668K (1.2%)	0.96

By comparing the extent of genome-wide sequence conservation to the neutral rate, the proportion of small (50–100 bp) segments in the mammalian genome that is under (purifying) selection can be estimated to be about 5%


Chiaramonte et al., 2003
Thomas et al., 2003 (MCSs)
Smith et al., 2004
Cooper et al., 2005 (GERP)
Siepel et al., 2005 (PhastCons)
Asthana et al., 2007 (SCONE)
Parker et al., 2009 (Topography)
Garber et al., 2009 (Si-Phy)
Eöry et al., 2010
Davydov et al., 2010 (GERP++)
ax
by
cy
by
bx
bz
bz
bz
ax
bx
0
2
4
6
8
10
12
α_{sel} (% of human genome)
a. Single pairwise alignment
b. Multiple species alignment
c. Multiple pairwise alignments
x. Whole genome
y. Partial genome (≤12Mb)
z. ENCODE pilot regions (30Mb)



Perspective

Raising the estimate of functional human sequences

Michael Pheasant and John S. Mattick[1]

ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, St Lucia, Queensland 4072, Australia

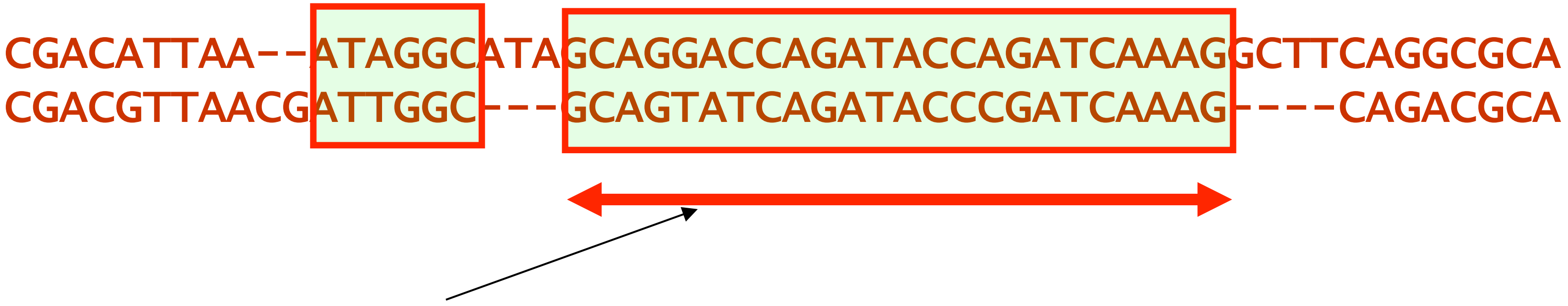
While less than 1.5% of the mammalian genome encodes proteins, it is now evident that the vast majority is transcribed, mainly into non-protein-coding RNAs. This raises the question of what fraction of the genome is functional, i.e., composed of sequences that yield functional products, are required for the expression (regulation or processing) of these products, or are required for chromosome replication and maintenance. Many of the observed noncoding transcripts are differentially expressed, and, while most have not yet been studied, increasing numbers are being shown to be functional and/or trafficked to specific subcellular locations, as well as exhibit subtle evidence of selection. On the other hand, analyses of conservation patterns indicate that only ~5% (3%–8%) of the human genome is under purifying selection for functions common to mammals. However, these estimates rely on the assumption that reference sequences (usually ancient transposon-derived sequences) have evolved neutrally, which may not be the case, and if so would lead to an underestimate of the fraction of the genome under evolutionary constraint. These analyses also do not detect functional sequences that are evolving rapidly and/or have acquired lineage-specific functions. Indeed, many regulatory sequences and known functional noncoding RNAs, including many microRNAs, are not conserved over significant evolutionary distances, and recent evidence from the ENCODE project suggests that many functional elements show no detectable level of sequence constraint. Thus, it is likely that much more than 5% of the genome encodes functional information, and although the upper bound is unknown, it may be considerably higher than currently thought.

Raising the estimate of functional human sequences

Michael Pheasant and John S. Mattick

Genome Res. 2007 17: 1245-1253; originally published online Aug 9, 2007;
Access the most recent version at doi:10.1101/gr.6406307

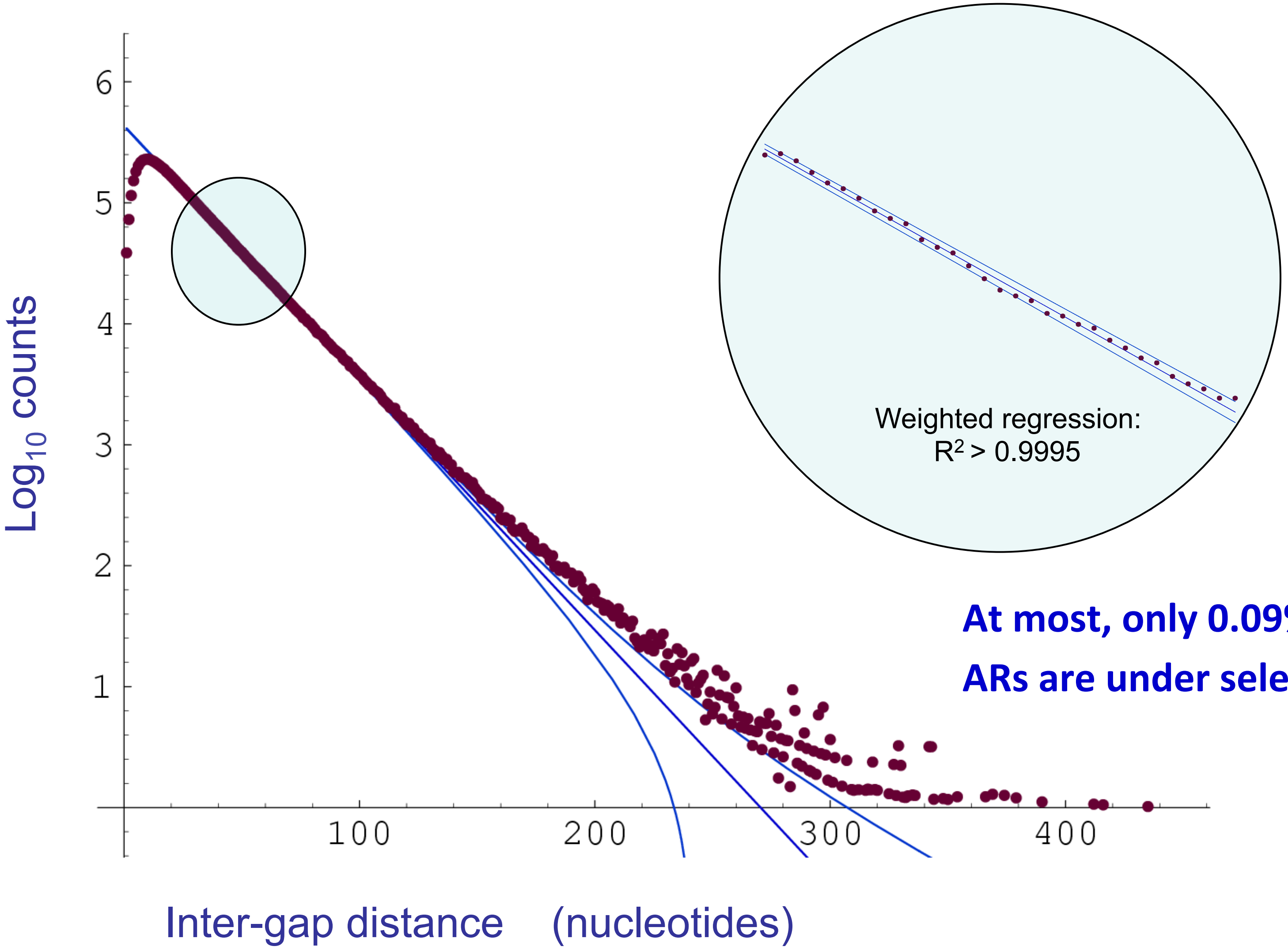
Gerton Lunter’s indel model: Insertions/ Deletions



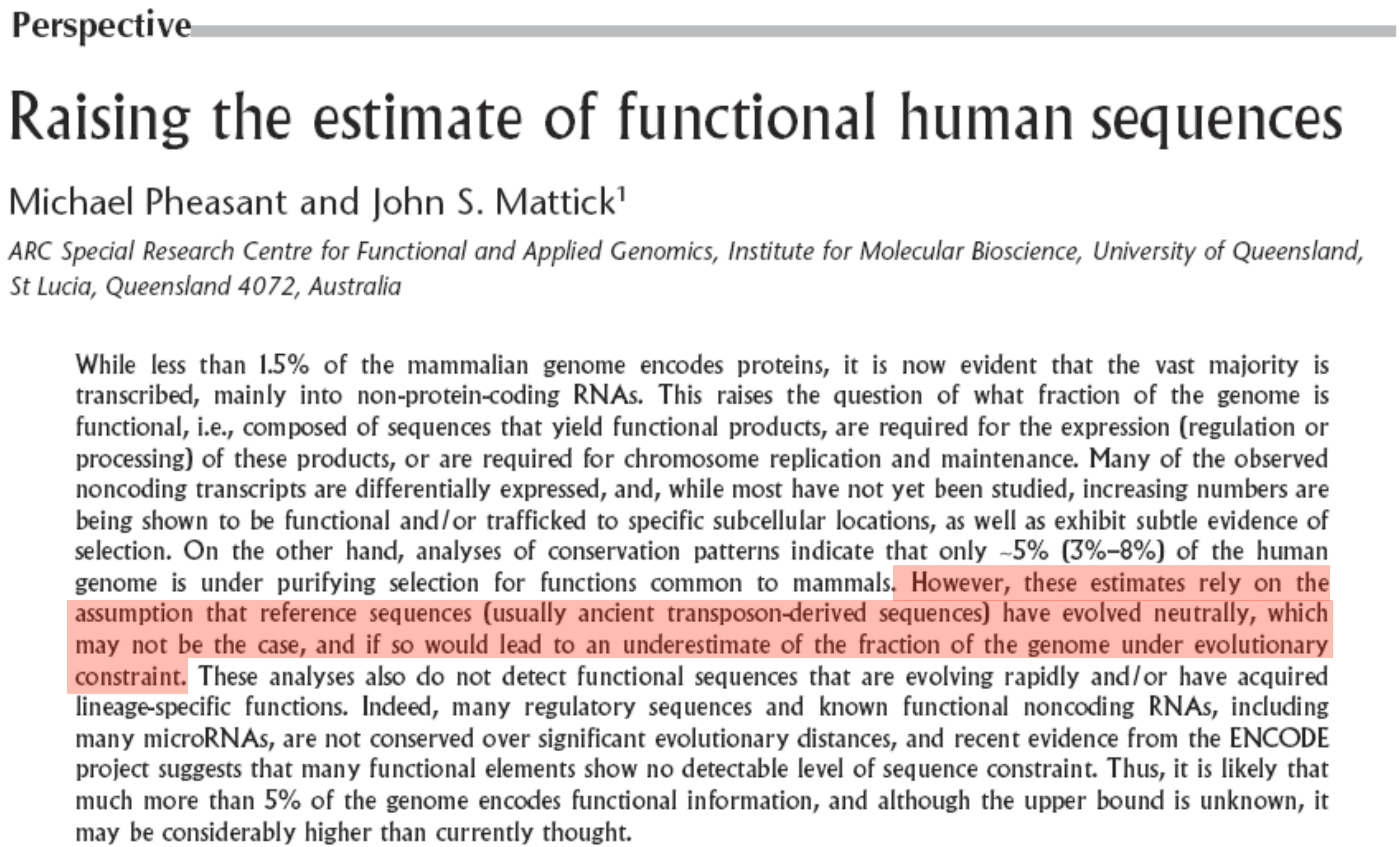
```
CGACATTAA--ATAGGCATAGCAGGACCAGATACCAGATCAAAGGCTTCAGGCGCA
CGACGTTAACGATTGGC---GCAGTATCAGATACCCGATCAAAG----CAGACGCA
```

- Consider lengths of ***inter-gap segments***
- **Do they follow a geometric distribution?**

Inter-gap distances within ancestral repeats



At most, only 0.09% of all ARs are under selection.



Perspective

Raising the estimate of functional human sequences

Michael Pheasant and John S. Mattick[1]

ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, St Lucia, Queensland 4072, Australia

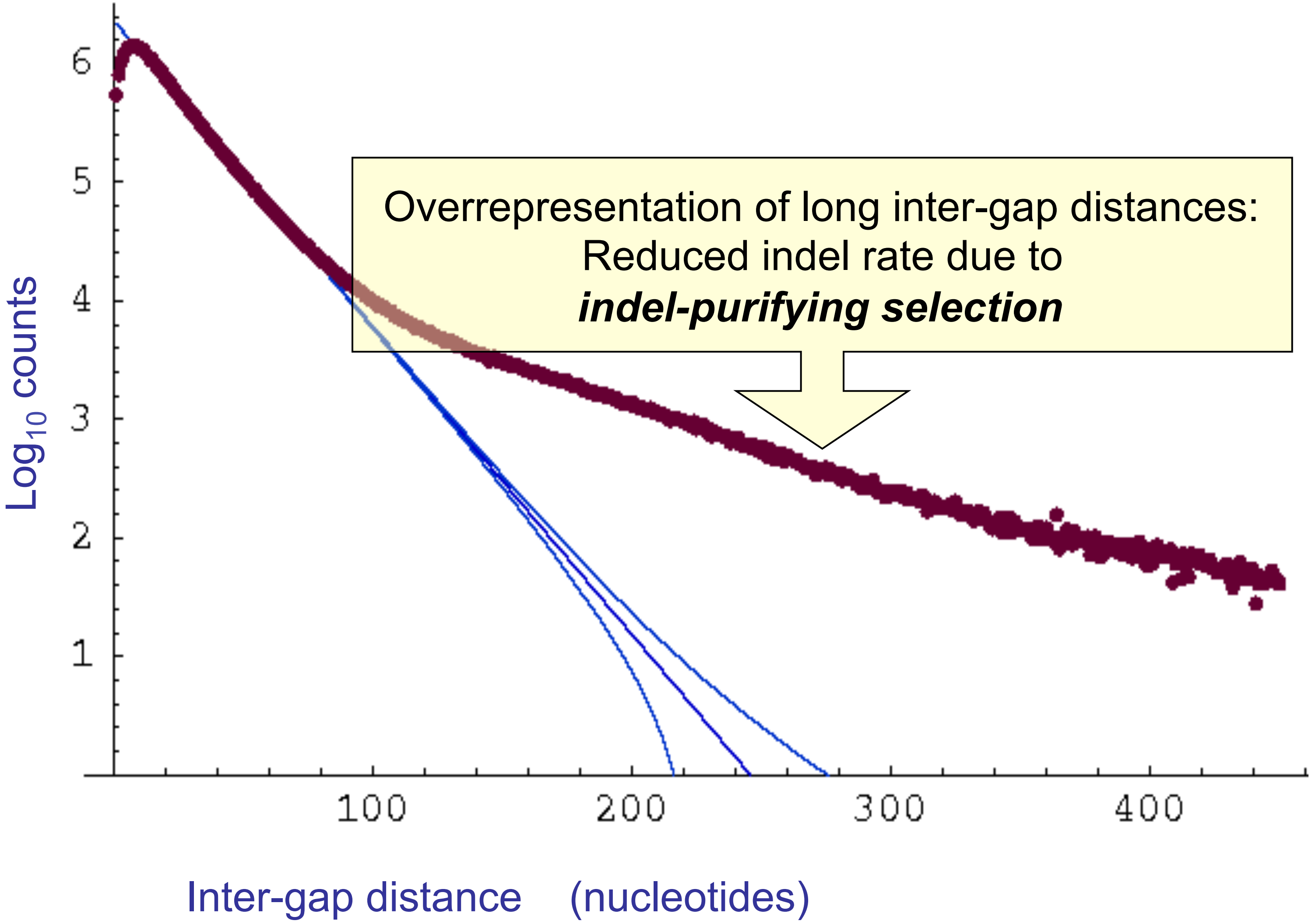
While less than 1.5% of the mammalian genome encodes proteins, it is now evident that the vast majority is transcribed, mainly into non-protein-coding RNAs. This raises the question of what fraction of the genome is functional, i.e., composed of sequences that yield functional products, are required for the expression (regulation or processing) of these products, or are required for chromosome replication and maintenance. Many of the observed noncoding transcripts are differentially expressed, and, while most have not yet been studied, increasing numbers are being shown to be functional and/or trafficked to specific subcellular locations, as well as exhibit subtle evidence of selection. On the other hand, analyses of conservation patterns indicate that only ~5% (3%–8%) of the human genome is under purifying selection for functions common to mammals. However, these estimates rely on the assumption that reference sequences (usually ancient transposon-derived sequences) have evolved neutrally, which may not be the case, and if so would lead to an underestimate of the fraction of the genome under evolutionary constraint. These analyses also do not detect functional sequences that are evolving rapidly and/or have acquired lineage-specific functions. Indeed, many regulatory sequences and known functional noncoding RNAs, including many microRNAs, are not conserved over significant evolutionary distances, and recent evidence from the ENCODE project suggests that many functional elements show no detectable level of sequence constraint. Thus, it is likely that much more than 5% of the genome encodes functional information, and although the upper bound is unknown, it may be considerably higher than currently thought.

Raising the estimate of functional human sequences

Michael Pheasant and John S. Mattick

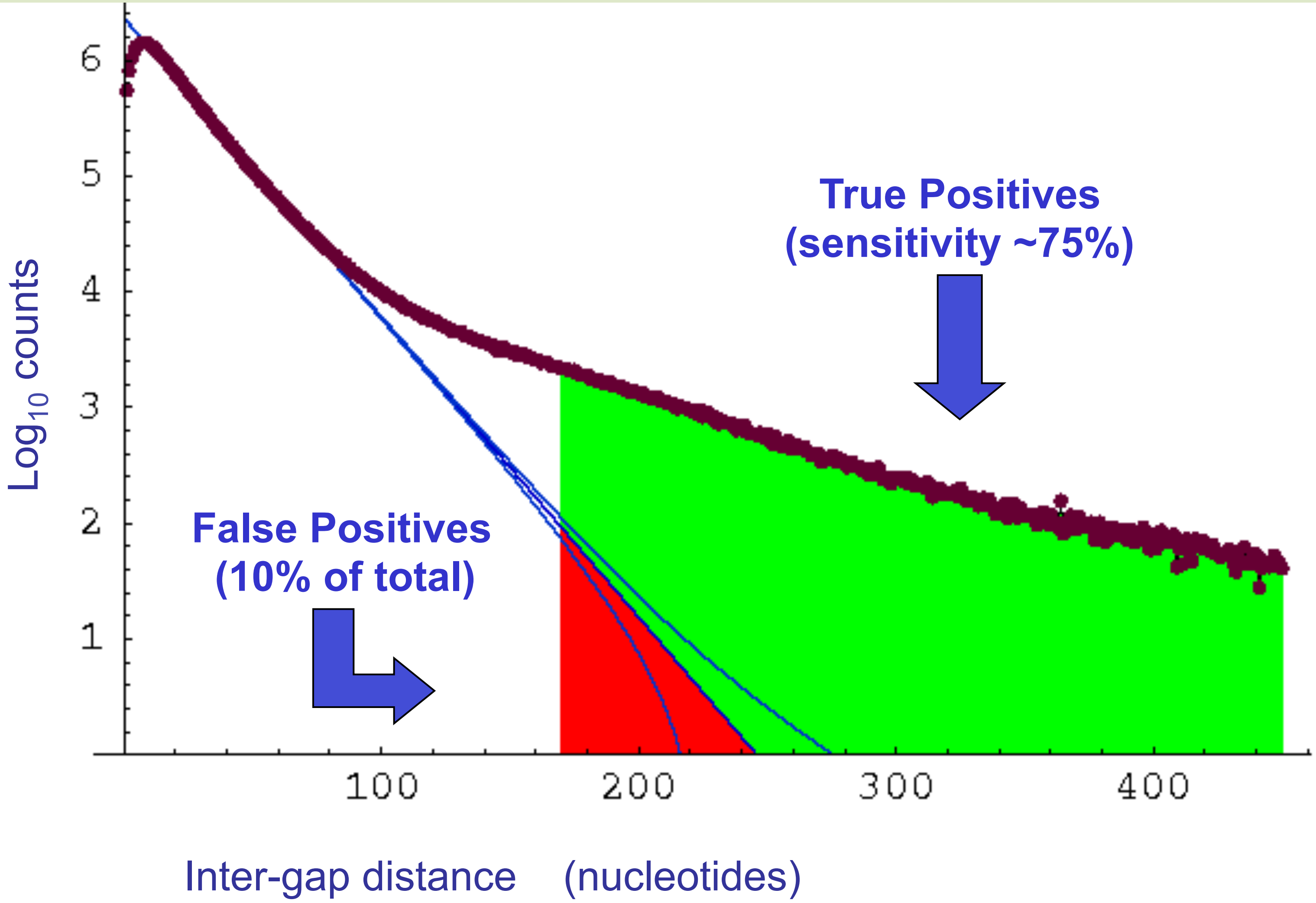
Genome Res. 2007 17: 1245-1253; originally published online Aug 9, 2007;
Access the most recent version at doi:10.1101/gr.6406307

Inter-gap distances: whole genome



Overrepresentation of long inter-gap distances:
Reduced indel rate due to
indel-purifying selection

Identifying sequence under indel purifying selection

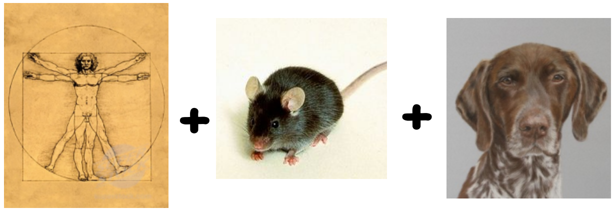


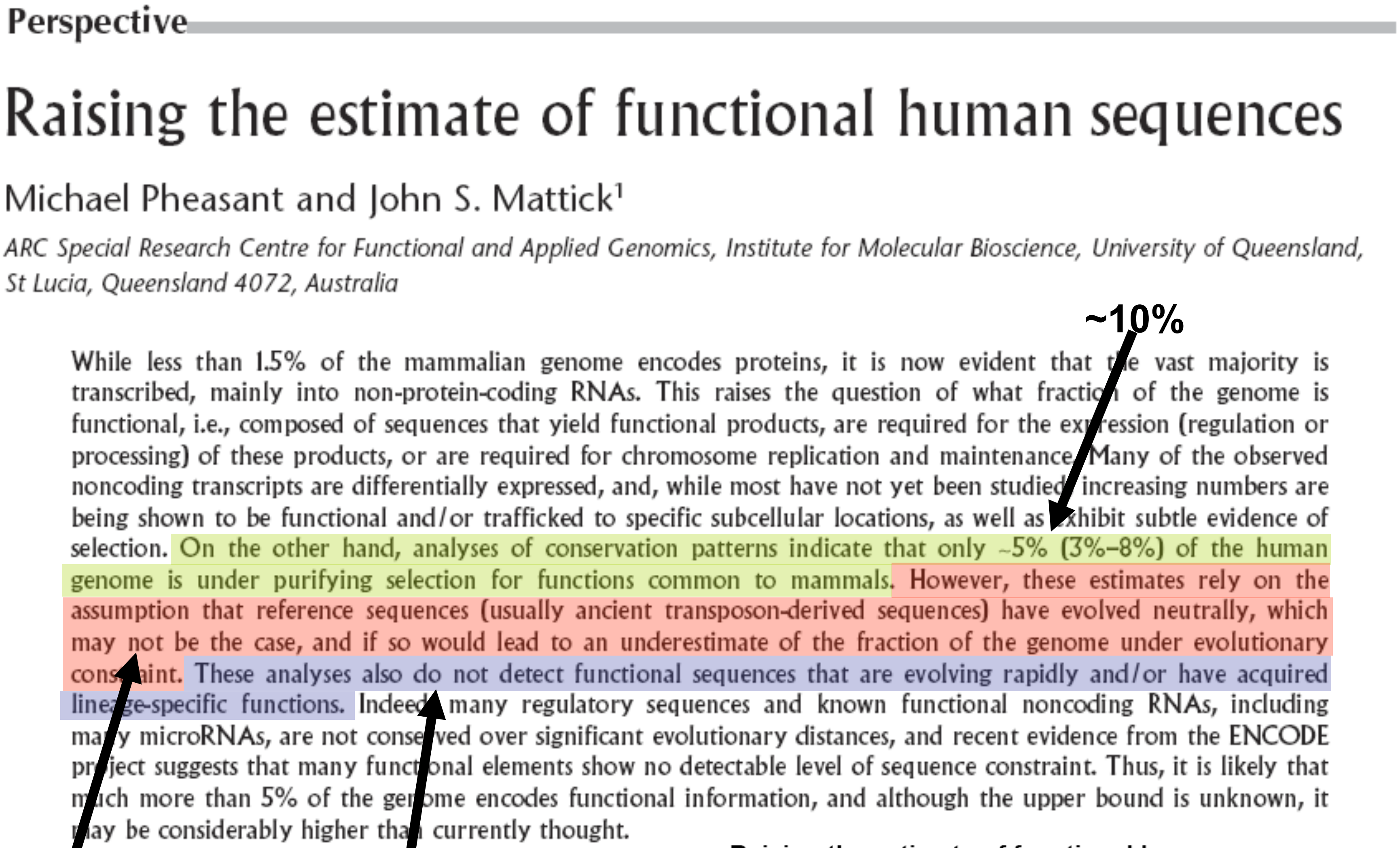
Fraction of conserved DNA

Lower bound: **~79 Mb**, or **~2.56 %** under indel purifying selection (human/mouse/dog)

Upper bound: **~100 Mb**, or **~3.25 %**

So: functional non-coding sequence represents over 1.56-2.25% of the human genome.





Perspective

Raising the estimate of functional human sequences

Michael Pheasant and John S. Mattick[1]

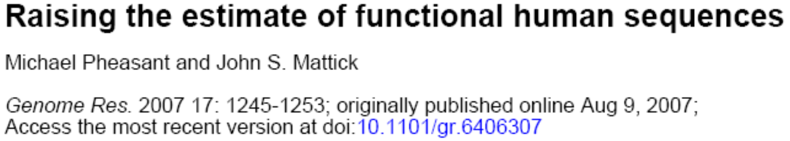
ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, St Lucia, Queensland 4072, Australia

While less than 1.5% of the mammalian genome encodes proteins, it is now evident that the vast majority is transcribed, mainly into non-protein-coding RNAs. This raises the question of what fraction of the genome is functional, i.e., composed of sequences that yield functional products, are required for the expression (regulation or processing) of these products, or are required for chromosome replication and maintenance. Many of the observed noncoding transcripts are differentially expressed, and, while most have not yet been studied, increasing numbers are being shown to be functional and/or trafficked to specific subcellular locations, as well as exhibit subtle evidence of selection. On the other hand, analyses of conservation patterns indicate that only ~5% (3%–8%) of the human genome is under purifying selection for functions common to mammals. However, these estimates rely on the assumption that reference sequences (usually ancient transposon-derived sequences) have evolved neutrally, which may not be the case, and if so would lead to an underestimate of the fraction of the genome under evolutionary constraint. These analyses also do not detect functional sequences that are evolving rapidly and/or have acquired lineage-specific functions. Indeed, many regulatory sequences and known functional noncoding RNAs, including many microRNAs, are not conserved over significant evolutionary distances, and recent evidence from the ENCODE project suggests that many functional elements show no detectable level of sequence constraint. Thus, it is likely that much more than 5% of the genome encodes functional information, and although the upper bound is unknown, it may be considerably higher than currently thought.

~10%

'TEs are predominantly neutral'

much 'turn-over' in functional sequence



Raising the estimate of functional human sequences

Michael Pheasant and John S. Mattick

Genome Res. 2007 17: 1245-1253; originally published online Aug 9, 2007;
Access the most recent version at doi:10.1101/gr.6406307

"the functional portion of the genome may ~~exceed 20%~~"

So: is the amount of functional material shared at different divergences?

For example,

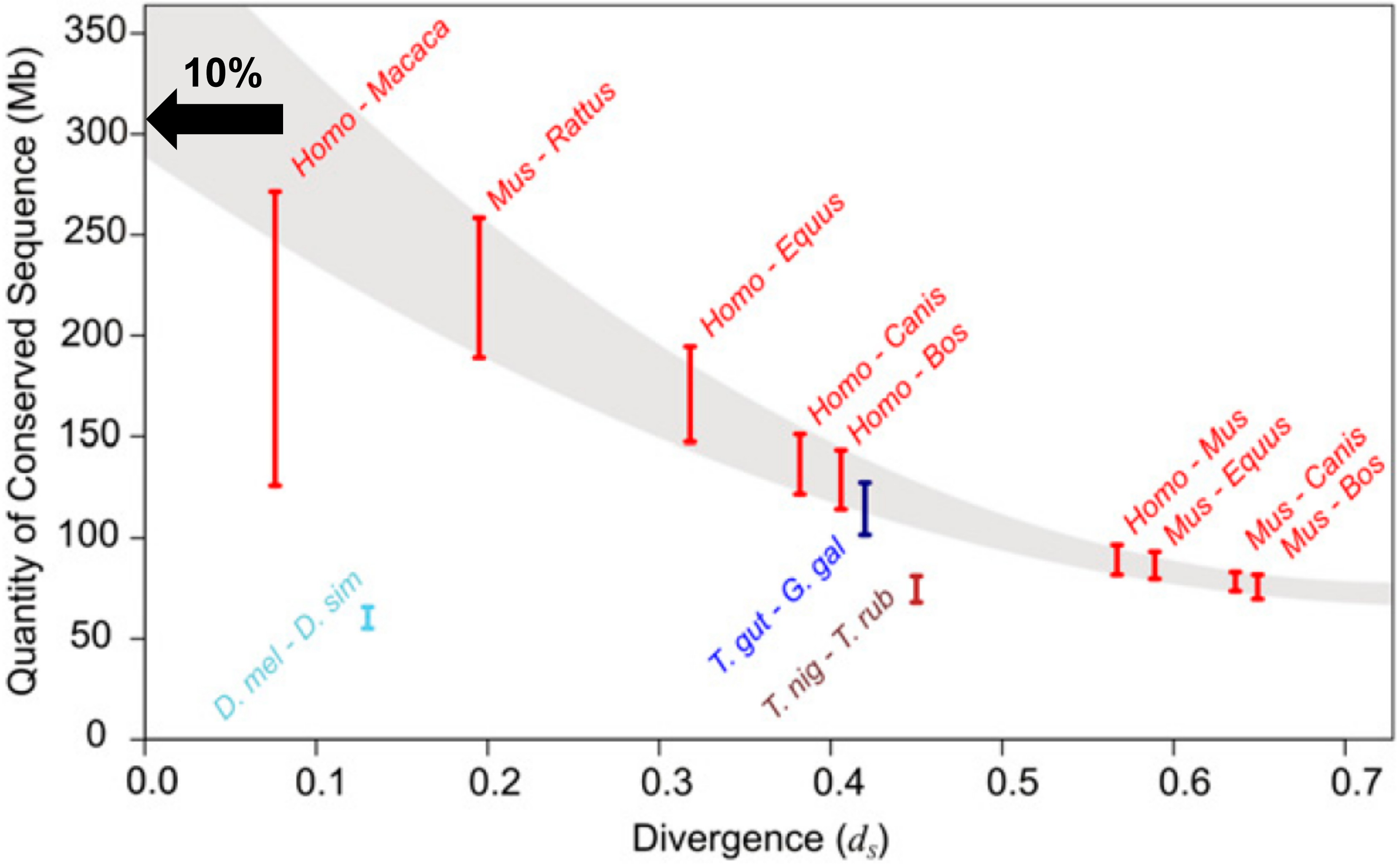
- human – mouse (75 My)
- human – macaque (25 My)
- mouse – rat (15 My)?

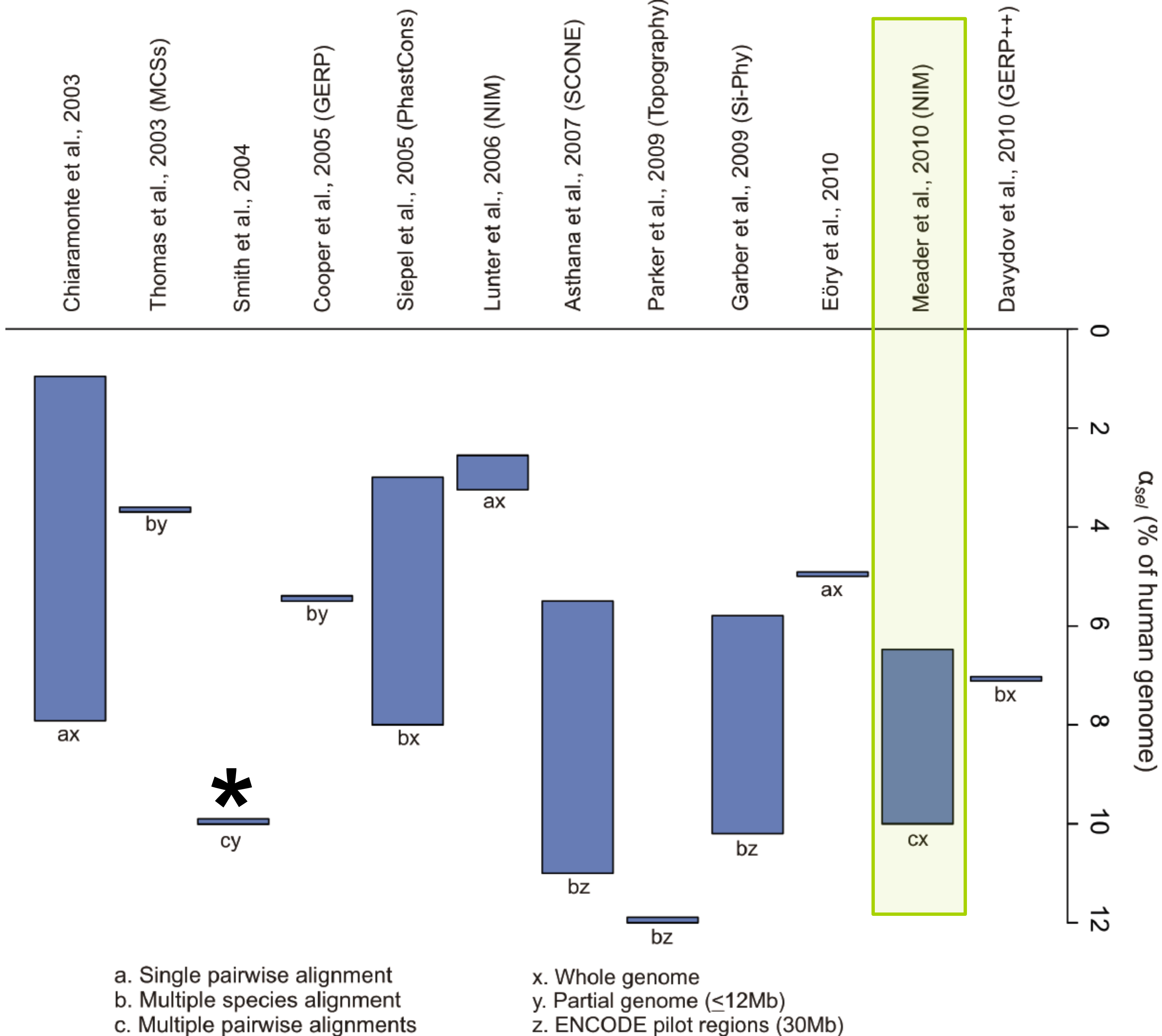
No

Massive turnover of functional sequence in human and other mammalian genomes

Stephen Meader, Chris P. Ponting and Gerton Lunter

Genome Res. 2010 20: 1335-1343 originally published online August 6, 2010




Chiaramonte et al., 2003
Thomas et al., 2003 (MCSs)
Smith et al., 2004
Cooper et al., 2005 (GERP)
Siepel et al., 2005 (PhastCons)
Lunter et al., 2006 (NIM)
Asthana et al., 2007 (SCONE)
Parker et al., 2009 (Topography)
Garber et al., 2009 (Si-Phy)
Eöry et al., 2010
Meader et al., 2010 (NIM)
Davydov et al., 2010 (GERP++)
0
2
4
6
8
10
12
α_{sel} (% of human genome)
ax
by
*
cy
by
bx
ax
bz
bz
bz
ax
cx
bx
a. Single pairwise alignment
b. Multiple species alignment
c. Multiple pairwise alignments
x. Whole genome
y. Partial genome (≤12Mb)
z. ENCODE pilot regions (30Mb)

ELSEVIER

GENOMICS

Evidence for turnover of functional noncoding DNA in mammalian genome evolution

Nick G.C. Smith*[,1], Mikael Brandström, Hans Ellegren

Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36 Uppsala, Sweden

Received 13 May 2004; accepted 20 July 2004
Available online 9 September 2004

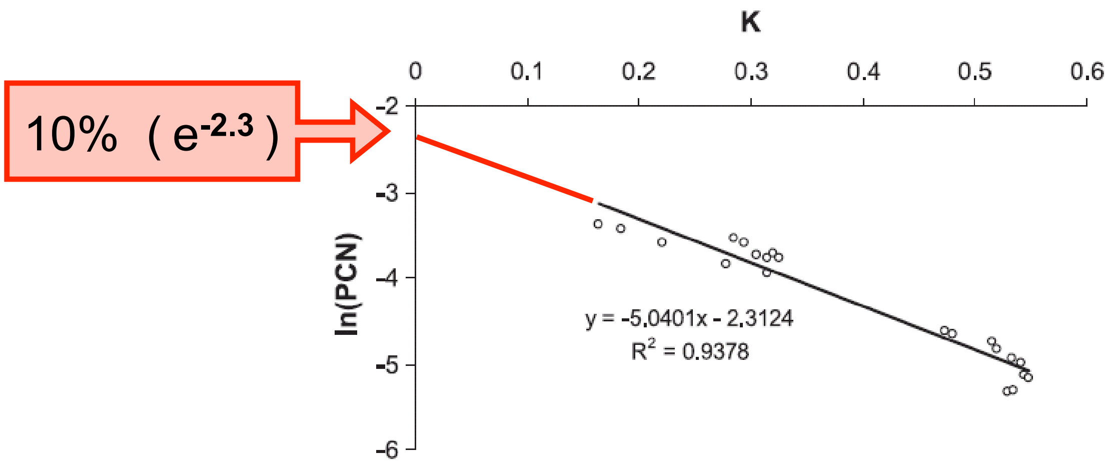


Fig. 1. The relationship between the proportion of the noncoding genome estimated to be conserved by negative selection, PCN, and the pairwise divergence, K, plotted for 21 different pairwise mammalian comparisons. The regression line for ln(PCN) versus K is also shown, along with its equation and the R^2 value.

Data:

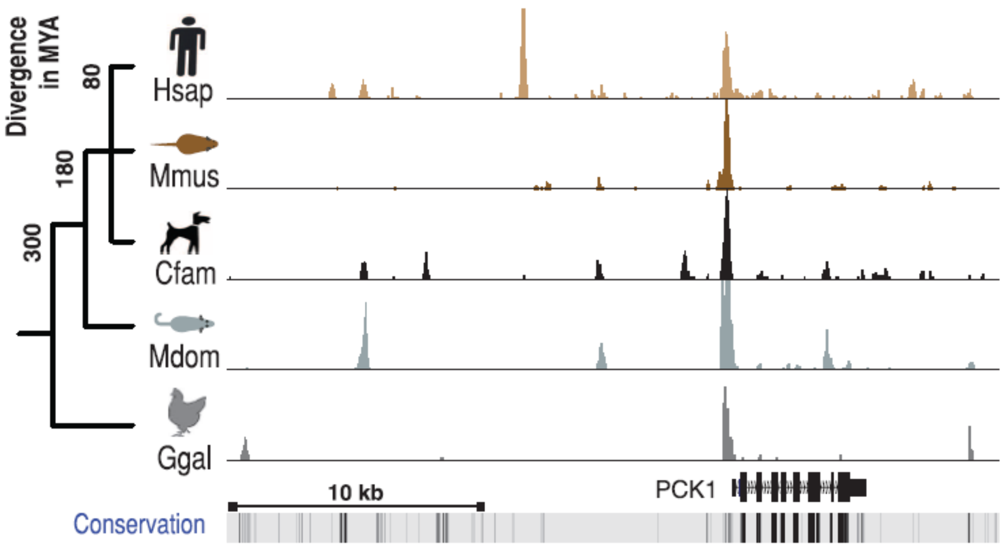
1.8 Mb non-exonic sequence
(CFTR + 9 other genes)

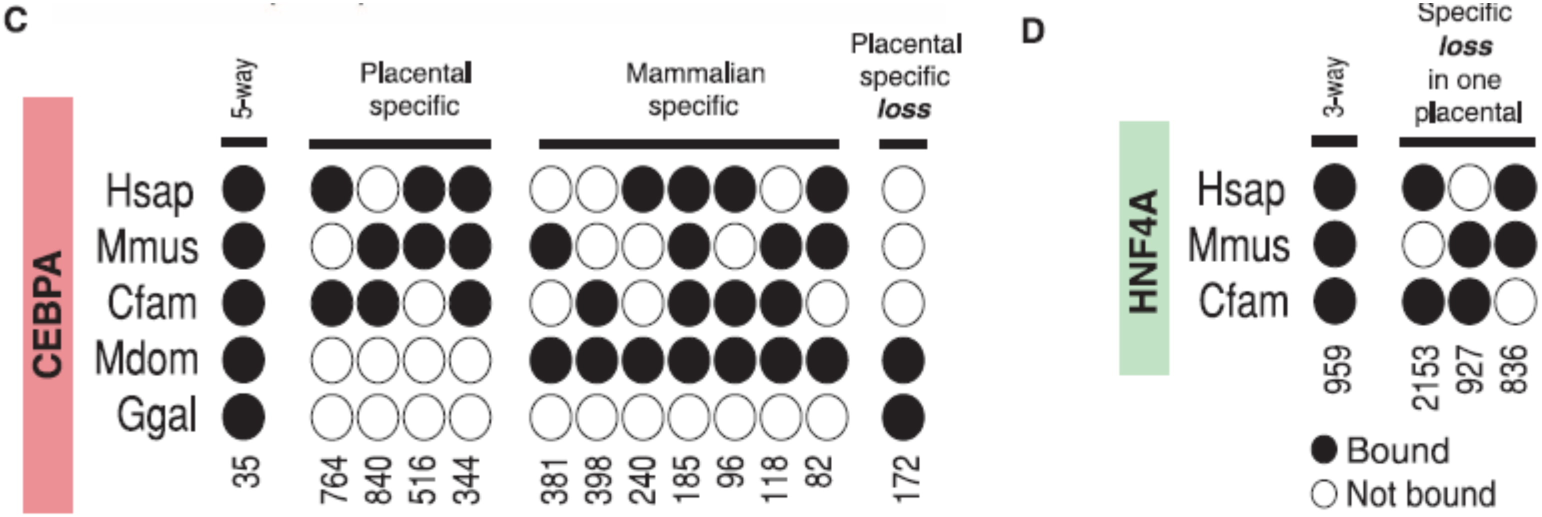
8 mammals
(human, baboon, cat, dog, pig, cow, rat, mouse)

Science AAAS

Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding
Dominic Schmidt, *et al.*
Science **328**, 1036 (2010);
DOI: 10.1126/science.1186176

CEBPA ChIP-seq of animal livers





CEBPA

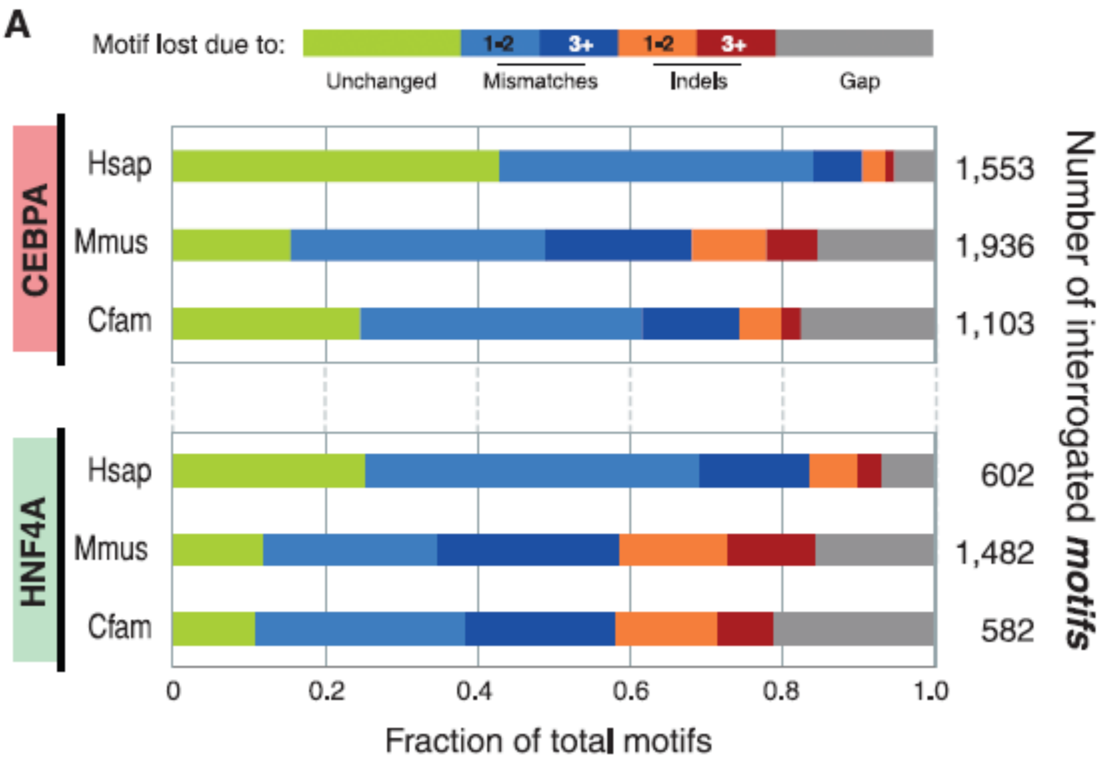
	5-way	Placental specific				Mammalian specific							Placental specific *loss*
Hsap	●	●	○	●	●	○	○	●	●	●	○	●	○
Mmus	●	○	●	●	●	●	○	○	●	○	●	●	○
Cfam	●	●	●	○	●	○	●	○	●	●	●	○	○
Mdom	●	○	○	○	○	●	●	●	●	●	●	●	●
Ggal	●	○	○	○	○	○	○	○	○	○	○	○	●
	35	764	840	516	344	381	398	240	185	96	118	82	172

HNF4A

	3-way	Specific *loss* in one placental		
Hsap	●	●	○	●
Mmus	●	○	●	●
Cfam	●	●	●	○
	959	2153	927	836

● Bound
○ Not bound

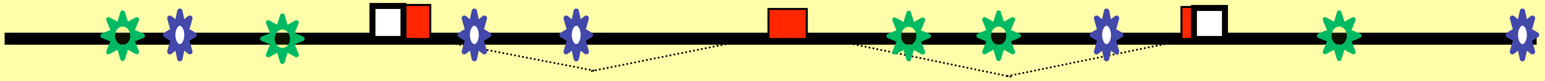
Fig. 4. Lineage-specific loss and turnover of TF binding events. (**A**) The unbound regions in each placental mammal that align to regions showing TF binding in the other two placental mammals were collected, and the mechanisms by which the underlying motifs were disrupted were summarized. (**B**) Turnovers occurred near lineage-specific lost binding events approximately half the time; shared turnovers represent cases where a cluster of binding events likely occurred in a common ancestor (fig. S16).



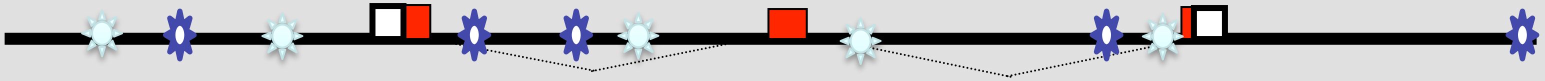
Approximately half of lineage-specific losses of TF binding showed evidence of nearby compensatory binding events (Fig. 4B). A quarter of species-specific losses had a nearby (±10 kb) gained binding event that is unique to the same lineage (unshared turnover), and an additional quarter of the losses had a nearby binding event that is shared in one or more other species (shared turnover) (fig. S16).

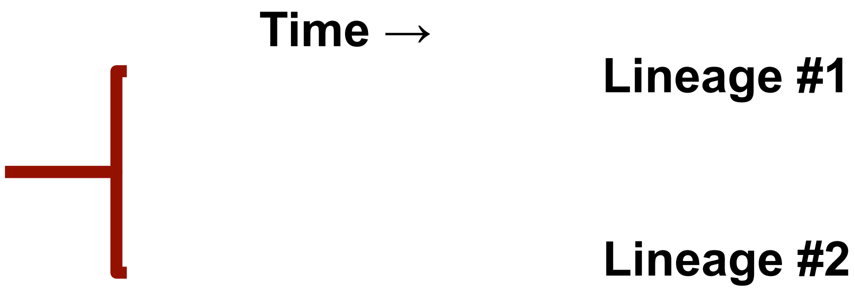
Ephemerality of lowly constrained functional elements

Lineage #1

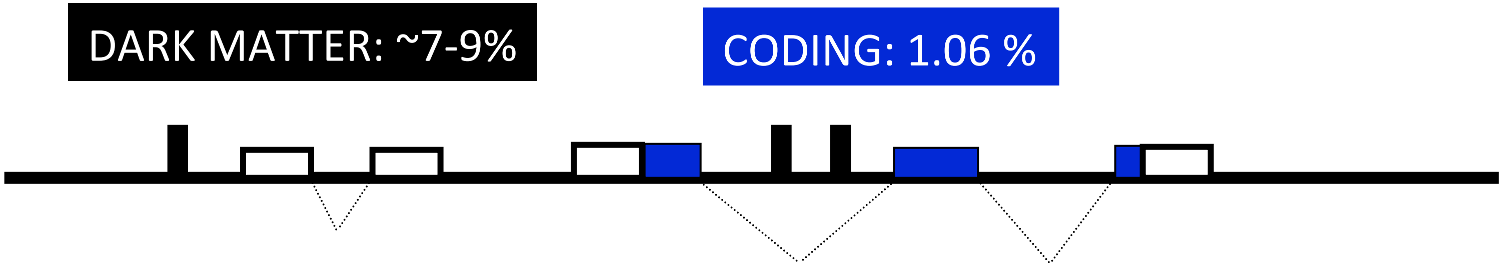


Lineage #2





Our View of the Human/Mouse Genome



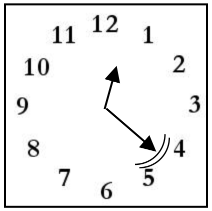
DARK MATTER: ~7-9%

CODING: 1.06 %

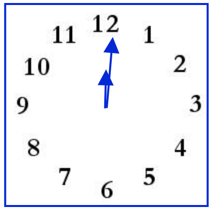
King & Wilson: *Human and Chimpanzee "macromolecules are so alike that regulatory mutations may account for their biological differences"*

Science (1975) 188, 107-116

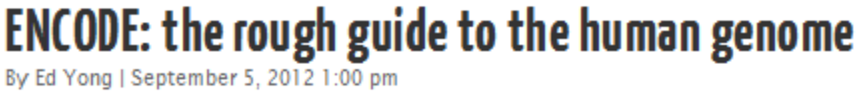
dark matter



proteins

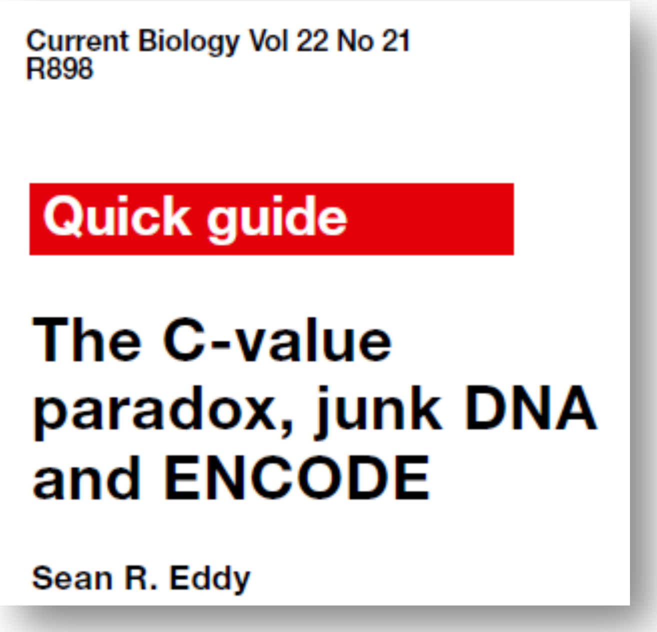


ENCODE’s 80%


ENCODE: the rough guide to the human genome

By Ed Yong | September 5, 2012 1:00 pm

According to ENCODE’s analysis, 80 percent of the genome has a “biochemical function”. “Almost every nucleotide is associated with a function of some sort or another, and we now know where they are, what binds to them, what their associations are, and more,” says Tom Gingeras, one of the study’s many senior scientists.


Current Biology Vol 22 No 21
R898

Quick guide

The C-value paradox, junk DNA and ENCODE

Sean R. Eddy

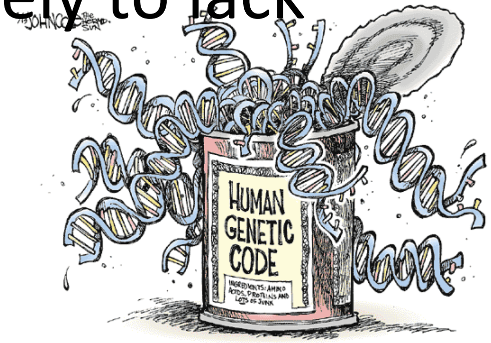
The Functional Portion

Approximately 10% of the human genome appears to be constrained with respect to insertions & deletions.

This compares with 1.2% that encodes protein sequences.

The amount of constrained but non-coding sequence is thus considerably larger (8-fold) than constrained coding sequence.

90% of the human genome, therefore, is likely to lack constraint and truly is *junk*.



The Future: Functional, unconserved, sequence

Functional, unconserved, human sequence is *TWICE* the amount of functional sequence that is conserved to mouse.

Deducing the functions of such lineage-specific sequence will require:

comparisons to many primate genomes; and,

diverse experimental approaches.

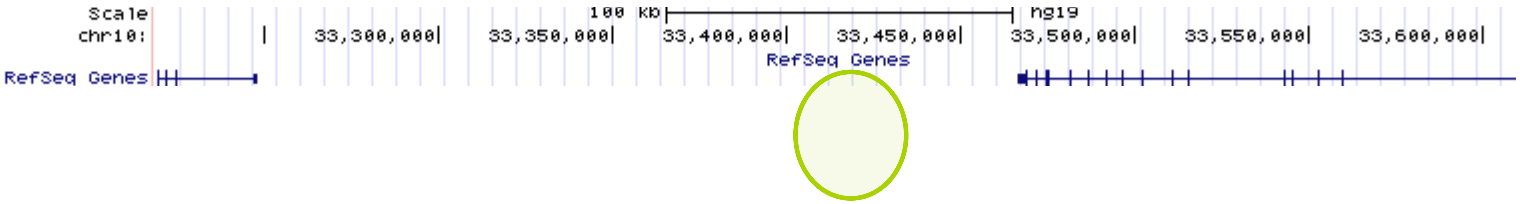
ENCODE

Encyclopedia of DNA Elements

nature.com/encode

ENCODE 2012

2001

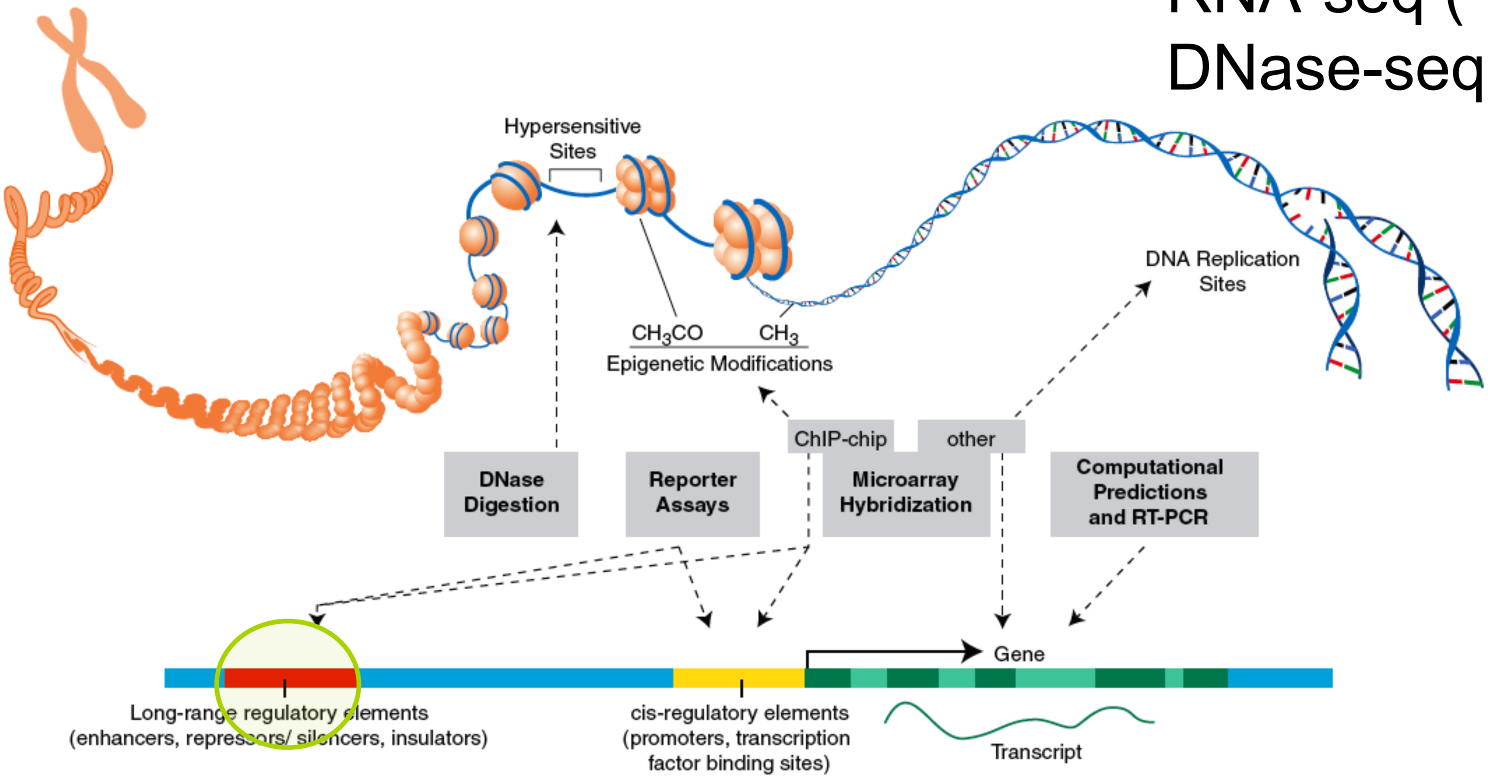

Scale
100 kb
hg19
chr10:
33,300,000
33,350,000
33,400,000
33,450,000
33,500,000
33,550,000
33,600,000
RefSeq Genes
RefSeq Genes

2012

Experimental assays

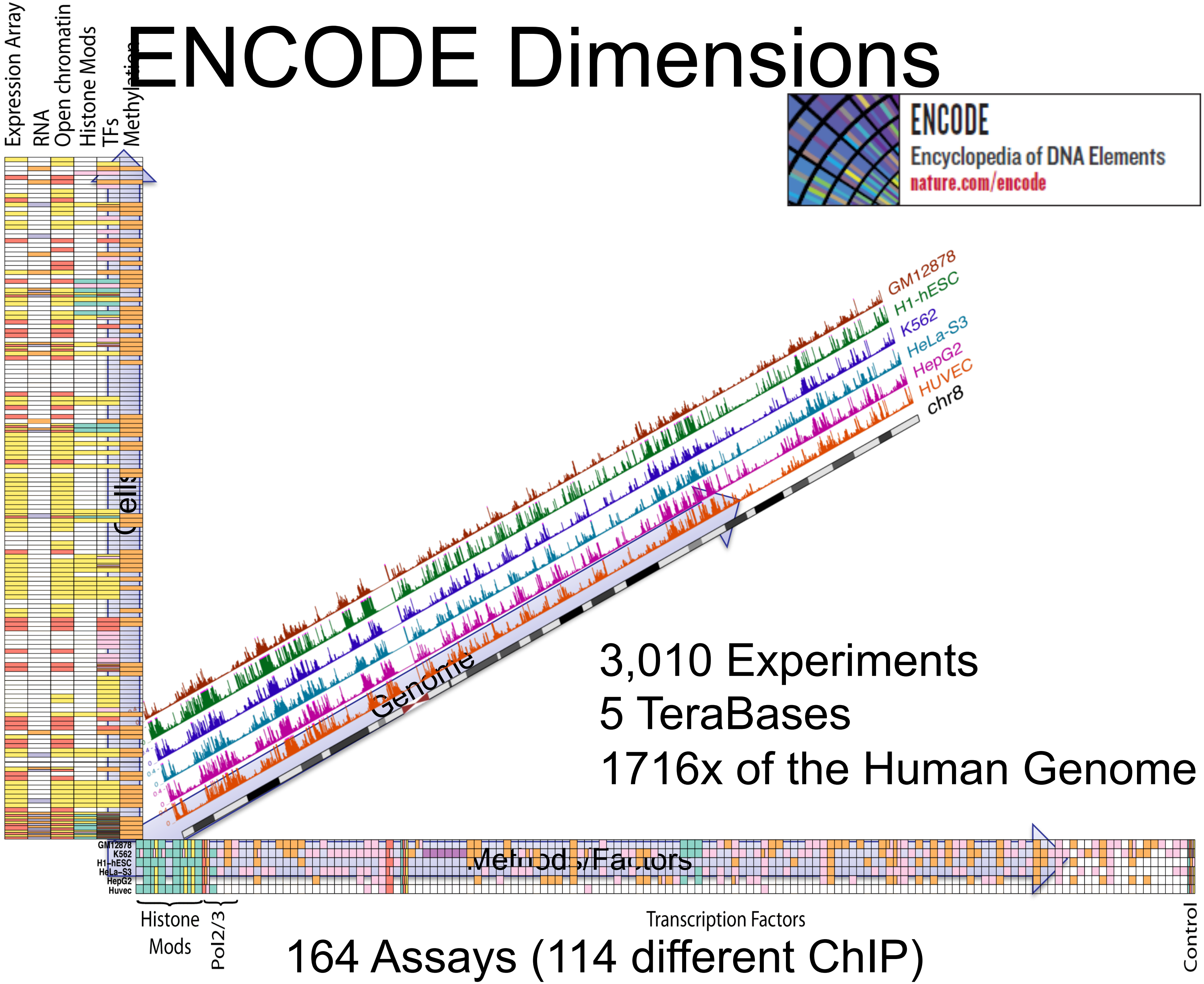
ENCODE
Encyclopedia of DNA Elements
nature.com/encode

ChIP-seq (~150)
RNA-seq (~100)
DNase-seq (~100)


Hypersensitive Sites
CH_3CO
CH_3
Epigenetic Modifications
DNA Replication Sites
ChIP-chip
other
DNase Digestion
Reporter Assays
Microarray Hybridization
Computational Predictions and RT-PCR
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

ENCODE Dimensions

ENCODE
Encyclopedia of DNA Elements
nature.com/encode



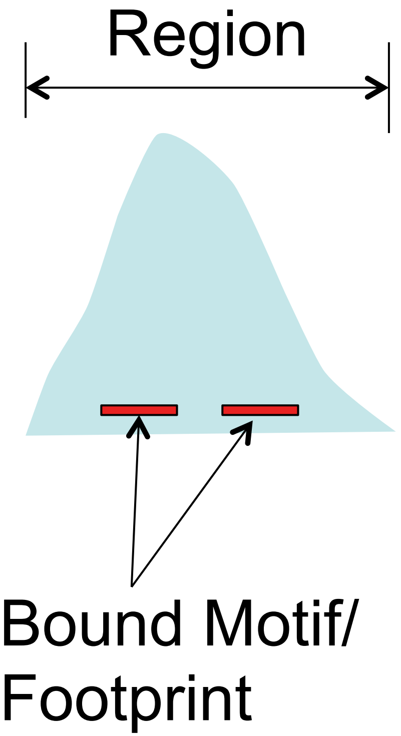
182 Cell Lines/ Tissues

3,010 Experiments
5 TeraBases
1716x of the Human Genome

164 Assays (114 different ChIP)

"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."

Element Type	Coverage	Cumulative Coverage
Exons	3%	3%
ChIP-seq bound motifs	4.5%	5%
DNaseI Footprints	5.7%	9%
ChIP-seq bound regions	8.1%	12%
DNaseI HS regions	15.2%	19.4%
Histone Modifications (*)	44%	49%
RNA	62%	80%
(* excluding broad marks)		



Region

Bound Motif/ Footprint

(Union over all experiments and cell types)

ENCODE
Encyclopedia of DNA Elements
nature.com/encode

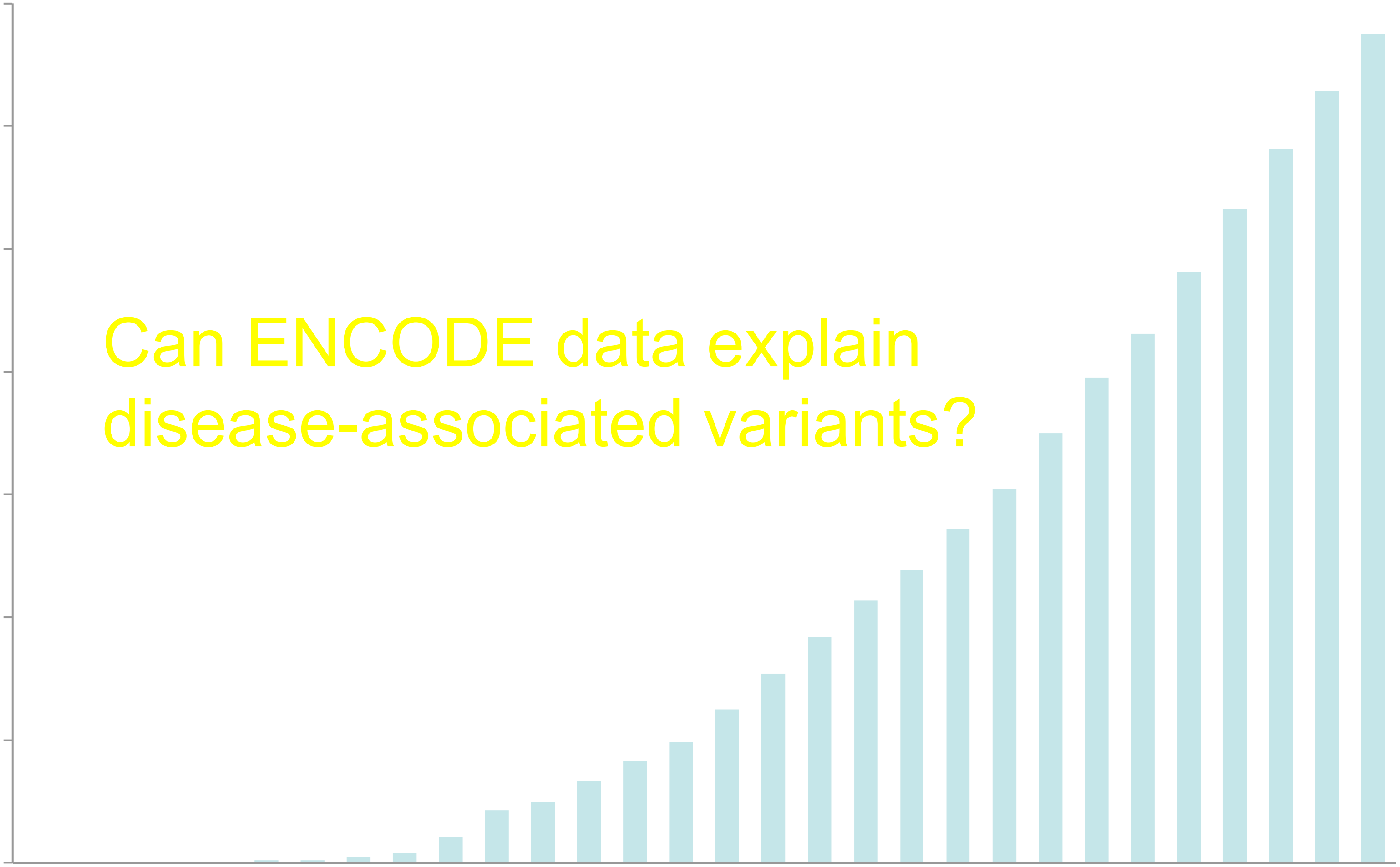
Elements are evenly spaced over the genome

99% of the genome is within 1.7 kb of a biochemical event

95% of the genome is within 8 kb of a bound motif or footprint

ENCODE
Encyclopedia of DNA Elements
nature.com/encode

Published GWA Reports, 2005 – 6/2012

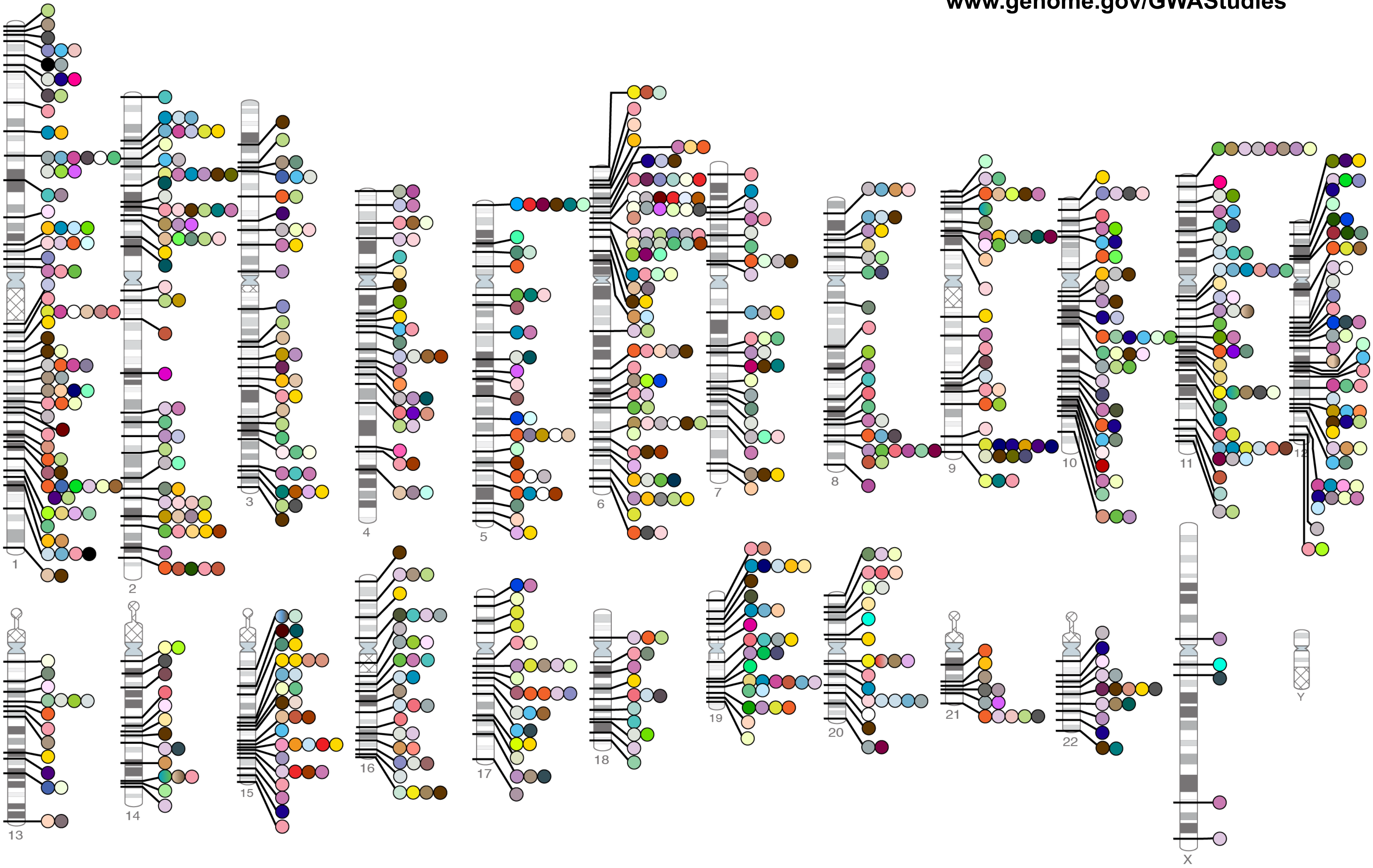


Can ENCODE data explain disease-associated variants?

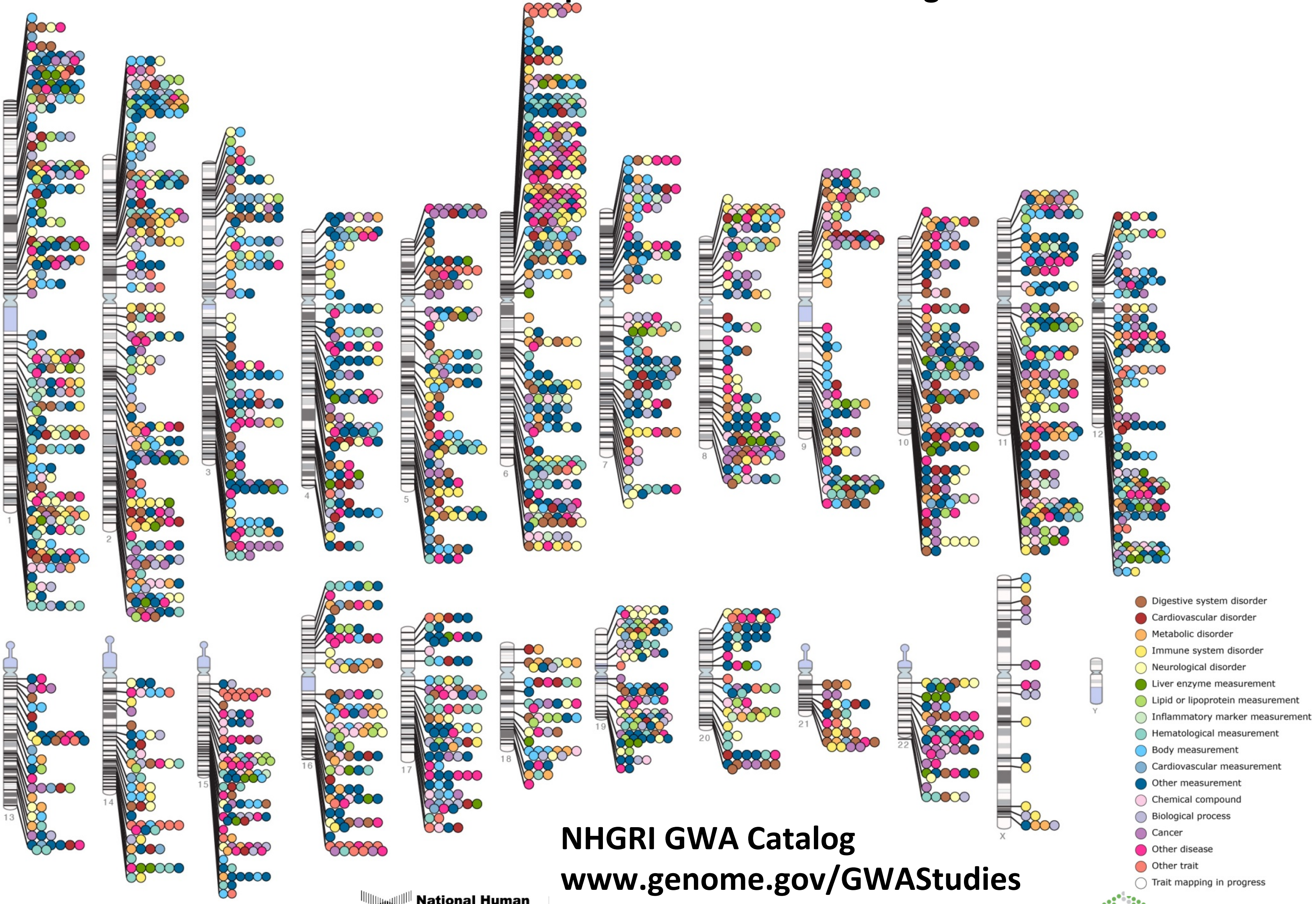
Through 6/30/12 postings

Published Genome-Wide Associations through 6/2010

NHGRI GWA Catalog
www.genome.gov/GWAStudies



Published Genome-Wide Associations through 07/2012
Published GWA at $p \leq 5X10^{-8}$ for 18 trait categories

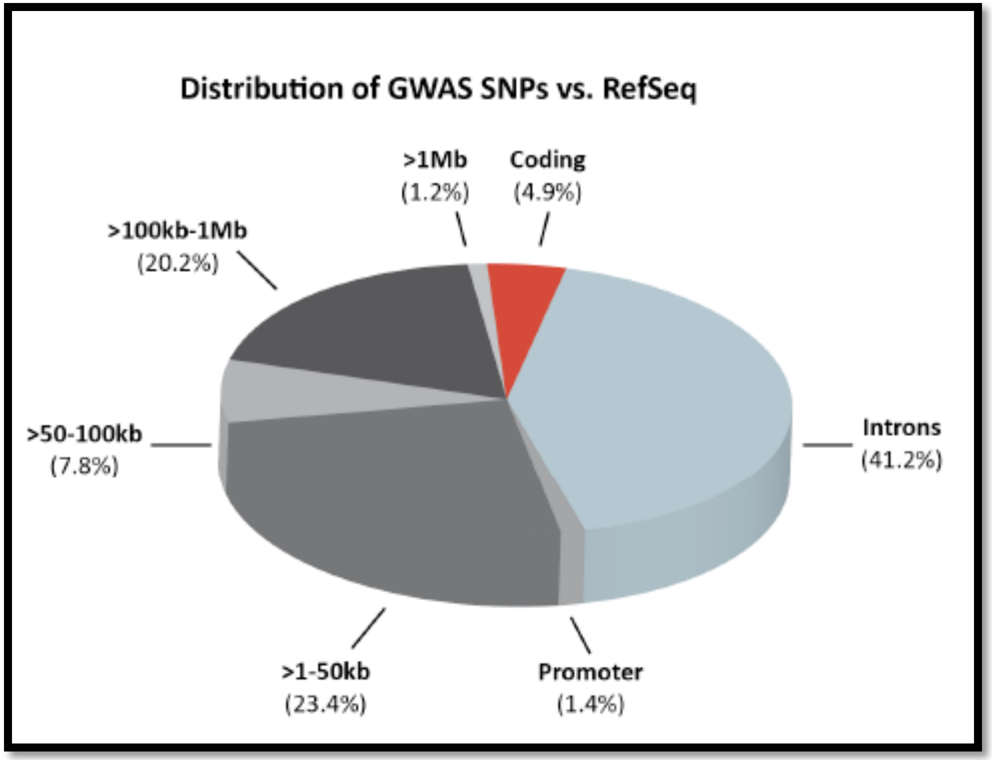


NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human
Genome Research
Institute

EMBL-EBI

95% of GWAS risk-associated SNPs lie *outside of* coding sequence



Science AAAS

Systematic Localization of Common Disease-Associated Variation in Regulatory DNA
Matthew T. Maurano *et al.*
Science **337**, 1190 (2012);
DOI: 10.1126/science.1222794

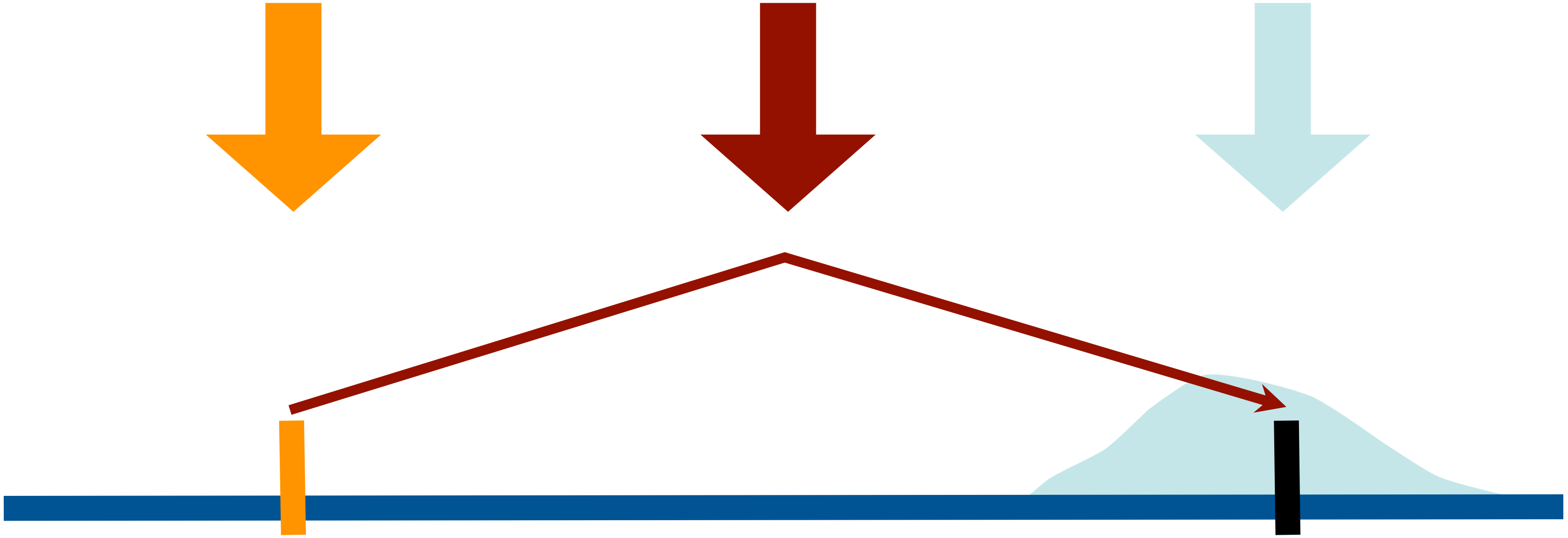
Functional SNPs (fSNPs)

Belinda Giardine, Marc Shaub, Ross Hardison, Mike Snyder, John Stam.

Genome Wide Association Studies (GWAS) Results

Linkage Disequilibrium

ENCODE Functional Region



Reported SNP

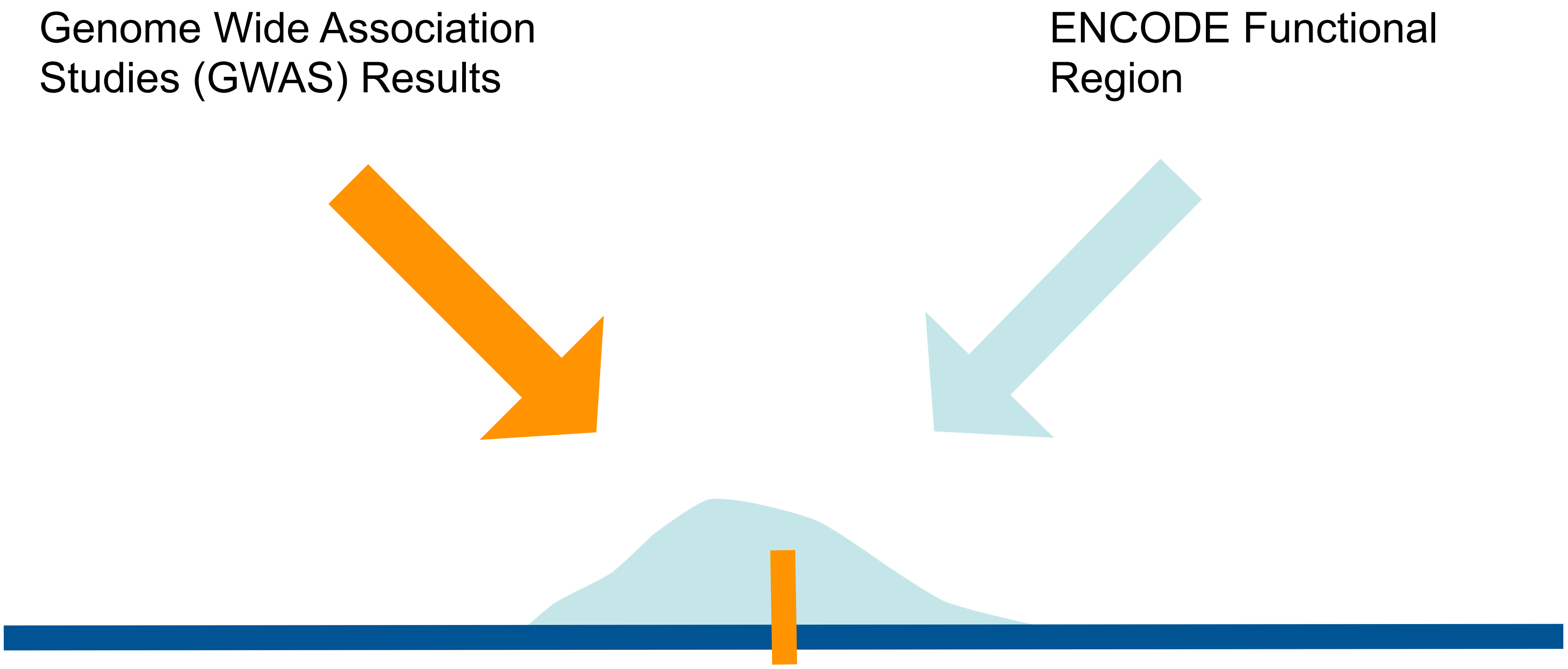
fSNP

Statistically associated with the phenotype

✔ Associated with the phenotype
✔ In a functional region

ENCODE
Encyclopedia of DNA Elements
nature.com/encode

Functional SNP - Direct Hit



Genome Wide Association Studies (GWAS) Results

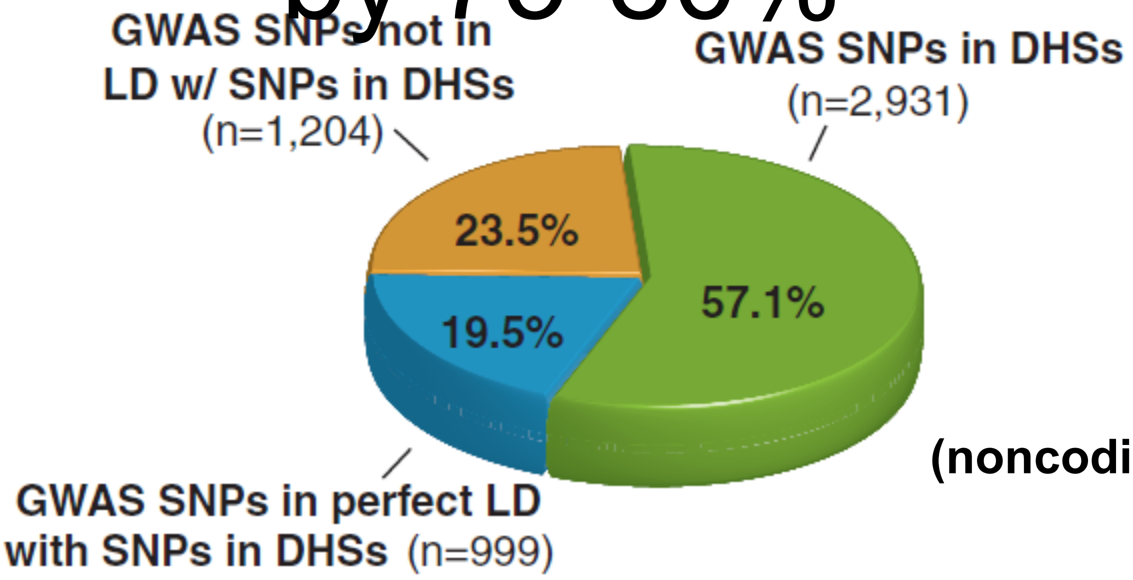
ENCODE Functional Region

<u>fSNP Direct Hit</u>

✔ Association reported in a GWAS
✔ In a functional region

ENCODE
Encyclopedia of DNA Elements
nature.com/encode

When you extend to SNPs in high LD, GWAS SNPs overlap by 75-80%



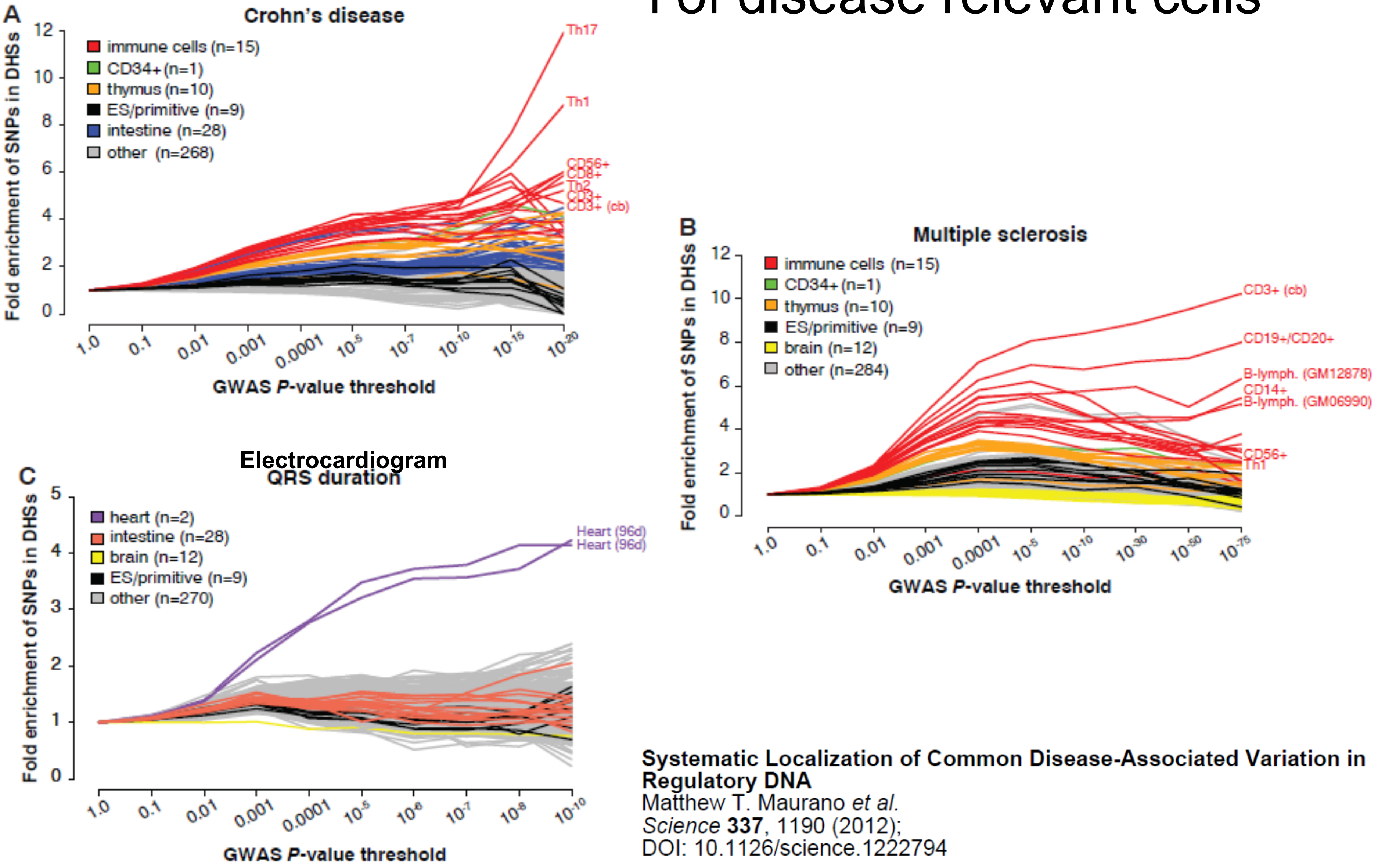
(noncoding SNPs only)

76.5% of GWAS SNPs are either within or in perfect LD with DHSs.

88.1% GWAS SNPs lie within DHSs active in fetal cells and tissues

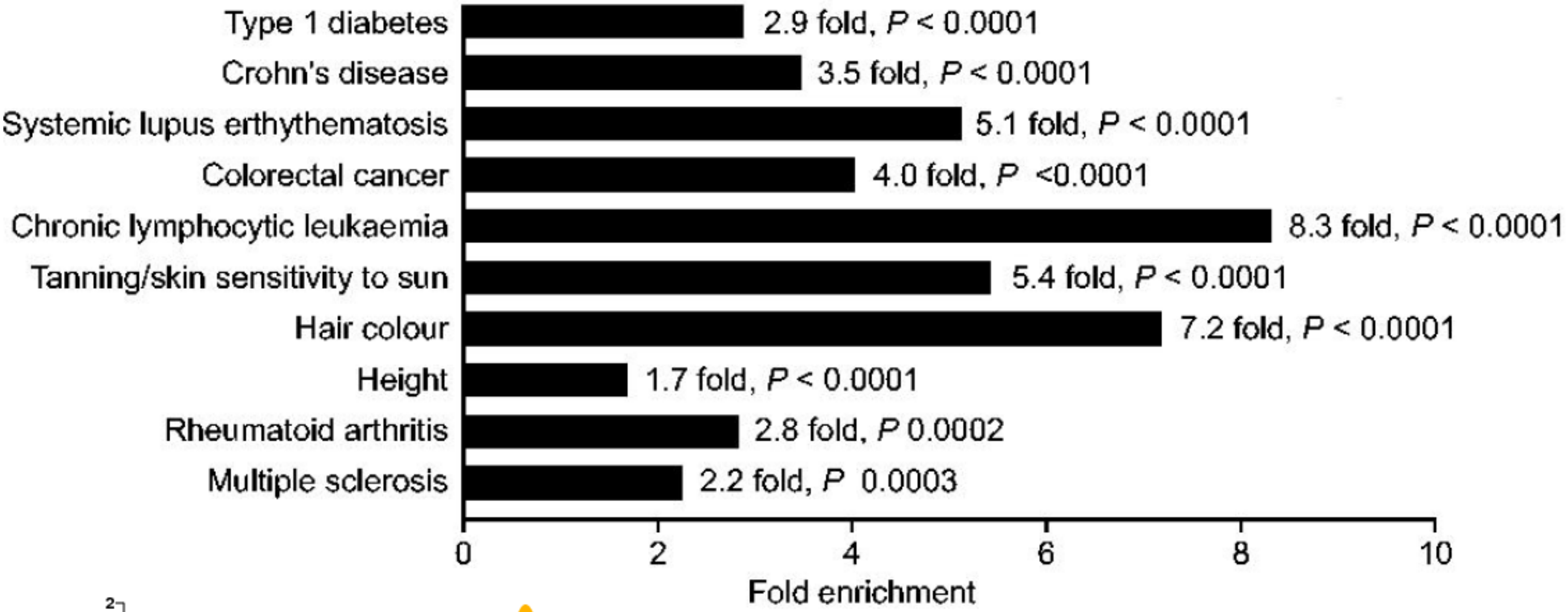
Systematic Localization of Common Disease-Associated Variation in Regulatory DNA
Matthew T. Maurano *et al.*
Science **337**, 1190 (2012);
DOI: 10.1126/science.1222794

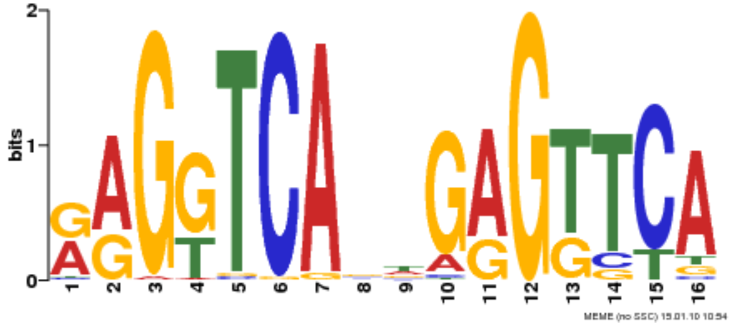
Objective identification of disease relevant cells



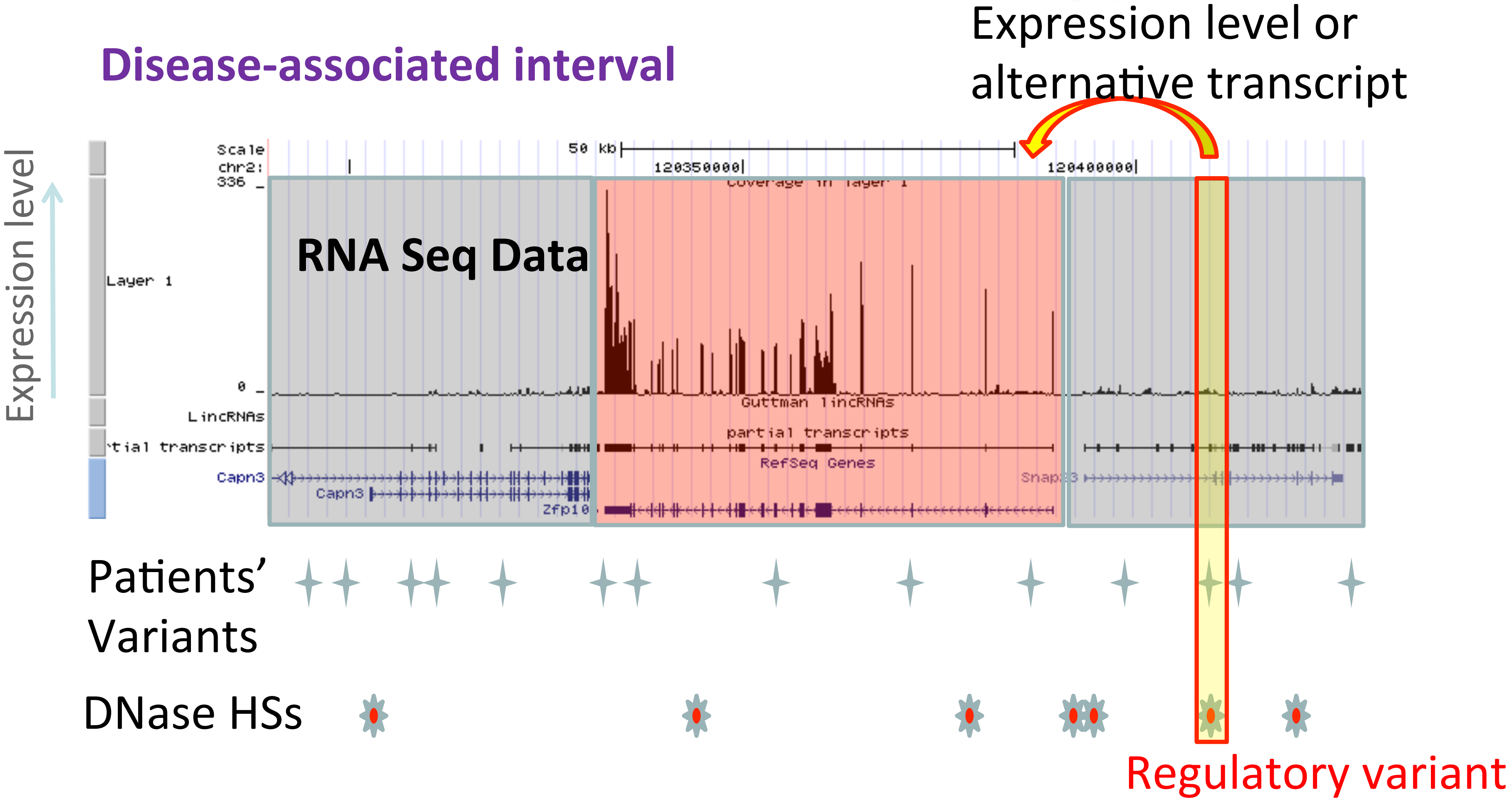
Systematic Localization of Common Disease-Associated Variation in Regulatory DNA
Matthew T. Maurano *et al.*
Science **337**, 1190 (2012);
DOI: 10.1126/science.1222794

Diseases/traits associated with VDR binding





Causal Variant Identification using a Disease-relevant cell type



Disease-associated interval

Expression level or alternative transcript

Expression level

RNA Seq Data

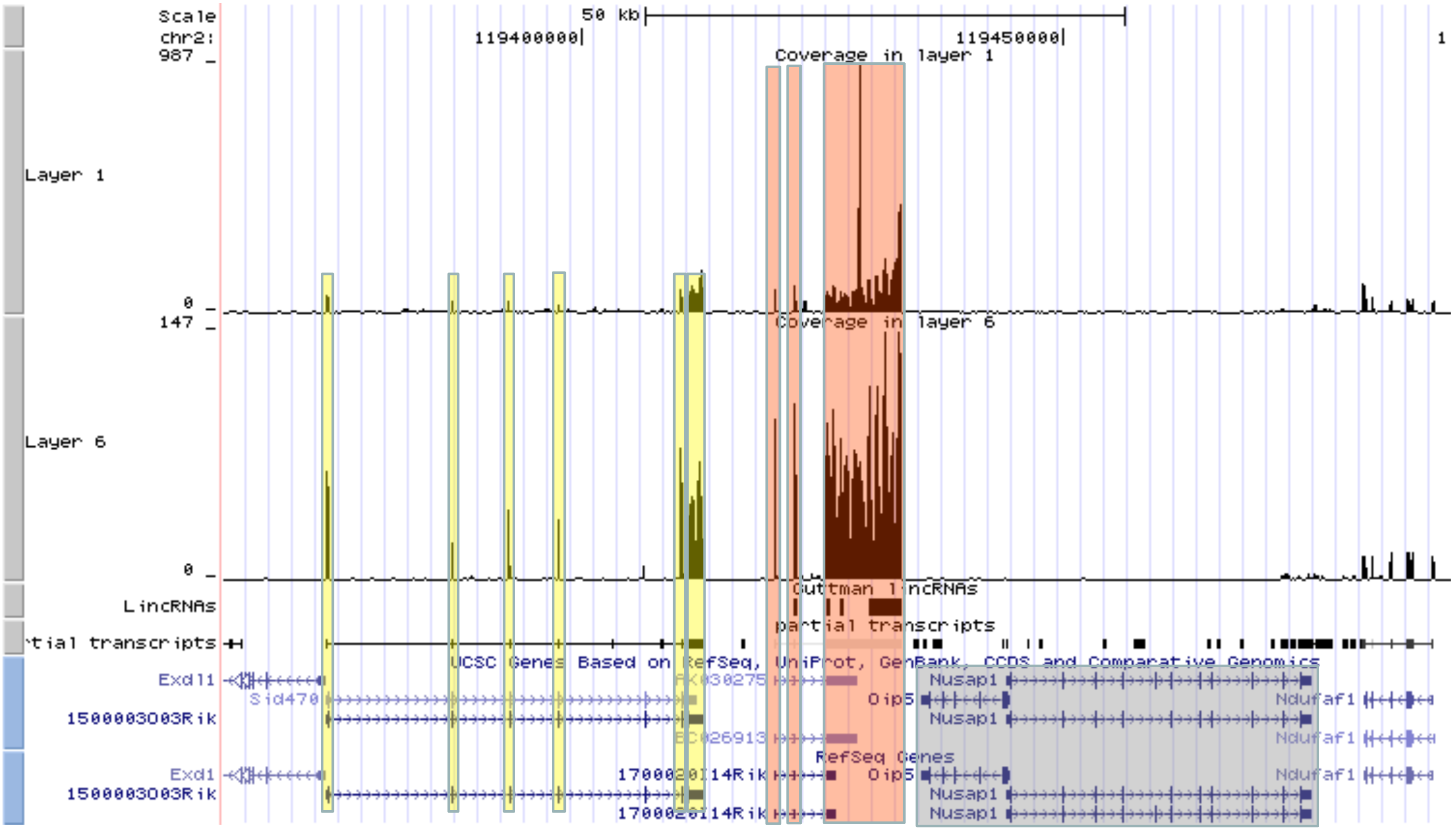
Patients' Variants

DNase HSs

Regulatory variant

Long noncoding RNA genes

Expression level



Expressed protein-coding gene

Expressed non-coding gene

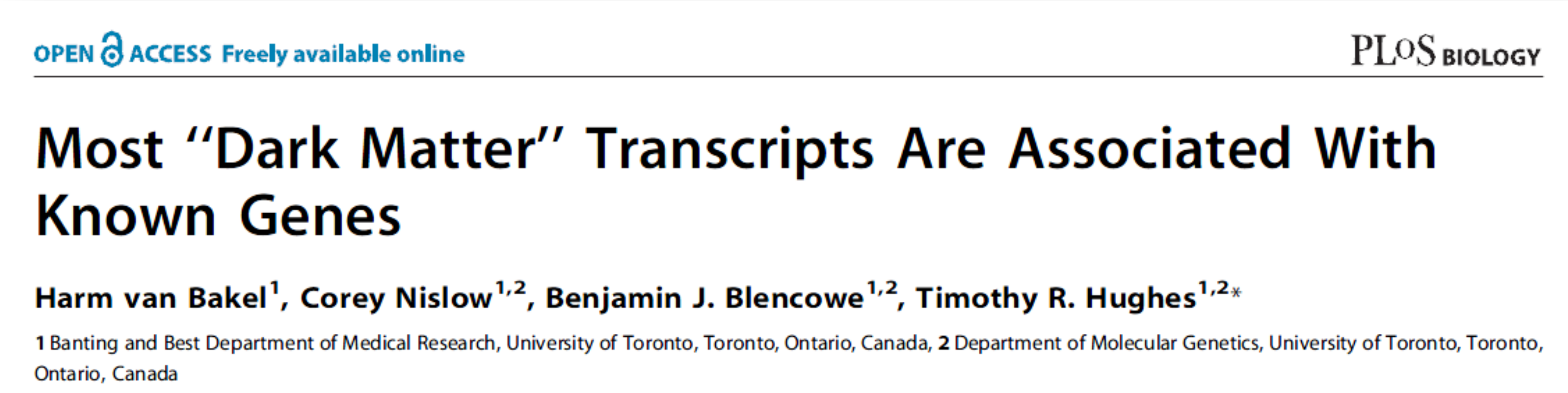
Non-expressed protein-coding genes

ENCODE
Encyclopedia of DNA Elements
nature.com/encode

"we annotated 9,640 manually curated long non-coding RNA (lncRNA) loci"

"80% of the detected lncRNAs are present in our samples in 1 or fewer copies per cell"

"62% of genomic bases are reproducibly represented in sequenced long (>200 nucleotides) RNA molecules or GENCODE exons. Of these bases, only 5.5% are explained by GENCODE exons. The majority of transcribed bases are within or overlapping annotated genes boundaries (*i.e.* intronic) and only 31% of bases in sequenced transcripts were intergenic"



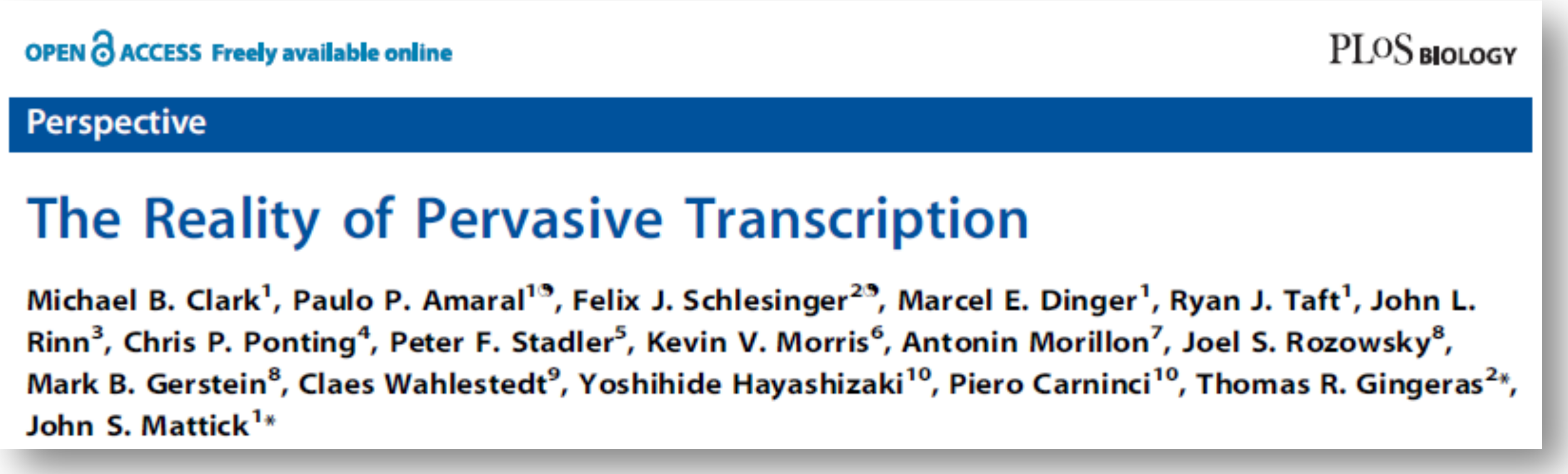
OPEN ACCESS Freely available online

PLoS BIOLOGY

Most "Dark Matter" Transcripts Are Associated With Known Genes

Harm van Bakel[1], Corey Nislow[1,2], Benjamin J. Blencowe[1,2], Timothy R. Hughes[1,2*]

1 Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada, 2 Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada



OPEN ACCESS Freely available online

PLoS BIOLOGY

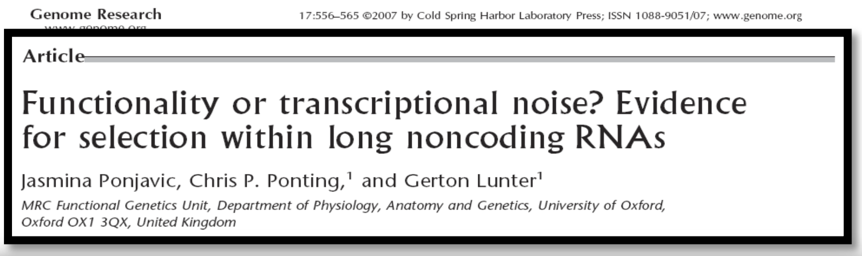
Perspective

The Reality of Pervasive Transcription

Michael B. Clark[1], Paulo P. Amaral[1☯], Felix J. Schlesinger[2☯], Marcel E. Dinger[1], Ryan J. Taft[1], John L. Rinn[3], Chris P. Ponting[4], Peter F. Stadler[5], Kevin V. Morris[6], Antonin Morillon[7], Joel S. Rozowsky[8], Mark B. Gerstein[8], Claes Wahlestedt[9], Yoshihide Hayashizaki[10], Piero Carninci[10], Thomas R. Gingeras[2*], John S. Mattick[1*]

Intergenic lncRNAs: objections to their functionality

- Sequence not conserved
- Little evidence for phenotypes
- Low expression levels
- Transcription not conserved?



Genome Research
www.genome.org

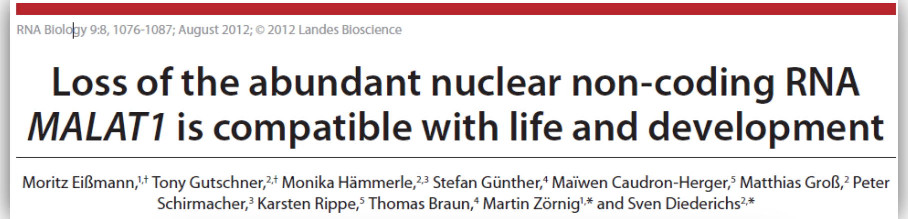
17:556–565 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

Article

Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs

Jasmina Ponjavic, Chris P. Ponting,[1] and Gerton Lunter[1]

MRC Functional Genetics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford OX1 3QX, United Kingdom



RNA Biology 9:8, 1076-1087; August 2012; © 2012 Landes Bioscience

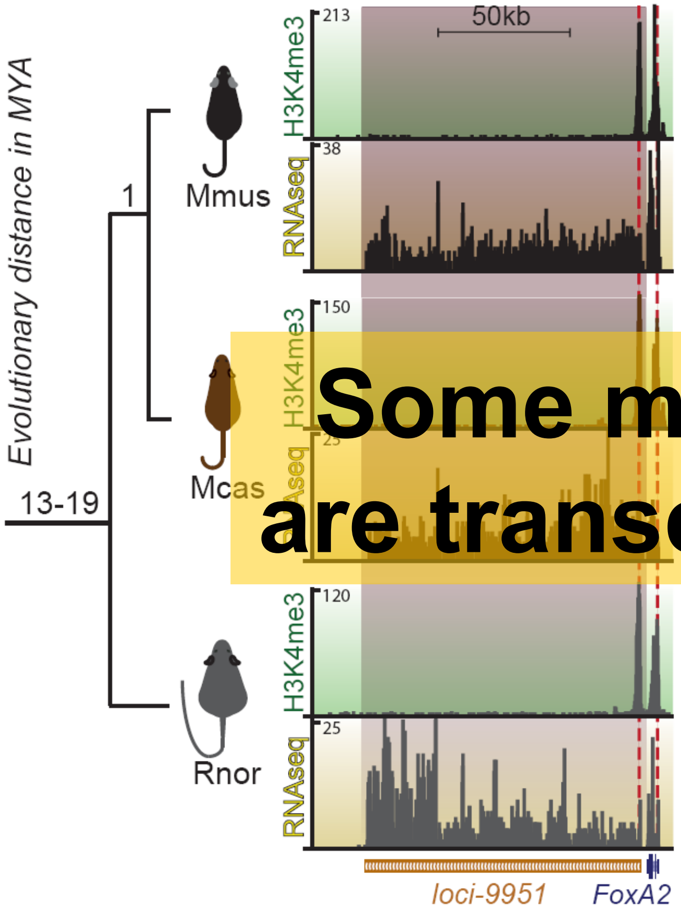
Loss of the abundant nuclear non-coding RNA *MALAT1* is compatible with life and development

Moritz Eißmann,[1,†] Tony Gutschner,[2,†] Monika Hämmerle,[2,3] Stefan Günther,[4] Maïwen Caudron-Herger,[5] Matthias Groß,[2] Peter Schirmacher,[3] Karsten Rippe,[5] Thomas Braun,[4] Martin Zörnig[1,*] and Sven Diederichs[2,*]

ENCODE
Encyclopedia of DNA Elements
nature.com/encode

?

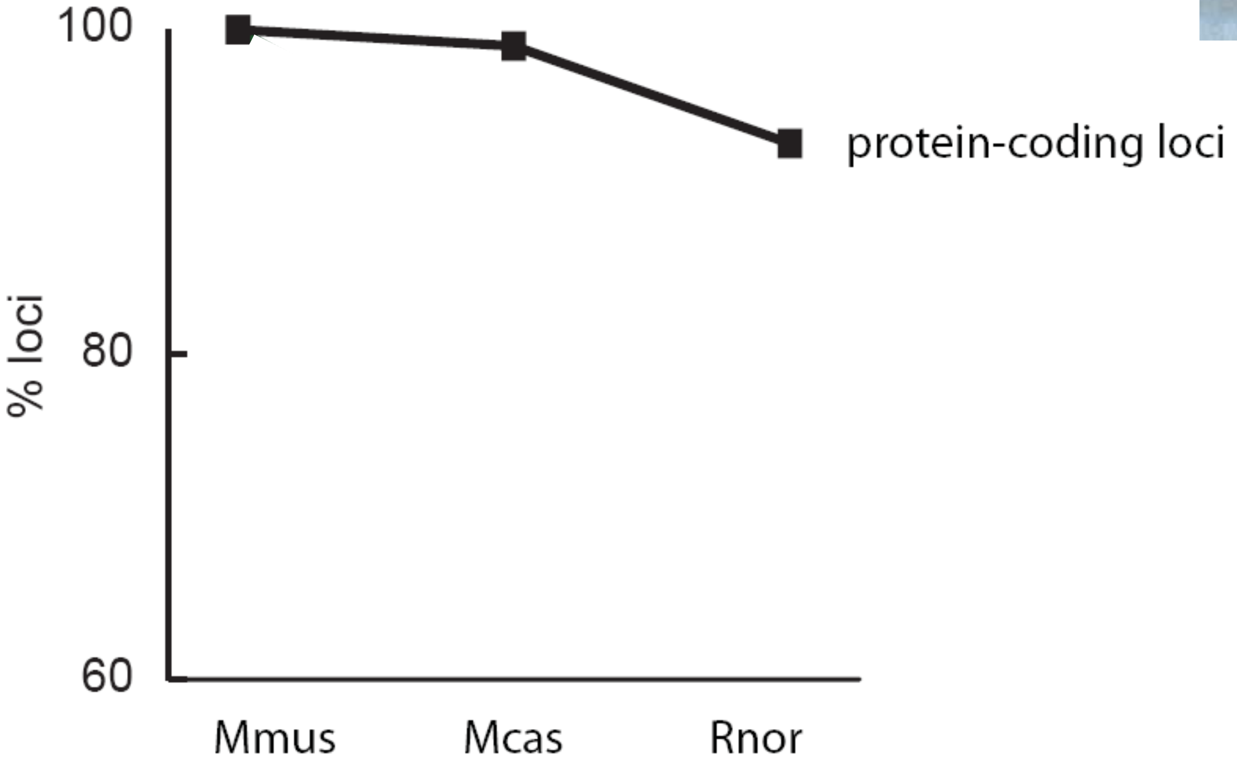
Loss of transcription but not of genomic sequence.



Some mammalian lincRNAs are transcribed only fleetingly.

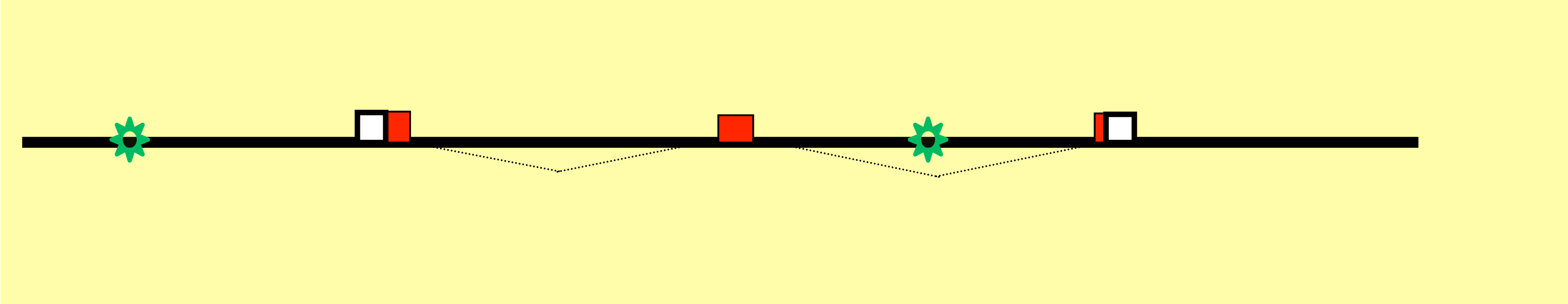
Conserved Rodents

LincRNA and protein-coding transcriptional turnover

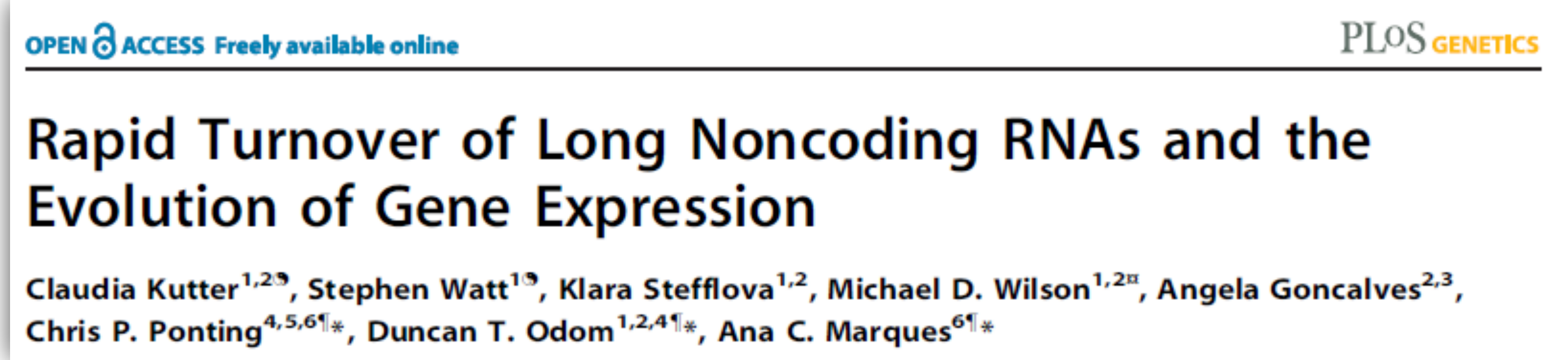


Ana Marques

Type 1: enhancer (e)RNAs – *Compensatory gains*



Coding Gene Expression



OPEN ACCESS Freely available online

PLoS GENETICS

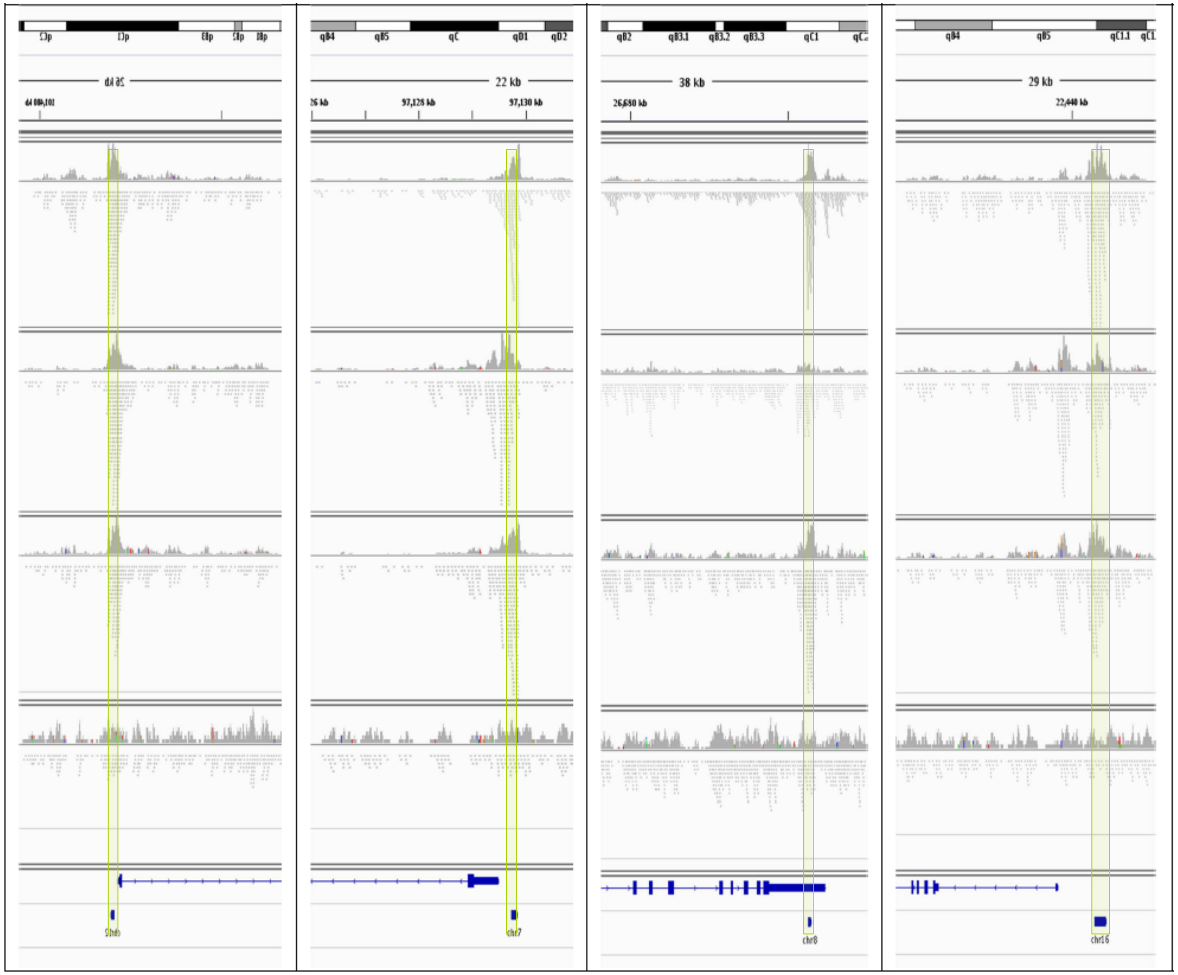
Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression

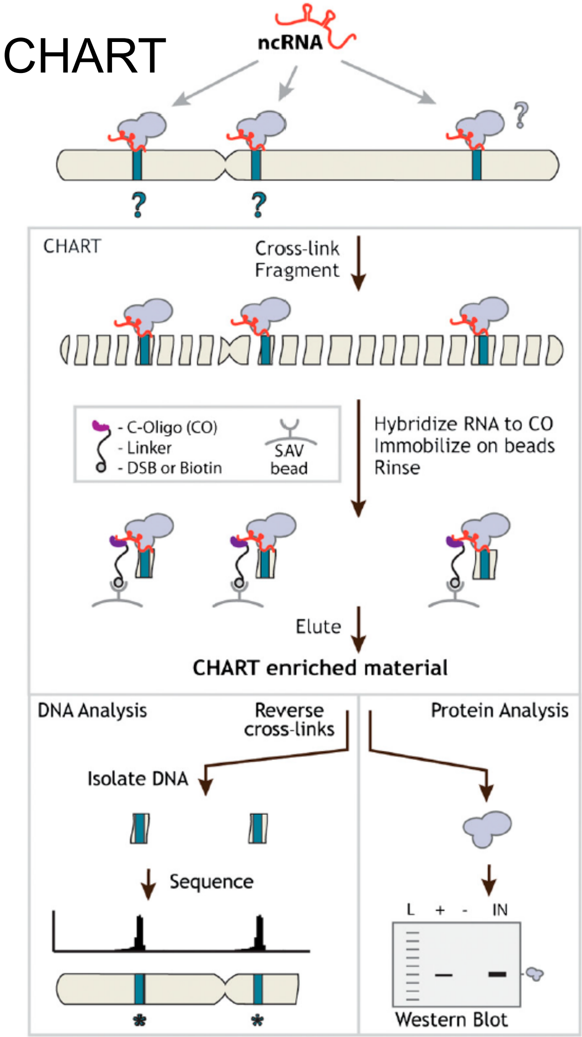
Claudia Kutter[1,2,9], Stephen Watt[1,9], Klara Stefflova[1,2], Michael D. Wilson[1,2¤], Angela Goncalves[2,3], Chris P. Ponting[4,5,6¶*], Duncan T. Odom[1,2,4¶*], Ana C. Marques[6¶*]

Type 2: ribonucleoprotein (rnp)RNAs

The genomic binding sites of a noncoding RNA

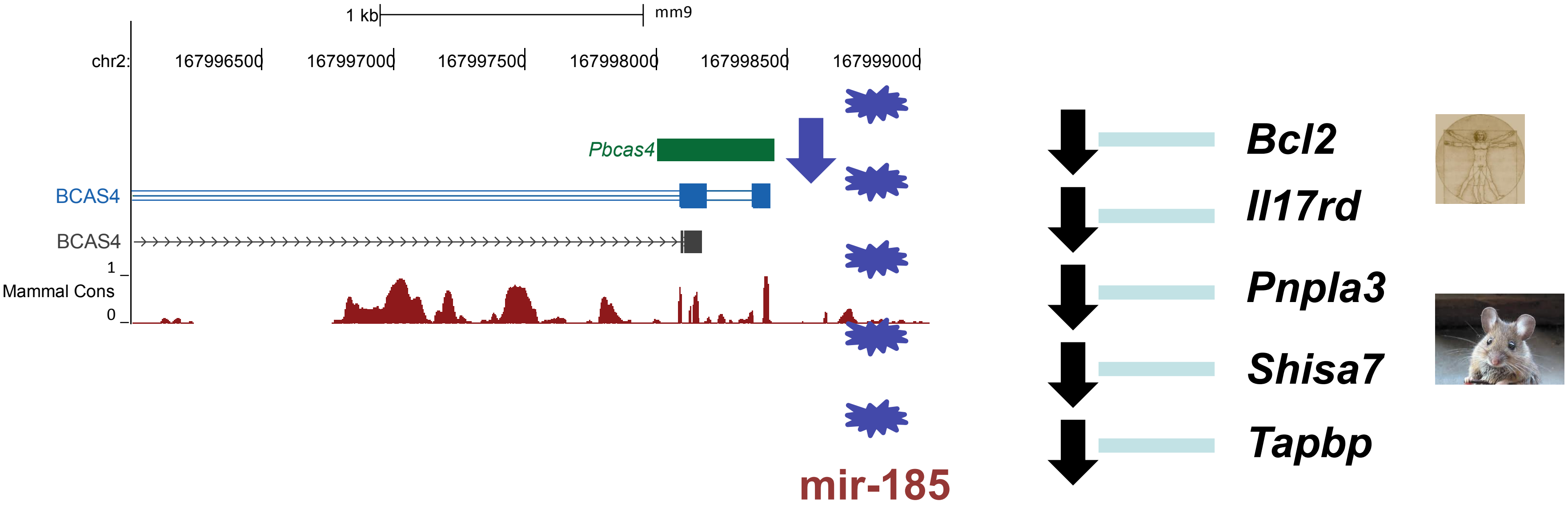
Matthew D. Simon[a], Charlotte I. Wang[b], Peter V. Kharchenko[c], Jason A. West[a], Brad A. Chapman[a], Artyom A. Alekseyenko[b], Mark L. Borowsky[a], Mitzi I. Kuroda[b], and Robert E. Kingston[a,1]





Vlada Chalei

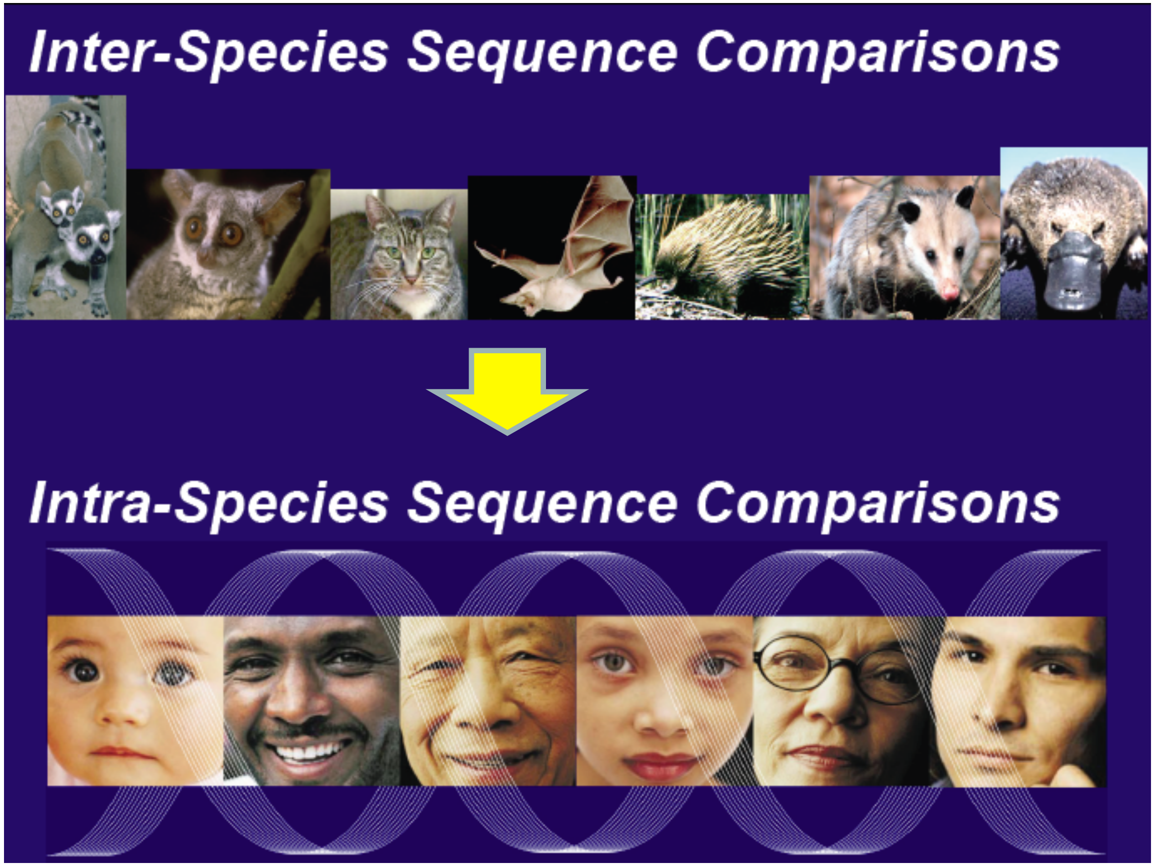
Type 3: competitive endogenous (decoy) RNAs (ceRNAs)



Mouse *Pbcas4*, a pseudogene of human *BCAS4*, is a conserved miRNA decoy

Ana Marques et al. "**Conservation of post-transcriptional roles of unitary pseudogenes suggests that mRNAs are often bifunctional" Genome Biology, 2012, 13:R102.**

Part 3: The future

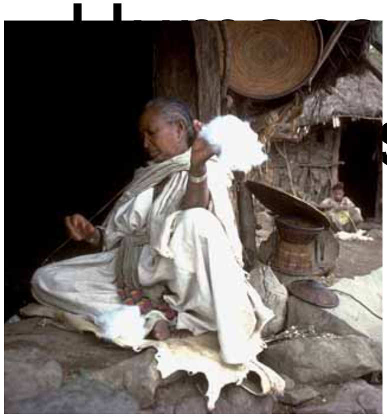


Inter-Species Sequence Comparisons

Intra-Species Sequence Comparisons

Different, but not that different

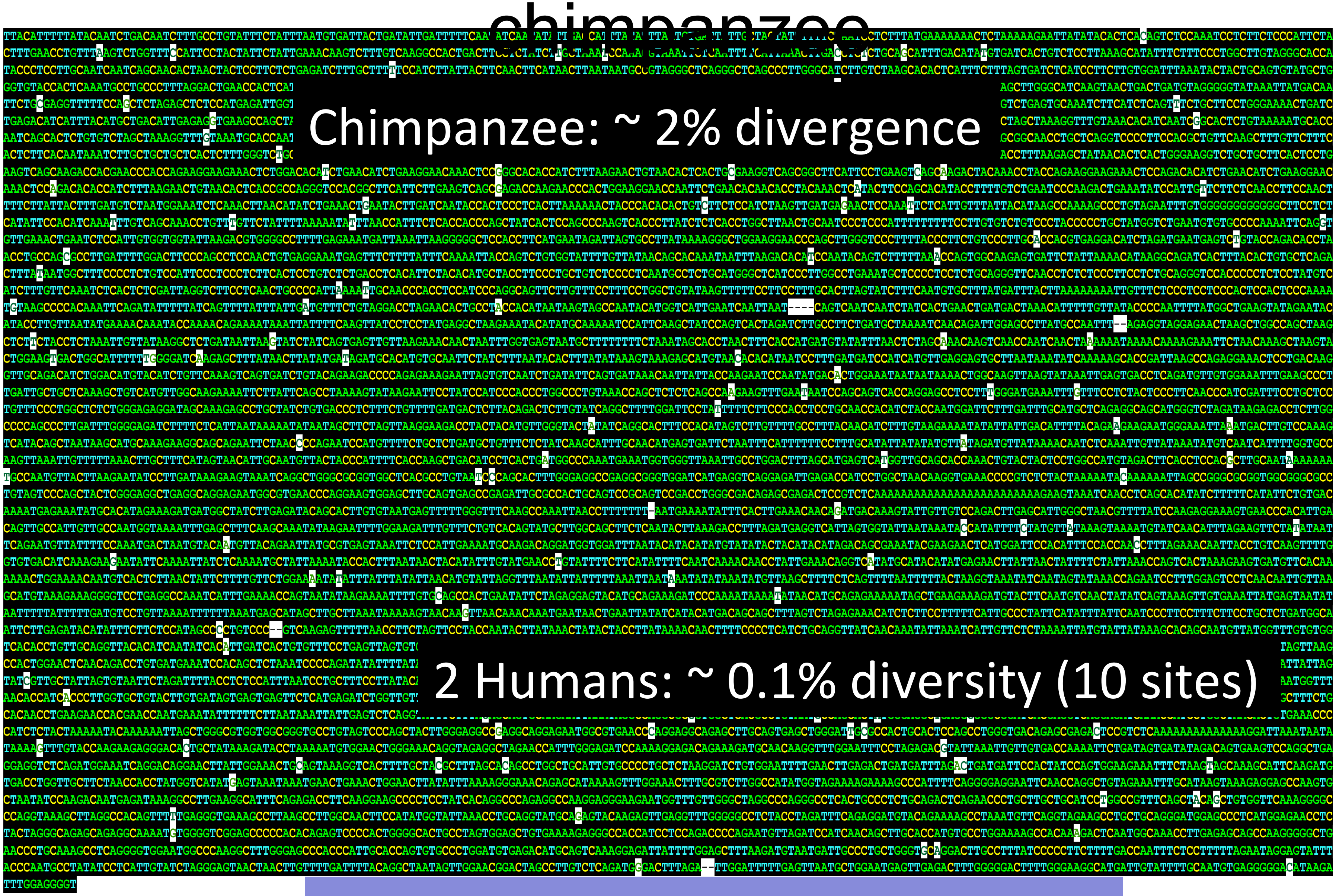
Humans are one of the least diverse organisms



Species	Diversity (percent)
Humans	0.08 - 0.1
Chimpanzees	0.12 - 0.17
Drosophila simulans	2
E. coli	5
HIV1	30

Photos from UN photo gallery www.un.org/av/photo

10,000 bases of human chr13 *vs* chimpanzee



Chimpanzee: ~ 2% divergence

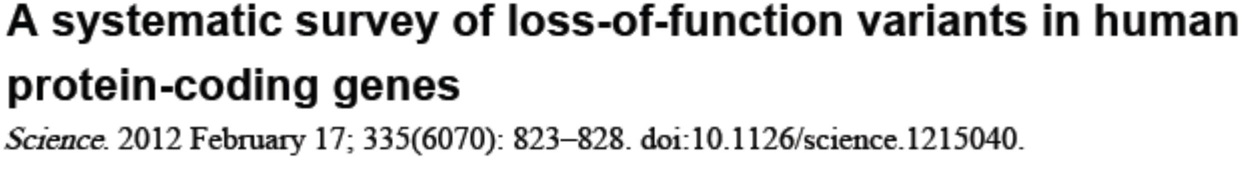
2 Humans: ~ 0.1% diversity (10 sites)

Page 1 of 300,000 of volume 1

No-one (genome) is perfect

Any European individual’s genomes are expected to carry:

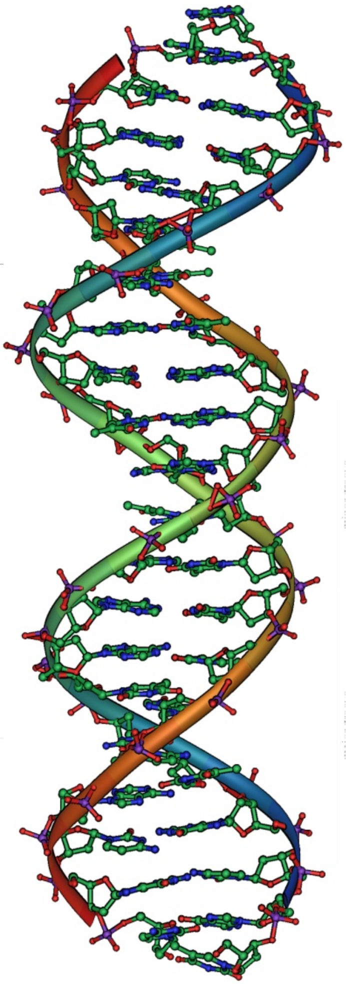
- 100 loss-of-function variants;
- of which 18 are in a homozygous state.
- Around a quarter of disrupting variants affect only a subset of transcripts.

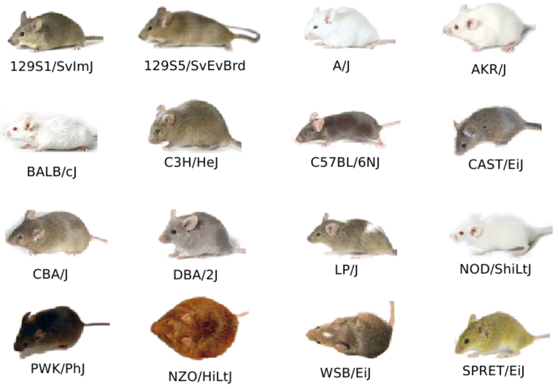


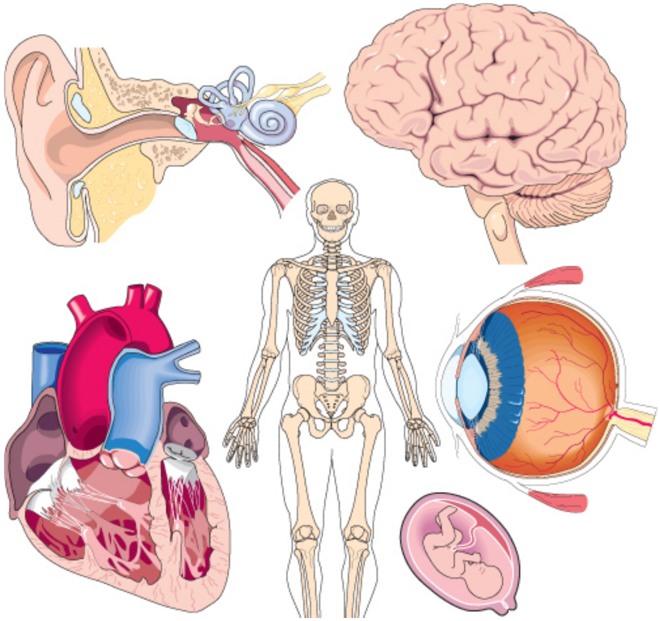
A systematic survey of loss-of-function variants in human protein-coding genes

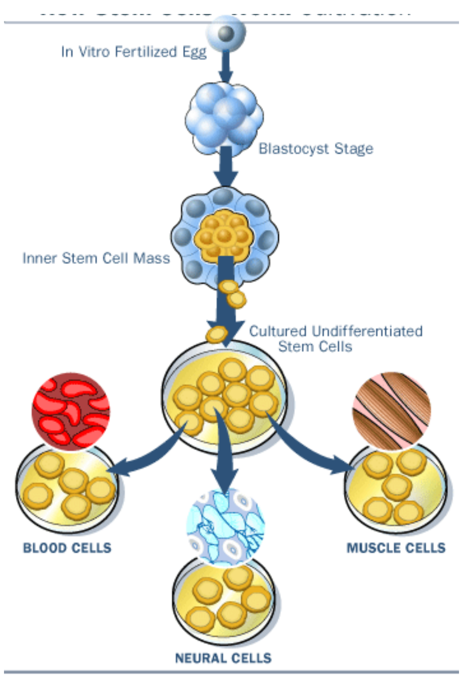
Science. 2012 February 17; 335(6070): 823–828. doi:10.1126/science.1215040.

The promise of the human genome is only being realised now with population genomics









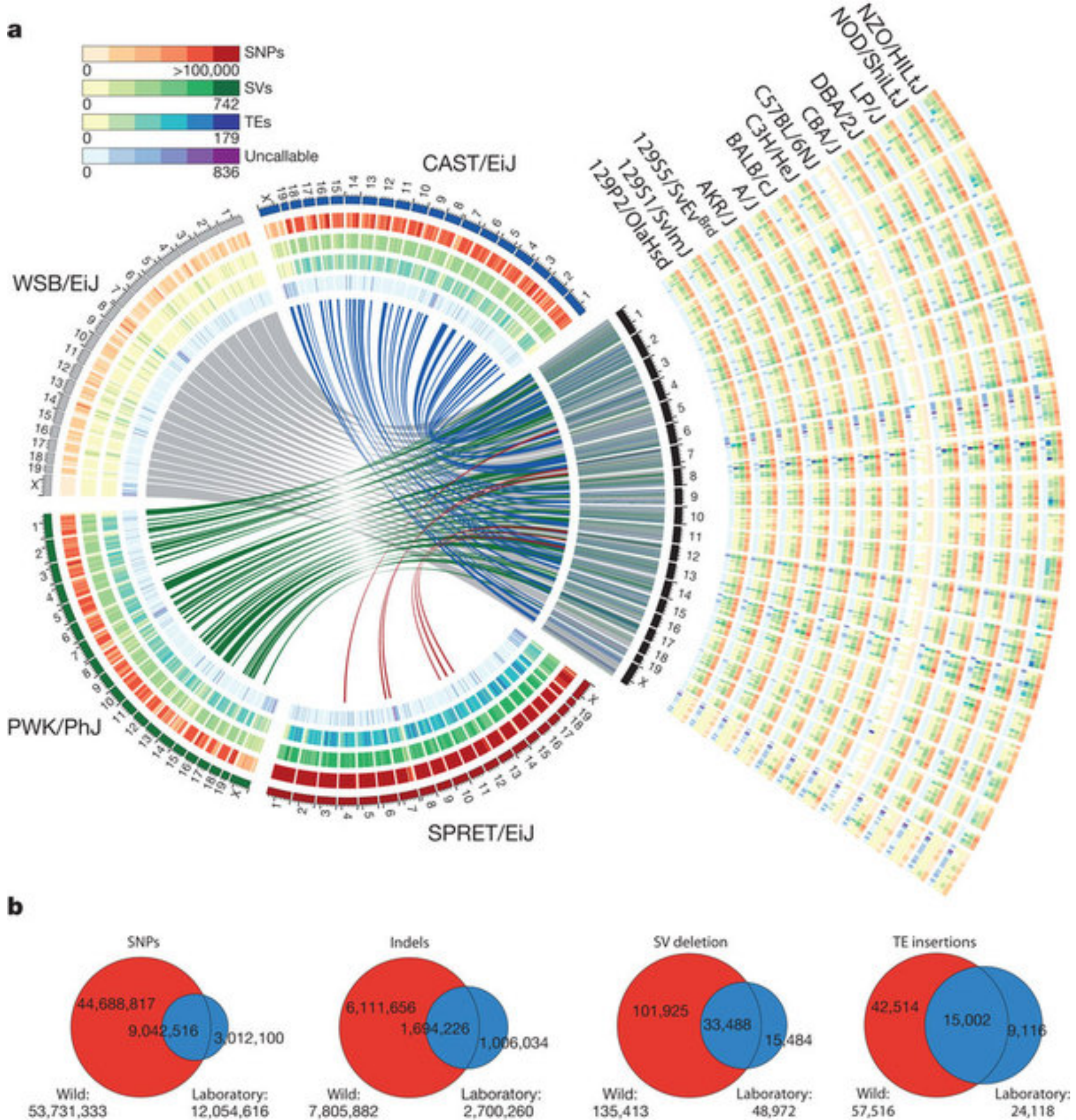
Major Issues in Population Genomics

- Genetic variation must underlie both pathological and non-pathological traits that show significant heritability
  - How do we locate these variants, and is there clinical use when they are found?
- Genetic variation must also underlie species differences.
  - How do we locate these variants?
- Do orthologous genes control equivalent traits in different species?
  - Can model organisms appropriately model human traits?
- How often do somatic variants cause disease (outside of cancer)?
  - How genomically mosaic is any person?

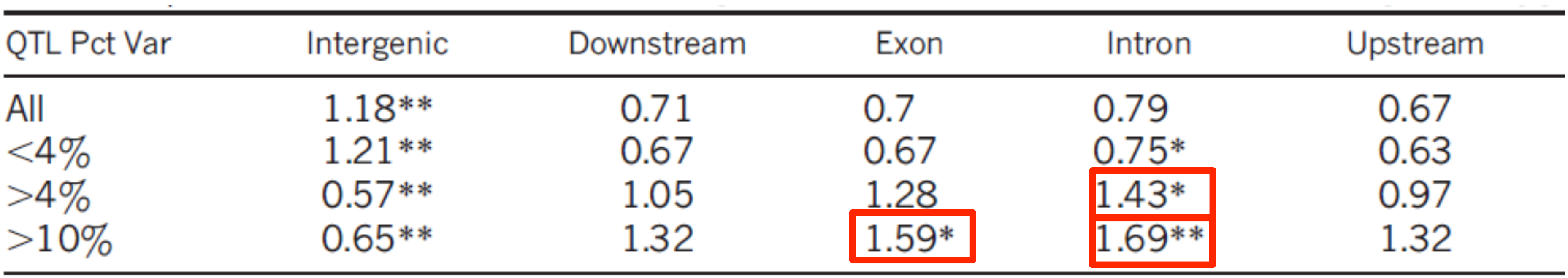
Population genomics requires detailed phenotyping

Vertebrate population genomics will initially study human and rodent species.

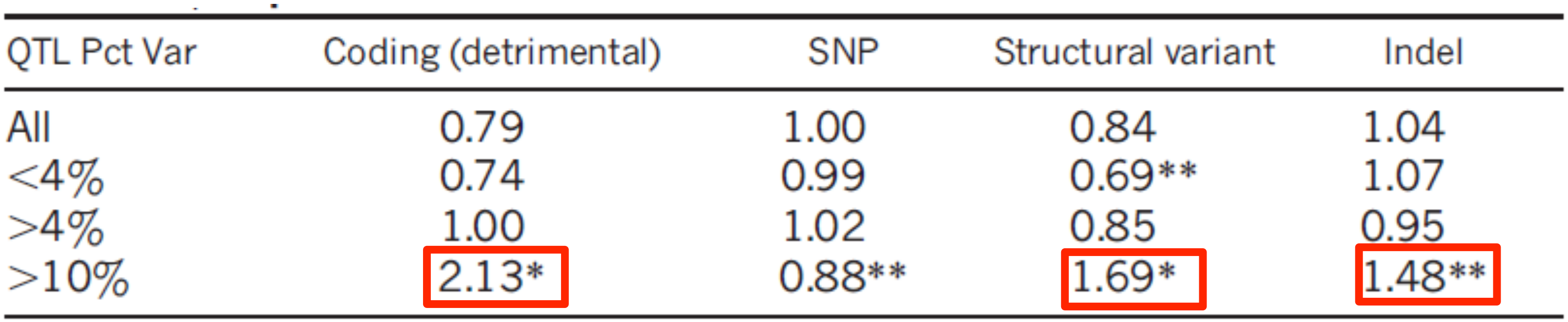
Genome Sequences of 17 Mouse Strains



Molecular Nature of Sequence Variants and their Effect on Phenotypic Variation



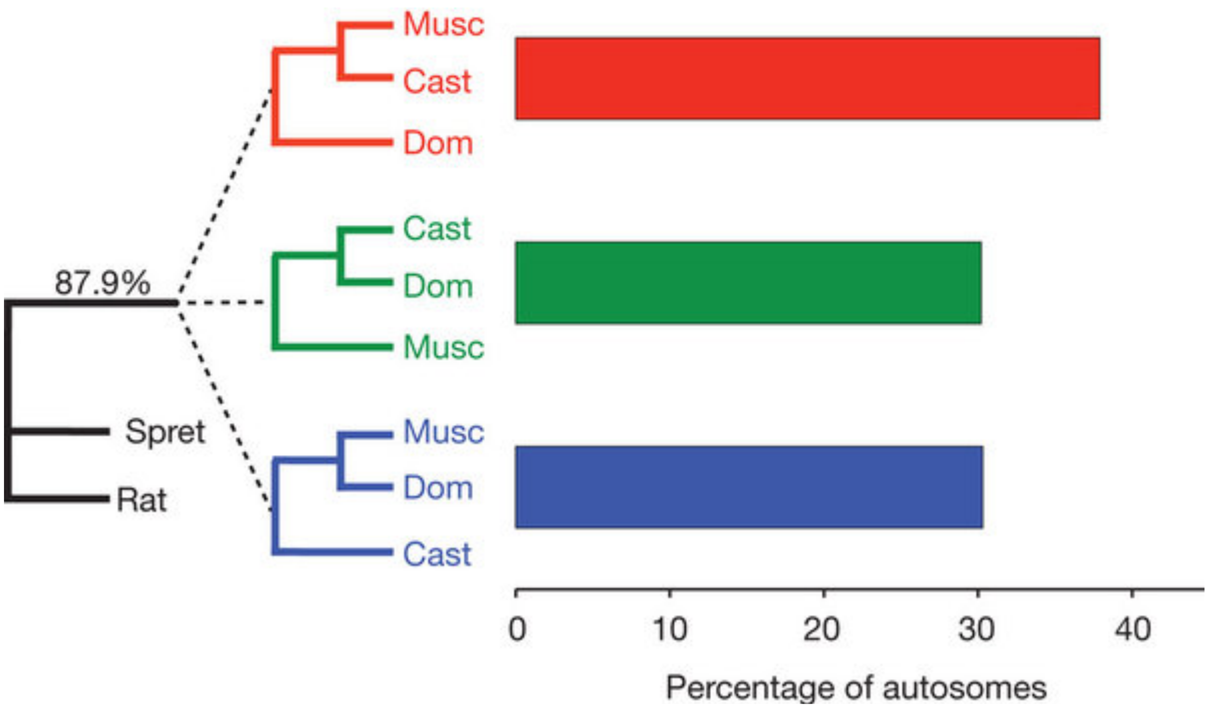
QTL Pct Var	Intergenic	Downstream	Exon	Intron	Upstream
All	1.18**	0.71	0.7	0.79	0.67
<4%	1.21**	0.67	0.67	0.75*	0.63
>4%	0.57**	1.05	1.28	1.43*	0.97
>10%	0.65**	1.32	1.59*	1.69**	1.32

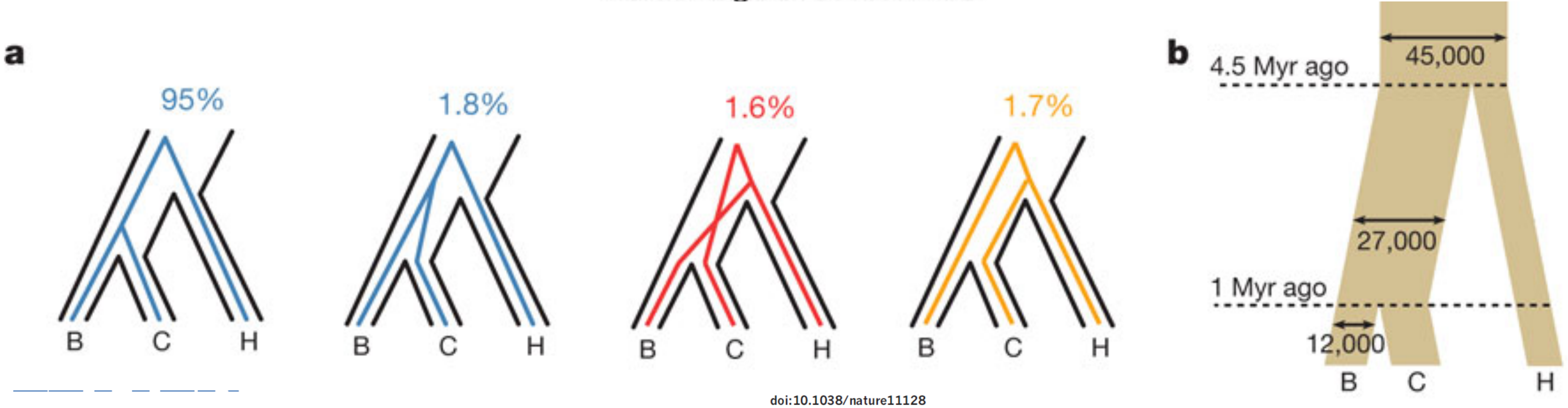


QTL Pct Var	Coding (detrimental)	SNP	Structural variant	Indel
All	0.79	1.00	0.84	1.04
<4%	0.74	0.99	0.69**	1.07
>4%	1.00	1.02	0.85	0.95
>10%	2.13*	0.88**	1.69*	1.48**

Red boxes indicate significantly large effect size variants

Incomplete Lineage Sorting





doi:10.1038/nature11128

The bonobo genome compared with the chimpanzee and human genomes

The genomes of many (most?) animal species will soon be sequenced

(David Haussler, Stephen O'Brien, & Oliver Ryder)

Porifera

Amphimedon queenslandica, a sponge (2009[1])

Placozoa

Trichoplax adhaerens, a Placozoan (2008[2])

Cnidaria

Hydra magnipapillata, a model medusozoan (2010[3])

Nematostella vectensis, a model anemone (starlet sea anemone) (2007[4])

Deuterostomia

Echinoderms

Strongylocentrotus purpuratus, a sea urchin and model deuterostome (2006[5])

Hemichordates

Saccoglossus kowalevskii, an acorn worm (2009)[6])

Urochordates

Ciona intestinalis, a tunicate (2003[7])

Ciona savignyi, a tunicate (2007[8])

Cephalochordates

Branchiostoma floridae, a lancelet (2008[9])

Cyclostomes

Petromyzon marinus, a lamprey (2009[10])

Cartilaginous Fish

Callorhinchus milii, an elephant shark (2007[11])

Bony Fish

Danio rerio, a zebrafish (2007[12]) (order Cypriniformes)

Gadus morhua, Atlantic cod (2011[13]) (order Gadiformes)

Gasterosteus aculeatus, Three-spined stickleback (2006, 2012[14]) (order Gasterosteiformes)

Latimeria chalumnae, West Indian Ocean coelacanth and oldest known living lineage of Sarcopterygii ([15]) (or

Oryzias latipes, medaka (2007)[16] (order Beloniformes)

Takifugu rubripes, a puffer fish ([17] International Fugu Genome Consortium[18] 2002[19]) (order Tetraodontiform

Tetraodon nigroviridis, a puffer fish (2004[20]) (order Tetraodontiformes)

Amphibians

Xenopus tropicalis, Western clawed frog (2010[21])

Sea-change in genomics?

For over 10 years, genomics has been dominated by the large, well-funded genome sequencing centers.

In the next 10 years, research may become more equitable with investment being placed increasingly on analysis (*people*) rather than on technology & hardware.

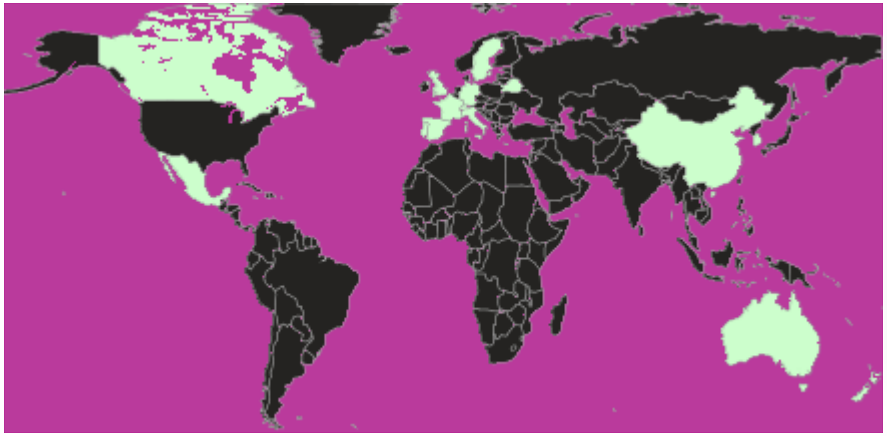
Sequencing is already cheap & analysis expensive, so you should already consider yourself, as a talented, now well-qualified, computational genomics researcher, to be a much sought-after individual.

The pace of change in genomics, once more, is a great leveller.

The most important commodity in genomics is ideas.

Good luck.

Thanks to:
- all group members past & present
- members of all genome consortia



Please contact me if ever you're interested in a post-doc / fellowship etc. in Oxford.
Chris.Ponting@dpag.ox.ac.uk

CGAT
COMPUTATIONAL GENOMICS
ANALYSIS AND TRAINING

elixir
UNITED
KINGDOM