Comparisons of Mammalian Genomes

Chris Ponting

MRC Functional Genomics Unit

University of Oxford

MRC Functional Genomics Unit

“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 in 1986, speaking in 1998.

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

News

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.

Dr. Thomas H. Roderick

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?**

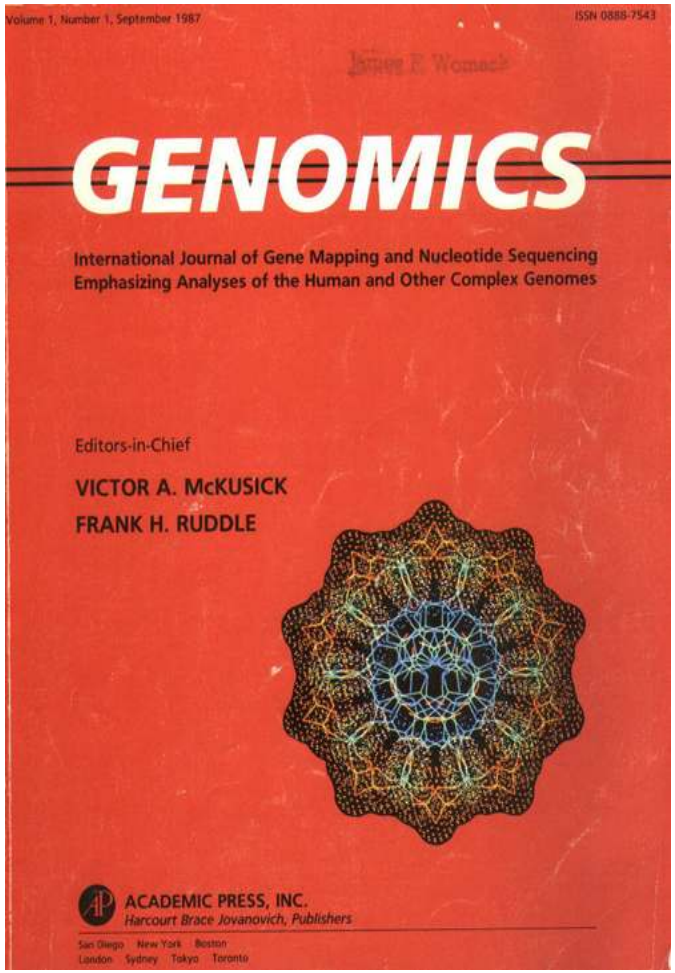
Roderick: Well, it certainly encompassed what the journal wanted to cover. It encompassed sequencing, mapping, and new technologies. But we felt it also had the comparative aspect of genomes of various species, their evolution, and how they related to each other. Although we didn't come up with the term "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



From Protein Domains to Genomes

October 1999

1998

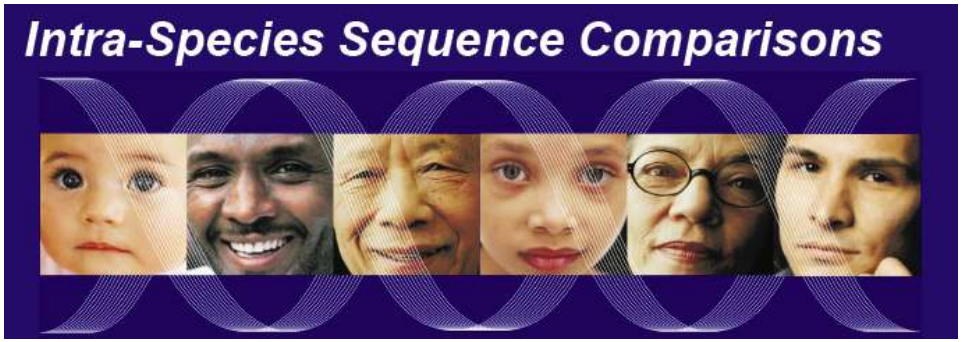
1998 | 1999 | 2006 | 2007 | 2008 | 2009 | 2010

Era of Genomes and Evolution

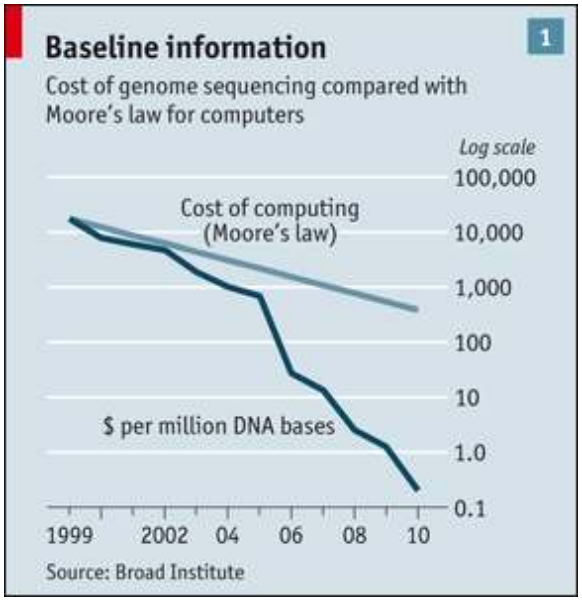
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010

Era of Population Genomics

2011 2012 2013 2014 2015 2016 2017 2018 2019 2020



the first human genome cost ~ $3b



Economist Jun 17th 2010

a genome every 4 days

Summary

Part 1.

- Human Genome Project
- Mouse Genome & Comparisons with Human

Part 2.

- The functional portion of the genome
- Transcript maps

Part 3.

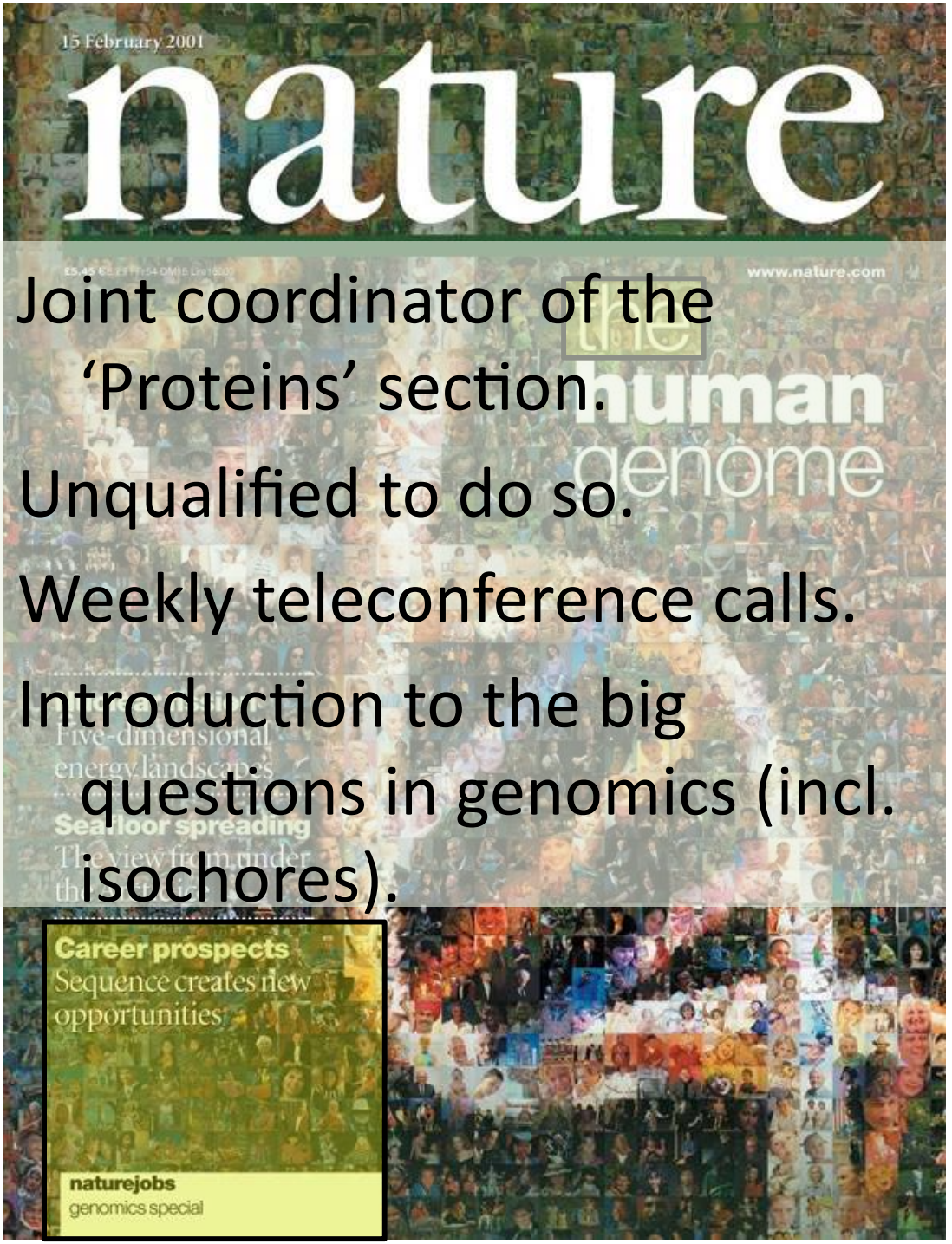
- Genome evolution
- Future

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.

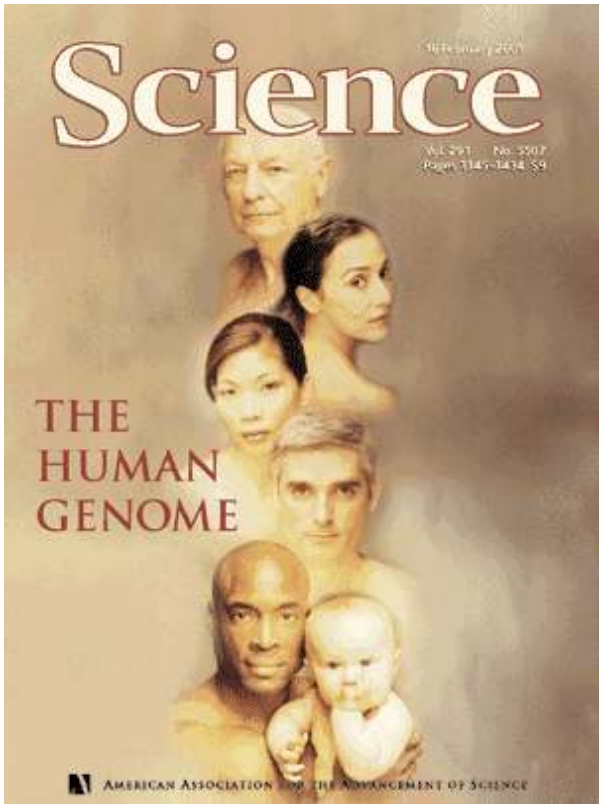
Introduction to the big questions in genomics (incl. isochores).

articles

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

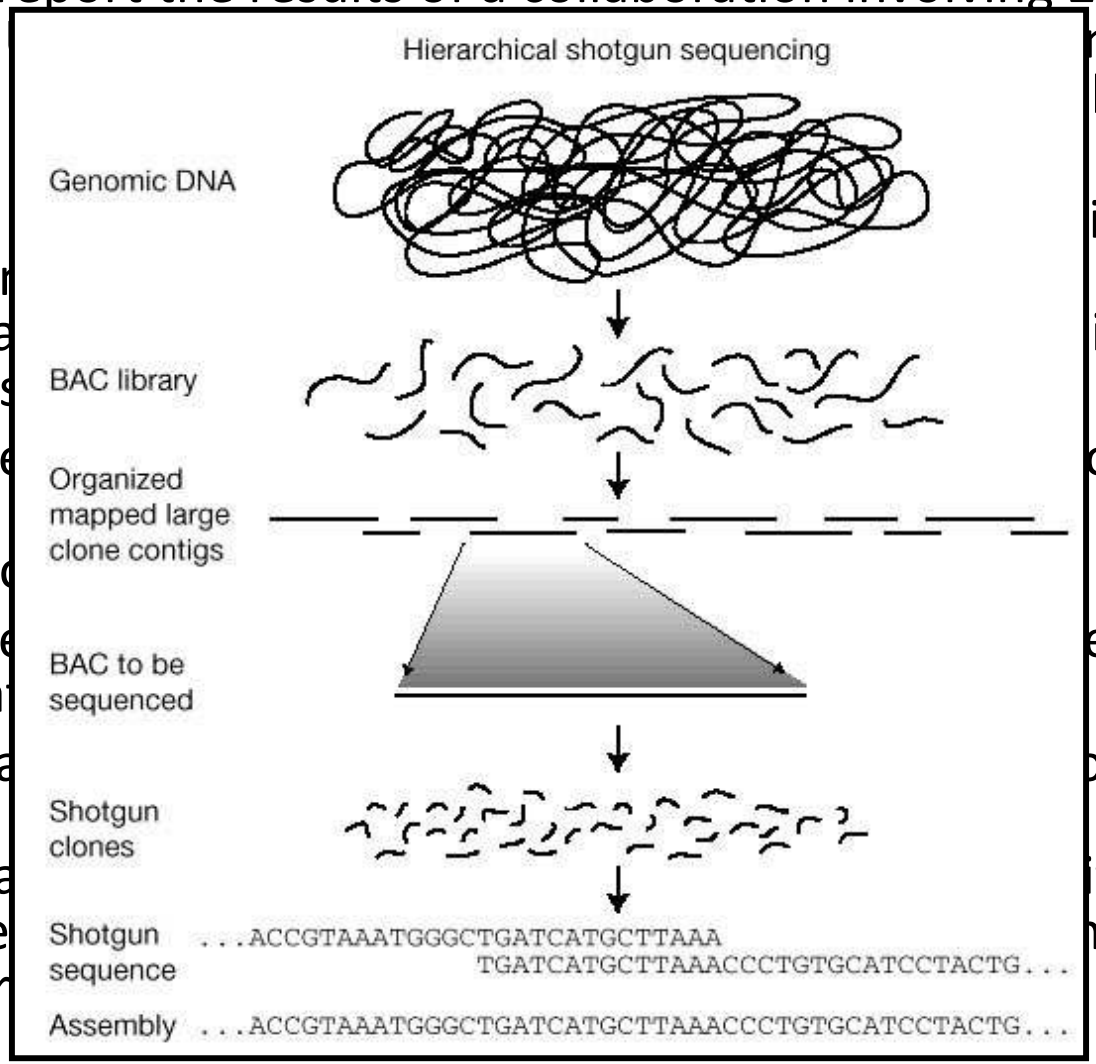
* *A partial list of authors appears on the opposite page. Affiliations are listed at the end of the pa*

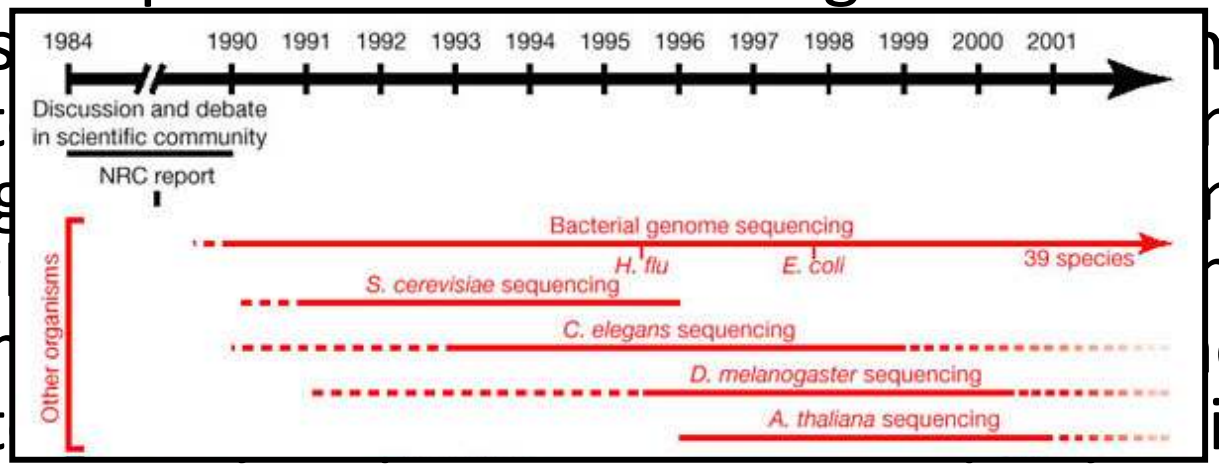


The Sequence of the Human Genome

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]

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



- The sequence of the human genome is of interest in s 

me to be ext mes as larg ne and eig n ger e 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

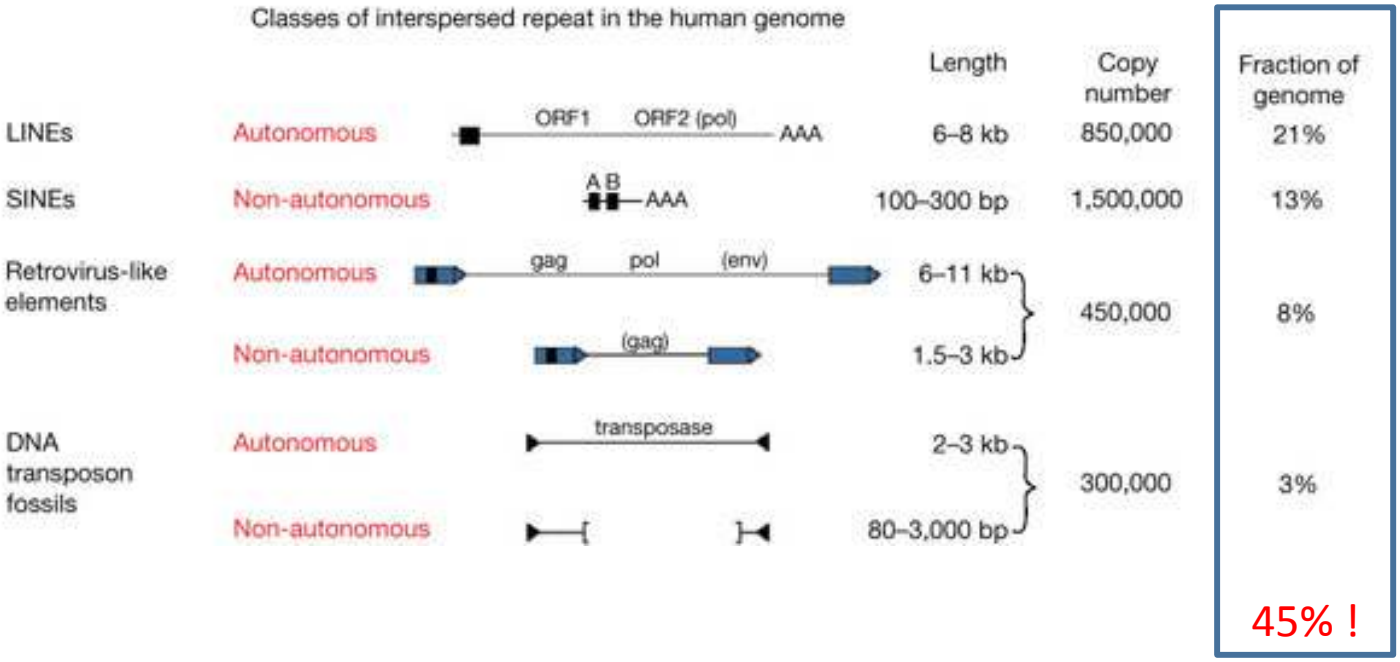


FIGURE 17. Almost all transposable elements in mammals fall into one of four classes.

Isochores

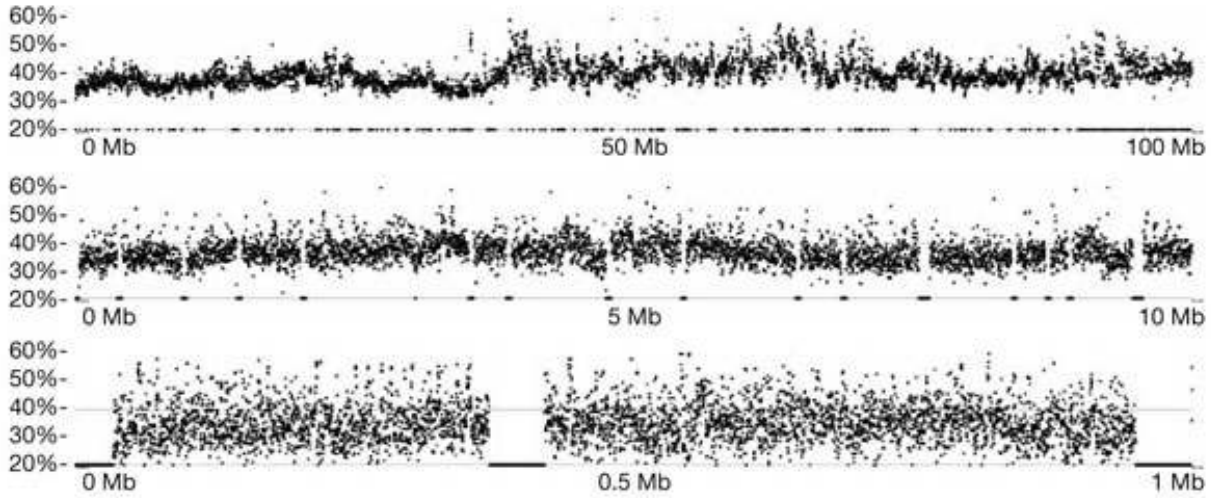
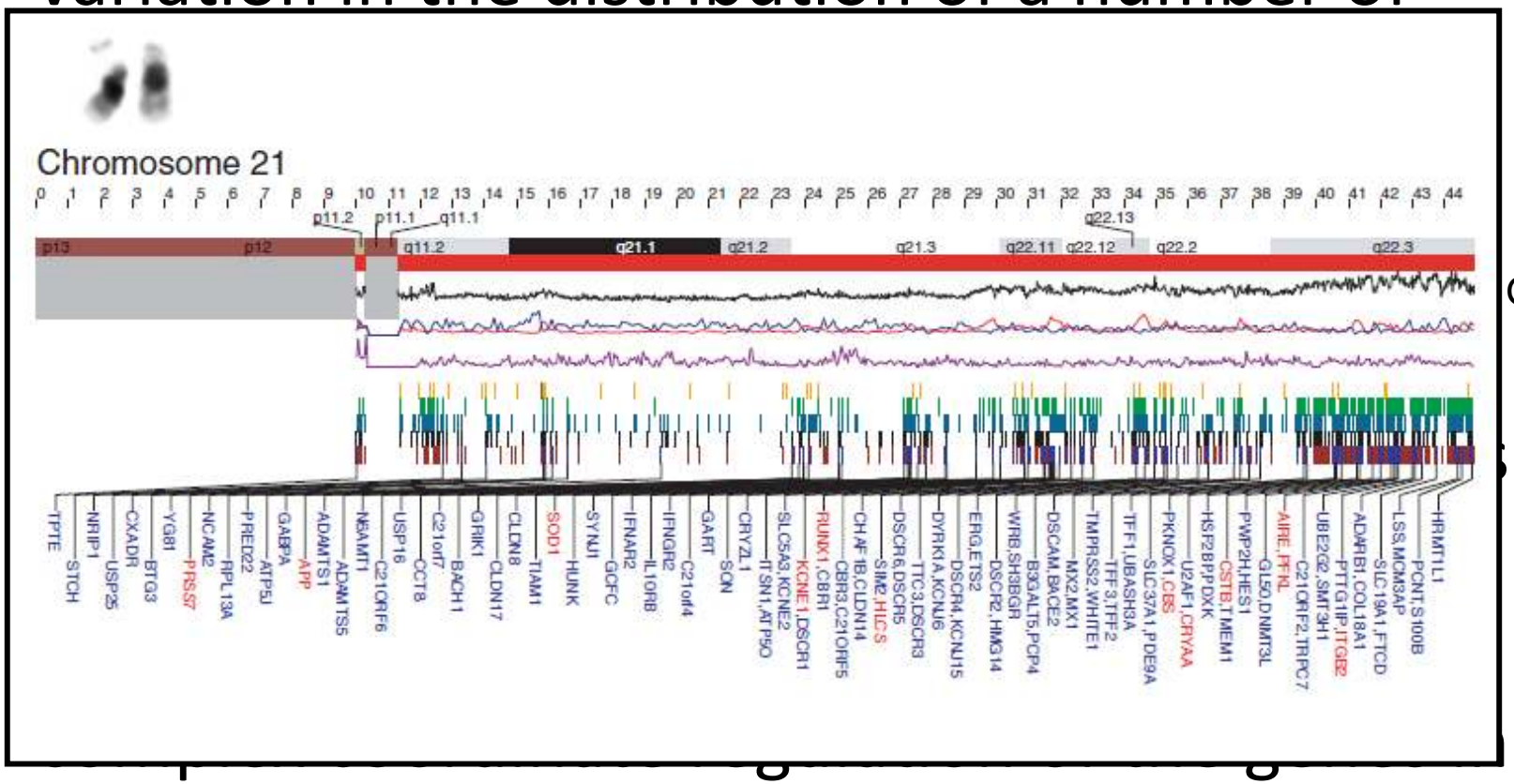


FIGURE 13. Variation in GC content at various scales

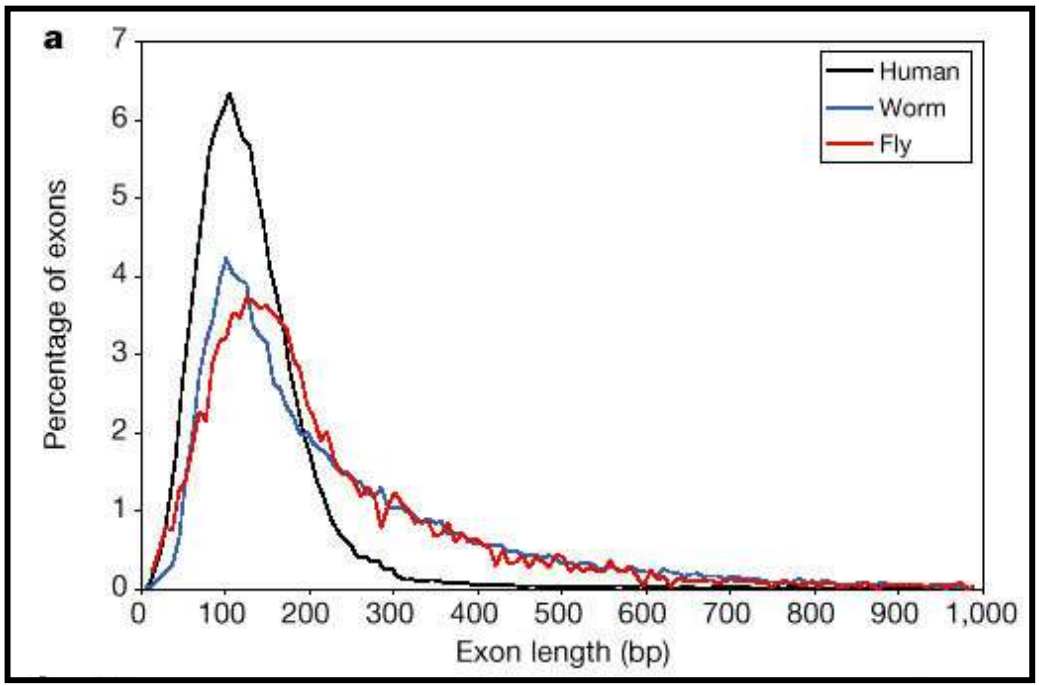
Bernardi et al. described the genome as being composed of a mosaic of compositionally homogeneous regions dubbed ‘isochores’.

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

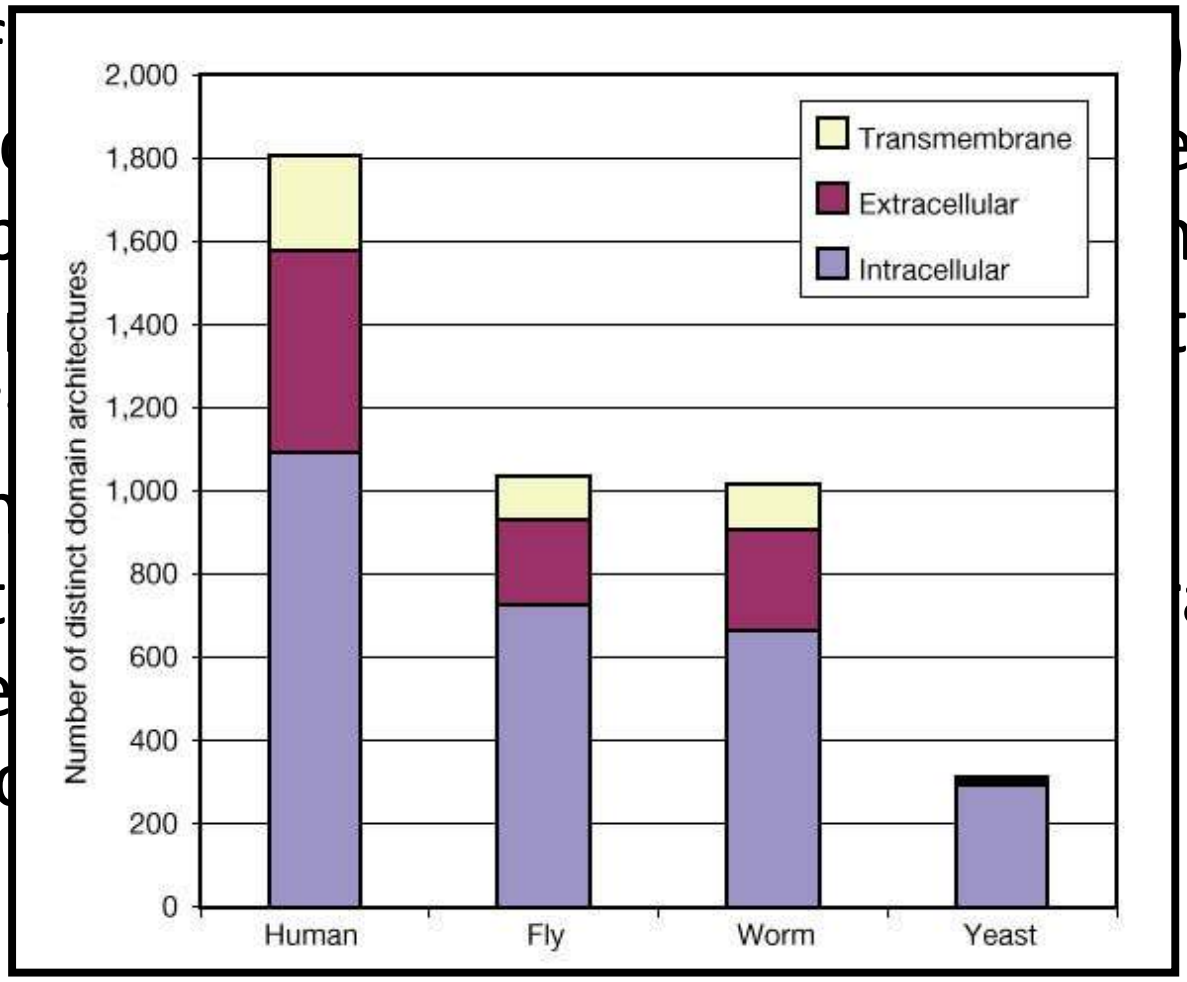


the clusters.

- 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.

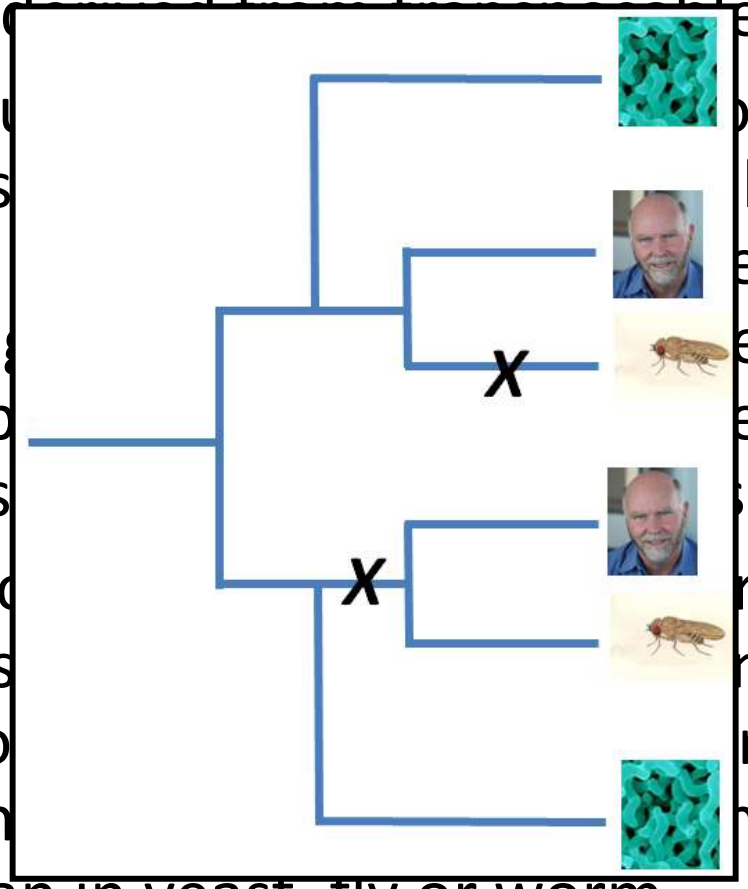


- The f
enco
comp
due i
speci
estim
fact t
pre-e
colle

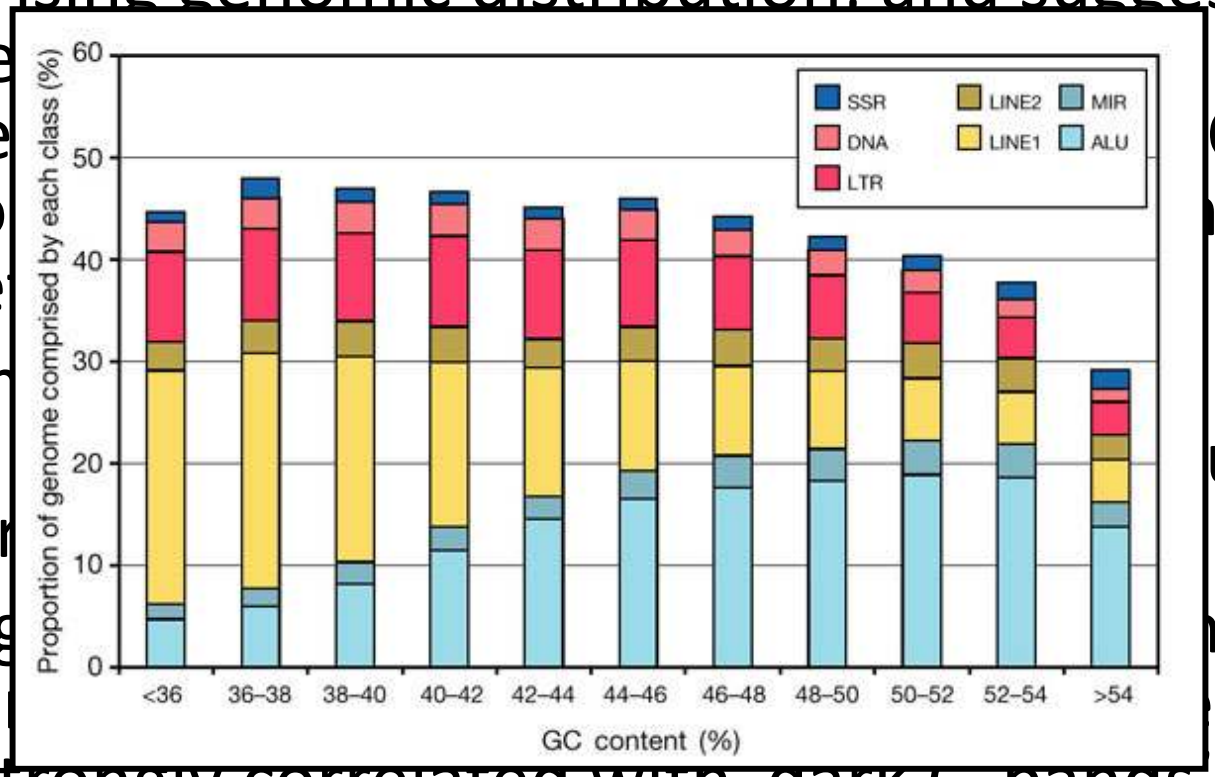


is is
e-
the
anged

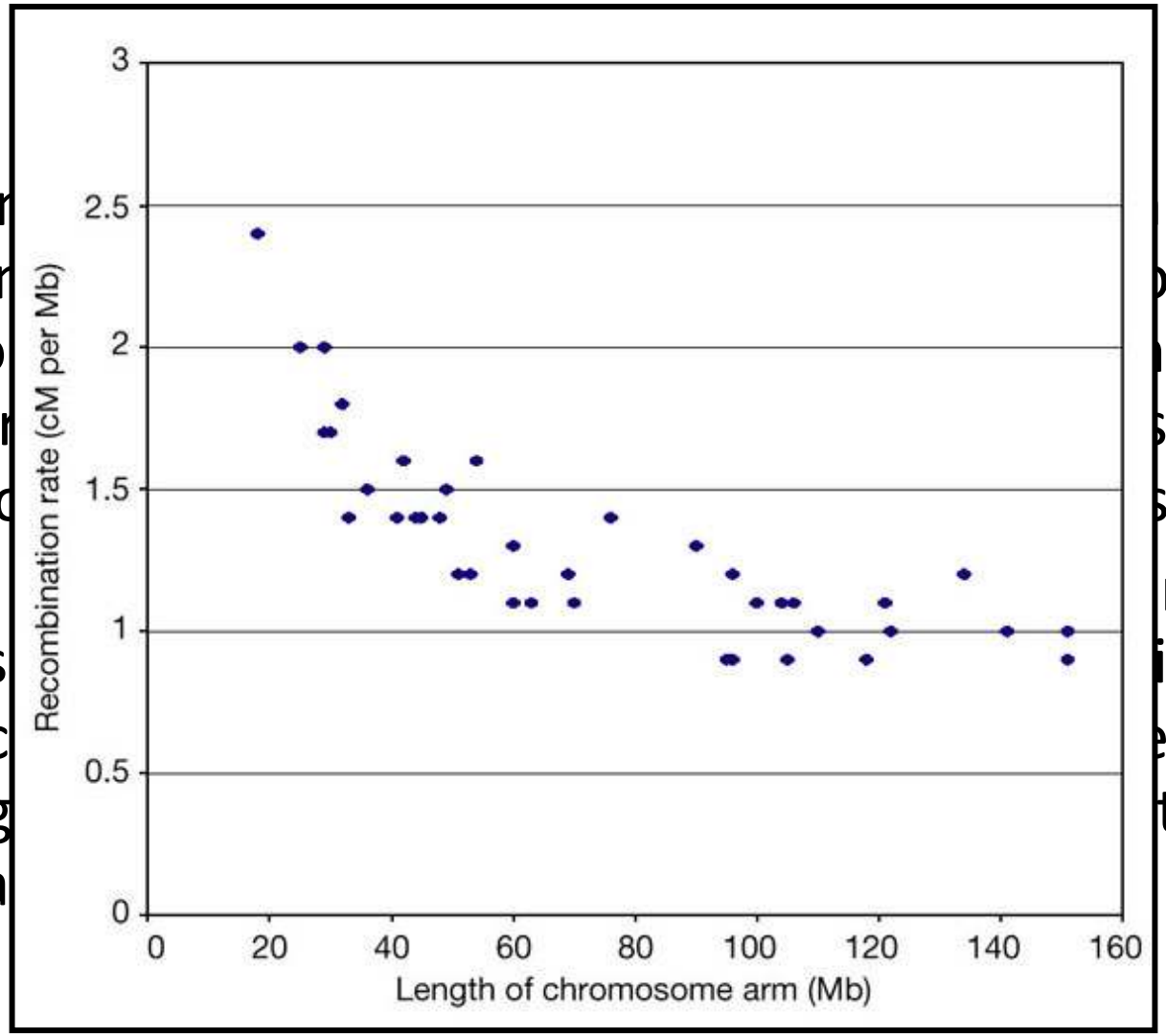
- Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been d[illegible] elements.
- Although abou[illegible]me derives from transpos[illegible] been a marked decline in the [illegible]ements in the hominid linea[illegible]ear to have become comp[illegible]erminal repeat (LTR) retropos[illegible]so.
- The pericentro[illegible] regions of chromosomes[illegible]nt segmental duplications o[illegible]re in the genome. Segm[illegible] 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
- The r
- Cytog



- Recor distal regior osomes and o a patter st one cross s.
- More rphisms (SNPs ied. This collec e-wide linkag the huma



2004

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

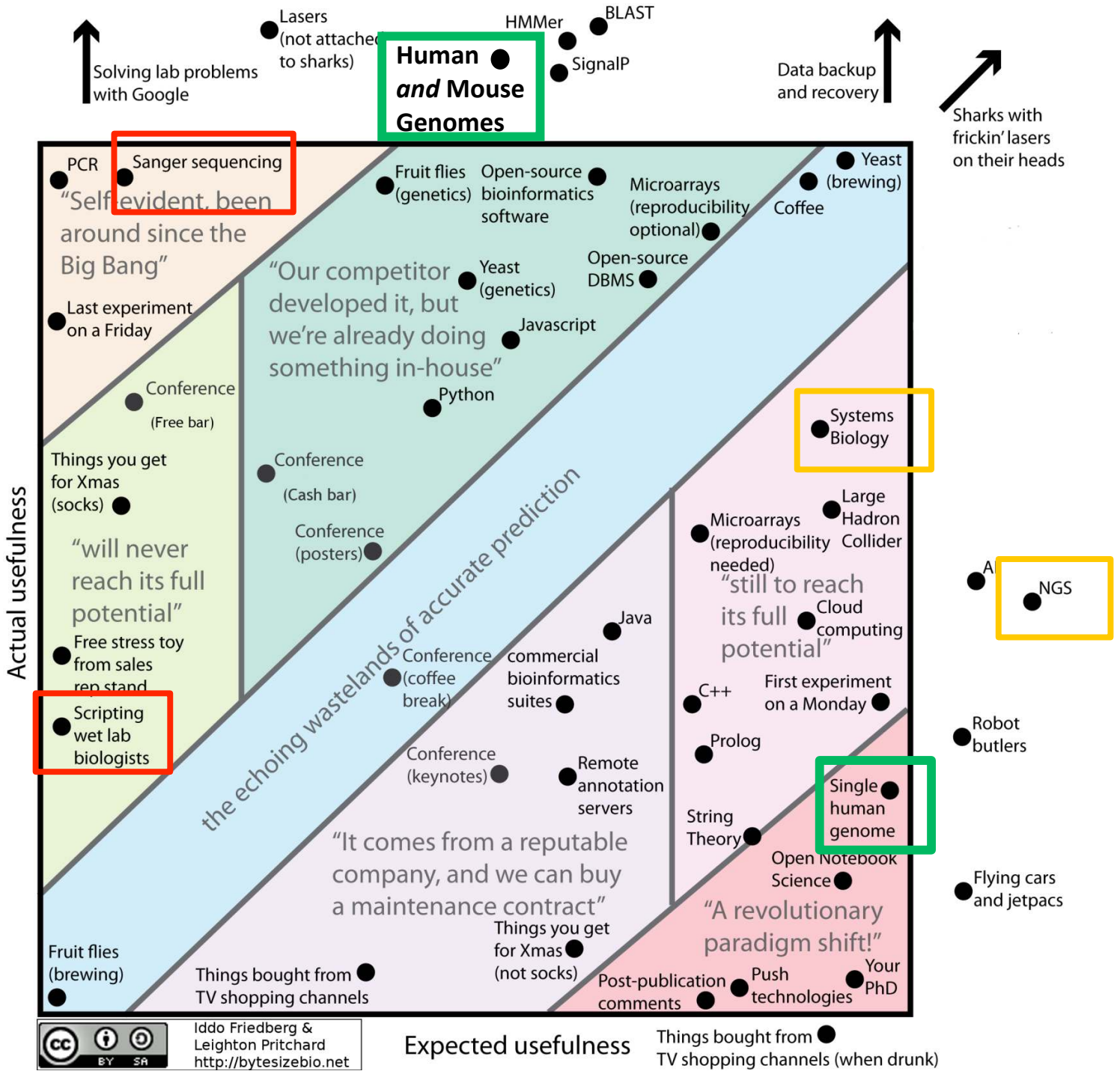
* A list of authors and their affiliations appears in the Supplementary Information

- "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."

The human genome: More questions than answers

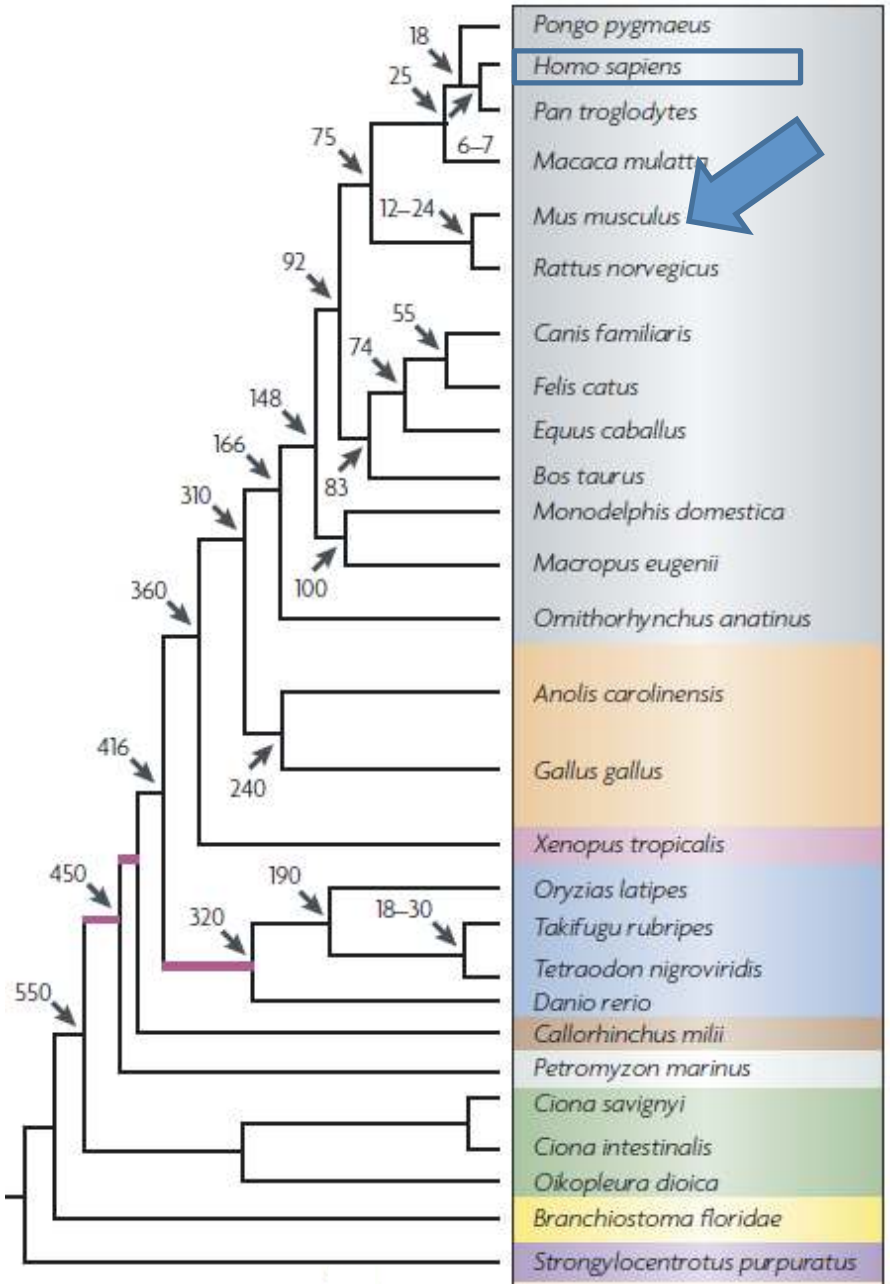
- How many genes? How much functional DNA?
- Are transposon-derived sequences functional?
- How did the heterogeneity in GC ('isochores') arise and how is it sustained?
- How is recombination controlled?
- How does the human genome, and its genes, differ from those of more closely-related genomes? Is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

How Useful is a Human Genome?



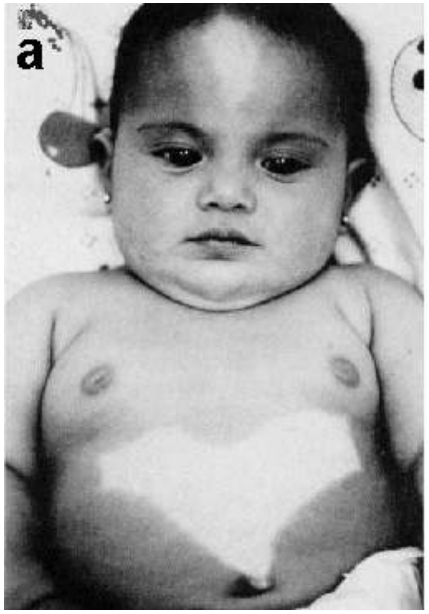
Evolutionary yardsticks

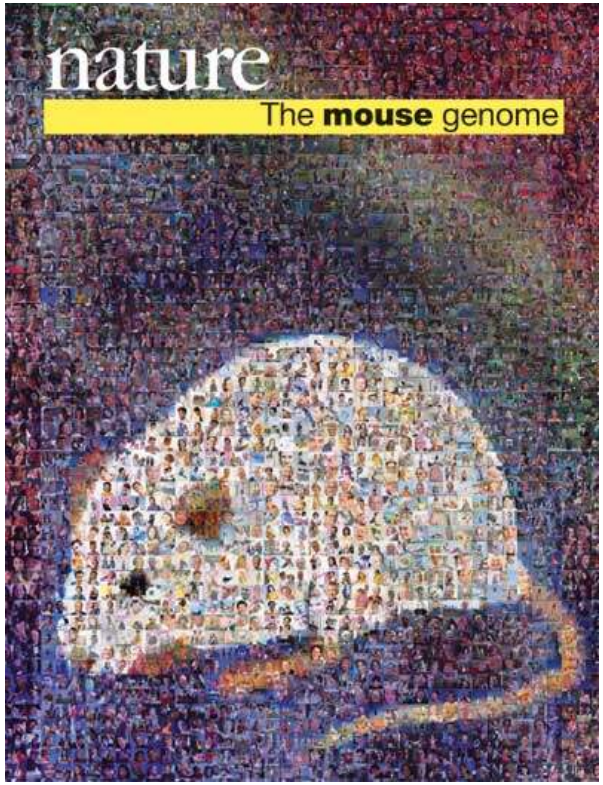
mammalian & invertebrate noncoding sequences do *not* align



c 2008

Mouse vs Human



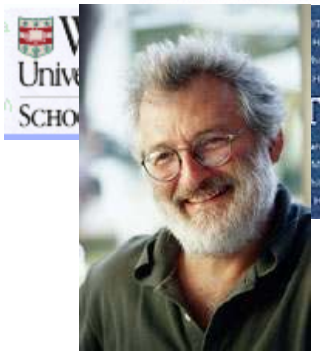


Human Genome Sequencing Center
Baylor College of Medicine

Genome Sequencing Center

Sequencing

Assembly



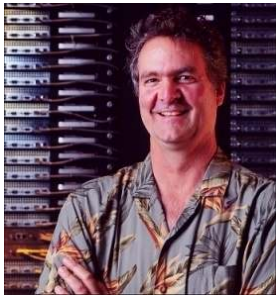
e! project Ensembl

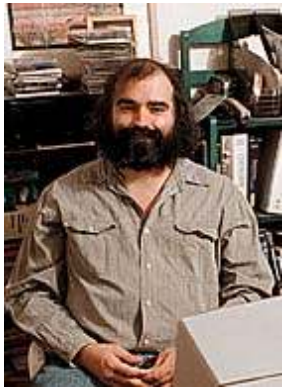
Transposon sequences

Gene Prediction

Genome Comparison

Gene Comparison





Orthologues and Paralogues

Cenancestor

SP1

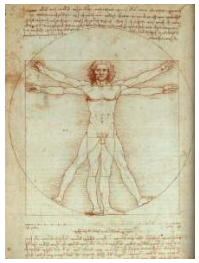
SP2

DP2

A1

B1

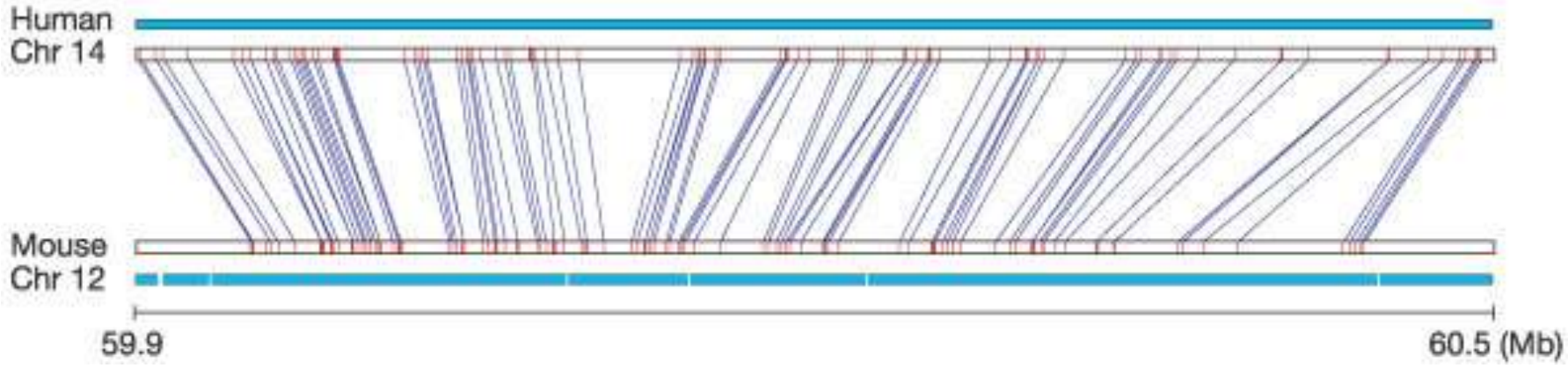
C1 *C2*



C1 and C2 are paralogues

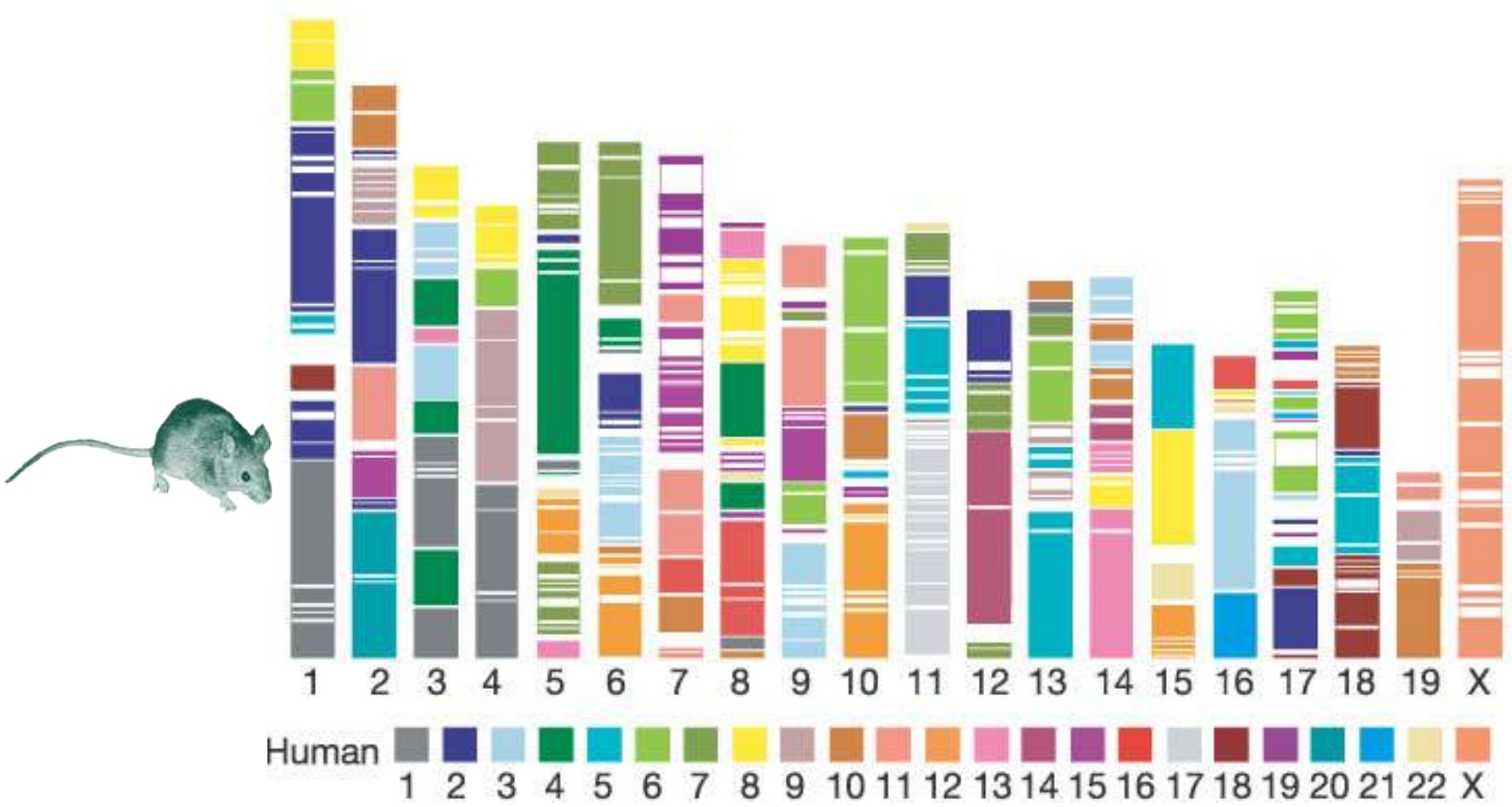
A1 and B1 and (C1 and C2) are orthologues

Human and mouse
“local synteny”

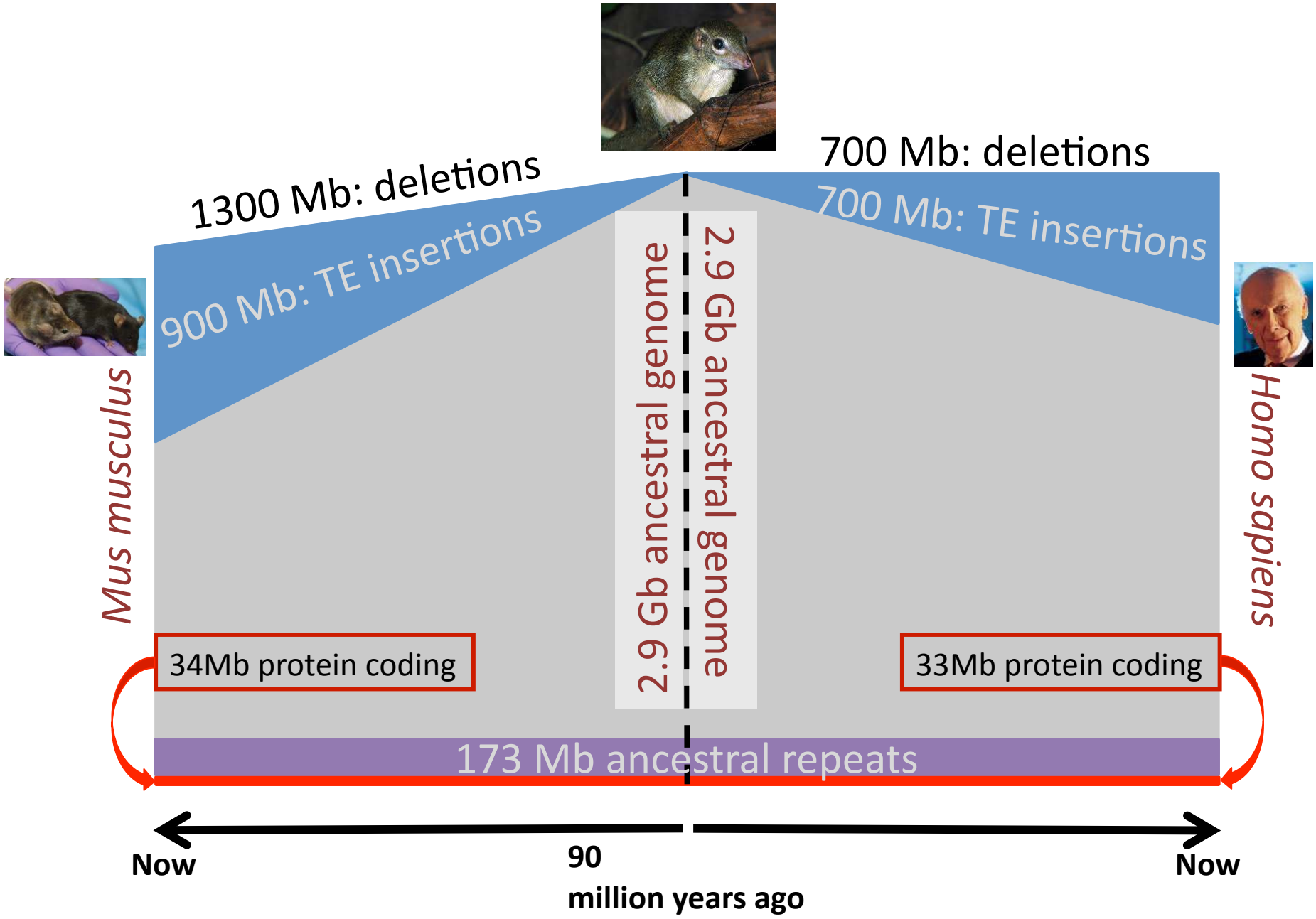


“Syntenic” regions contain orthologues!

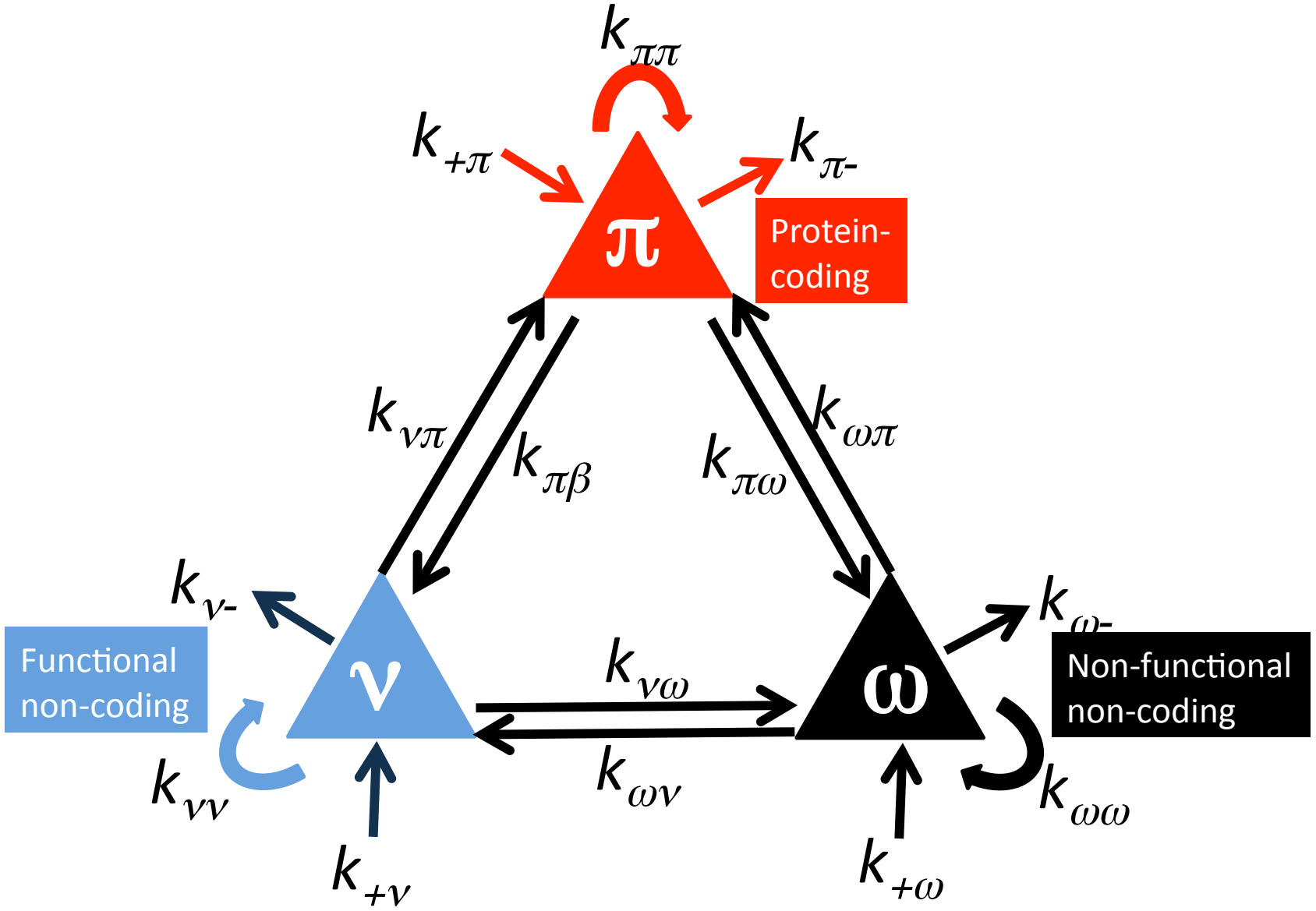
Human and mouse chromosomes: global orthology



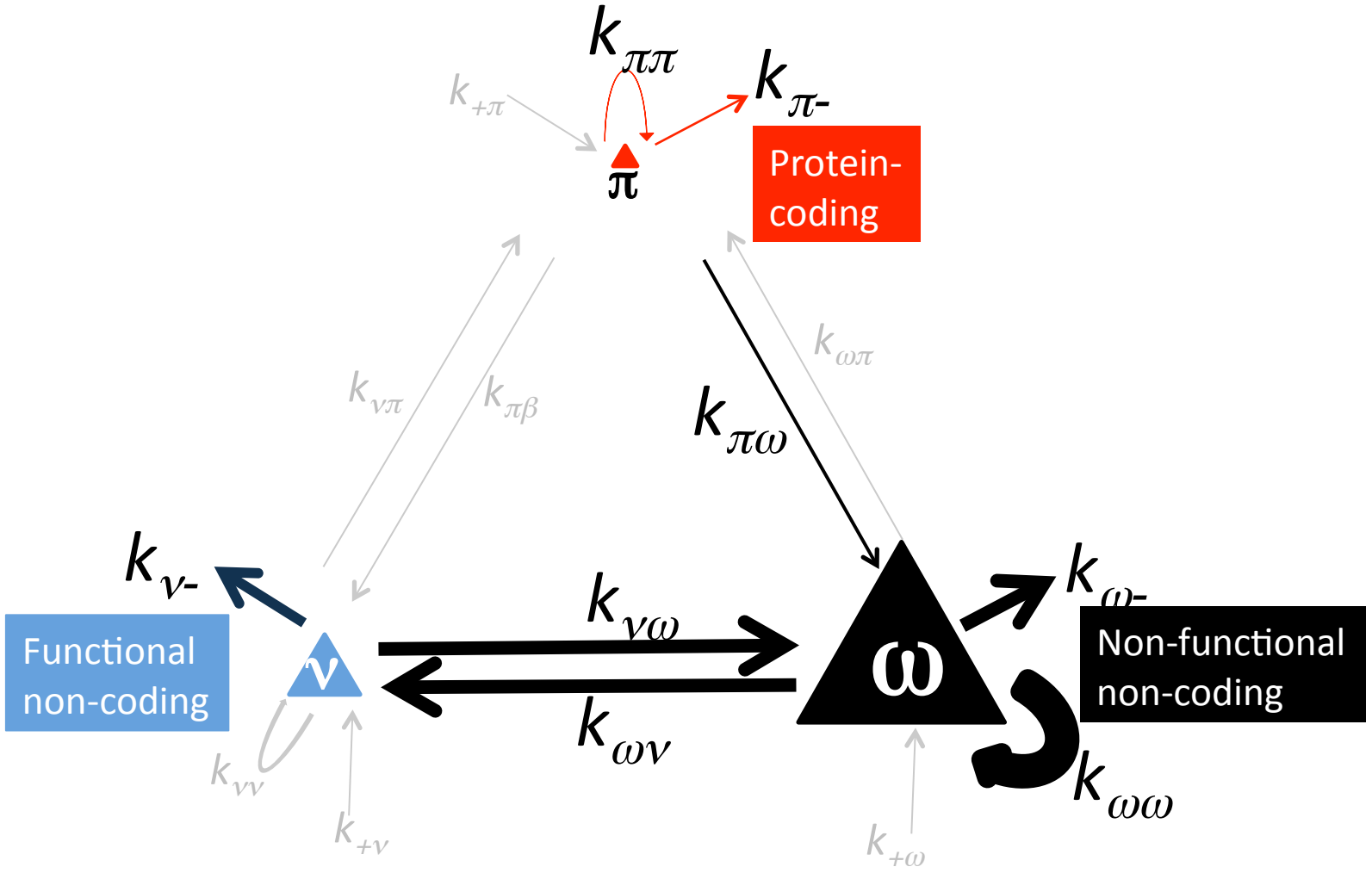
Only 40% of the human genome aligns to the mouse genome.


1300 Mb: deletions
900 Mb: TE insertions
700 Mb: deletions
700 Mb: TE insertions
2.9 Gb ancestral genome
2.9 Gb ancestral genome
Mus musculus
Homo sapiens
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



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.

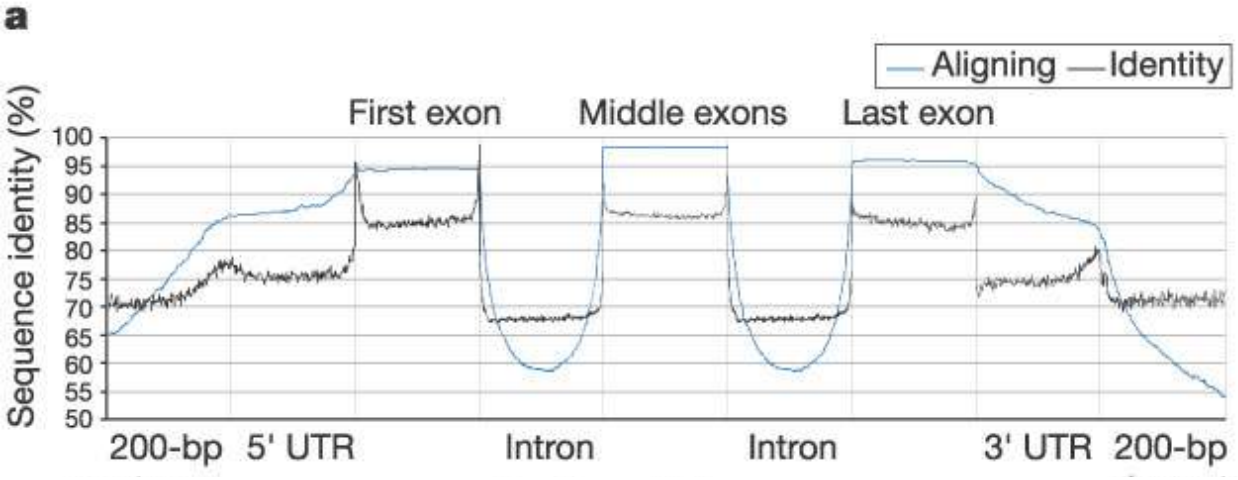
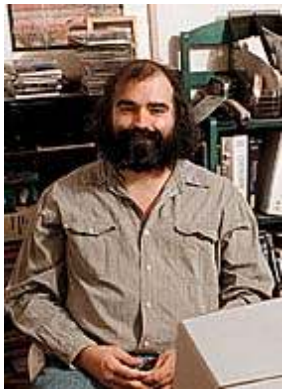
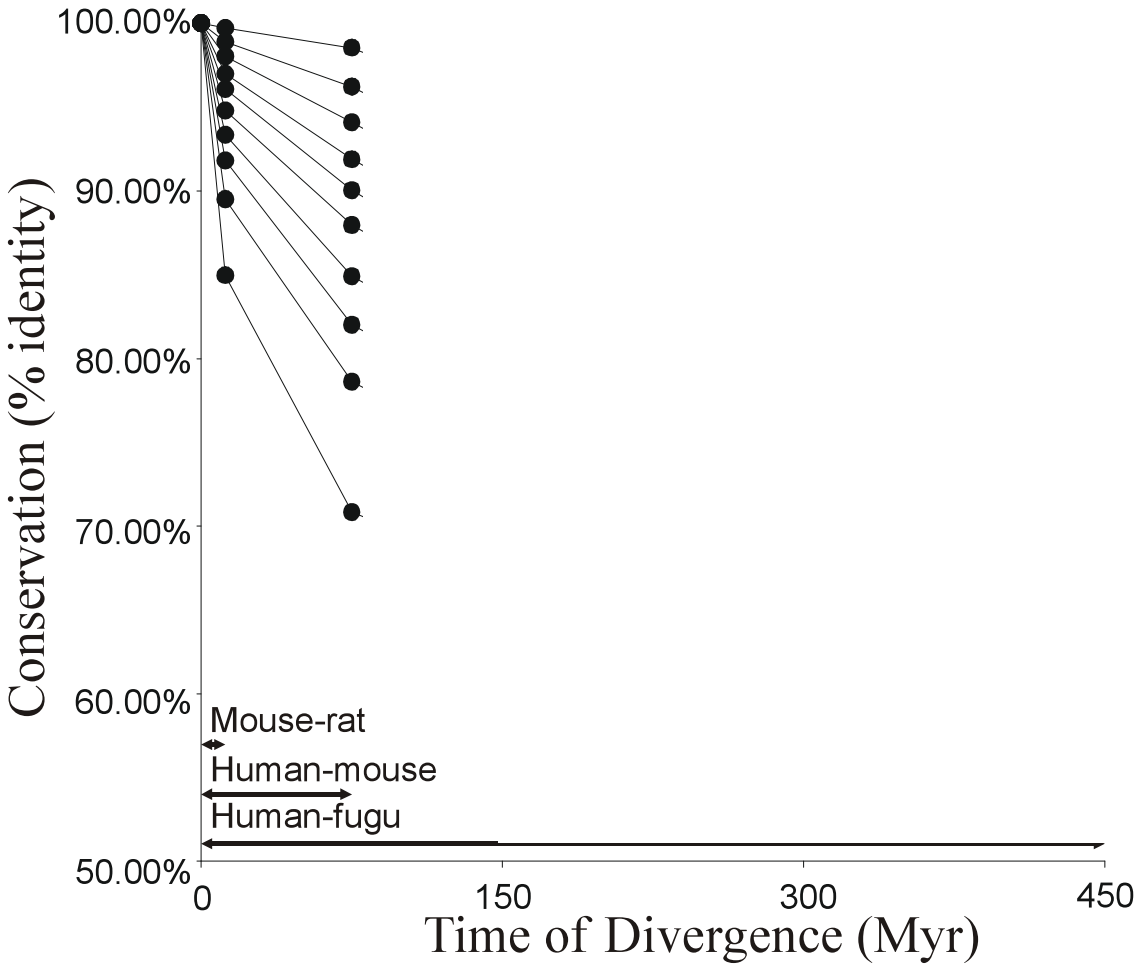


Figure 25. Sequence conservation between mouse and human genes
Mouse genome paper *Nature* 420, 520-562

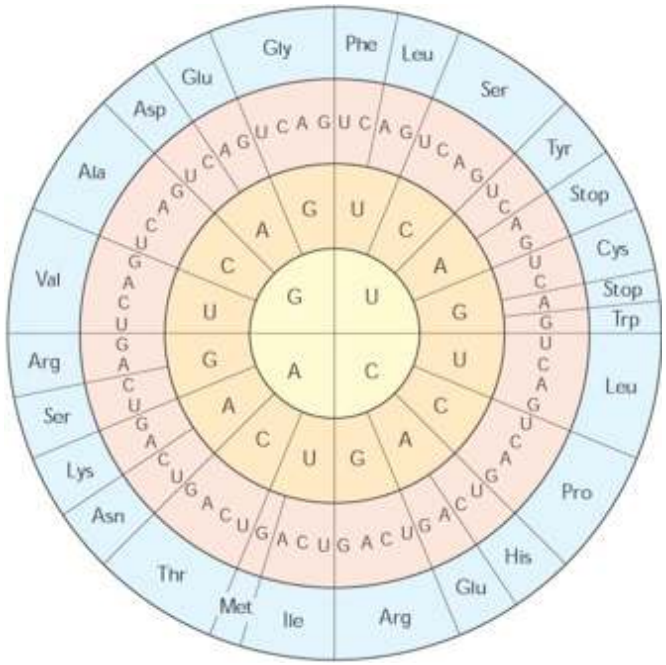


Gene Sequence Conservation is Clock-like



A model of neutral evolution

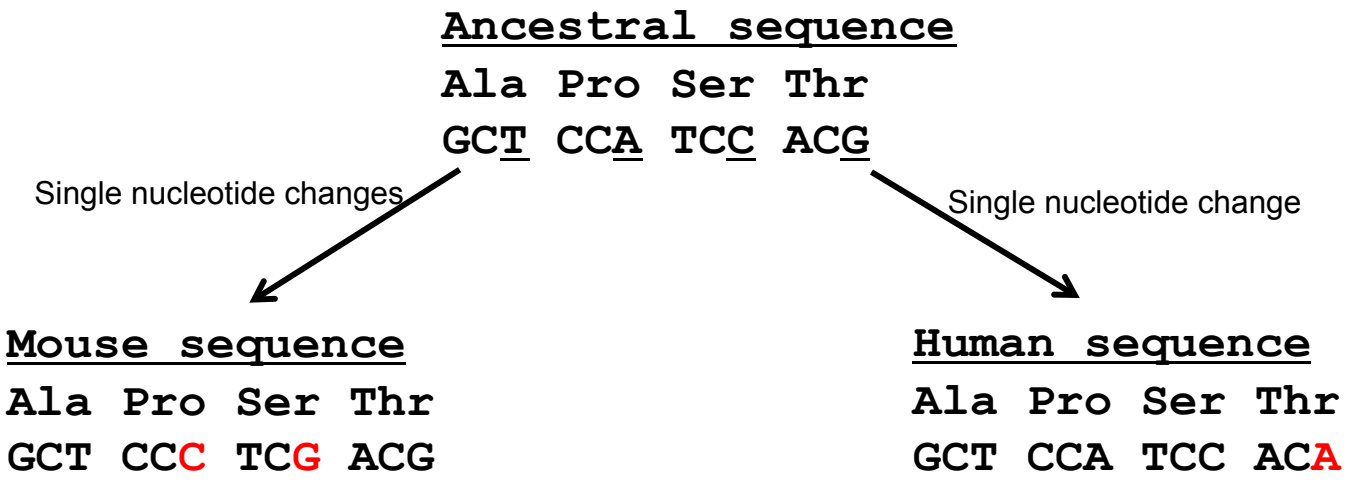
- d_S – the number of synonymous substitutions per synonymous site
- takes advantage of the redundant genetic code
- 4D sites GCx (ALA), CCx (PRO), TCx (SER), ACx (THR), CGx (ARG), GGx (GLY), CTx (LEU), GTx (VAL)
- "how much would a gene's sequence have changed if selection had not acted upon it?"



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

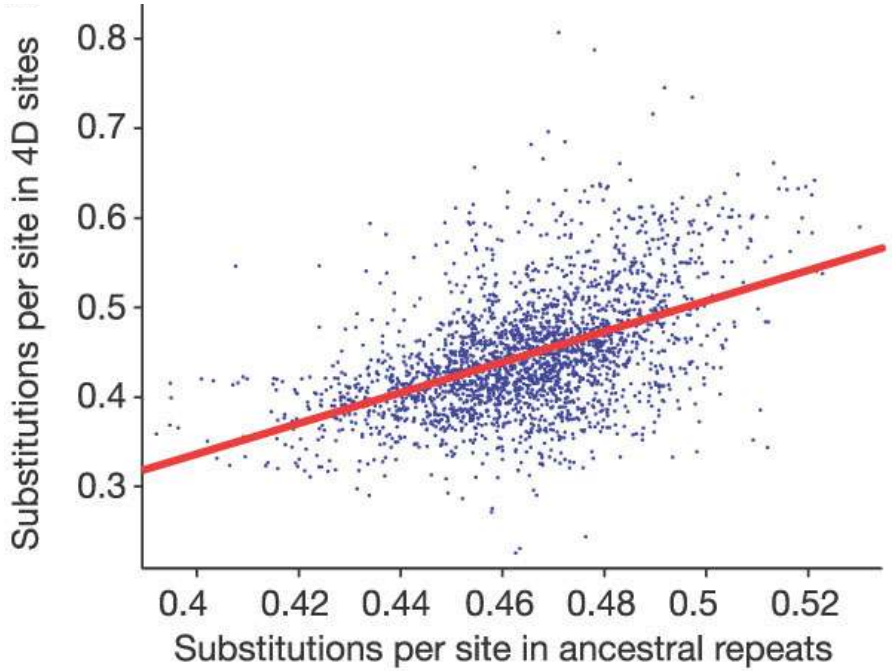
d_S = ¾ or 0.75

“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

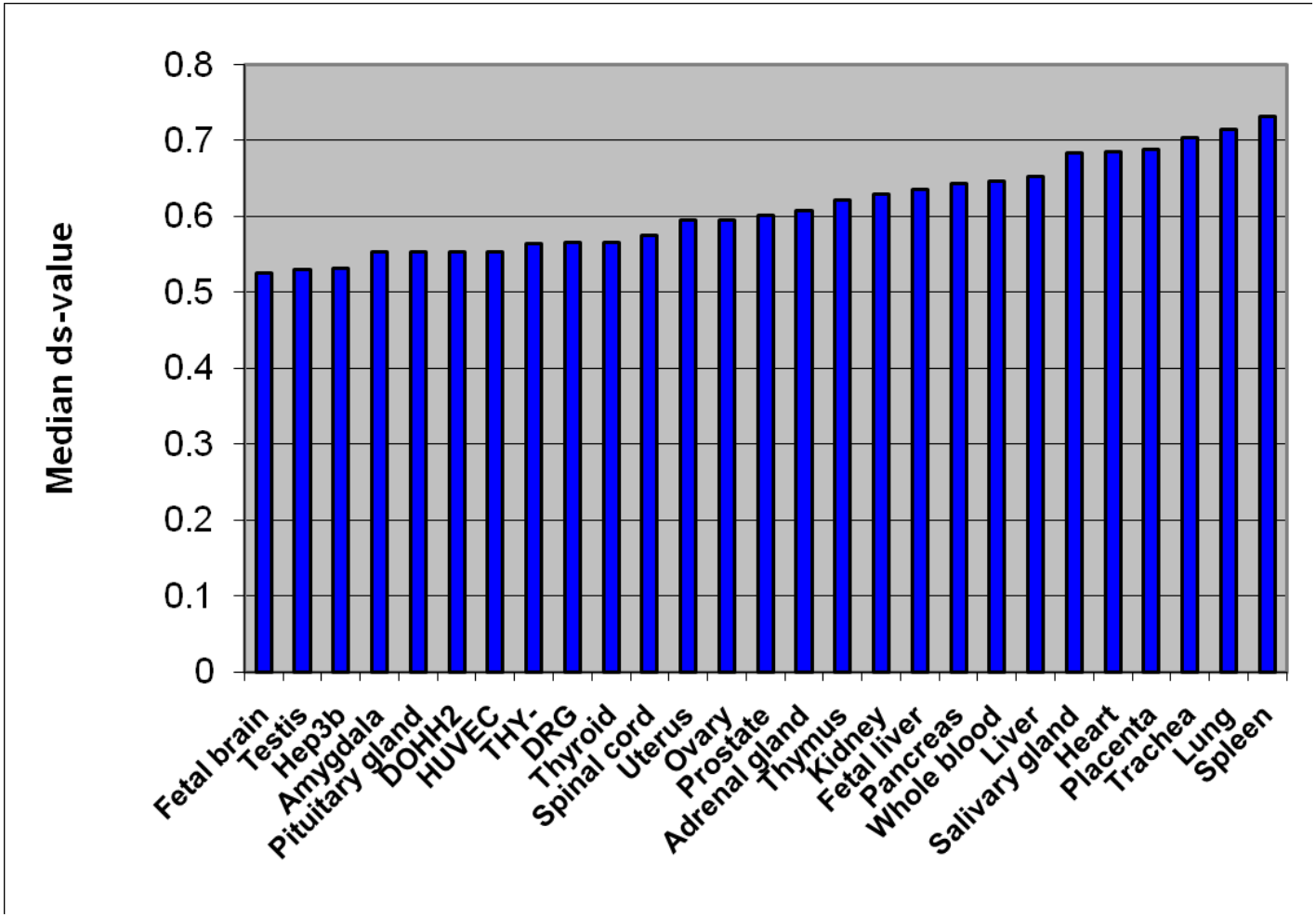


see also
Hardison et al.
Genome Res. 2003
13: 13-26.

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

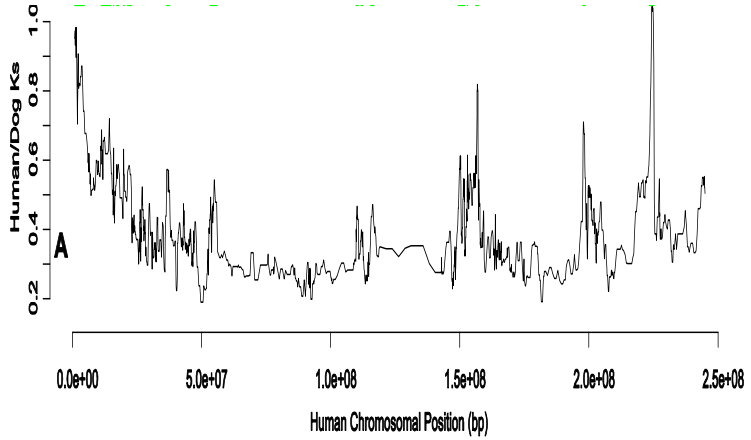
- Transcription-associated mutational strand asymmetry (Phil Green et al. Nature Genetics 33: 514-7)
- Associated with transcription-coupled repair processes (Majewski, Am J Human Genet 73, 688-692)
- Genes transcribed in the germline at high levels, when mutated, are repaired more readily, than those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

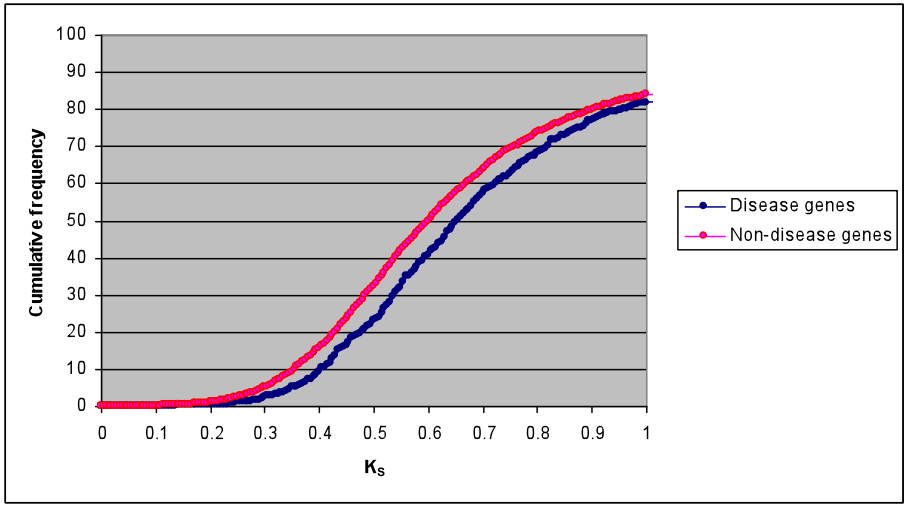
Tissue-specific genes' d_s



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

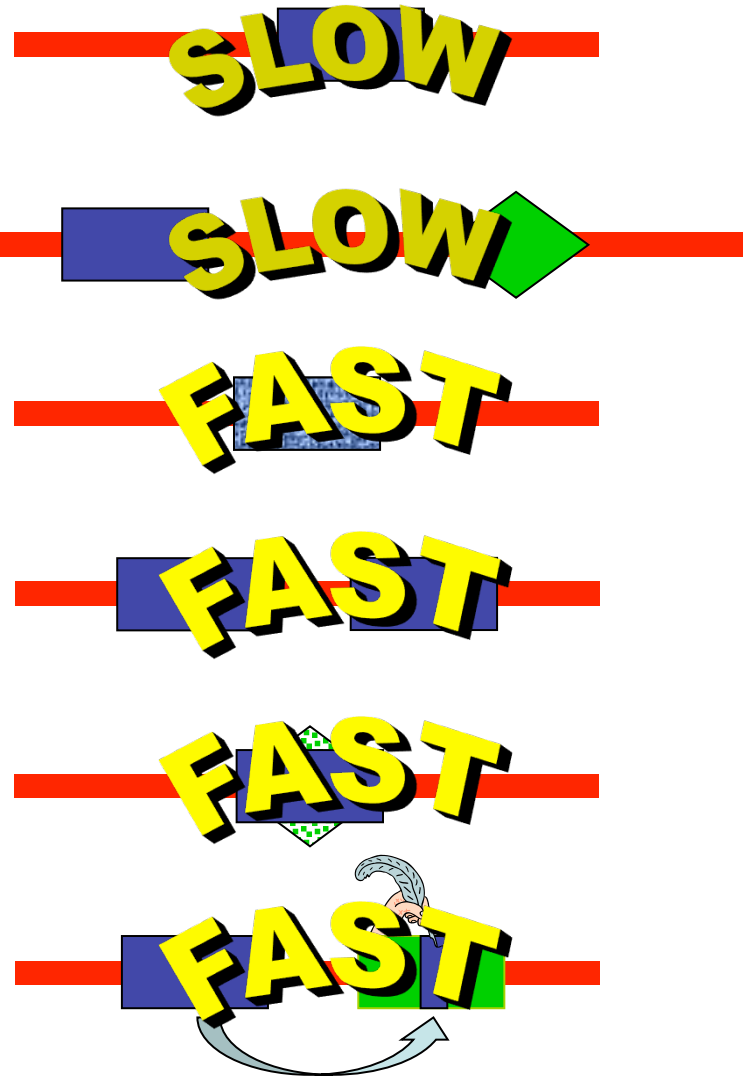
d_S variation





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

Positive Selection, dN/dS > 1

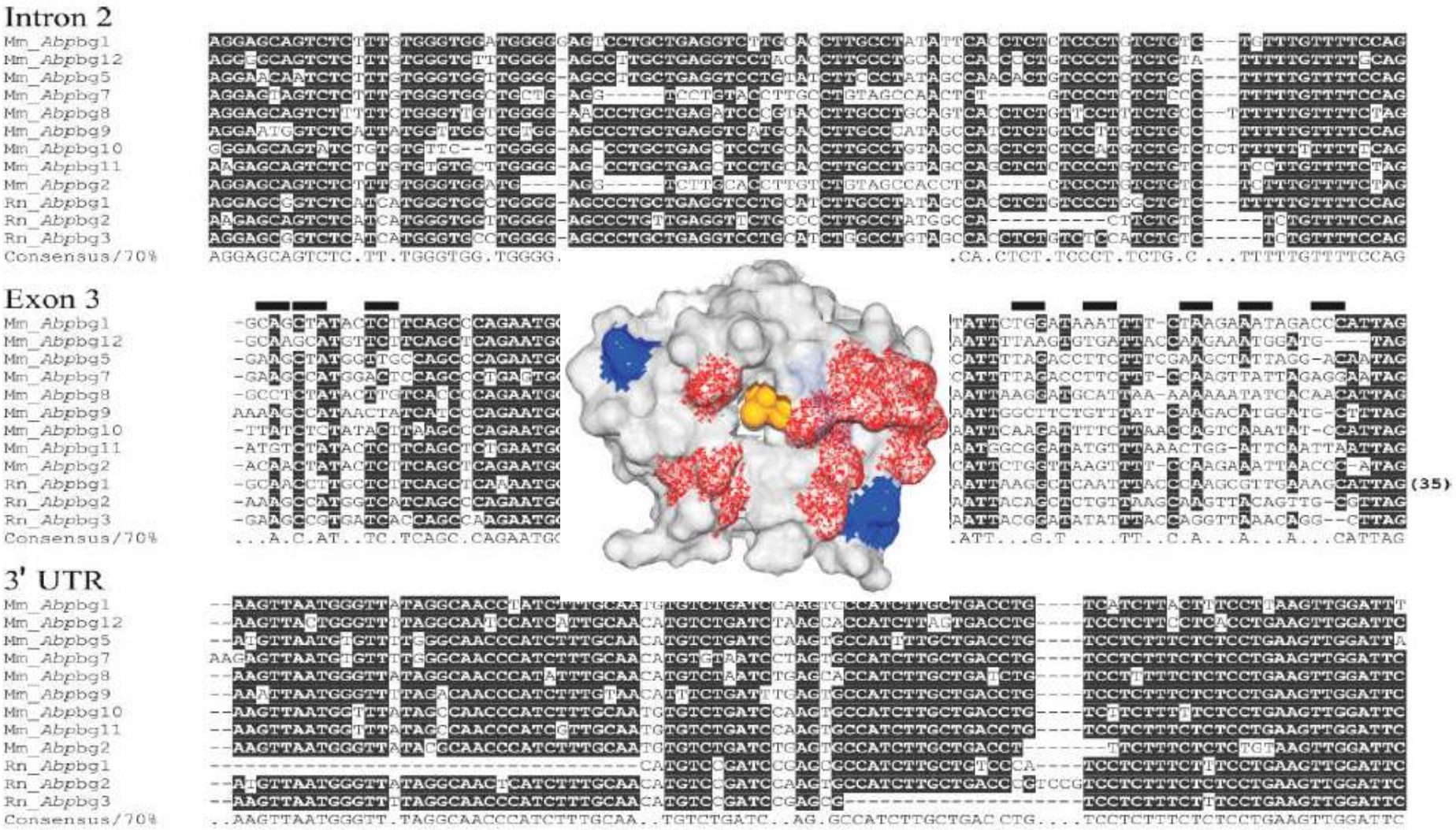
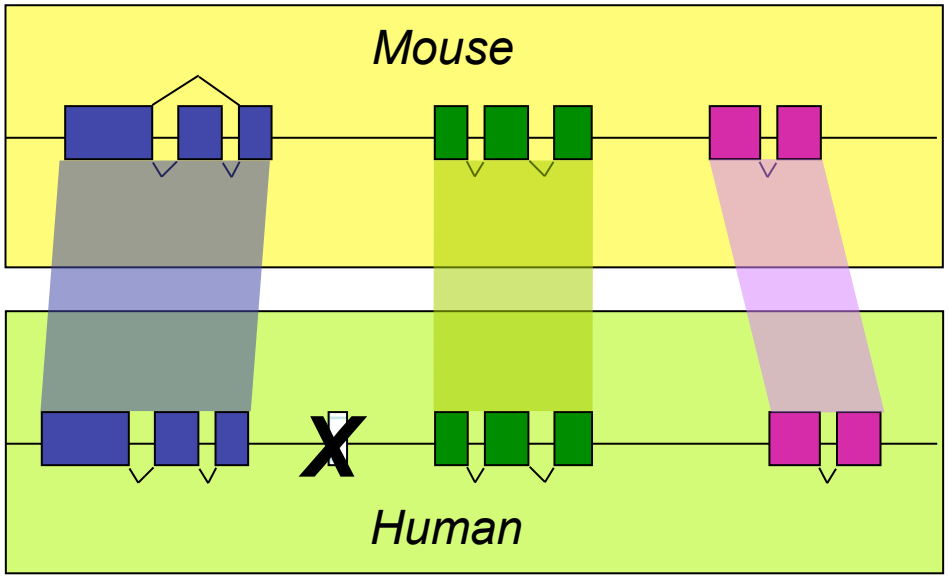


Figure 4 Multiple nucleotide sequence alignment of mouse and rat *Abpbg*-like exons 3 and surrounding genomic DNA. Genomic DNA corresponding to exon 3 (98 positions) and 100 nucleotide positions of both flanking intronic and 3′-UTR sequence was aligned with HMMER, and manually adjusted. We found that 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

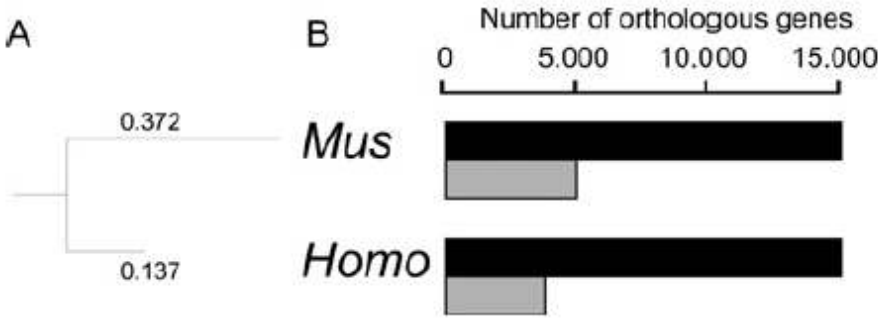
the mouse **genome**

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 the human genome itself. But why? The laboratory mouse is hailed

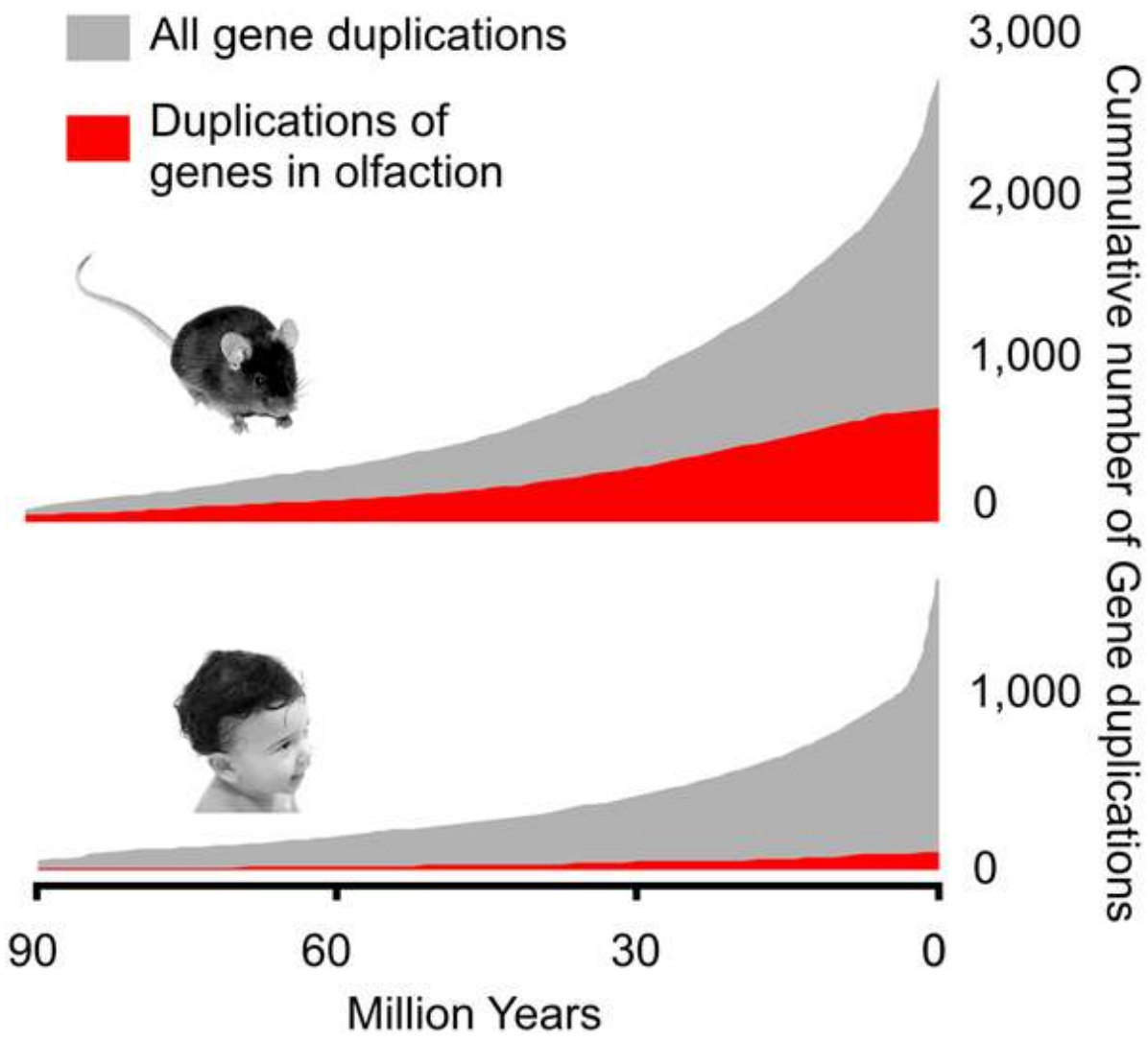
Figure 2 | Mouse genes have a higher synonymous nucleotide substitution rate (d_S) and have accumulated more lineage-specific duplicates than human genes

(**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).

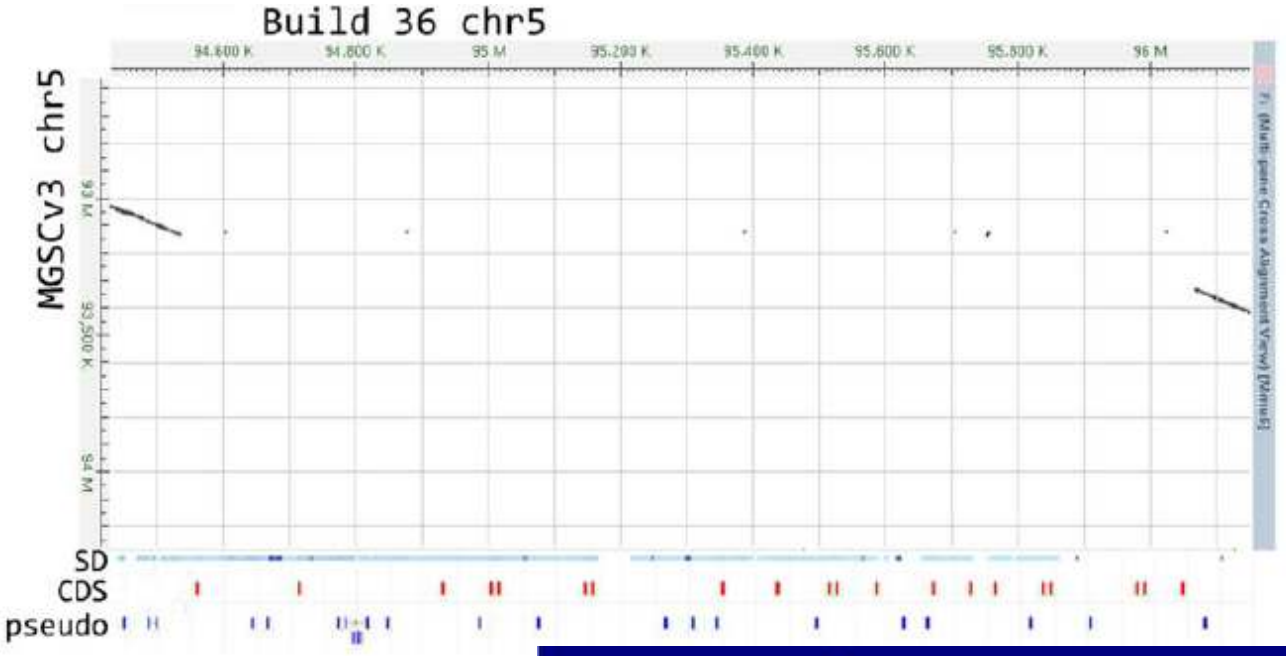


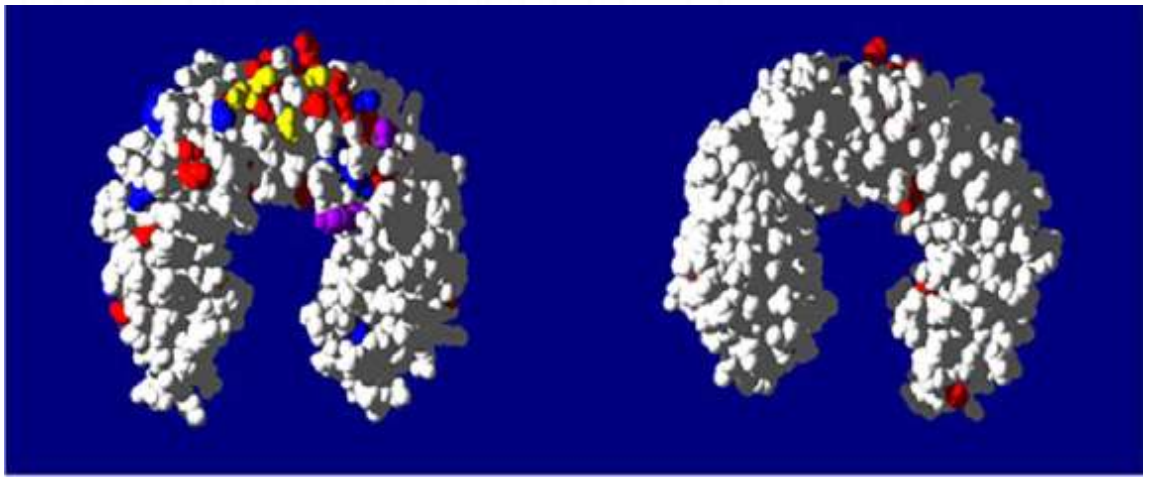
Ritu Dhand Chief Biology Editor

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

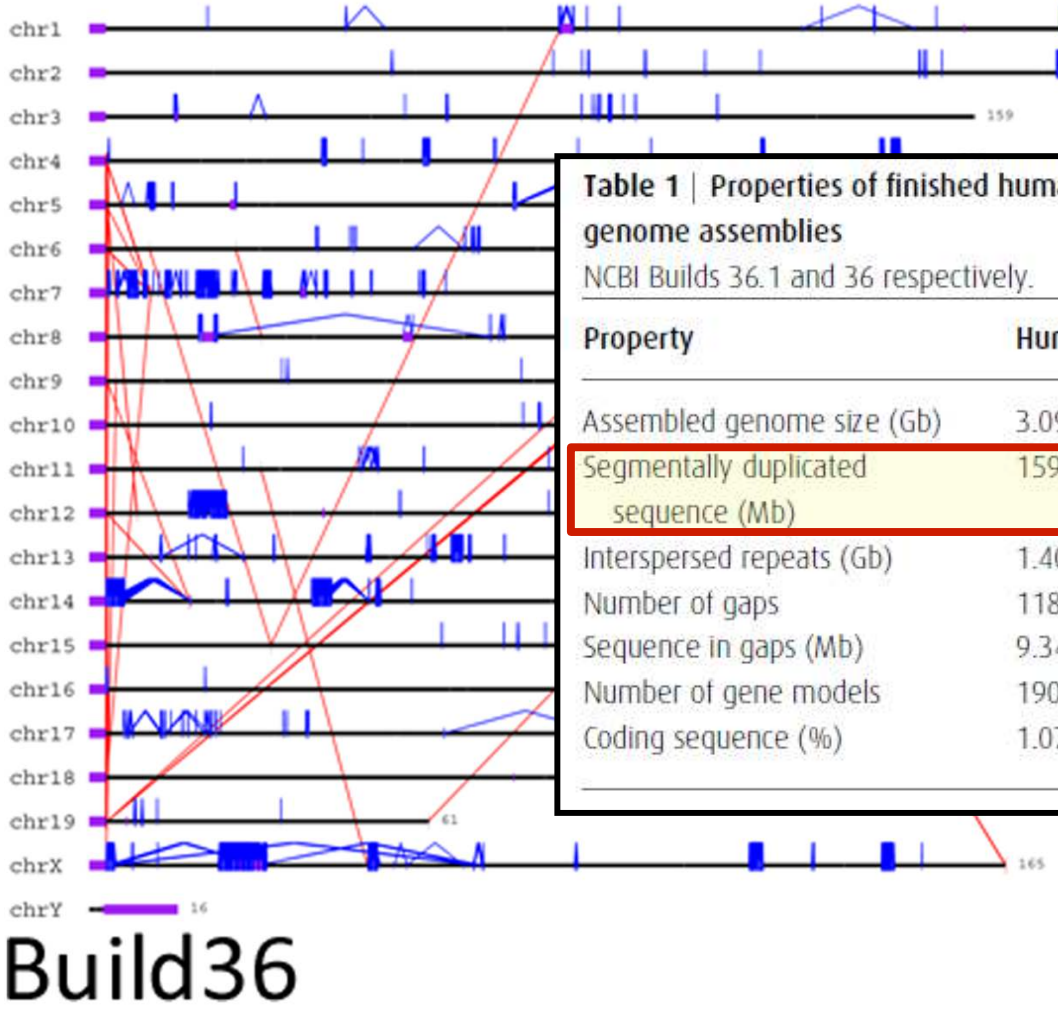
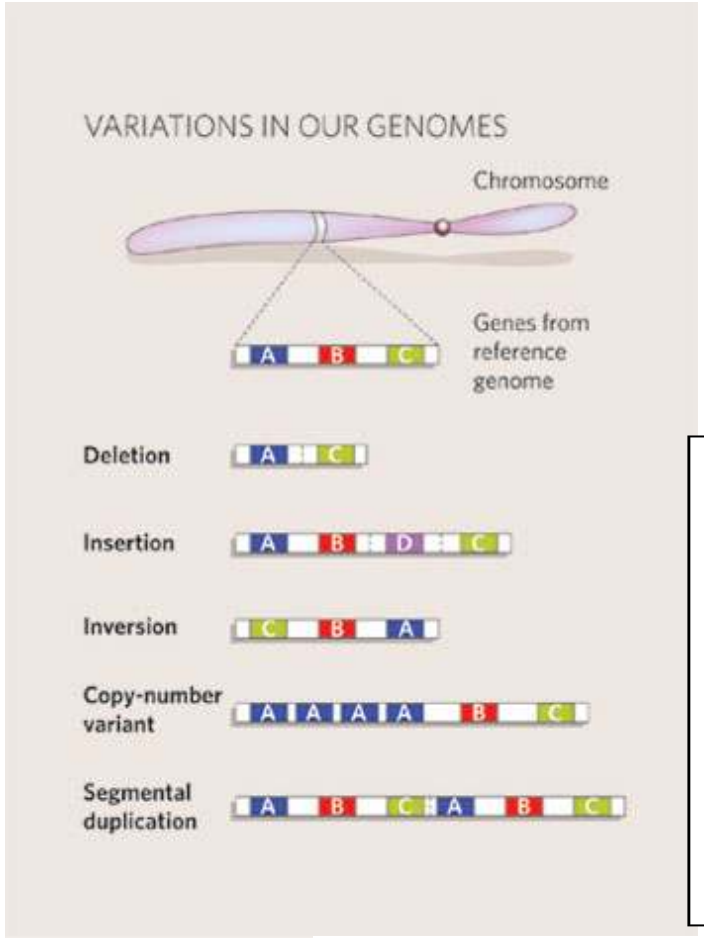


Table 1 | Properties of finished human and mouse reference genome assemblies

NCBI Builds 36.1 and 36 respectively.

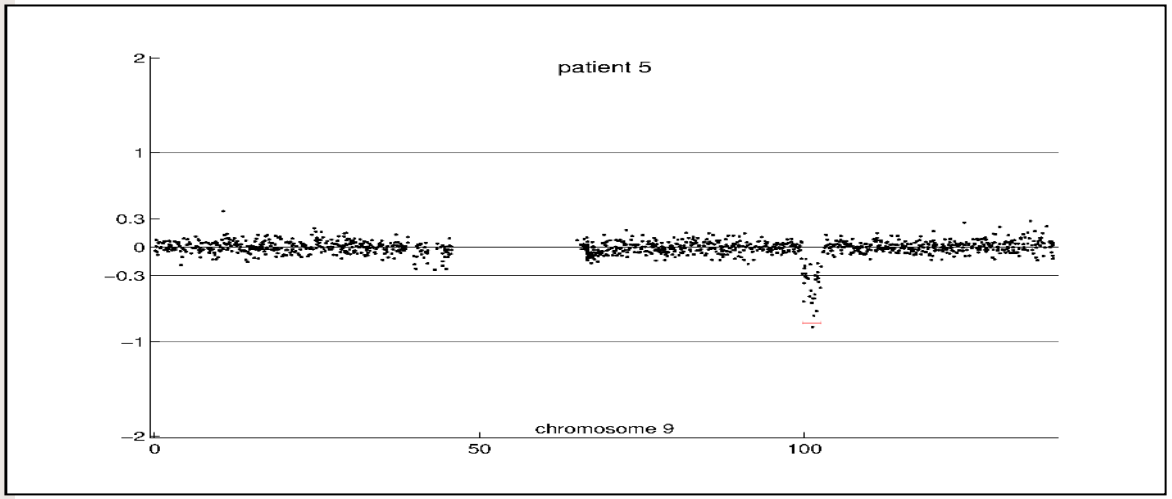
Property	Human	Mouse
Assembled genome size (Gb)	3.091	2.661
Segmentally duplicated sequence (Mb)	159.2 (5.52%)	126.0 (4.94%)
Interspersed repeats (Gb)	1.406	1.091
Number of gaps	118	1218
Sequence in gaps (Mb)	9.343	6.088
Number of gene models	19042	20210
Coding sequence (%)	1.07	1.27

Copy Number Variants (CNVs)



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

Segmental duplications and CNVs often coincide.

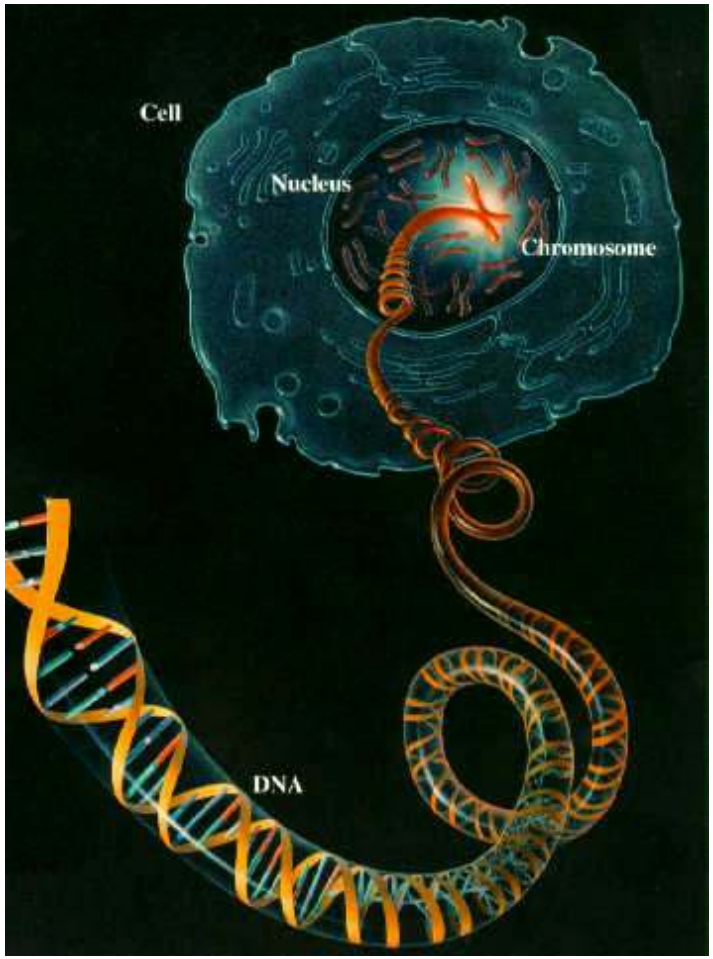


Immunity, defence, chemosensation genes

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]

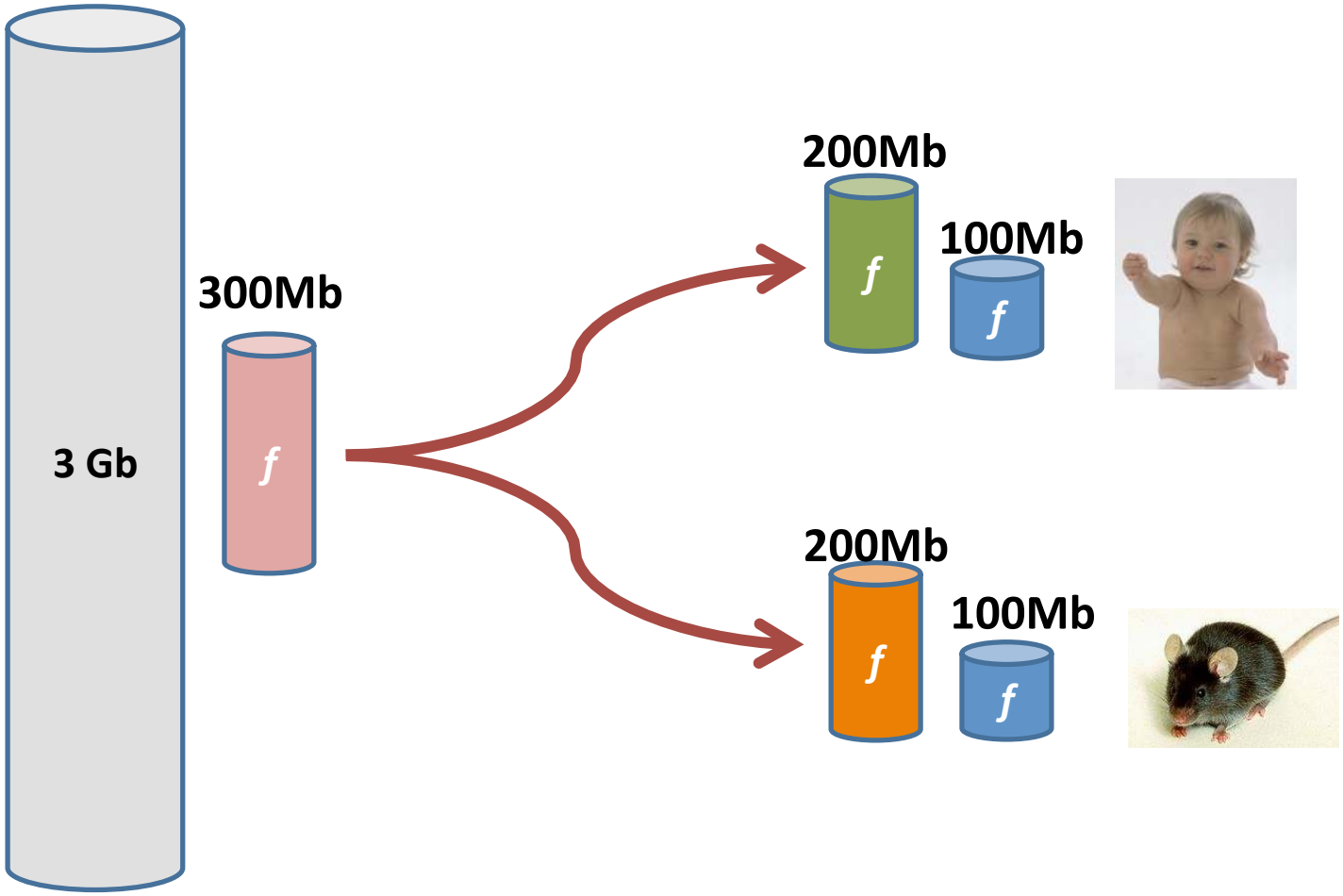
Little evidence that common CNVs are either adaptive or are associated with disease.

Part 2: functional DNA & transcript maps



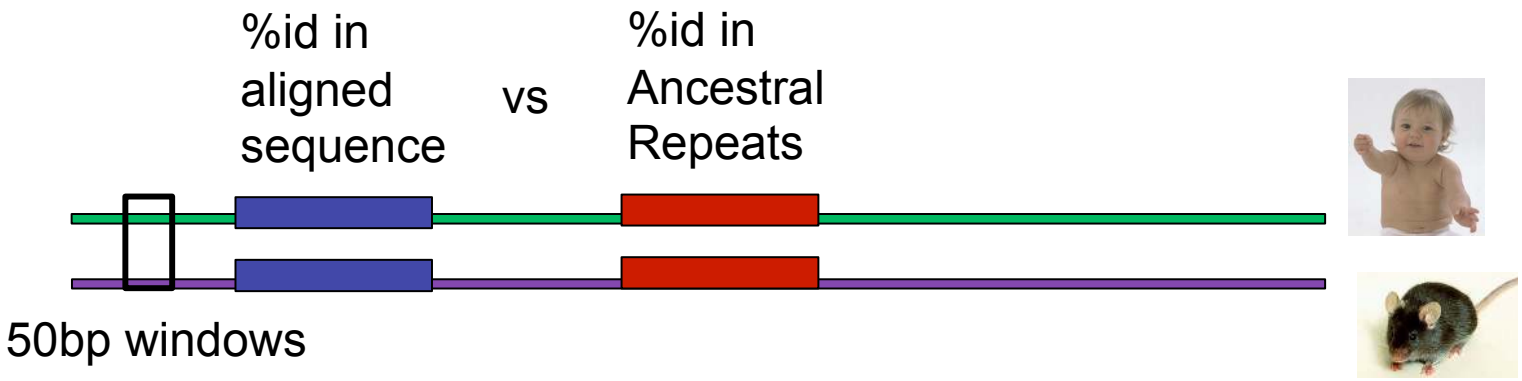
How much of our genome is biologically functional?

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.

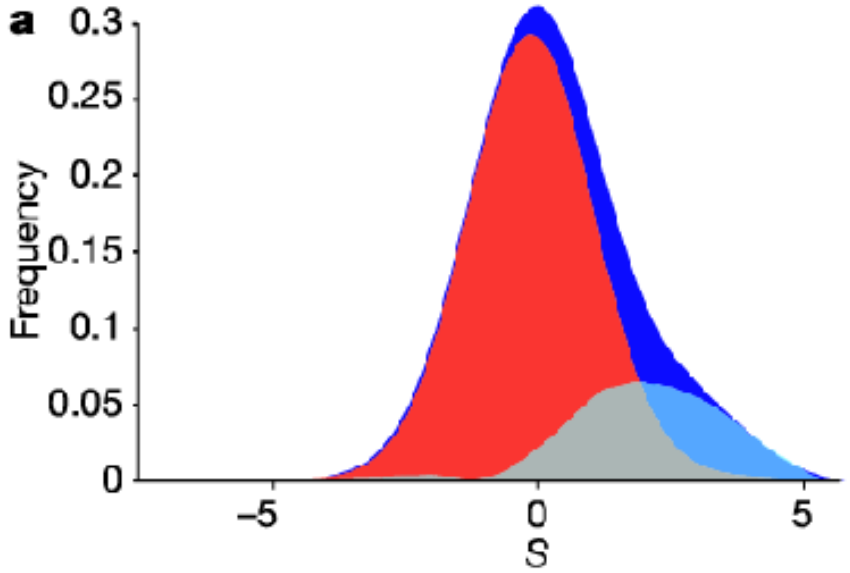
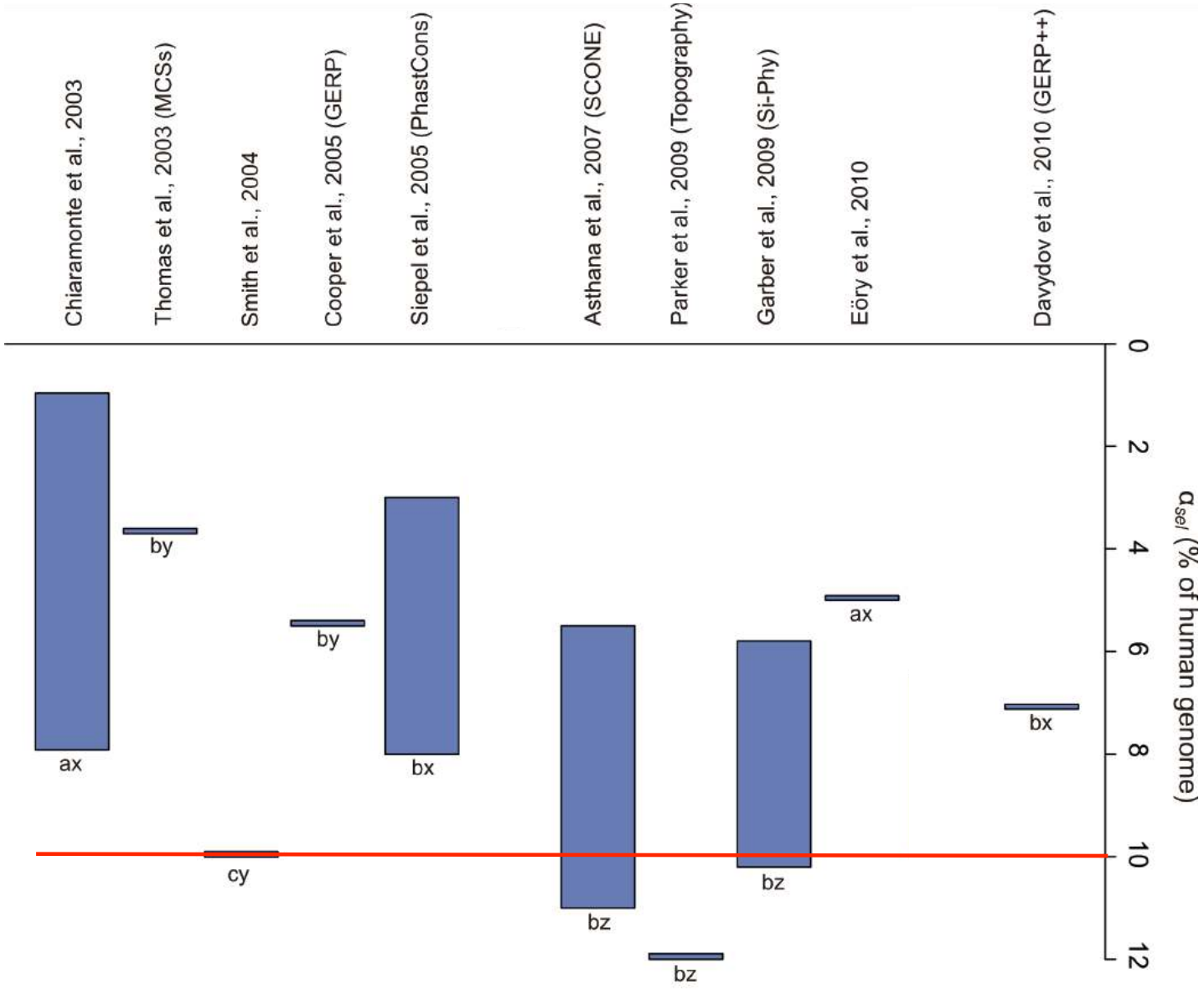


Table 1. Estimates of the Share of the Human Genome under Selection for Different Window Sizes (W) and Required Number of Aligned Bases (T)

W	T	$p_1 = (1 - p_0)$	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%


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

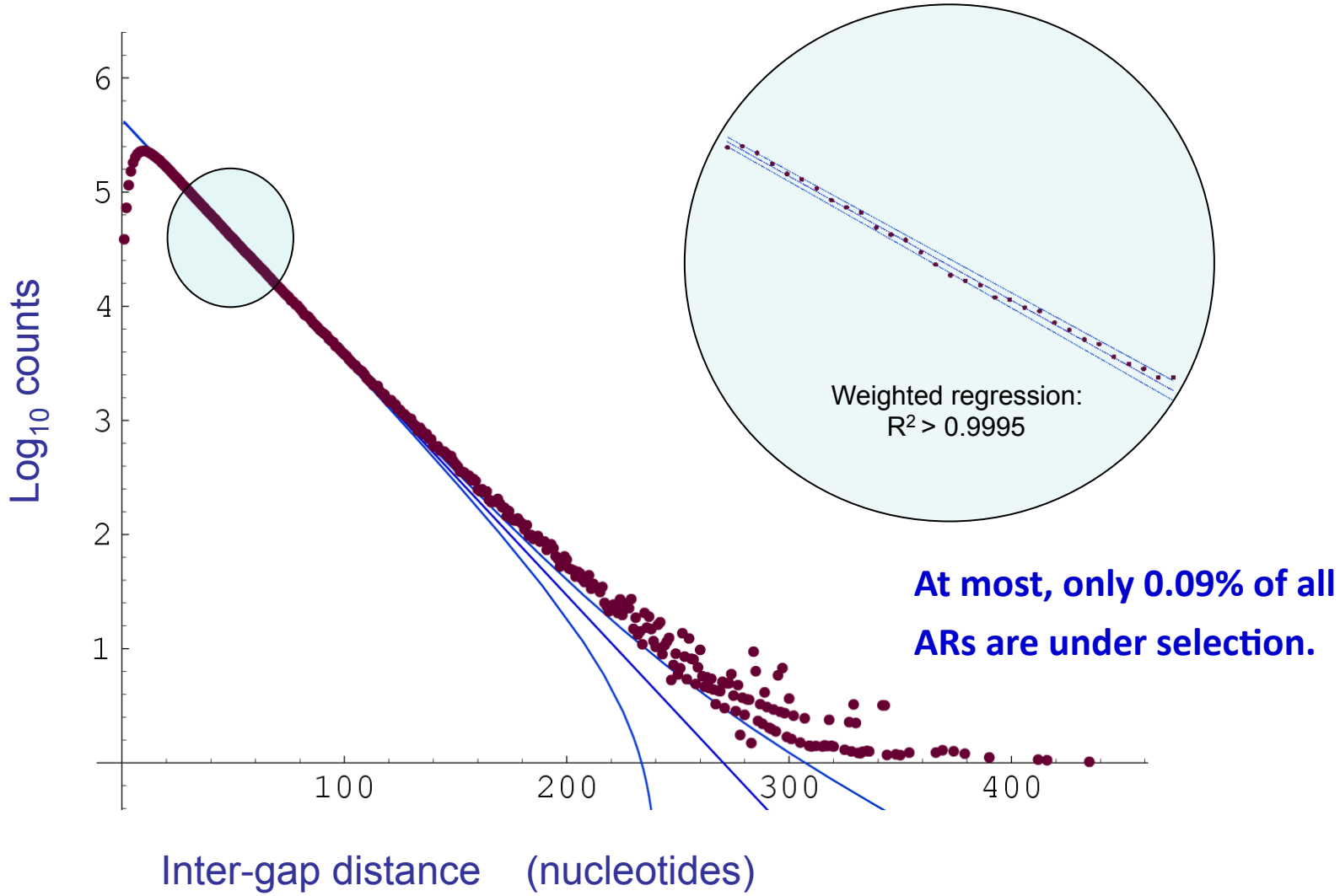
Michael Pheasant and John S. Mattick

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

- 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

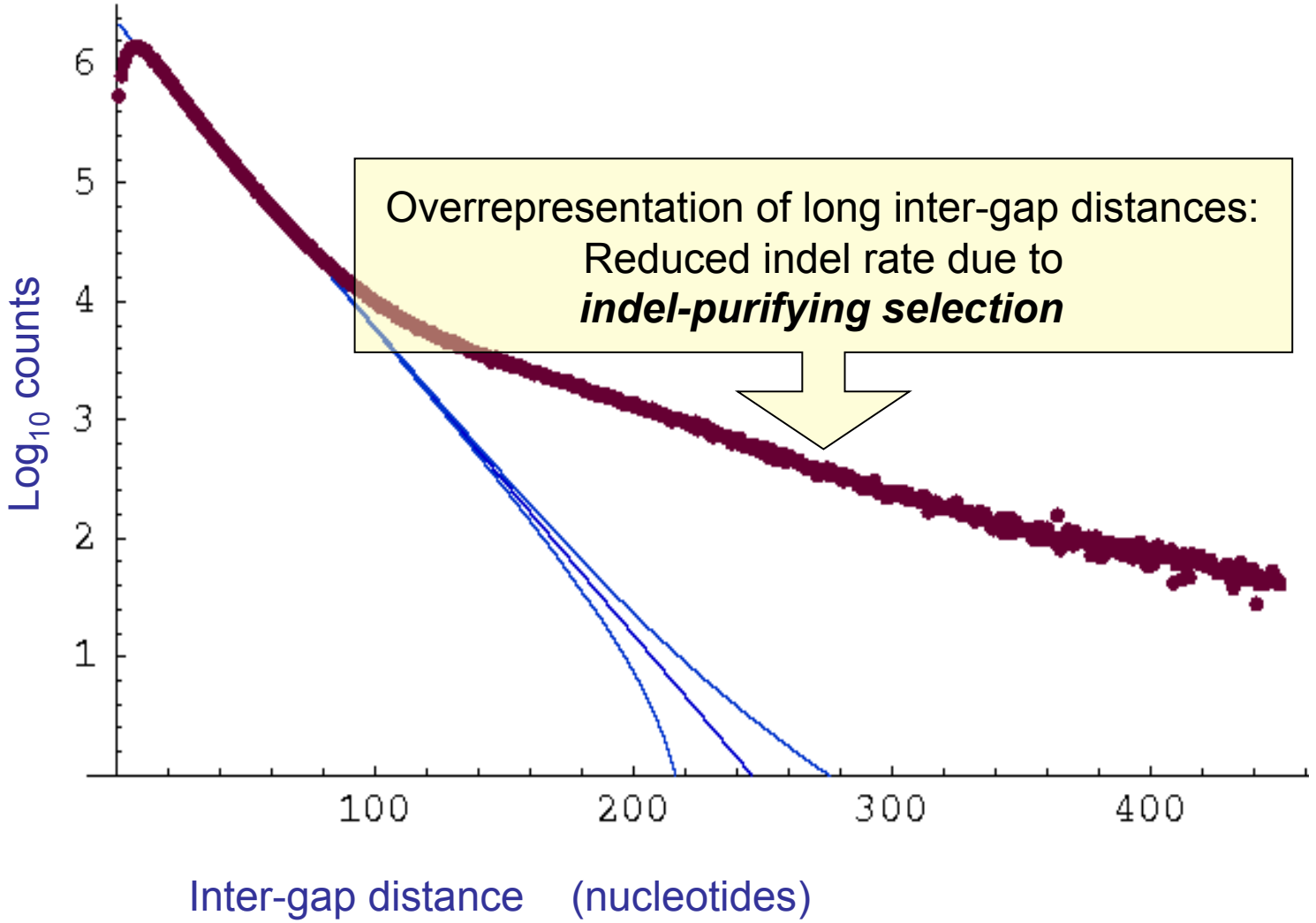
While less than 1.5% of the mammalian genome encodes proteins, it is now evident that the vast majority is transcribed, mainly into non-protein-coding RNAs. This raises the question of what fraction of the genome is functional, i.e., composed of sequences that yield functional products, are required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

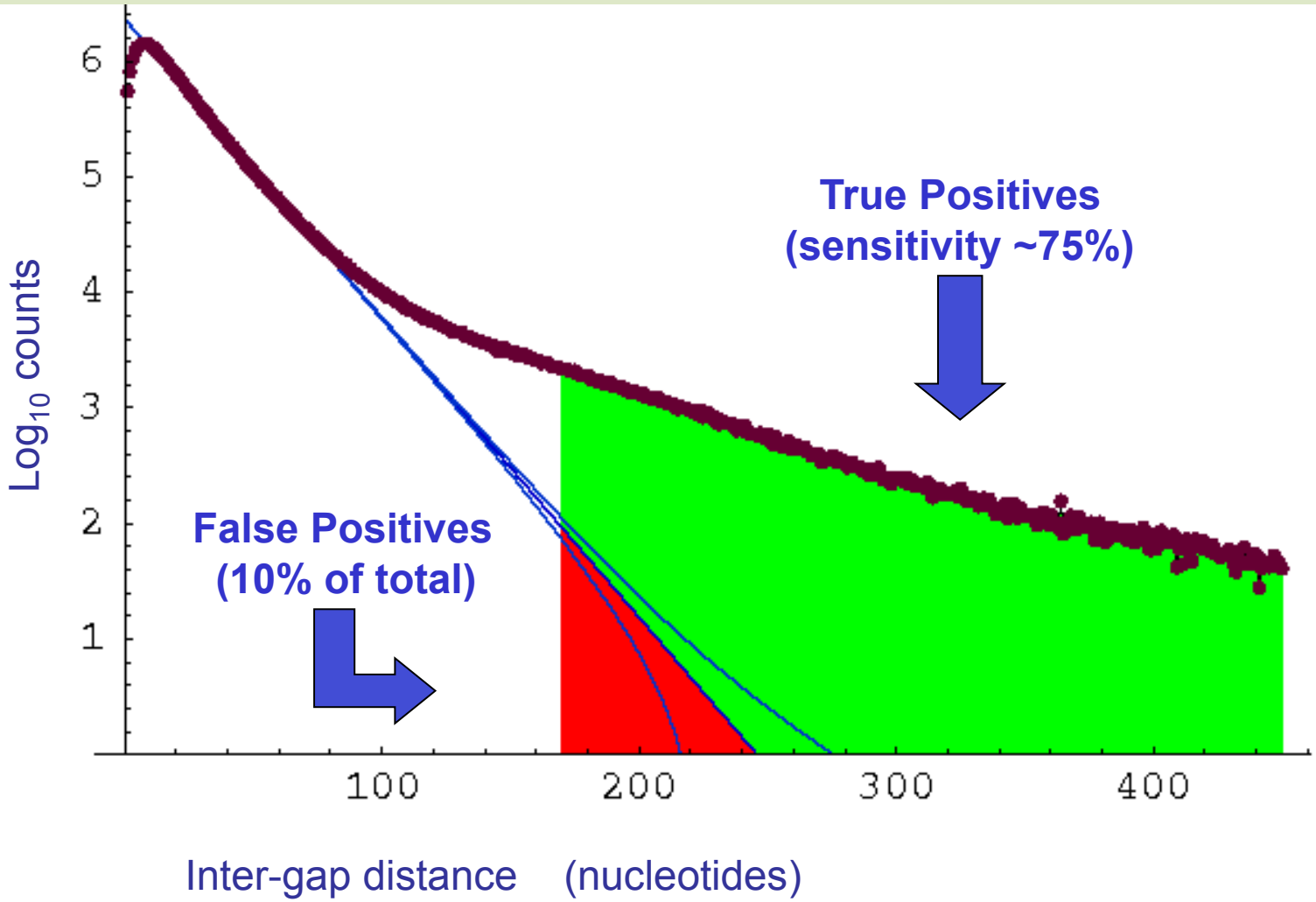
Michael Pheasant and John S. Mattick

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

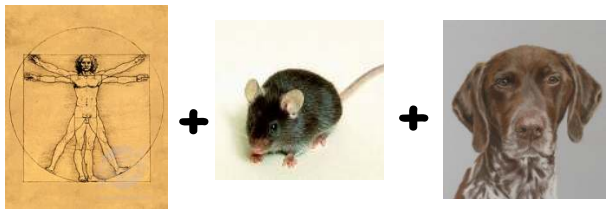


Identifying sequence under indel purifying selection



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

~10%

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

'TEs are predominantly neutral'

much 'turn-over' in functional sequence

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

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

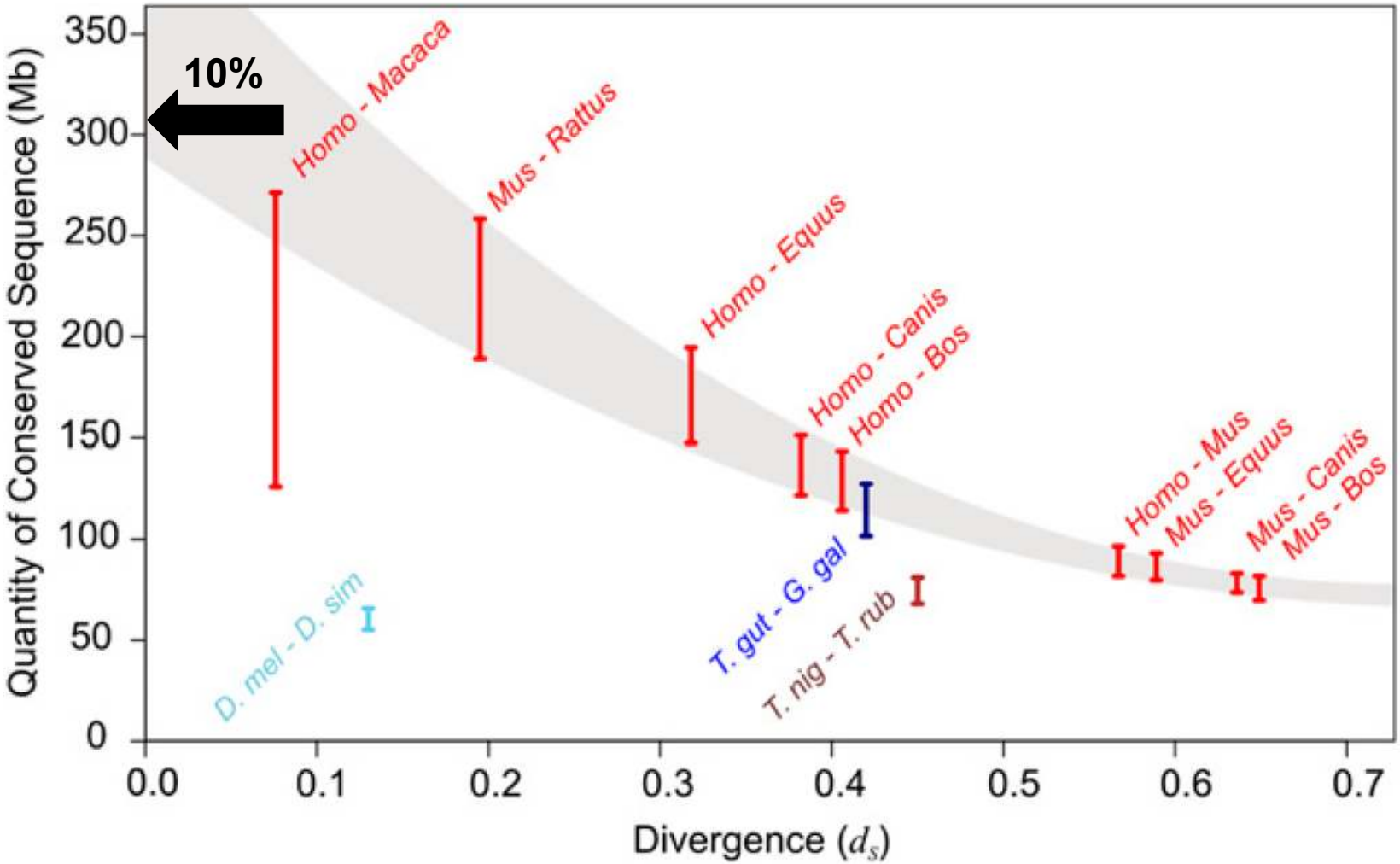
- For example,
  - human – mouse (75 My)
  - human – macaque (25 My)
  - mouse – rat (15 My)?

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++)
ax
by
cy
*
by
bx
ax
bz
bz
ax
bz
cx
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)

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

10% ($e^{-2.3}$)

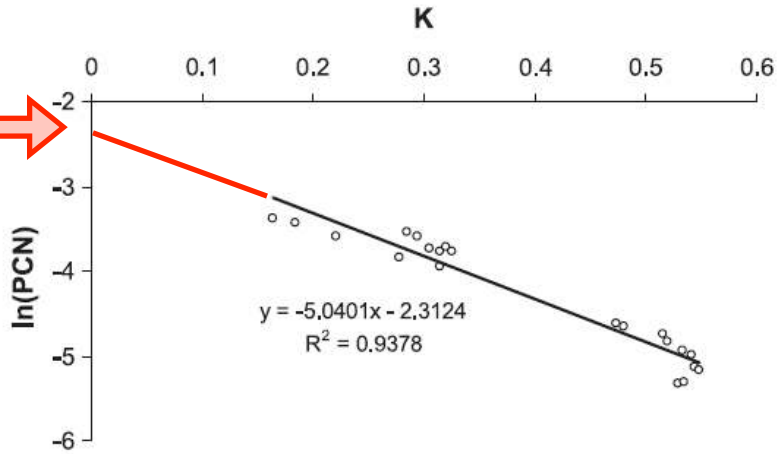


Fig. 1. The relationship between the proportion of the noncoding genome estimated to be conserved by negative selection, PCN, and the pairwise divergence, K, plotted for 21 different pairwise mammalian comparisons. The regression line for ln(PCN) versus K is also shown, along with its equation and the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

Data:

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

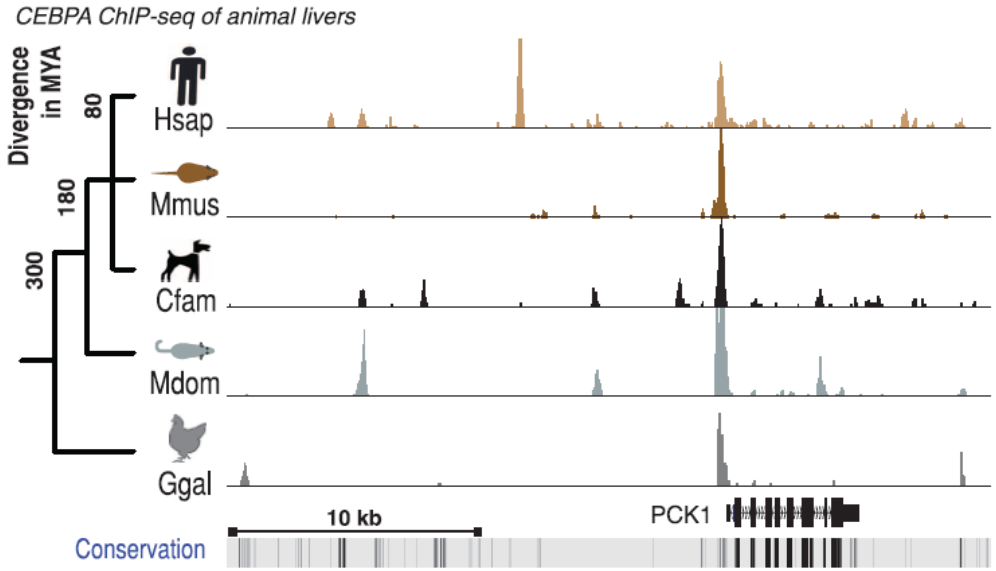
8 mammals
(human, baboon, cat, dog, pig, cow, rat, mouse)

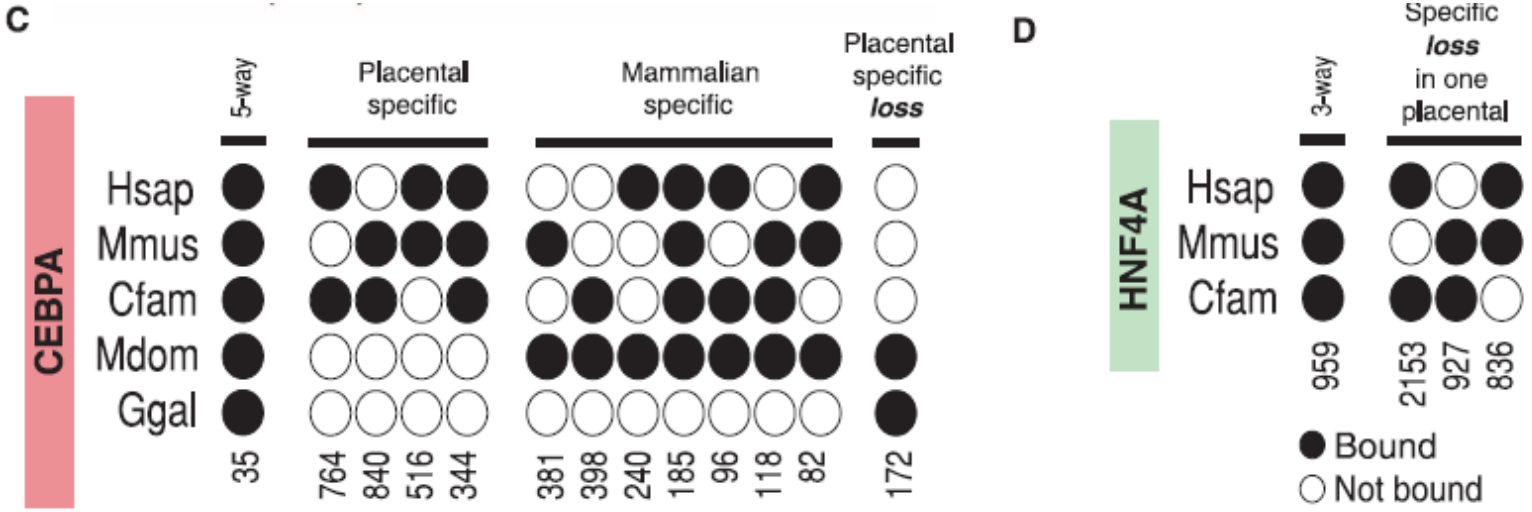
Science
AAAS

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





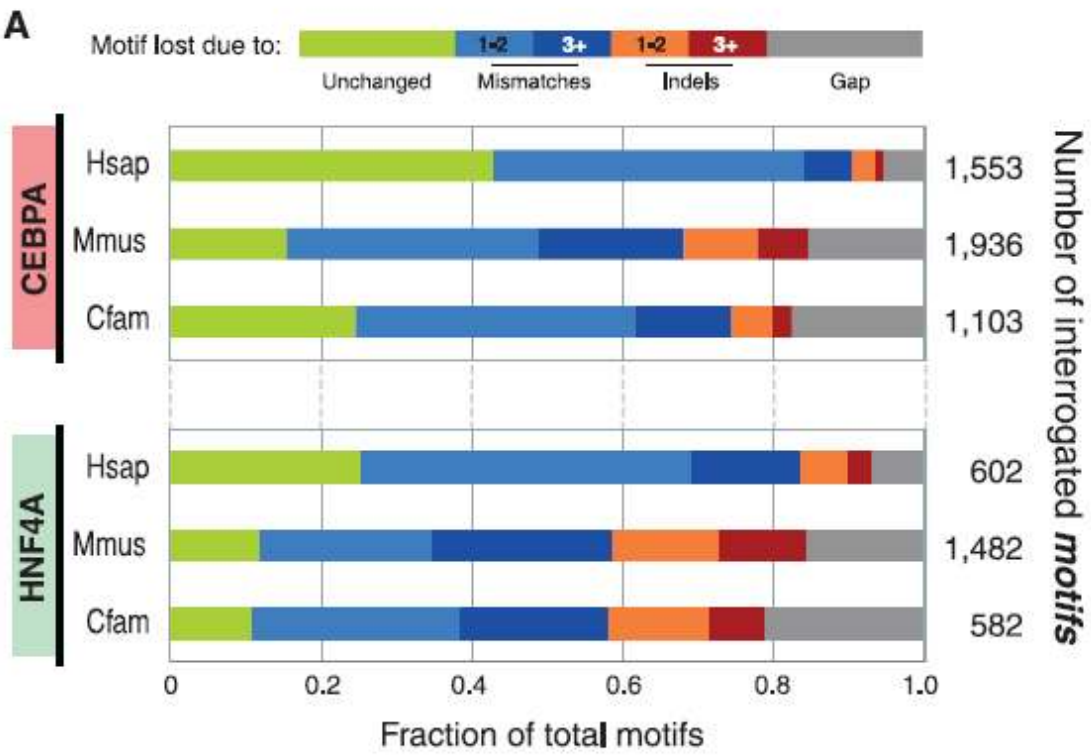
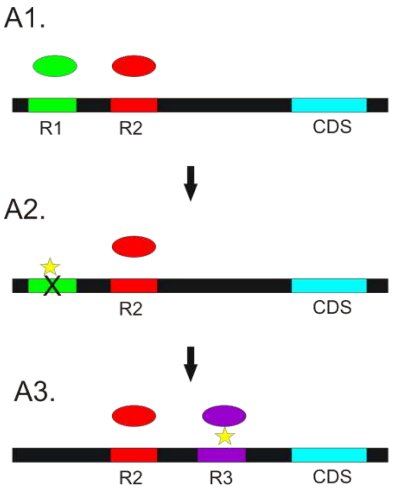


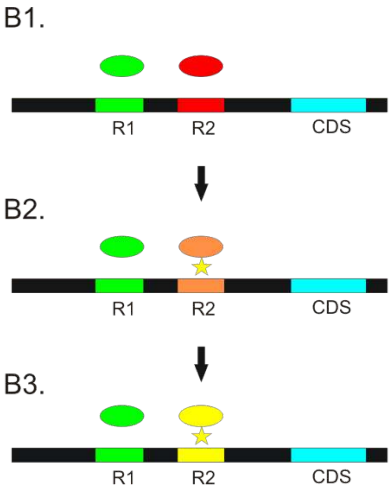
Fig. 4. Lineage-specific loss and turnover of TF binding events. (**A**) The unbound regions in each placental mammal that align to regions showing TF binding in the other two placental mammals were collected, and the mechanisms by which the underlying motifs were disrupted were summarized. (**B**) Turnovers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

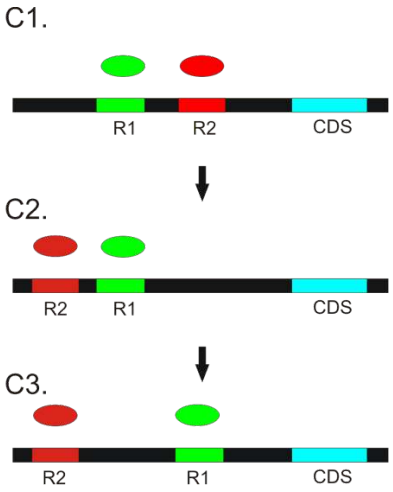
Compensation by new TFBS



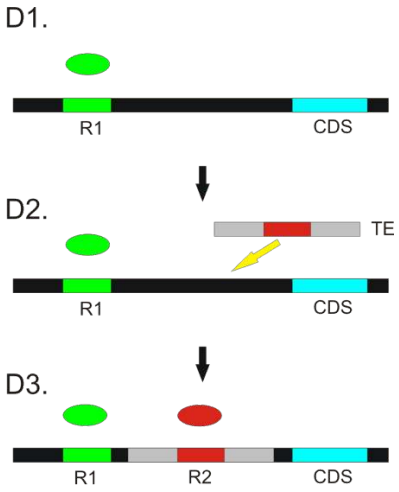
Compensation by co-evolution



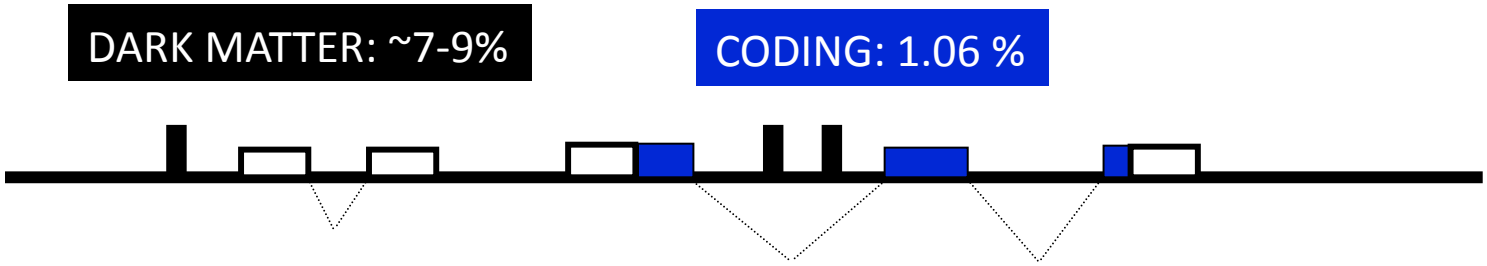
Shuffling of sites



Exaptation of TE

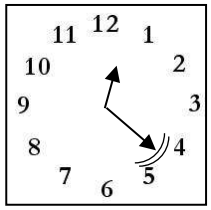


Our View of the Human/Mouse Genome

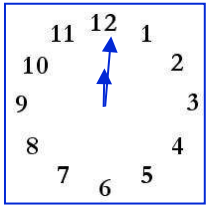


King & Wilson: *Human and Chimpanzee "macromolecules are so alike that regulatory mutations may account for their biological differences"*
Science (1975) 188, 107-116

dark matter

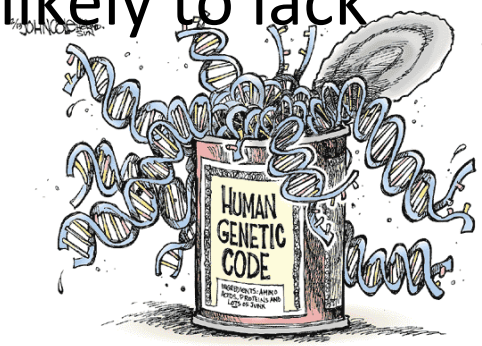


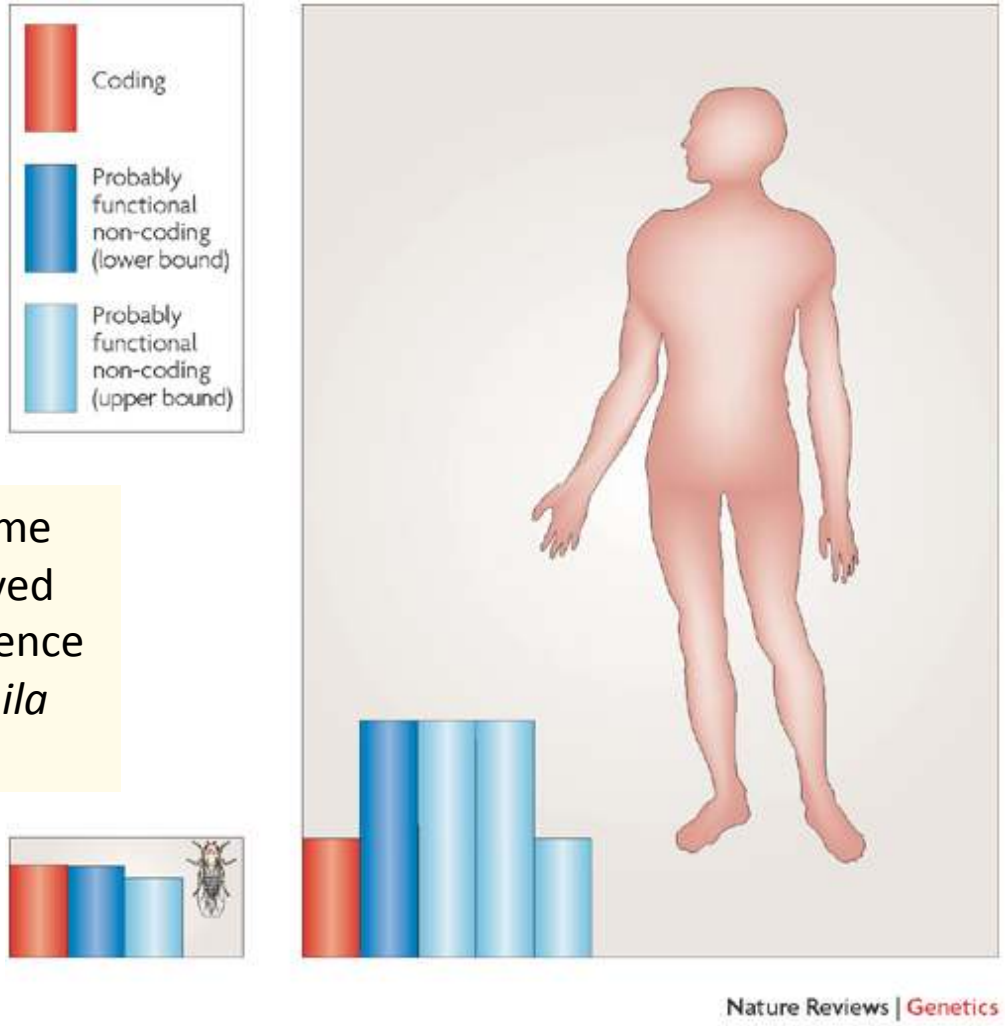
proteins



The Functional Portion

- Approximately 10% of the human genome appears to be constrained with respect to insertions & deletions.
- This compares with 1.0% that encodes protein sequences.
- The amount of constrained but non-coding sequence is thus considerably larger (9-fold) than constrained coding sequence.
- 90% of the human genome, therefore, is likely to lack constraint and truly is *junk*.





The human genome has more conserved (functional) sequence than the *Drosophila* genome.

The functional repertoires of metazoan genomes
Chris P. Ponting
Nature Reviews Genetics 9, 689-698 (September 2008)

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.

Pervasive Transcription

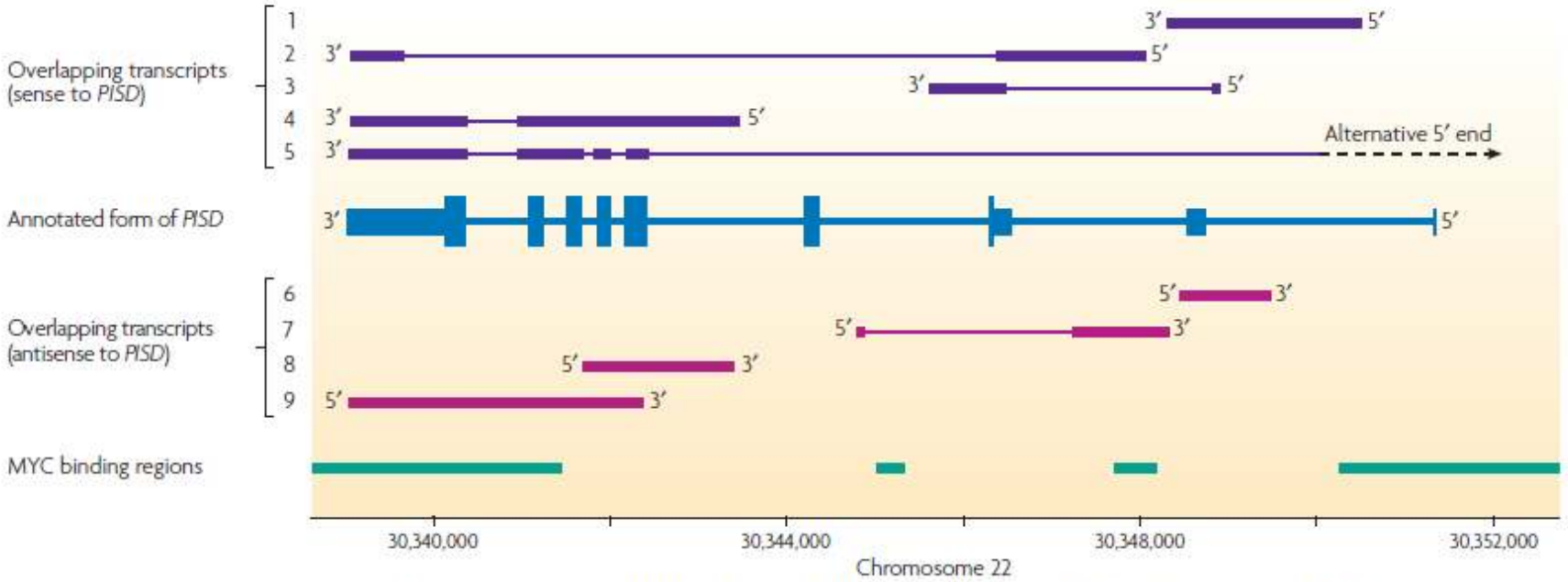


Figure 1 | **Overlapping transcriptional architecture — the *PISD* example.** Five transcripts (in purple) that overlap the RefSeq-annotated form (in blue) of the phosphatidylserine decarboxylase gene (*PISD*) on the same strand, and four transcripts (in pink) that overlap the gene on the opposite strand are shown. The overlapping transcripts were characterized using a combination of RACE and tiling arrays[16]. Binding sites of the transcriptional factor MYC (in green) were determined using a ChIP–chip assay[23]. The coordinates are taken from the hs.NCBIv35 version of the genome.

Genome-wide transcription and the implications for genomic organization

Philipp Kapranov, Aarron T. Willingham and Thomas R. Gingeras

cDNA sequencing

- 11,665 noncoding TUs (FANTOM2)
- >3,500 noncoding txs (FANTOM3)

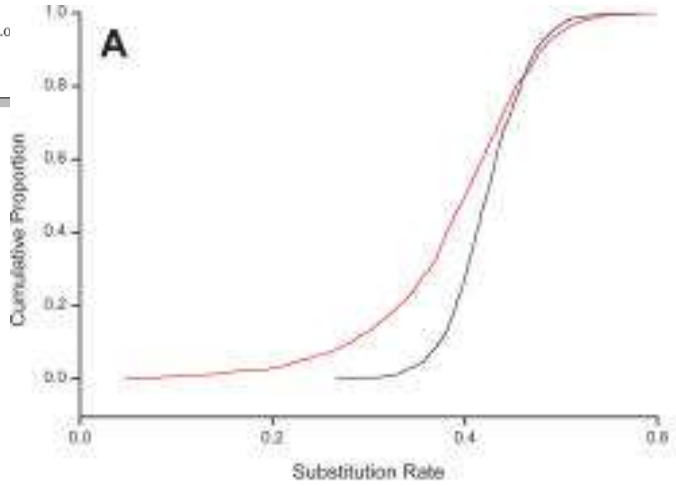
Genome Research
www.genome.org
17:556–565 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.o

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



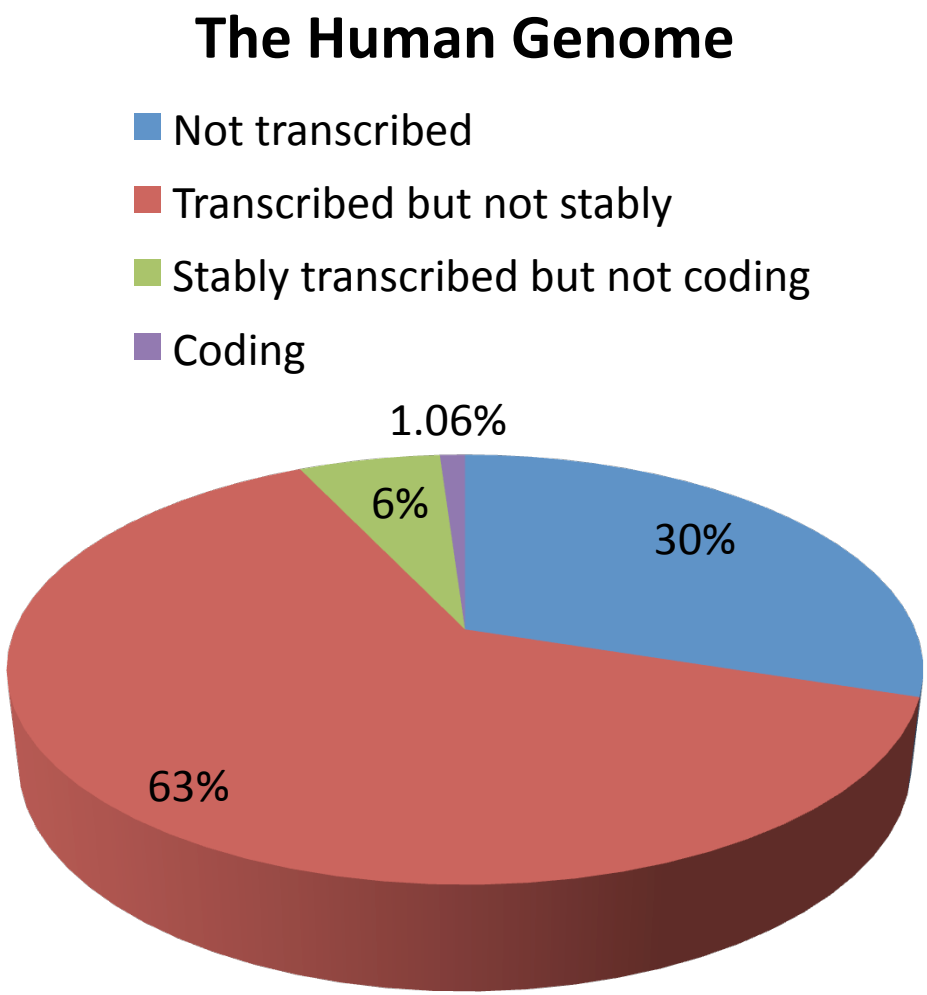
Abstract

consistent with neutralist explanations. Rather, our results indicate that purifying selection has acted on the macroRNAs' promoters, primary sequence, and consensus splice site motifs. Promoters have experienced the greatest elimination of nucleotide substitutions, insertions, and deletions. The proportion of conserved sequence (4.1%–5.5%) in these macroRNAs is comparable to the density of exons within protein-coding transcripts (5.2%). These

Tiling arrays

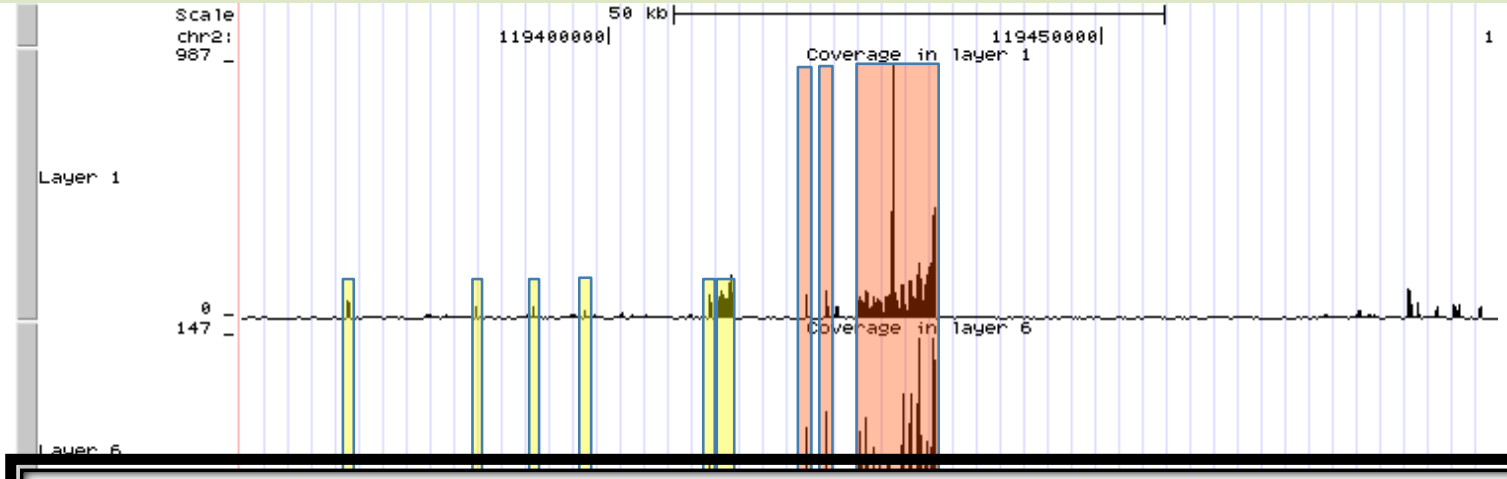
- ENCODE: 1% sample of the human genome
- 93% of bases appeared to be transcribed, three-quarters of which could be verified by an alternative technique.
- But transcription often not reproduced in other tissues/cells.
- Could such transcription often be tissue/cell-specific?

RNA-Seq


The Human Genome
Not transcribed
Transcribed but not stably
Stably transcribed but not coding
Coding
1.06%
6%
30%
63%

Whole genome RNA Seq: Cortical Cell Layers

Expression level



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

Expressed protein-coding gene

Expressed non-coding gene

Non-expressed protein-coding genes

Belgard, Ponting et al., unpublished data

OPEN ACCESS Freely available online

PLoS BIOLOGY

Synopsis

Dark Matter Transcripts: Sound and Fury, Signifying Nothing?

Richard Robinson*

Freelance Science Writer, Sherborn, Massachusetts, United States of America

"The emerging picture of RNA polymerase is of an inherently imprecise, not to say promiscuous, copyist, one whose output includes some mistakes along with lots of valuable product. In this view, most dark matter transcripts are not signals emerging from a hidden universe within the genome, but instead simply the noise emitted by a busy machine."

- “We conclude that, while there are bona fide new intergenic transcripts, their number and abundance is generally low in comparison to known exons, and the genome is not as pervasively transcribed as previously reported.”
- “We conclude that analysis of data from tiling arrays leads to vast overestimates of the proportion of transcriptional dark matter.”
- “we do not find evidence for widespread interleaved transcripts as previously described” [ENCODE]
- “Pervasive transcription of intergenic regions as described in previous studies occurs at a significantly reduced level and is of a random character.”

Issues

- Most study employ polyA+ samples only
- Tiling arrays (Cheng et al.): RNA transcripts lacking polyadenylation represent approximately half of the transcriptome, and a third of all transcribed sequences were detected in both poly A- and poly A+ forms.
- Most ncRNAs may be transcribed in specific cells only

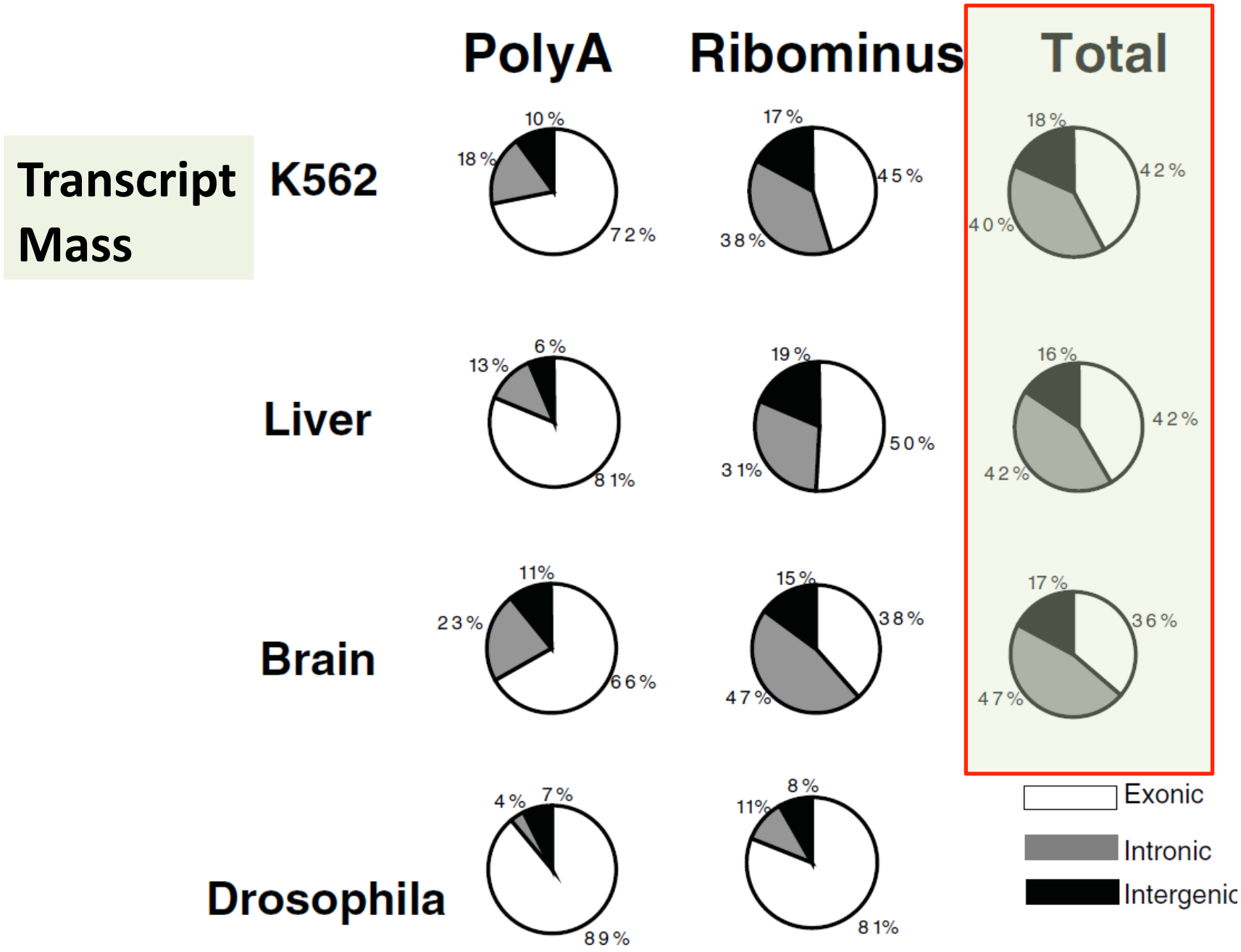
Kapranov et al.'s response

The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA

BMC Biology 2010, **8**:149 doi:10.1186/1741-7007-8-149

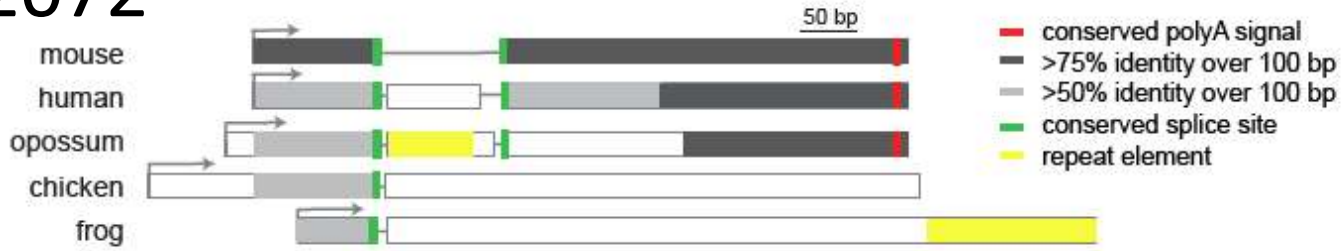
Philipp Kapranov (philippk08@gmail.com)
Georges St. Laurent (georgest98@yahoo.com)
Tal Raz (tal.raz@gmail.com)
Fatih Ozsolak (fozsolak@helicosbio.com)
C. Patrick Reynolds (Patrick.Reynolds@TTUHSC.edu)
Poul HB Sorensen (psor@interchange.ubc.ca)
Gregory Reaman (greaman@childrensoncologygroup.org)
Patrice Milos (PMilos@helicosbio.com)
Robert J Arceci (Arcecro@jhmi.edu)
John F Thompson (jthompson@helicosbio.com)
Timothy J Triche (triche@usc.edu)

- Helicos technology
- Total RNA


PolyA
Ribominus
Total
Transcript Mass
K562
Liver
Brain
Drosophila
10%
18%
72%
17%
45%
38%
18%
42%
40%
6%
13%
81%
19%
50%
31%
16%
42%
42%
11%
23%
66%
15%
38%
47%
17%
36%
47%
4%
7%
89%
8%
11%
81%
Exonic
Intronic
Intergenic

Are ncRNA loci conserved across different species?

AK082072

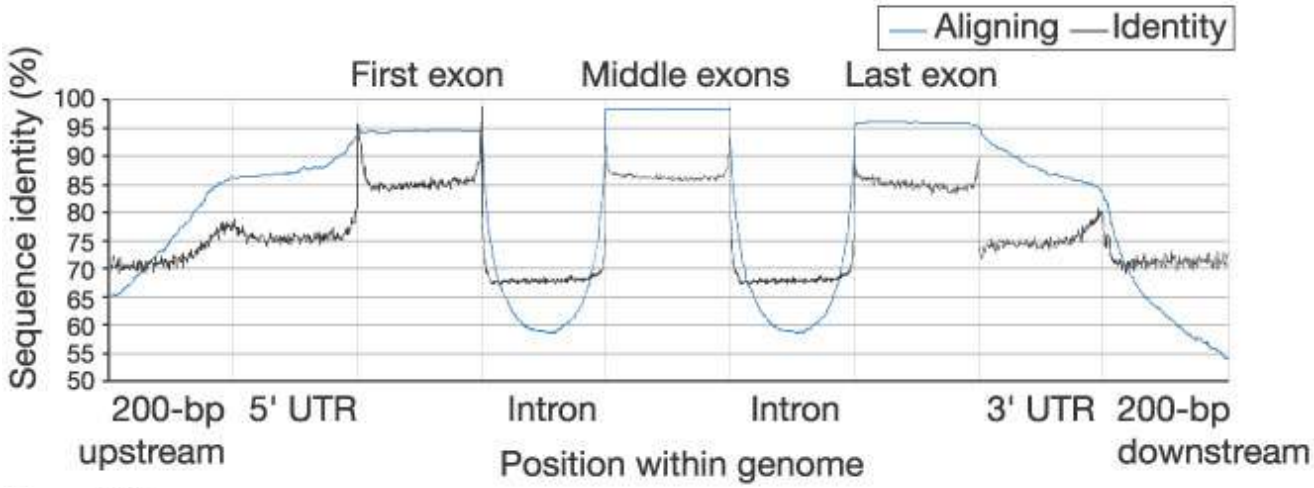


AK082467

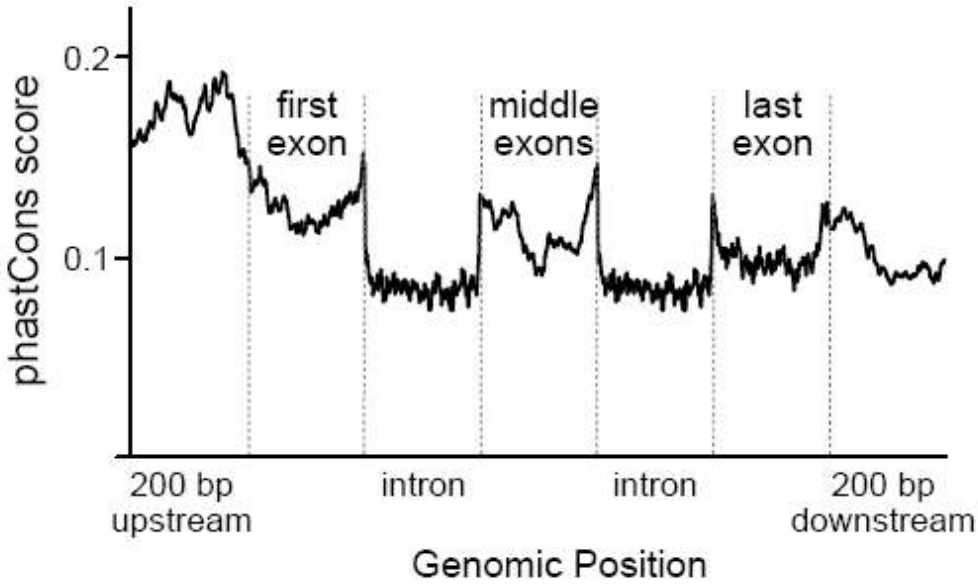
Only 14% of mouse long intergenic transcripts have evidence for transcription in human

Variation in conservation across coding and noncoding genes

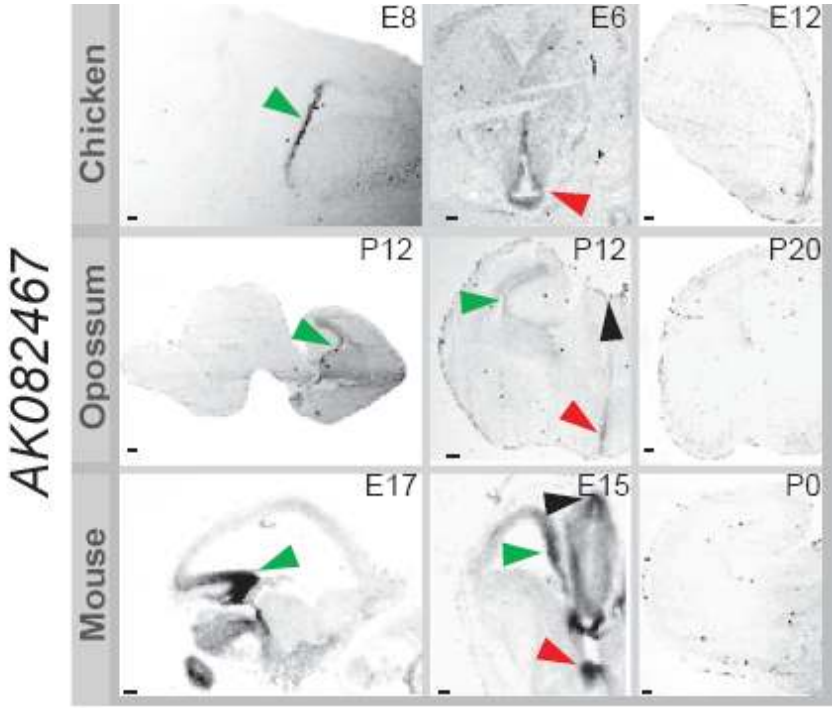
PROTEINS
Mouse Genome paper



Multiexonic long ncRNAs (Chodroff et al.)



Conserved expression across amniotes

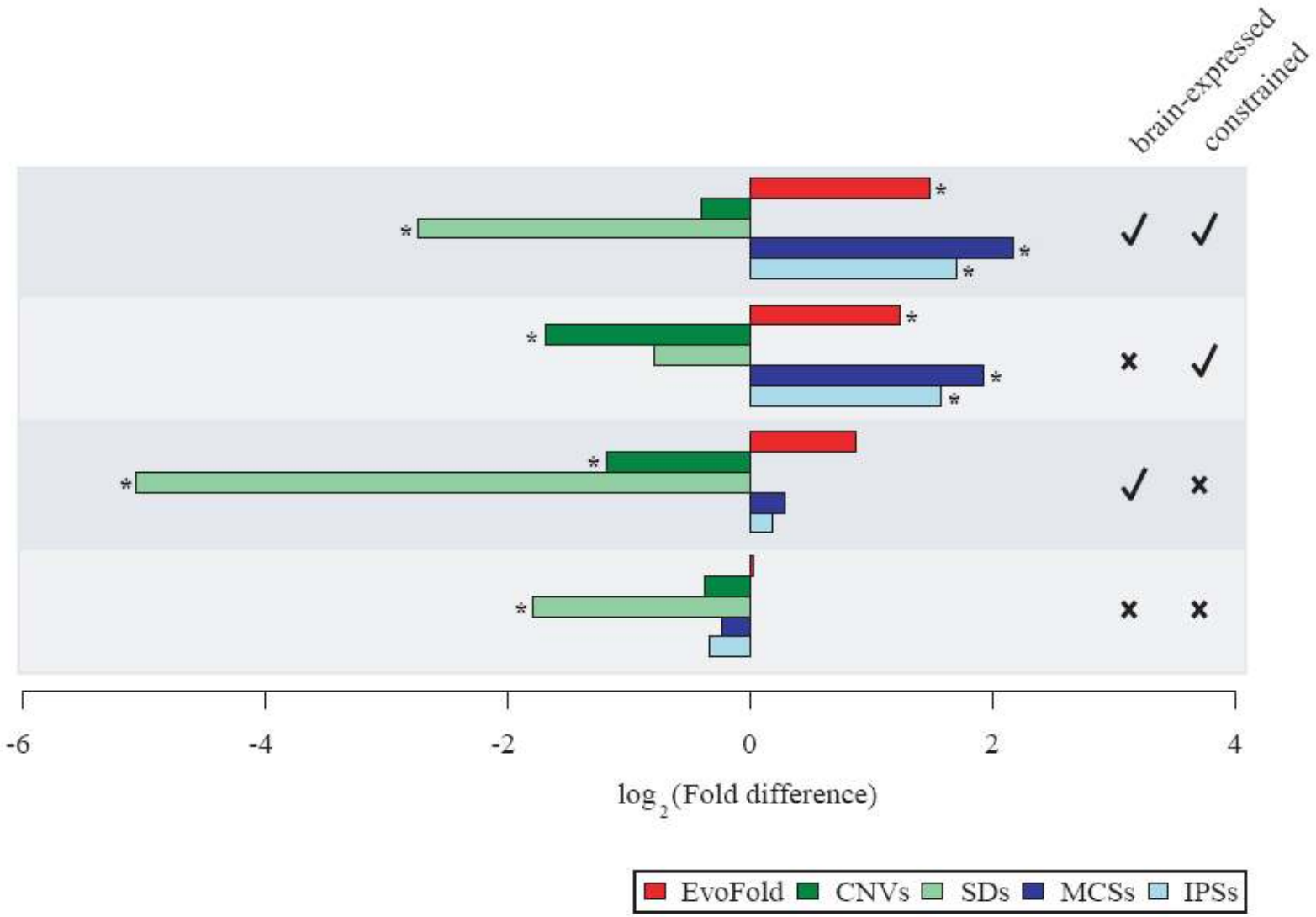


Ventricular zone of the hippocampal formation (green arrowheads), the preoptic area of the hypothalamus (red arrowheads), and the epithalamus (black arrowheads). Signal was undetectable at later developmental stages. Scale bars = 200 μm

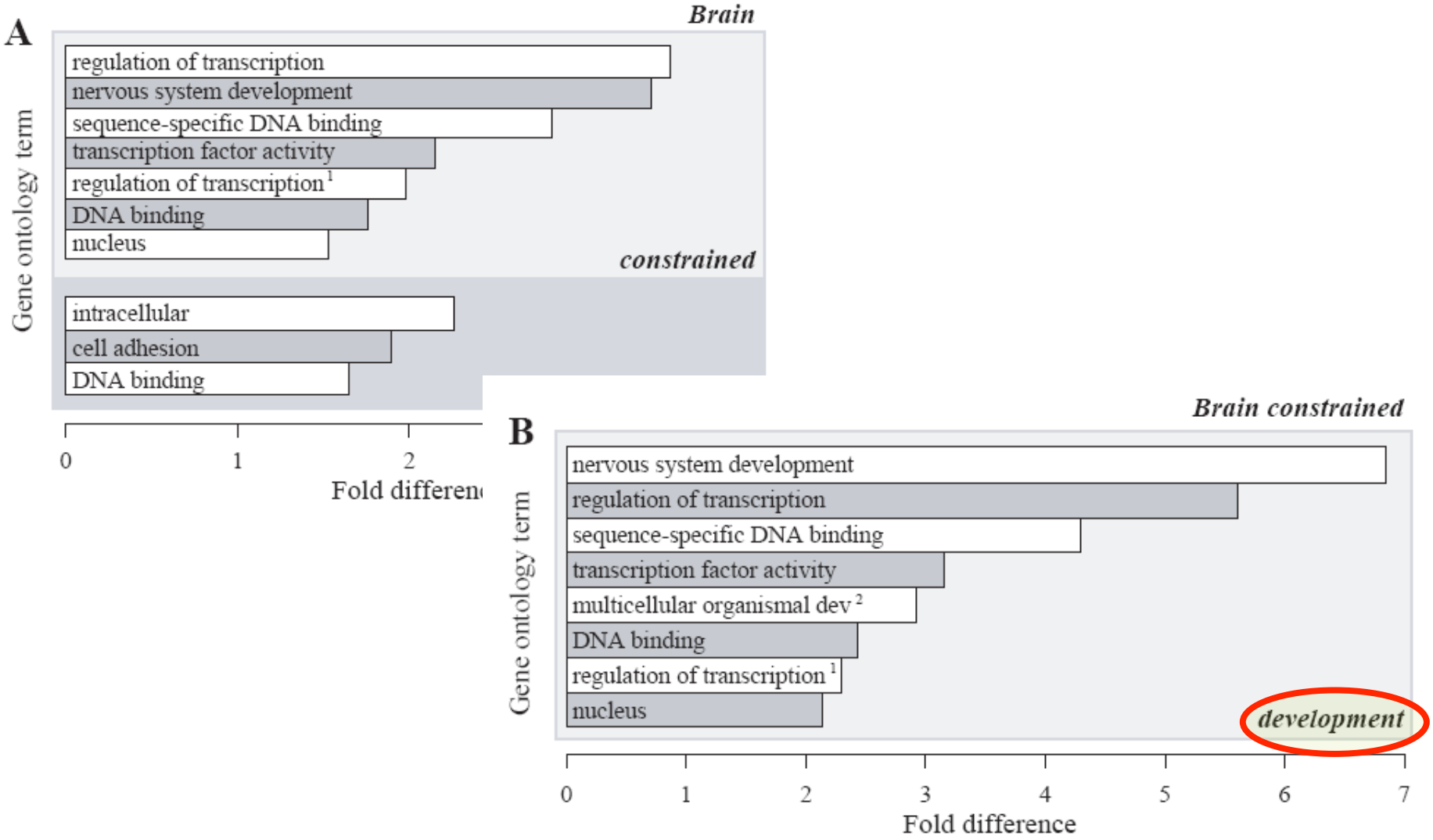
LincRNA gene mechanisms?

- *Cis* versus *trans*?
- Functional RNA or DNA?
- Transcriptional interference?
- Contribution to brain biology & development?
- Conservation of transcription (yes & no).

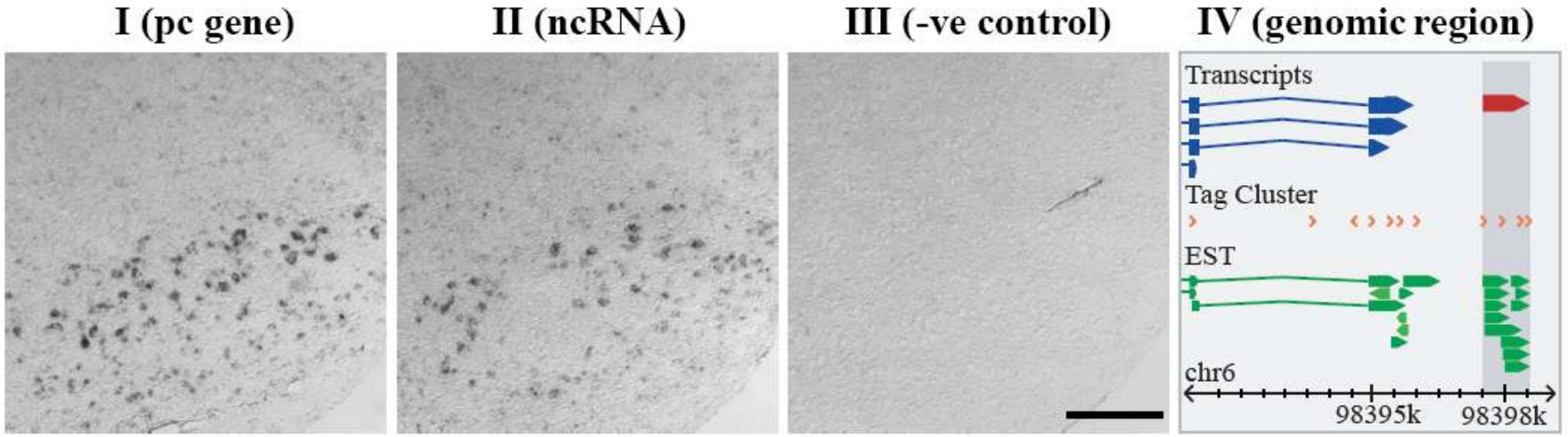
ncRNA, rather than underlying DNA, appears functional



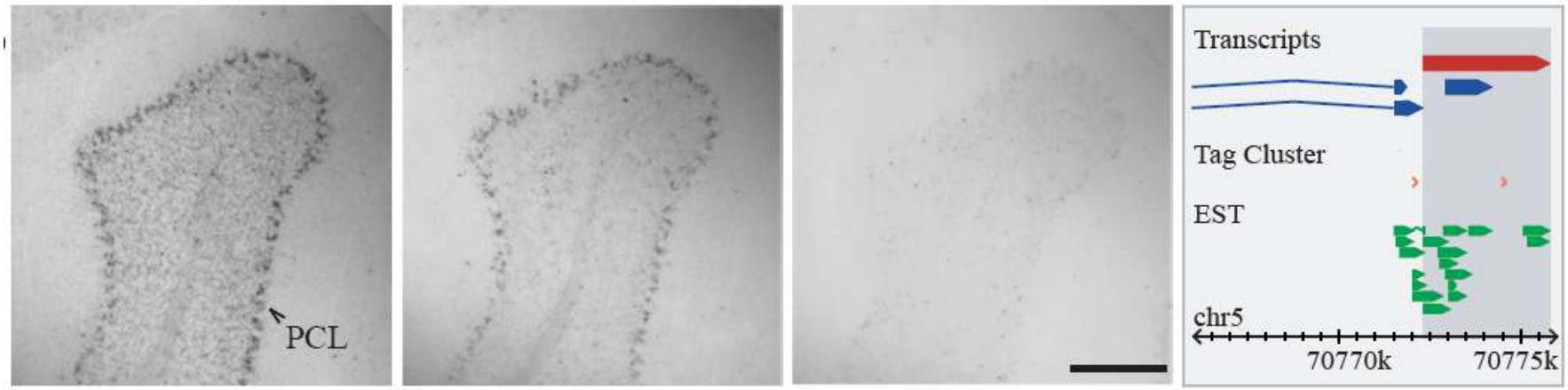
Brain-expressed ncRNAs tend to neighbour transcription factor genes



Overlapping expression of adjacent pc & nc transcripts



facial nuclei of the medulla



Purkinje cell layer of the cerebellum

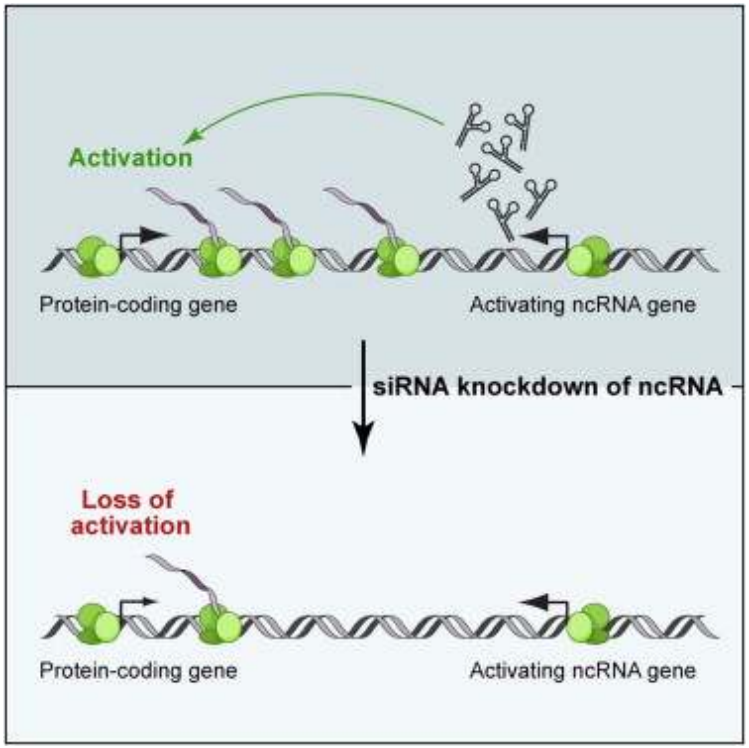
OPEN ACCESS Freely available online

PLoS GENETICS

Genomic and Transcriptional Co-Localization of Protein-Coding and Long Non-Coding RNA Pairs in the Developing Brain

Jasmina Ponjavic, Peter L. Oliver, Gerton Lunter, Chris P. Ponting*

MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, United Kingdom



Cell

Long Noncoding RNAs with Enhancer-like Function in Human Cells

Ulf Andersson Ørom,[1] Thomas Derrien,[2] Malte Beringer,[1] Kiranmai Gumireddy,[1] Alessandro Gardini,[1] Giovanni Bussotti,[2] Fan Lai,[1] Matthias Zytnicki,[2] Cedric Notredame,[2] Qihong Huang,[1] Roderic Guigo,[2] and Ramin Shiekhattar[1,2,3,*]

[1]The Wistar Institute, 3601 Spruce Street, Philadelphia, PA 19104, USA

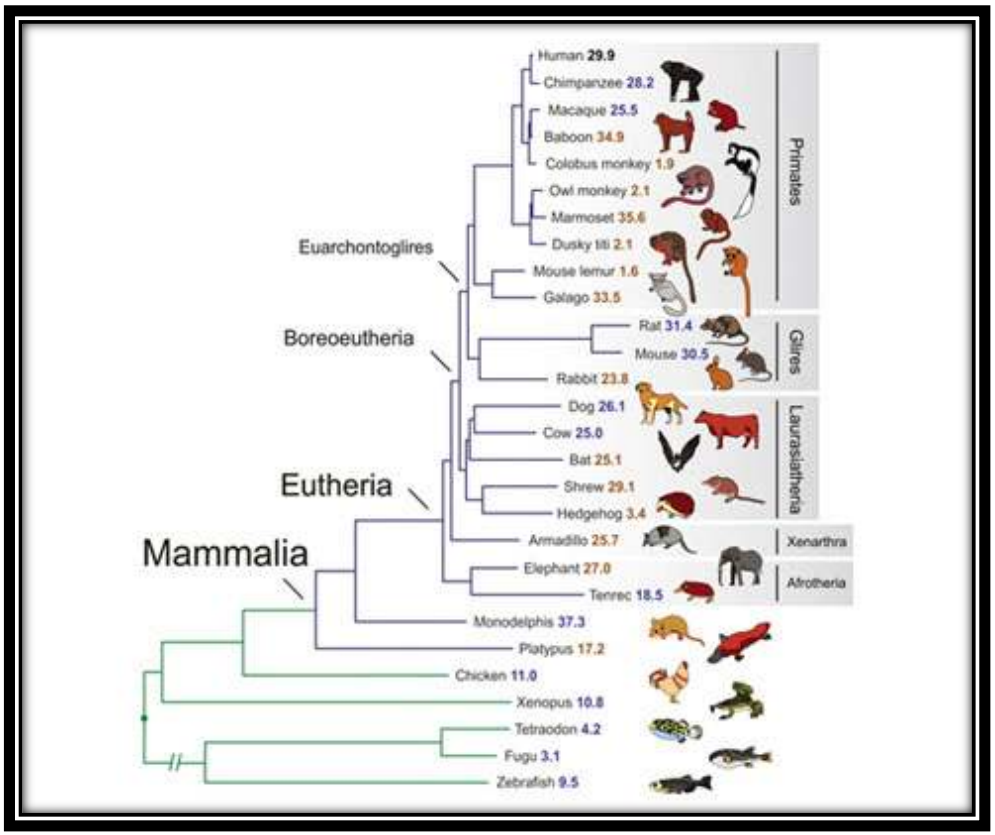
[2]Centre for Genomic Regulation (CRG), UPF, Barcelona, Spain

[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

*Correspondence: shiekhattar@wistar.org

DOI 10.1016/j.cell.2010.09.001

Part 3: Genome Evolution & the future (shorter!)



Isochores

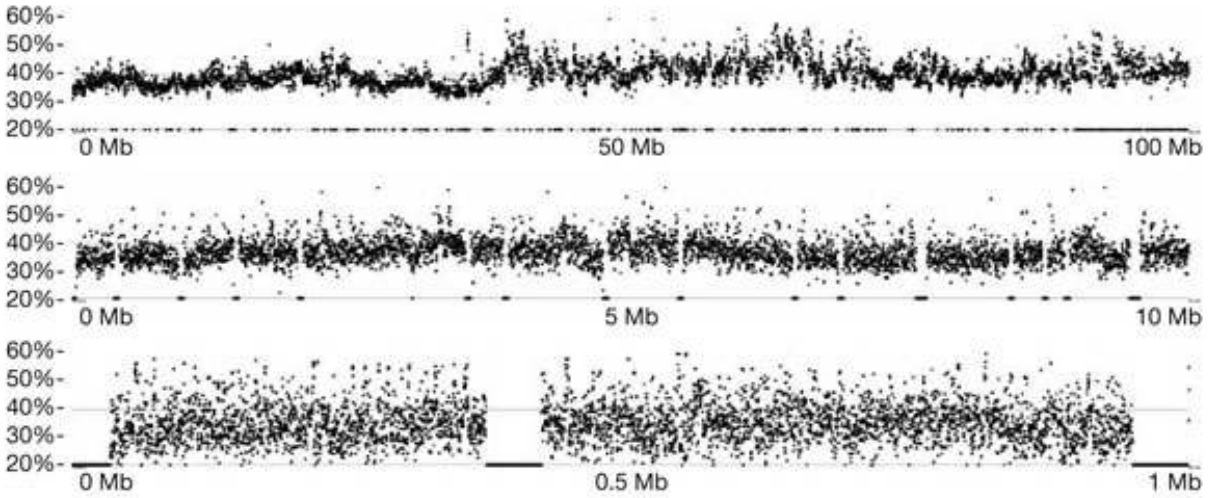
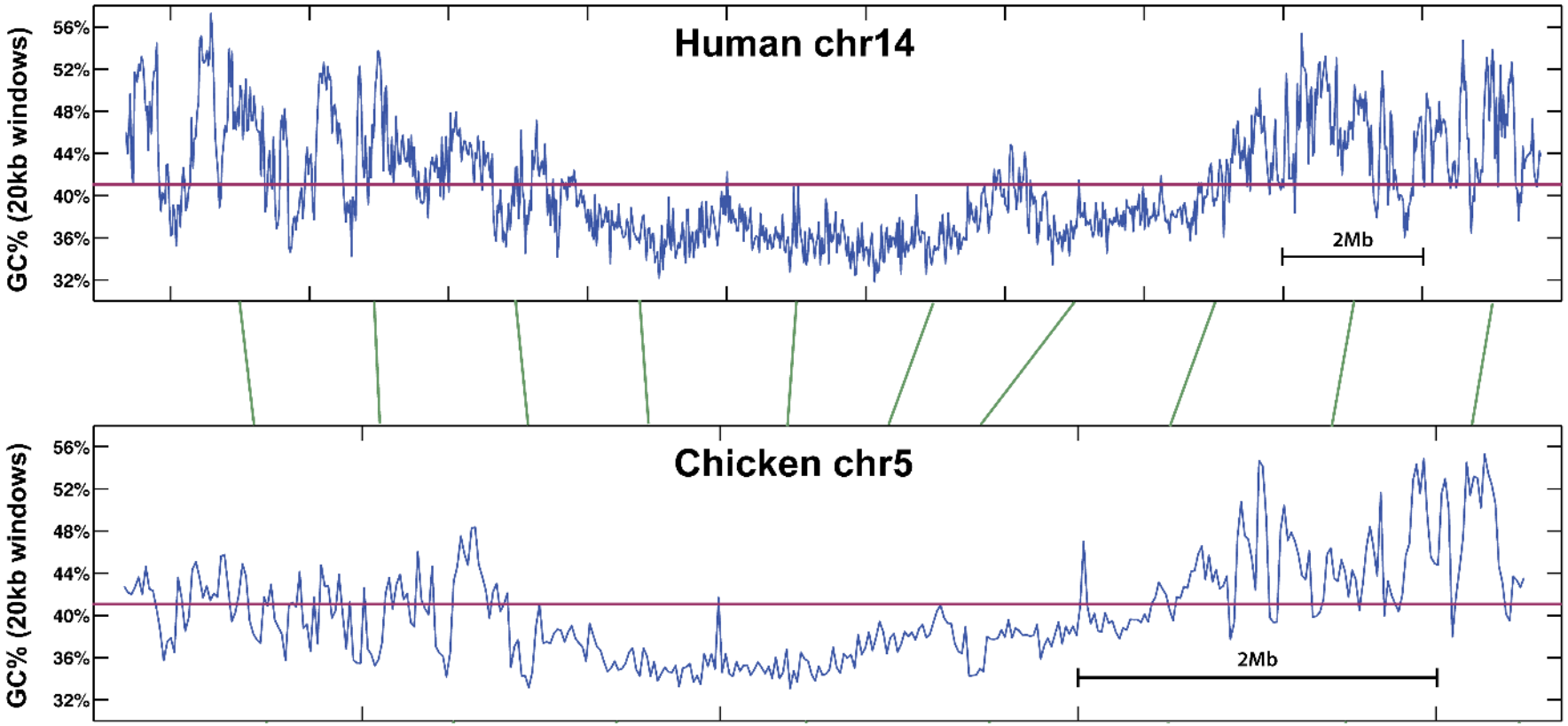
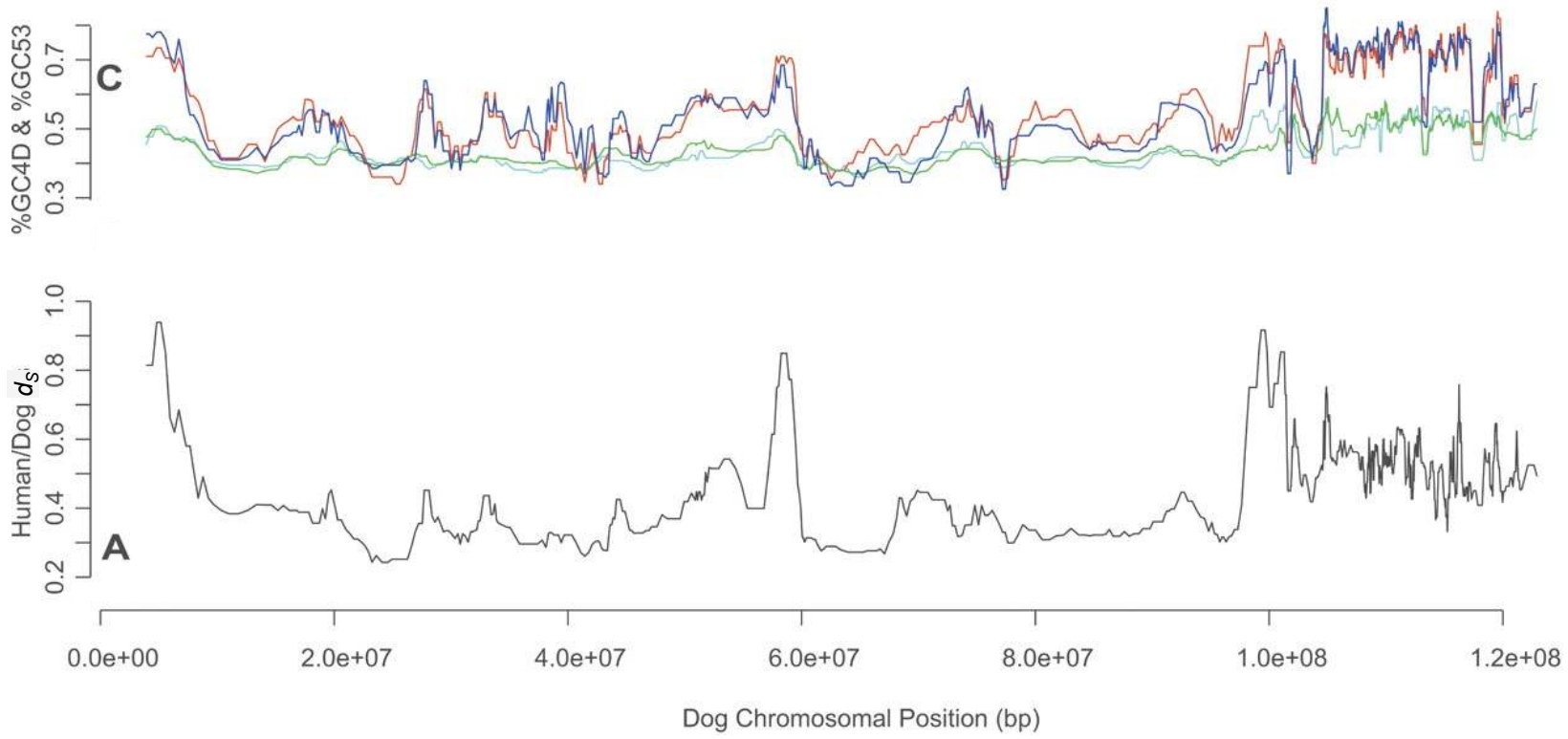


FIGURE 13. Variation in GC content at various scales

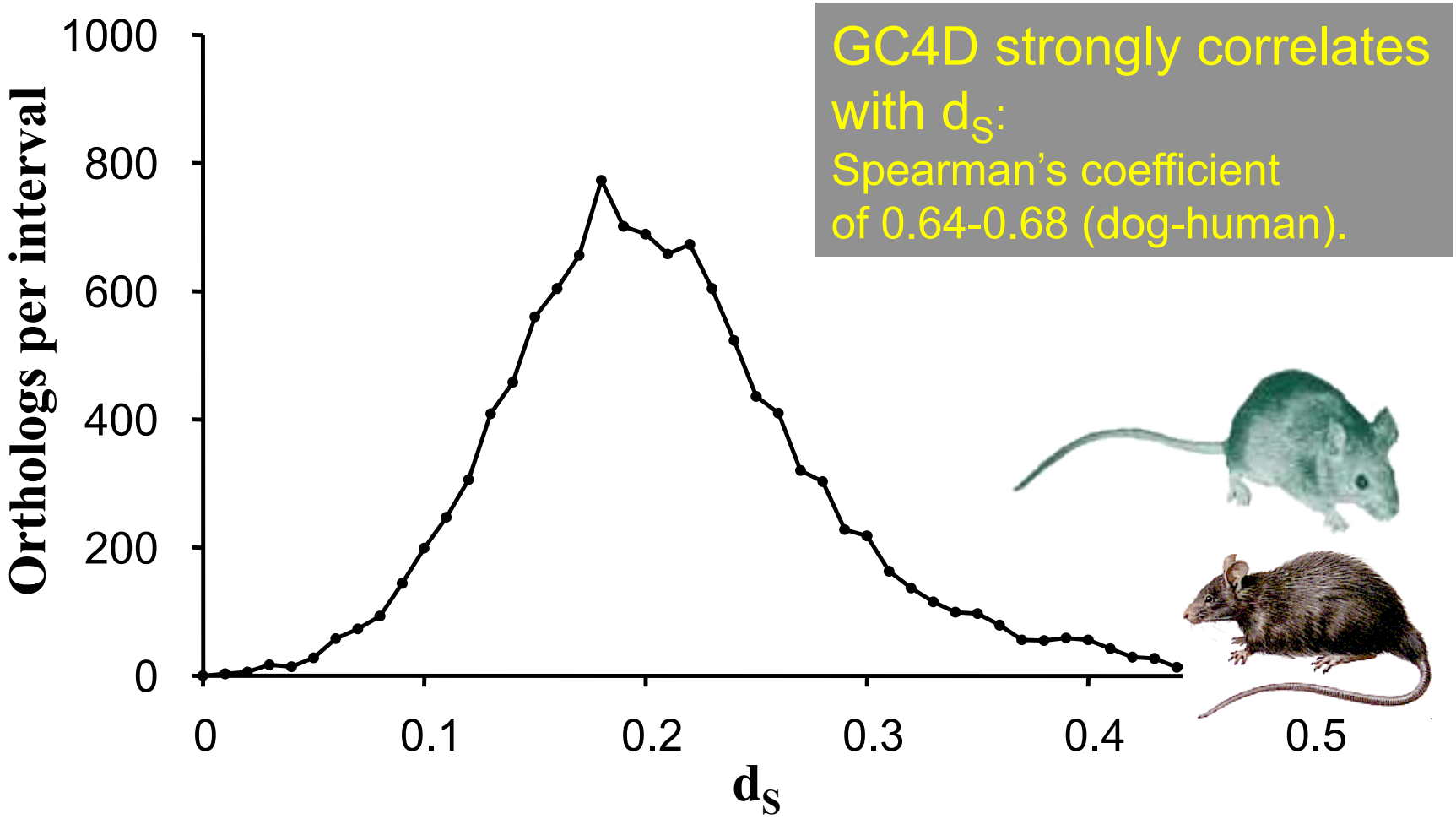
Conservation of Isochores



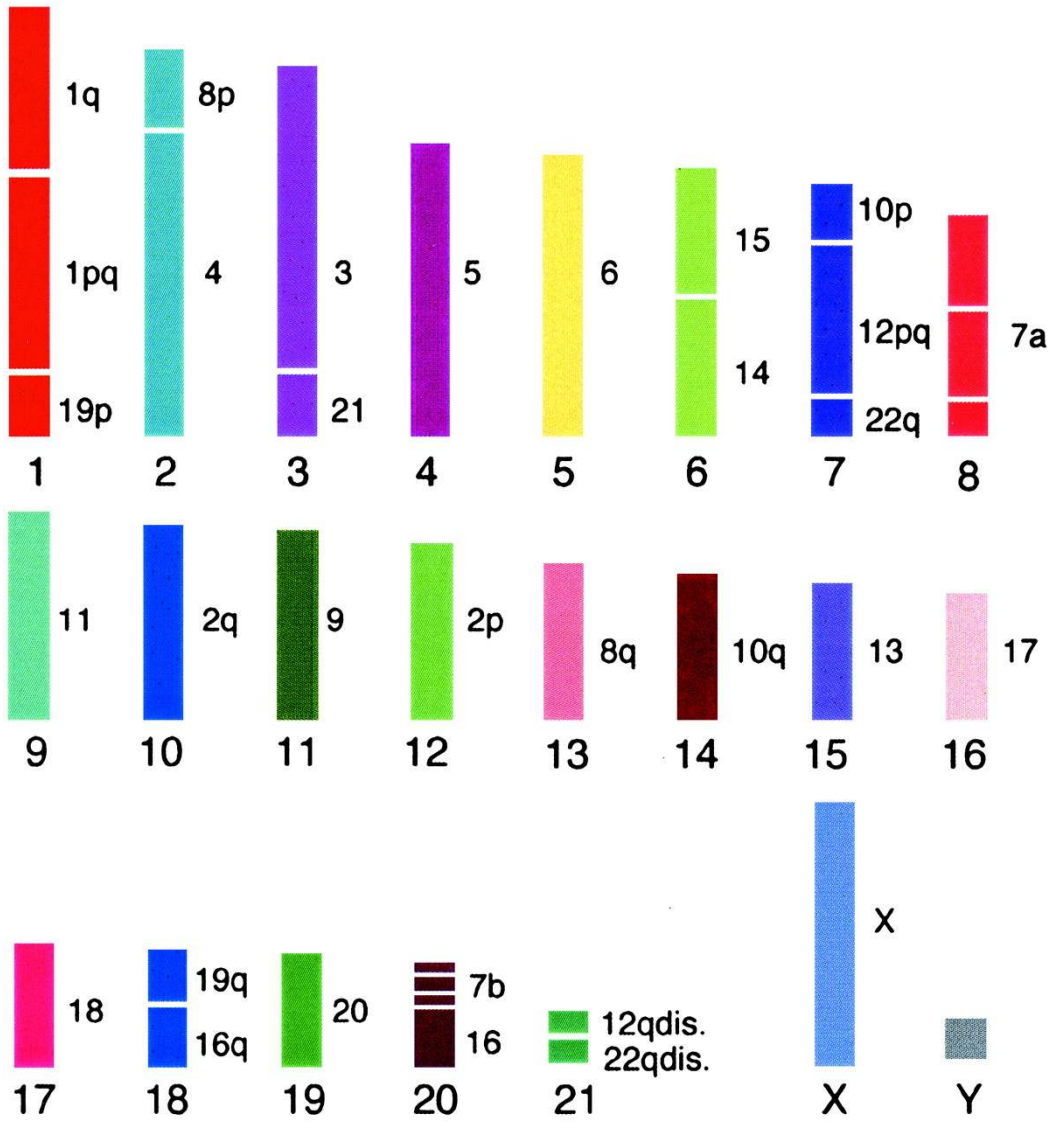
GC4D, GC53 and d_S correlate



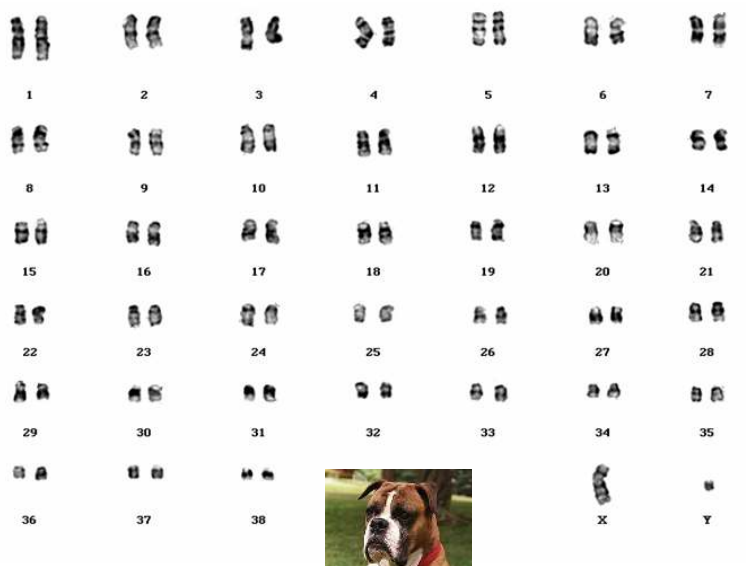
Variation in d_S values

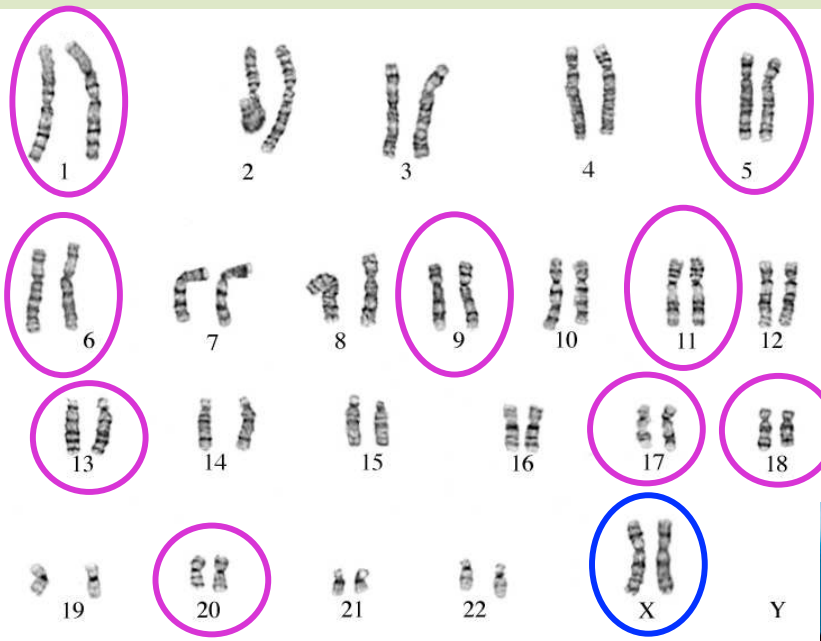


Eutherian ancestral karyotype



Mammalian Karyotypes



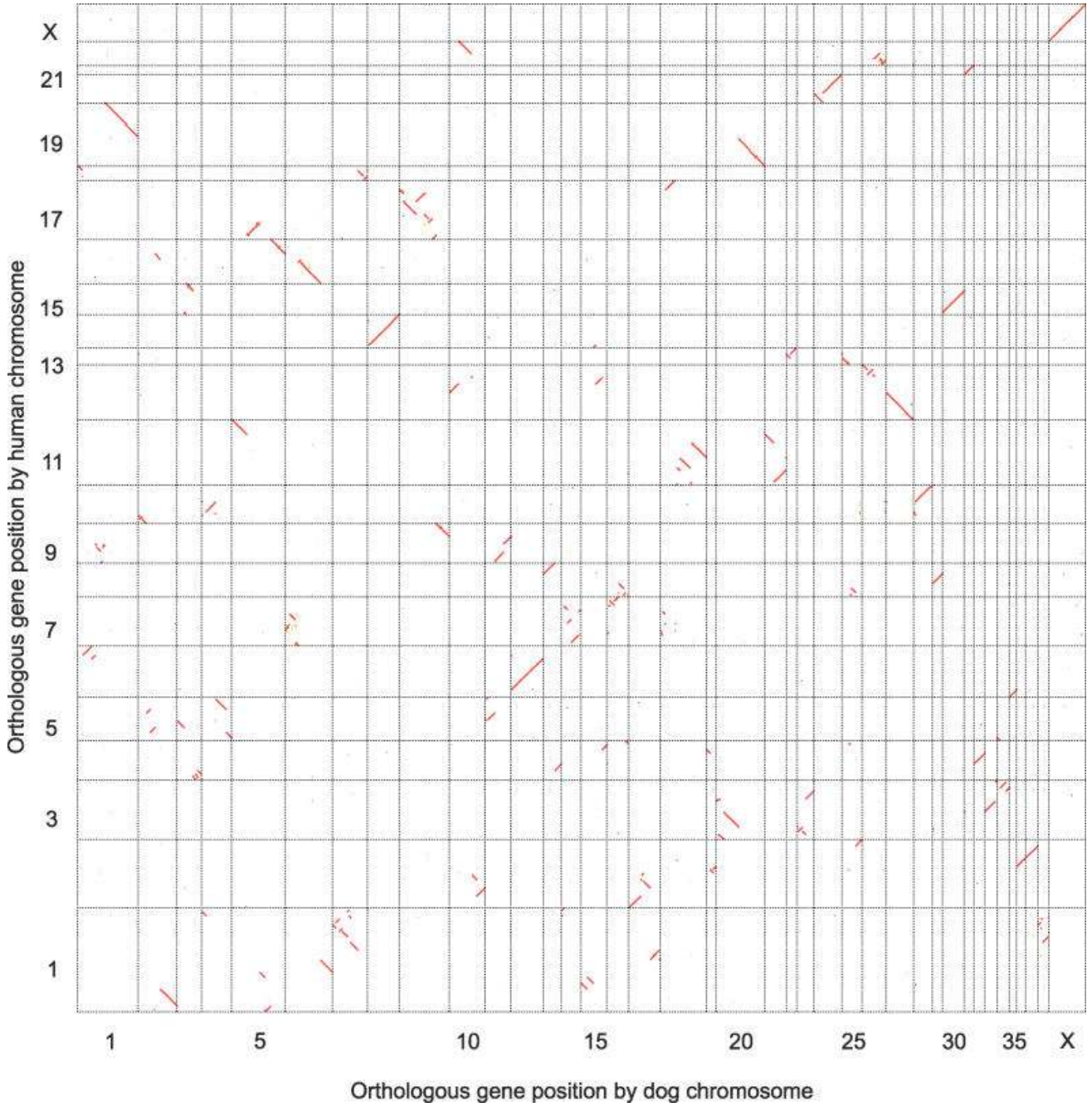


Ancestral chromosomes

Two Models of Rearrangement:

(1) Random : Nadeau and Taylor, 1984; Trinh et al., 2004
(2) Non-random : Pevzner and Tesler, 2003; Bailey et al., 2004

Karyotype Evolution

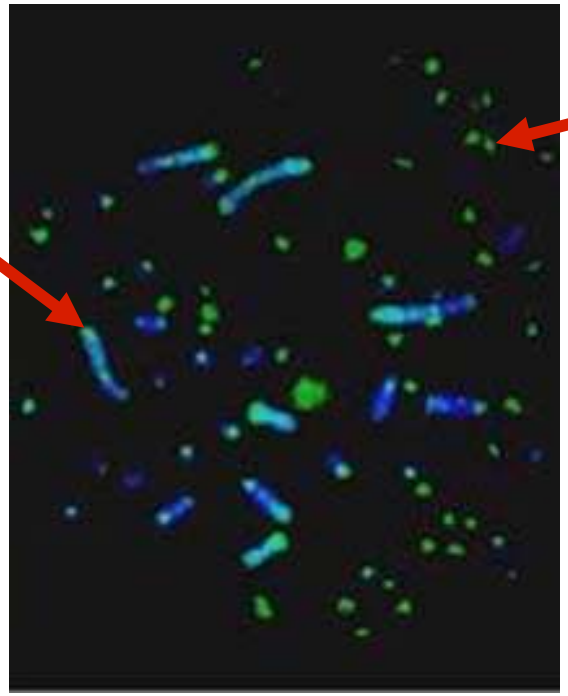


Chicken (avian)
Macro- and Micro-Chromosomes

Large (macro-) chromosomes:

more DNA but lower gene density;

lower mutation rate; lower G+C

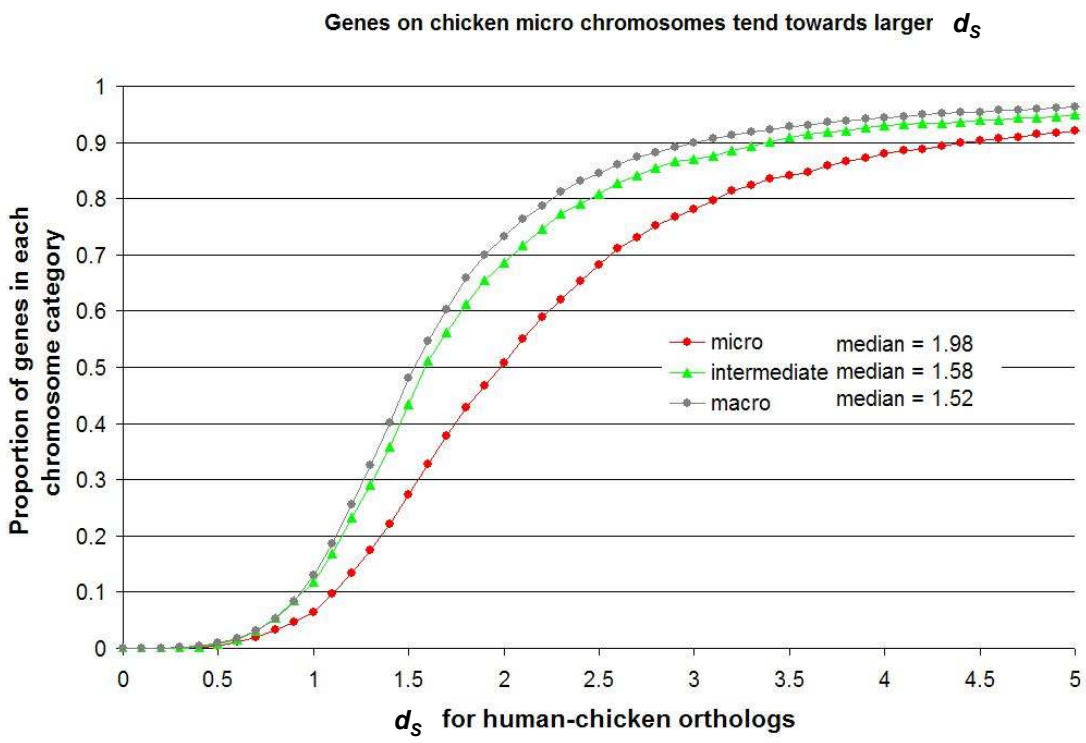


Small (micro-) chromosomes:

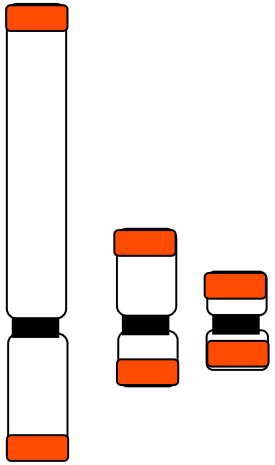
25% of the DNA but half the genes;

higher mutation rate; higher G+C

Chicken-human neutral rates vary between chicken chromosome types



10Mb subtelomeric genes indistinguishable from microchromosomal genes, with respect to d_S



d_S variation

Species	**Subtelomeric / Subtelomeric**	**Subtelomeric / Interstitial**	**Interstitial / Subtelomeric**	**Interstitial / Interstitial**
Chicken / Human	3.15	1.95	2.40	1.53
Dog / Human	0.69	0.35	0.67	0.31
Dog / Chicken	2.15	2.08	1.98	1.64

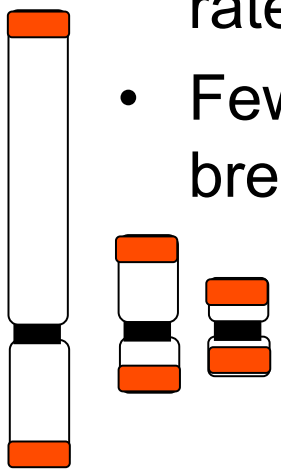
“Hotspots”

Hot

- High G+C content
- High dS values
- High recombination rates
- Strong purifying selection (low dN/dS rates)
- Increased chromosomal breakages

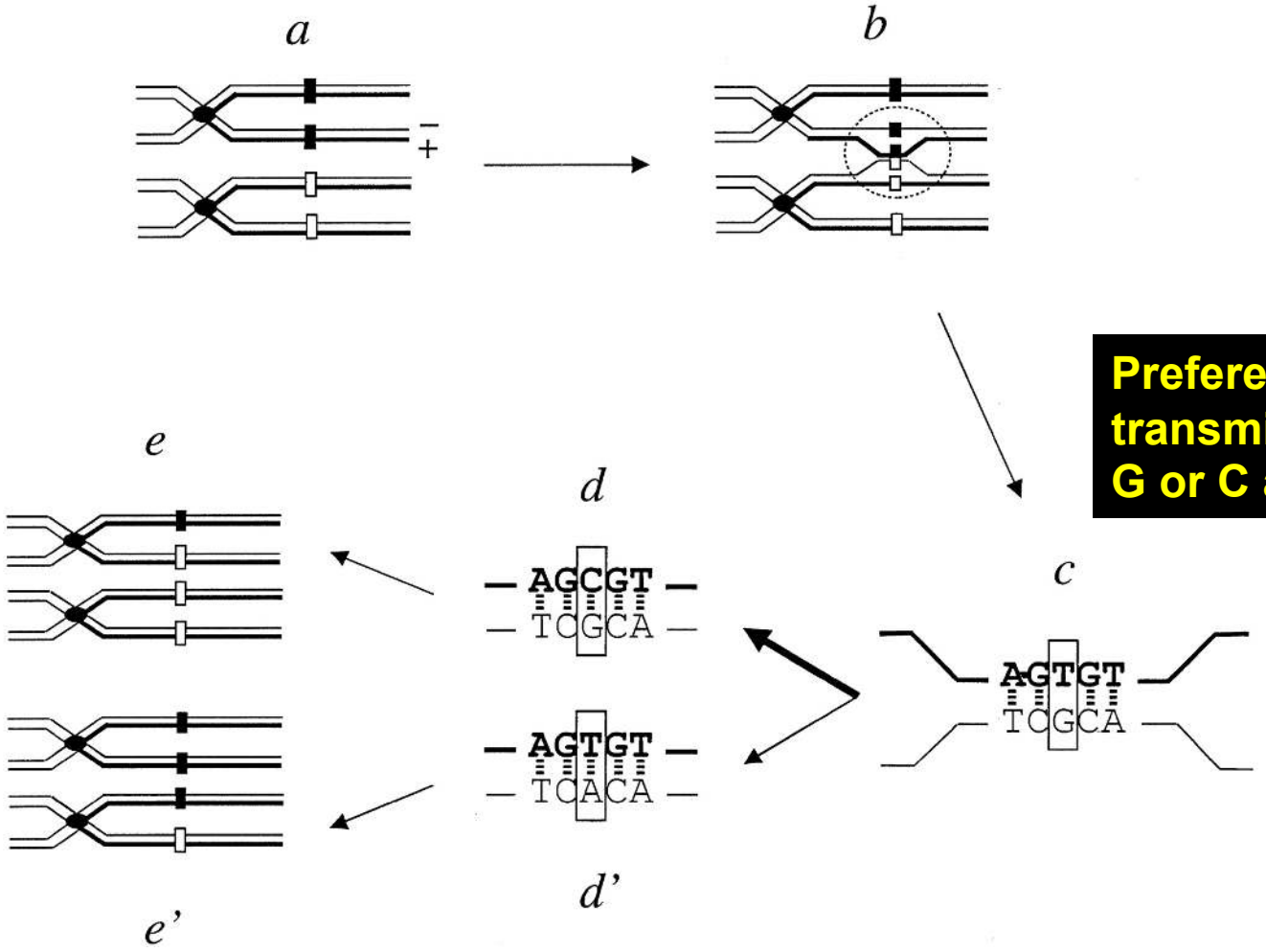
Cold

- Low G+C content
- Low dS values
- Low recombination rates
- Weaker purifying selection (higher dN/dS rates)
- Fewer chromosomal breakages



GC-biased gene conversion

Figure 1. Gene conversion during a recombination event

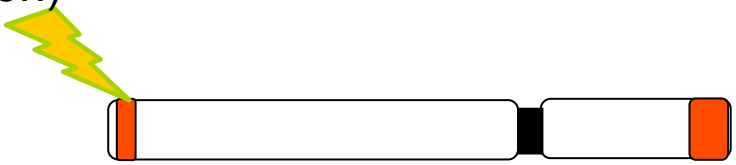


Galtier, N. et al. Genetics 2001;159:907-911

GENETICS

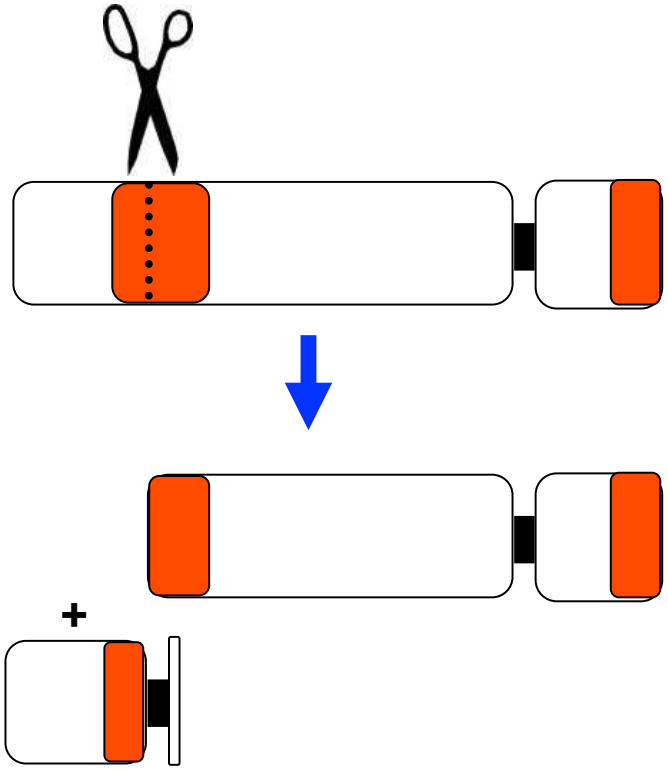
Proposed origin of isochores

Recombination
(GC-biased gene
conversion)



Arose prior to the emergence of both sauropsids (birds, lizards etc.) and mammals.

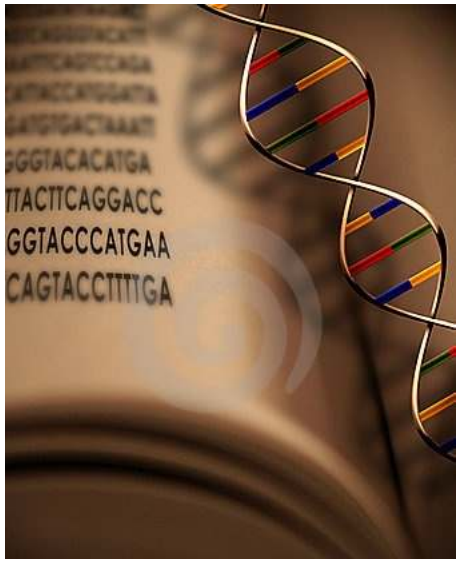
High GC / recombination / dS hotspots at subtelomeres because of breakage?



WHY are the karyotypes of humans, squirrels and cats so stable, whereas those of dogs, mice and rats are not?

June 26, 2000: G-Day

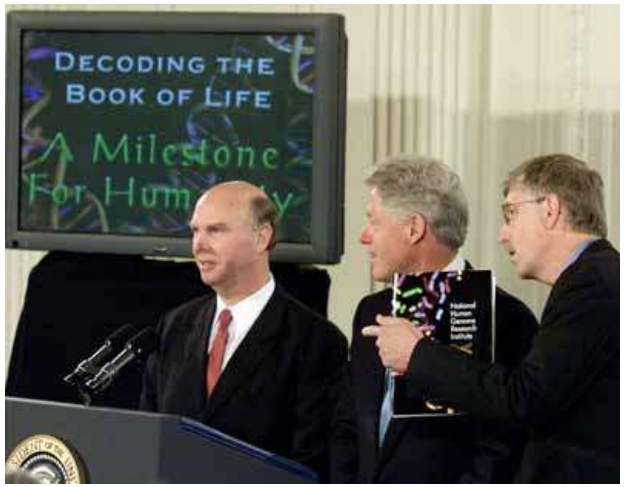
Completion of the Working Draft of the Human Genome



'Day for ages' dawns in medical science

Clinton hails map of human genome

Scientists Announce DNA Mapping



'Milestone for Humanity'

Scientists Decode Blueprint of the Human Body

Scientists complete genetic milestone

'Human alphabet' complete, next step is to make sentences

Tony Blair
"Let us be in no doubt about what we are witnessing today: A revolution in medical science whose implications far surpass even the discovery of antibiotics, the first great technological triumph of the 21st century."

Craig Venter
"Having the genetic code is not a very important moment other than it's the beginning of what we can do with it".

Francis Collins
the benefits of human genome mapping will include "a new understanding of genetic contributions to human disease" and "the development of rational strategies for minimizing or preventing disease phenotypes altogether."

2010

The New York Times

June 12, 2010

A Decade Later, Genetic Map Yields Few New Cures

By NICHOLAS WADE

Ten years after President Bill Clinton announced that the first draft of the human genome was complete, medicine has yet to see any large part of the promised benefits.

Francis Collins (more recently)
"it is fair to say that the Human Genome Project has not yet directly affected the health care of most individuals."

Family under the microscope

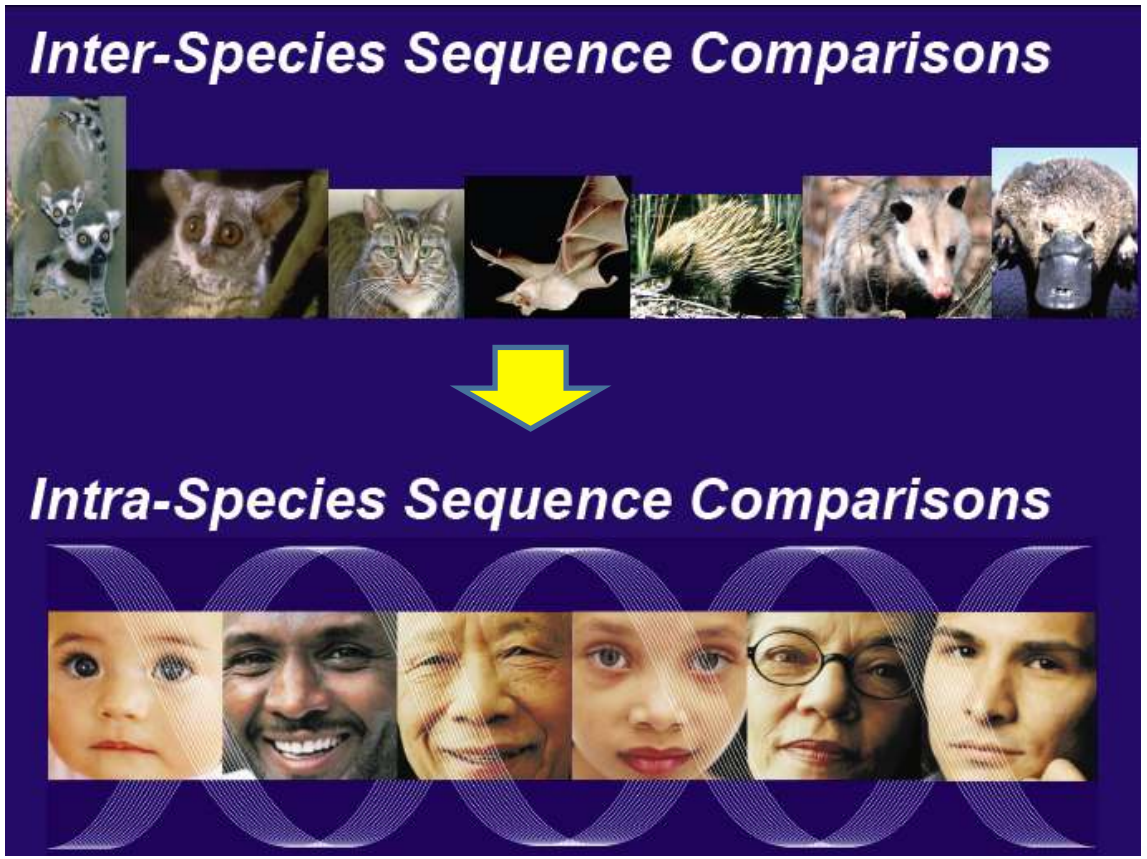
Nature v nurture – what are the latest genetic findings?

Oliver James
The Guardian, Saturday 23 January 2010
Article history

"To the horror of geneticists, Craig Venter, one of the main researchers, pointed out that the fact that we only have about 25,000 genes meant psychological differences between individuals could not be much determined by them – "our environments are critical," he concluded.

Initially, geneticists disputed this, but the last decade has seen an increasingly rapid retreat. After many millions of pounds and thousands of studies, attempts to identify genes that have much effect on our psychology have failed. The most distinguished researchers now admit that it is extremely unlikely that there are single genes for major mental illnesses such as schizophrenia. After decades of hearing from such people that there would be a gene for almost everything, I admit to having felt a twinge of smugness.

Their fallback position is that it's lots of different genes interacting together that matters, but that much remains to be seen."



Differences among us.

Human Diversity

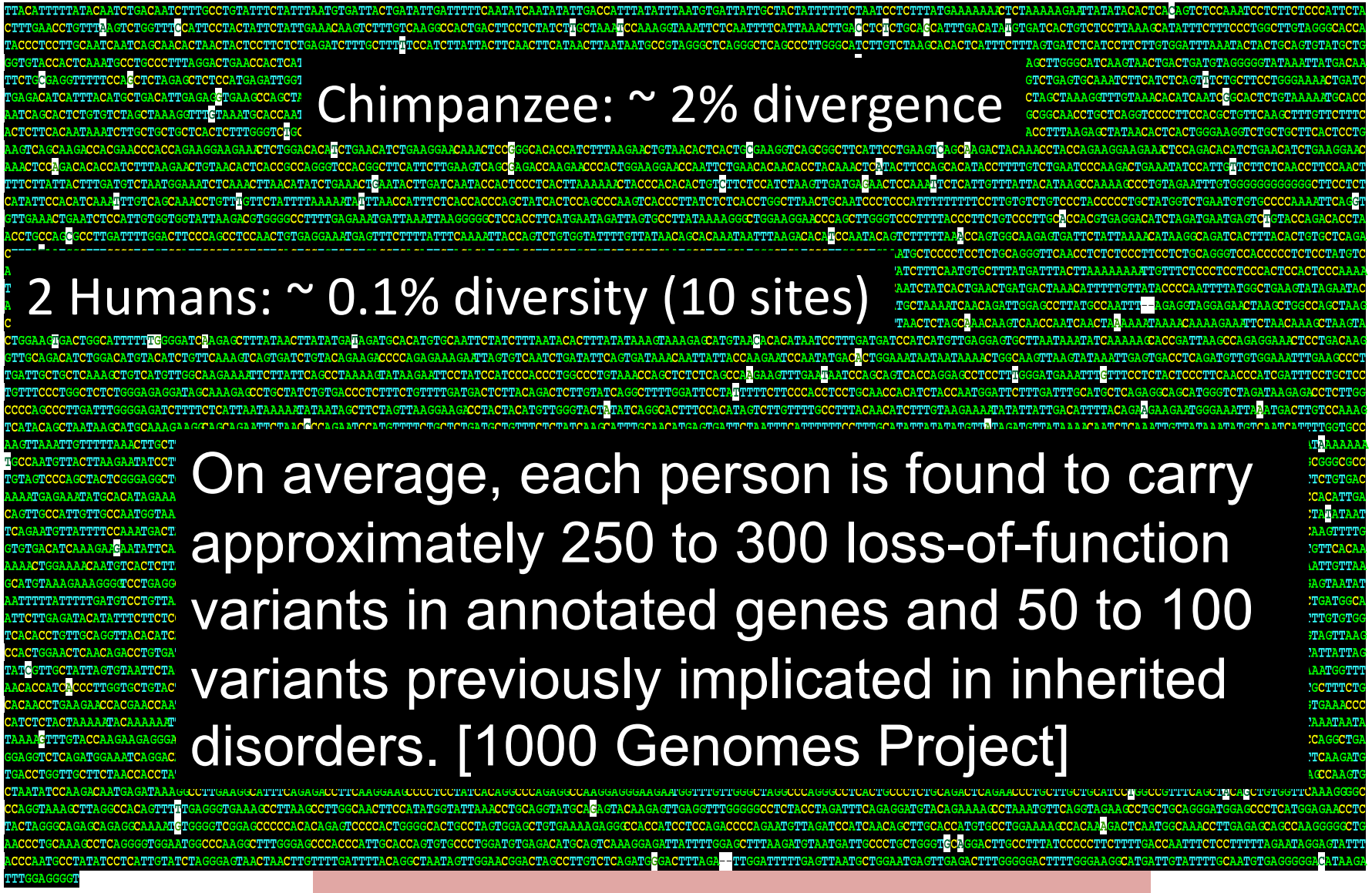
Different, but not that different

- 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

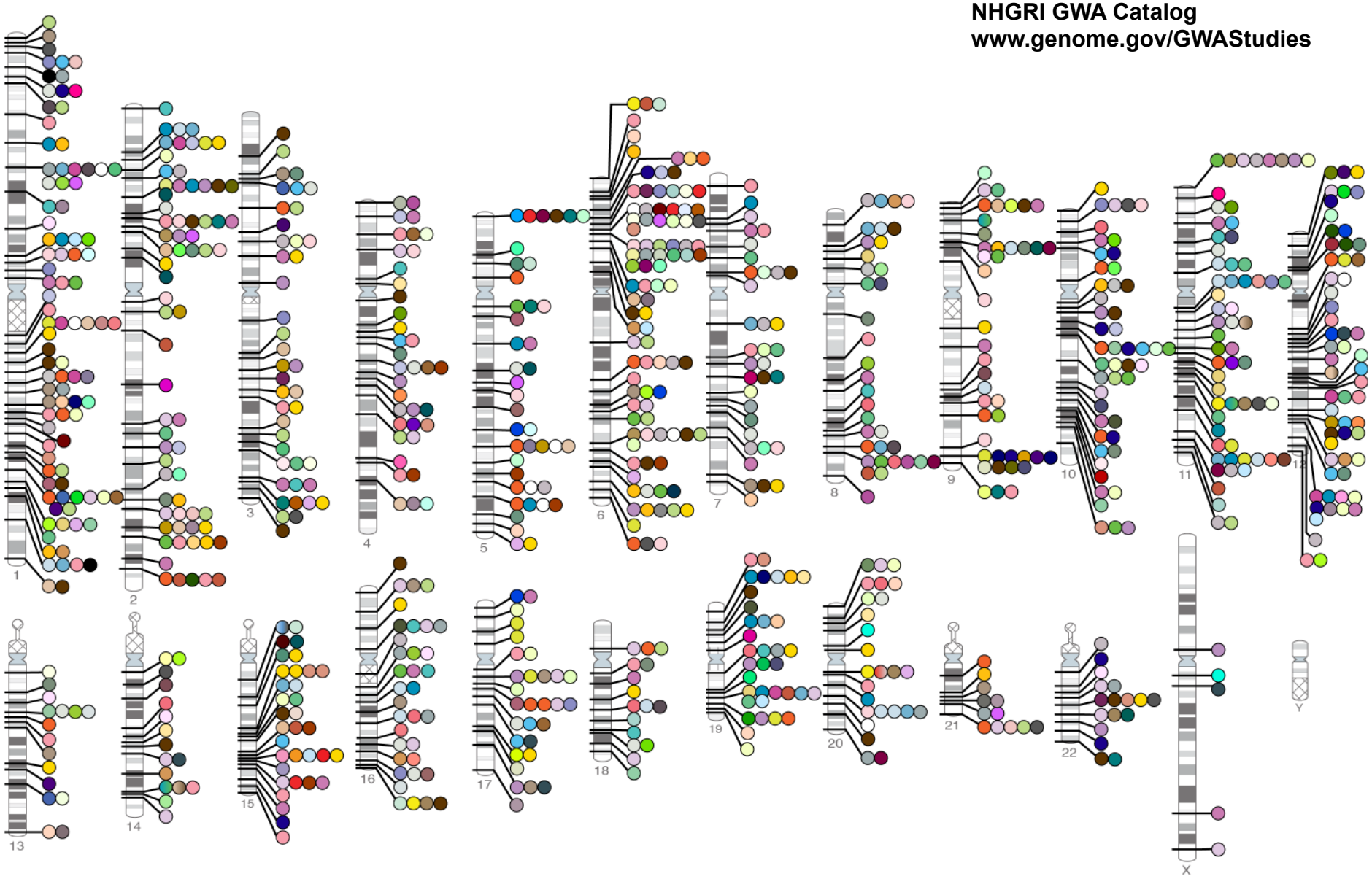
Photos from UN photo gallery www.un.org/av/photo

10,000 bases of human chr13 *vs* chimpanzee



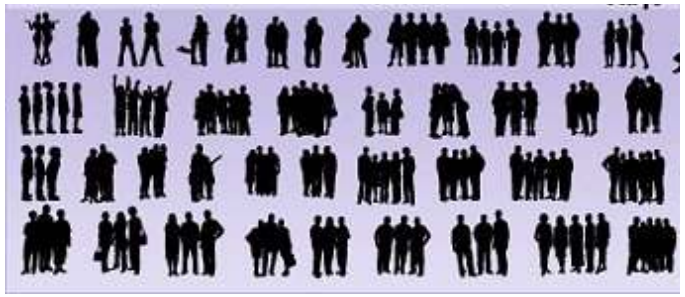
Published Genome-Wide Associations through 6/2010, 904 published GWA at $p \leq 5 \times 10^{-8}$ for 165 traits

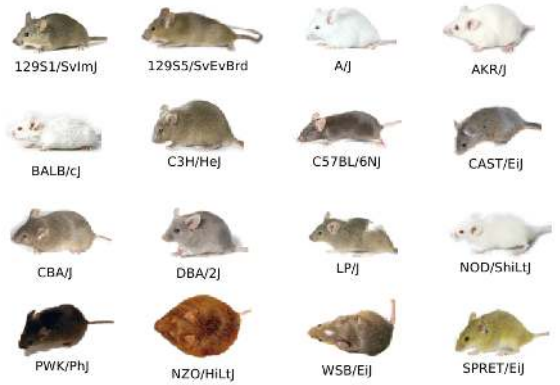
NHGRI GWA Catalog
www.genome.gov/GWAStudies




Acute lymphoblastic leukemia
Adhesion molecules
Adiponectin levels
Age-related macular degeneration
AIDS progression
Alcohol dependence
Alzheimer disease
Amyotrophic lateral sclerosis
Angiotensin-converting enzyme activity
Ankylosing spondylitis
Arterial stiffness
Asthma
Atherosclerosis in HIV
Atrial fibrillation
Attention deficit hyperactivity disorder
Autism
Basal cell cancer
Bipolar disorder
Biliary atresia
Bilirubin
Birth weight
Bladder cancer
Blond or brown hair
Blood pressure
Blue or green eyes
BMI, waist circumference
Bone density
Breast cancer
C-reactive protein
Cardiac structure/function
Carnitine levels
Carotenoid/tocopherol levels
Celiac disease
Chronic lymphocytic leukemia
Cleft lip/palate
Cognitive function
Conduct disorder
Colorectal cancer
Corneal thickness
Coronary disease
Creutzfeldt-Jakob disease
Crohn's disease
Cutaneous nevi
Dermatitis
Drug-induced liver injury
Eosinophil count
Eosinophilic esophagitis
Erythrocyte parameters
Esophageal cancer
Essential tremor
Exfoliation glaucoma
Eye color traits
F cell distribution
Fibrinogen levels
Folate pathway vitamins
Freckles and burning
Gallstones
Glioma
Glycemic traits
Hair color
Hair morphology
HDL cholesterol
Heart failure
Heart rate
Height
Hemostasis parameters
Hepatitis
Hirschsprung's disease
HIV-1 control
Homocysteine levels
Idiopathic pulmonary fibrosis
IgE levels
Inflammatory bowel disease
Intracranial aneurysm
Iris color
Iron status markers
Ischemic stroke
Juvenile idiopathic arthritis
Kidney stones
LDL cholesterol
Leprosy
Leptin receptor levels
Liver enzymes
LP (a) levels
LpPLA(2) activity and mass
Lung cancer
Major mood disorders
Malaria
Male pattern baldness
Matrix metalloproteinase levels
MCP-1
Melanoma
Menarche & menopause
Multiple sclerosis
Myeloproliferative neoplasms
Narcolepsy
Nasopharyngeal cancer
Neuroblastoma
Nicotine dependence
Obesity
Open angle glaucoma
Open personality
Optic disc parameters
Osteoarthritis
Osteoporosis
Otosclerosis
Other metabolic traits
Ovarian cancer
Pancreatic cancer
Pain
Paget's disease
Panic disorder
Parkinson's disease
Periodontitis
Peripheral arterial disease
Phosphatidylcholine levels
Phytosterol levels
Platelet count
Primary biliary cirrhosis
PR interval
Prostate cancer
Protein levels
Psoriasis
Pulmonary funct. COPD
QRS interval
QT interval
Quantitative traits
Recombination rate
Red vs.non-red hair
Renal function
Response to antidepressants
Response to antipsychotic therapy
Response to hepatitis C treat
Response to statin therapy
Restless legs syndrome
Rheumatoid arthritis
Schizophrenia
Serum metabolites
Skin pigmentation
Smoking behavior
Speech perception
Sphingolipid levels
Statin-induced myopathy
Stroke
Systemic lupus erythematosus
Systemic sclerosis
Telomere length
Testicular germ cell tumor
Thyroid cancer
Tooth development
Total cholesterol
Triglycerides
Type 1 diabetes
Type 2 diabetes
Ulcerative colitis
Urate
Venous thromboembolism
Vertical cup-disc ratio
Vitamin B12 levels
Vitamin D insuffiency
Vitiligo
Warfarin dose
Weight
White cell count
YKL-40 levels

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



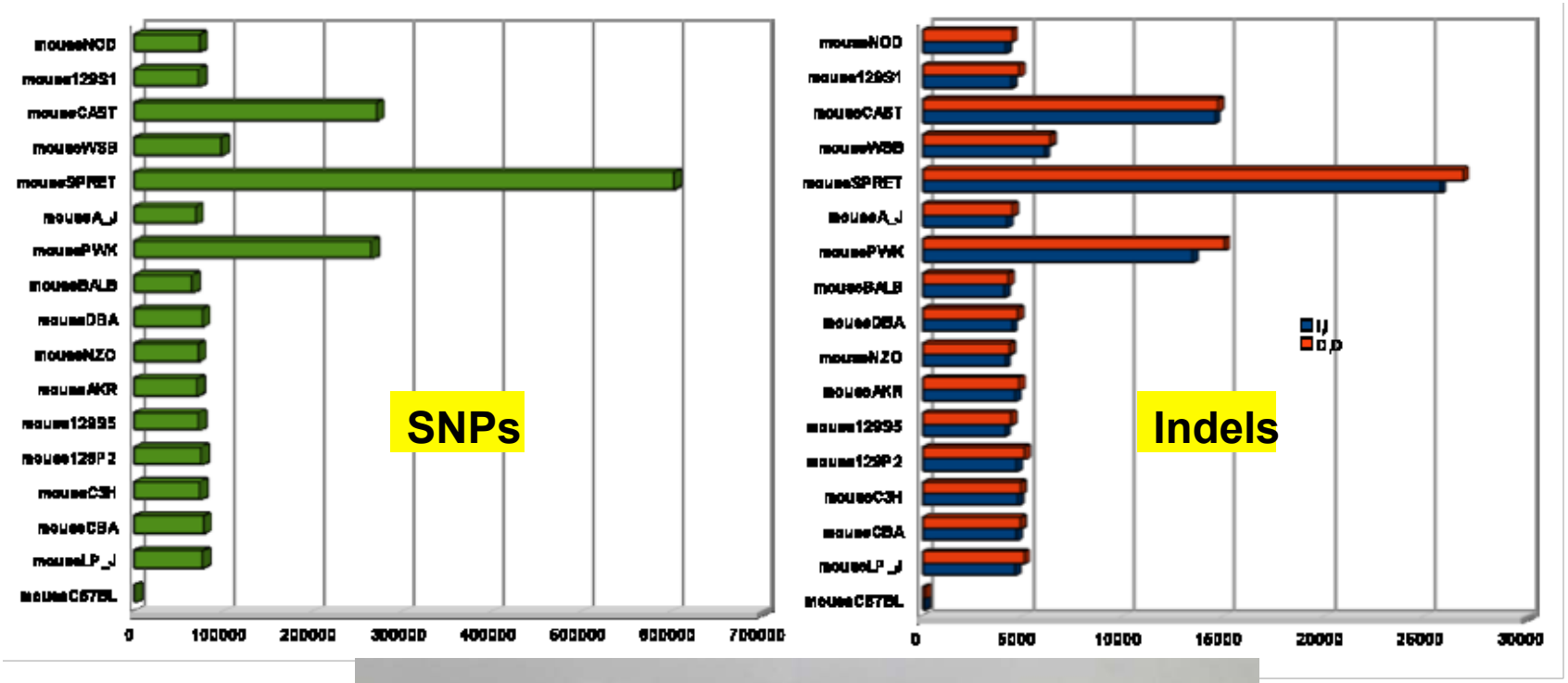


- Genetic variation must underlie both pathological and non-pathological traits that show significant heritability
  - How do we locate these variants, and is there clinical use when they are found?
- Genetic variation must also underlie species differences.
  - How do we locate these variants?
- Do orthologous genes control equivalent traits in different species?
  - Can model organisms appropriately model human traits?

Population genomics requires detailed phenotyping

Vertebrate population genomics will initially study human and rodent species.

Genome Sequences of 17 Mouse Strains



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

GENOME 10K©

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

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
training future leaders in genomic research: hiring from 2011 for 2012
MRC Functional Genomics Unit