

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.



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



Dr. Thomas H. Roderick

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.

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

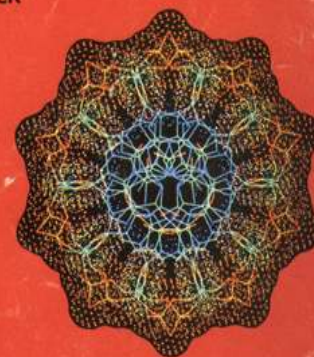
ISSN 0888-7543

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

15 February 2001

nature

£5.45 €8.25 FRF54 DM16 Yen1800

www.nature.com

the human genome

Nuclear fission
Five-dimensional
energy landscapes

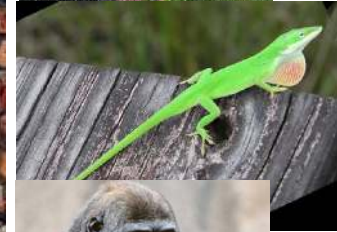
Seafloor spreading
The view from under
the Arctic ice

Career prospects
Sequence creates new
opportunities

naturejobs
genomics special



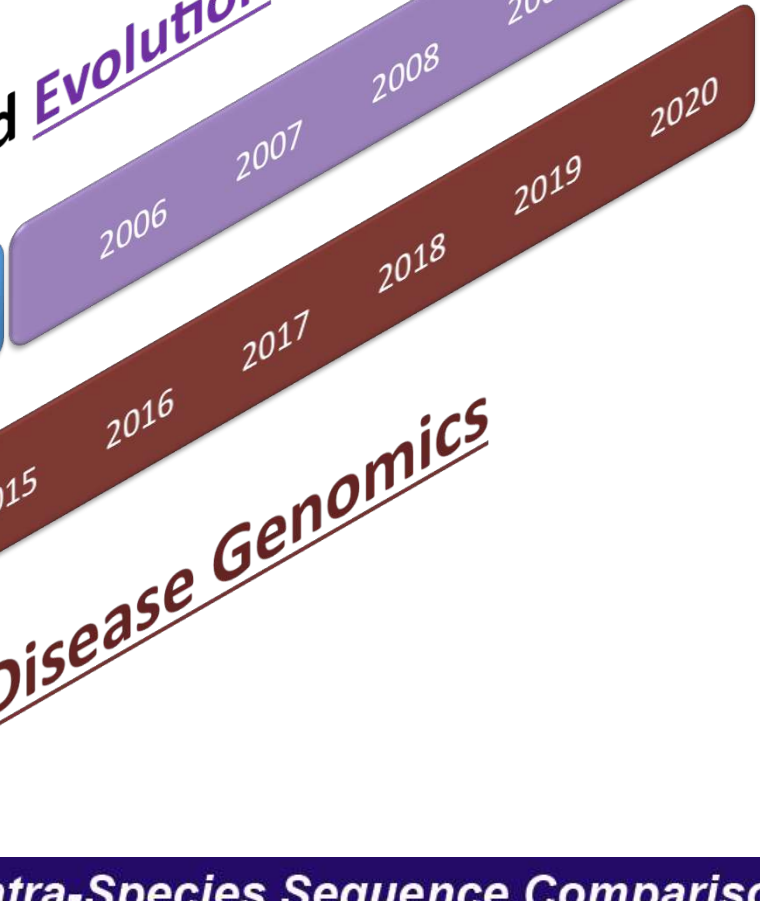
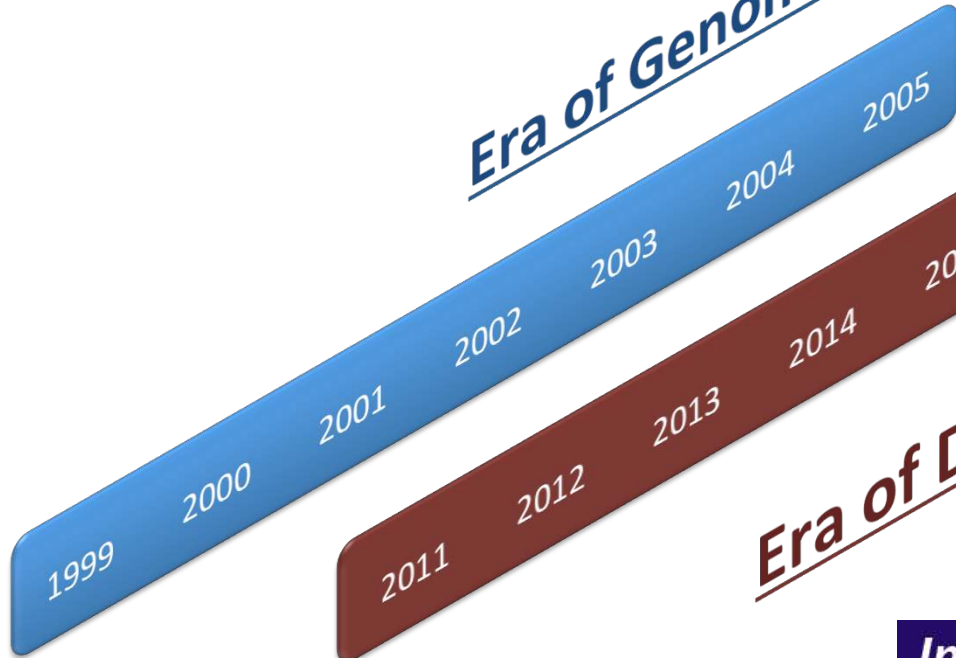
Finished



Inter-Species Sequence Comparisons



Era of Genomes and Evolution

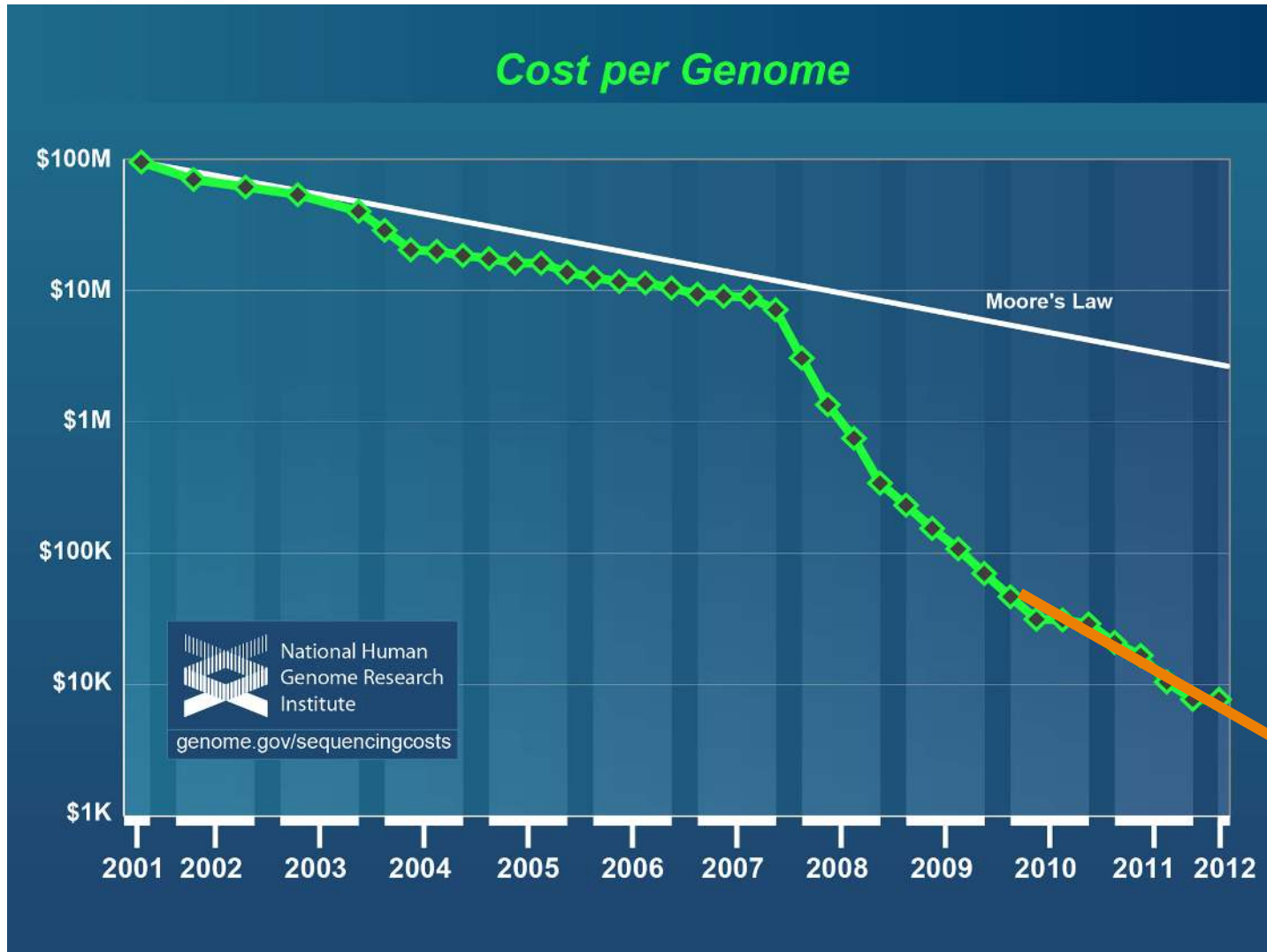


Era of Disease Genomics

Intra-Species Sequence Comparisons



the first human genome cost ~ \$3b



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

15 February 2001

nature

ISSN 0950-8688 (print) 0950-8689 (online) DOI: 10.1038/35001

www.nature.com

Joint coordinator of the
'Proteins' section.

Unqualified to do so.

Weekly teleconference calls.

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

Career prospects

Sequence creates new
opportunities

naturejobs
genomics special

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

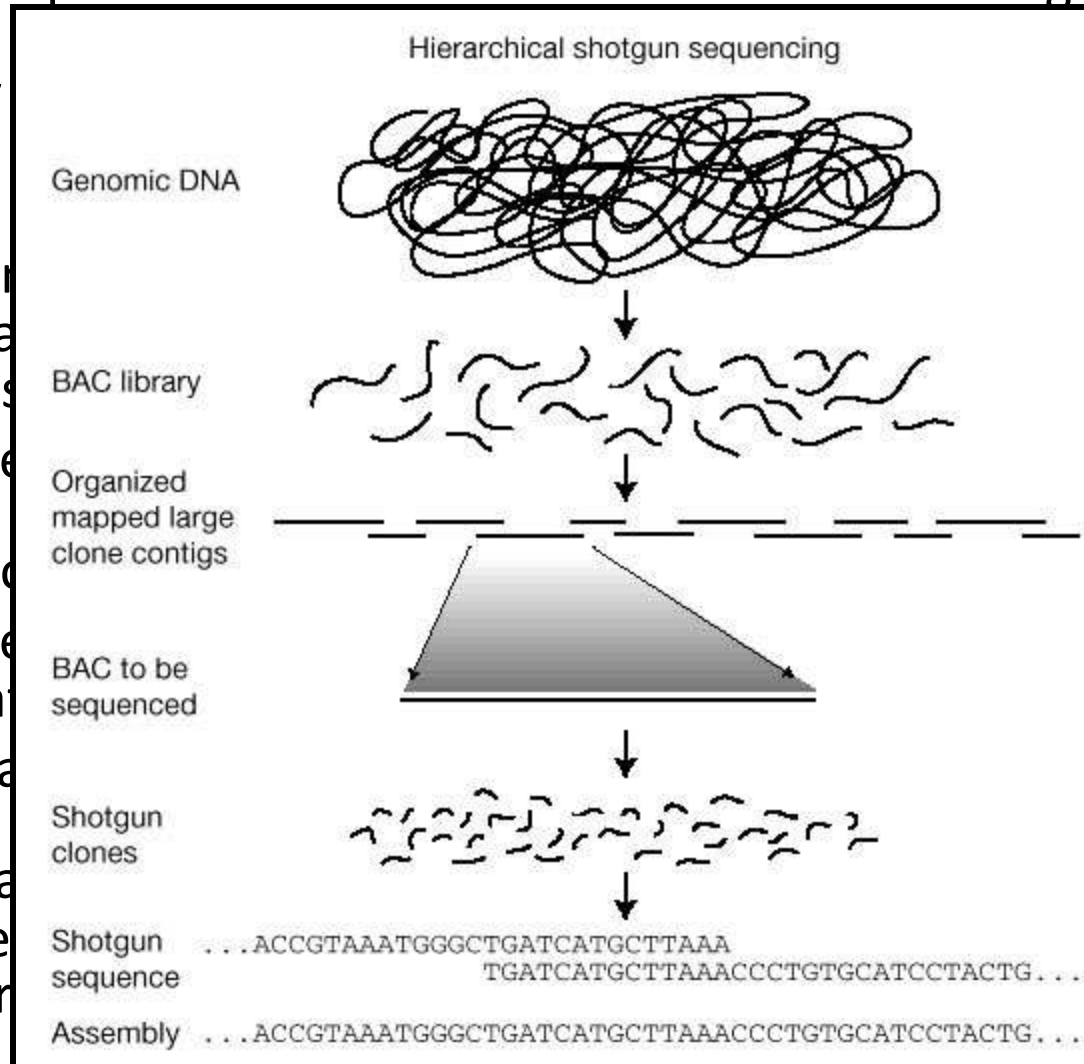
* A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.



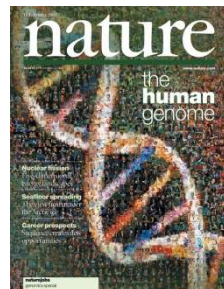
The Sequence of the Human Genome

J. Craig Venter,^{1a} 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. Huxon,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,² 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 Flanagan,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. Morkulov,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 Wenyang Zhong,¹ Shiaoqing 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 Ferreira,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 Kerriellen 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 Foster,¹ Harold Gire,¹ Stephen Glanowski,¹ Kenneth Glasser,¹ Anna Glodok,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

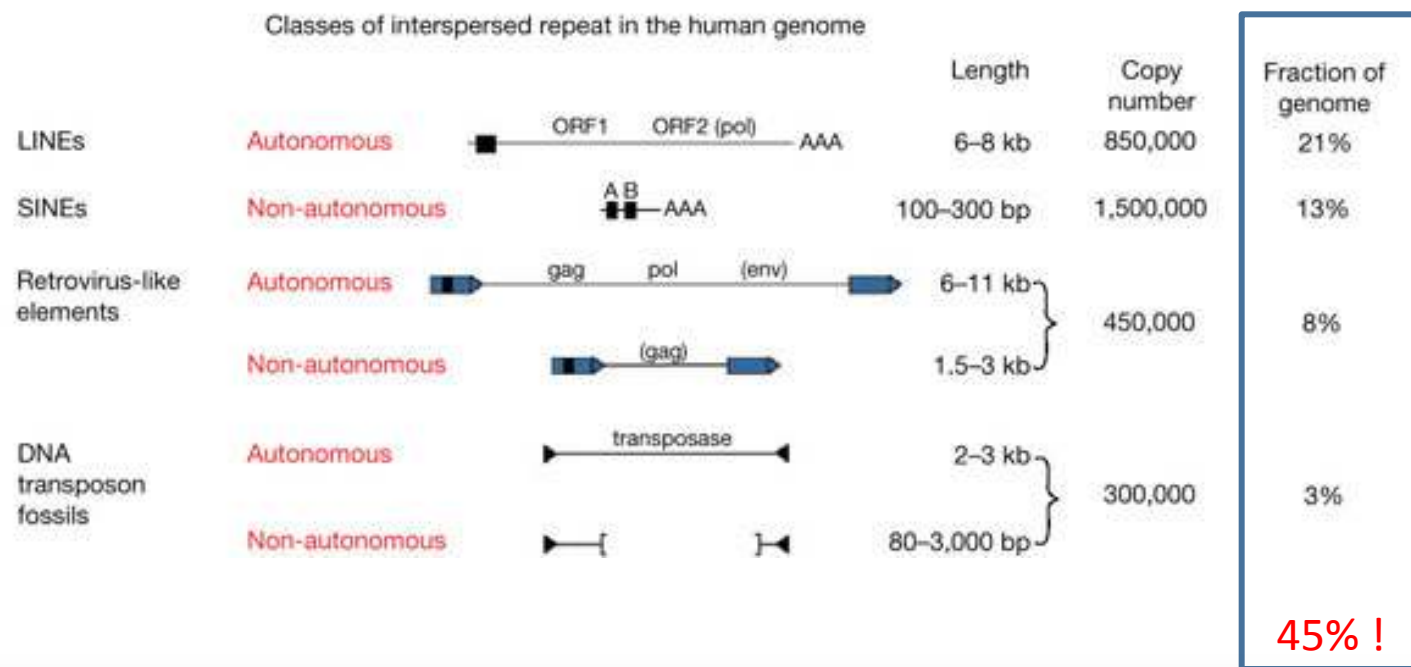
- Here we report the results of a collaboration involving 20 groups from the German Genome Project, the first draft of the human genome.
- The draft sequence covers approximately 1.5% of the human genome and is available in public databases.
- The sequence is based on a hierarchical shotgun sequencing strategy, with roughly 100,000 clones sequenced.
- The sequence is being updated as more clones are sequenced.
- The task is to fill the gaps and complete the sequence.
- The task is to complete the sequence and proceed to the next stage of the project.



ence,
human
sical map
human
ic
od, with
roughly
restriction
osing all
illion bases
ity of the
ould



Transposable elements dominate the human genome

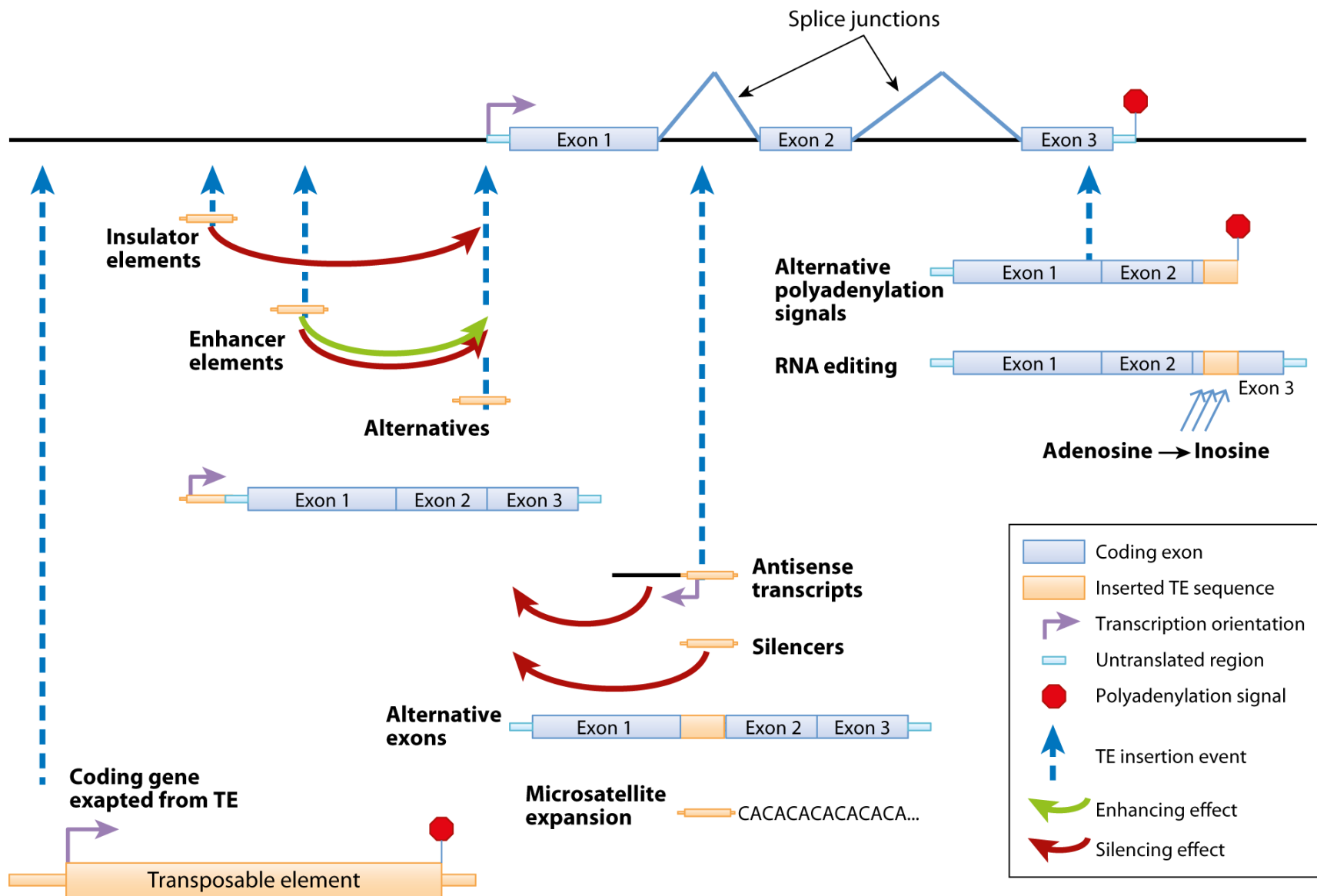


OPEN ACCESS Freely available online

PLOS GENETICS

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

A. P. Jason de Koning¹, Wanjun Gu^{1*}, Todd A. Castoe¹, Mark A. Batzer², David D. Pollock^{1*}



Isochores

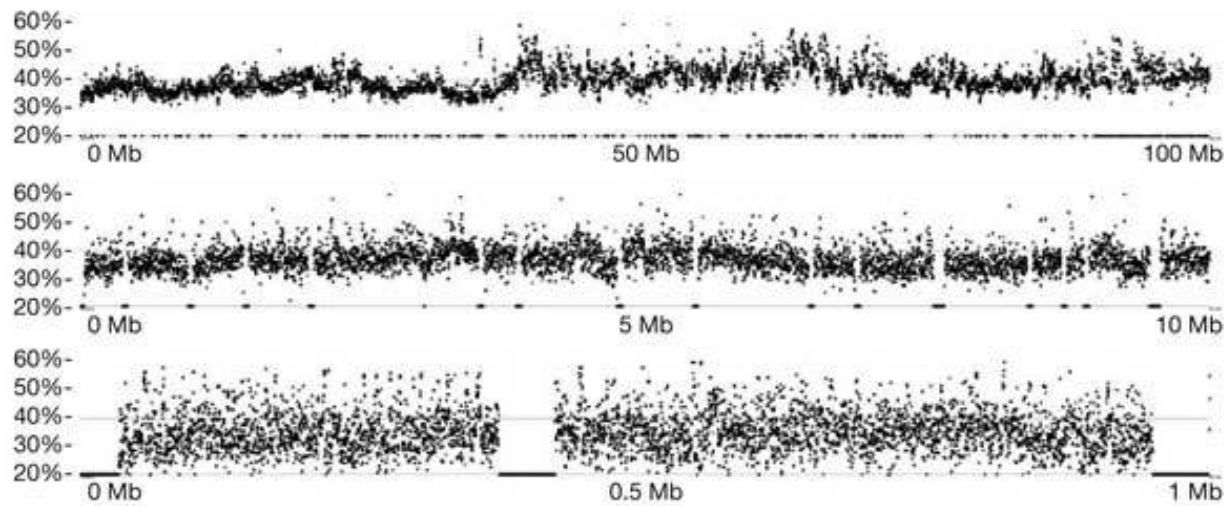
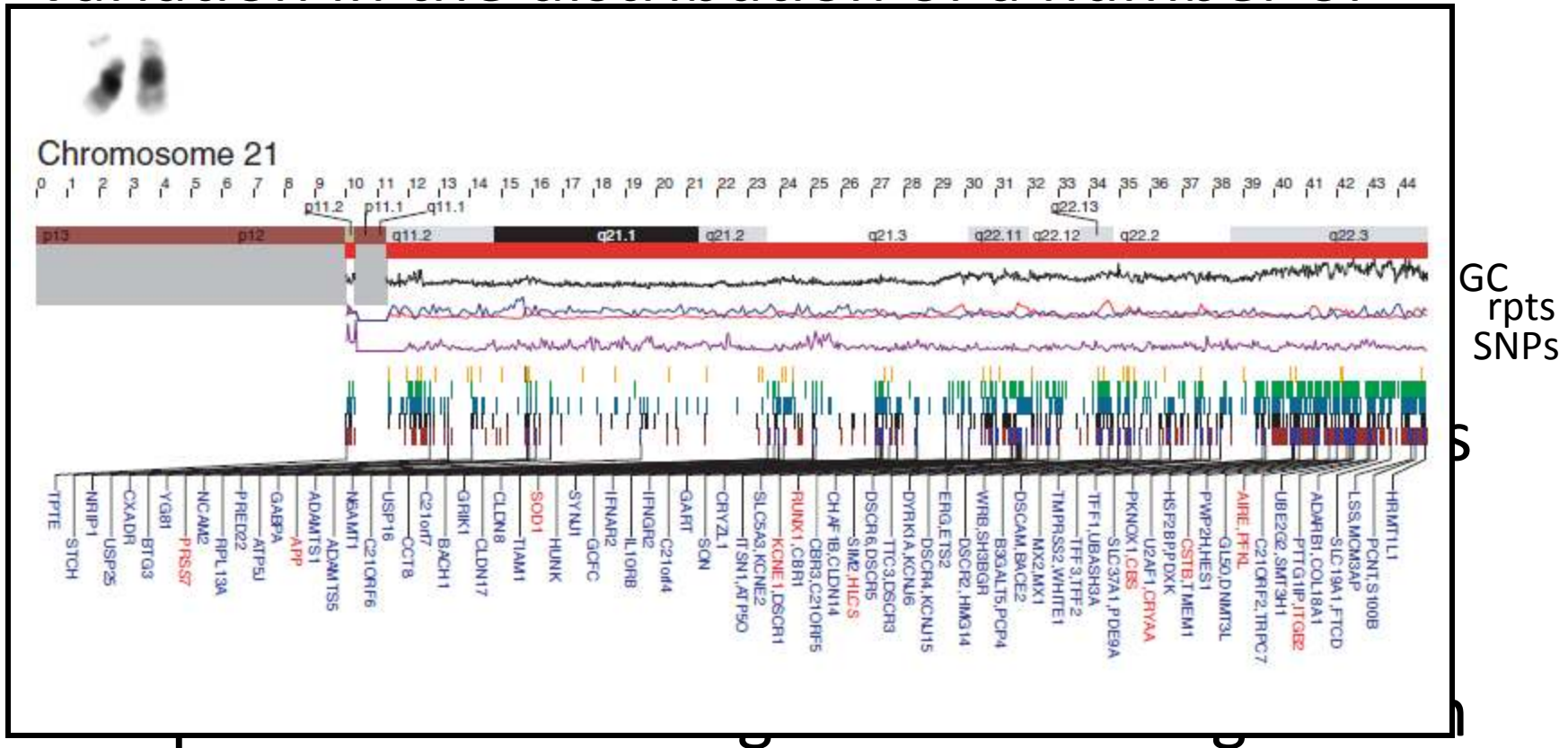


FIGURE 13. Variation in GC content at various scales

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



GC
rpts
SNPs

3

1

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**



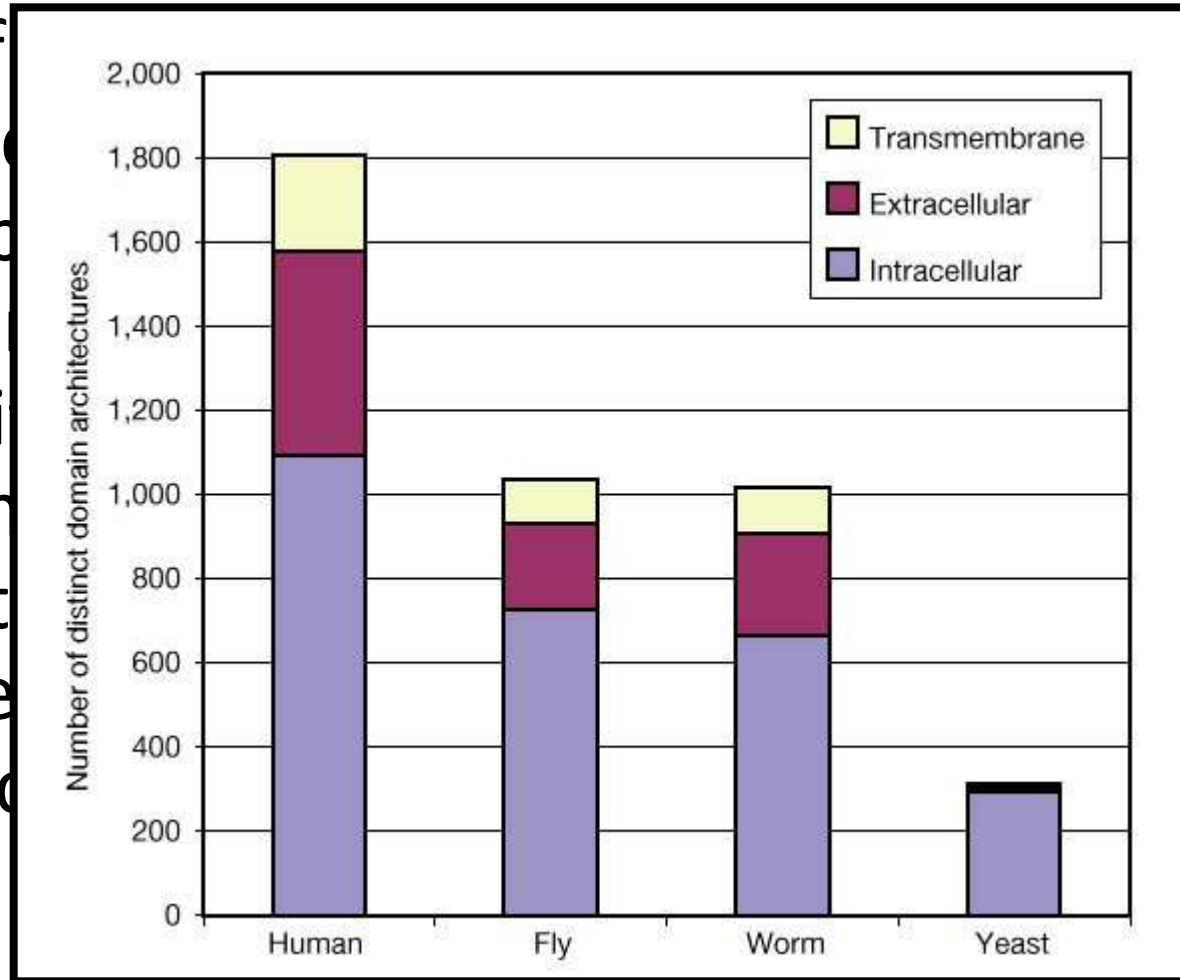
The Evolutionary Landscape of Alternative Splicing in Vertebrate Species

Nuno L. Barbosa-Morais *et al.*

Science 338, 1587 (2012);

DOI: 10.1126/science.1230612

- The f
enco
comp
due i
speci
estim
fact t
pre-e
collec

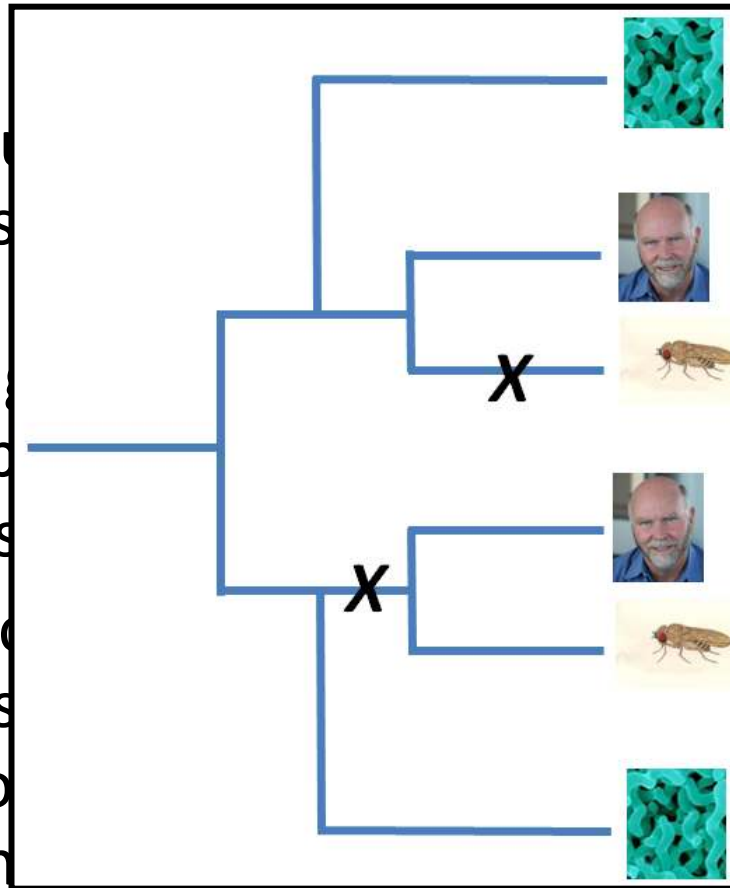


his is
te-
the
ranged

- 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 about 1% of the human genome is derived from transposons, there has been a marked decline in the number of transposons in the hominid lineage since we became competent for (LTR) retroposons.

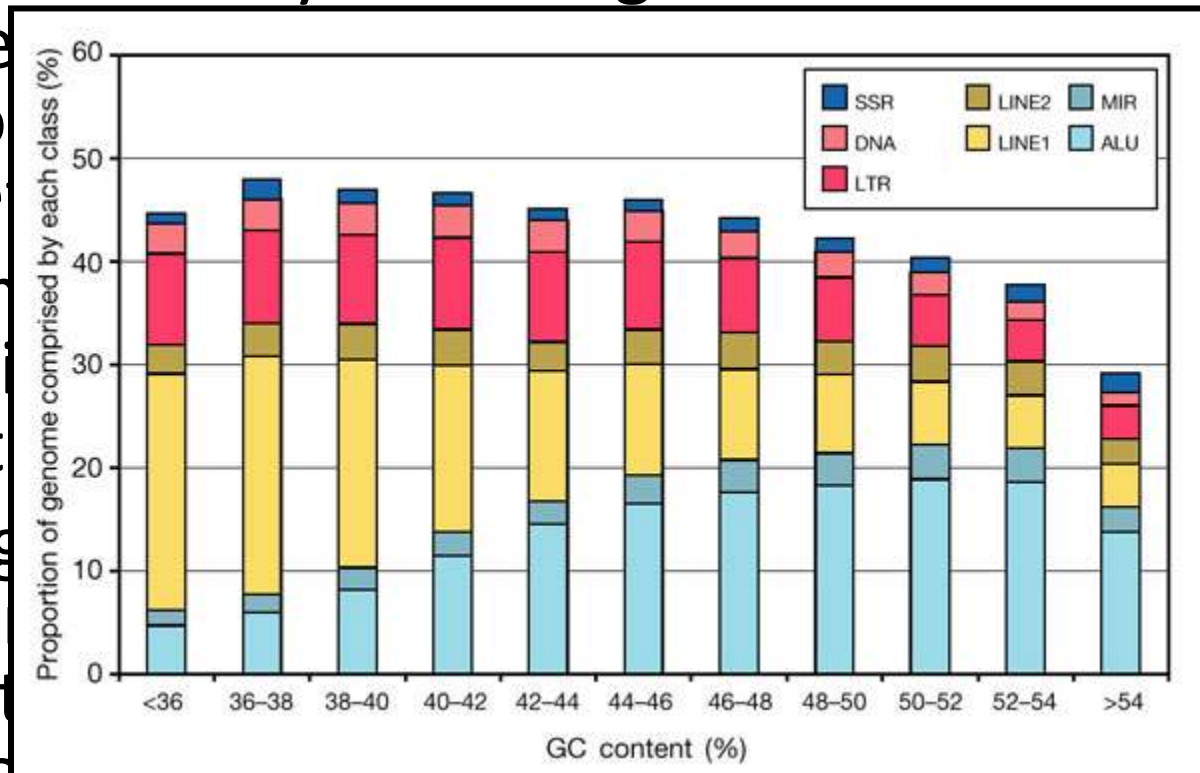
- The pericentromeric regions of human chromosomes contain duplications of genes that are not found in the genome. Segments of the genome that are duplicated in humans but not in yeast, fly or worm.



Some derives from transposons, but there has been a marked decline in the number of transposons in the hominid lineage since we became competent for (LTR) retroposons.

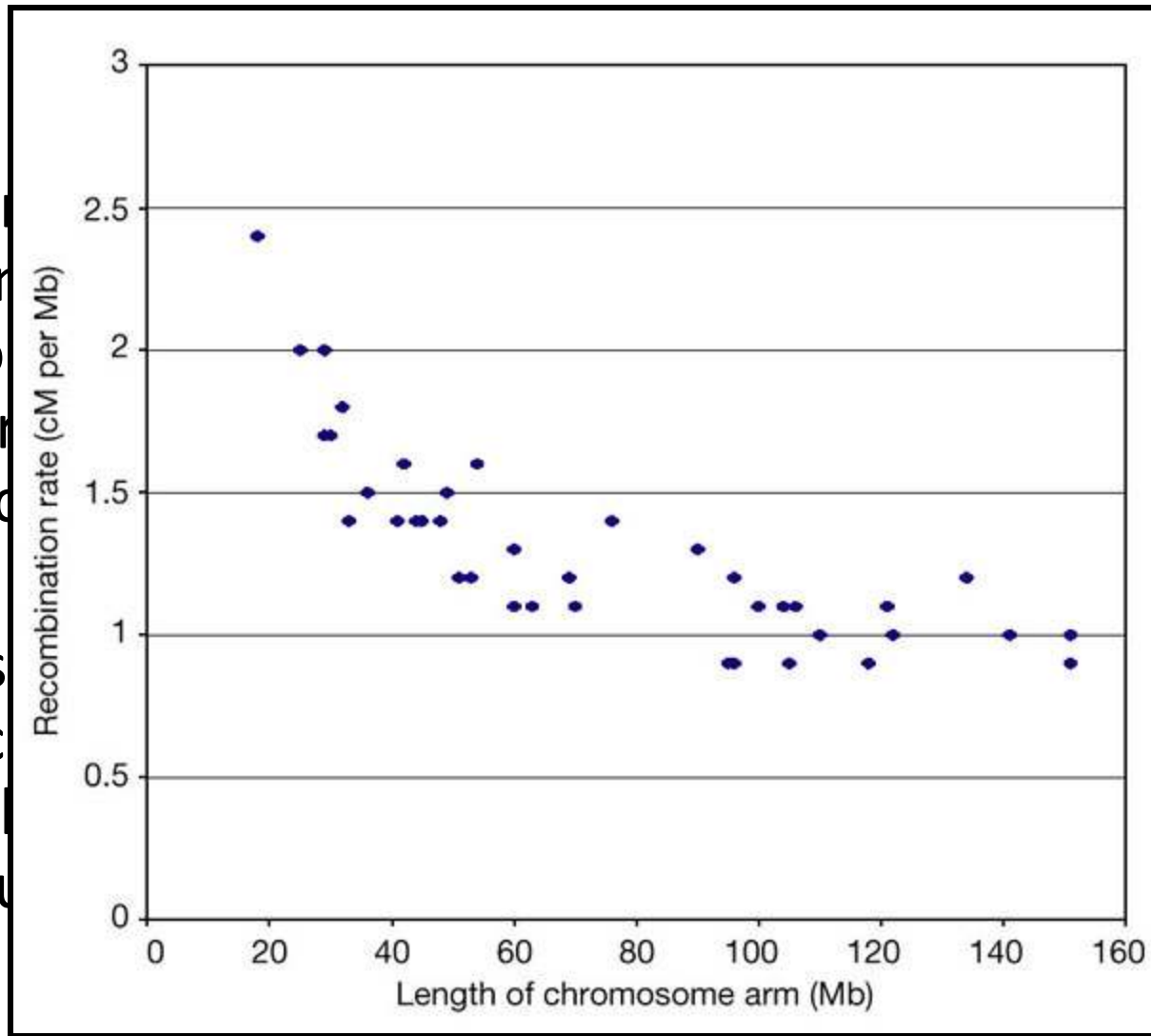
The pericentromeric regions of human chromosomes contain duplications of genes that are not found in the genome. Segments of the genome that are duplicated in humans but not in yeast, fly or worm.

- 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



- prefer C-rich regions may
- The ratio of high to low GC content is twice as high as that
- Cytosine configurations are stable in GC-rich regions, but unstable in AT-rich regions.

- Recombination rate is higher in the proximal region and decreases towards the distal end of the chromosome arm.
- More SNPs are found in the proximal region. This can be explained by the fact that the recombination rate is higher in the proximal region, leading to a higher frequency of recombination events and thus a higher frequency of SNPs.



h distal
osomes
n a
st one
s.
rphisms
ied.
homo-
es in

2004

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*

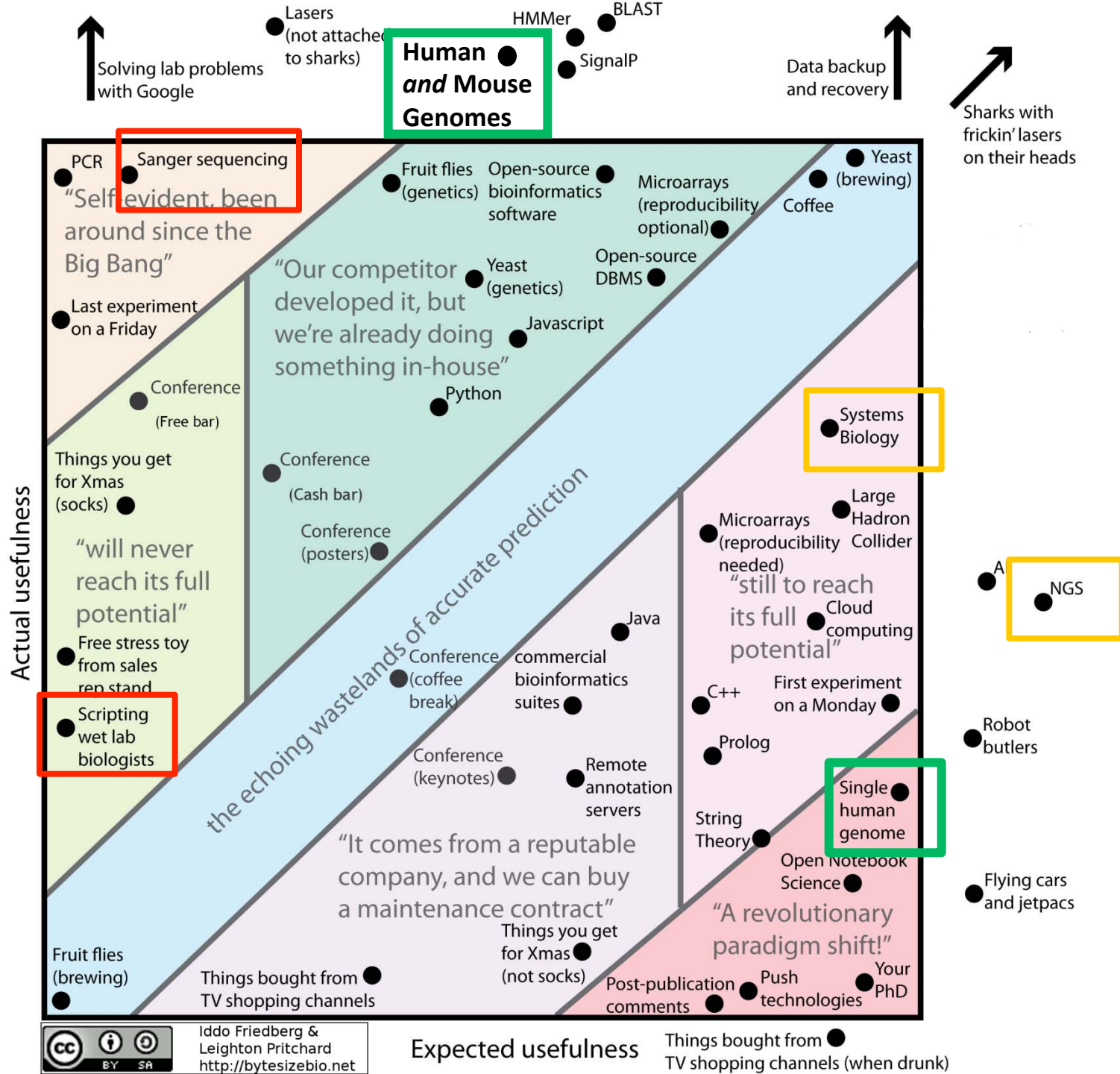
- “The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps.
- It covers 99% of the euchromatic genome and is accurate to an error rate of 1 event per 100,000 bases.
- Notably, the human genome seems to encode only 20,000–25,000 protein-coding genes.”

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?

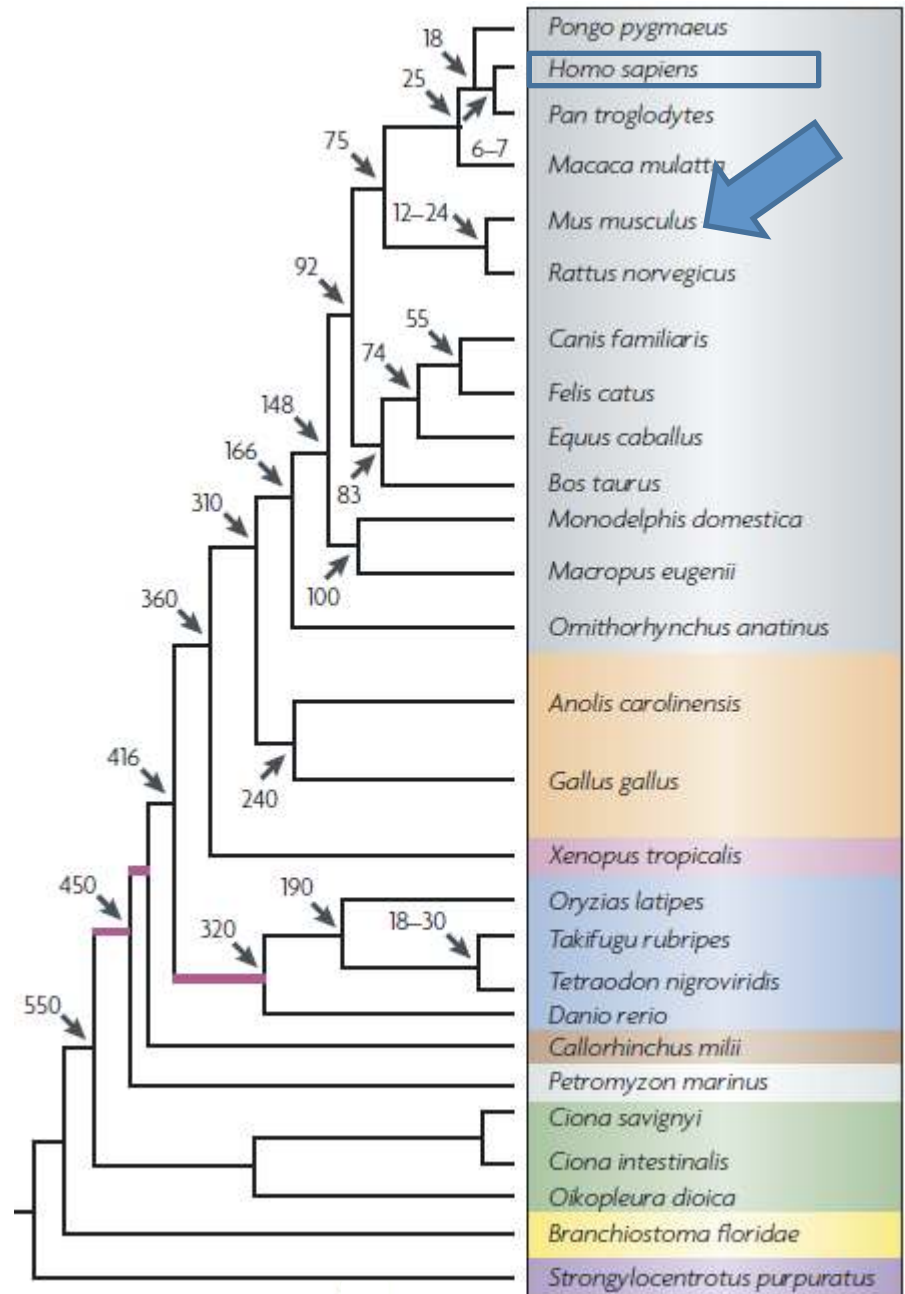
How Useful is a Human Genome?



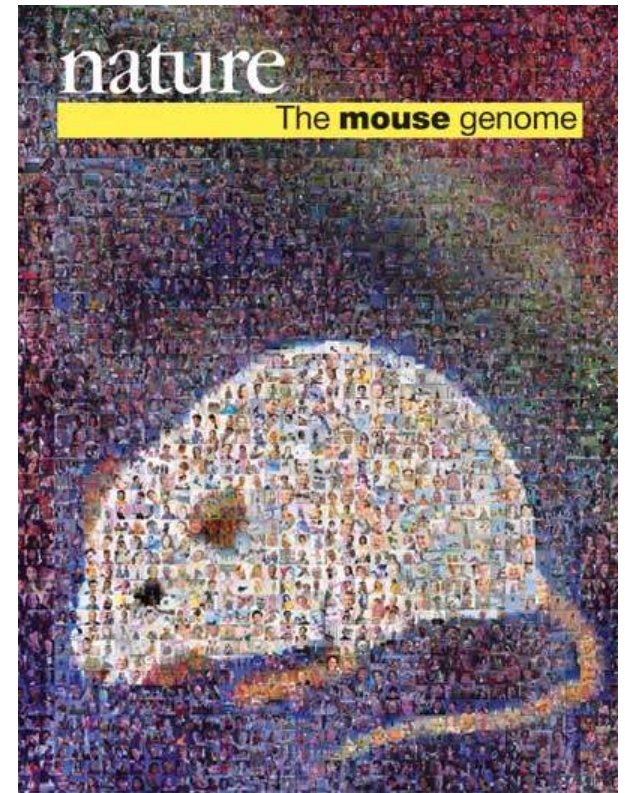
Evolutionary yardsticks

mammalian & invertebrate
noncoding sequences do
not align

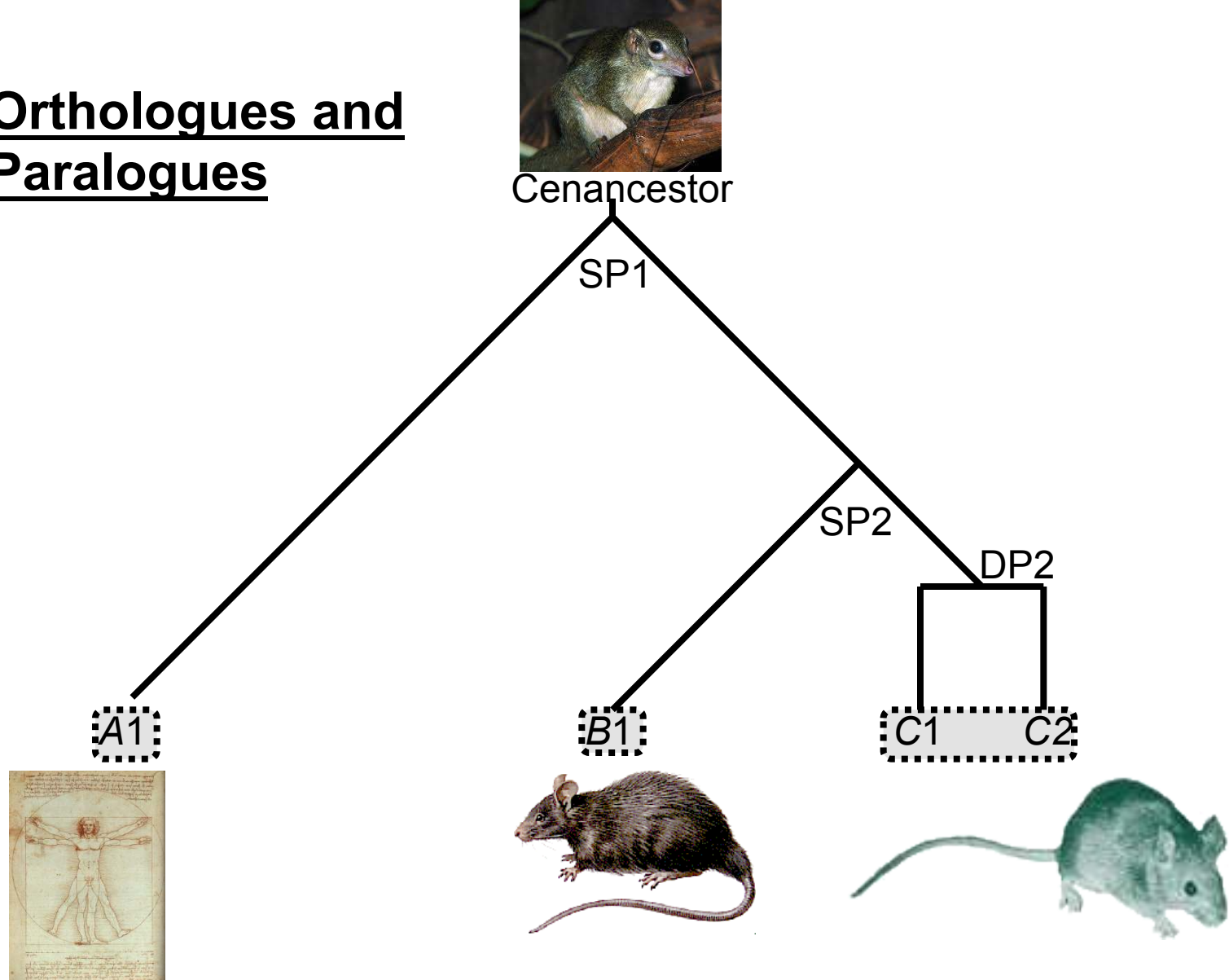
c 2008



Mouse vs Human

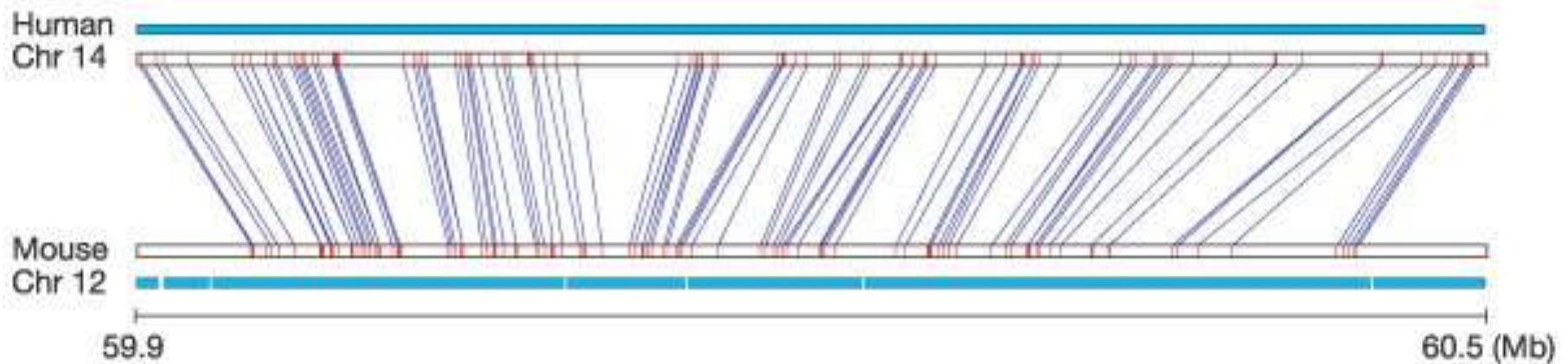


Orthologues and Paralogues



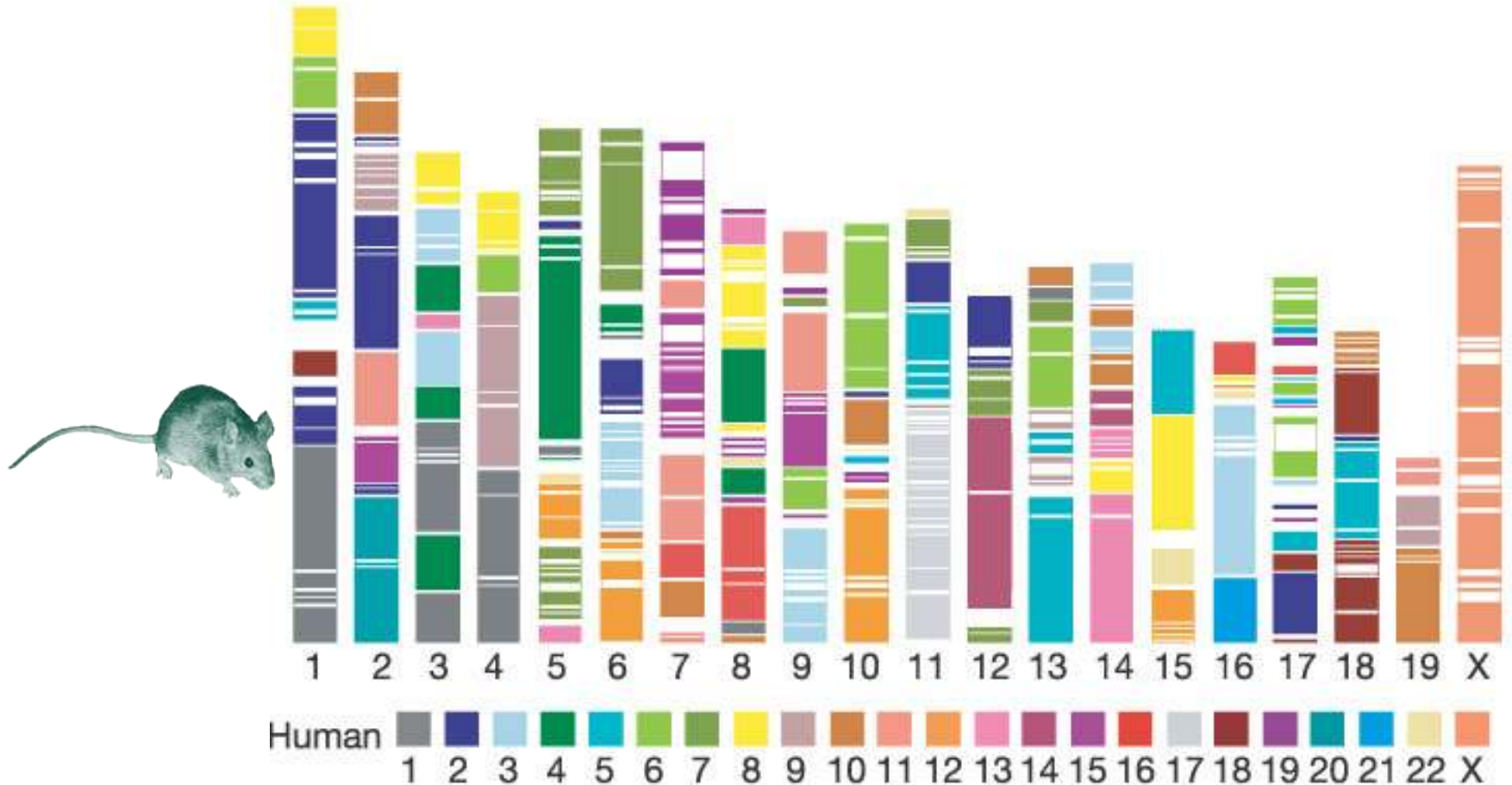
C1 and C2 are paralogues
A1 and B1 and (C1 and C2) are orthologues

Human and mouse “local synteny”

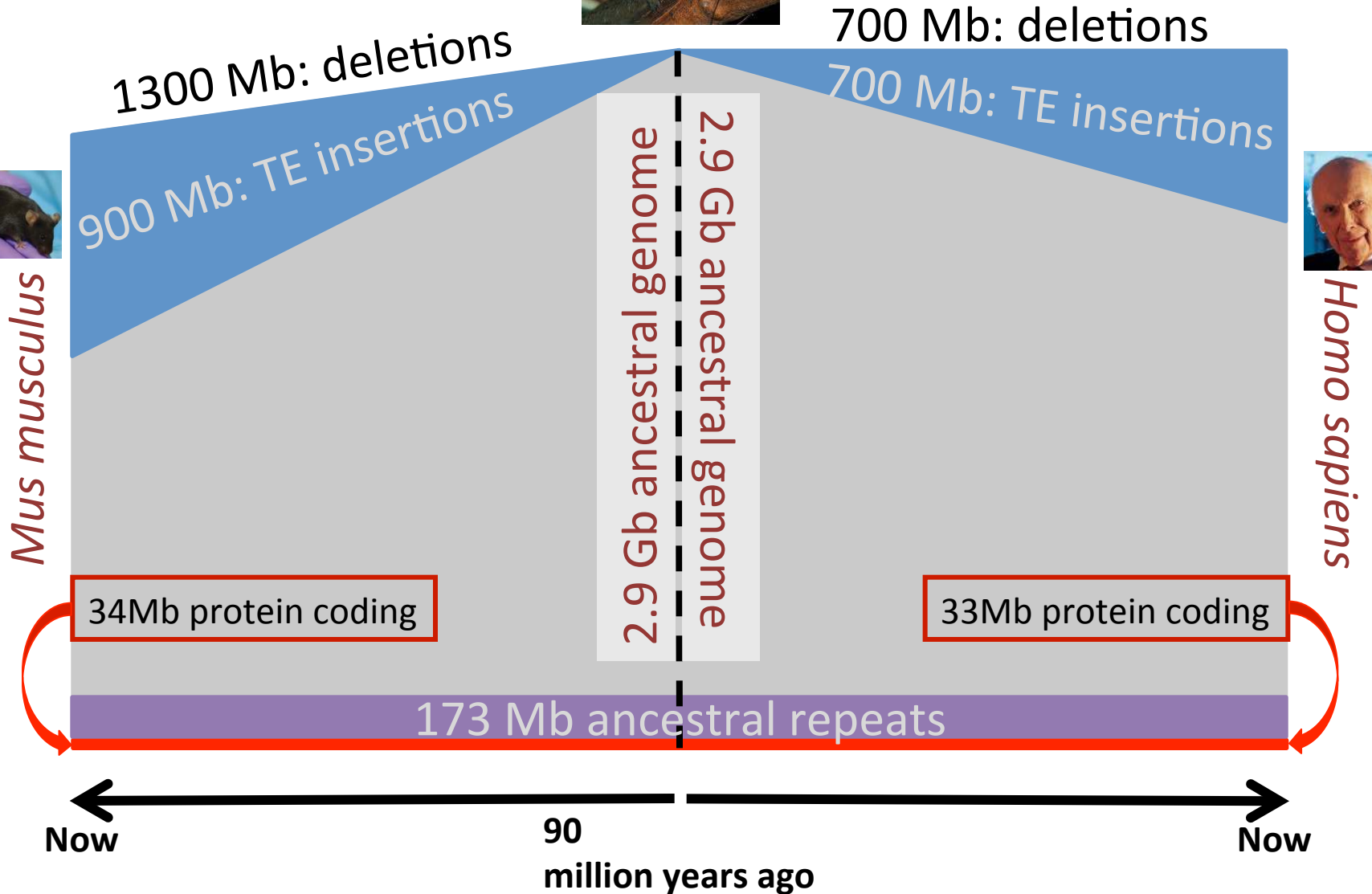
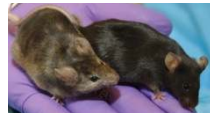


“Syntenic” regions contain orthologues!

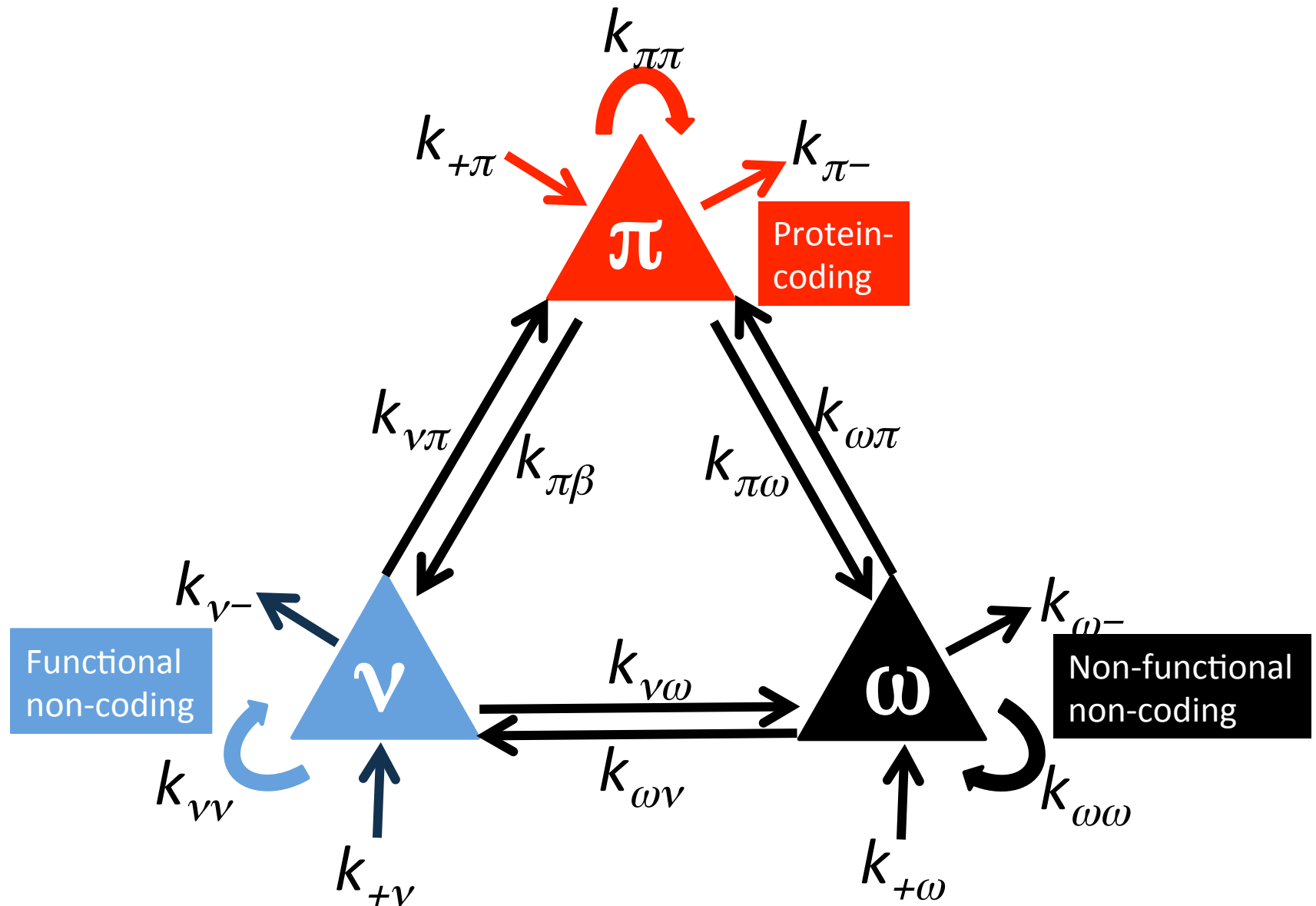
Human and mouse chromosomes: global orthology



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



Three-state model of a mammalian genome

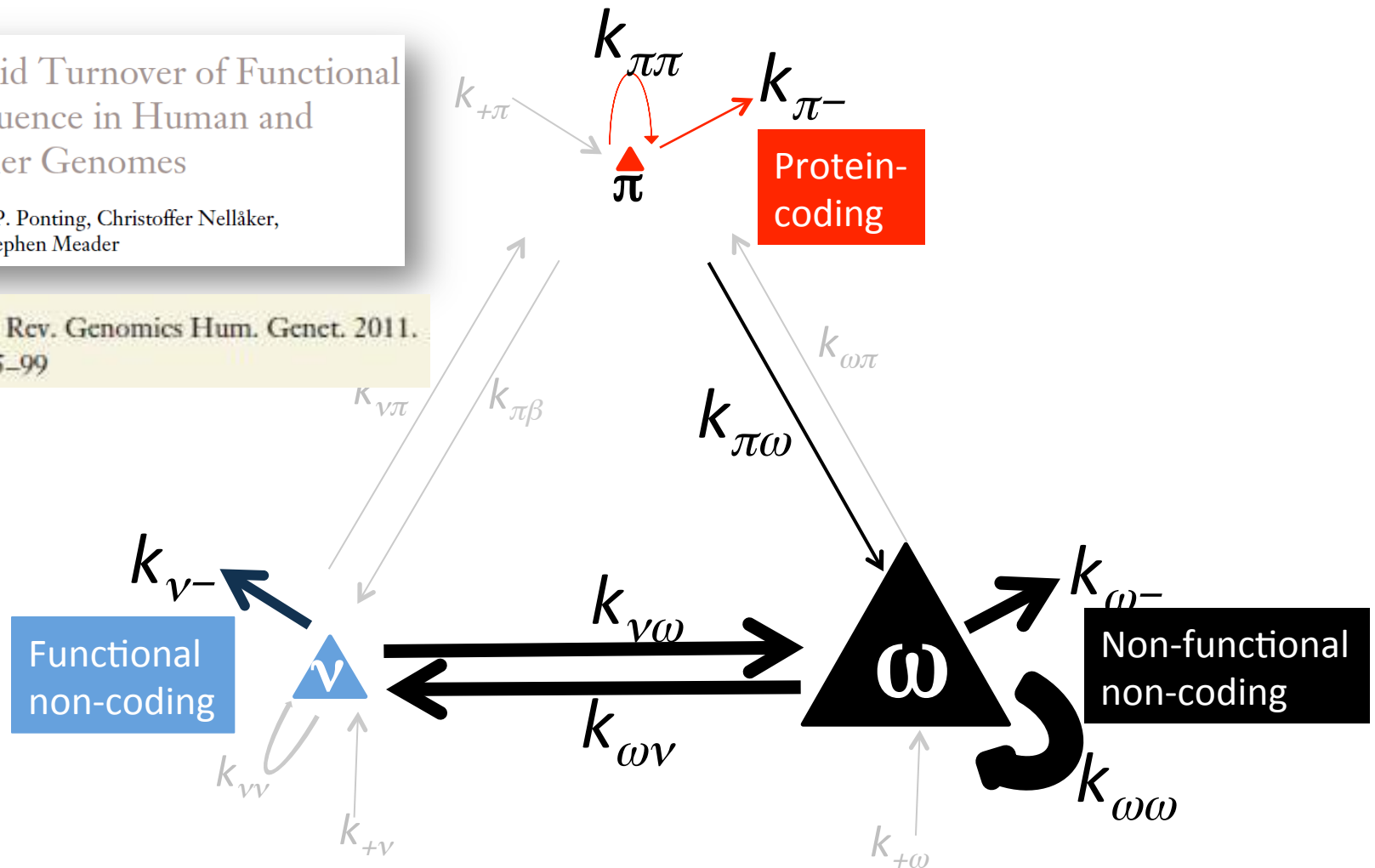


The dynamics of a mammalian genome are dominated by transposable elements

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.

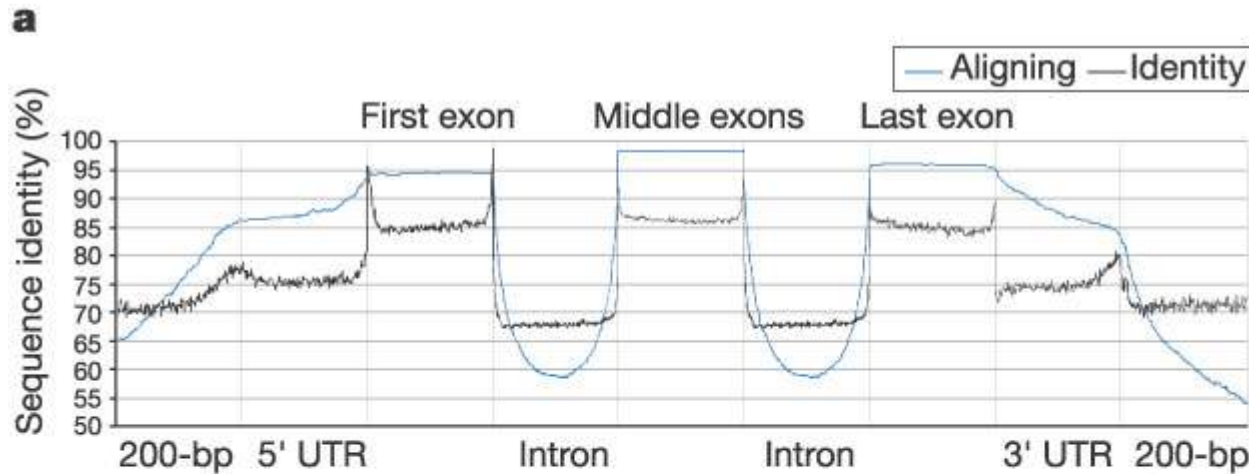
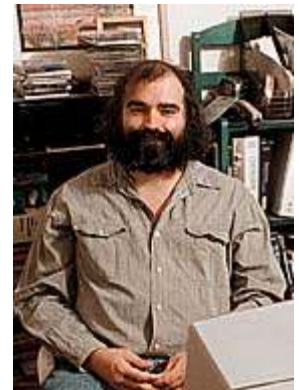
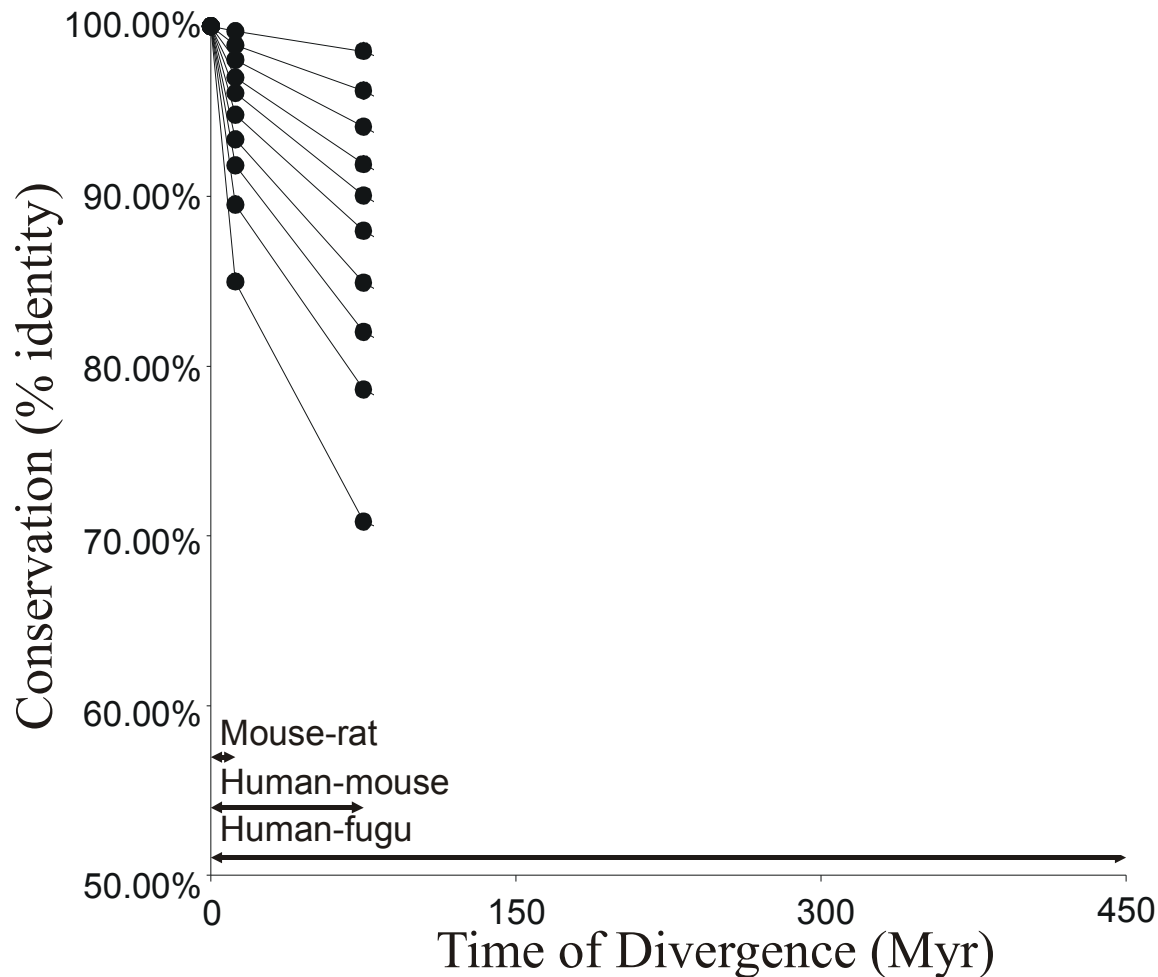


Figure 25. Sequence conservation between mouse and human genes
Mouse genome paper *Nature* 420, 520-562



Gene Sequence Conservation is Clock-like



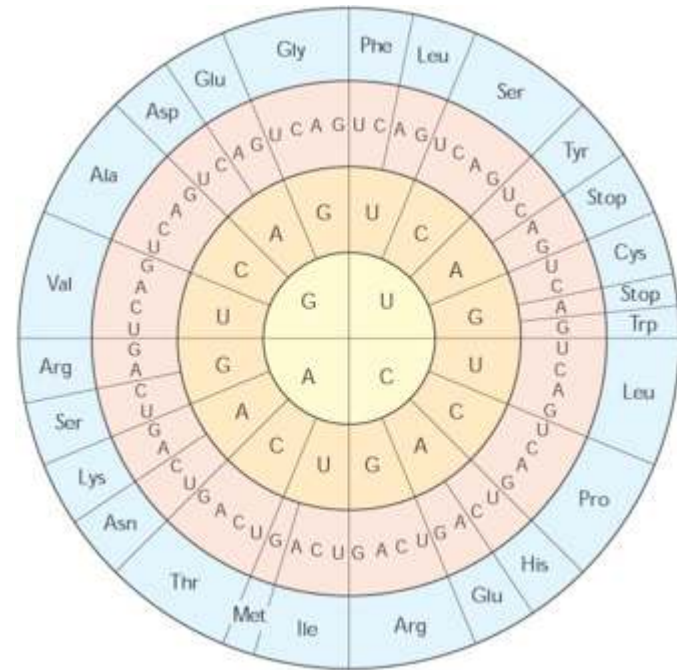
A model of neutral evolution

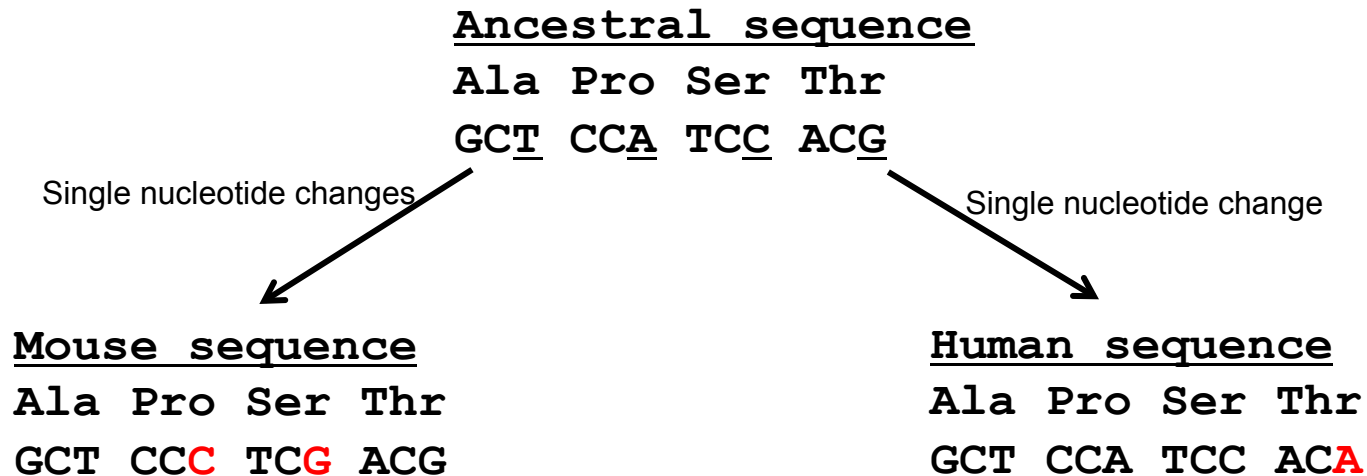
- d_s – the number of synonymous substitutions per synonymous site
- takes advantage of the redundant genetic code
- 4D sites GCx (ALA), CCx (PRO), TCx (SER),
 ACx (THR), CGx (ARG), GGx (GLY),
 CTx (LEU), GTx (VAL)
- “how much would a gene’s sequence have changed if selection had not acted upon it?”

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





3 nucleotide substitutions at 4 synonymous sites
 $d_s = \frac{3}{4}$ or 0.75

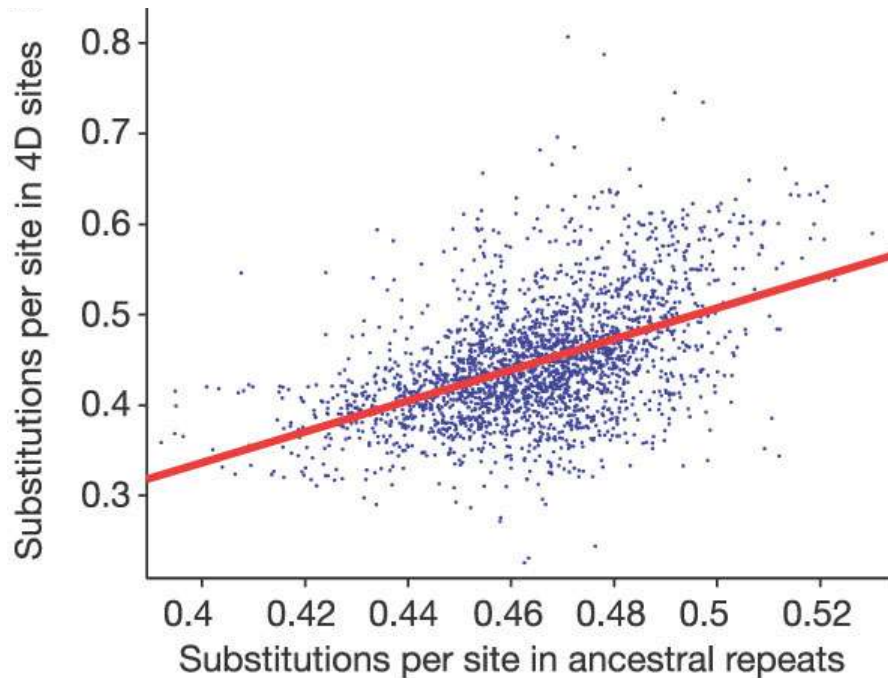
“Ancestral repeats”

- 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



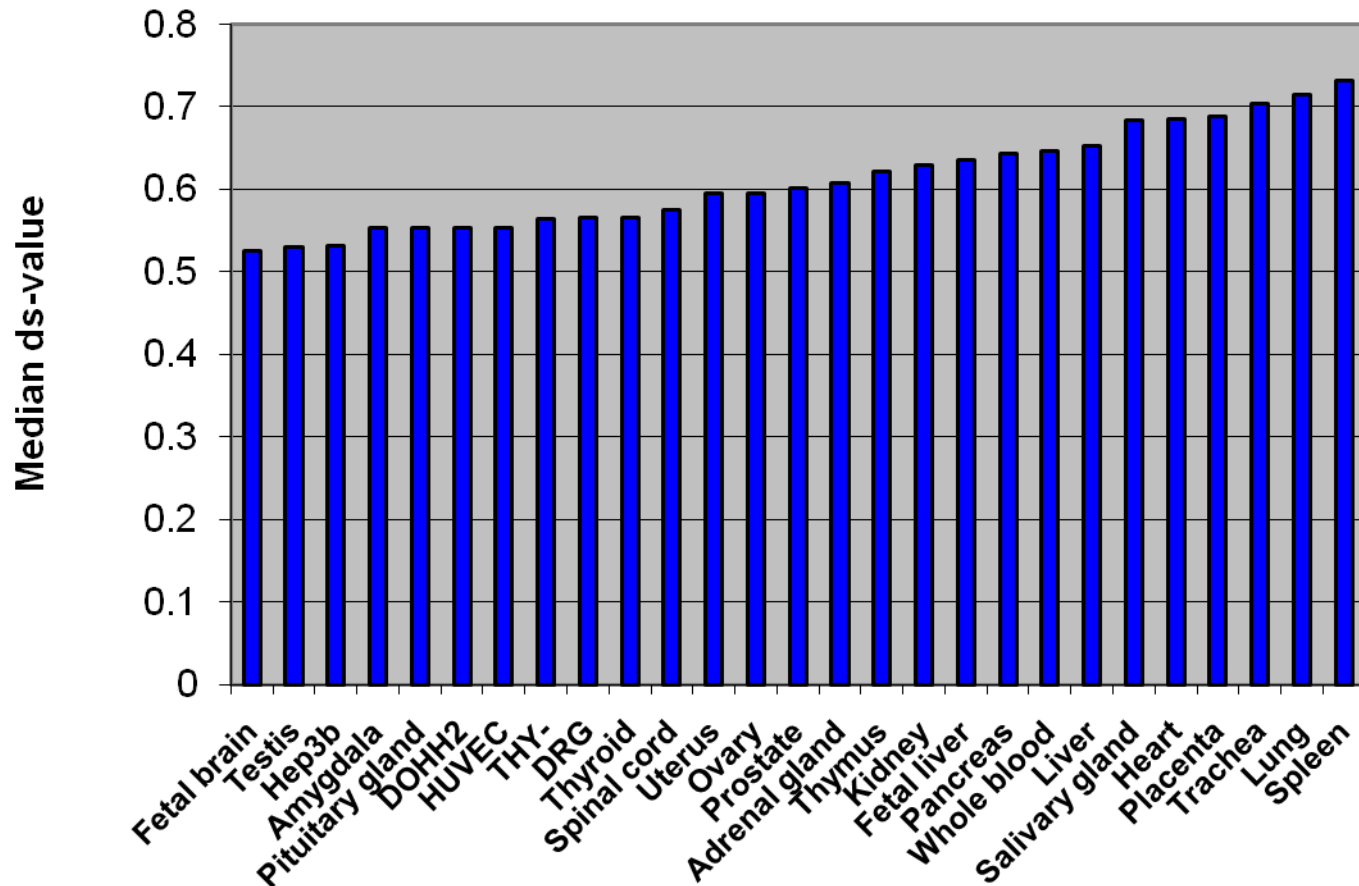
see also
Hardison et al.
Genome Res. 2003
13: 13-26.



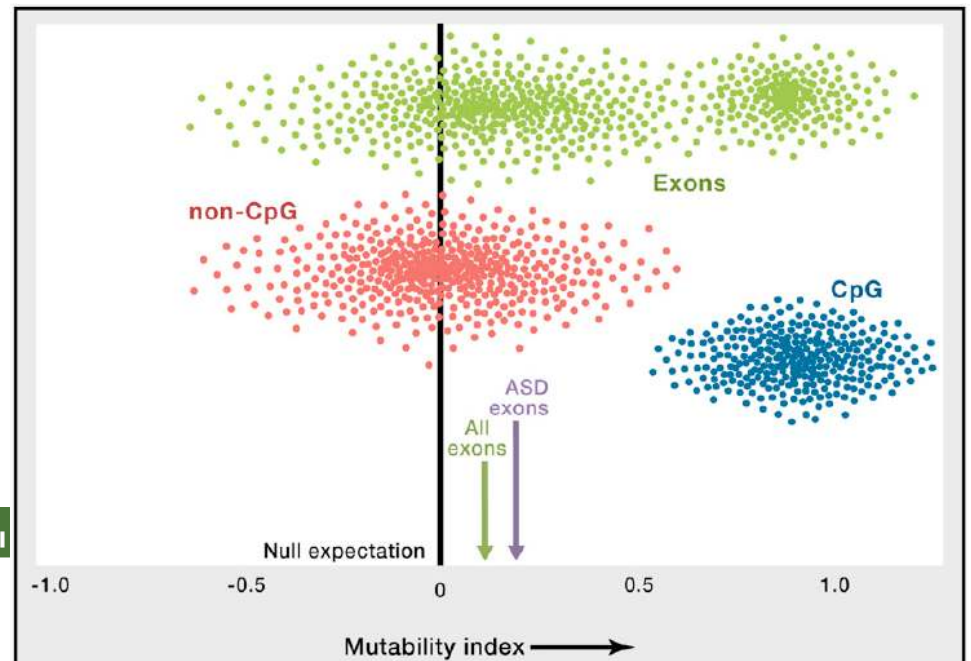
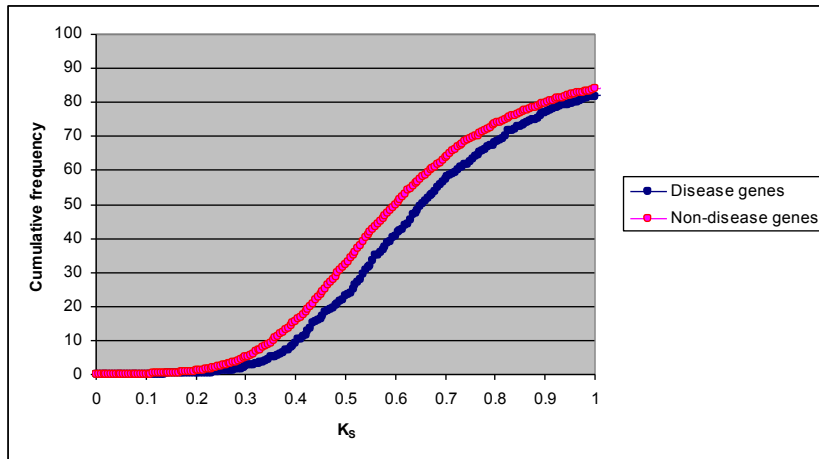
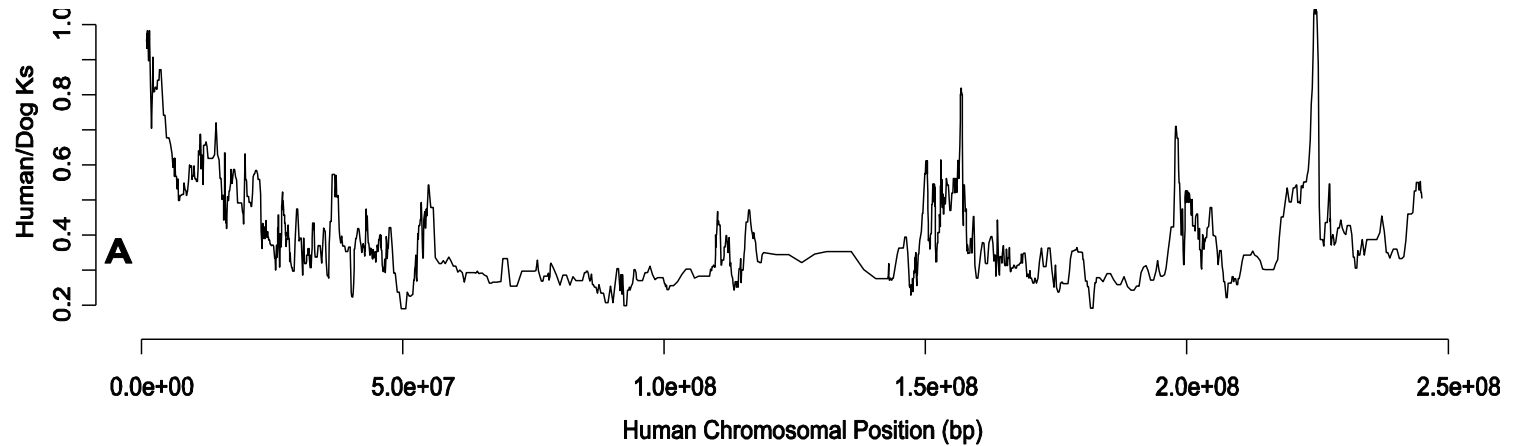
Variation in rates of mutation and/or rates of repair?

- 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



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?
- d_N/d_S , 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

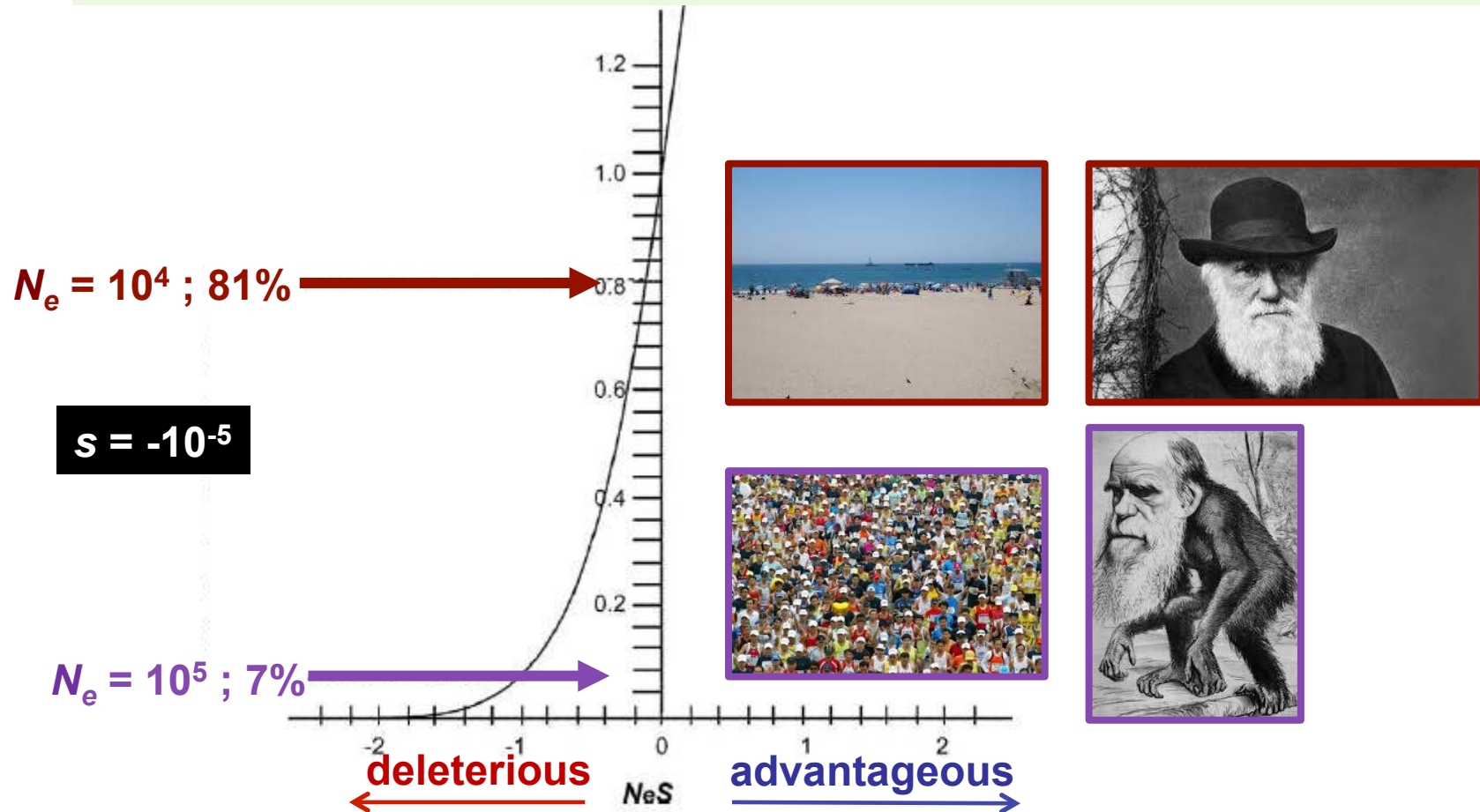


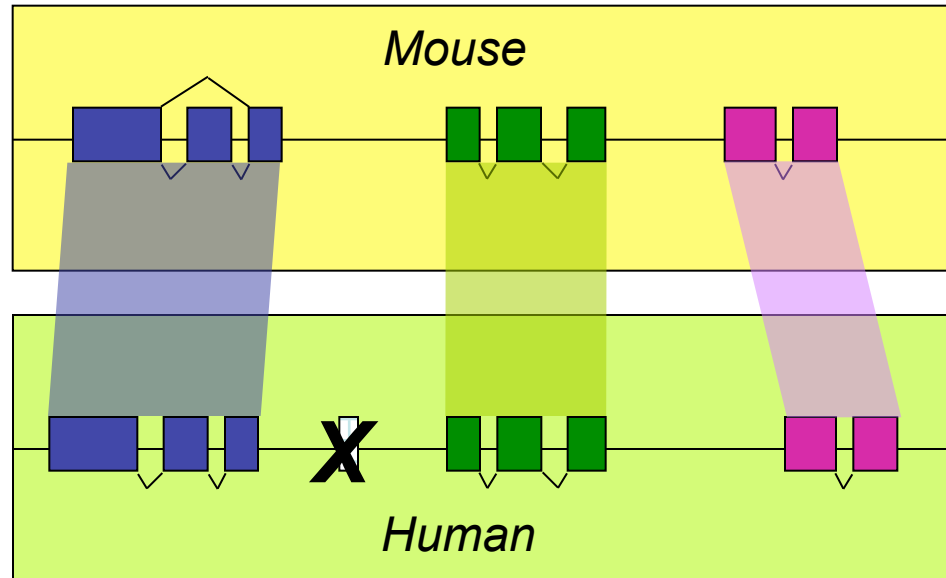
Fig. 1. The probability of fixation of a new variant with respect to the neutral expectation of $1/(2N)$ graphed as a function of the product of the effective population size and the selection coefficient of the variant ($N_e \times s$). The dashed lines represent cases in which $s = -10^{-5}$ but N_e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

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 substitution; 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. Ponting¹ 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 the human genome itself. But what? The laboratory mouse is hailed

timeline

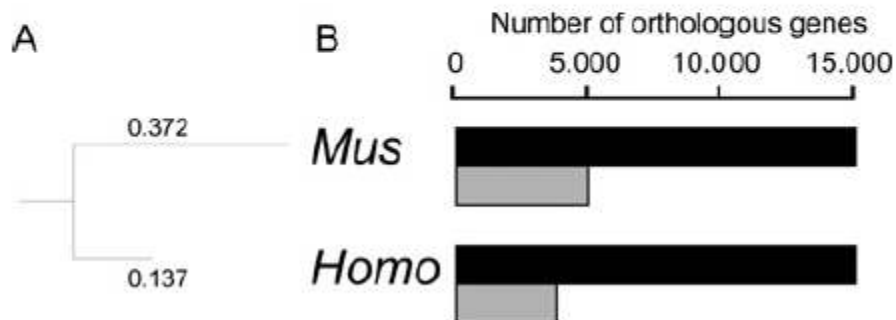
510 The mouse genome

commentary

512 Mining the mouse genome

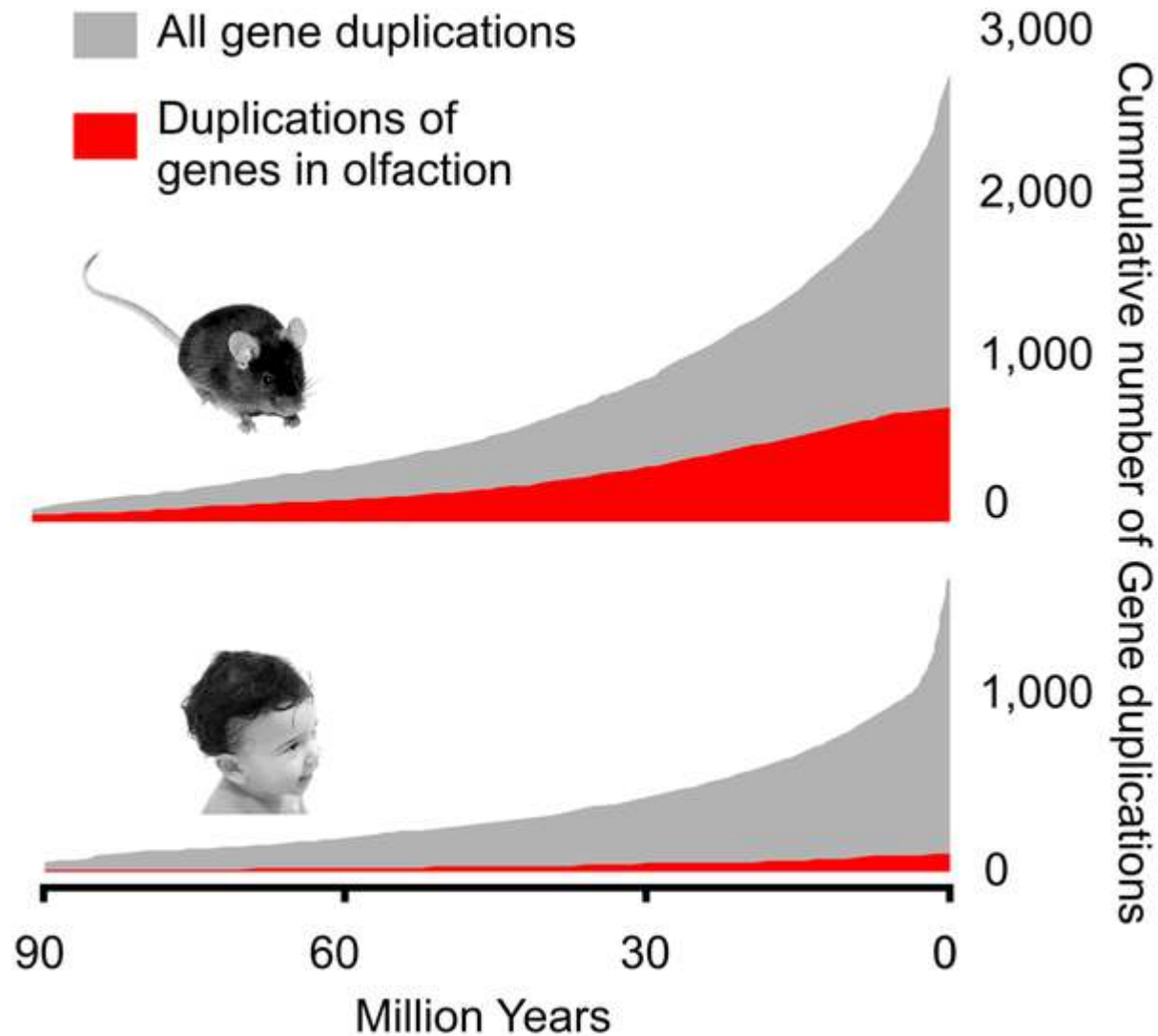
Figure 2 | Mouse genes have a higher synonymous nucleotide substitution rate (d_s) and have accumulated more lineage-specific duplicates than human genes

(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).

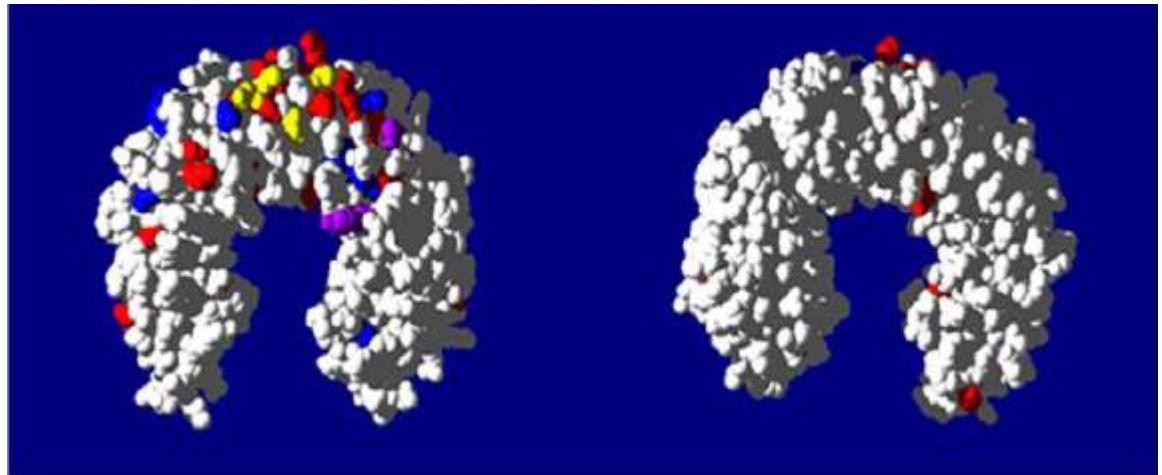
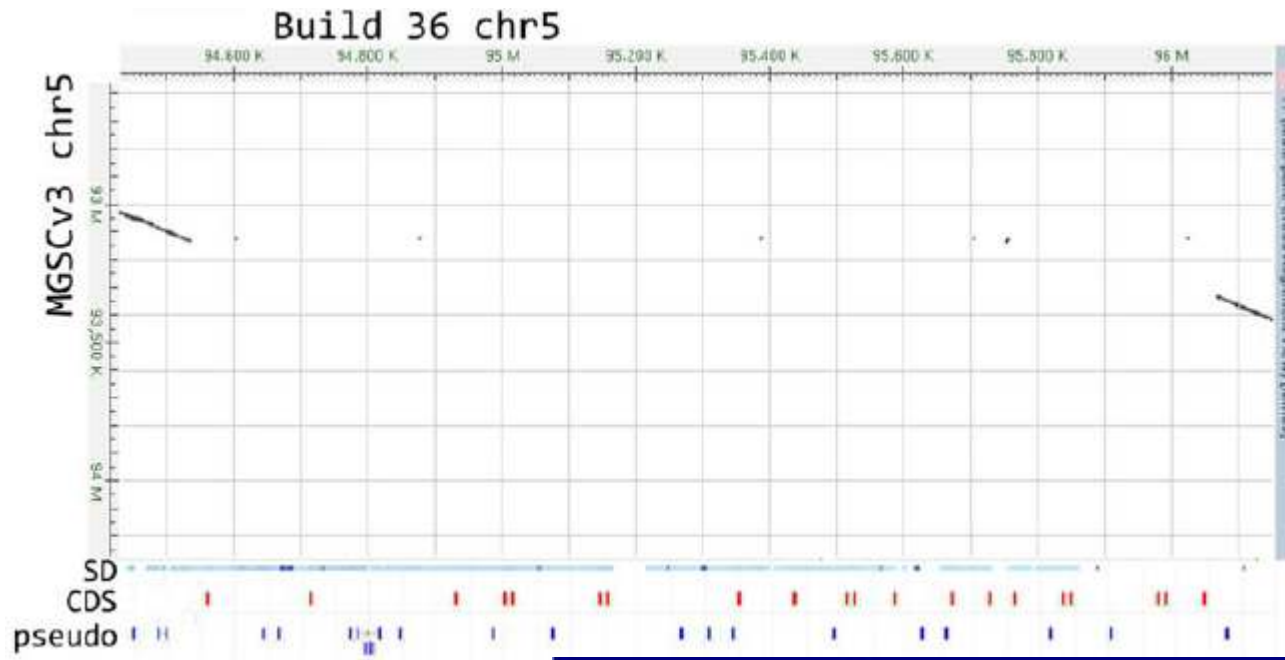


Ritu Dhand Chief Biology Editor

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

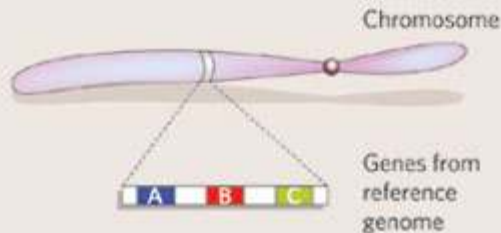
NCBI Builds 36.1 and 36 respectively.

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)

VARIATIONS IN OUR GENOMES



Deletion



Insertion



Inversion



Copy-number variant

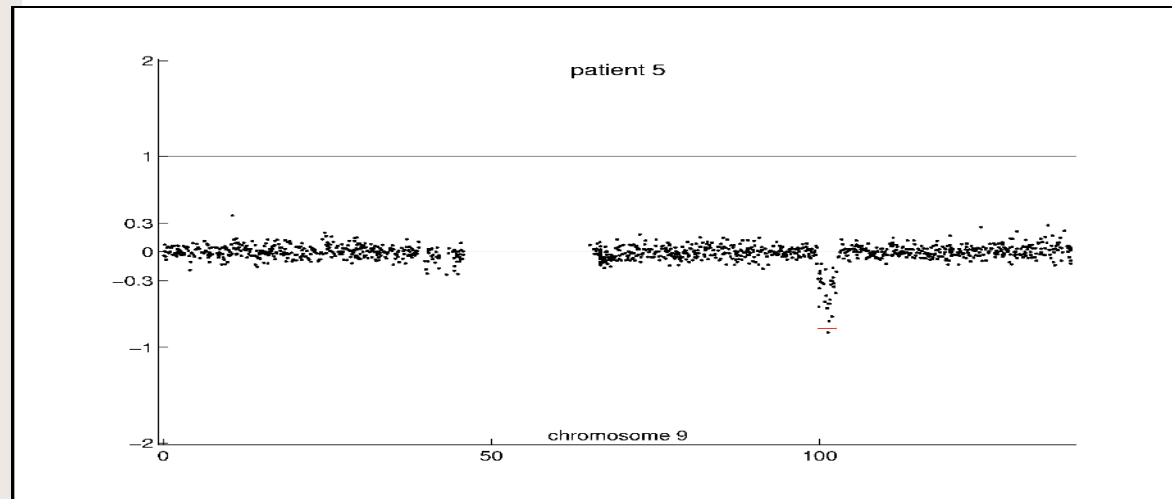


Segmental duplication



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

Segmental duplications and CNVs often coincide.



chr9 (q22.33-q31.1)



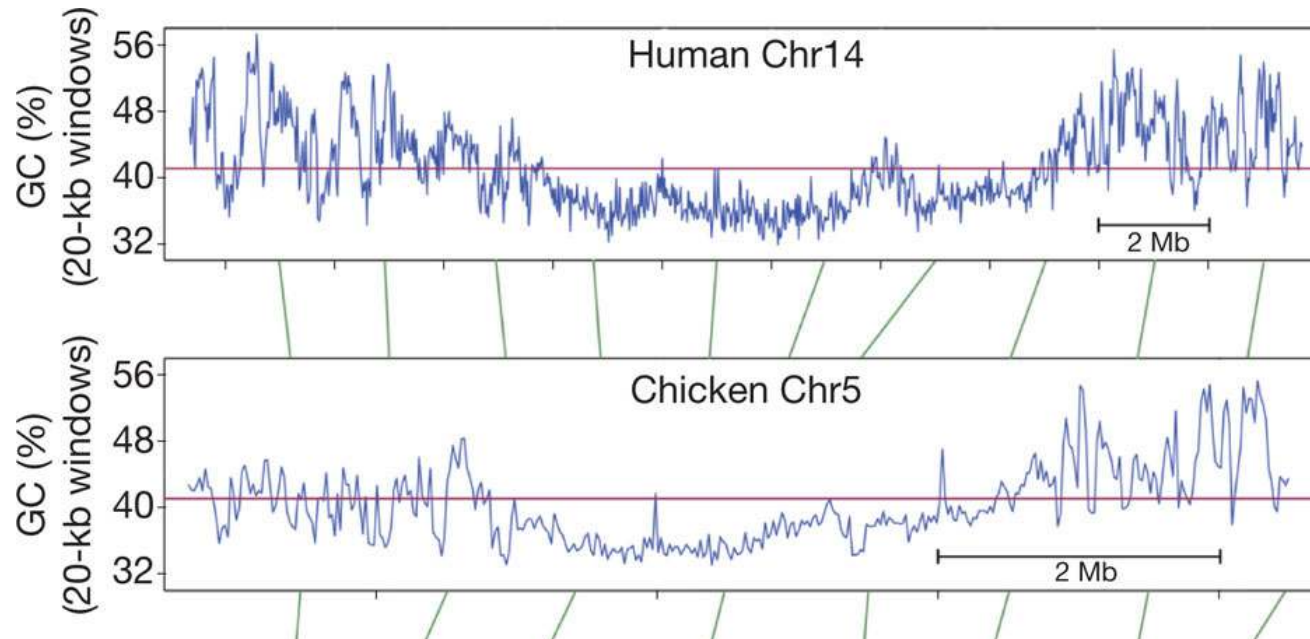
Immunity, defence, chemosensation genes

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

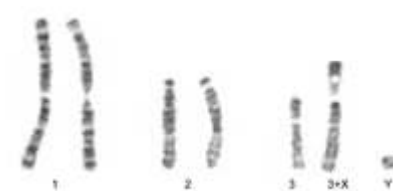
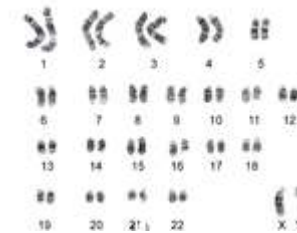
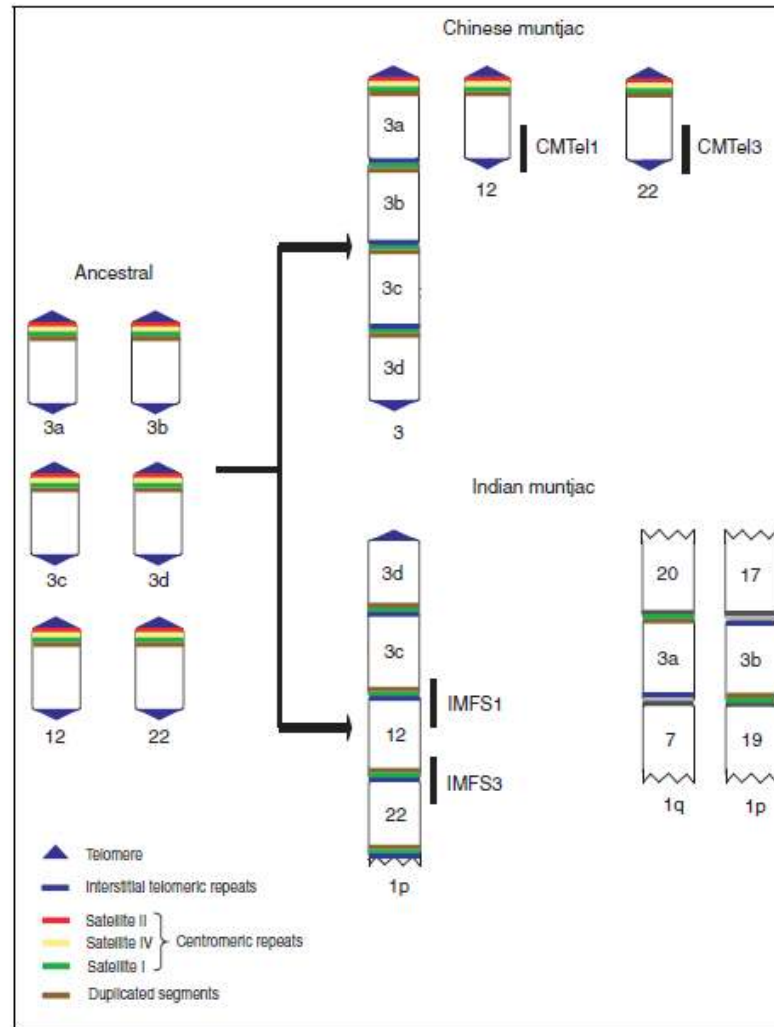
Little evidence that common CNVs are either adaptive or are associated with disease.

Evolutionary questions without
adequate answers

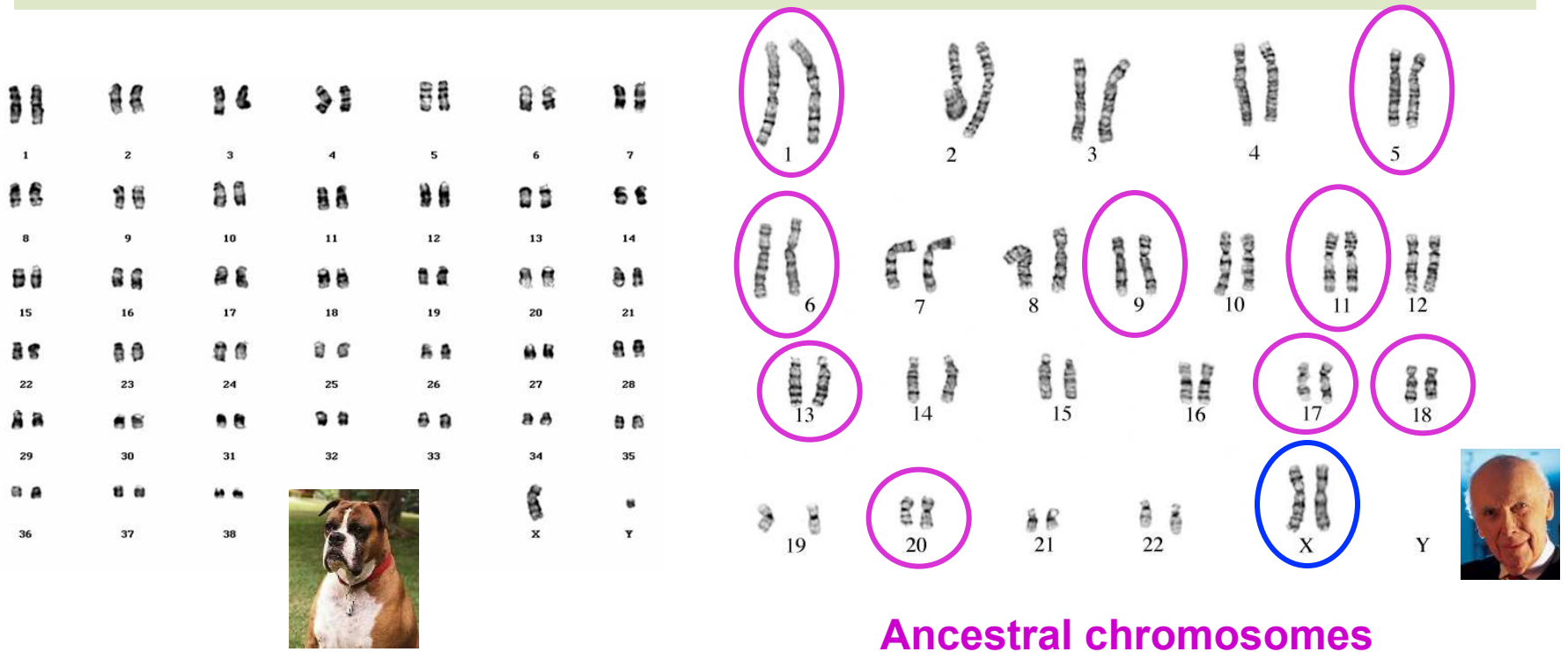
(Non-)Conservation of Isochores – Why?



Why rapid variations in karyotypes?



Mammalian Karyotypes

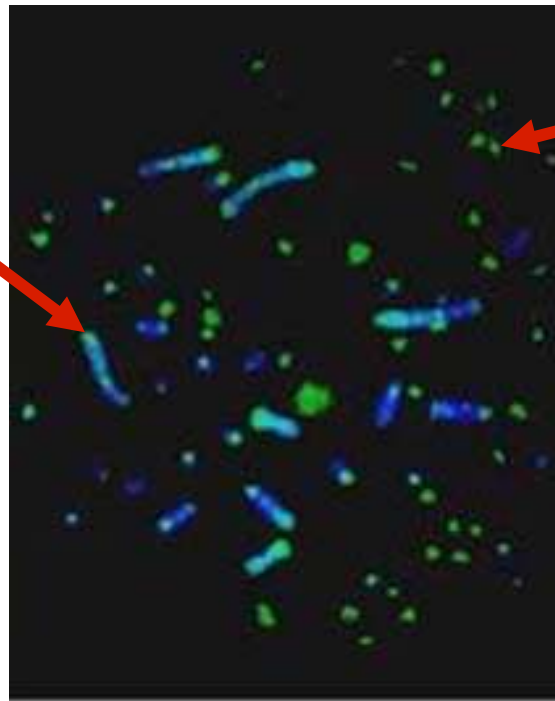


Why do birds & lizards have small (micro) chromosomes?

Large (macro-) chromosomes:

more DNA but lower gene density;

lower mutation rate; lower G+C

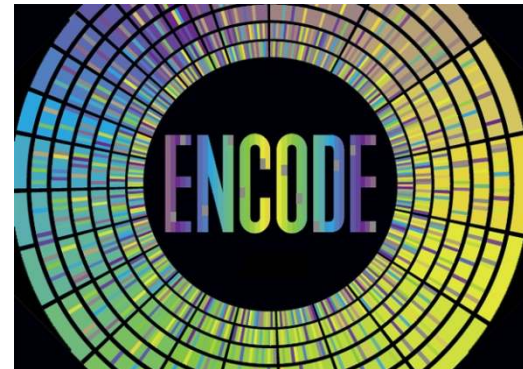
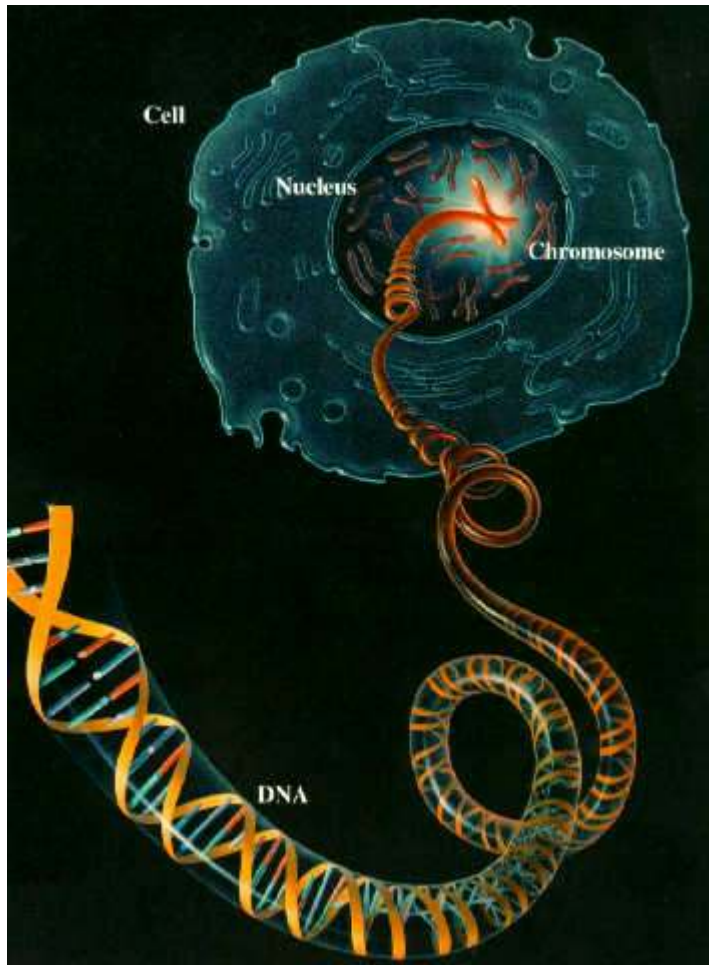


Small (micro-) chromosomes:

25% of the DNA but half the genes;

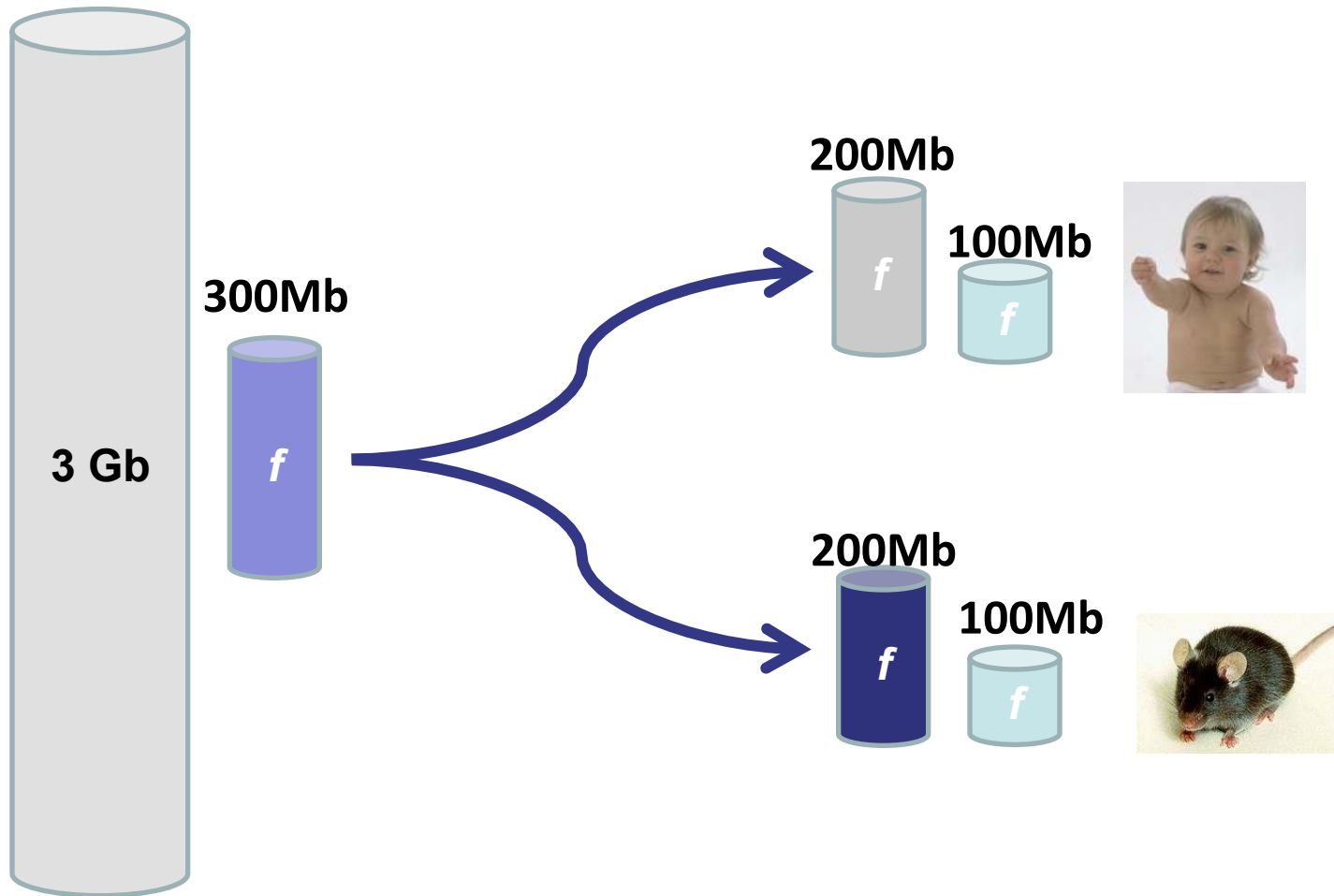
higher mutation rate; higher G+C

Part 2: functional DNA & transcript maps



How much of our genome is biologically functional?

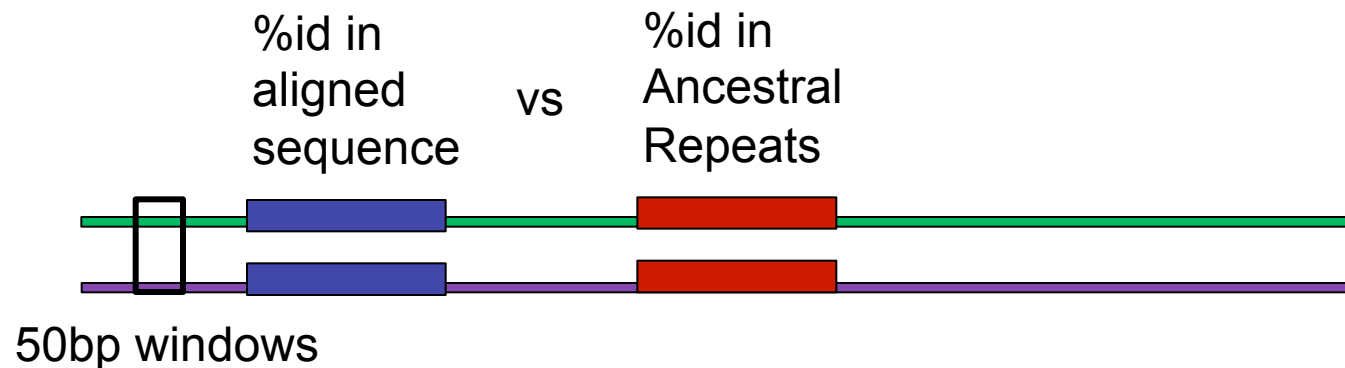
My view: 10% ...



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%



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

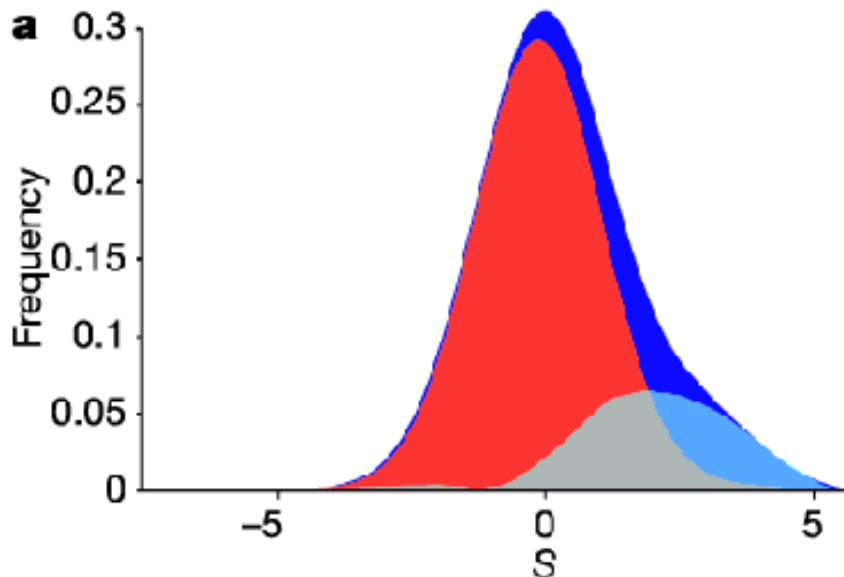
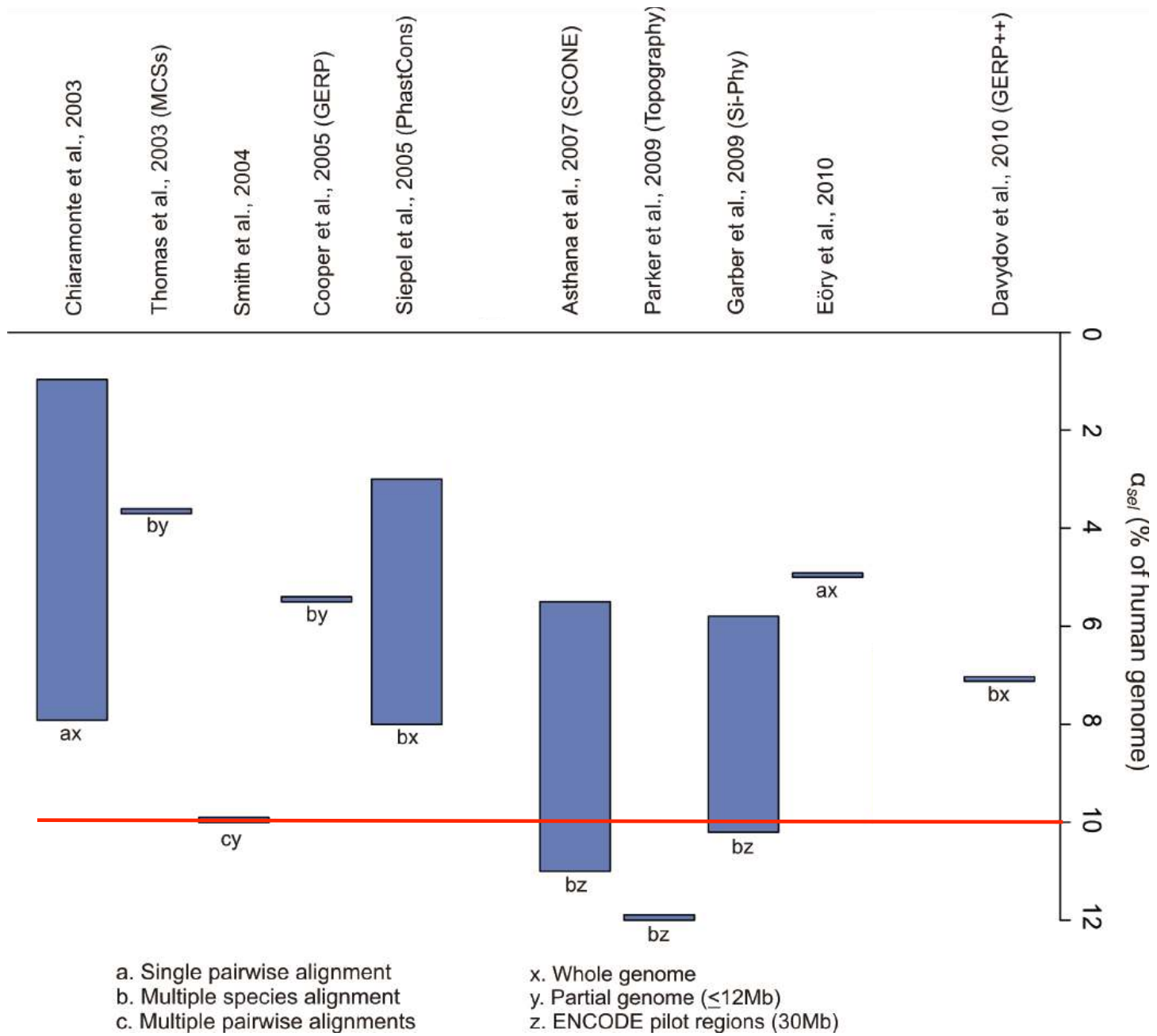


Table 1. Estimates of the Share of the Human Genome under Selection for Different Window Sizes (W) and Required Number of Aligned Bases (T)

W	T	$p_I = (1 - p_O)$	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

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

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%



Raising the estimate of functional human sequences

Michael Pheasant and John S. Mattick¹

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](https://doi.org/10.1101/gr.6406307)

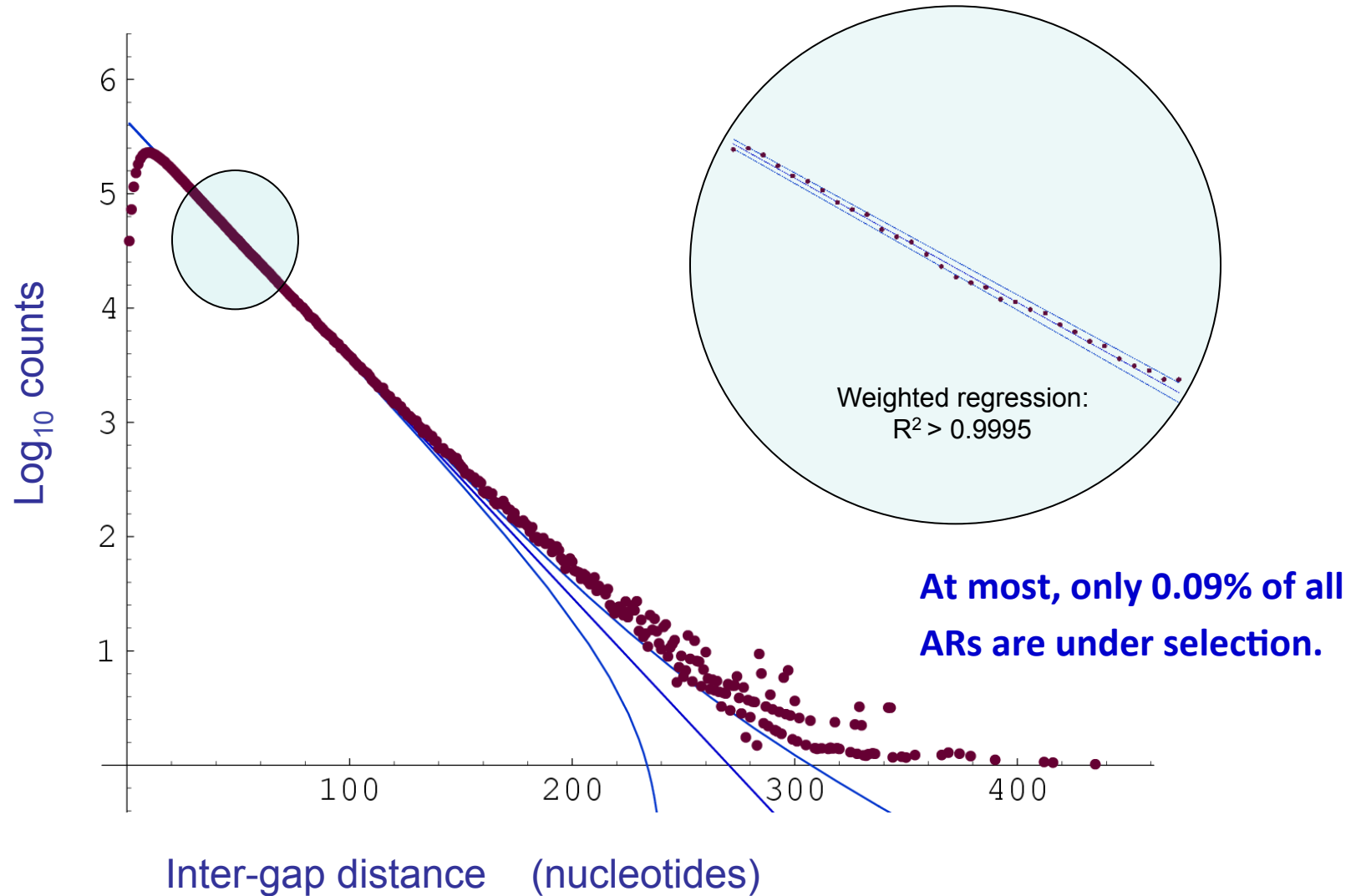
Gerton Lunter's indel model: Insertions/Deletions

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



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

Inter-gap distances within ancestral repeats



Raising the estimate of functional human sequences

Michael Pheasant and John S. Mattick¹

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

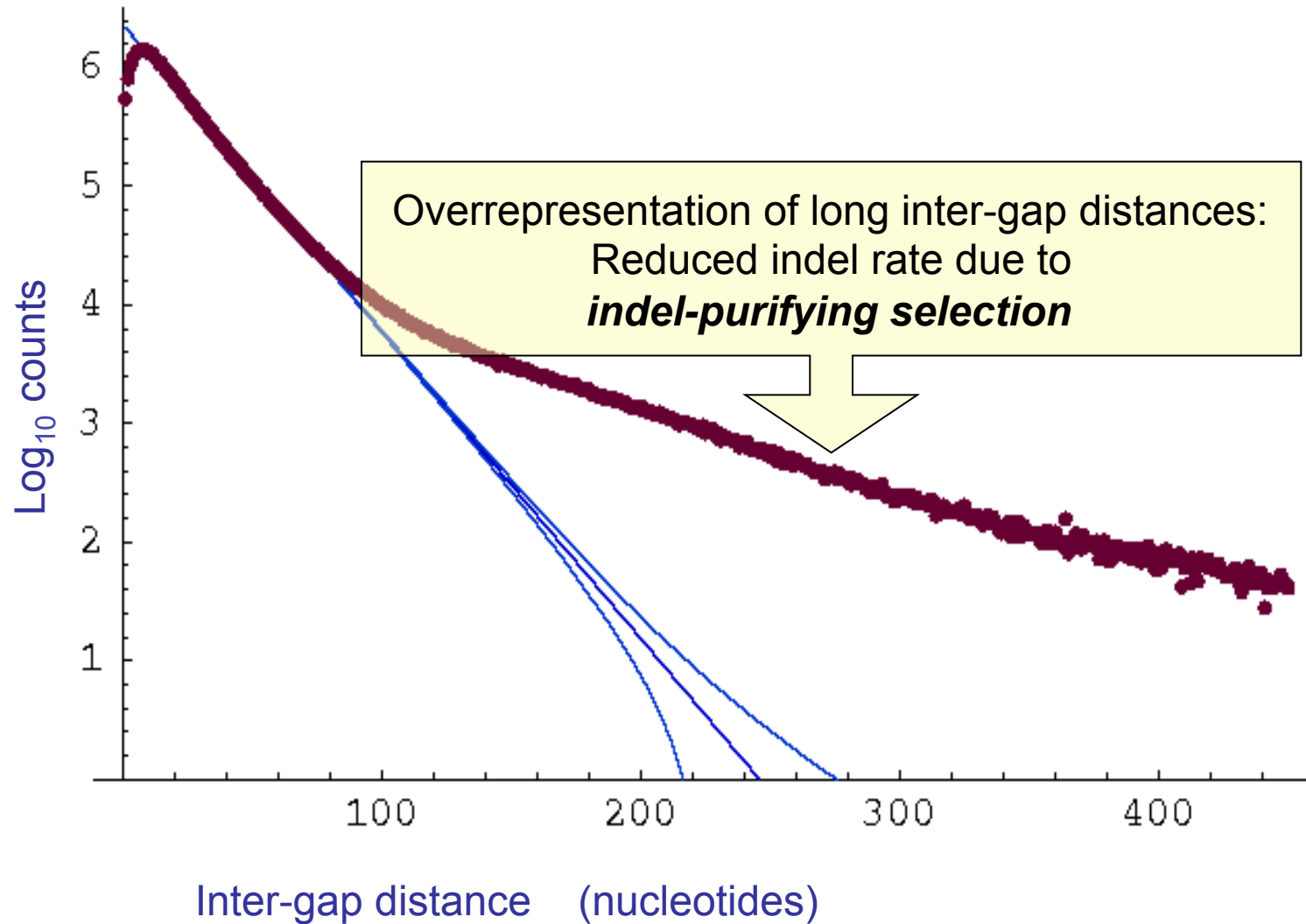
While less than 1.5% of the mammalian genome encodes proteins, it is now evident that the vast majority is transcribed, mainly into non-protein-coding RNAs. This raises the question of what fraction of the genome is functional, i.e., composed of sequences that yield functional products, are required 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

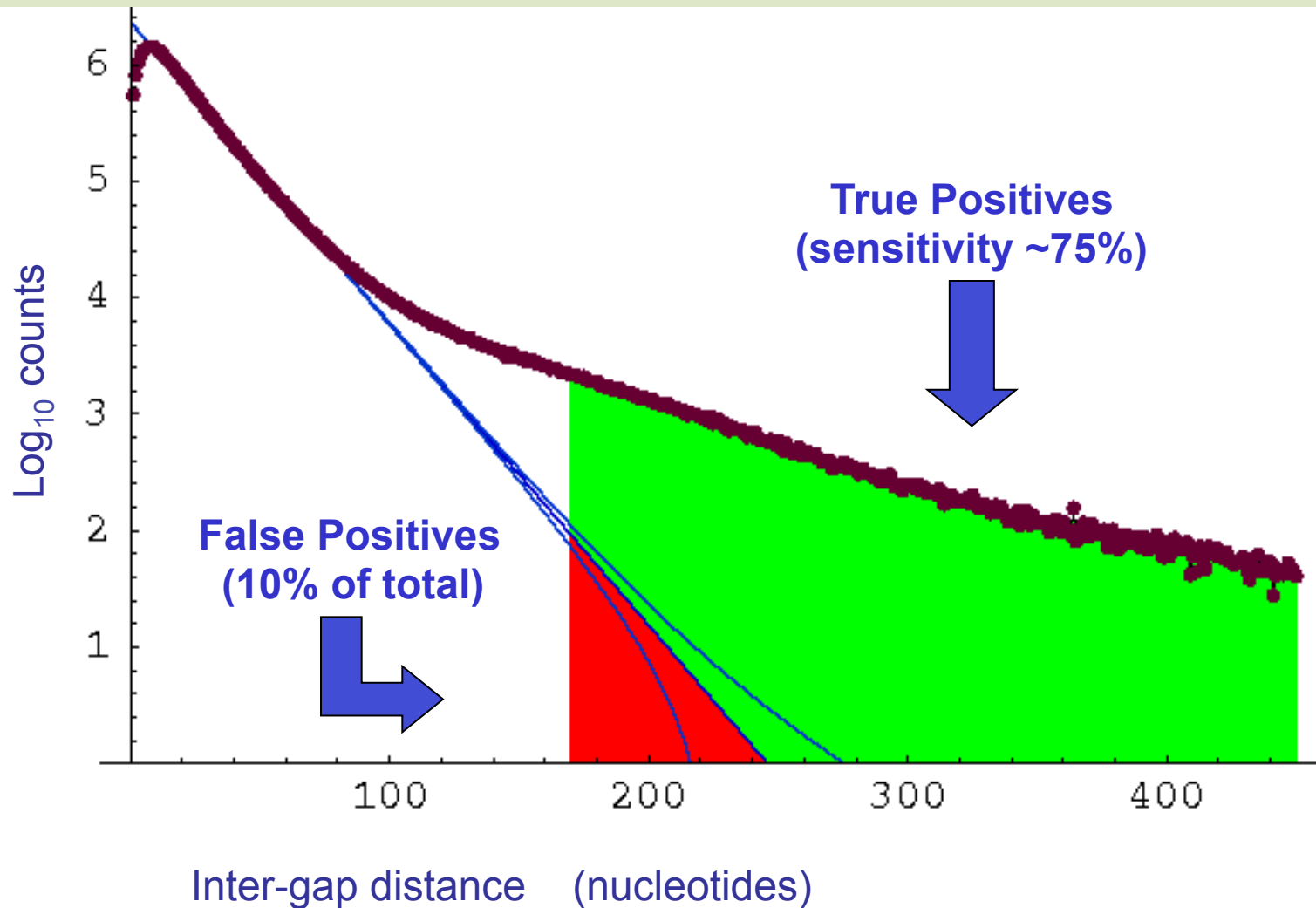
Michael Pheasant and John S. Mattick

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

Inter-gap distances: whole genome



Identifying sequence under indel purifying selection

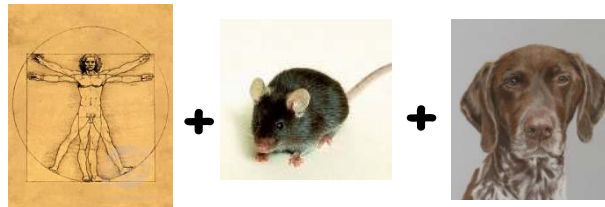


Fraction of conserved DNA

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.



Raising the estimate of functional human sequences

Michael Pheasant and John S. Mattick¹

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](https://doi.org/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,

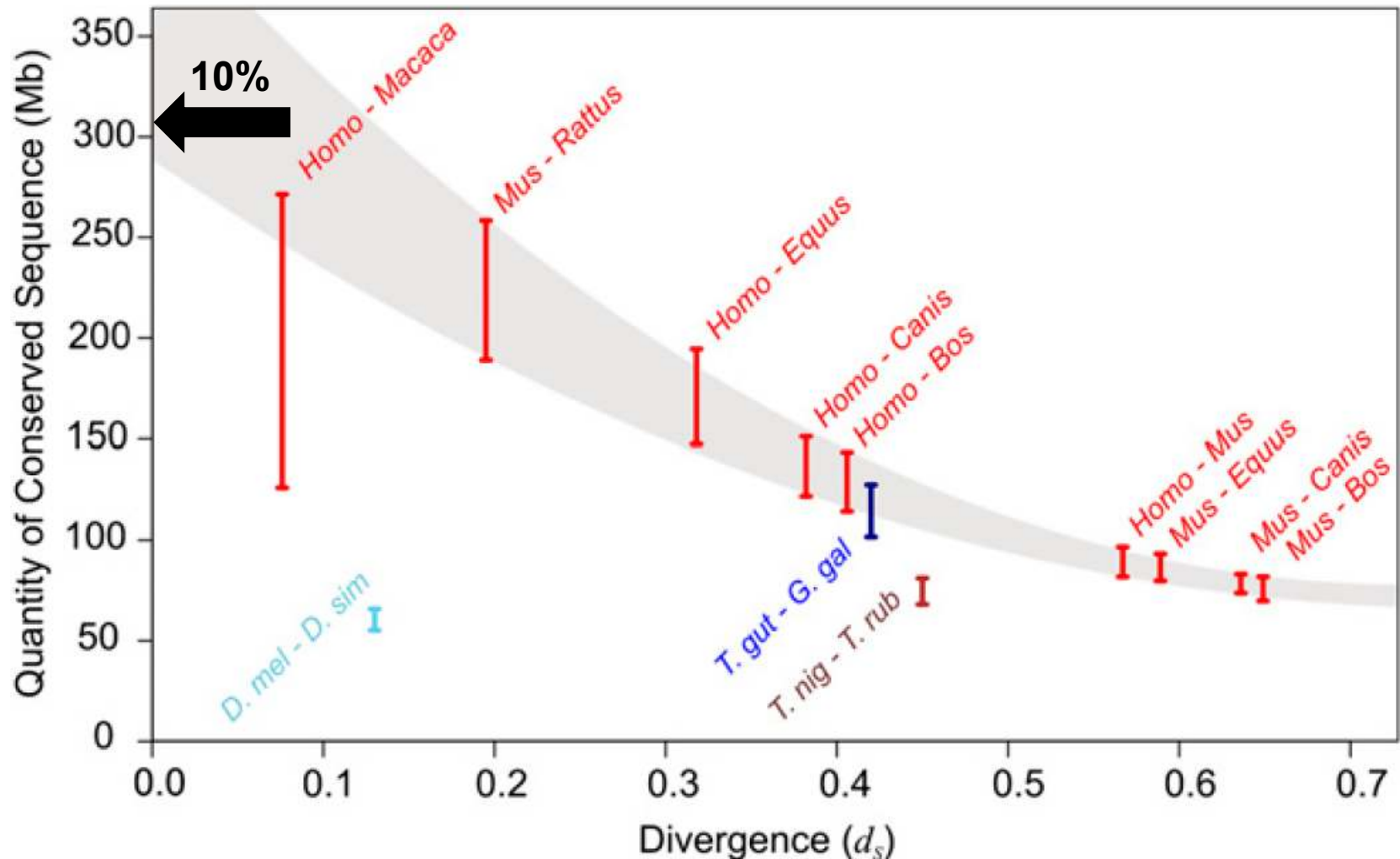
- human – mouse (75 My)
- human – macaque (25 My)
- mouse – rat (15 My)?

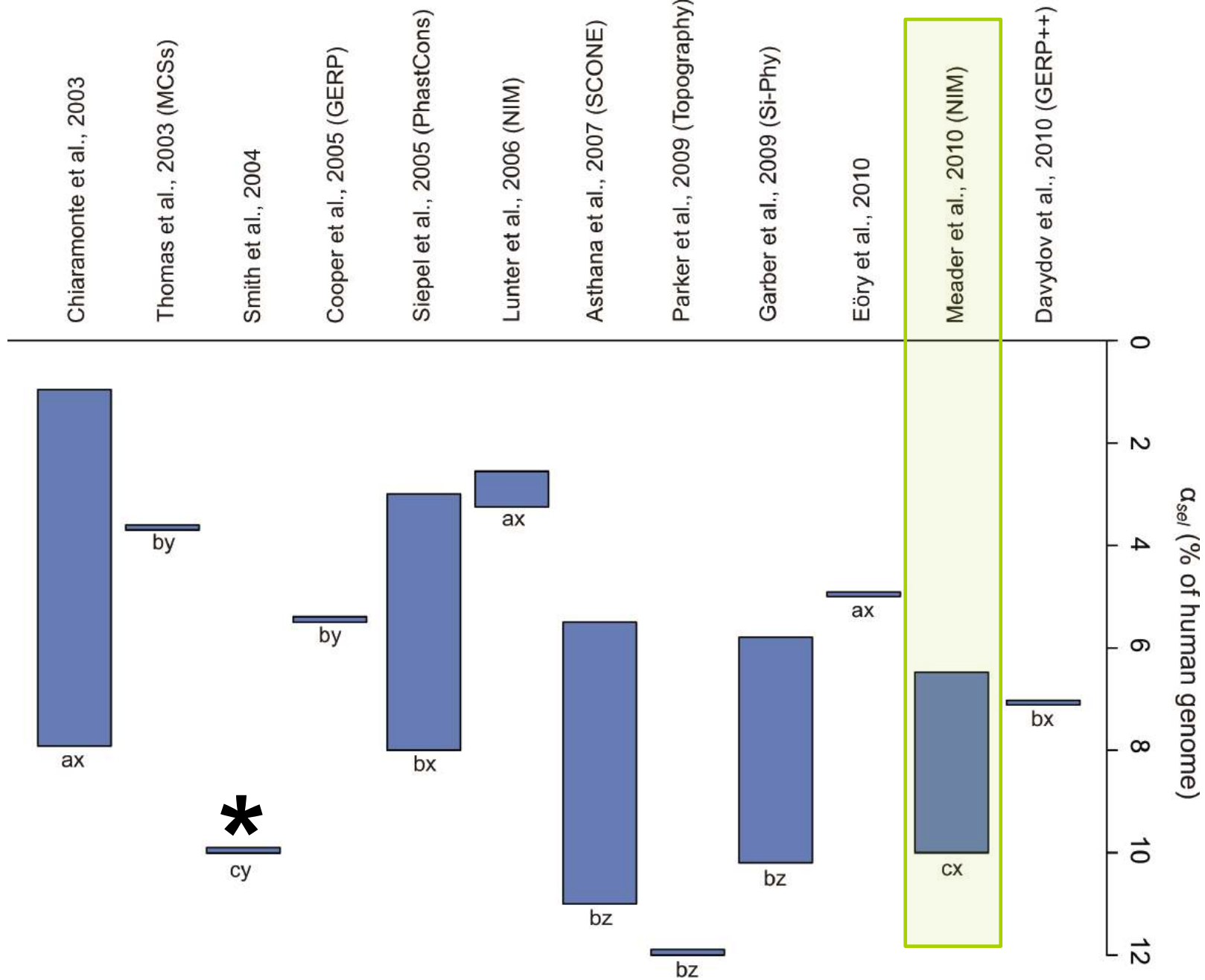
No

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





a. Single pairwise alignment
 b. Multiple species alignment
 c. Multiple pairwise alignments

x. Whole genome
 y. Partial genome ($\leq 12\text{Mb}$)
 z. ENCODE pilot regions (30Mb)



ELSEVIER

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

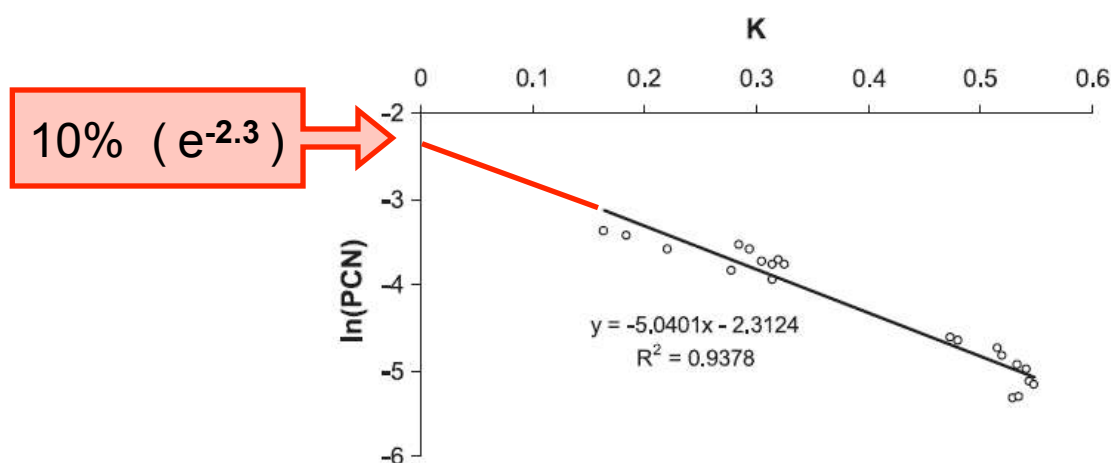


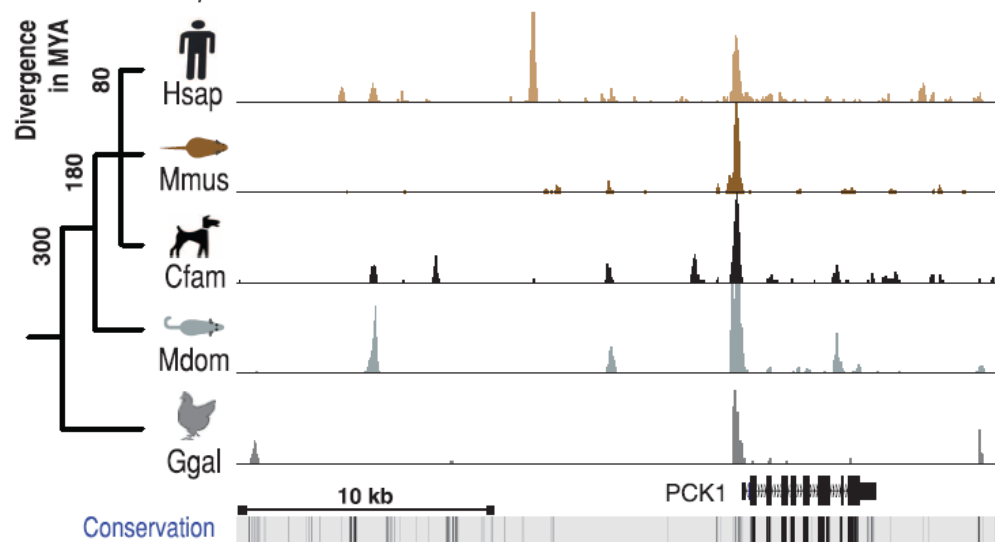
Fig. 1. The relationship between the proportion of the noncoding genome estimated to be conserved by negative selection, PCN, and the pairwise divergence, K , plotted for 21 different pairwise mammalian comparisons. The regression line for $\ln(PCN)$ versus K is also shown, along with its equation and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

Data:

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

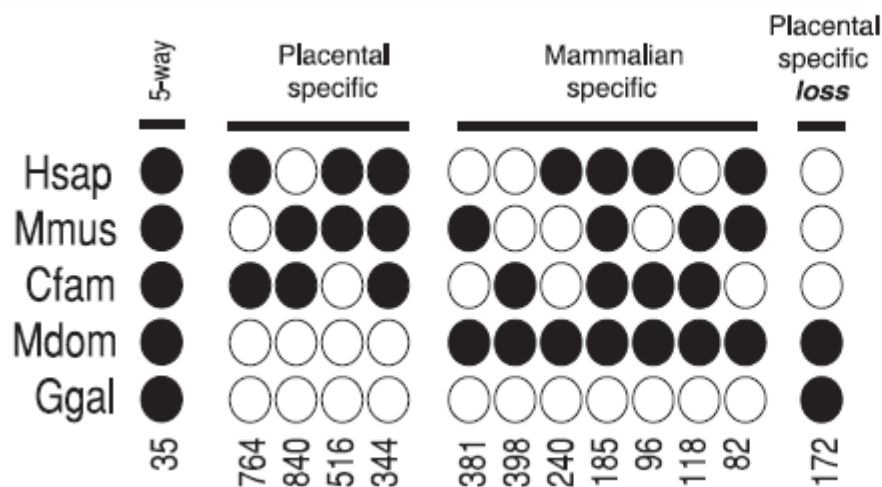
8 mammals
(human, baboon, cat,
dog, pig, cow, rat, mouse)

CEBPA ChIP-seq of animal livers



C

CEBPA



D

HNF4A

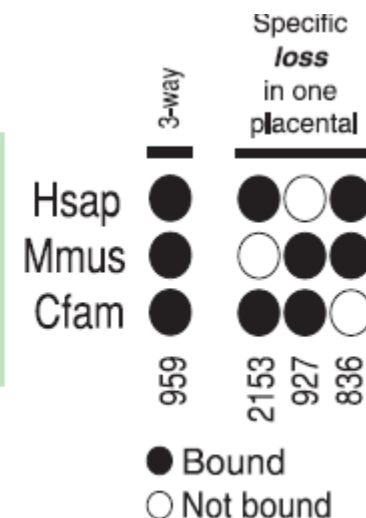
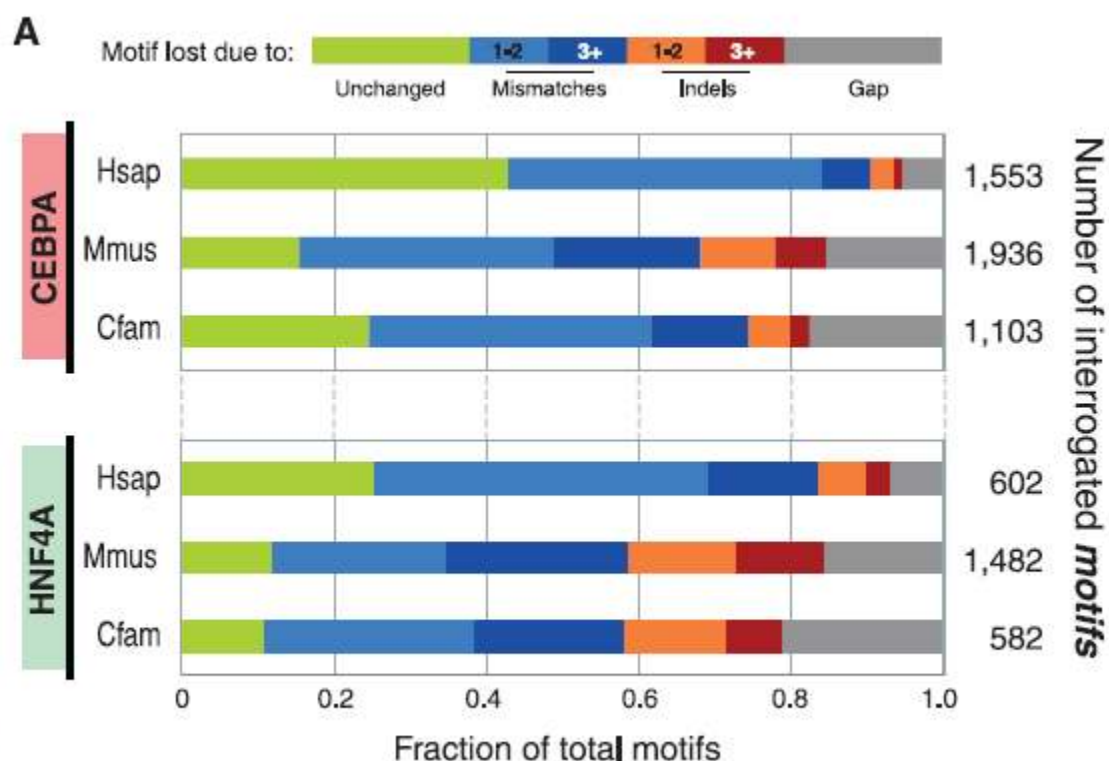


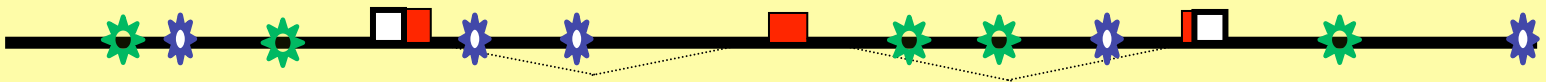
Fig. 4. Lineage-specific loss and turnover of TF binding events. **(A)** The unbound regions in each placental mammal that align to regions showing TF binding in the other two placental mammals were collected, and the mechanisms by which the underlying motifs were disrupted were summarized. **(B)** Turnovers 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).



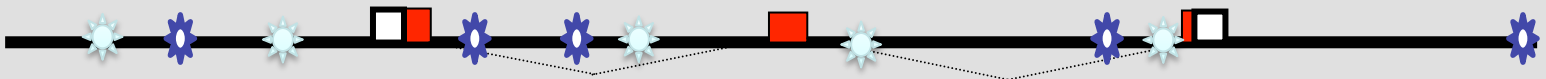
Approximately half of lineage-specific losses of TF binding showed evidence of nearby compensatory binding events (Fig. 4B). A quarter of species-specific losses had a nearby (± 10 kb) gained binding event that is unique to the same lineage (unshared turnover), and an additional quarter of the losses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



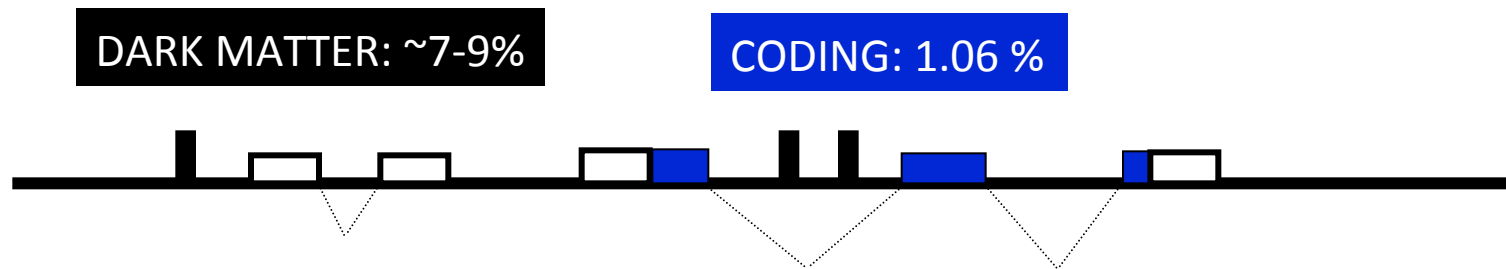
Time →



Lineage #1

Lineage #2

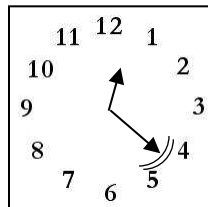
Our View of the Human/Mouse Genome



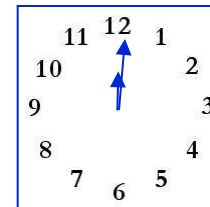
King & Wilson: *Human and Chimpanzee "macromolecules are so alike that regulatory mutations may account for their biological differences"*

Science (1975) 188, 107-116

dark matter



proteins

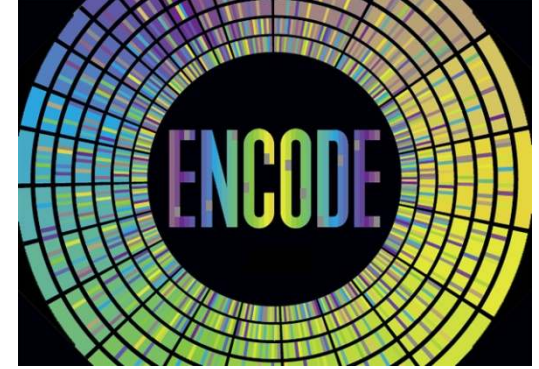


ENCODE's 80%

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

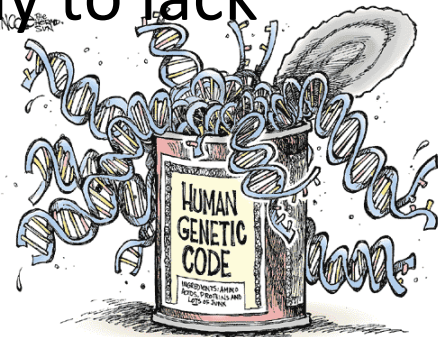
The Functional Portion

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*.

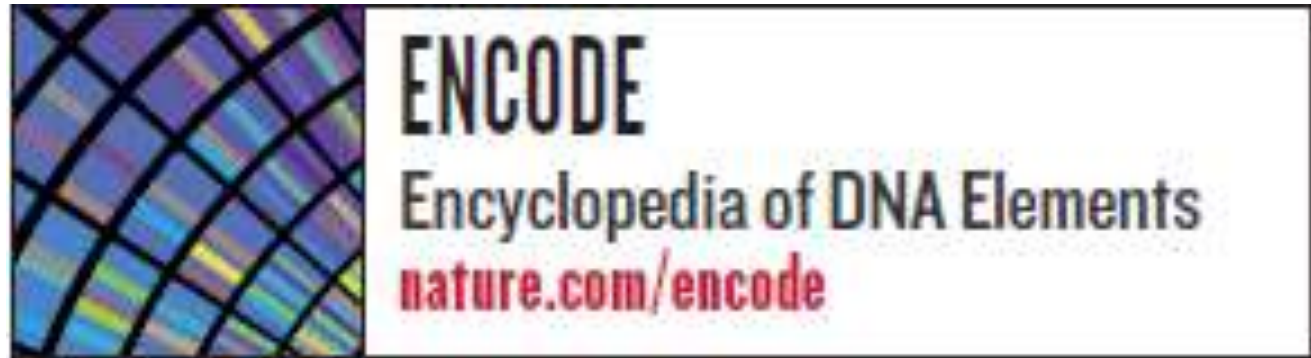


The Future: Functional, unconserved, sequence

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

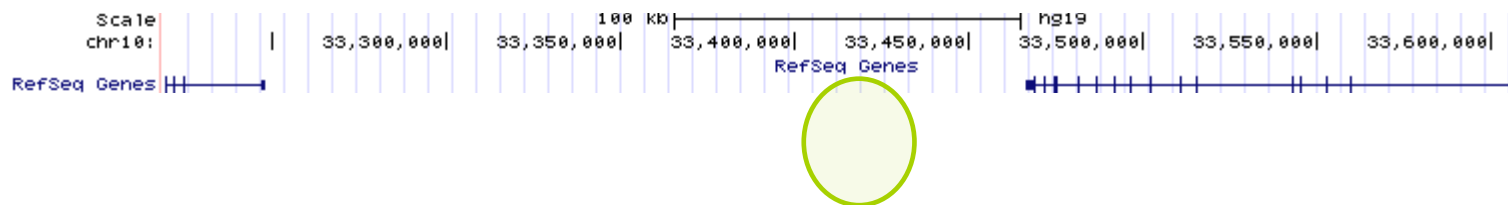
Deducing the functions of such lineage-specific sequence will require:

comparisons to many primate genomes; and, diverse experimental approaches.



ENCODE 2012

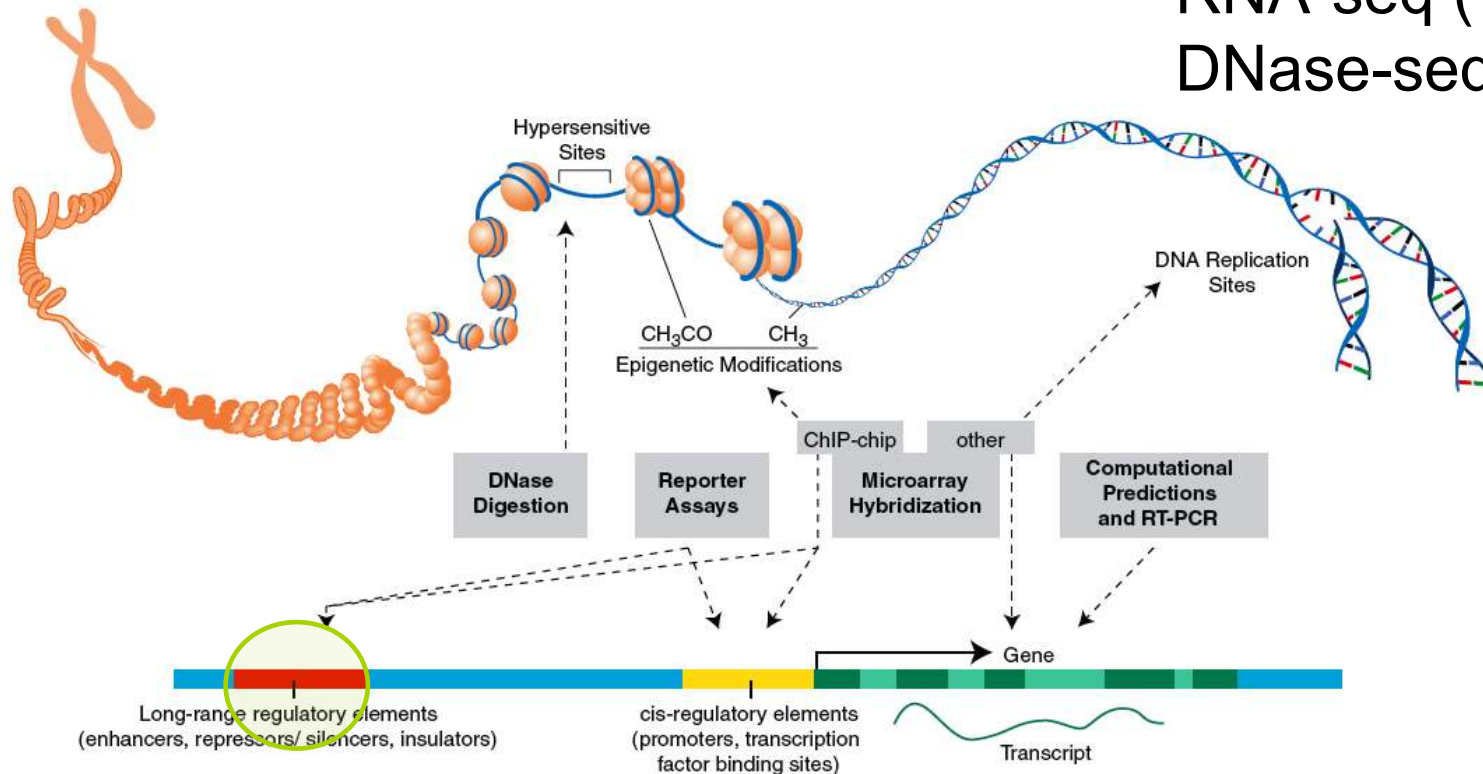
2001



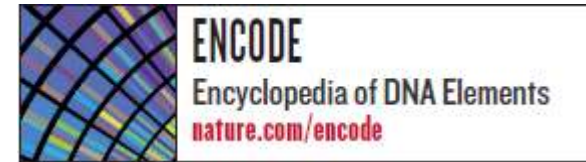
2012

Experimental assays

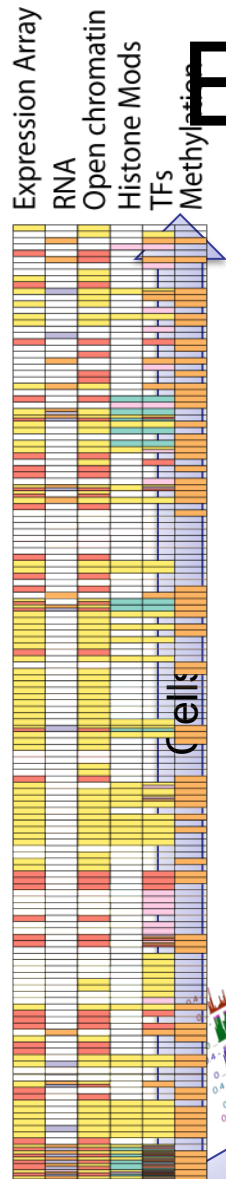
ChIP-seq (~150)
 RNA-seq (~100)
 DNase-seq (~100)



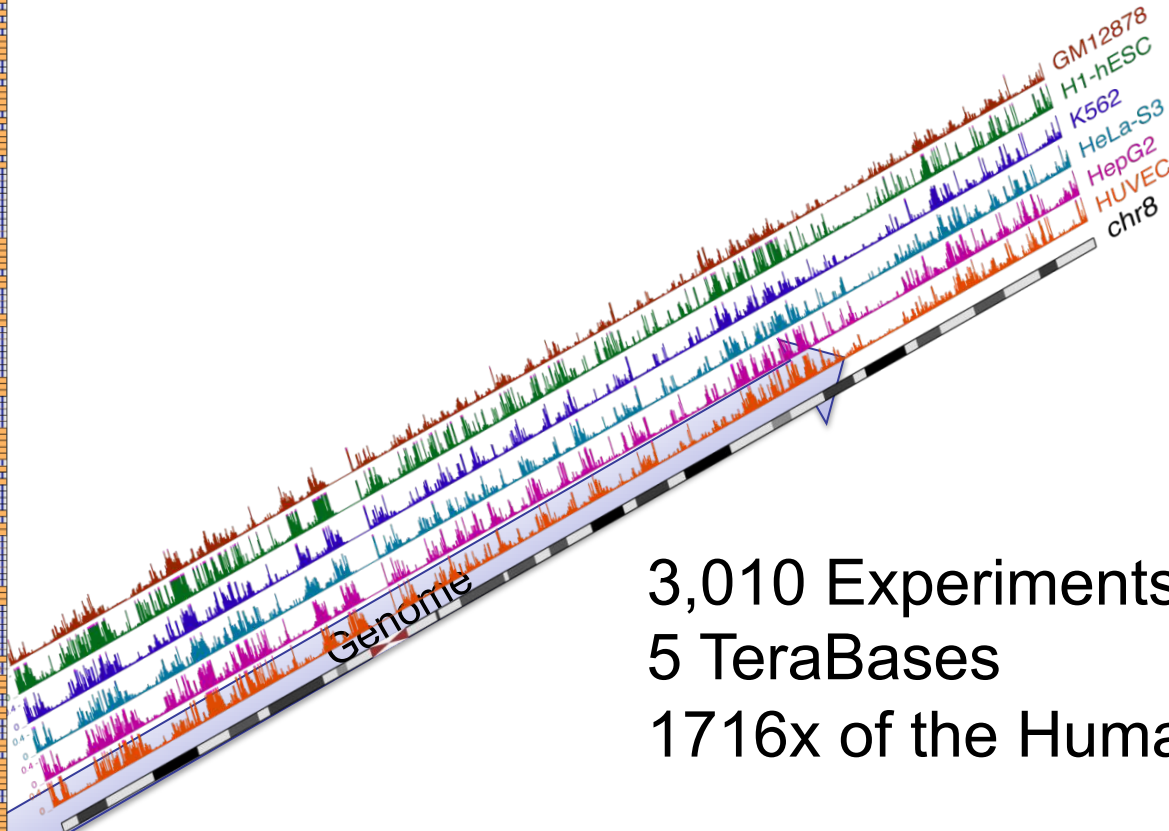
ENCODE Dimensions



182 Cell Lines/ Tissues



Cells



Genome

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

GM12878
K562
H1-hESC
HeLa-S3
HepG2
Huvec

Histone
Mods
Pol2/3

Method, Factors

Transcription Factors

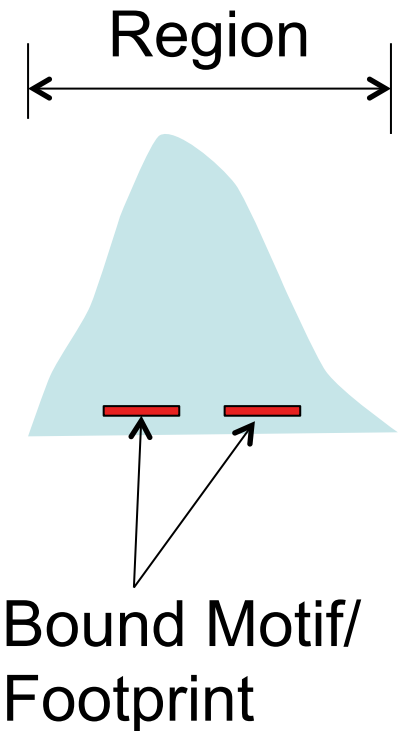
164 Assays (114 different ChIP)

Control

“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)		

(Union over all experiments and cell types)



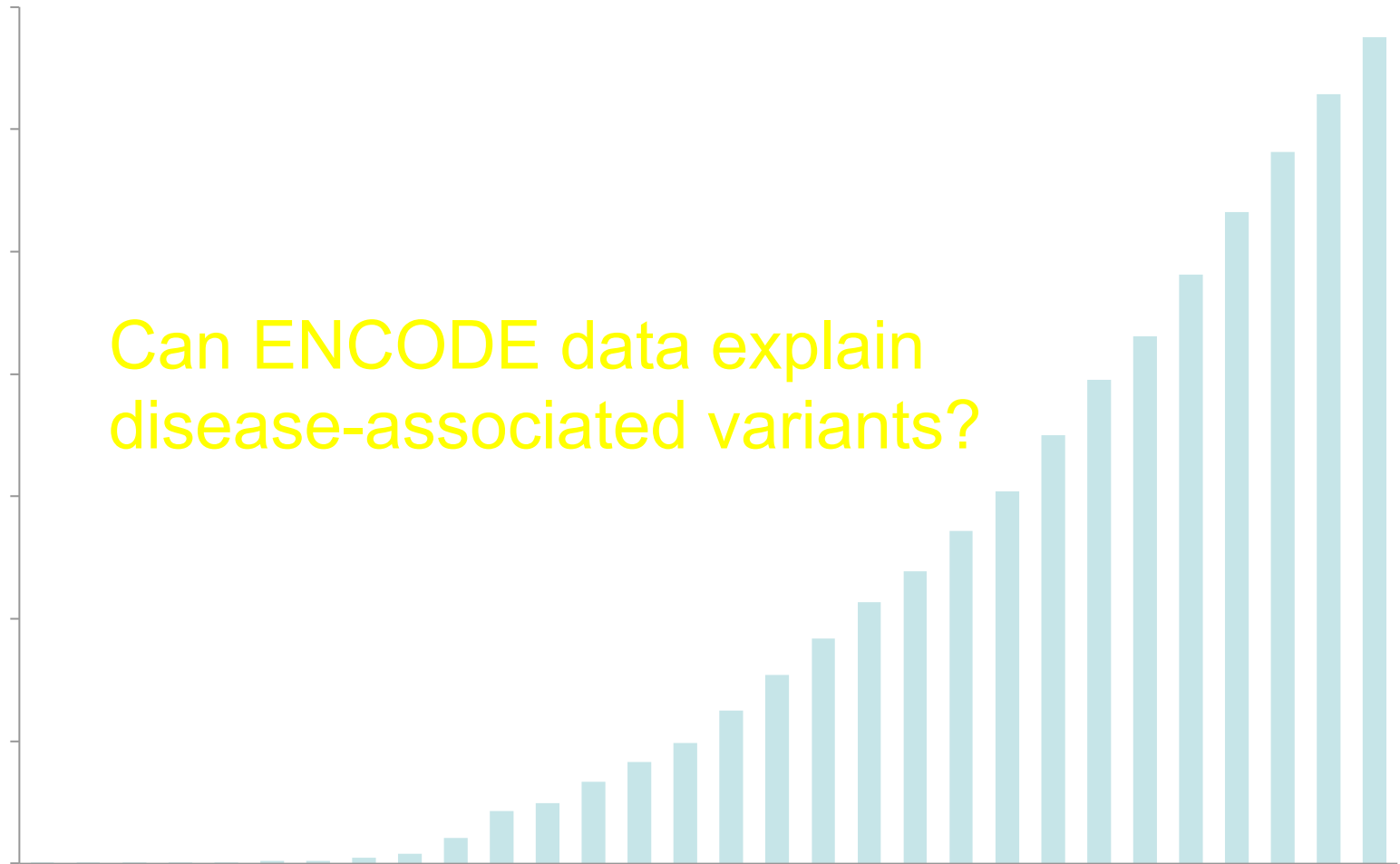
Elements are evenly spaced over the genome



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

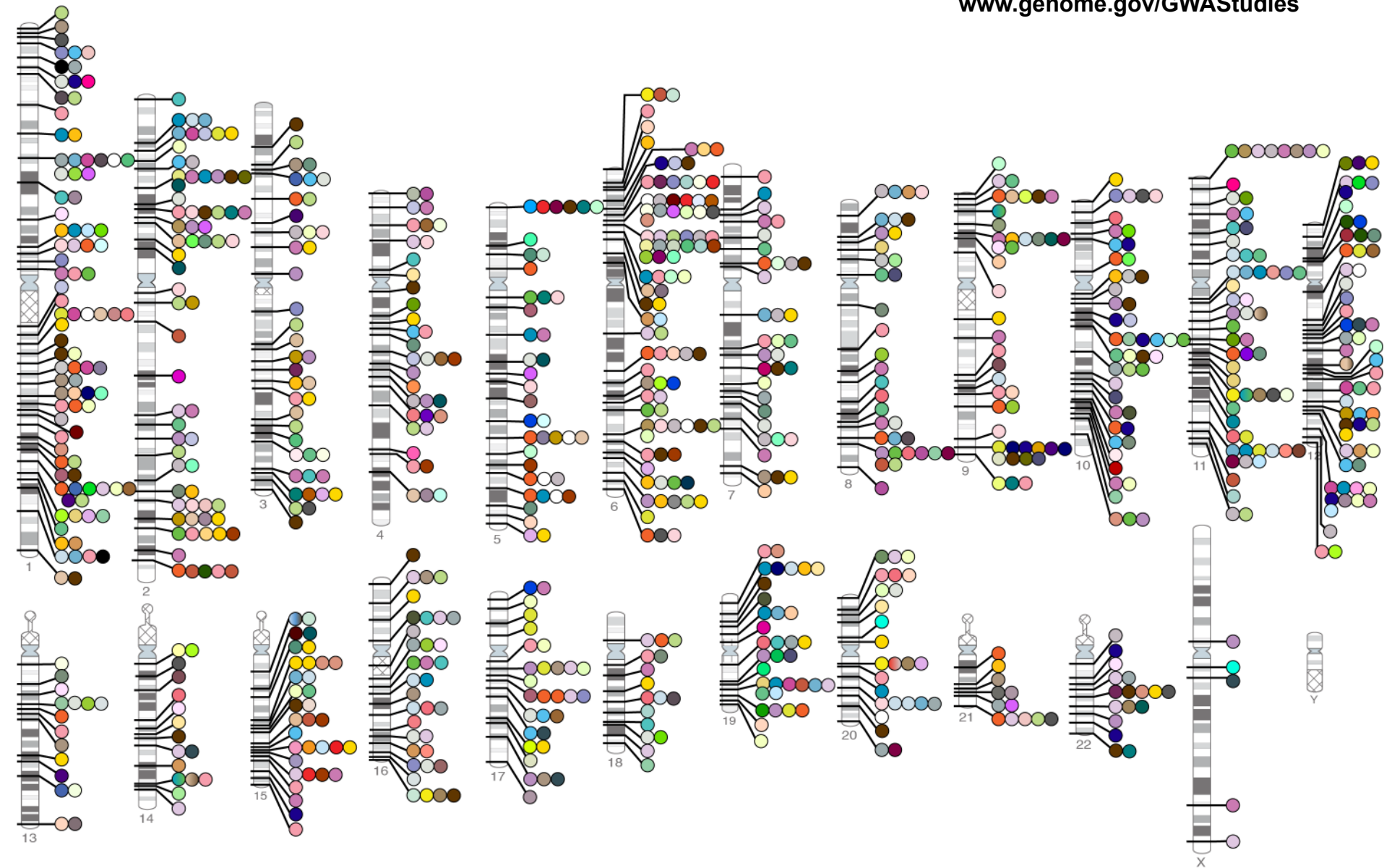
Published GWA Reports, 2005 – 6/2012



Published Genome-Wide Associations through 6/2010

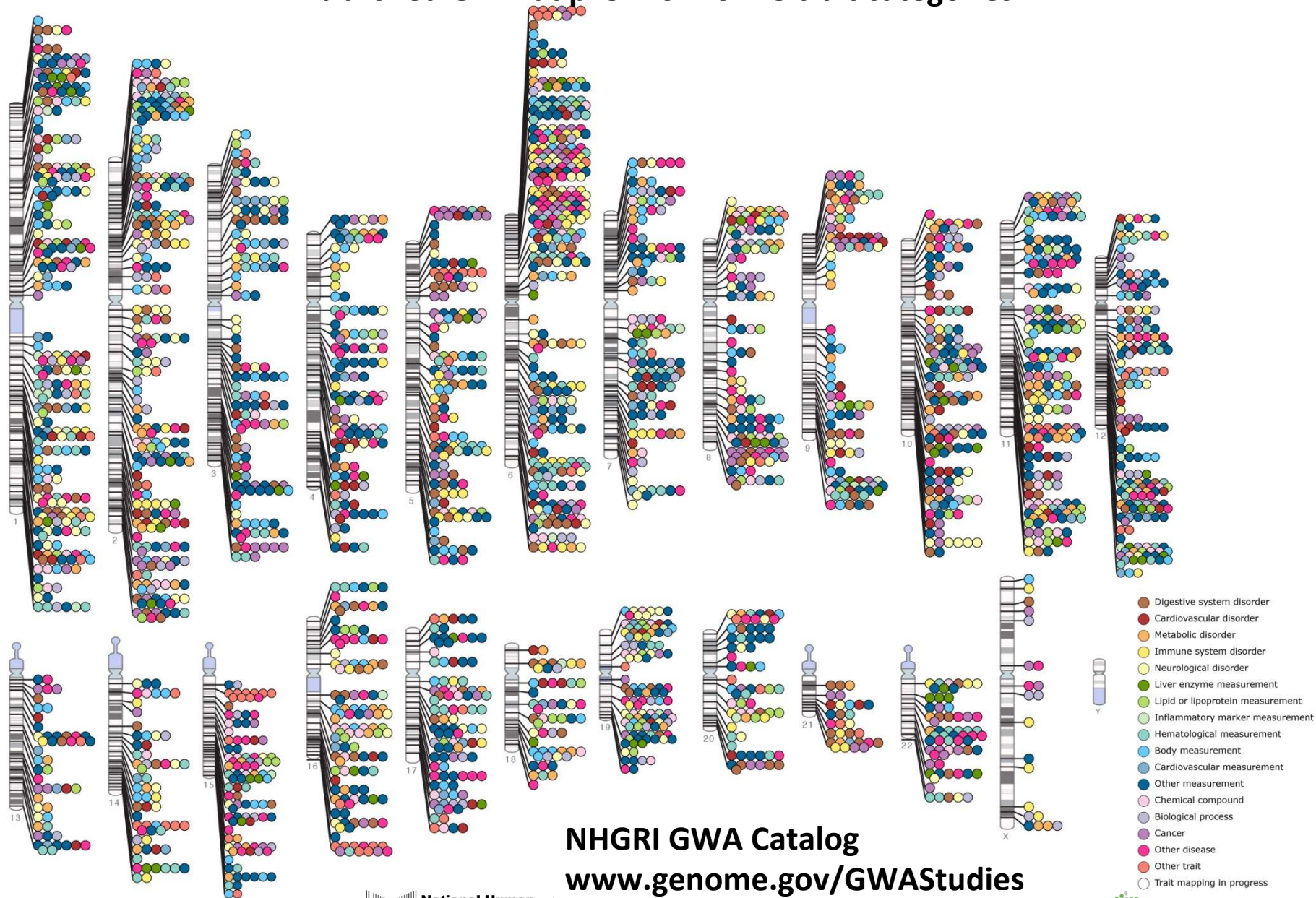
NHGRI GWA Catalog

www.genome.gov/GWASudies



Published Genome-Wide Associations through 07/2012

Published GWA at $p \leq 5 \times 10^{-8}$ for 18 trait categories



NHGRI GWA Catalog

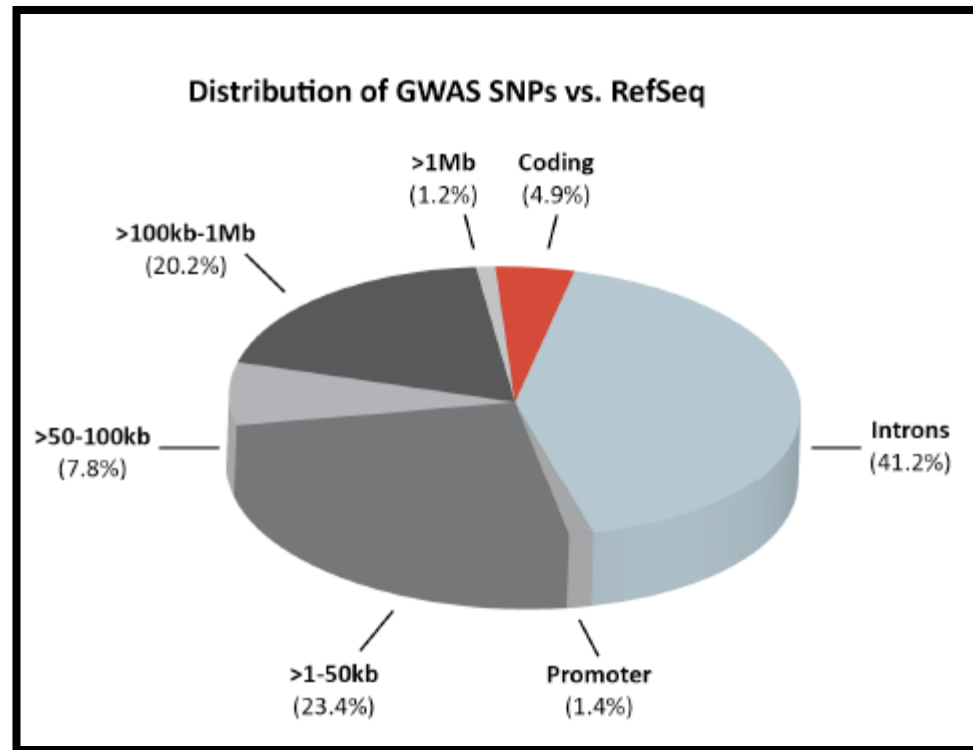
www.genome.gov/GWASStudies

www.ebi.ac.uk/fgpt/gwas/

EMBL-EBI

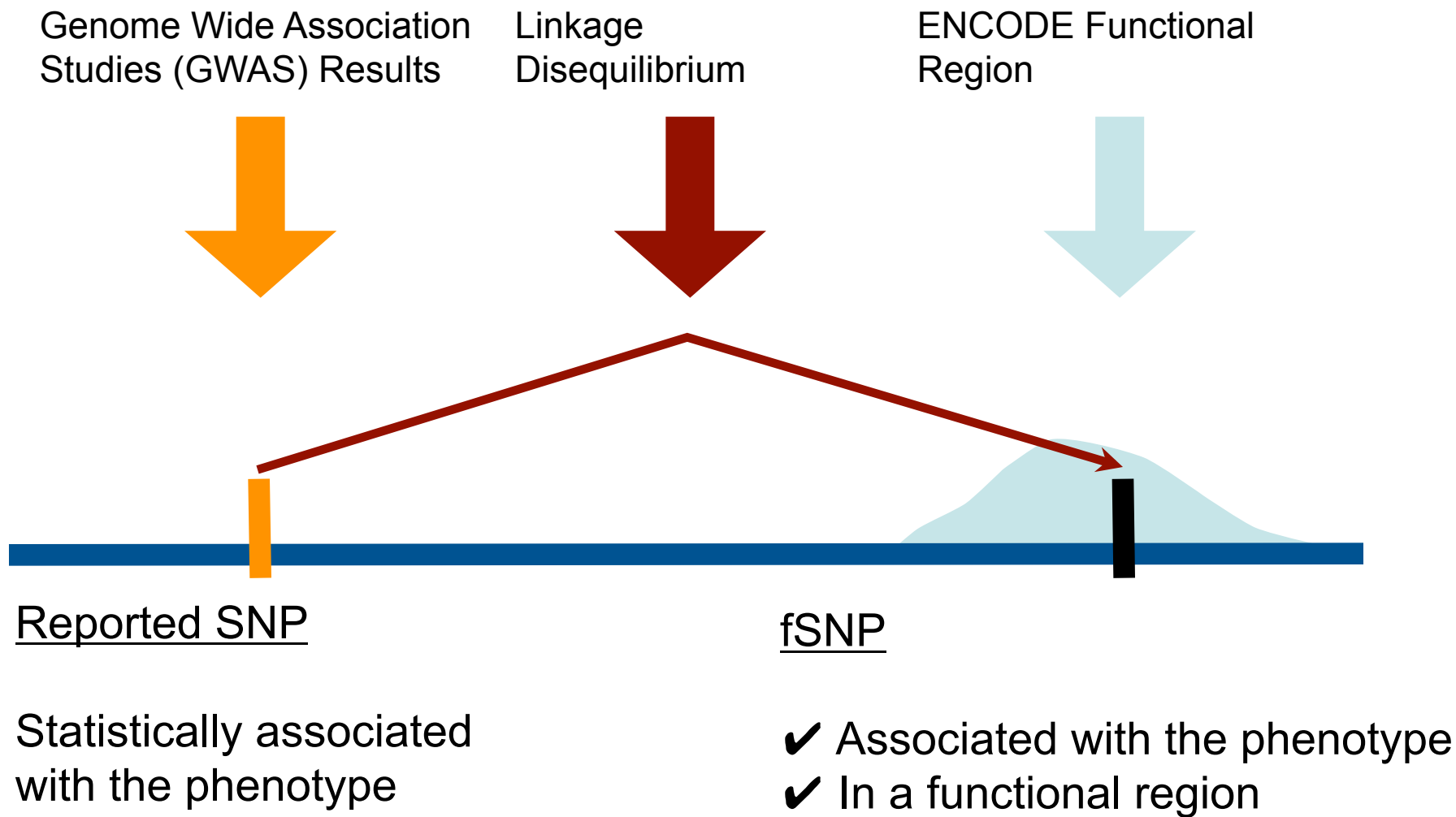


95% of GWAS risk-associated SNPs lie *outside* of coding sequence



Functional SNPs (fSNPs)

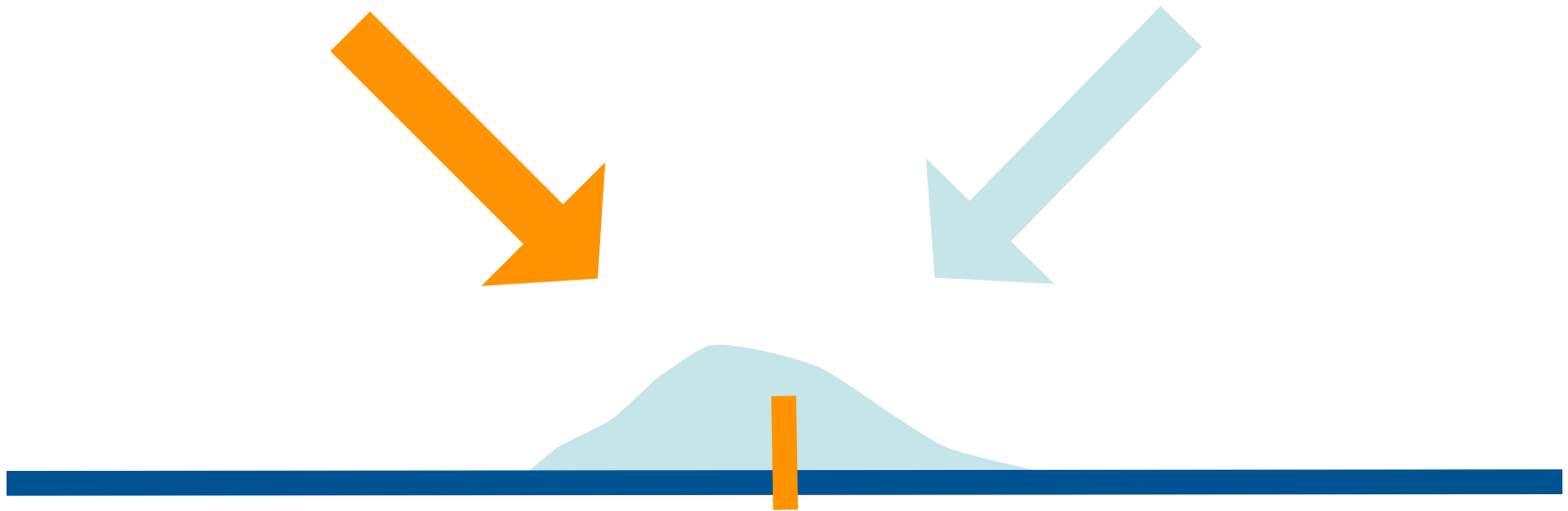
Belinda Giardine, Marc Shaub, Ross Hardison, Mike Snyder, John Stam.



Functional SNP - Direct Hit

Genome Wide Association
Studies (GWAS) Results

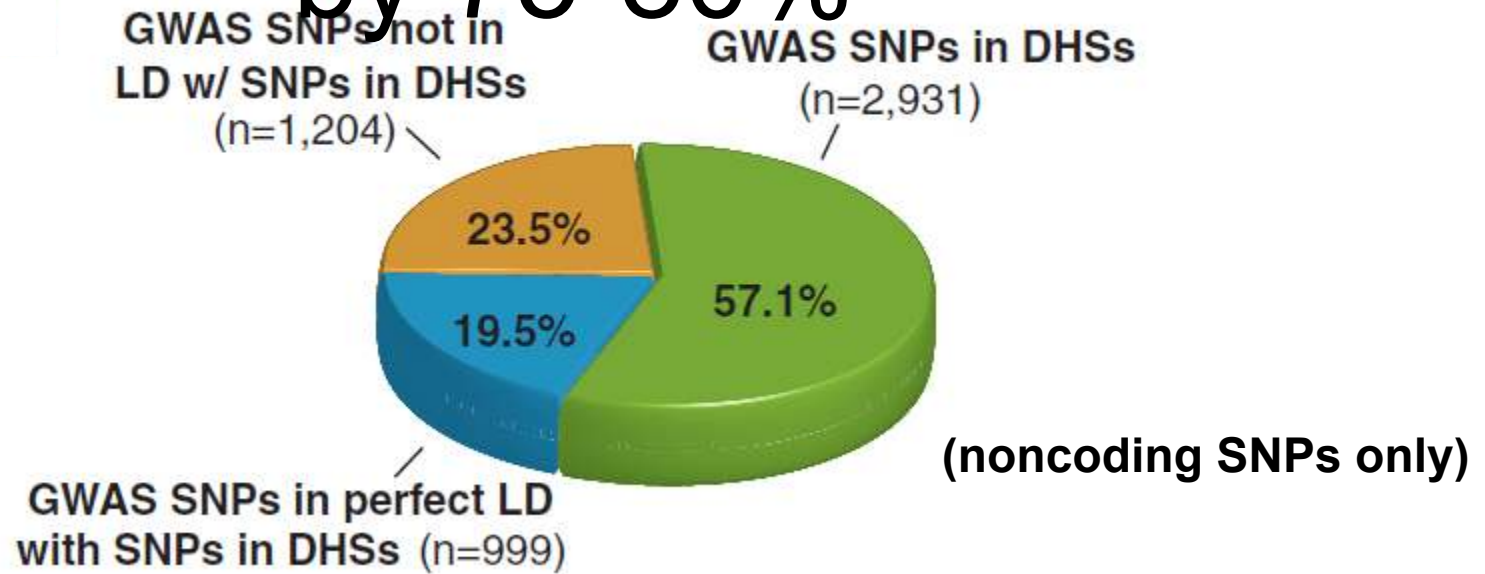
ENCODE Functional
Region



fSNP Direct Hit

- ✓ Association reported in a GWAS
- ✓ In a functional region

high LD, GWAS SNPs overlap
by 75-80%



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

Science

AAAS

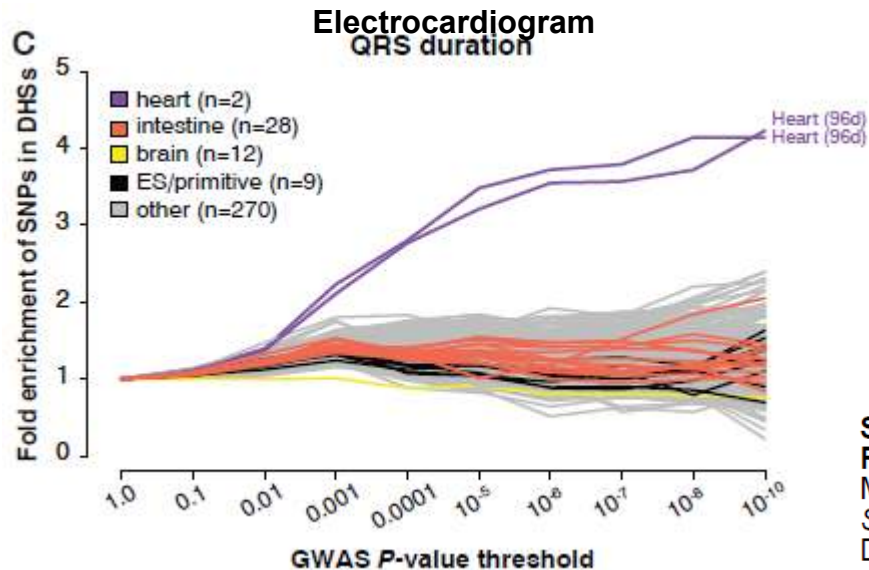
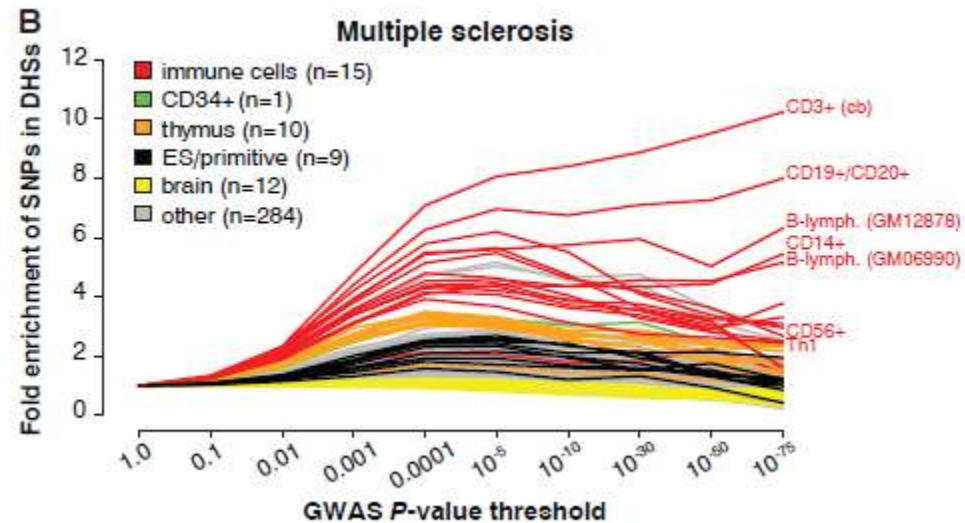
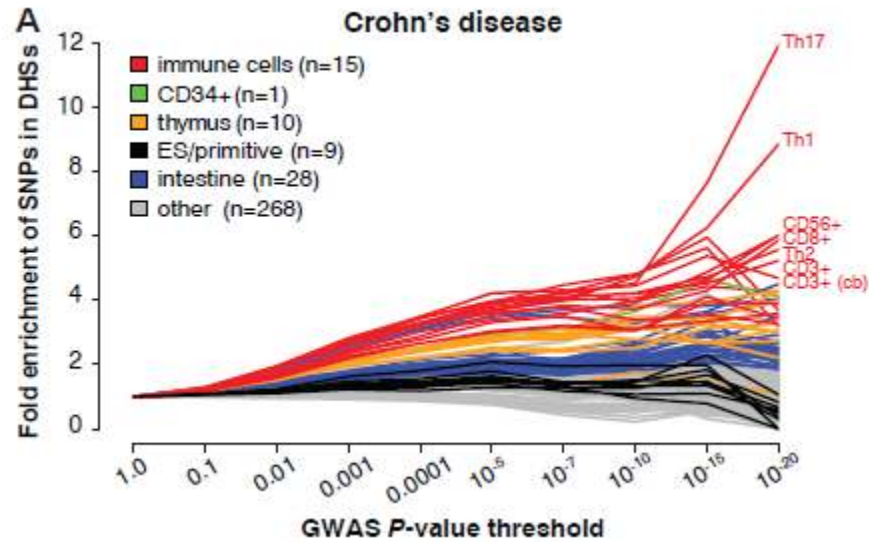
Systematic Localization of Common Disease-Associated Variation in
Regulatory DNA

Matthew T. Maurano *et al.*

Science **337**, 1190 (2012);

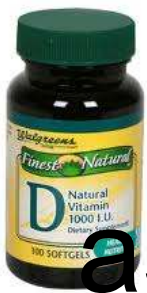
DOI: 10.1126/science.1222794

of disease relevant cells

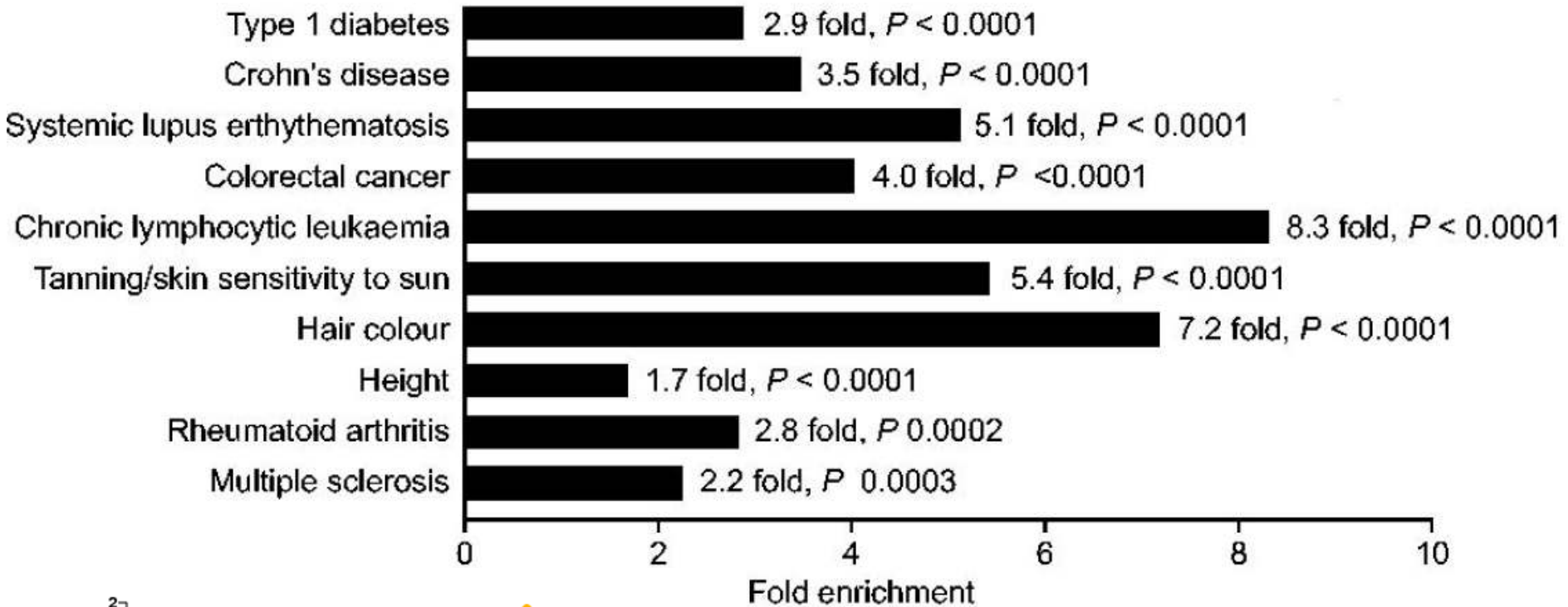


Systematic Localization of Common Disease-Associated Variation in Regulatory DNA

Matthew T. Maurano *et al.*
Science **337**, 1190 (2012);
 DOI: 10.1126/science.1222794



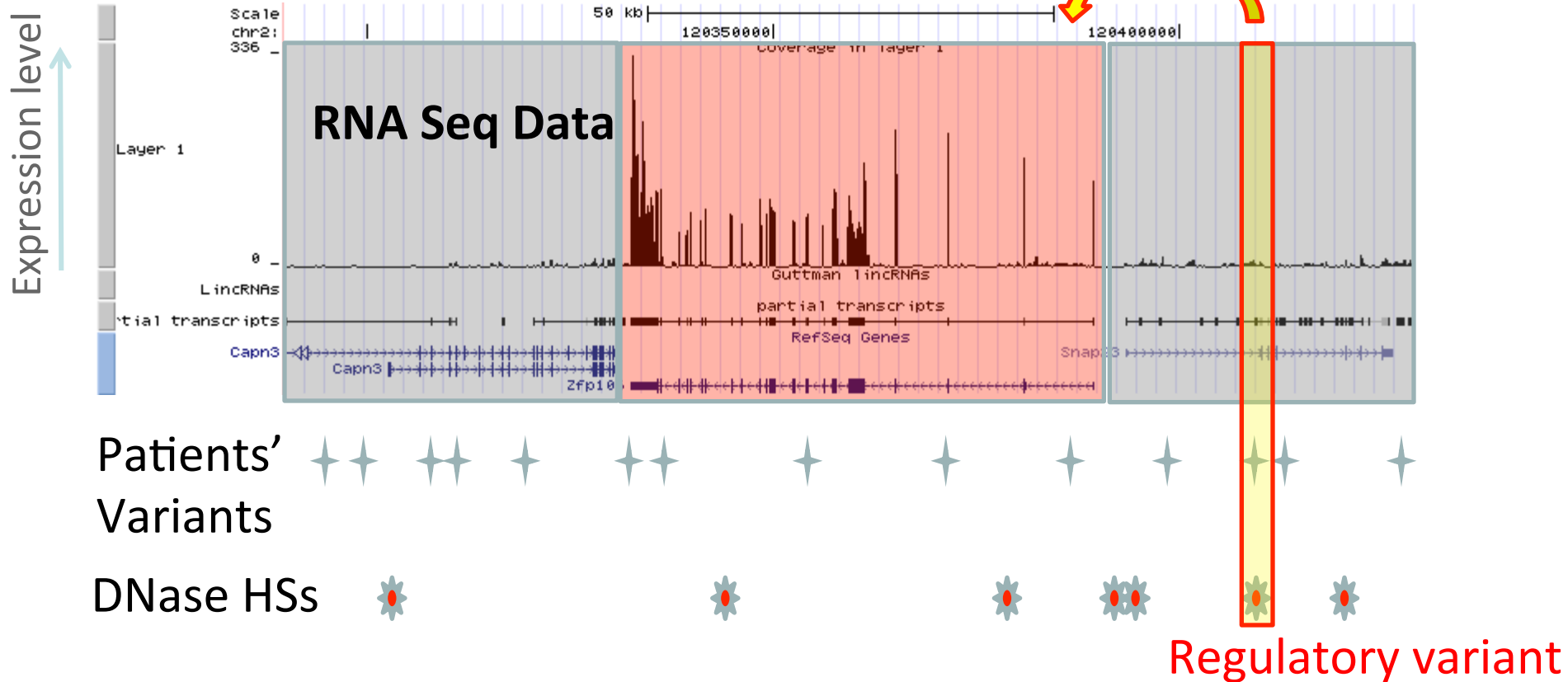
Diseases/traits associated with VDR binding

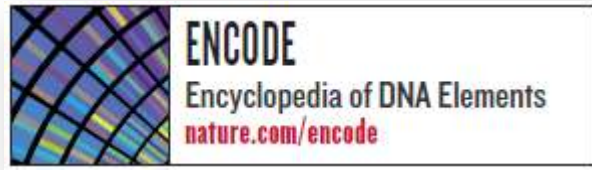


Causal Variant Identification using a Disease-relevant cell type

Disease-associated interval

Expression level or alternative transcript





- “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”

Most “Dark Matter” Transcripts Are Associated With Known Genes

Harm van Bakel¹, 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

Perspective

The Reality of Pervasive Transcription

Michael B. Clark¹, Paulo P. Amaral^{1,9}, Felix J. Schlesinger^{2,9}, Marcel E. Dinger¹, Ryan J. Taft¹, John L. Rinn³, Chris P. Ponting⁴, Peter F. Stadler⁵, Kevin V. Morris⁶, Antonin Morillon⁷, Joel S. Rozowsky⁸, Mark B. Gerstein⁸, Claes Wahlestedt⁹, Yoshihide Hayashizaki¹⁰, Piero Carninci¹⁰, 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

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 Lunter¹

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 Maiwen Caudron-Herger,⁵ Matthias Groß,² Peter Schirmacher,³ Karsten Rippe,⁵ Thomas Braun,⁴ Martin Zörnig^{1,*} and Sven Diederichs^{2,*}



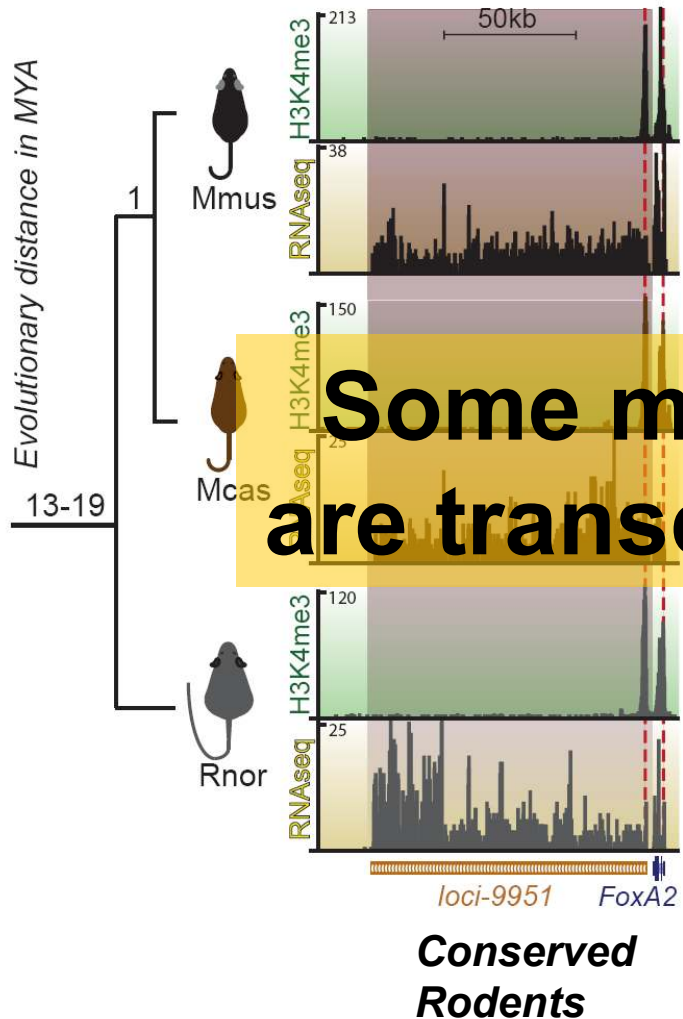
ENCODE

Encyclopedia of DNA Elements

nature.com/encode

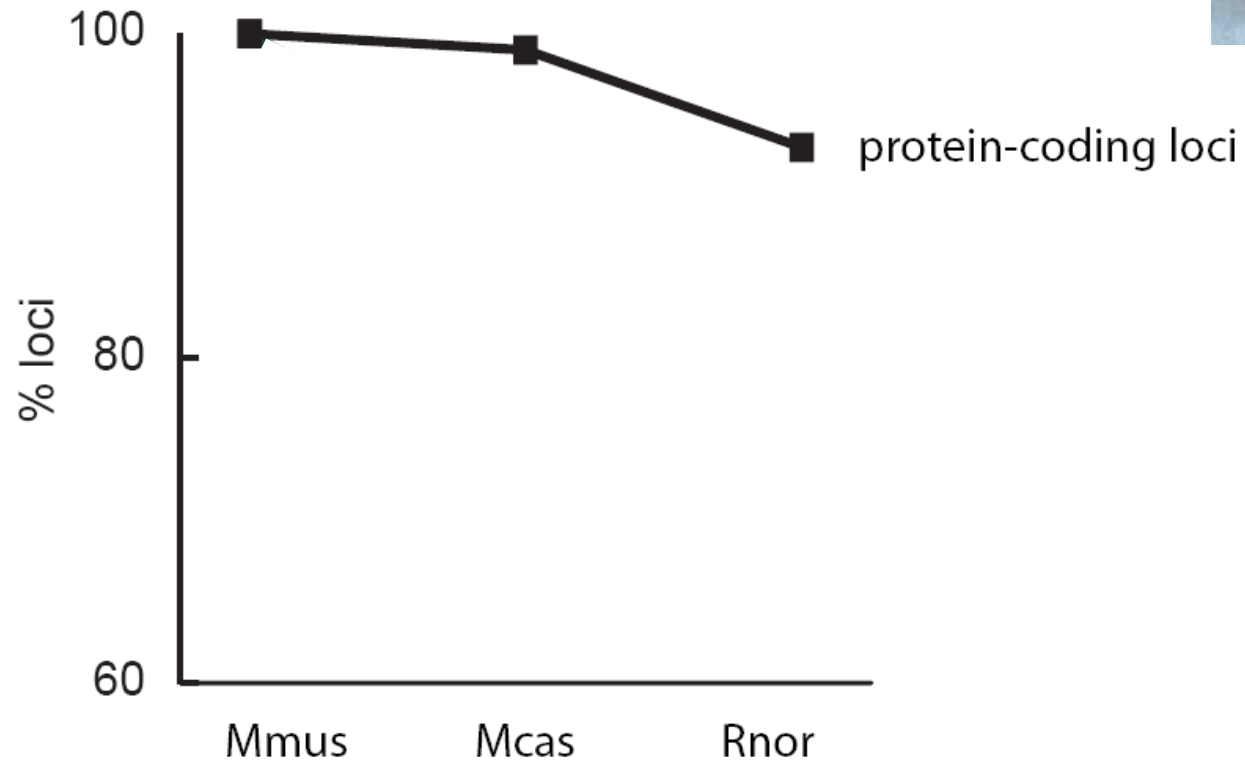
?

Loss of transcription but not of genomic sequence.

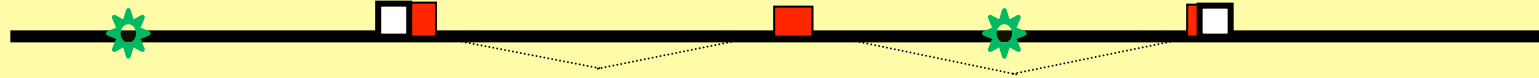


**Some mammalian lincRNAs
are transcribed only fleetingly.**

LincRNA and protein-coding transcriptional turnover



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



Coding
Gene
Expression

OPEN ACCESS Freely available online

PLOS GENETICS

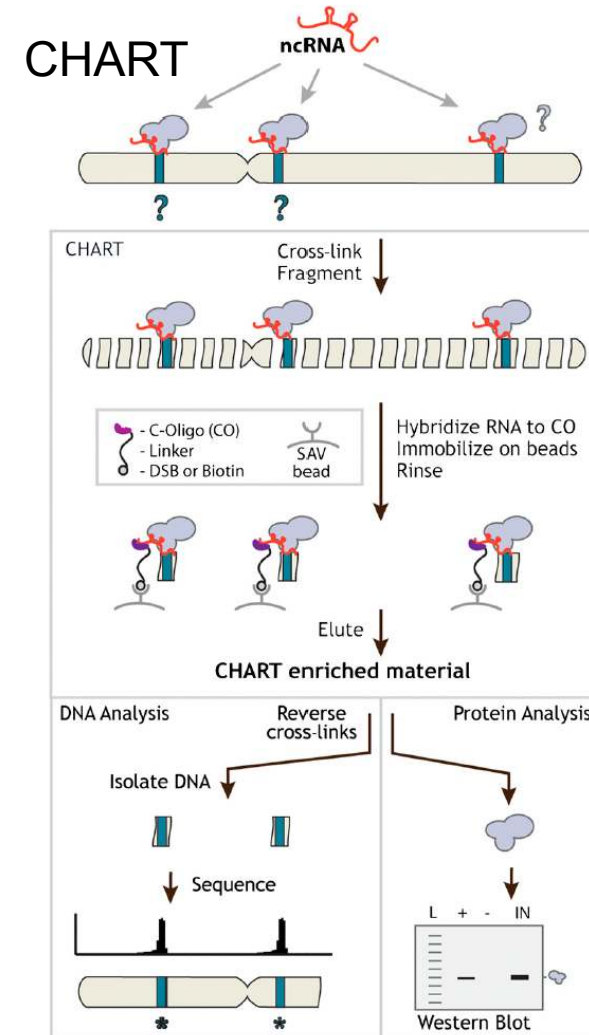
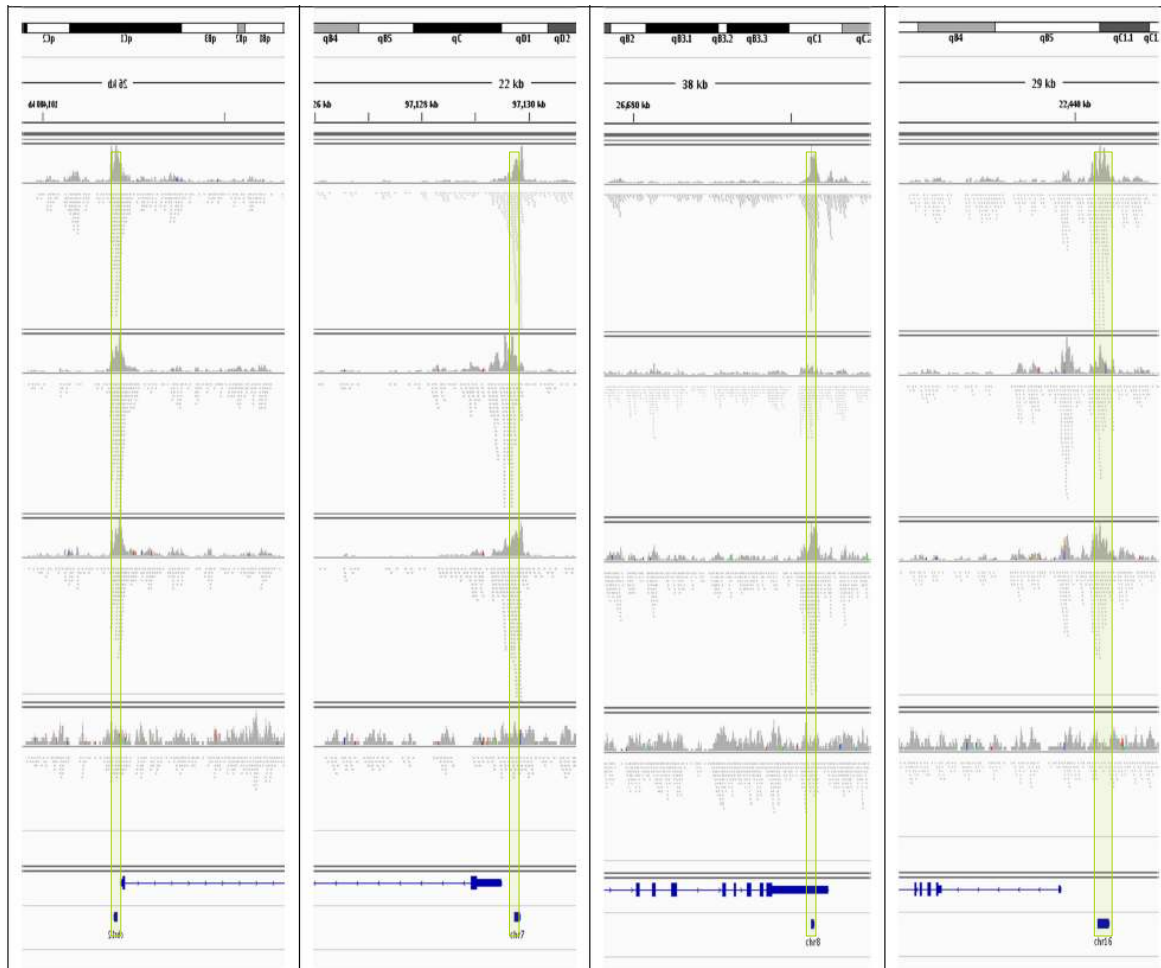
Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression

Claudia Kutter^{1,2,3}, Stephen Watt^{1,3}, Klara Stefflova^{1,2}, Michael D. Wilson^{1,2,4}, Angela Goncalves^{2,3},
Chris P. Ponting^{4,5,6,†,*}, Duncan T. Odom^{1,2,4,†,*}, Ana C. Marques^{6,†,*}

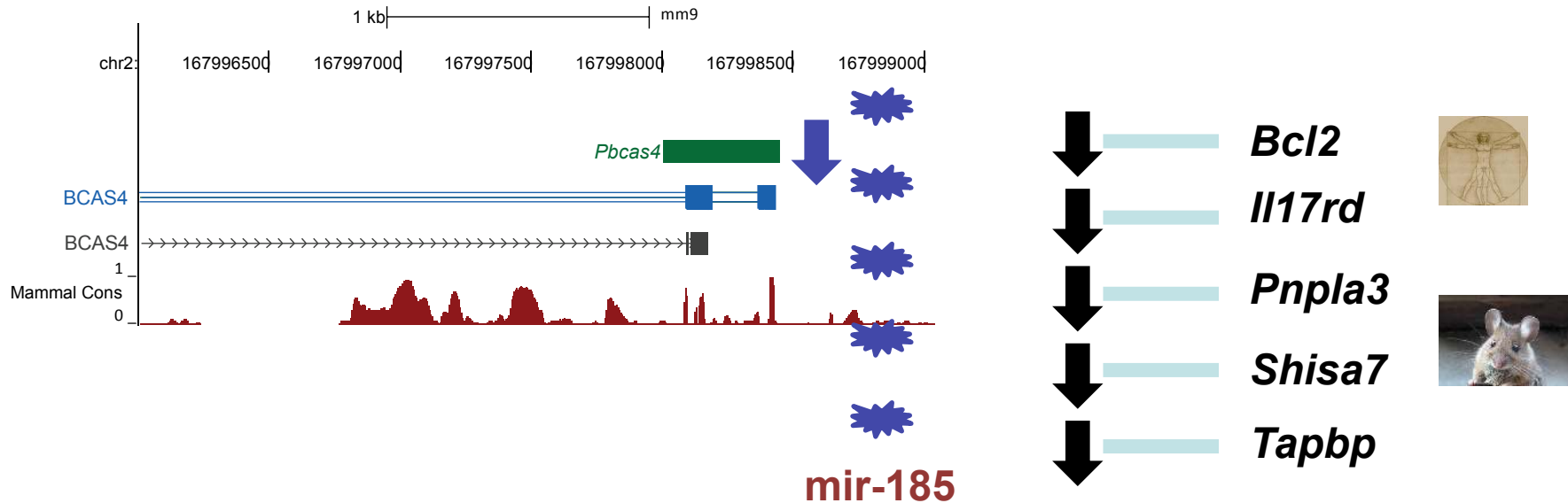
Type 2: ribonucleoprotein (rnp)RNAs

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}



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

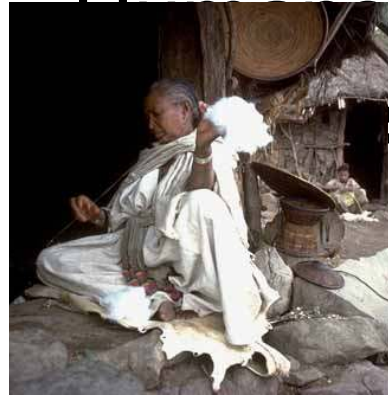
Inter-Species Sequence Comparisons



Intra-Species Sequence Comparisons



Different, but not that different



Humans are one of the least diverse
species

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



10,000 bases of human chr13 vs chimpanzee

Chimpanzee: ~ 2% divergence

2 Humans: ~ 0.1% diversity (10 sites)

No-one (genome) is perfect

Any European individual's genomes are expected to carry:

- 100 loss-of-function variants;

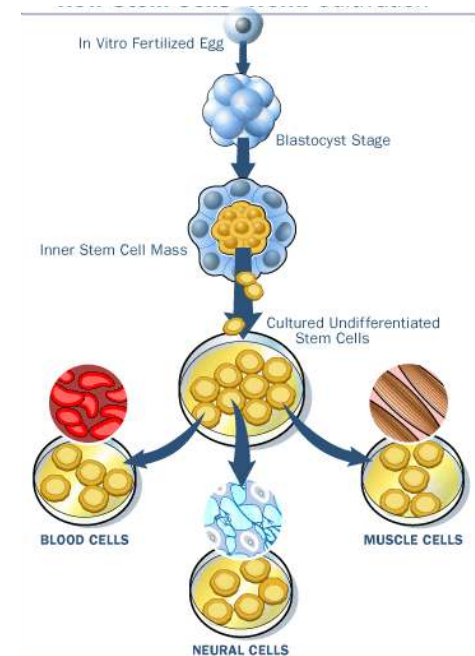
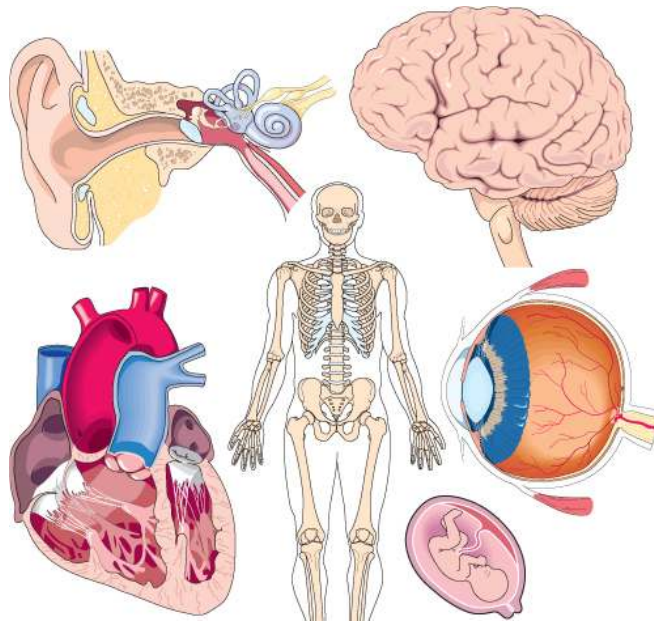
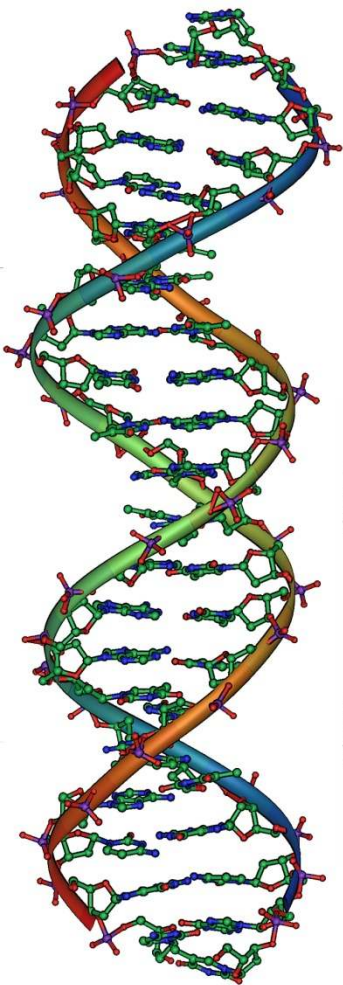
- of which 18 are in a homozygous state.

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

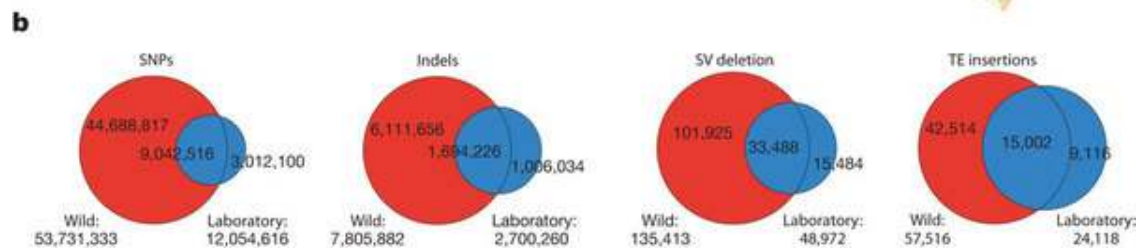
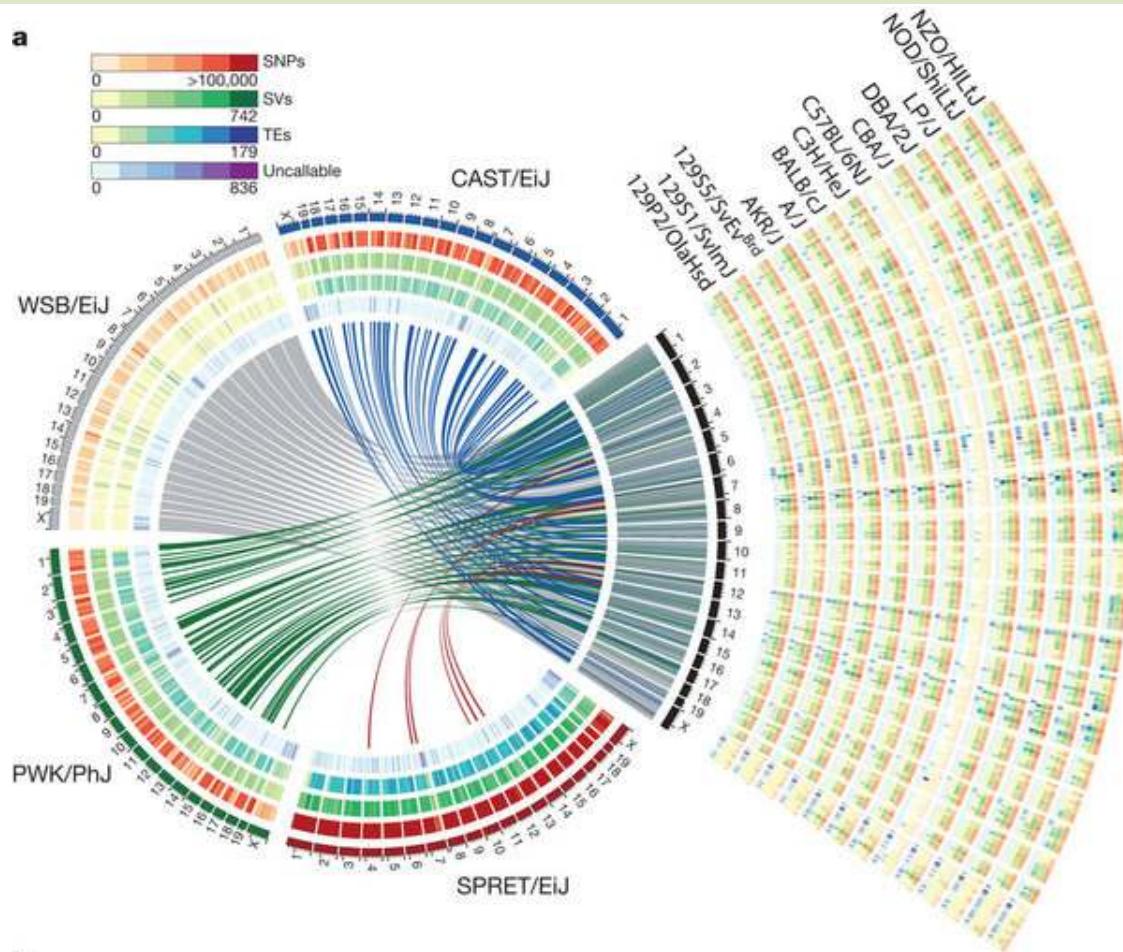


1000 Genomes
A Deep Catalog of Human Genetic Variation



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

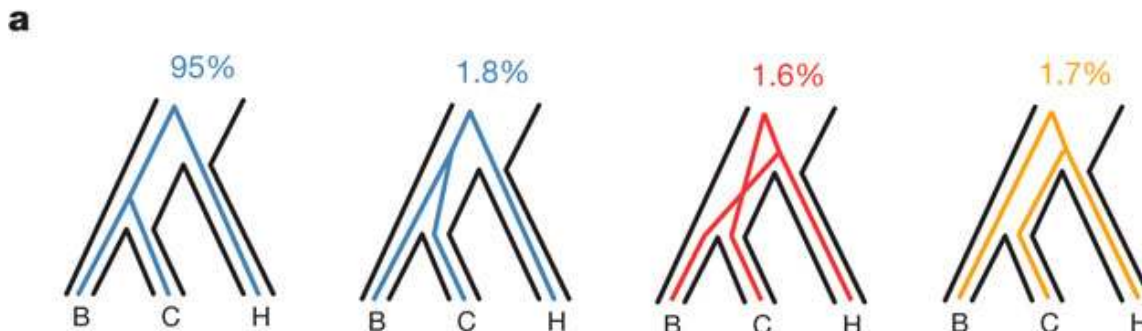
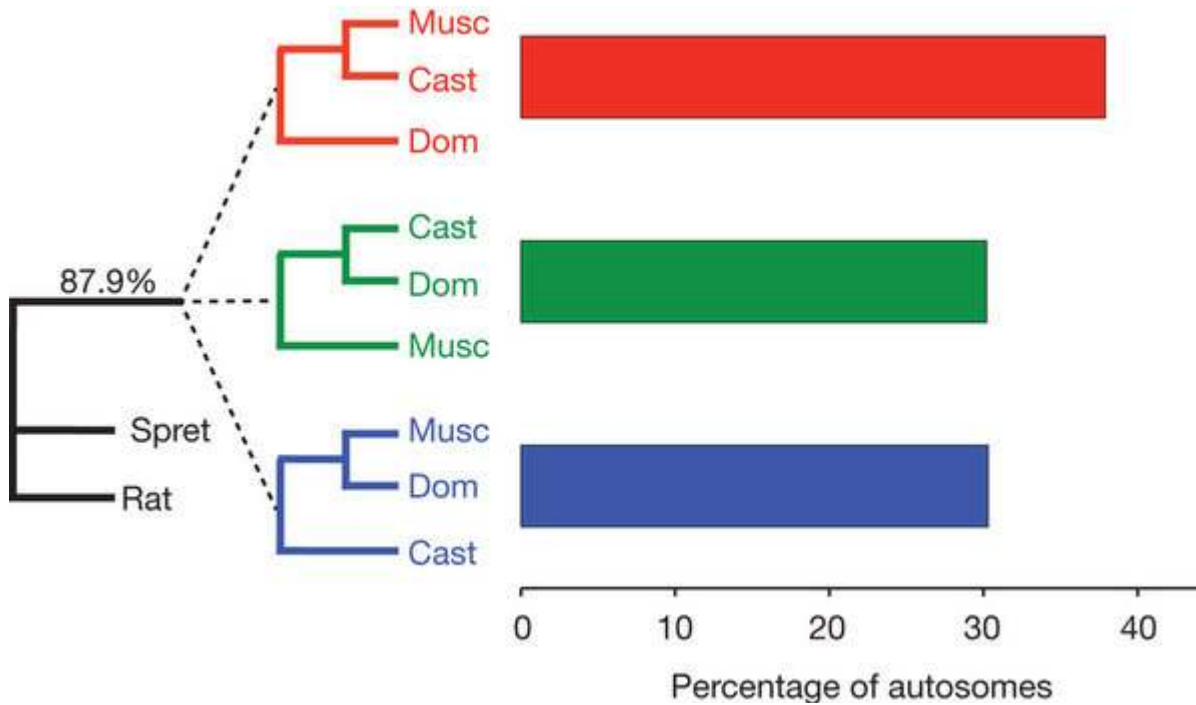
QTL Pct Var	Intergenic	Downstream	Exon	Intron	Upstream
All	1.18**	0.71	0.7	0.79	0.67
<4%	1.21**	0.67	0.67	0.75*	0.63
>4%	0.57**	1.05	1.28	1.43*	0.97
>10%	0.65**	1.32	1.59*	1.69**	1.32

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**

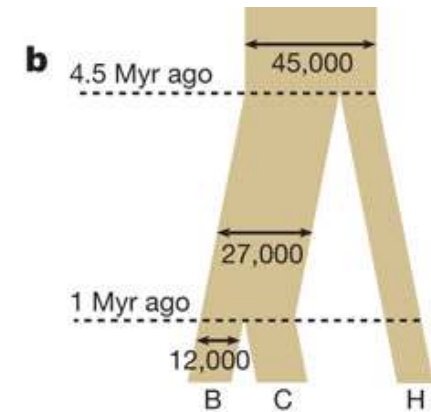
Red boxes indicate significantly large effect size variants



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

[*Callorhynchus 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]) (order [Sarcopterygii](#))

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

[*Takifugu rubripes*](#), a [puffer fish](#) (^[17] International Fugu Genome Consortium^[18] 2002^[19]) (order [Tetraodontiformes](#))

[*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



COMPUTATIONAL GENOMICS
ANALYSIS AND TRAINING

