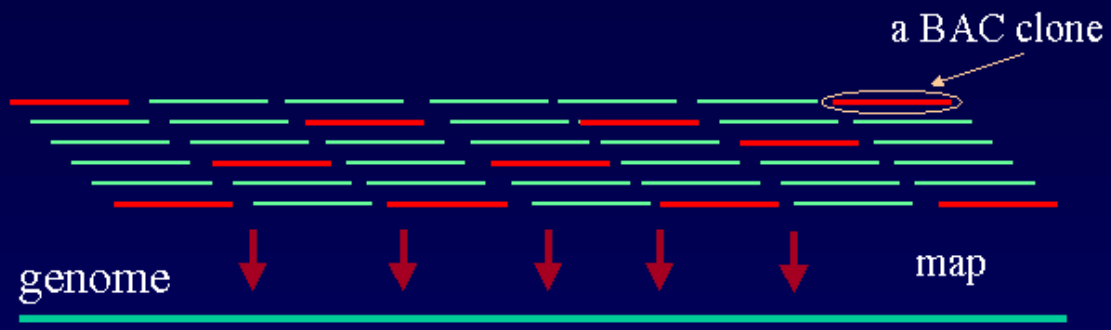
Lecture 6: introduction to human genome and mammalian comparative genomics

Outline

1. How to sequence a genome ?
2. Basic facts of mammalian genomes
3. Whole genome alignment and synteny
4. Genome arrangement, duplication
5. Ultra-conserved elements
6. Detecting functional elements: Phylogenetic footprinting and shadowing
7. Evolution of genes
8. Evolution of gene expression

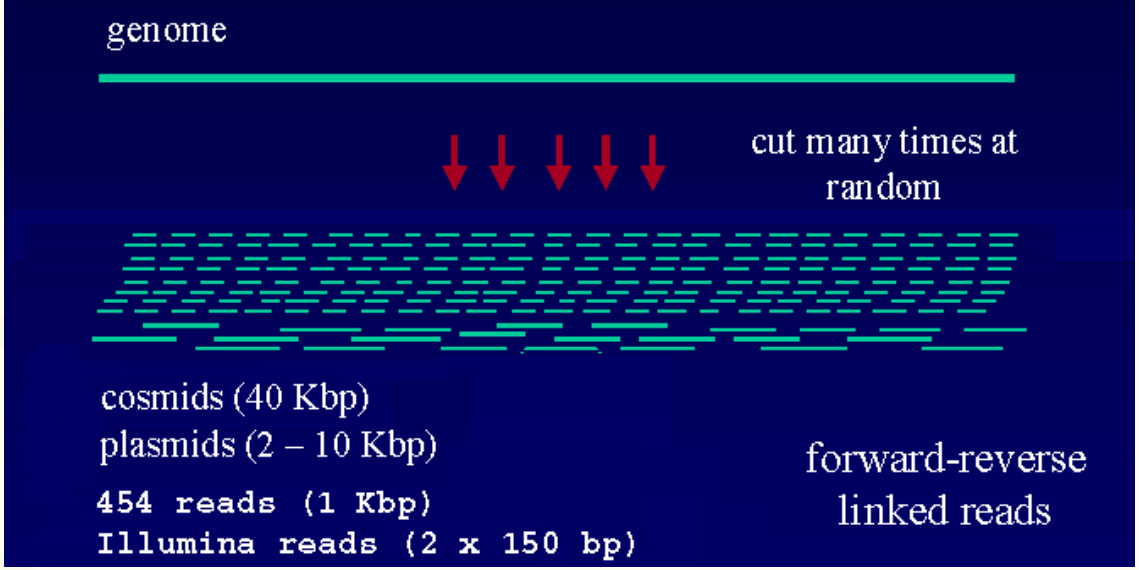
Clone-by-Clone: Physical Mapping



- Get a set of large clones (BAC, ~150 Kbp)

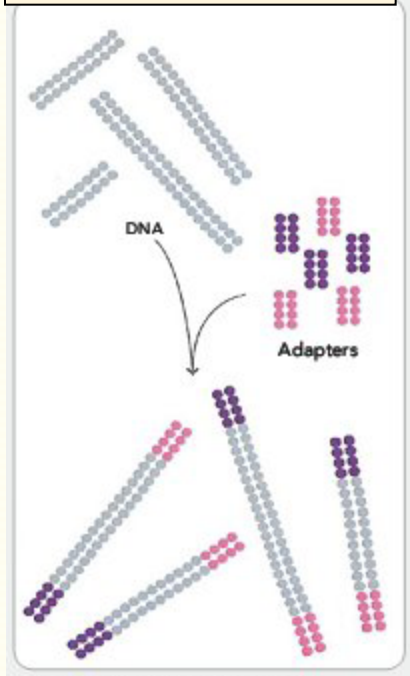
Genome Sequencing and Assembly

Whole Genome Shotgun Sequencing

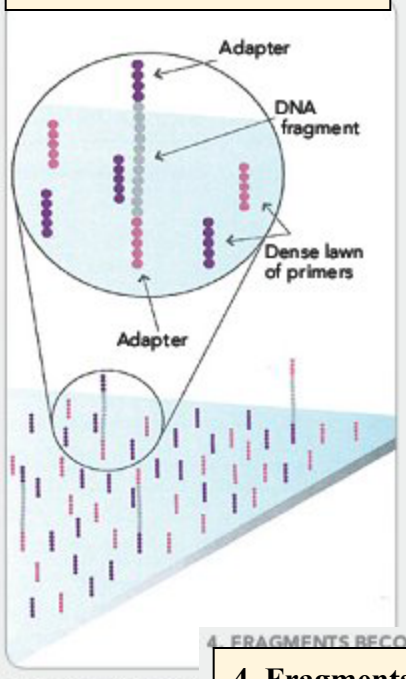


Modified from M. Kellis

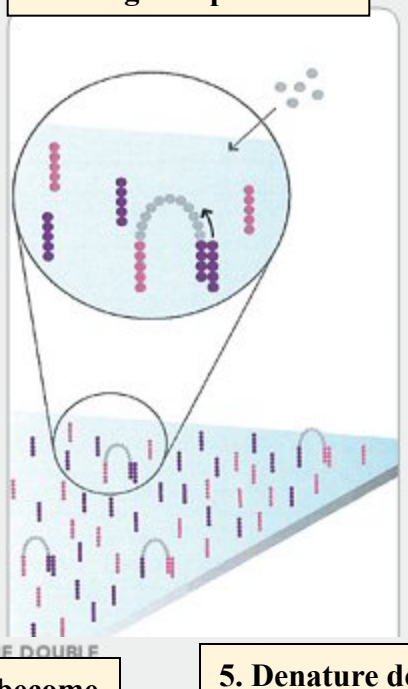
1. Prepare Genomic DNA



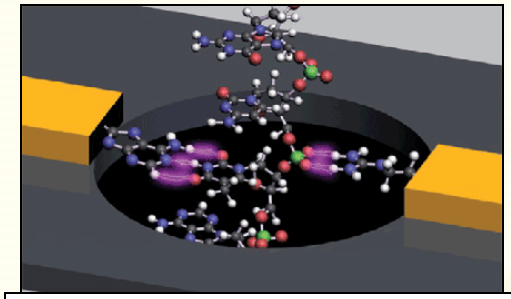
2. Attach DNA to surface

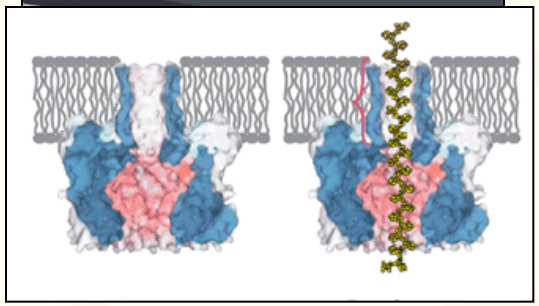


3. Bridge Amplification



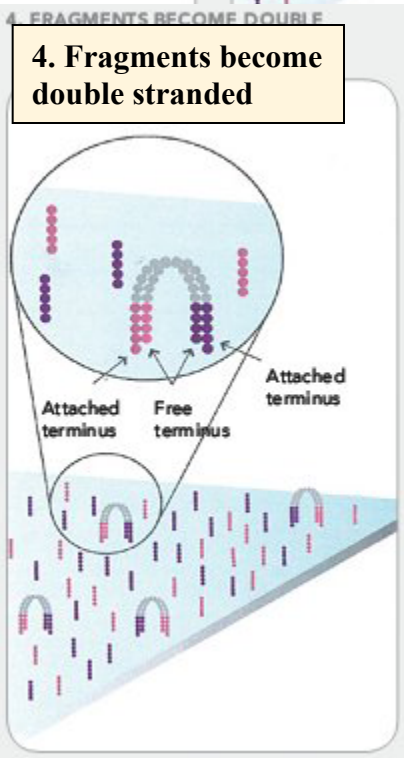
Illumina Solexa platform: Sequencing by synthesis



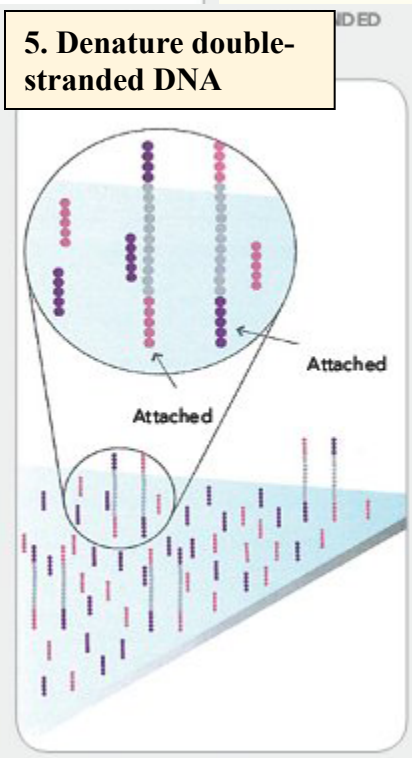


Nanopore: next-next generation

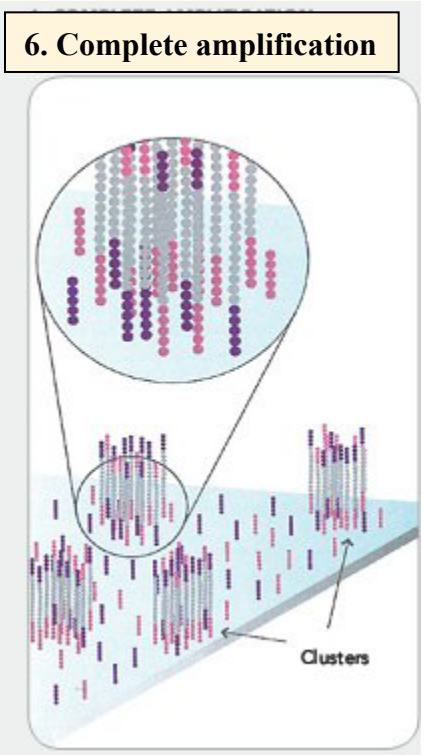
4. Fragments become double stranded



5. Denature double-stranded DNA



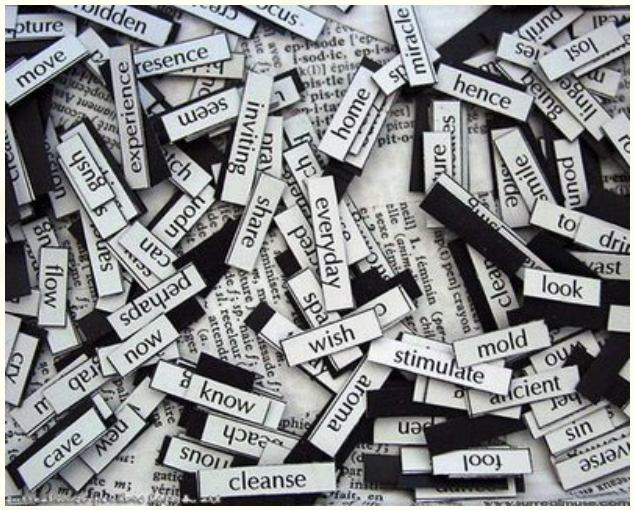
6. Complete amplification



Genome assembly

- De novo assembly
  - It is like putting words back together into a 3 billion letter book

- Use a reference genome
  - If we already know one person's genome sequence and now sequenced another person's genome

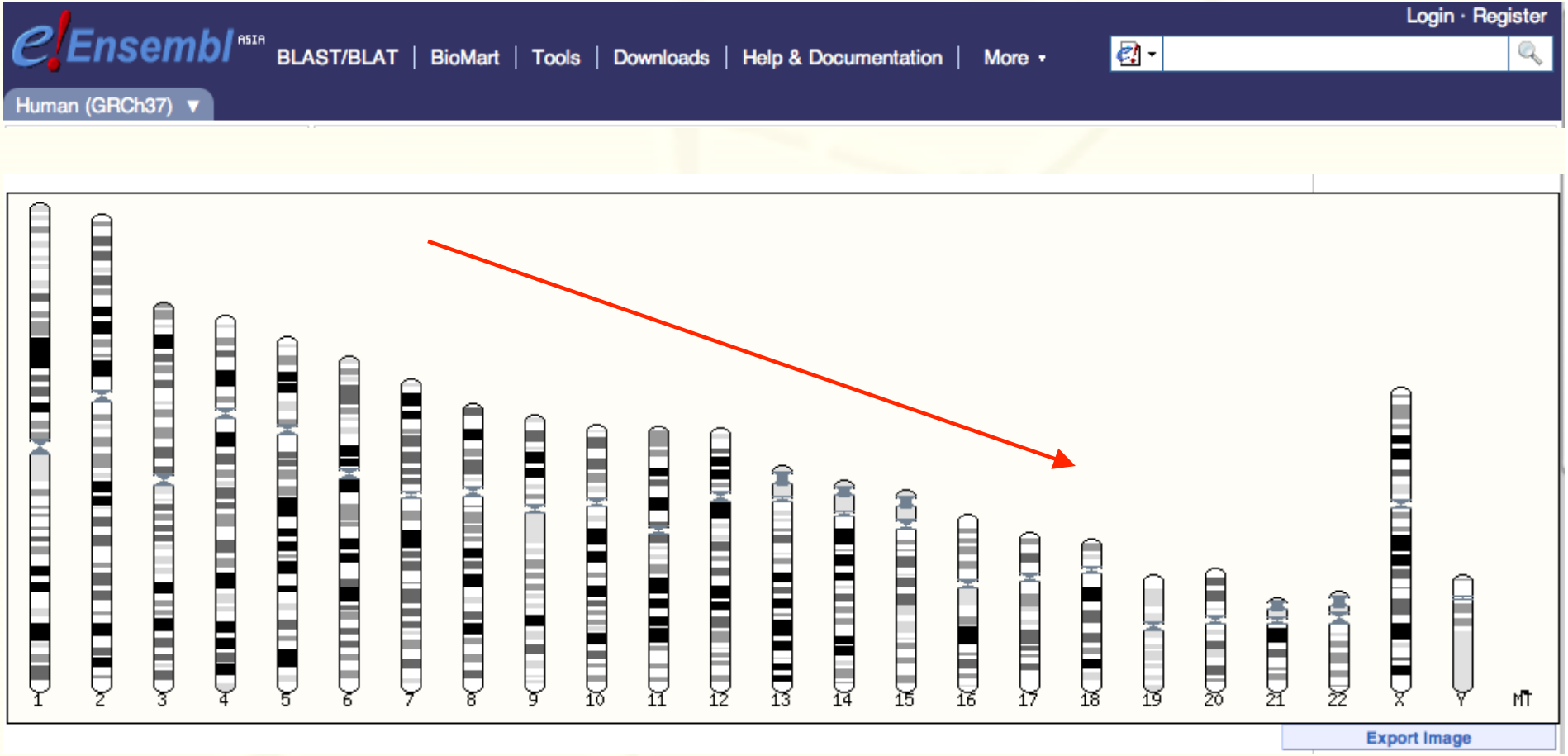


What to do with a genome (3 billions of A, C, G, T) ?

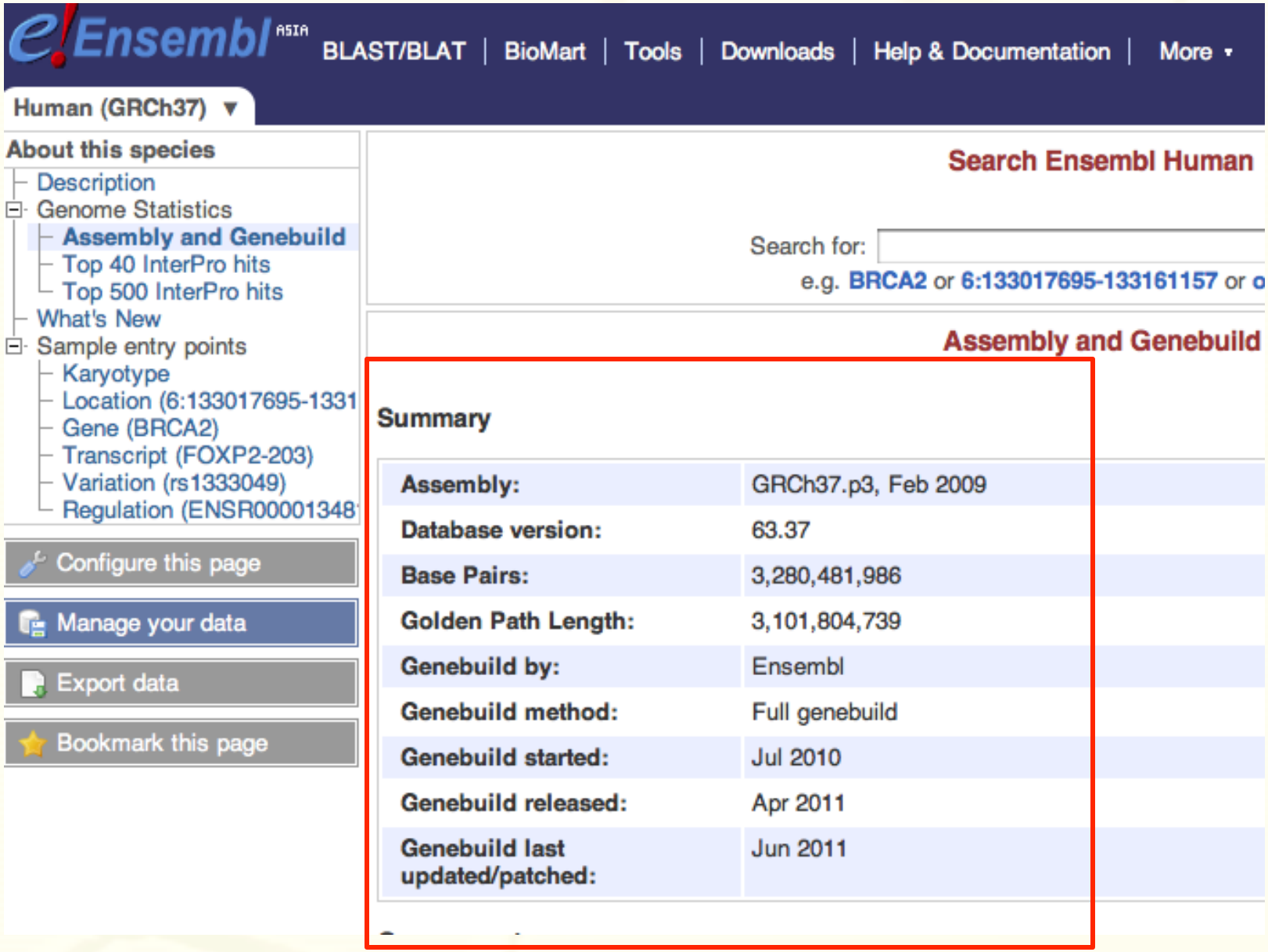
- **Quality control:**
  - Sanger re-sequencing, checking error rate, polymorphisms
- **Repetitive elements:**
  - More than 50% of the mammalian genome
- **Gene prediction:**
  - by sequence homology, *ab initio* (HMM), transcript mapping
- **Noncoding RNA:**
  - rRNA, tRNA, snoRNA, microRNA …
- **Genome alignment:**
  - synteny, conservation, duplication…
- **Genome evolution:**
  - Rate of substitution, % of sequence under conservation and selection …
  - Lineage-specific gene gain and loss, adaptation
  - Pseudogene: loss of *umami* taste receptor in panda

Human intervention is necessary,
data management and presentation (database) is a big challenge

www.ensembl.org



human


Ensembl ASIA
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | More
Human (GRCh37)
About this species
Description
Genome Statistics
Assembly and Genebuild
Top 40 InterPro hits
Top 500 InterPro hits
What's New
Sample entry points
Karyotype
Location (6:133017695-1331
Gene (BRCA2)
Transcript (FOXP2-203)
Variation (rs1333049)
Regulation (ENSR00001348
Configure this page
Manage your data
Export data
Bookmark this page
Search Ensembl Human
Search for:
e.g. BRCA2 or 6:133017695-133161157 or o
Assembly and Genebuild
Summary
Assembly: GRCh37.p3, Feb 2009
Database version: 63.37
Base Pairs: 3,280,481,986
Golden Path Length: 3,101,804,739
Genebuild by: Ensembl
Genebuild method: Full genebuild
Genebuild started: Jul 2010
Genebuild released: Apr 2011
Genebuild last updated/patched: Jun 2011

Ensembl "Genes"

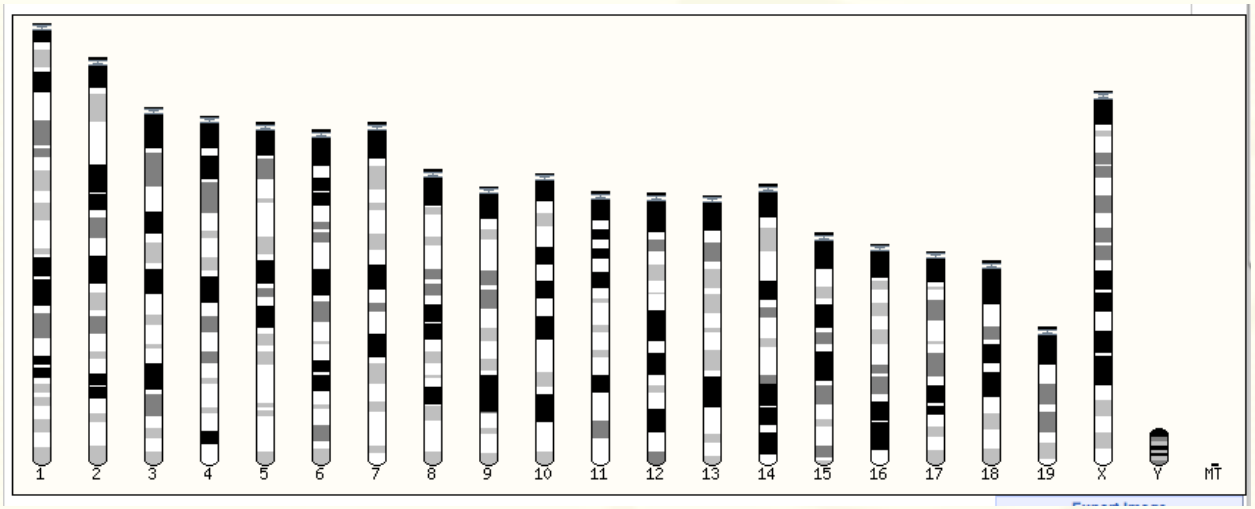
Gene counts

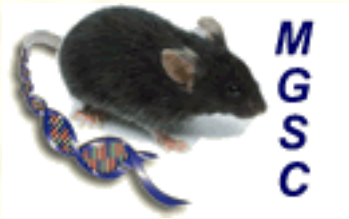
Known protein-coding genes:	20,599
Novel protein-coding genes:	895
Pseudogenes:	14,012
RNA genes:	8,563
Immunoglobulin/T-cell receptor gene segments:	556
Gene exons:	631,122
Gene transcripts:	174,416

Other

Genscan gene predictions:	46,737
Short Variants (SNPs, indels, somatic mutations):	30,095,750

- Known genes:
  - Predicted genes that have good experimental evidence
- Novel genes:
  - Have sequence similarity to known genes, but not 100% identical
- Pseudogenes:
  - dead genes
- RNA genes:
  - tRNA, rRNA, microRNA, snoRNA, noncoding RNAs,
- Gene transcripts
  - multiple transcripts per gene
- Genscan prediction:
  - –ab initio prediction based on HMM model





Summary

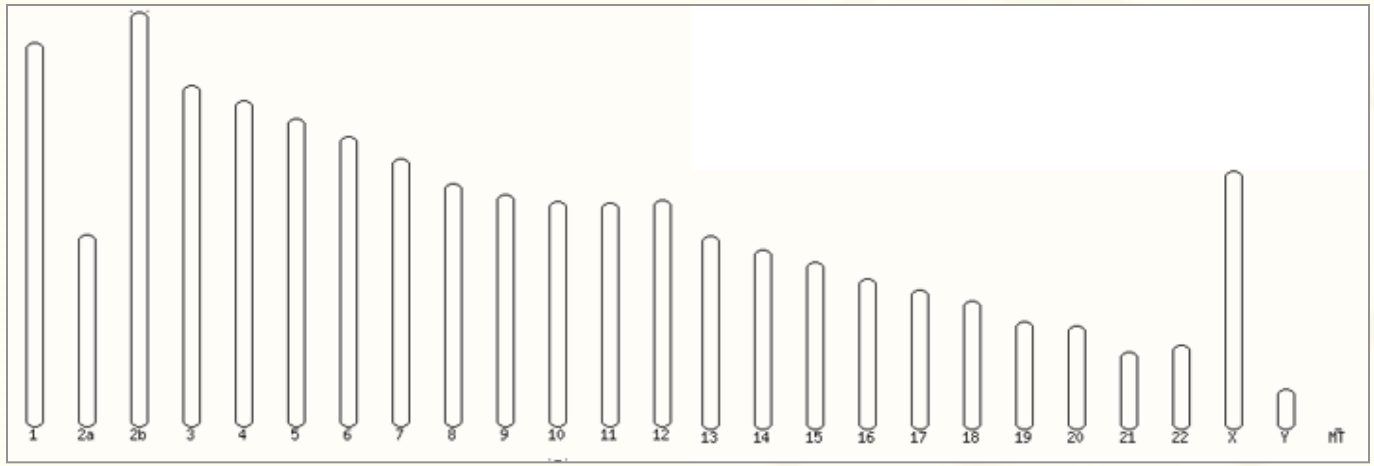
Assembly:	NCBIM37, Apr 2007
Database version:	63.37
Base Pairs:	3,420,842,930
Golden Path Length:	2,716,965,481
Genebuild by:	Ensembl
Genebuild method:	Full genebuild
Genebuild started:	Apr 2010
Genebuild released:	Jan 2011
Genebuild last updated/patched:	Apr 2011

Gene counts

Known protein-coding genes:	21,873
Novel protein-coding genes:	794
Pseudogenes:	4,948
RNA genes:	6,256
Immunoglobulin/T-cell receptor gene segments:	481
Gene exons:	404,826
Gene transcripts:	93,805

Other

Genscan gene predictions:	46,375
Short Variants (SNPs, indels, somatic mutations):	15,429,547



Summary

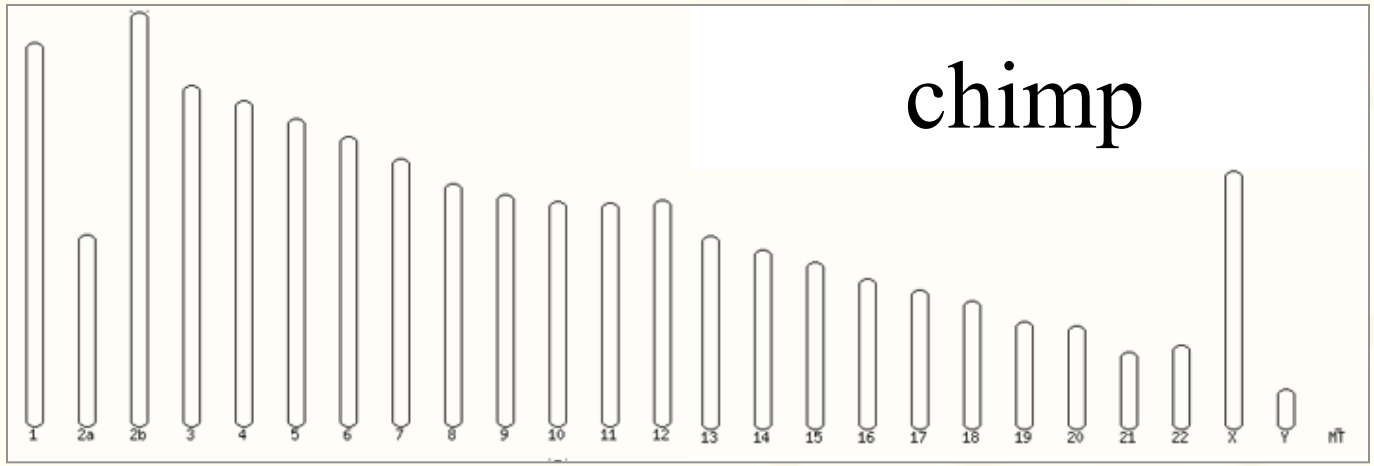
Assembly:	CHIMP2.1, Mar 2006
Database version:	63.21
Base Pairs:	2,928,559,526
Golden Path Length:	3,350,413,343
Genebuild by:	Ensembl
Genebuild method:	Projection build
Genebuild started:	Feb 2008
Genebuild released:	Jul 2008
Genebuild last updated/patched:	May 2010

Gene counts

Known protein-coding genes:	16,152
Projected protein-coding genes:	2,360
Novel protein-coding genes:	1,317
Pseudogenes:	471
RNA genes:	6,865
Gene exons:	18,743
Gene transcripts:	9,826

Other

Genscan gene predictions:	126,538
Short Variants (SNPs, indels, somatic mutations):	1,520,076



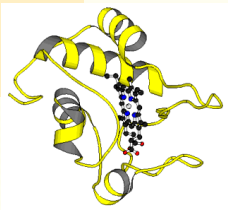
Genebuild method:	Projection build
Genebuild started:	Feb 2008
Genebuild released:	Jul 2008
Genebuild last updated/patched:	May 2010

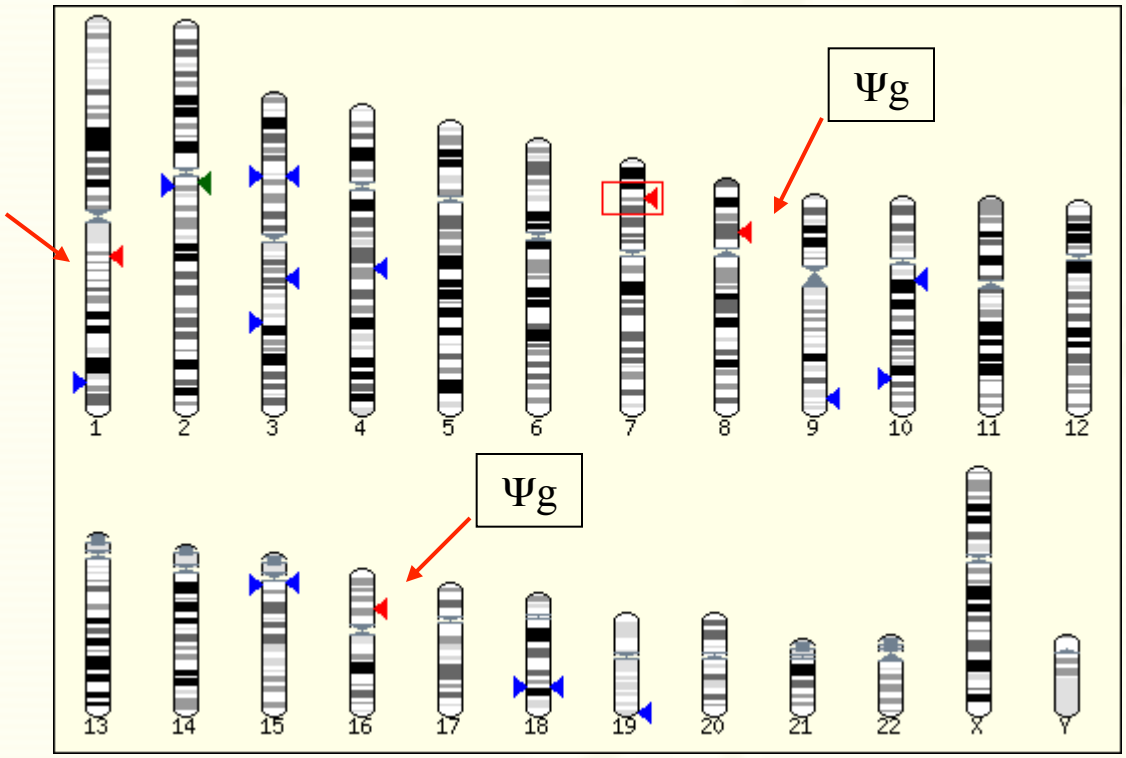
ding genes:	16,152
coding	2,360
ing genes:	1,317
	471
	6,865
	18,743
Gene transcripts:	9,826

Other

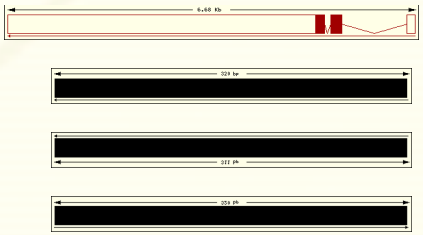
Genscan gene predictions:	126,538
Short Variants (SNPs, indels, somatic mutations):	1,520,076

Pseudogene: genomic fossil
an example cytochrome c pseudogenes



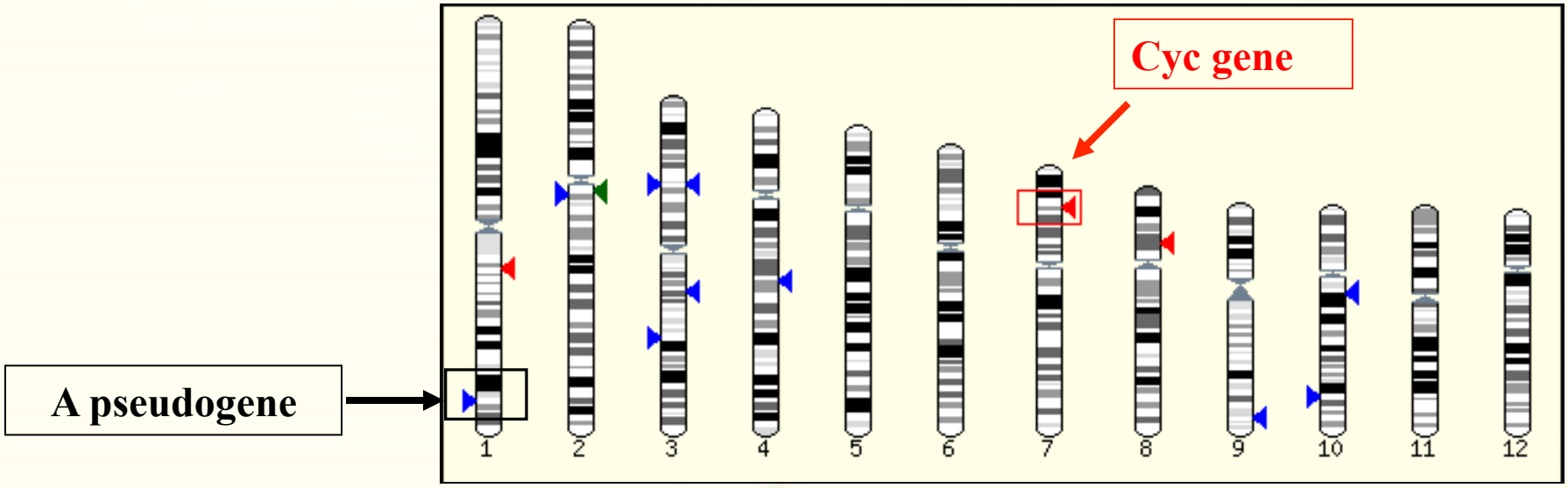


Transcript Structures




```
Cyc_gene    AAGTGTTCCCAGTGCCACACCGTTGAAAAGGGAGGCAAGCACAAGACTGGGCCAAATCTCCATGGTCTCT
Pseudogene  AAGTGTGCCCAATGCCACACCATGGTAAAGCGAGGCAAGTACAAGAGTGAGCCAAATCTCCATGGTCTAT

Cyc_gene    TTGGGCGGAAGACAGGTCAGGCCCCTGGATACTCTTACACAGCCGCCAATAAGAACAAAGGCATCATCTG
Pseudogene  TTATGCAGAAGACAGGTCAGGCCACTGAATTCTCT--CACAGACGCCAATGAGAACAAAGGCATCACCTG

Cyc_gene    GGGAGAGGATACACTGATGGAGTATTTGGAGAATCCCAAGAAGTACATCCCTGGAACAAAAATGATCTTT
Pseudogene  AGGAGAGGAGACACTGATGGAGTATTTGCAGAATCCCAAGAAGTACATCCCTGGAACAAAAATGACCATT

Cyc_gene    GTCGGCATTAAGAAGAAGGAAGAAAGGGCAGACTTAATAGCTTATCTCAAAAAGCTACTAATGAGTAA
Pseudogene  GTCAGCACTAAGAAGAAGGCAGAAAGGGCAGACTTGATAGCTTATCTCAGAAAAGCTAATAATCAG
```

Pseudogene: genomic fossil
an example cytochrome c pseudogenes



Cyc gene	KCSQCHTVEKGGKHKTGPNLHGLFGRK TGQAP-G- YSYTAANKNKGIIWG
Pseudogene	KCAQCHTMVKRGKYKSEPNLHGLFMQKTGQAT/G/YSLTDANENKGITXG

Cyc gene	EDTLMEYLENPKKYIPGTKMIFVGIKK KEERADLIAYLKKA TNE
Pseudogene	EETLMEYLQNPKKYIPGTKMTIVSTKKKAERADLIAYLRKANNQ

Human has much fewer genes than expected

- Before 2000s, it was estimated that human has > 35,000 genes.

Nat Genet. **2000** Jun;25(2):232-4.

Analysis of expressed sequence tags indicates 35,000 human genes.

Ewing B, **Green** P.

Department of Molecular Biotechnology, University of Washington, Seattle, Washington, USA.

- The initial annotation (2001) indicates 30,000 genes, which later was reduced to ~21,000 genes after removing pseudogenes and others.

Estimated gene numbers of selective species

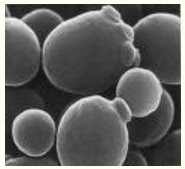
- Dog (Canis familiaris): ~14,000
- Platypus (鸭嘴兽): ~17,000
- Chicken: (Gallus gallus): ~17,000
- Frog (Xenopus tropicalis): ~17,000
- Zebrafish (Danio rerio): ~17,700
- Sea squirt 海鞘(Ciona intestinalis): ~14,000
- 线虫 (Caenorhabditis elegans)：20,000 genes
- Fruitfly (Drosophila melanogaster): 14,000
- 草履虫 (Paramecium tetraurelia): 40,000
- Yeast (S. cerevisiae): 5800











What makes human so complex if we have similar number of genes as a fruitfly ?

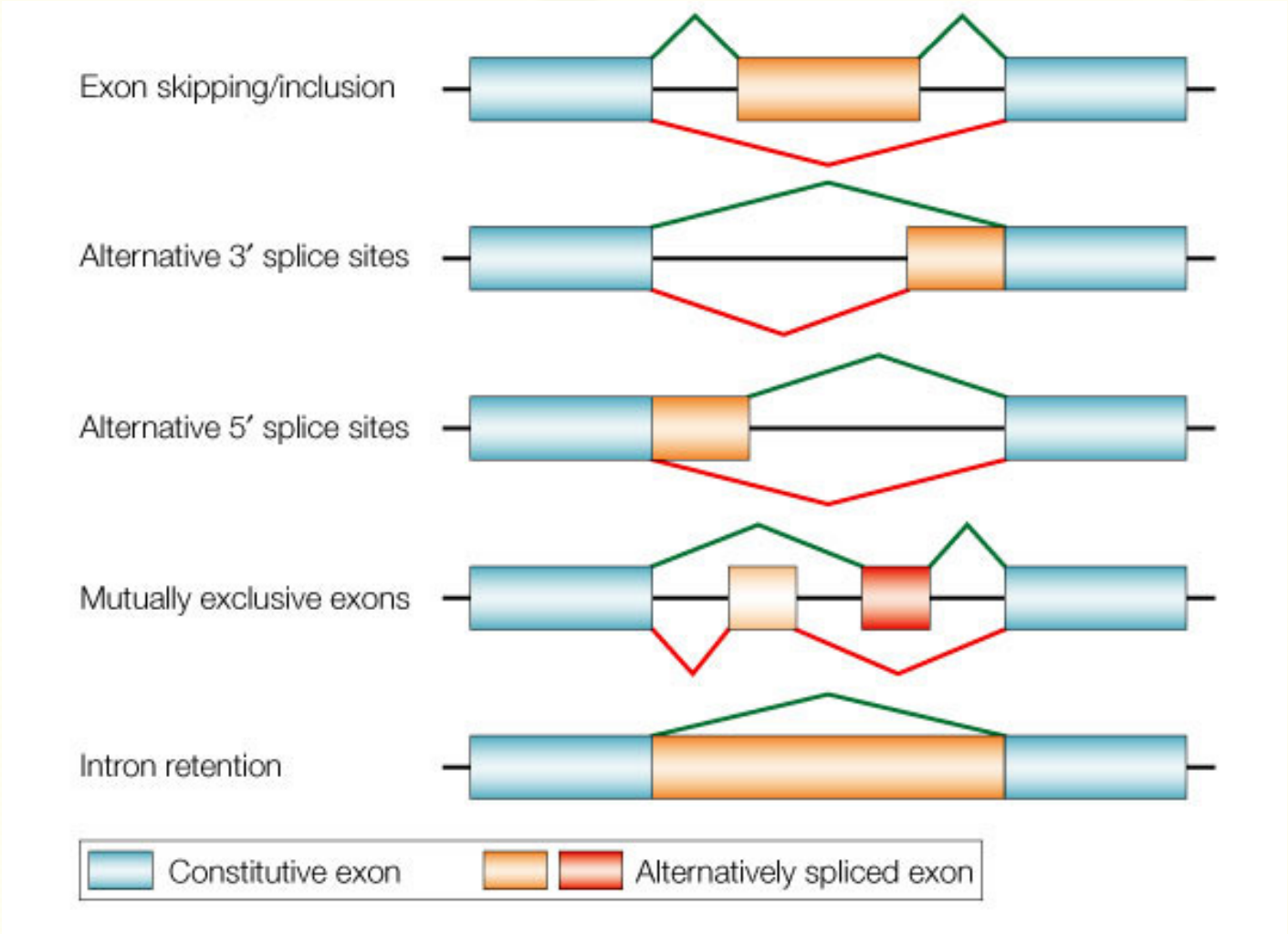
Possible reasons:

- Human proteins are longer and have more domains, thus can interact with more proteins.

- Human genes undergo alternative splicing, thus one gene can generate multiple proteins

- The regulation of human genes is more complex: by transcription factors, microRNAs, phosphorylations etc. The same or slightly different form of the same protein can be made at different abundance, at different time, and in different tissues.

The majority of the human genes undergo alternative splicing (AS)

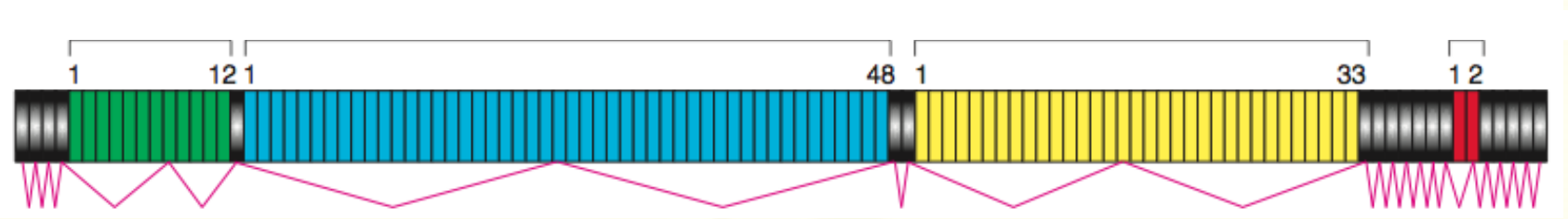


Alternative splicing: an extreme case

Exon cluster 4 (12 exons)

Exon cluster 5 (48 exons)

Exon cluster 6 (33 exons)



Drosophila Down syndrome cell adhesion molecule (*Dscam*)

Possible distinct mRNA transcripts: 12 x 48 x 33 = 19,008

Anastassiou et al Genome Biology 2006

Some basic facts about mammalian genome (human)

- Very small portion (~5%) of the genome encode for proteins, the vast majority of the regions are repetitive elements, intergenic sequences, pseudogenes, introns and potential regulatory elements.

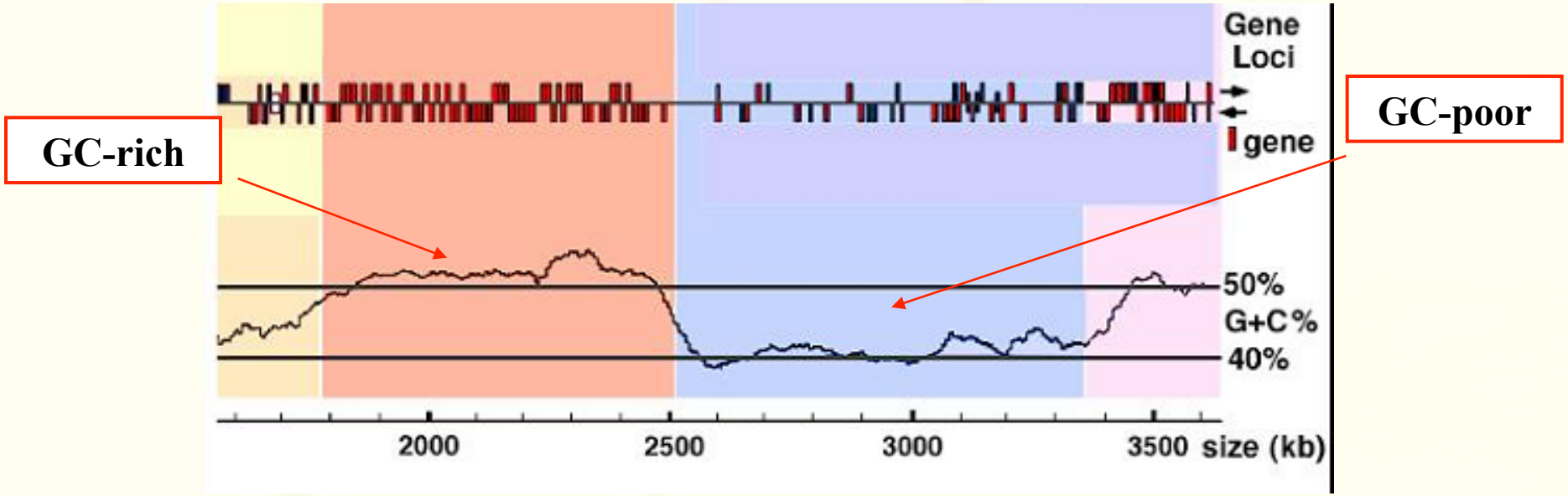
- **Junk DNA** may not be “junk”, a large fraction of the intergenic DNA is actually transcribed into RNA, but their potential function is not clear.
- Some of the **“junk DNA”** contains regulatory elements, finding them is difficult.

- Mammalian genomes are mosaic of isochores: ~300 Kb long DNA with homogeneous G+C%, caused by bias in mutation or recombination.

- Genes are more concentrated in G+C rich regions.

Isochores in the genome

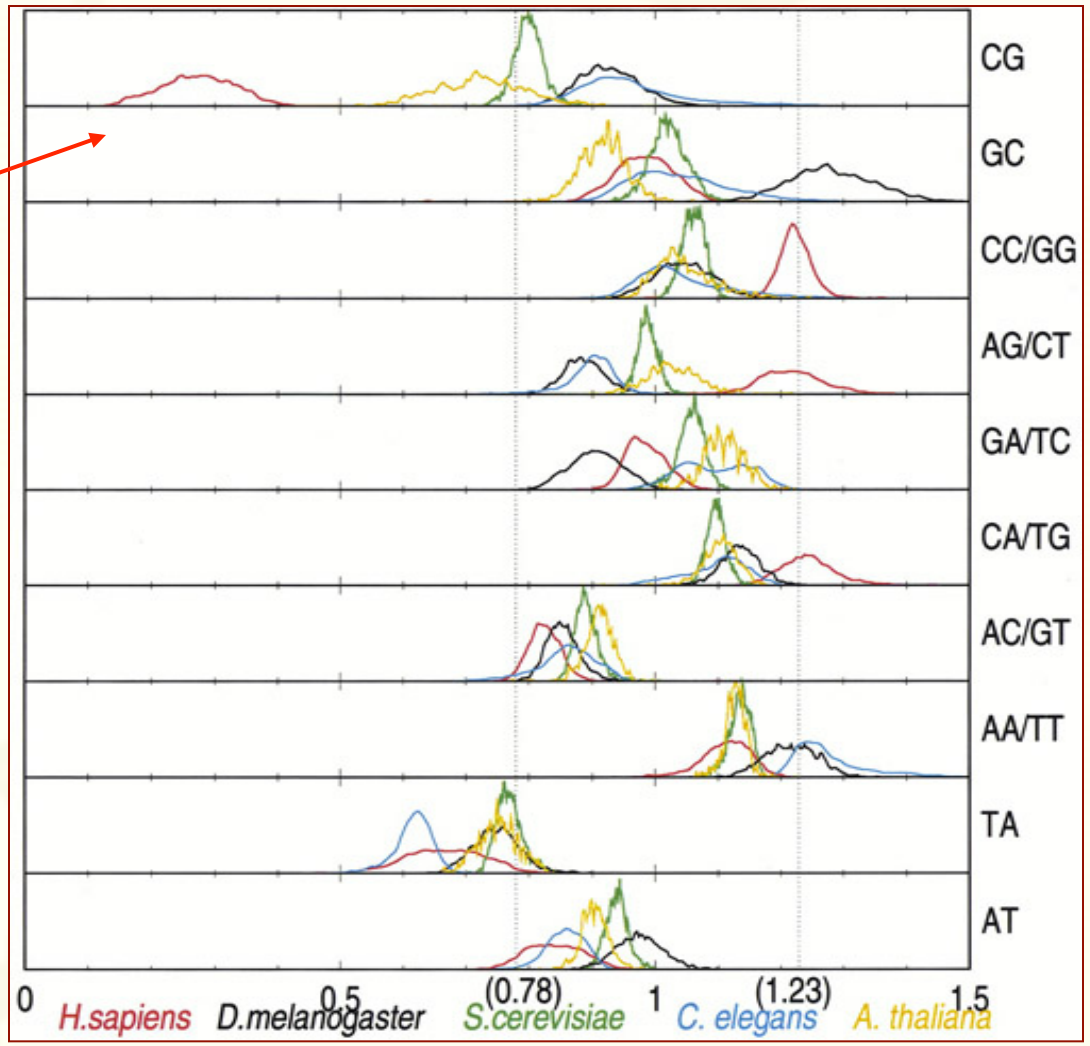
Giemsa staining bands correspond to isochores



Di-nucleotide frequency bias

Human genome contains 21% C and 21% G, therefore the frequency of CpG should be 0.21x0.21 = **4%.**

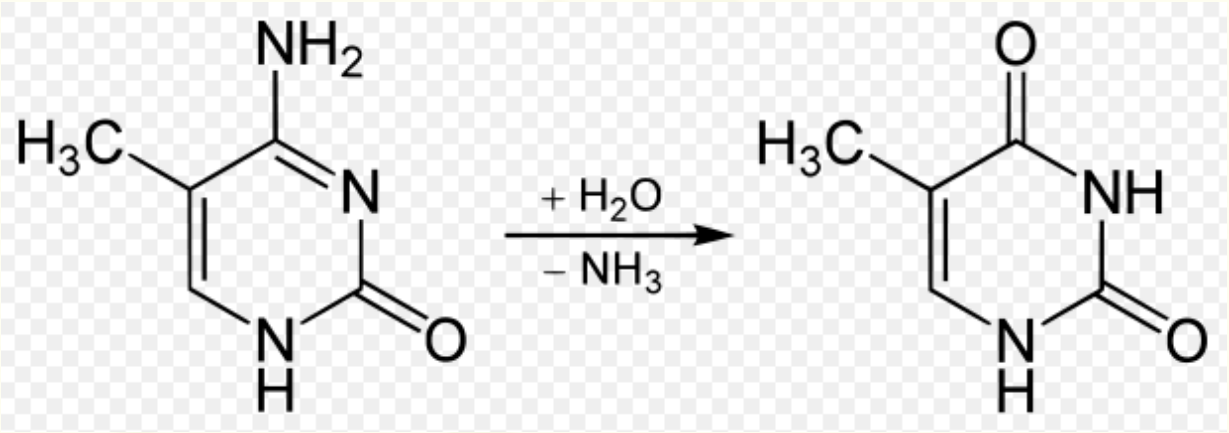
However the real CpG frequency is only **1%** , less than 25% of the expected.



Genome-scale compositional comparisons in eukaryotes. Gentles AJ, Karlin S, Genome Res. 2001 Apr;11(4):540-6.

Why vertebrate genomes contain few CpG ?

- C (cytosine) base followed immediately by a G (guanine) base (a CpG) is rare in vertebrate DNA.
- This is because the cytosines followed by G tend to be methylated. The methylated cytosine can undergo deamination and becomes U (Uracil).



CpG islands are often present in the promoters of genes

- CpG dinucleotides are depleted in the genome, however large segments of CpG are found in 40% of the human genes.
- The methylation state of these CpG islands can regulate the expression of the genes.
- DNA methylation can be measured by microarrays or deep sequencing.

Repetitive Elements in the Human Genome

- LINE retrotransposon
  - Retro-: going through an RNA intermediate
  - LINE: Long Interspersed Elements
  - The complete sequence is 6000 – 8000 bp long, contains genes for an RNA binding protein and an endonuclease and reverse transcriptase.

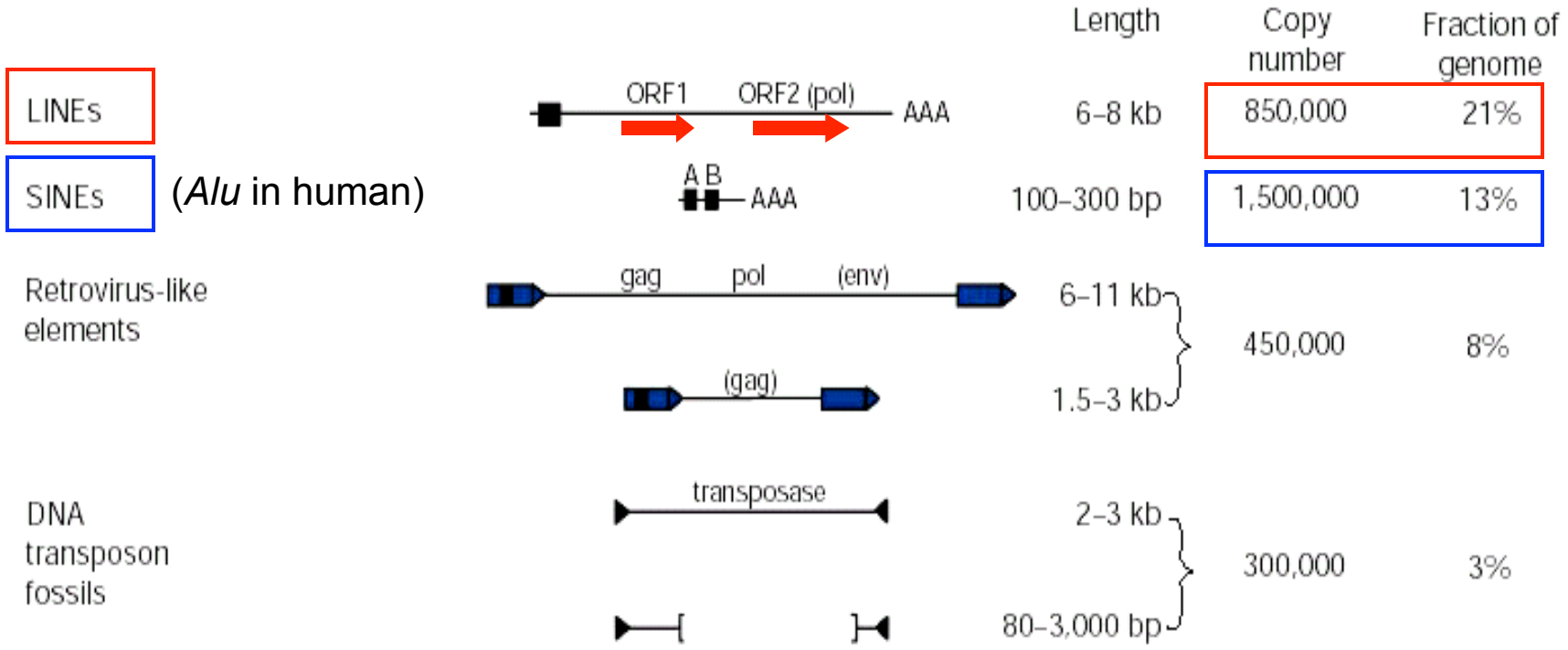
- SINE retrotransposon
  - SINE: Short Interspersed Elements
  - It is a “parasite’s parasite”, depends on LINE for its propagation
  - ***Alu* elements** is the most abundant in human, 300 bp long

- Retrovirus
- DNA transposon

[International Human Genome Sequencing Consortium, ***Nature*** 409, 2001]

Repetitive Elements in the Human Genome

Classes of interspersed repeat in the human genome



In total, 46% of the human genome are **recognizable** interspersed repeats

[International Human Genome Sequencing Consortium, ***Nature*** 409, 2001]

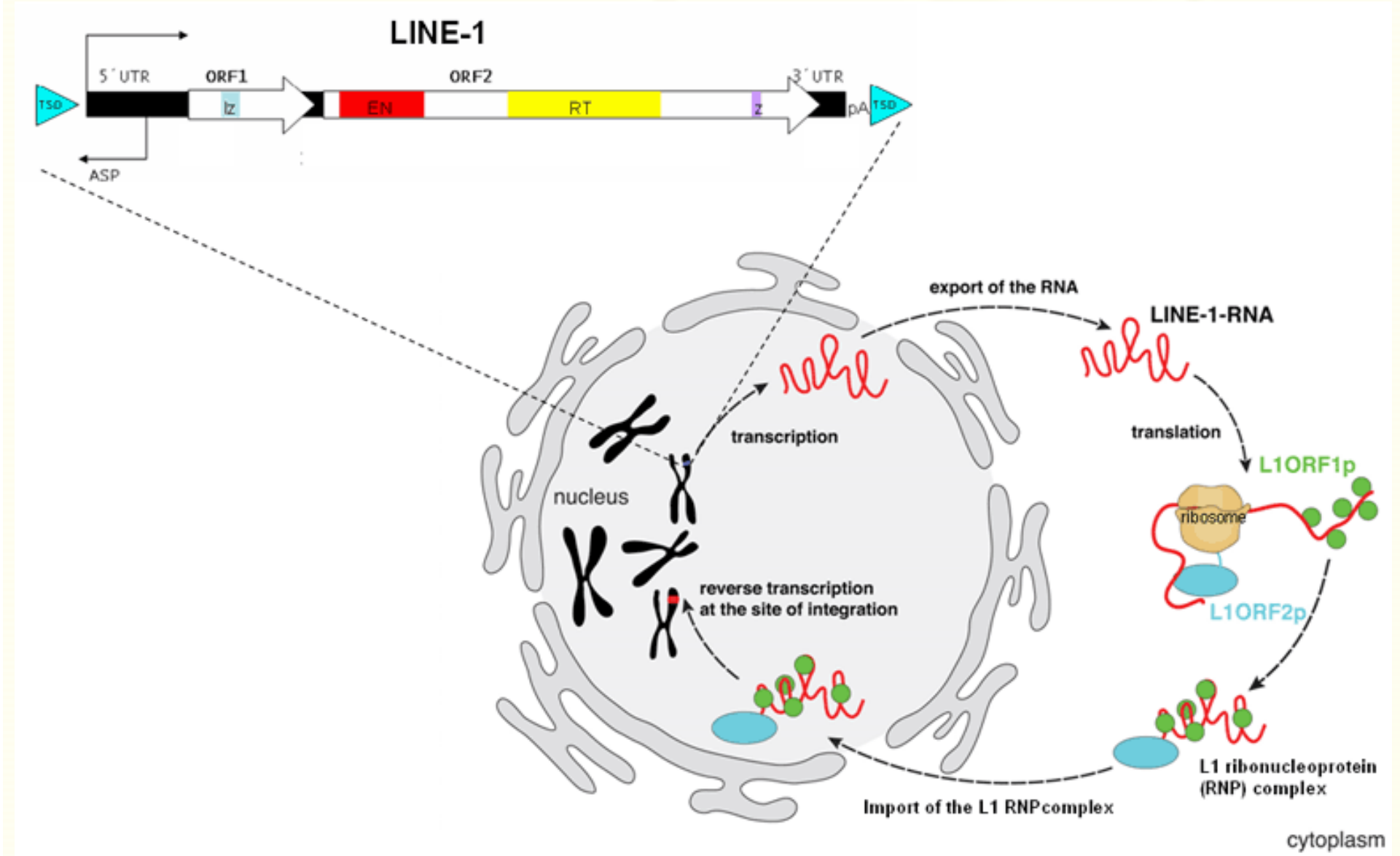
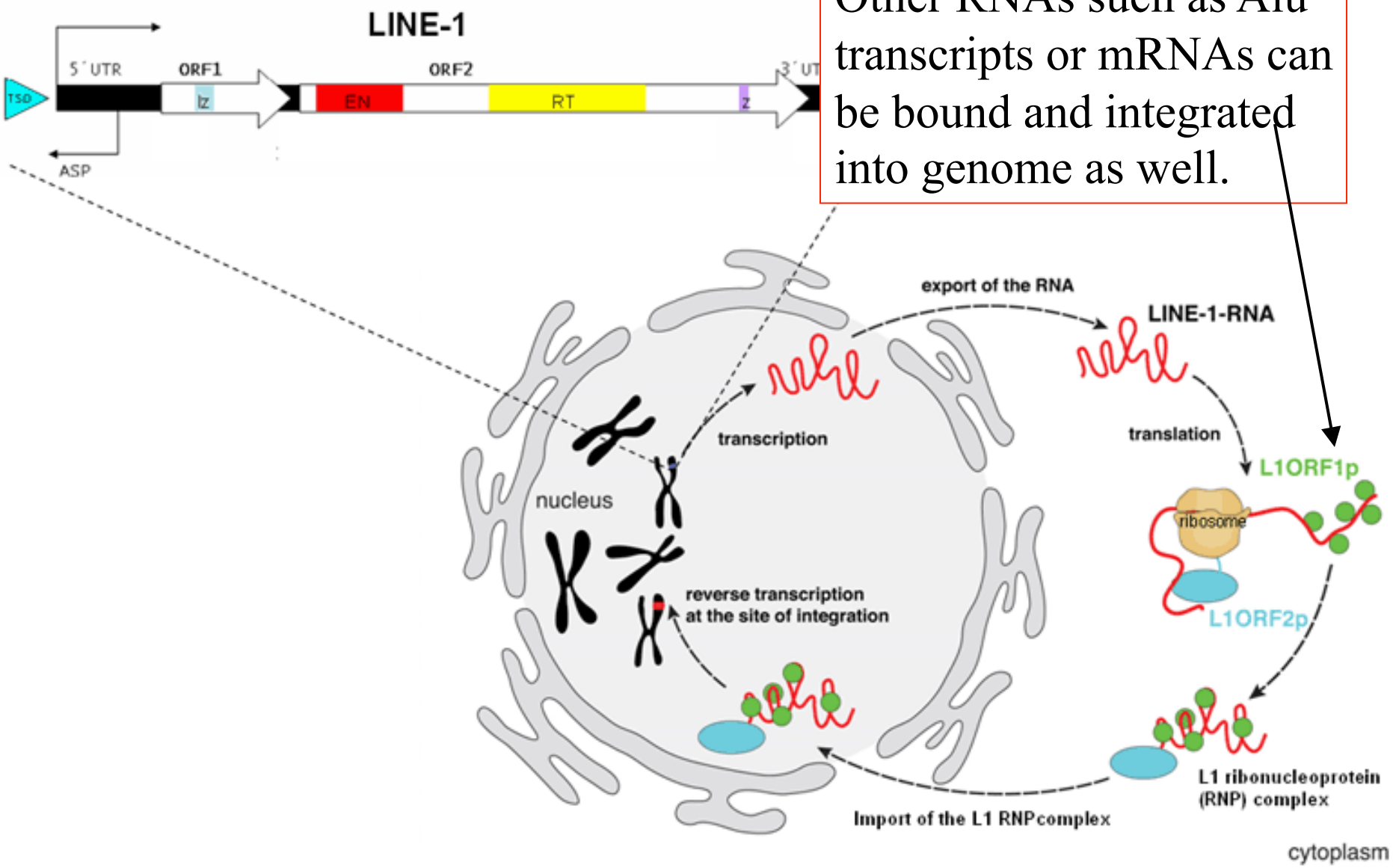
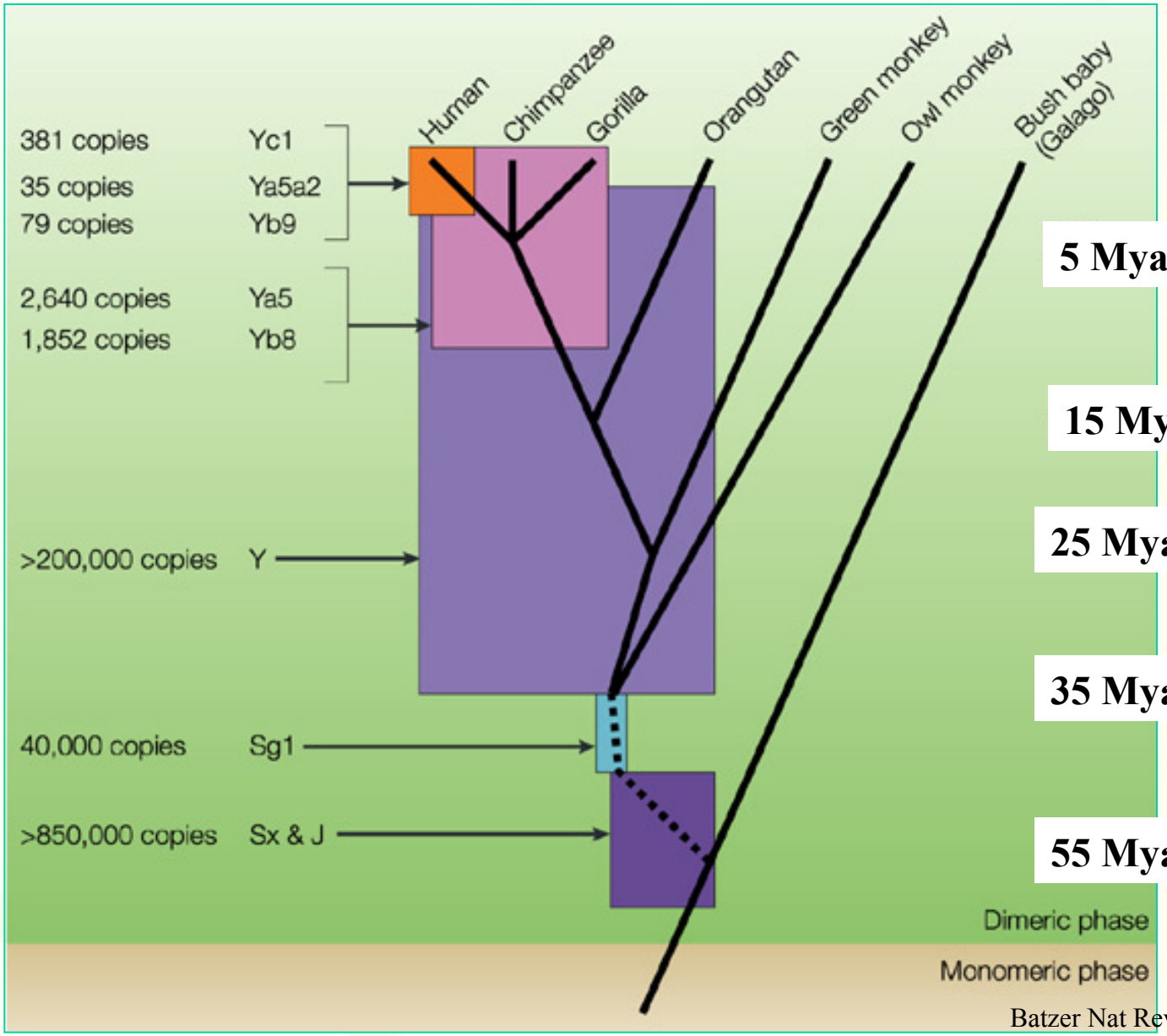


Image: Elena Khazina and Oliver Weichenrieder; Max Planck Institute for Developmental Biology

Other RNAs such as Alu transcripts or mRNAs can be bound and integrated into genome as well.

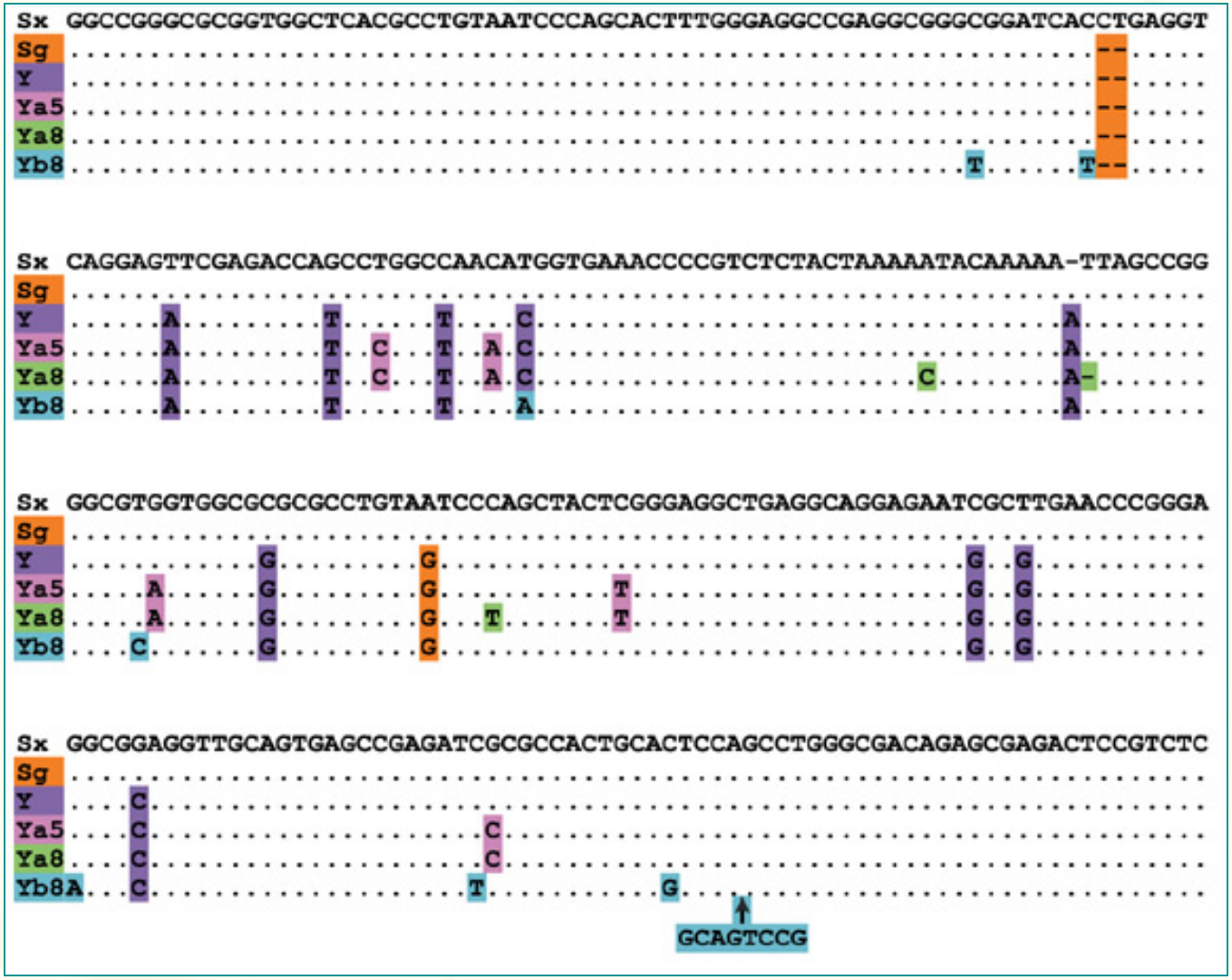


Family of Alu elements



Batzer Nat Rev Gen 2002

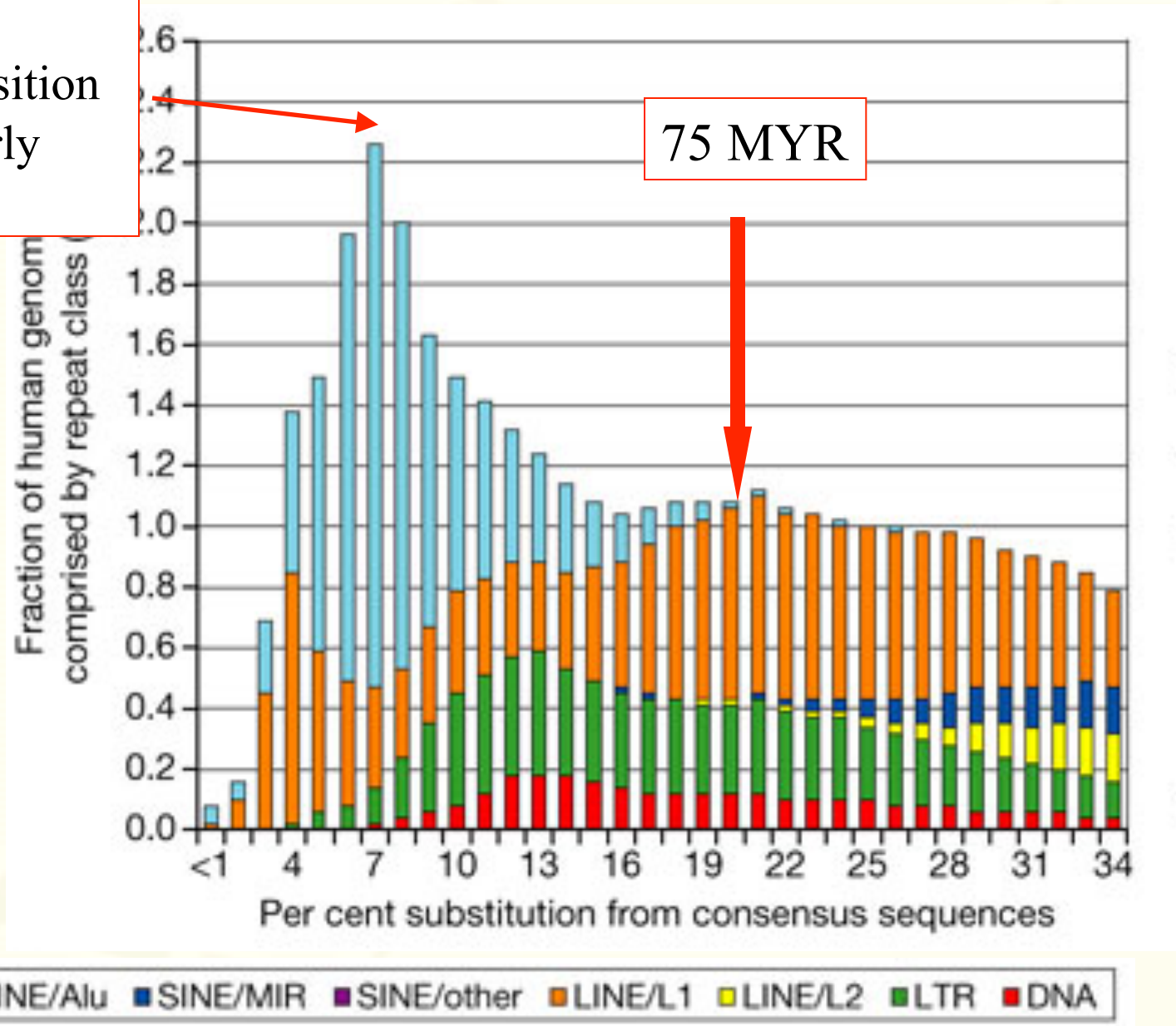
Family of Alu elements



Batzer Nat Rev Gen 2002

Dating Repetitive Elements by Sequence Divergence

A burst of retrotransposition events in early primates



[International Human Genome Sequencing Consortium, ***Nature*** 409, 2001]

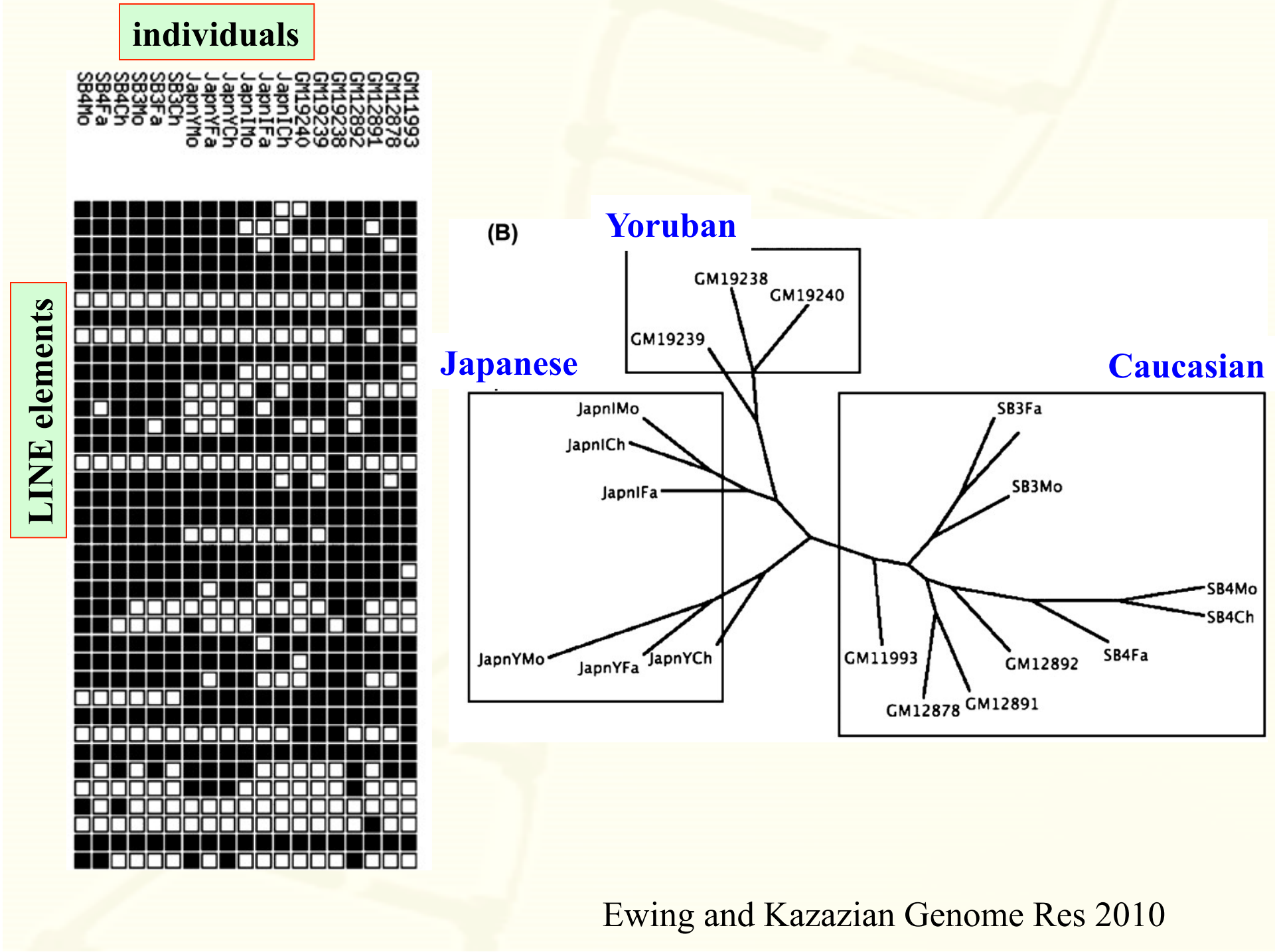
Some LINE elements are still active in human and are polymorphic among individuals and populations

Genome Res. 2010 Sep;20(9):1262-70. Epub 2010 May 20.

High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes.

Ewing AD, Kazazian HH Jr.

- “In total, we assayed 25 individuals at 1139 sites”
- “We find that any two individual genomes differ at an average of 285 sites (25%) with respect to L1 insertion presence or absence. “
- “We estimate that the rate of L1 retrotransposition in humans is between 1/95 and 1/270 births”



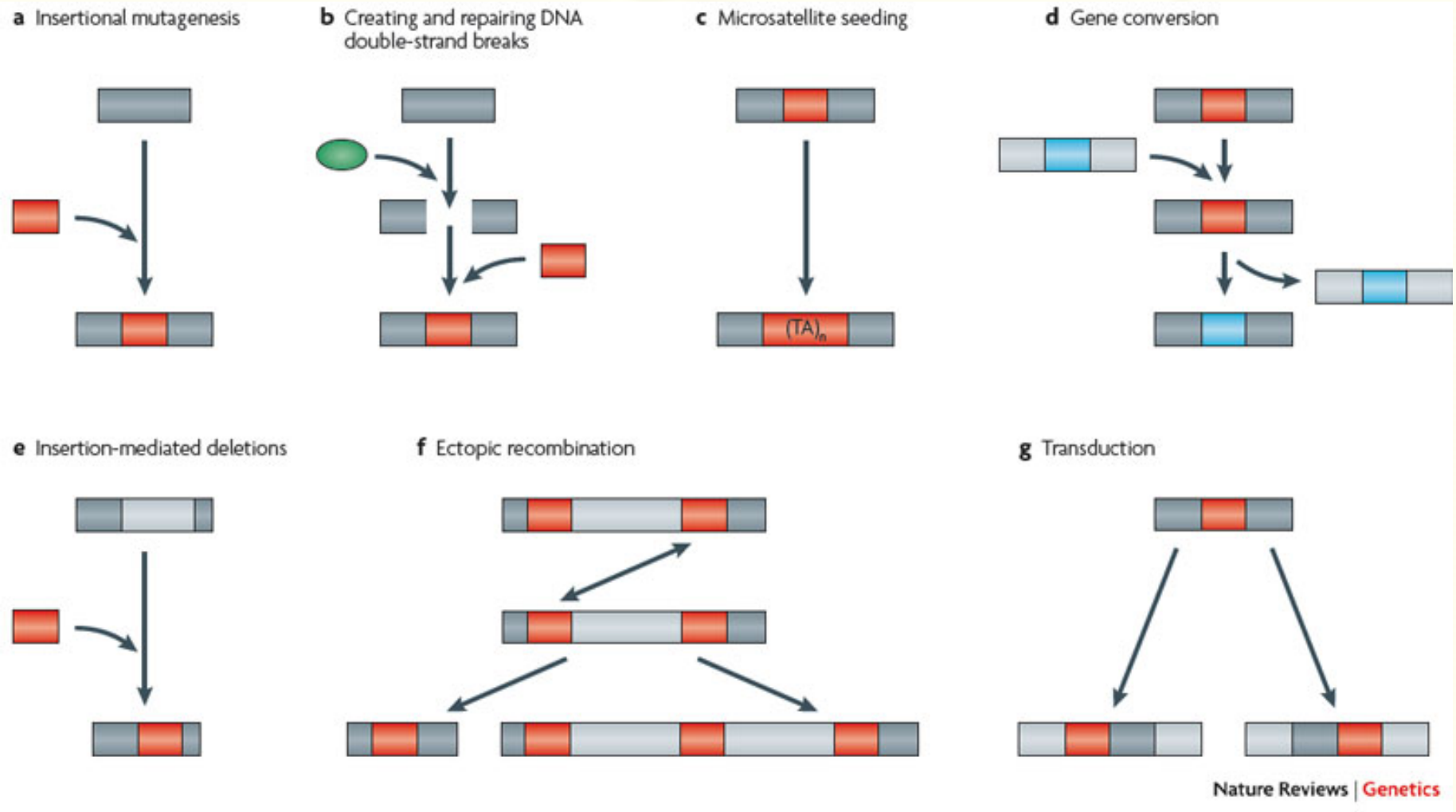
Ewing and Kazazian Genome Res 2010

Why so many repetitive elements ?

- **Why are repeats bad ?**
  - Repetitive elements wastes energy: replication, transcription etc… birds have fewer repeats because they have high metabolic rates
  - Insertion of repetitive elements can be harmful.

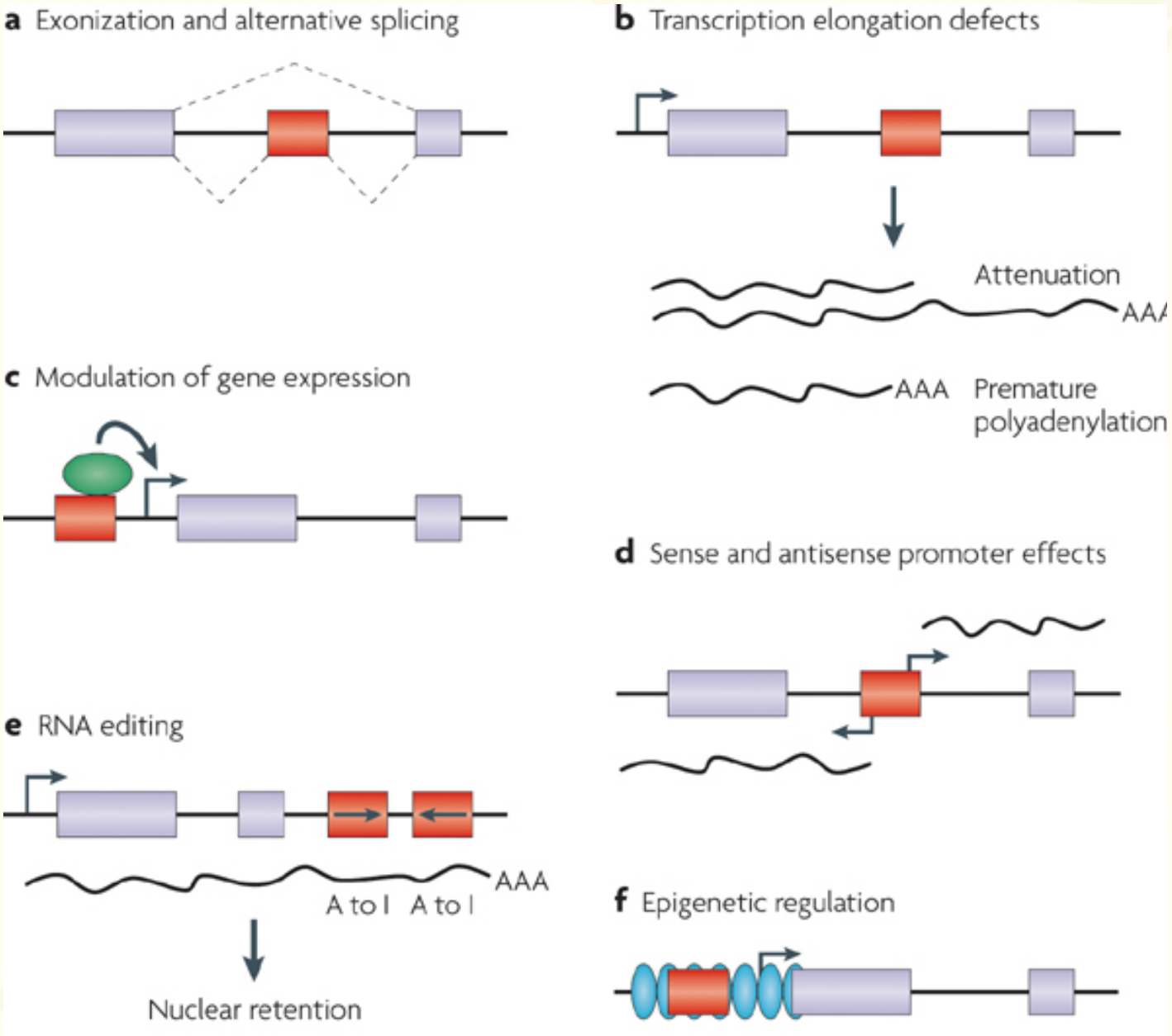
- **Why are they still here ?**
  - Mammalian genomes can not rid them because of small population size which allows accumulation of junk DNA.
  - The generation and deletion of repetitive elements have reached an equilibrium
  - Mammalian genomes can tolerate them because we developed mechanism to control them such as histone modification, and siRNA

Effect of repeats on genes



Richard Cordaux & Mark A. Batzer, Nat Rev Gen 2009

Effects of repeats on gene expression



Richard Cordaux & Mark A. Batzer, Nat Rev Gen 2009

Using repetitive elements to infer phylogeny

Proc Natl Acad Sci U S A. 1999 Aug 31;96(18):10261-6.

Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: hippopotamuses are the closest extant relatives of whales.

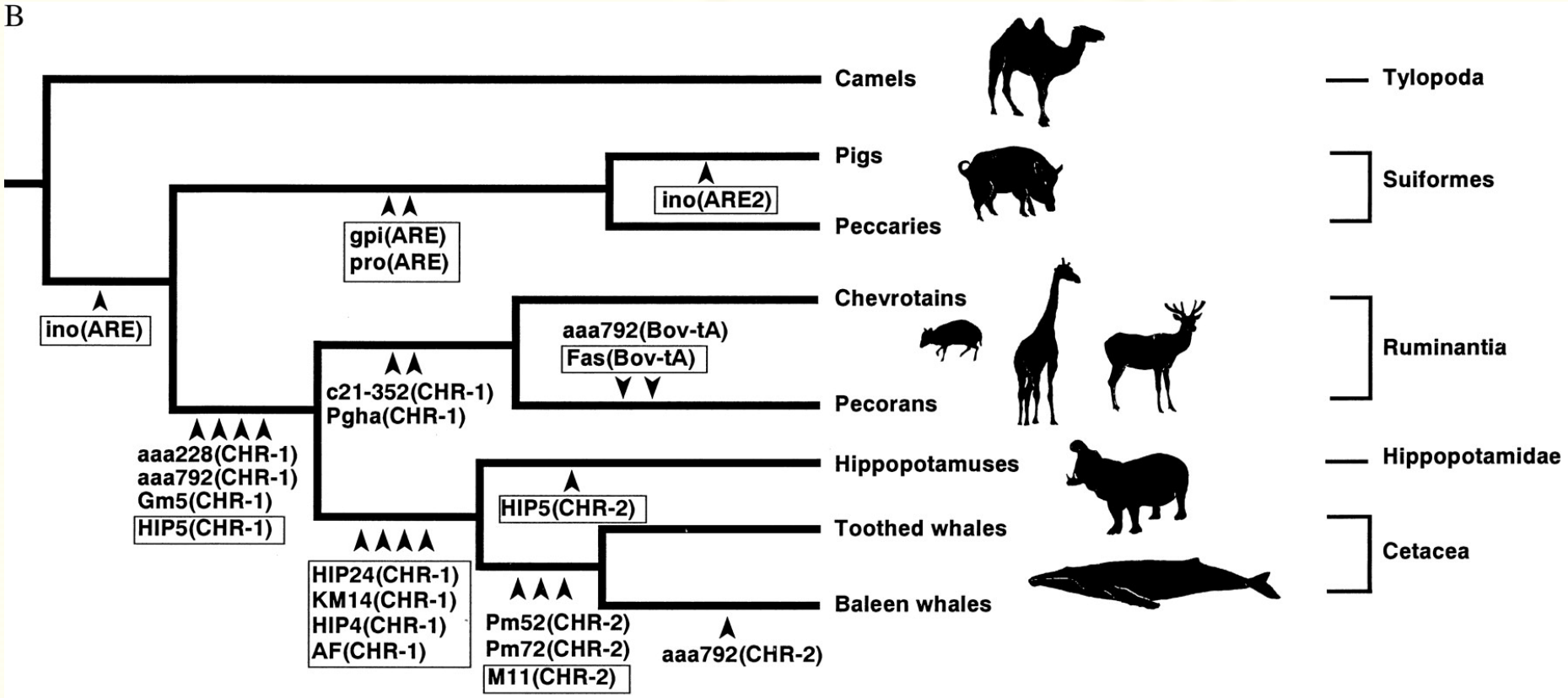
Nikaido M, Rooney AP, Okada N.

A

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Camel	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	?	0	0	0
Pig	0	0	0	?	0	0	0	0	?	0	0	0	?	?	0	0	?	1	1	1
Peccary	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	1	1
Chevrotain	?	0	?	?	?	?	?	?	?	1	0	?	?	?	1	1	0	?	0	0
Deer	0	0	0	0	0	0	0	1	?	1	1	1	1	1	1	?	1	1	0	0
Graffe	?	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0
Sheep	0	0	0	0	0	?	0	1	1	1	1	1	1	1	1	1	1	1	0	0
Cow	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0
Hippo	0	?	0	1	1	1	1	0	1	1	0	1	1	0	0	0	?	1	0	0
Humpback	1	1	1	1	1	1	1	0	1	1	0	1	1	0	0	0	?	?	0	0
Beaked	1	1	1	1	1	1	1	0	?	1	0	1	1	0	0	0	?	1	0	0

1	PM52	11	aaa792(Bov tA)
2	PM72	12	Gm5
3	M11	13	HIP5(CHR-1)
4	HIP24	14	HIP5(Bov A)
5	KM14	15	c21-352
6	HIP4	16	pgha
7	AF(CHR-1)	17	Fas
8	AF(MER)	18	INO
9	aaa228	19	pgi
10	aaa792(CHR-1)	20	pro

Using repetitive elements to infer phylogeny



Human genome has thousands of pseudogenes

Millions of Years of Evolution Preserved: A Comprehensive Catalog of the Processed Pseudogenes in the Human Genome

Zhaolei Zhang, Paul M. Harrison, Yin Liu, and Mark Gerstein[1]

Identification and Analysis of Over 2000 Ribosomal Protein Pseudogenes in the Human Genome

Zhaolei Zhang, Paul Harrison, and Mark Gerstein[1]

Different Types of Pseudogenes

- Duplicated Pseudogenes:
  - Created from tandem duplication or unequal-crossover
  - Segment duplication is prevalent (5% of the genome)
- Retropseudogenes (Processed Pseudogenes)
  - by mRNA retrotransposition (反专录）
- Other types:
  - Spontaneous loss of function: e.g. Olfactory Receptors
  - Numt (Nuclear mitochondria DNA)

63% of the human olfactory receptor genes are pseudogenes

- Human has 900 olfactory receptor genes and pseudogenes, 63% has a disrupted open reading frame.
-
- Other primates have similar fraction of pseudogenes, probably the result of decreased dependence on olfaction due to arising of color vision

Genome Res. 2001 May;11(5):685-702.

The complete human olfactory subgenome.

Glusman G, Yanai I, Rubin I, Lancet D.

Loss of Olfactory Receptor Genes Coincides with the Acquisition of Full Trichromatic Vision in Primates

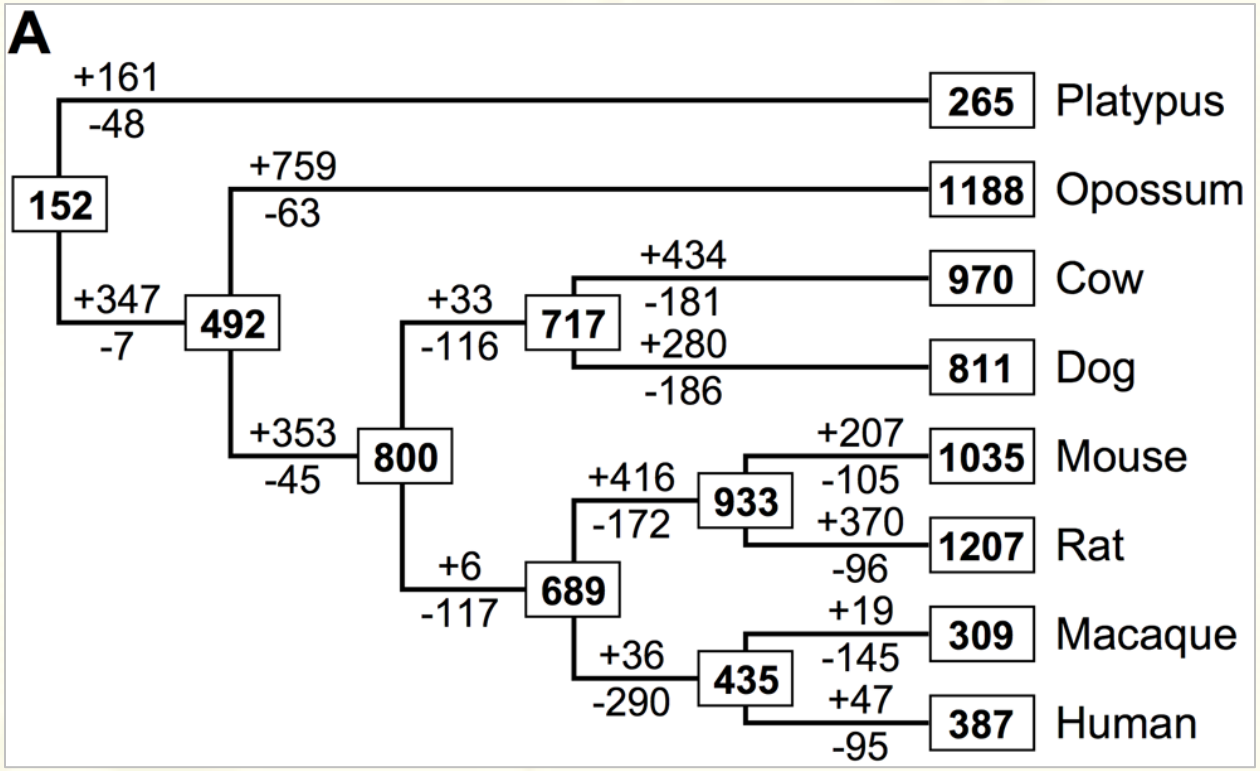
Yoav Gilad[1,2*], Victor Wiebe[1], Molly Przeworski[1], Doron Lancet[2], Svante Pääbo[1]

Evolution of olfactory receptor genes

PLoS One. 2007 Aug 8;2(8):e708.

Extensive gains and losses of olfactory receptor genes in mammalian evolution.

Niimura Y, Nei M.



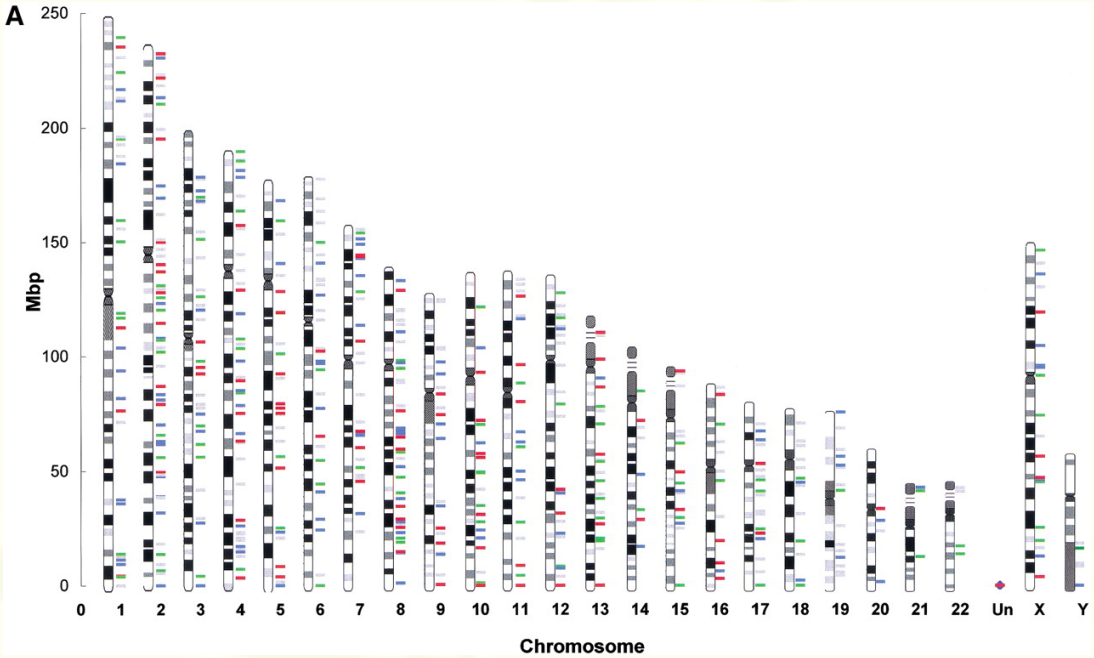
NUMT: Nuclear mitochondrial pseudogenes

- A total of > 600 insertions, 500, 000 bp, 0.016% of the nuclear genome

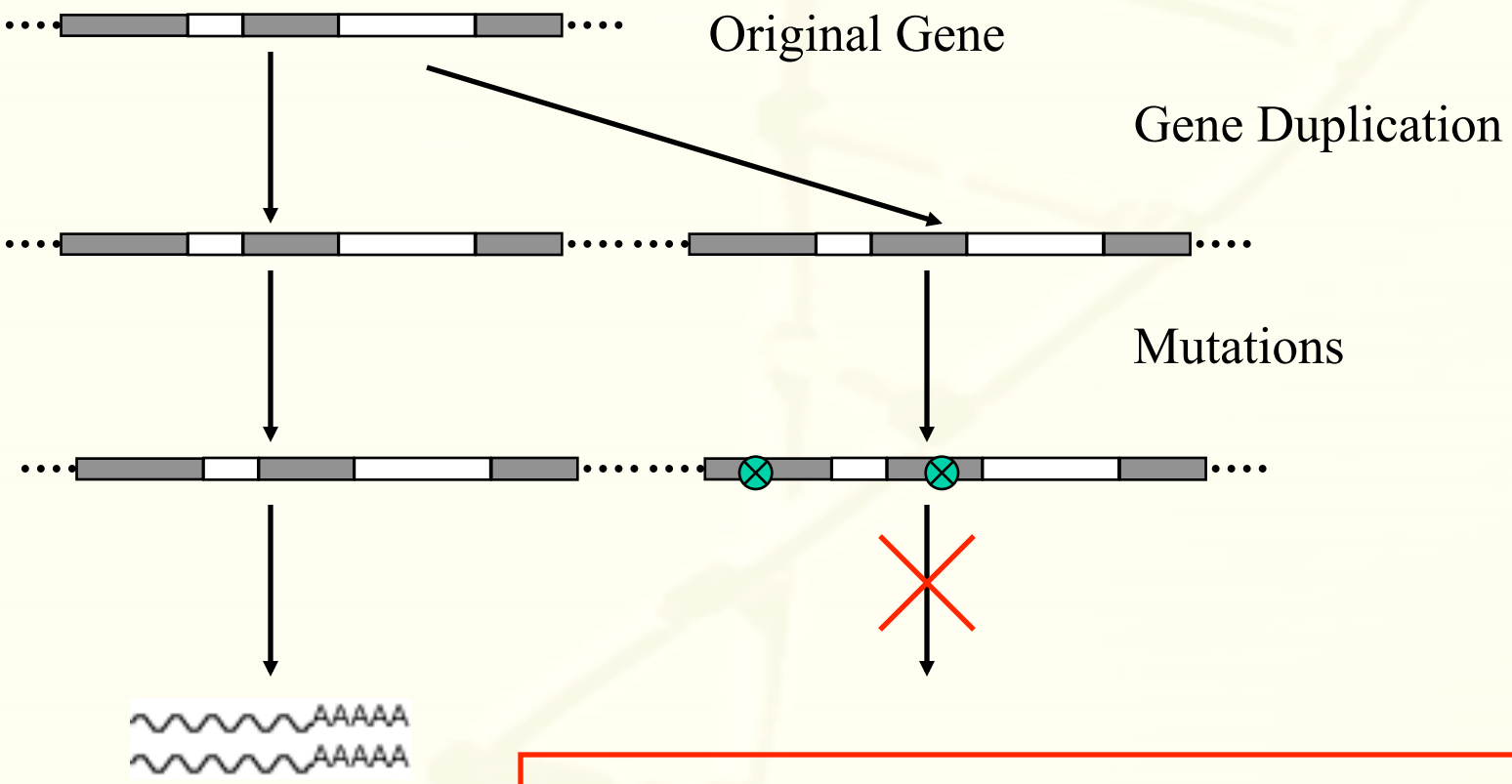
Genome Res. 2002 Jun;12(6):885-93.

Pattern of organization of human mitochondrial pseudogenes in the nuclear genome.

Woischnik M, Moraes CT.

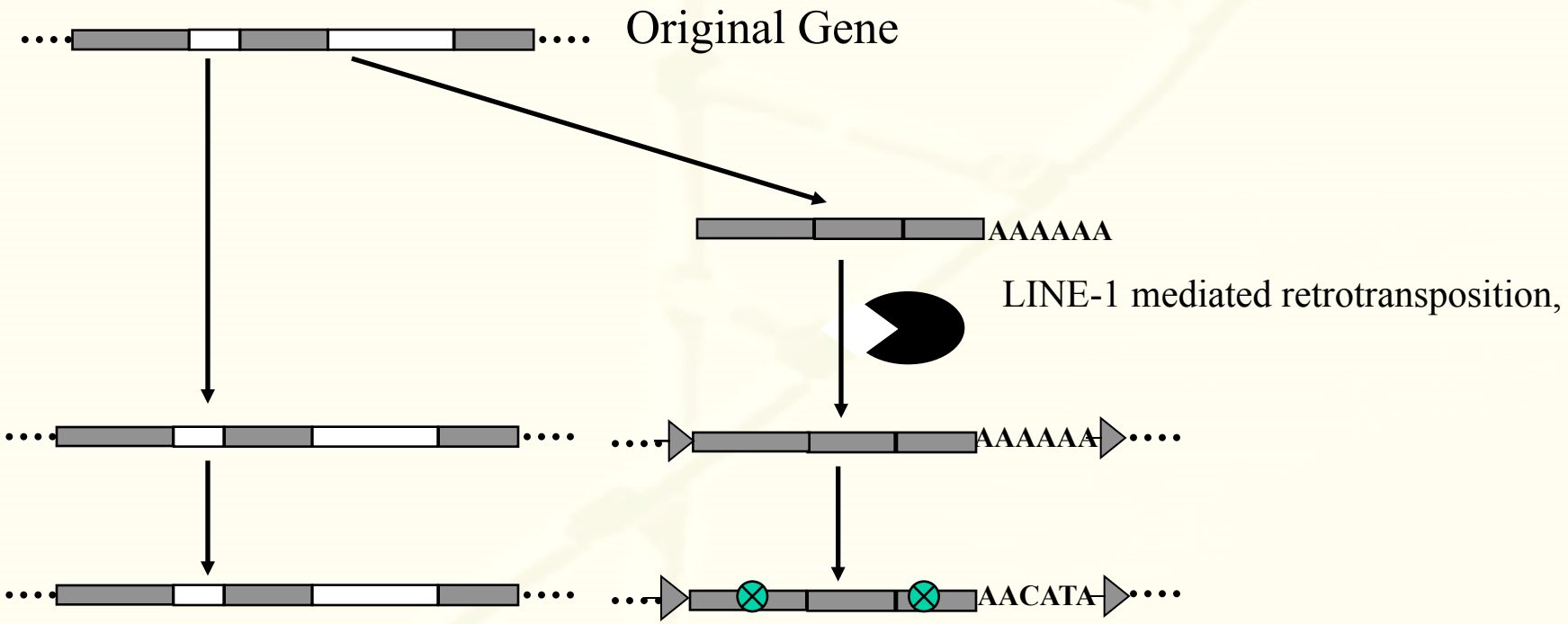


Duplicated Pseudogenes



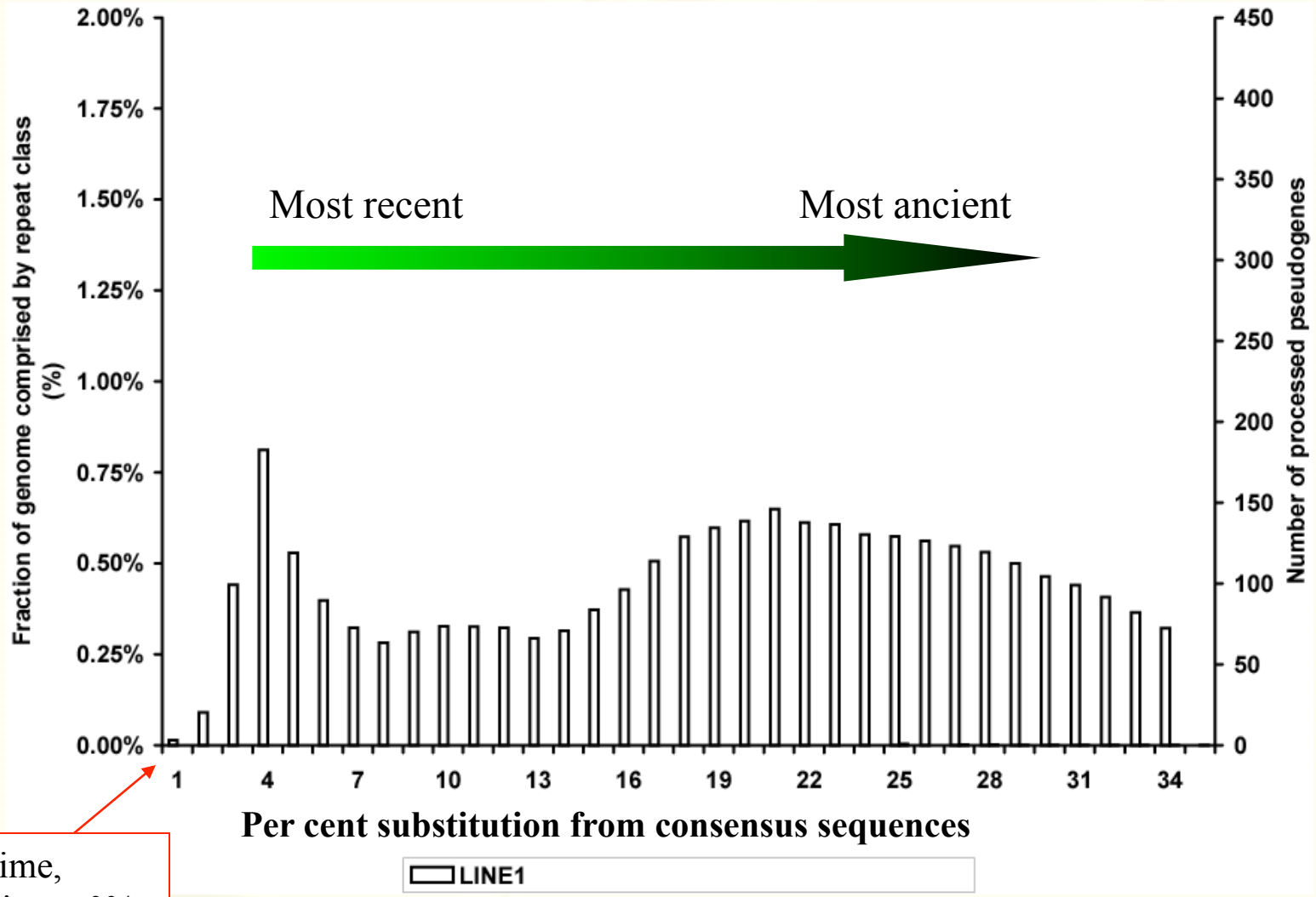
- ✓retains original intron/exon structure
- ✓e.g. globins, *Hox* cluster and *Arabidopsis* genome
- ✓sometimes can be transcribed

Retro-pseudogenes (Processed pseudogenes)



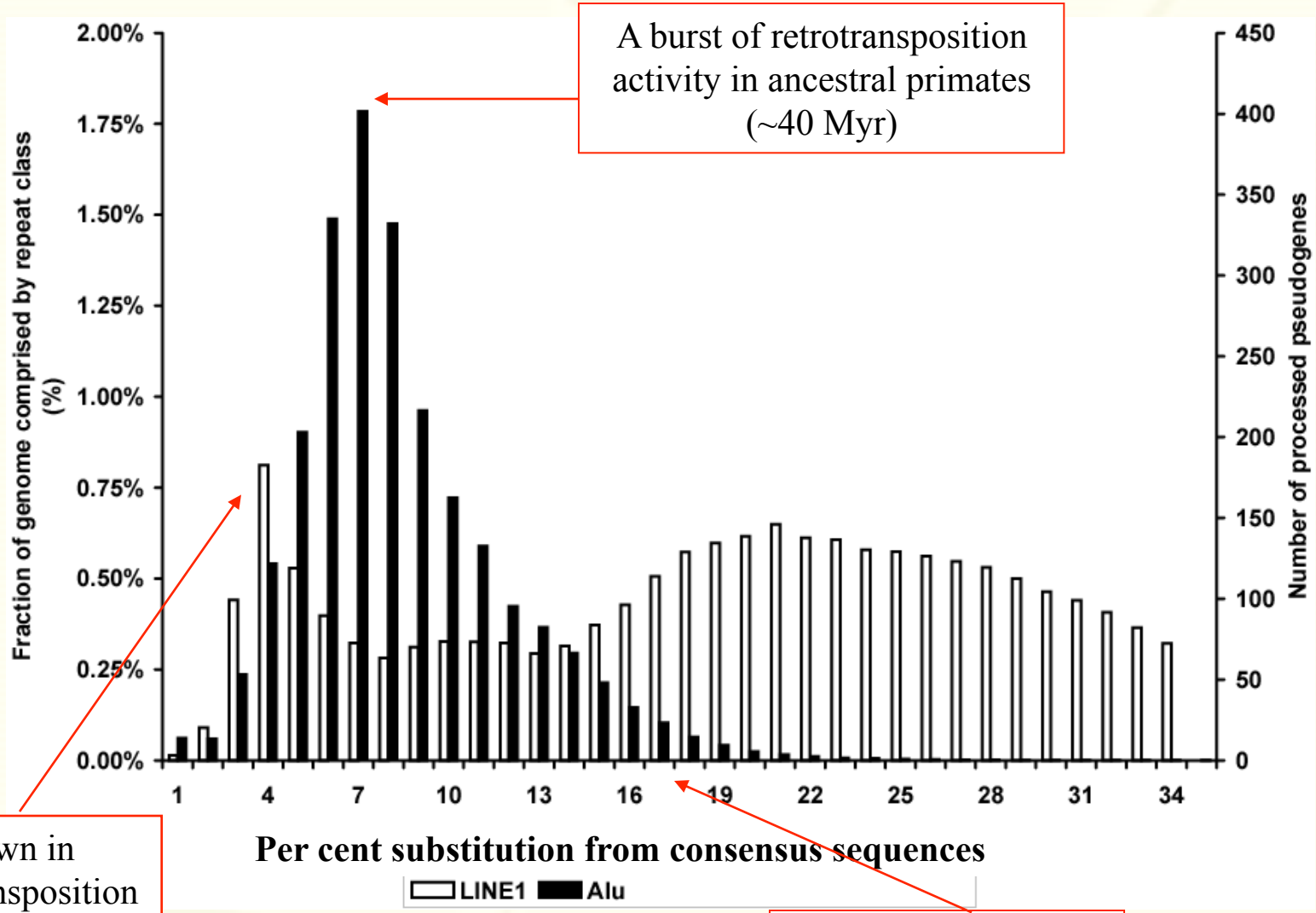
- ✓Mostly dead-on-arrival (DOA)
- ✓Features: intronless, poly-A tail, direct repeats
- ✓Target-primed reverse-transcription: -TT|AAA-

We can calculate the age of the pseudogenes by the number of substitutions

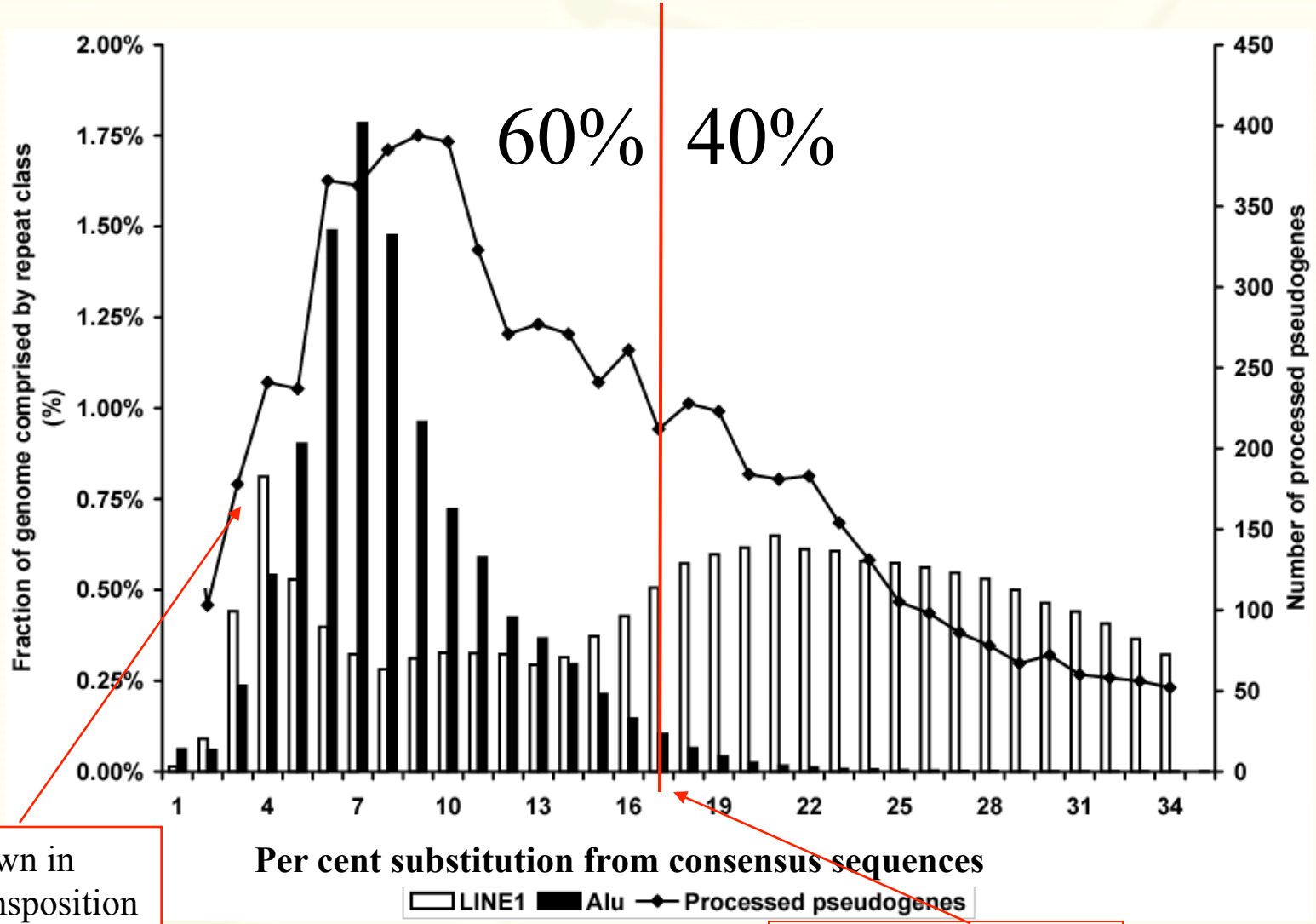


Present time,
Substitution = 0%

We can calculate the age of the pseudogenes by the number of substitutions



We can calculate the age of the pseudogenes by the number of substitutions

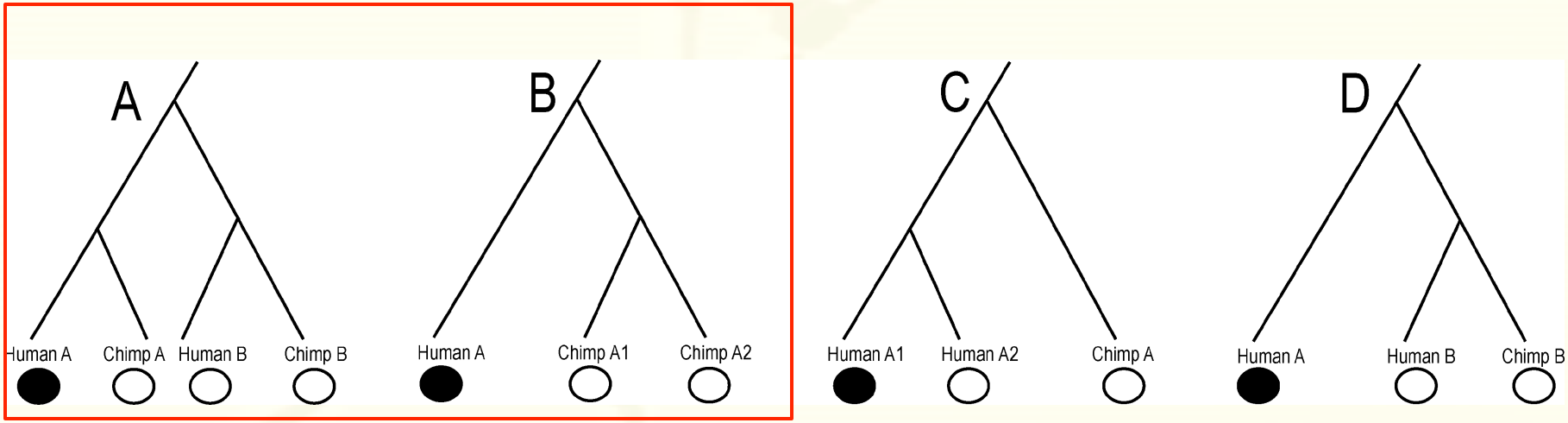


Slowdown in retrotransposition

Human and mouse diverge, ~75 Myr

Human specific pseudogenes

- Wang et al identified 80 pseudogenes that were inactivated after human split from the chimpanzees".
- They used phylogenetic tree to distinguish human specific loss from chimp specific gene gain.



Human specific

Genes losses during human origins, Wang, Grus, Zhangm PLOS 2005

Retrotransposition can generate new genes too

Emergence of Young Human Genes after a Burst of Retroposition in Primates

Ana Claudia Marques[1☯], Isabelle Dupanloup[1☯], Nicolas Vinckenbosch[1], Alexandre Reymond[1,2], Henrik Kaessmann[1*]

- *"We estimate that at least one new retrogene per million years emerged on the human lineage during the past ~63 million years of primate evolution."*

OPEN ACCESS Freely available online PLOS BIOLOGY

Birth and Rapid Subcellular Adaptation of a Hominoid-Specific CDC14 Protein

Lia Rosso[1☯], Ana Claudia Marques[1☯], Manuela Weier[1], Nelle Lambert[2], Marie-Alexandra Lambot[2], Pierre Vanderhaeghen[2], Henrik Kaessmann[1*]

1 Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland, **2** Institut de Recherches en Biologie Humaine et Moleculaire (IRIBHM), University of Brussels, Brussels, Belgium

Gene duplication was prevalent during hominoid evolution, yet little is known about the functional fate of new ape gene copies. We characterized the *CDC14B* cell cycle gene and the functional evolution of its hominoid-specific daughter gene, *CDC14Bretro*. We found that *CDC14B* encodes four different splice isoforms that show different subcellular localizations (nucleus or microtubule-associated) and functional properties. A microtubular *CDC14B* variant spawned *CDC14Bretro* through retroposition in the hominoid ancestor 18–25 million years ago (Mya). *CDC14Bretro* evolved brain-/testis-specific expression after the duplication event and experienced a short period of intense positive selection in the African ape ancestor 7–12 Mya. Using resurrected ancestral protein variants, we demonstrate that by virtue of amino acid substitutions in distinct protein regions during this time, the subcellular localization of CDC14Bretro progressively shifted from the association with microtubules (stabilizing them) to an association with the endoplasmic reticulum. CDC14Bretro evolution represents a paradigm example of rapid, selectively driven subcellular relocalization, thus revealing a novel mode for the emergence of new gene function.

OPEN ACCESS Freely available online PLOS BIOLOGY

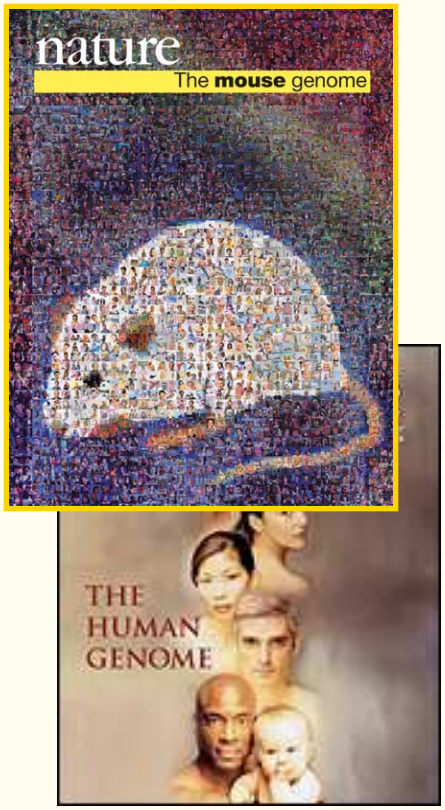
Loss of Egg Yolk Genes in Mammals and the Origin of Lactation and Placentation

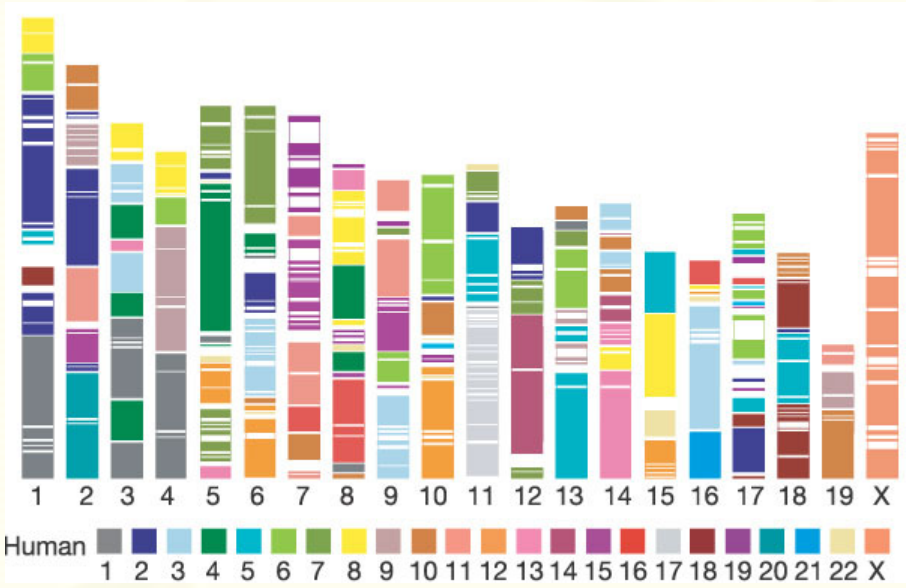
David Brawand[1], Walter Wahli[1,2¶*], Henrik Kaessmann[1¶*]

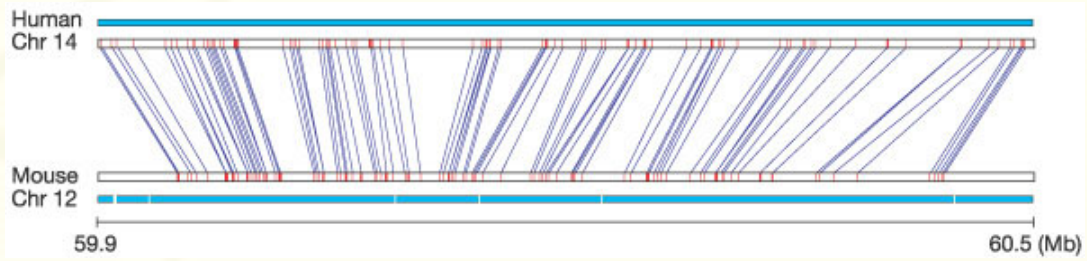
1 Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland, **2** National Research Center Frontiers in Genetics, University of Lausanne, Lausanne, Switzerland

Embryonic development in nonmammalian vertebrates depends entirely on nutritional reserves that are predominantly derived from vitellogenin proteins and stored in egg yolk. Mammals have evolved new resources, such as lactation and placentation, to nourish their developing and early offspring. However, the evolutionary timing and molecular events associated with this major phenotypic transition are not known. By means of sensitive comparative genomics analyses and evolutionary simulations, we here show that the three ancestral vitellogenin-encoding genes were progressively lost during mammalian evolution (until around 30–70 million years ago, Mya) in all but the egg-laying monotremes, which have retained a functional vitellogenin gene. Our analyses also provide evidence that the major milk resource genes, caseins, which have similar functional properties as vitellogenins, appeared in the common mammalian ancestor ~200–310 Mya. Together, our data are compatible with the hypothesis that the emergence of lactation in the common mammalian ancestor and the development of placentation in eutherian and marsupial mammals allowed for the gradual loss of yolk-dependent nourishment during mammalian evolution.

Genome Alignment and Synteny



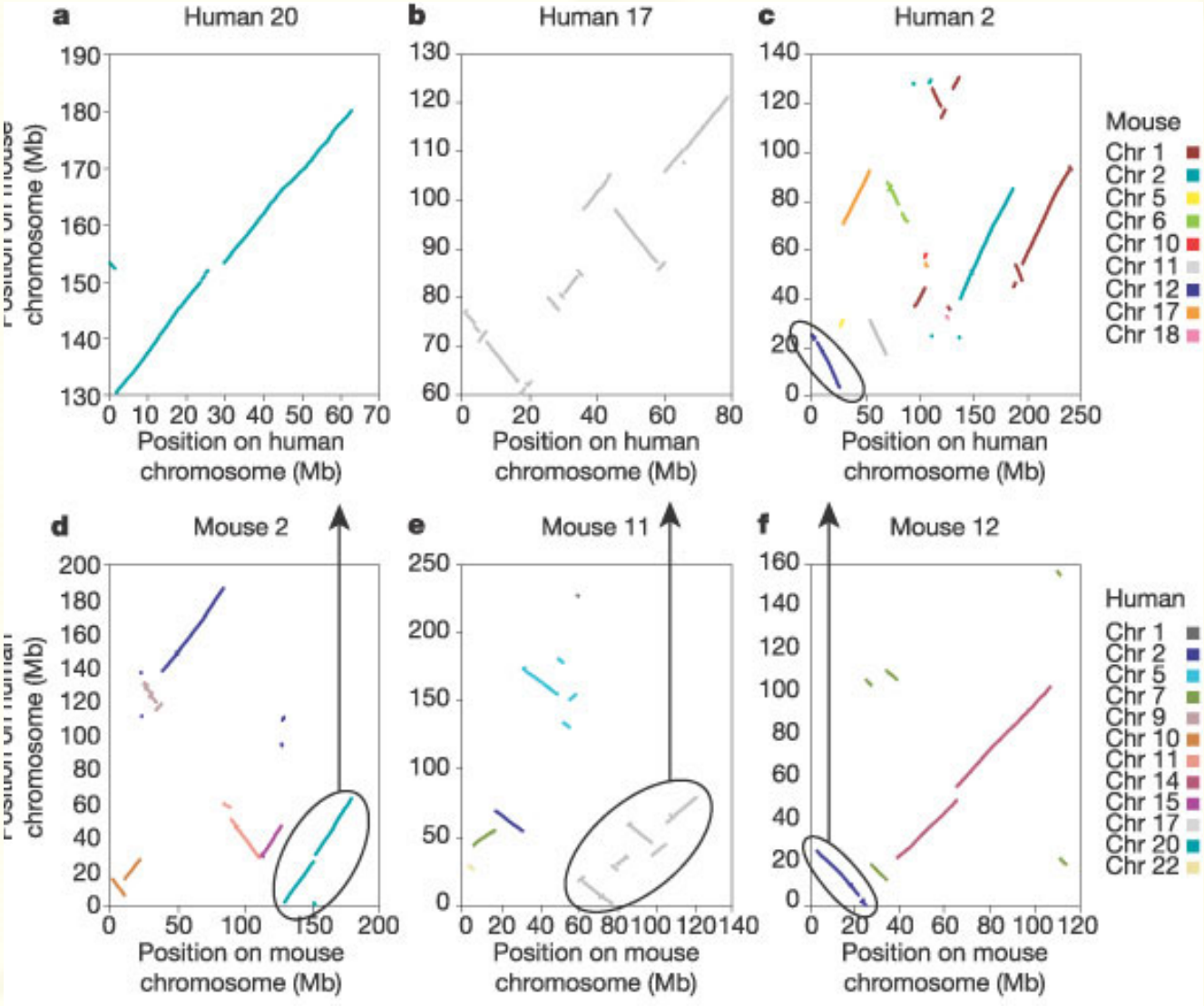




About 90% of the mouse and human genomes are in syntenic blocks.

[Waterston *et al.* *Nature* 2002]

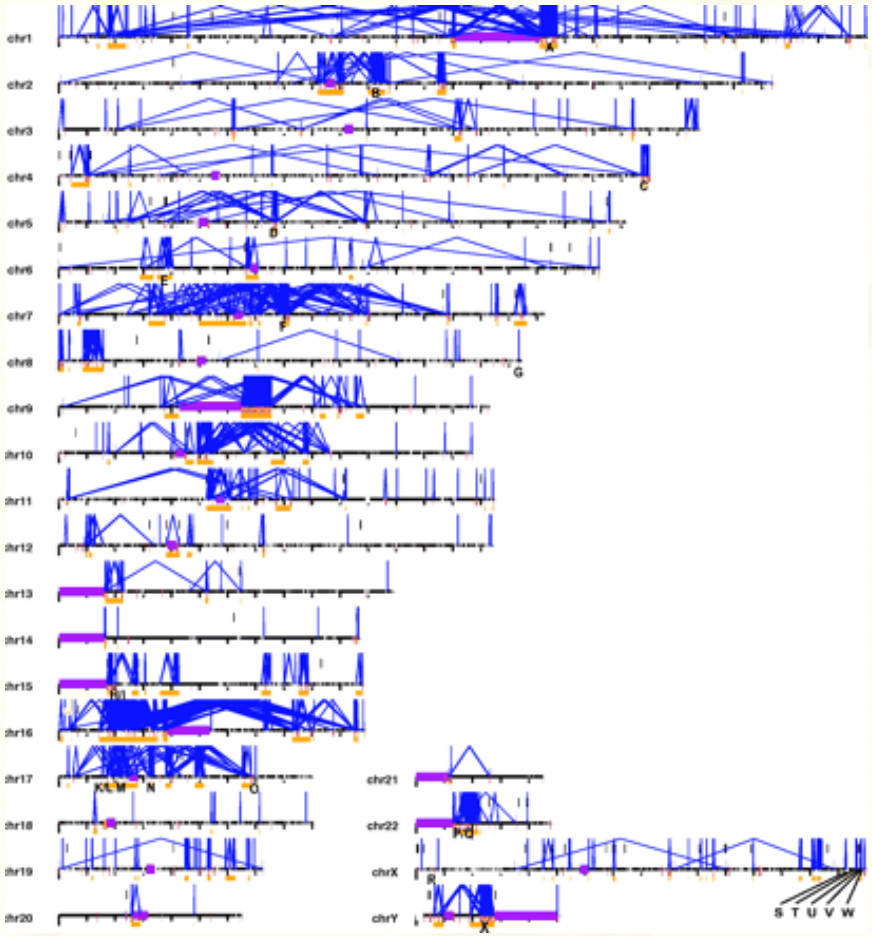
Human-mouse rearrangement events

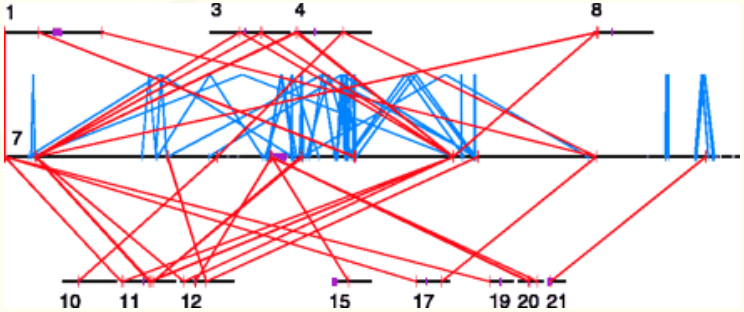


[Waterston *et al.* *Nature* 2002]

Genome Duplication

- **Segmental Duplication: 5% of the human genome** (E. Eichler etal)





chr7

"*Recent Segmental Duplications in the Human Genome*" Bailey, Science 2002

Human genome contains many "ultra-conserved regions"

- More than 500 regions (> 200 bp) that are absolutely conserved (100% identity) between orthologous regions of the human, rat, and mouse genomes.
- In addition, found > 5000 shorter ultraconserved sequence (100-200 bp).
- Nearly all of these segments are also conserved in the chicken and dog genomes, on average of 95 and 99% identity

Science. 2004 May 28;304(5675):1321-5. Epub 2004 May 6.

Ultraconserved elements in the human genome.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D.

Science. 2003 Nov 7;302(5647):1033-5. Epub 2003 Oct 2.

Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs).

Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, Antonarakis SE.

CNG-high

```
Human          CACACAAAGC ATAGGCTGCA AAATTATCCC CTGTCAAAAG AAAGAGCAGC
Green monkey   .......... .G........ .......... .......... ..........
Lemur          .......... .......... .......... .......... ..........
Mouse          .......... .......... .......... .......... ..........
Rabbit         .......... .......... .......... .......... ..........
Pig            .......... .......... .......... .......... ..........
Cat            .......... .......... .......... .......... ..........
Bat            .......... .......... .......... .......... G.........
Shrew          .......... .......... .......... T......... ..........
Armadillo      .......... .......... .......... .......... ..........
Elephant       .......... .......... .......... .......... ..........
Wallaby        .......... .......... .......... .......... ..........
Platypus       .......... .......... .......... .......... ..........

Human          TGCGGGTGCC AATTACAGCA ACCTTTCAAC CCTTTAGGTA CTGGAAACTA
Green monkey   .......... .......... .......... .......... ..........
Lemur          .......... .......... .......... .......... ..........
Mouse          .......... .......... .......... .......... ..........
Rabbit         .......... .......... .......... .......... ..........
Pig            .......... .......... .......... .......... ..........
Cat            .......... .......... .......... .......... ..........
Bat            .......... .......... .......... .......... ..........
Shrew          .......... .......... .......... .......... ..........
Armadillo      .......... .......... .......... .......... ..........
Elephant       .......... .......... .......... .......... ..........
Wallaby        ..T.A..... .......... .......... .......... .......G.
Platypus       .......... .......... .......... .......... ..........
```

CNG

```
Human          AACGTTAACC TGCCTGGGTC CC-GGGGTAT GGGAGCGCTA AATCTTCCGT
Green monkey   .......... .......... ..-....... .......... ..........
Lemur          ...C...... .......... ..-....... .......... ..........
Mouse          .......... .......... ..-....... .......... .G......C.
Porcupine      ...A...... ......A... ..-....... .......... ..........
Pig            ....G..... ....C...G. ..-....... .......... ........A.
Cat            .GG.C..... ....---.C. GG-C..C.C. ..CG..C.CC ..........
Shrew          ....CG.... ....CT-.CT .T-...C... ........A. ..G..CT..G
Armadillo      ....G..... ....CT.... ..-...C... .......... ..........
Elephant       .C....C... C...C..... ..--..C... ........-- .C........
Wallaby        ..T....... ...A...... TT-....... .......... ........T.
Platypus       .....G.... ..A..A..CT .AG.....G. .....A.... .CC.....T.

Human          CCTACCACTG TATAATGAAC AGACTGTTTC ATTAGTGGGA CAATTCACTC
Green monkey   .......... .......... .......... .......... ........—
Lemur          .....A.... .......... .......... .......... .........-
Mouse          ...T.A.... .......... .......... ....A..A.. ..........
Porcupine      .....A.... .......... .......... .......... ..........
Pig            .....G.... .......... .......... .......... ........—
Cat            ..C..A.... A......... .......... .....C.... ..........
Shrew          ..C....... .......... .......... .....G.A.. ..........
Armadillo      ....AA.... .......... .......... .......... ..........
Elephant       .....A.... .G........ .......... .....C.... ..........
Wallaby        A....A.... .......... .......... .......... ..........
Platypus       A.G..A.... .......... .......... ......C... .T......T
```

Coding

```
Human          GCTCAGTCAC TCCAGAATCC CTGAGAAAAG CAATAGAGGC TGTATCACCG
Lemur          .......... .......G.. T......... .......... .........A
Mouse          ...T...... .......ATG ........G. .......... CA.T......
Porcupine      .T....C... .......A.A .......... .......... ...T..T..A
Rabbit         ...T...... .......A.T .....G.... .......A.. .........A
Pig            .T.T...... .......A.T .......... .........A CA.......A
Cat            .......... .......A.. ..T....... .......... CA.C.....A
Bat            ...T...... .......A.G .......... .......... CA.......A
Shrew          A......... .......A.T T...A..G.. ....C..... CA.......A
Armadillo      ..AT...... .......A.. ........G. .......... CA.......A
Elephant       ...T...... .......A.. ........G. .T........ A........A
Opossum        AA.T.A.... .......G.. T...A..... .T....A.. CA..G....A

Human          GGGCTATATA GAGTTAGTAT CACAAGTGAA GTTGAG---A GTACCTCAAA
Lemur          ....A..... .......C.. TG........ ...A..---. ......TG..
Mouse          ....A...C. .......... TG........ .....A---. ....G.C.G
Porcupine      ....AG.... C......... T.........C .....---. A......C..
Rabbit         ....A..... .......... TG........ .C...---. ..T..C.C..
Pig            ....A..... .....C..G T.........C A.....---T ..........
Cat            ...AA...C. A......... T......... .....A---. ..........
Bat            ..A.A..... .......... TG.....G. A.C...---. ..........
Shrew          ...AA...C. .......C.. T....A.... .....---. .....G....
Armadillo      ....A....C .......... ..T...A... .A...A---G .....C....
Elephant       ....A...C. .T........ T.........T .....---. .....C...G
Opossum        ....AG.... A.......C. TG.C....G AG.C.T---G .C.G.C.T..
```

ncRNA

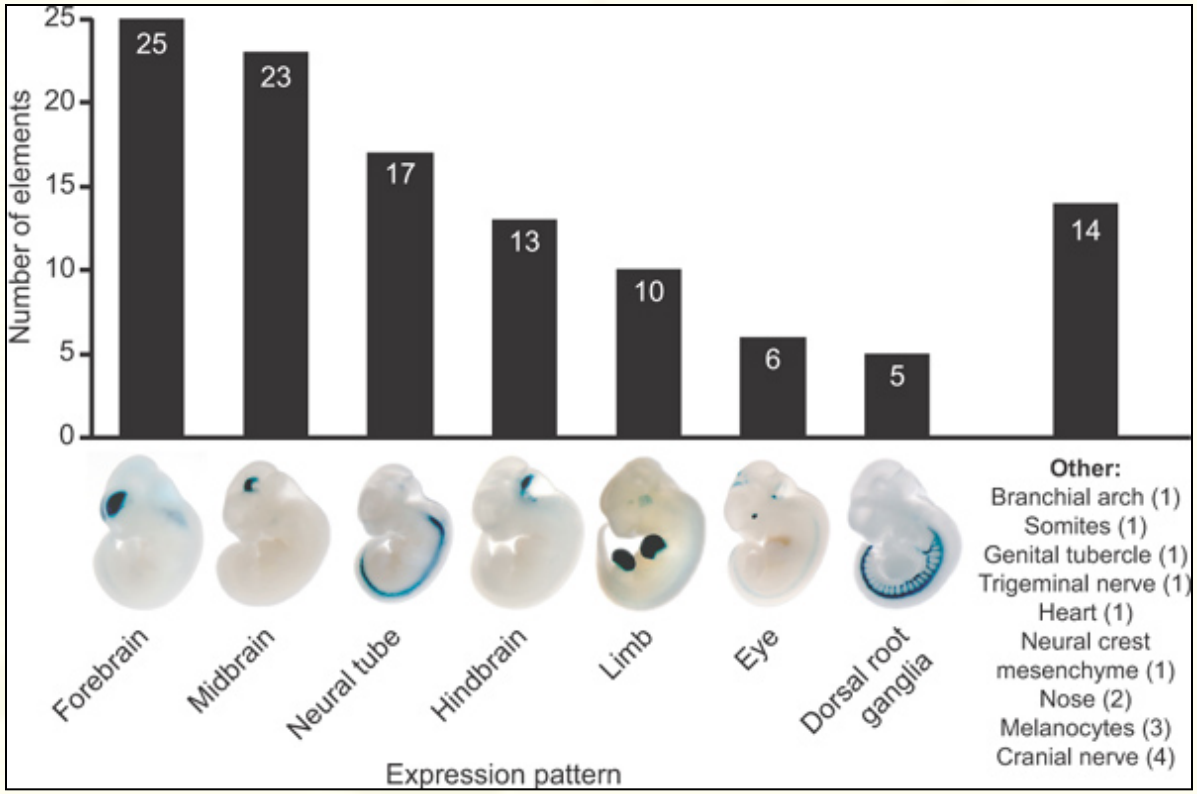
```
Human          TTCTGCCTAC CCTGT-TGGT ATAAA-GATA TTTTGAGCAG ACTGTAAACA
Greenmonke     .......... .....-.... ...---.... .......... ..........
Lemur          .......... .....-.... .....-.... .......... ...C......
Mouse          .......C.. .....-.... .....-..C. .........T ..........
Porcupine      .......C.. ...T.-.... ...C.-.... .......... .A........
Pig            .......C.. .....-.... .....-.G.. .......... .T...G....
Cat            .......C.. .....-.... .....-.... .......... .T...G....
Bat            .......C.. .....-.... .....-...G .......... .T...GT...
Shrew          ....AT.C.. .T...-.... .....-T... .......... .T...G....
Armadillo      ....A..C.- .T...-.... .....-..C. .......... .T......G.
Elephant       .......C.. .....-.... .....-.G.. .......... .T........
Wallaby        ......T... .....G.... ...GGT.... .......... .T....T.A.

Human          AG-AAAAAAA AAATCATGCA TTCTTAGCAA AATTGCCTAG TATGTTAATT
Greenmonke     ..-....... .......... .......... .......... ..........
Lemur          ..-....... .--....... ..G....... .........T ..........
Mouse          .AC...C... .........T ..G....T.. .......... ......G...
Porcupine      ..-....... ---....... ..G....T.. .......... ..........
Pig            ..-....... ..-......T ..G....... .......... ..........
Cat            ..-....... ..-......T ..G....... .......... ..........
Bat            .AC.....C. .........T ..G....... .......... ..........
Shrew          C.-....... ---.....— ..T....... ....AT.... ..........
Armadillo      CA-....... ---....... ..T....T.. .......... ..........
Elephant       ..G....... .......... ..G....... .-...T.... .......C..
Wallaby        TA-..T..T. ..-......G ..G....... .G..----- ---------
```

Ultra-conserved regions are more conserved than coding regions

E. T. Dermitzakis, Science 2003

Using reporter assay to test the function of these ultraconserved regions in transgenic mice

- "***In vivo* enhancer analysis of human conserved non-coding sequences**", Pennacchio, et al Nature 2006
- Tested 167 elements, 75 (45%) functioned reproducibly as tissue-specific enhancers at embryonic day 11.5, most involved in developing nervous system.



Phylogenetic Shadowing

- **Objective:** we want to compare genomic sequences to find functionally important sequences such as exons or regulatory elements.

- **Hypothesis:** these regions should evolve slower than other background regions.

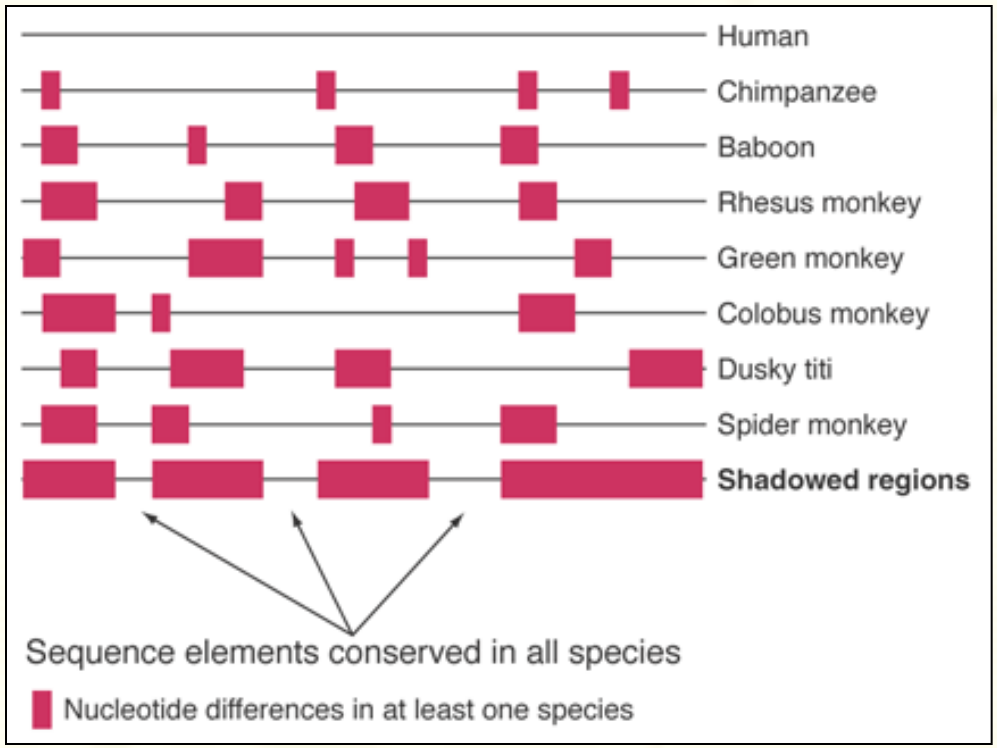
- **Approach**: We can align sequences and look for regions that have slower evolutionary rate (valleys).

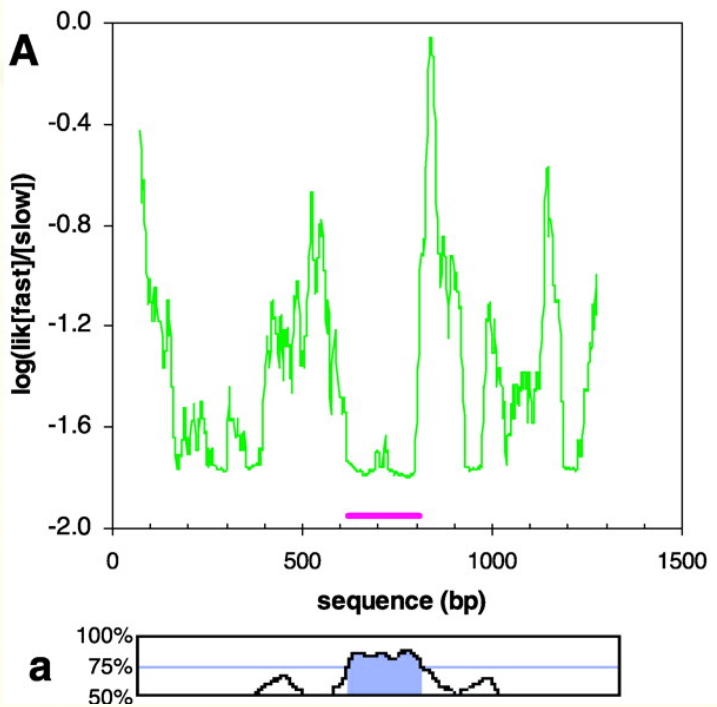
Phylogenetic Shadowing

Science. 2003 Feb 28;299(5611):1391-4.

Phylogenetic shadowing of primate sequences to find functional regions of the human genome.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM.





Evolution and divergence of gene expression

- How divergent are gene expression among animals ?
  - Do the orthologous genes in human, chimpanzee and monkeys have similar expression breadth and expression level ?
  - What about expression levels of orthologous genes from more distantly related species such as chicken and frog ?
- How divergent are gene regulatory mechanisms among orthologs ?
- How divergent are gene expression and regulatory mechanisms among individual humans ?

Evolution of gene expression among primates

PLoS Biol. 2004 May;2(5):E132. Epub 2004 May 11.

A neutral model of transcriptome evolution.

Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, Pääbo S.

"neutral model": divergence in expression level proportional to time of divergence

Science. 2005 Sep 16;309(5742):1850-4. Epub 2005 Sep 1.

Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Pääbo S.

Nature. 2006 Mar 9;440(7081):242-5.

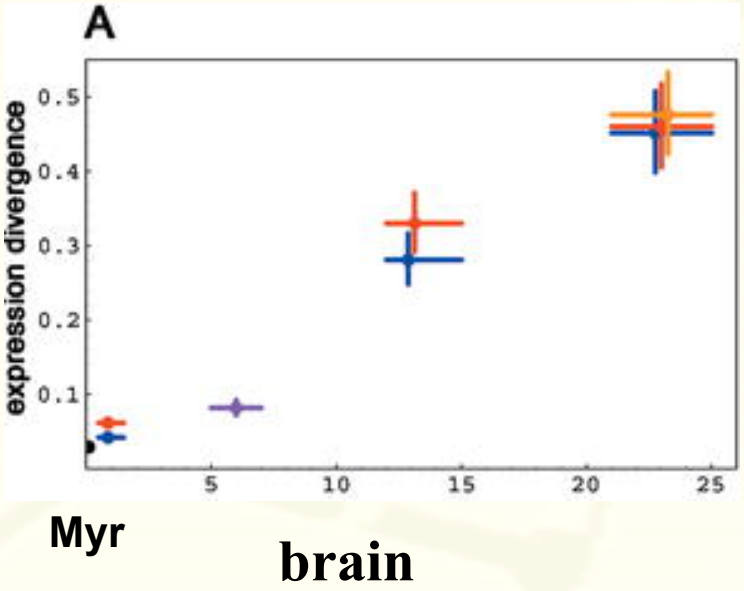
Expression profiling in primates reveals a rapid evolution of human transcription factors.

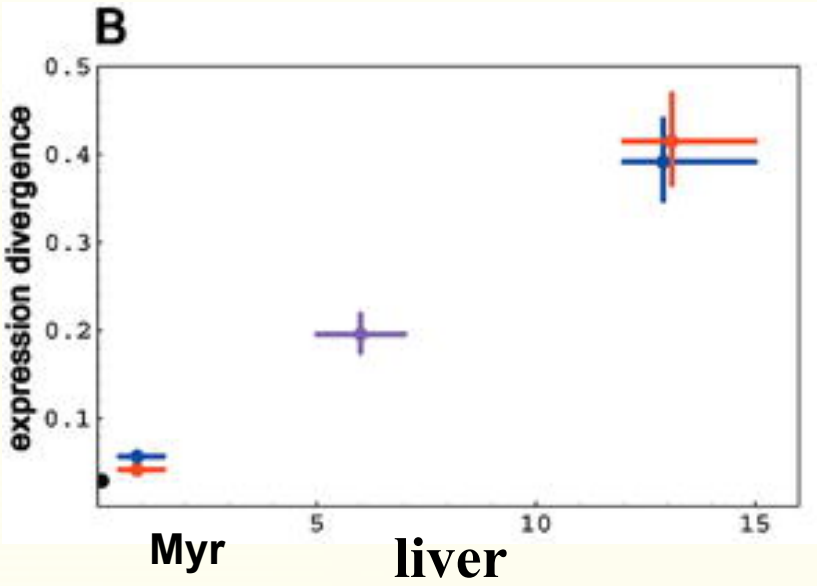
Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP.

Evolution of gene expression among primates

Major observations:

- "…changes in expression proportional to divergence time", suggesting a neutral model
- " … brain shows the least differences between the species whereas liver shows the most."
- "transcription factors show the biggest difference"



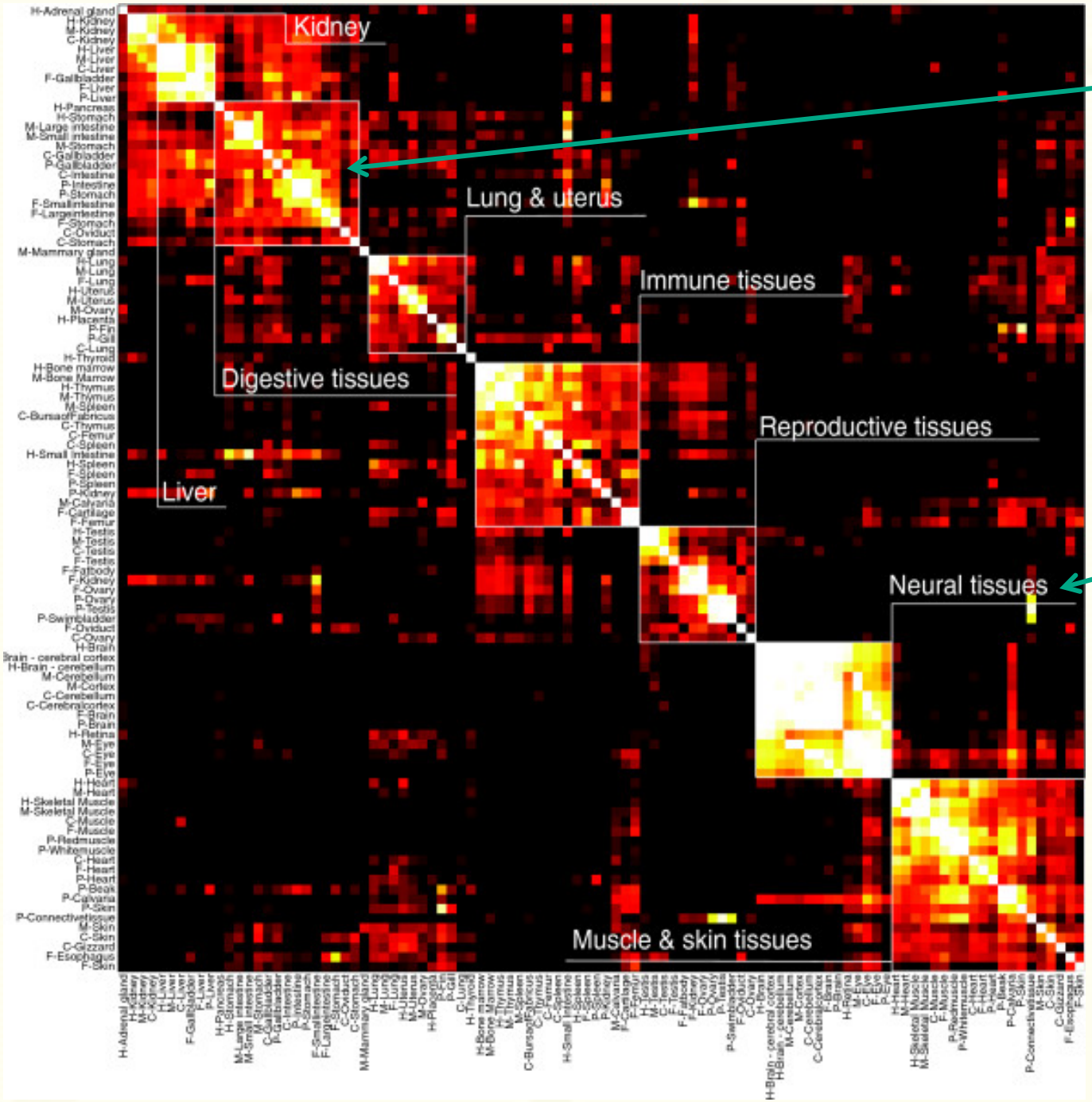


Evolution of gene expression among human / mouse / chicken / fish / frog

- Chan et al designed custom microarray to survey the expression level of 20 tissues in five vertebrate species.
- Strikingly, conservation of expression correlates poorly with the amount of conserved regulatory elements. Many genes show conserved human/fish expression despite having almost no conserved regulatory sequences.
- “We find that, on average, transcription factor gene expression is neither more nor less conserved than that of other genes.”

It is important to control for confounding factors such as seasonal effects, individual variations, age, sex etc.

Chan et al . Conservation of core gene expression in vertebrate tissues Journal of biology 2009



Digestive tissues

Correlation between orthologous tissues from 5 animals

Neural tissues

0 0.1 0.2 0.3 0.4 >0.5

Pearson correlation coefficient

Divergence of TF binding sites in 5 vertebrates

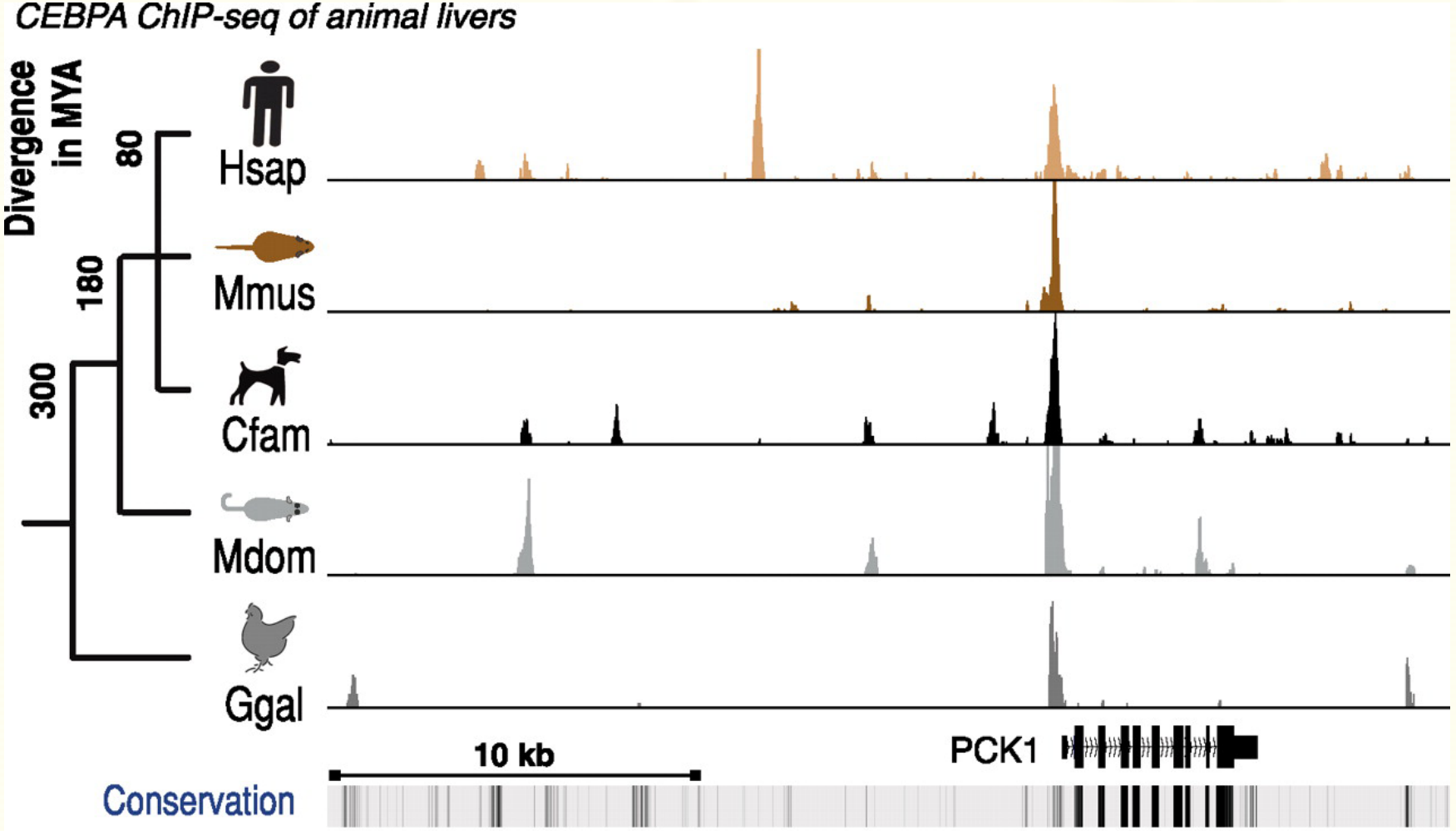
- Schmidt et al collected liver samples from human, mouse, rat, dog and chicken, and determined the binding sites of two transcription factors: CEBPA and HNF4A.

- *“binding specificity of the TFs are mostly unchanged among the five species”*
- “however, *most binding is species-specific, and aligned binding events present in all five species are rare”*

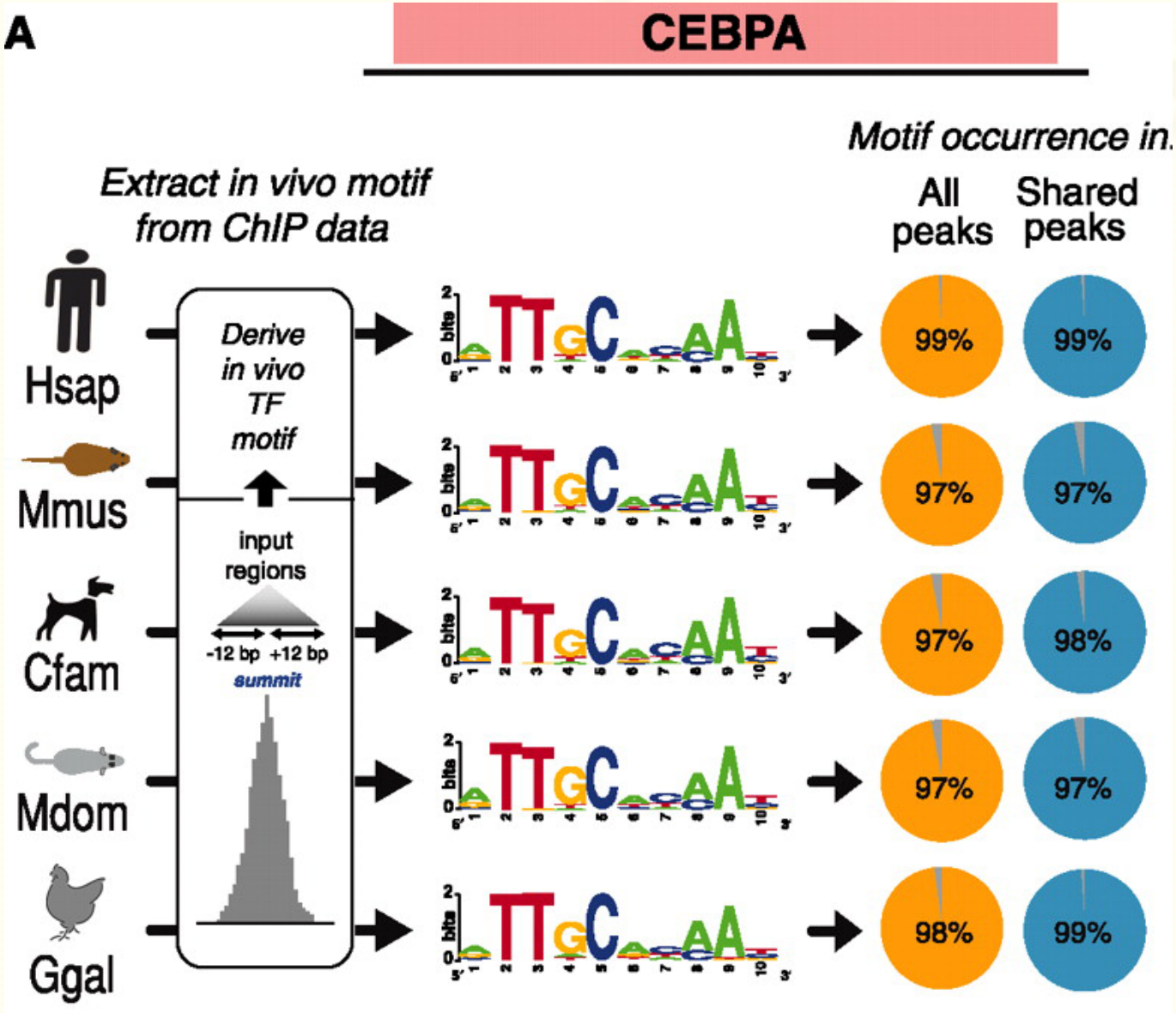
Science. 2010 May 21;328(5981):1036-40. Epub 2010 Apr 8.

Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding.

Schmidt D, **Wilson** MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, Talianidis I, Flicek P, Odom DT.



Divergence and conservation of binding by CEBPA



“binding specificity of the TFs are mostly unchanged among the five species”

How variable are the gene expression among human individuals ?

- **Stranger** et al used microarray to measure genome-wide gene expression of 270 individuals from Chinese, Japanese, Caucasian, and Africans.
- They used Epstein-Barr virus–transformed lymphoblastoid (淋巴) cell line

- **Questions**:
- (i) how variable is human gene expression?
- (ii) What is the proportional contribution of cis- vs trans-regulatory element?
- (iii) What is the proportional contribution of SNPs and CNVs ?

Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes

Barbara E. Stranger,[1] Matthew S. Forrest,[1] Mark Dunning,[2] Catherine E. Ingle,[1] Claude Beazley,[1] Natalie Thorne,[2] Richard Redon,[1] Christine P. Bird,[1] Anna de Grassi,[3] Charles Lee,[4,5] Chris Tyler-Smith,[1] Nigel Carter,[1] Stephen W. Scherer,[6,7] Simon Tavaré,[2,8] Panagiotis Deloukas,[1] Matthew E. Hurles,[1*] Emmanouil T. Dermitzakis[1*]

Population genomics of human gene expression

Barbara E Stranger[1], Alexandra C Nica[1], Matthew S Forrest[1], Antigone Dimas[1], Christine P Bird[1], Claude Beazley[1], Catherine E Ingle[1], Mark Dunning[2], Paul Flicek[3], Daphne Koller[4], Stephen Montgomery[1], Simon Tavaré[2], Panos Deloukas[1] & Emmanouil T Dermitzakis[1]

Observations:

cis- regulatory variation is more dominant than **trans-** effects, which is the primary effect contributing to phenotypic variation in humans.

SNPs and CNVs captured 83.6% and 17.7% of the total detected genetic variation in gene expression

Variation in TF binding among individual humans

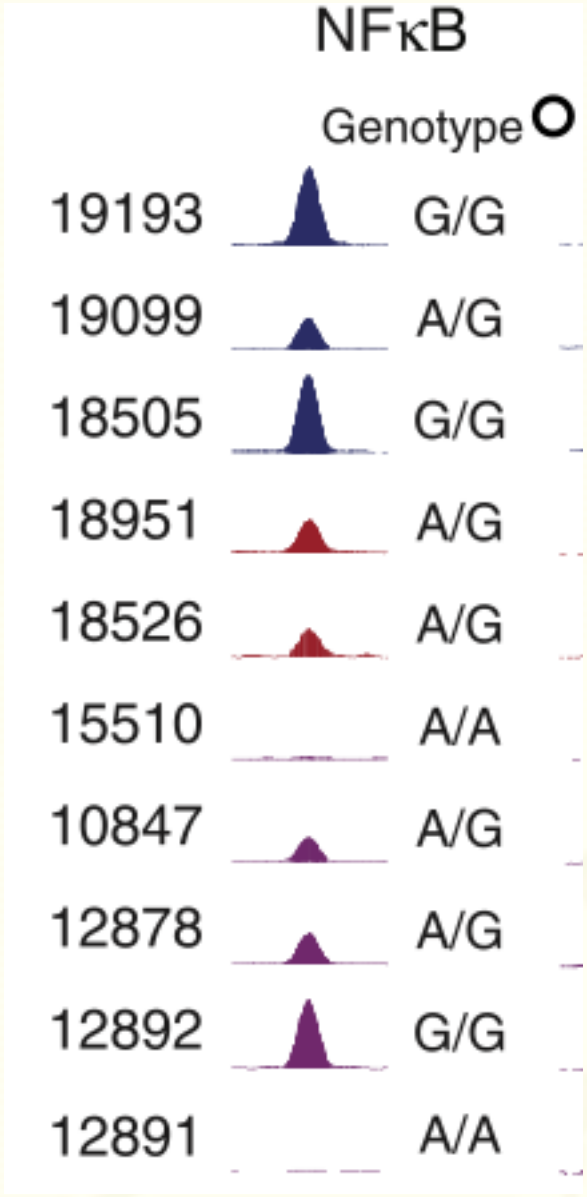
- Kasowski et al used ChIP-seq to measure binding of RNA polymerase II (Pol II) and a nuclear factor kappa B (NF*K*B) among 10 individuals (using lymphoblastoid cell lines)

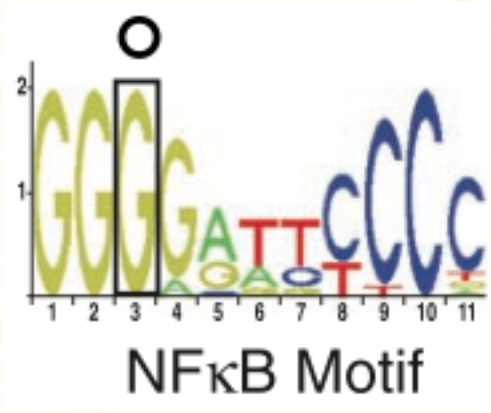
- They found 25% of the Pol II binding sites and 7.5% of the NF*K*B binding sites are variable among individuals.

- They conclude variation of TF binding are responsible for phenotypic differences among individuals.

Variation in Transcription Factor Binding Among Humans

Maya Kasowski,[1*] Fabian Grubert,[1,2*] Christopher Heffelfinger,[1] Manoj Hariharan,[1,2] Akwasi Asabere,[1] Sebastian M. Waszak,[3,4] Lukas Habegger,[5] Joel Rozowsky,[6] Minyi Shi,[1,2] Alexander E. Urban,[1,7] Mi-Young Hong,[1] Konrad J. Karczewski,[2] Wolfgang Huber,[3] Sherman M. Weissman,[7] Mark B. Gerstein,[5,6,8] Jan O. Korbel,[3,9]† Michael Snyder[1,2]†





Effect of polymophisms on the binding of transcription factor NF*K*B

Summary

- We discussed the basic facts on human genome, human genes, and their evolution.
- Repetitive elements make up more than half of the human genome. Only very small fraction of the genome code for proteins.
- There are state-of-art technologies to investigate the divergence and variation of gene expression, and gene regulation.

The END