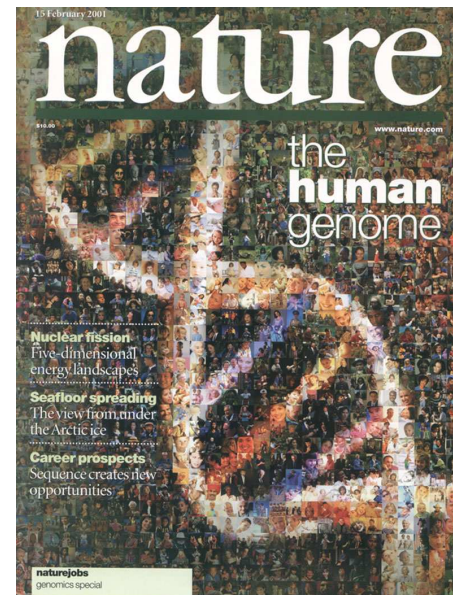
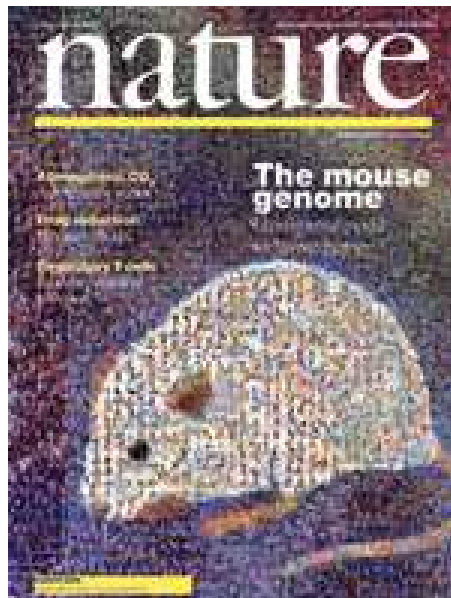


Non-genic evolution and selection in the human genome

or:

“Junk DNA”



Two puzzling observations

About 1.5% of our genome codes for protein. What is remaining 98.5% for? Just “junk”?

Roughly half of our genome consists of *transposable elements*.

Thought to be ‘selfish genes’; no function, apart from promoting own existence.

Rest could be old, unrecognizable TEs, or just DNA tagging along.

So yes?

But:

- Comparison of mouse and human genome resulted in estimate that 5% of our genome is under purifying selection, i.e. is functional.
- The distribution of the Alu transposable element seems in contradiction with its mechanism of transposition. Selection pressure? Function?

Outline of talk

This talk:

1. Overview of neutral evolution (mutation, no selection)
2. Transposable elements, their evolution, and their effect on us
3. CpG methylation: function, and consequences for genetic mutations
4. Ongoing effects; diseases caused by TEs and CpG methylation
5. Parasites or symbionts:
Are transposable elements biologically functional?
6. The proportion of our genome under selection
7. Conclusions

1. Mutations

Evolution is result of two opposing forces:

- **Mutation**, “proposing” changes to DNA, and
- **Selection**, either or not accepting those proposals

Non-functional DNA is (by definition) not under any selection pressure.
Gives opportunity to study **process of mutation** by itself.

Mutations - microscopic view

To see mutations, compare **homologous non-functional DNA** from e.g. human and mouse. Here is a 318 bp example:

human chr10 276805-277123, mouse chr13 106554329-106554631

```
GGCTAAAGTAGTTTCTTTCTTTTTCAGCTGGATGAACTGCAGCTTTGCAAGAATTGCTTTTACTTGTCAAAT
GCTCGTCCTGACAACTGGTTCTGTTATCCTTGTGTATGTGAAATTTTACCTCAAGTGTGTAACATACAGCTC
TAAGGAA--GTTTATT-----TCCAGTTTGGTTAAAGATTATAATTTCTTCCTAAAT-----TTTTAAAAT
ATCATAGTATATATGGGTTTAATGAGTATAAACAACTGAAAA-----AAAGTAACAGGGCTTGTTATCCATTT
CTGATTTTAAAATGAGATACTGAATAAAAATAAATATTGGGTCct
```

```
GGTTCAAACCGTTGCTTTGCTTTT-CAGCTGGACGAGCTGCAGCTTTGCAAGAACTGCTTCTATCTGTCAAAC
GCACGGCCCGACAACTGGTTCTGCTACCCTTGCGTATGTGGGATTTT-----GTGATTTTCCTTTA
TAGAGAACTGTTTGTTAGTTATCCCTGTTTACTTACGGATGGTAAAG--CCTGTGAAATATACTTCTTTACAT
CATATAGTTCATGTGATCTTAATGAGTATGCACAGATGAAACCTGACAAAGCAACAA---CCATTCTCAGTTC
CGAATGTCAACA----GCACTGAGTAA-----GACCAT
```

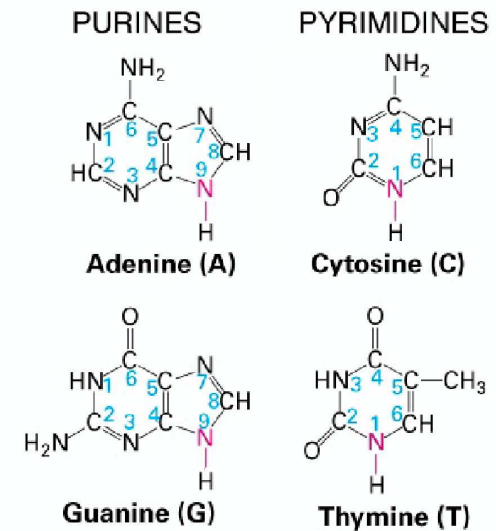
BLASTZ reciprocal best hits from human, mouse, rat database at UCSC; see <http://mgc.ucsc.edu>

- On non-functional DNA, **66.7% of nucleotides are identical**.
- **Many small deletions** have occurred (not insertions it turns out)

Mutations - nucleotide substitutions

Substitution process usually modelled by **rate matrix**. Entries denote **probability per unit of time** of particular substitutions.

		To:			
		A	C	G	T
From:	A	.	0.070	0.220	0.060
	C	0.090	.	0.068	0.285
	G	0.280	0.070	.	0.090
	T	0.060	0.230	0.073	.



Estimated from human-mouse DNA (chromosome 21), time unit = human-mouse divergence time, $\approx 2 \times 75$ Myr

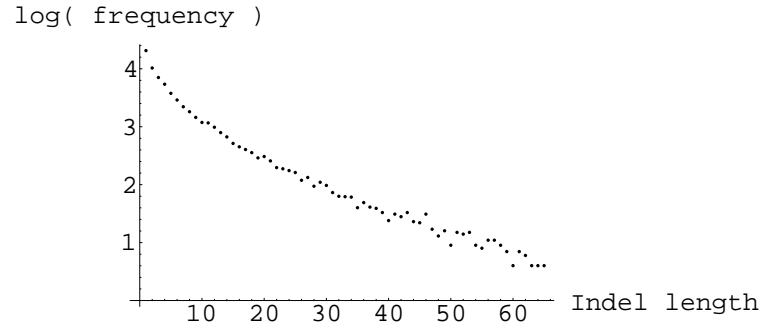
- **Transitions** ($A \leftrightarrow G$, $C \leftrightarrow T$) about **3×** more frequent than transversions.
- Total average substitution rate: $\approx 2.5 \times 10^{-9}$ substitutions per site per year
- Current substitution rate for humans about **5×** higher than for mice.

Mutations - small indels

Second important class of mutations: **indels**.

In alignments, **insertions** and **deletions** cannot be distinguished, and are together referred to as **indels**. However deletions in fact outnumber insertions by a factor 2 – 3. Nature vol 420, 5 Dec. 2002

- Most indels are **short** (1/3 are **single-nucleotide** indels, 60% are ≤ 3 nucleotides.)



- **Average indel length** in human-mouse alignments: 5.4
- Indels closely follow **Poisson distribution**. Mean distance ~ 22 nucleotides.
- Indel rate 3.0×10^{-10} indels per site per year; 1.6×10^{-9} nucleotides per site per year.

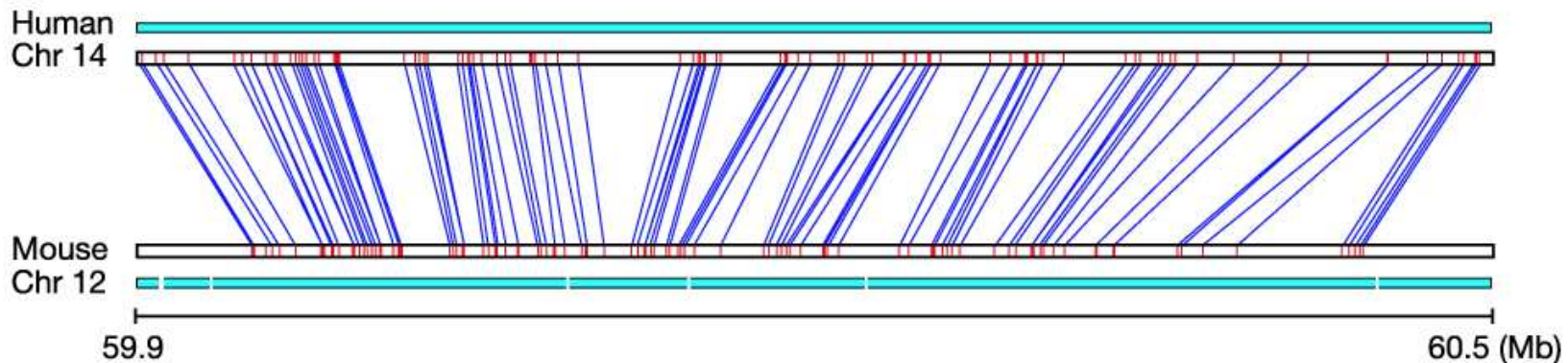
Indel rate same order of magnitude as substitution rate.

(Very different in coding regions, where indels are more often deleterious than substitutions.)

(Numbers estimated on **local alignments**; this ignores large indels (more than ~ 300 nucleotides))

Mutations - somewhat larger scale

This picture shows *reciprocal best hits* (homologous regions) between a **600 kb stretch** in the human and mouse genomes.



- Hits predominantly **linearly** distributed: **conserved syntenic**.
- At this scale, **long indels** occur.

These **long indels** are often caused by **transposable elements**.

2. Transposable elements

Transposable elements are pieces of genetic information that somehow manage to **multiply themselves** and **move around** in the genome.


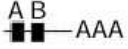
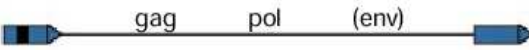
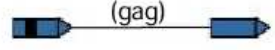


History: First suspected in **1940** from work by Barbare McClintock on genomic instability in maize. Existence of transposable elements was proven experimentally in **1970s**. She received Nobel prize in **1983**.

Four classes of transposable elements live in our genome:

- DNA **transposons**
- LINEs (long interspersed nuclear elements), **retroposons**
- SINEs (short interspersed nuclear elements), **non-autonomous retroposons**
- Retroviruses and retrovirus-like LTR (long terminal repeat) **retrotransposons**

Transposable elements in human genome

Classes of interspersed repeat in the human genome

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

- LINEs and SINEs were first distinguished by their length. Turned out to have different ‘lifestyle’ and are now distinguished by that.
- DNA **trans**posons and retro**trans**posons code for *transposase* (or related *integrase*). Insert double-stranded DNA into host genome.
- LINE **retro**posons and retrovirus-like **retro**transposons code for *reverse transcriptase*. Go through intermediate RNA phase.

Transposable elements in human genome

In our genome:

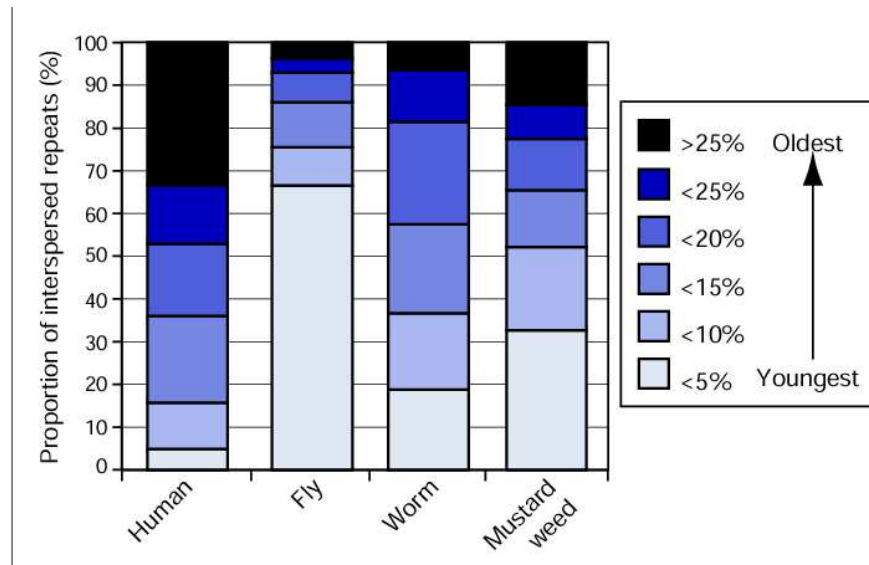
- One LINE element (**LINE1**) is particularly abundant and active
- One SINE element (**Alu**) is particularly abundant and active
- DNA transposons have been active, but not active now.

Repeat class	Fraction of genome (%)	Copy number
LINEs	20.99	850,000
LINE1	(17.39)	
SINEs	13.64	1,500,000
Alu	(10.74)	
LTRs	8.55	450,000
DNA elements	3.03	300,000
Unclassified	0.15	
Total transposable elements	46.36	

Activity of transposable elements

Activity varies greatly per organism:

- **Humans:** Rather quiet, ≈ 50 active LINEs, no or very few active DNA transposons, no LTRs through to be active.
- **Mice:** ≈ 3000 active LINEs, many active DNA transposons, many active LTRs.
- **Maize:** Genome size doubled in last ≈ 3 Myr because of transposon insertions.

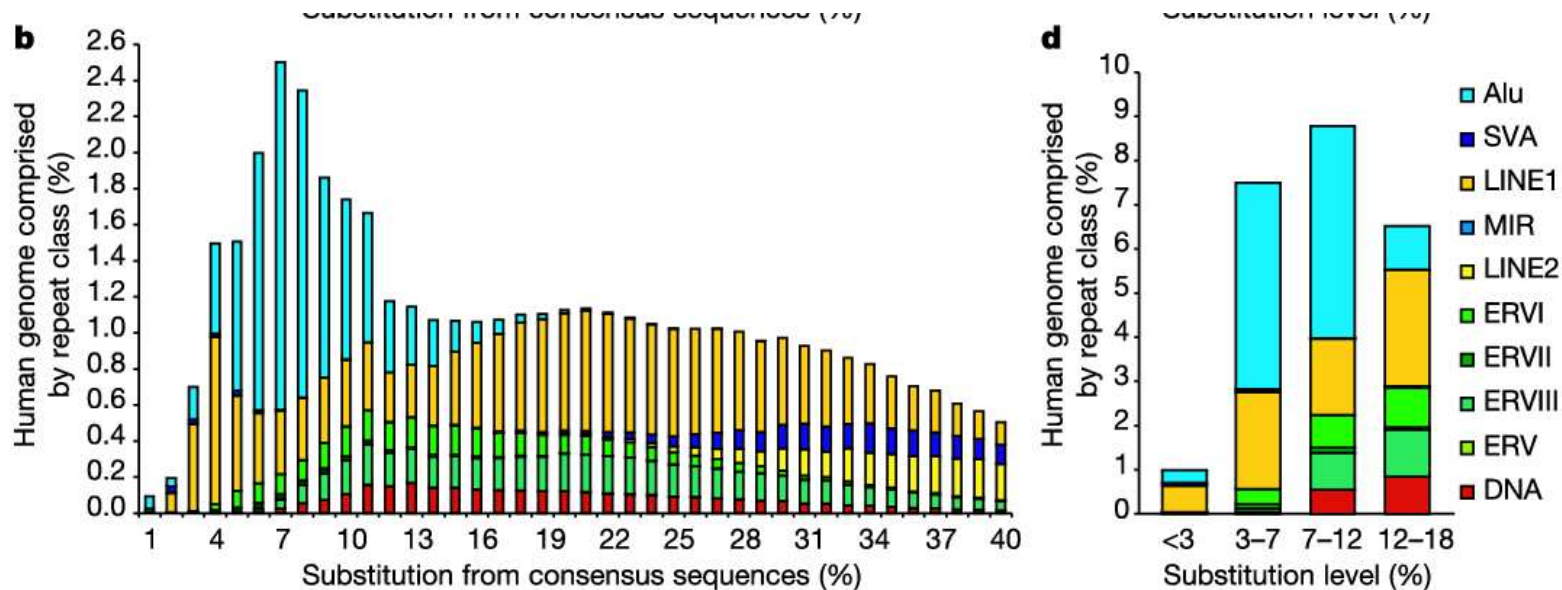


From: Nature vol 420, 5 Dec. 2002

In **fruitfly**, most TEs have few mutations (relative to consensus = ancestor): young.

In **human** DNA, there are relatively few young transposable elements.

Human transposable elements - activity over time



Frequency of various TEs, against proportion of substitutions from consensus sequence (~ age). Right-hand side: Same, larger bins.

Key: blueish SINE; yellowish LINE; greenish LTR, red, DNA transposon.

Activity **Alu** peaked at 7% divergence ~ 40 Myr, then dropped.

Activity **DNA transposons** and **LTR retrotransposons** started diminishing earlier.

(Dropoff at high-sequence-diversity bins (partly?) due to difficulty detecting highly diverged repeats.)

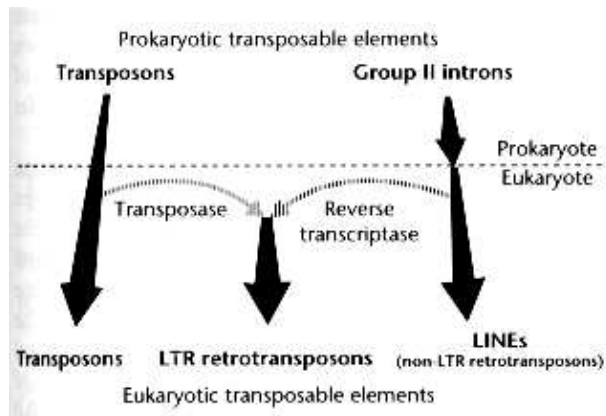
Evolution of transposable elements

Transposable elements exist in **all living organisms**.

All eukaryotes have LINE elements; most recent common ancestor >600 Myr old.

- Eukaryotic DNA transposons most related to **prokaryotic transposons**.
- Reverse transcriptase of LINE elements are most related to reverse transcriptase of **prokaryotic group-II introns** (mobile elements).
- LTR elements use **two-step mechanism** (both reverse transcription and DNA integration). No prokaryotic elements known with similar mechanism.

This suggests following evolutionary history:

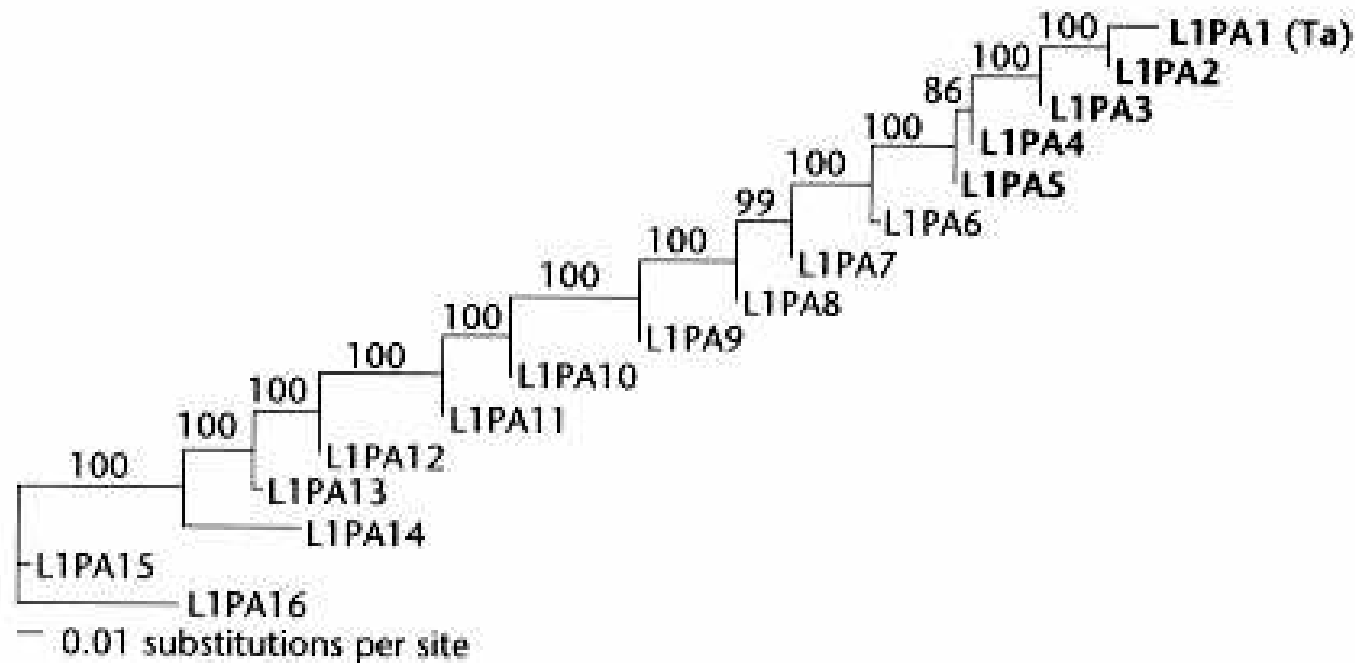


From: Nature Encyclopedia of the Human Genome (EHG), T 617

Evolution of transposable elements: LINEs

Because of high copy numbers, old transposable elements can be reconstructed.

Result for LINE elements, over last 60 Myr (\approx up to human-mouse split):

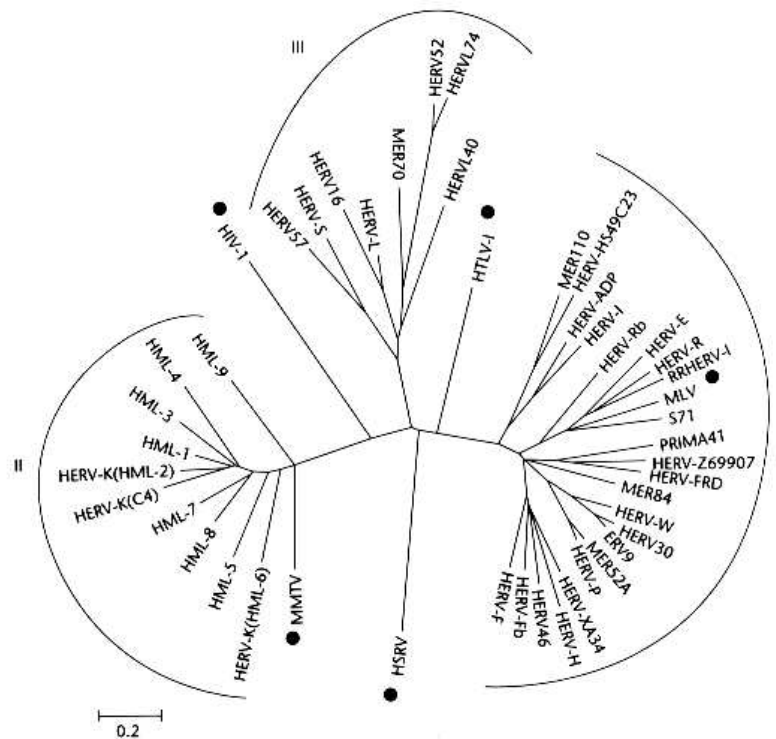


Generally, at any given time in last 60 Myr, one LINE family is dominant.

Evolution of transposable elements: LTRs

LTRs are related to retroviruses. Difference: retroviruses have functional *env* gene (envelope protein, protection and binding/infecting cell), which LTRs lack.

Chicken and egg: Did LTRs evolve from retroviruses, or vice versa? Retrotransposons probably evolved back and forth.

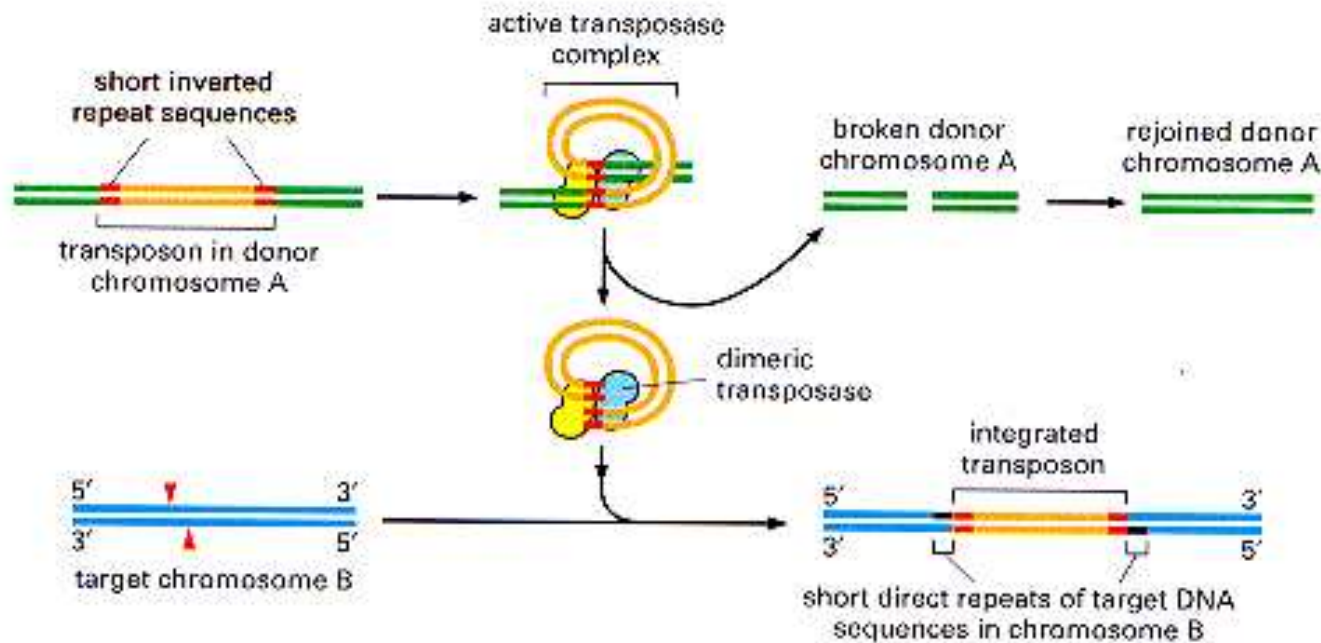


Phylogeny based on reverse transcriptase. Retroviruses MMTV, MLV are similar to current retrotransposons; HIV, HRS and HTLV are not.

- **HIV**: hum. immunodef. virus
- **HSRV**: hum. spumavirus
- **HTLV**: hum. T-cell leukemia v.
- **MMTV**: mouse mammary tumor v.
- **MLV**: mouse Moloney leukemia v.

DNA transposons: mechanism

Transposons move by a **cut-and-paste** mechanism.



From: Alberts et al.,
The Cell

Multiply when excising themselves during mitosis, when DNA repair mechanisms can recover removed portion from newly duplicated strand.

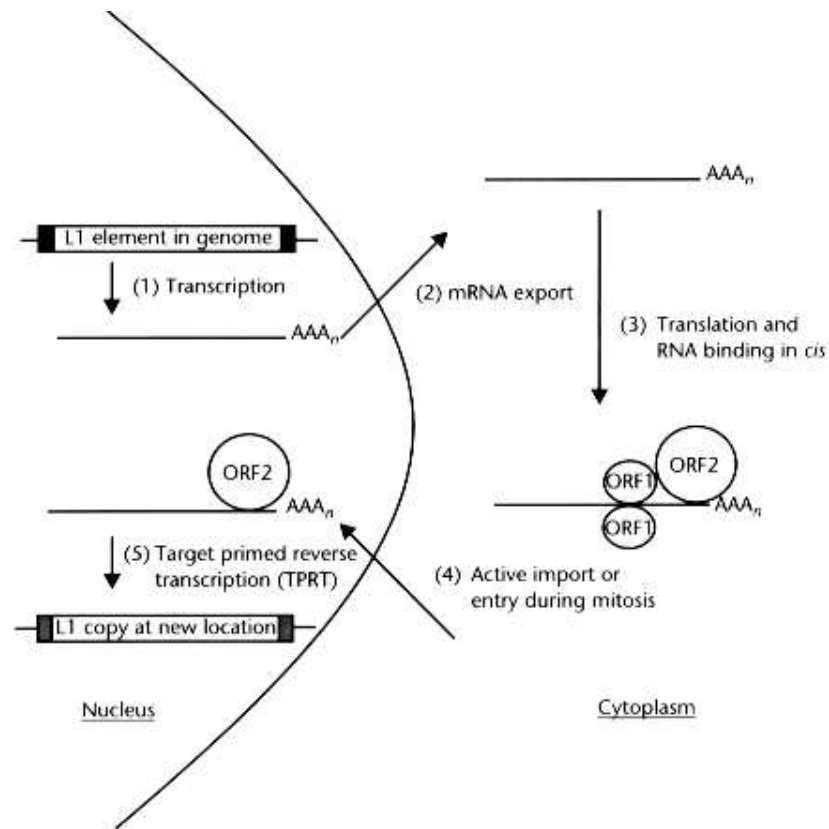
Work **in trans**, i.e. gene gets translated, then transposase looks for “itself” in genome. Recognises itself by 10 – 30 bp stretch, so often binds to inactive transposon. Result: mutations accumulate, copying becomes less efficient.

LINE elements: mechanism

LINEs have

- their own (pol II) **promoters**,
- two **ORFs** coding for protein,
- 3' binding site for ORF2 protein,
- **poly-A tail**

Act **in cis**, i.e. proteins coded by LINE bind to **own** mRNA.



After translation and binding to own mRNA, the LINE element:

- Gets **transported** back into nucleus;
- **Cleaves** host DNA, preferentially at **TT|AAAA**;
- **Transcribes** a DNA copy from RNA directly into genome. New copy is flanked by a 7–20 bp **target site duplication** from cleaved-and-repaired host DNA.

From: EHG R 53

LINE elements: more mechanism

LINEs contain weak **transcription termination signal**, translation by pol II may continue until poly-A signal (TTTT) found in host genome.

(Evolutionary beneficial, otherwise insertion into a gene leads to premature end of translation, probably deleterious)

Result: LINEs may copy much more than LINE element themselves (**5' extension**).

Reverse transcription is **inefficient** and may not complete.

Result: Incomplete LINE elements (**3' truncation**), or DNA duplications without any apparent LINE contribution.

LINEs also work **in trans**.

Result: Through to be responsible for **processed pseudogenes**, by reverse transcribing mRNA that has its introns already removed.

Result: Existence of **SINEs**.

SINE elements

SINEs are a pastiche of

- **tRNA** or other small RNA gene, contributing (pol III) promotor
- **LINE 3' end**, contributing binding site for LINE ORF2

Small (80-300 bp) and successful. Needs corresponding LINE to live:

- **tRNA promotor** ensures SINEs are efficiently transcribed.
- **ORF2 binding site** makes SINE RNA compete with LINE RNA. Makes the LINE proteins copy SINEs instead.

Examples:

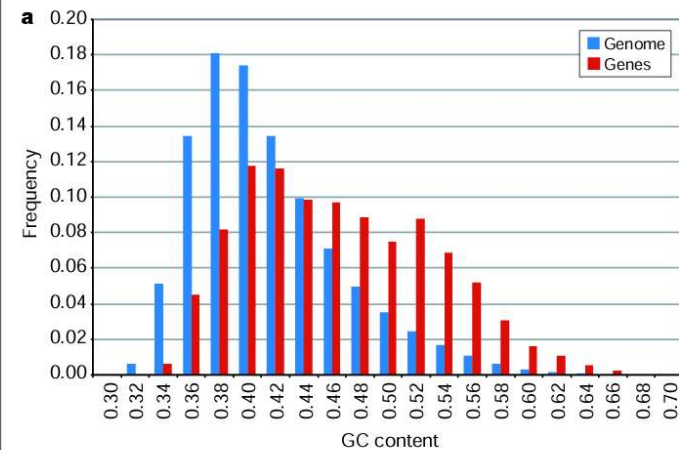
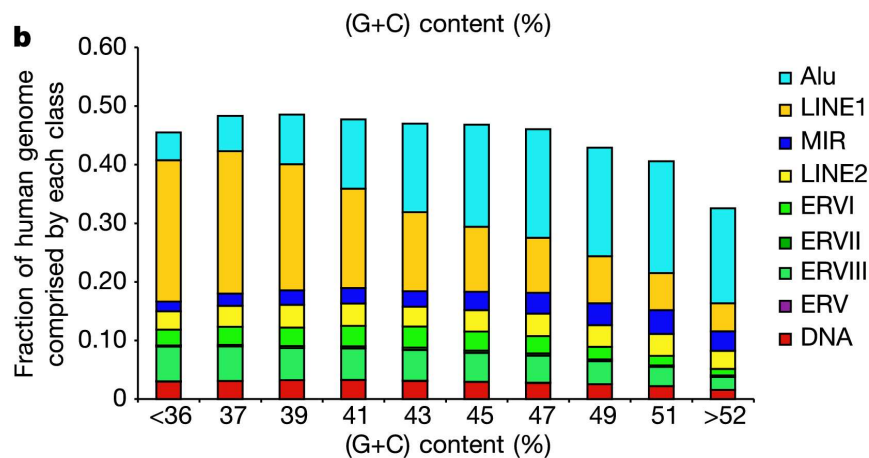
- human/primate **Alu** (282 bp), and rodent **B1** SINES derive from the 7SL signal-recognition particle (SRP) RNA, and uses LINE1 machinery.
- **MIR2** (Mammalian-wide interspersed repeat) most similar to 50 bp of reptilian LINE-like element.

LINE/SINE genomic distribution

Since **Alu** and **LINE1** use same machinery, one expects same distribution over genome. LINE1 is found **more often in A+T rich regions**. This is reasonable:

- Target site preference for **AAAA|TT** may explain bias
- Selection pressure: A+T rich regions are gene-poor, less deleterious.

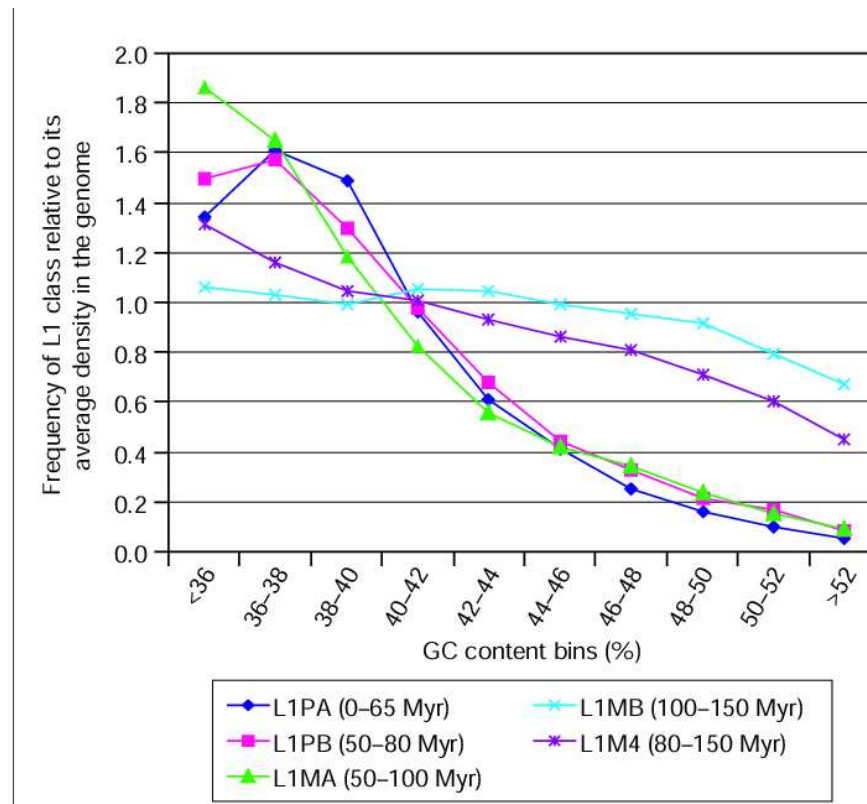
But: **Alu** (and mouse **B1**) **accumulate in G+C rich regions!**



Left: **Alu**, **LINE1** (and other transposable elts) distribution as function of C+G content.

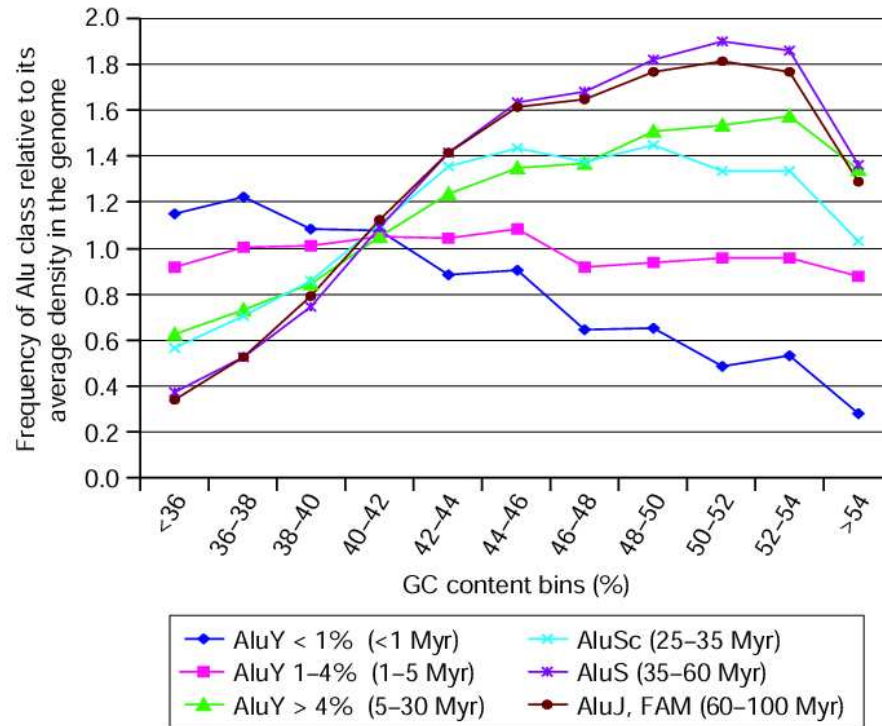
Right: C+G histogram, and gene density per C+G content bin.

LINE distribution in genome, over time



Young LINE1s have clear preference for A+T rich DNA.
Older LINE1s show a flatter distribution as function of G+C content.

SINE distribution in genome, over time



Young Alus also have preference for A+T rich DNA, but less pronounced.

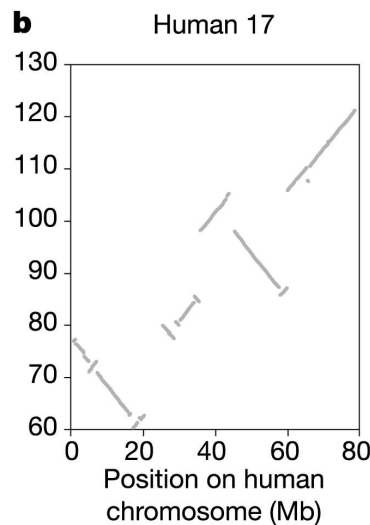
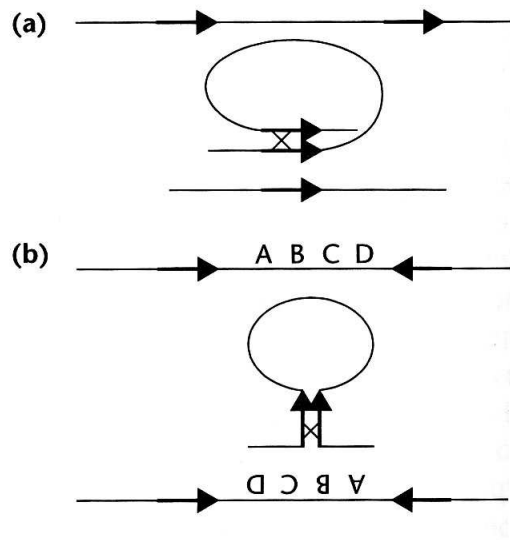
Older Alus very clearly **accumulate in G+C rich** (and gene-rich) DNA.

The different distributions of Alu and LINE1 in the genome suggests that **selection pressure** may be involved. **Biological function?**

Transposable elements: Effect on genome

High copy number of transposable elements provide many opportunities for **unequal homologous recombination**.

When this happens **within** a chromosome, leads to **deletions** or **inversions**.



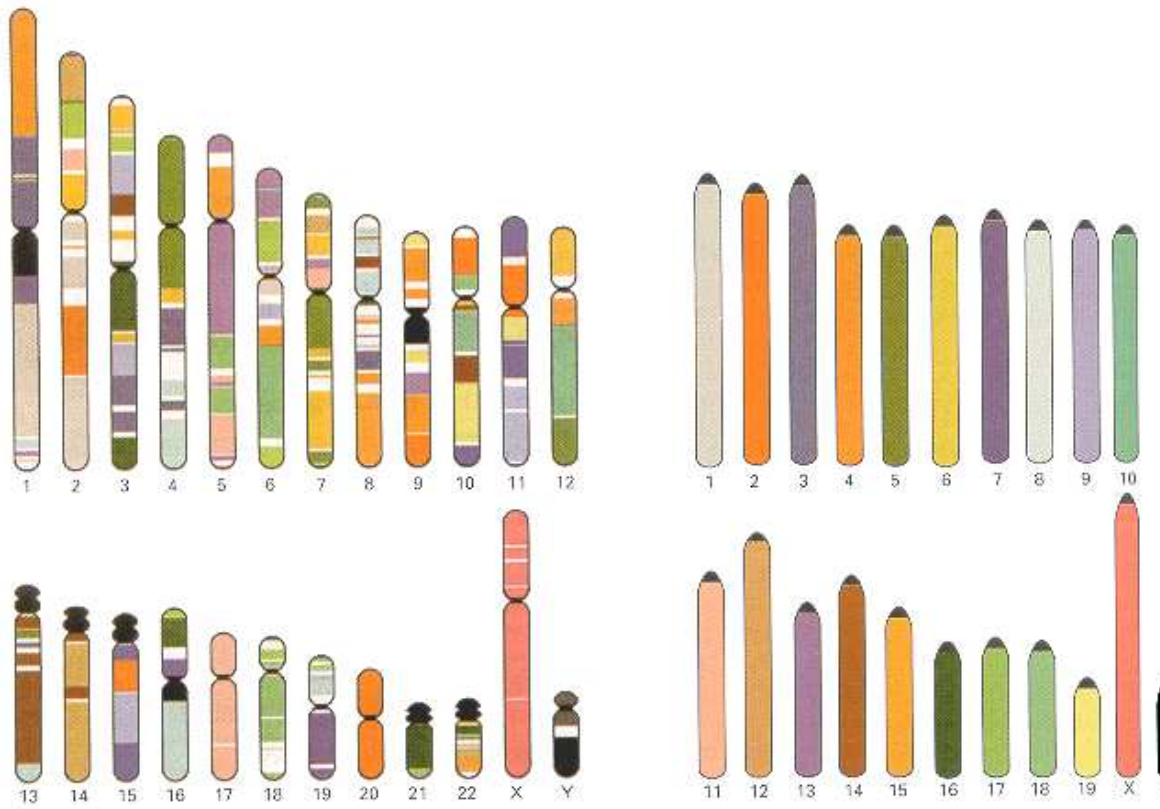
EHG T 622; Nature Feb. 2001

Direct evidence:

- **Existence of solo-LTRs**, result of recombination between two LTRs flanking one (or two) LTR-retrotransposon(s). EHG T 622
- **20% of Alus have no flanking target-site repeats.** CW Schmid, Nuc Acids Res 1998 26(20) 4541

Transposable elements: Effect on genome

When unequal homologous recombination occurs *between* chromosomes, **chromosome rearrangements** occur.



From: Alberts et al., The Cell,
after Nature vol 420, 5 Dec. 2002

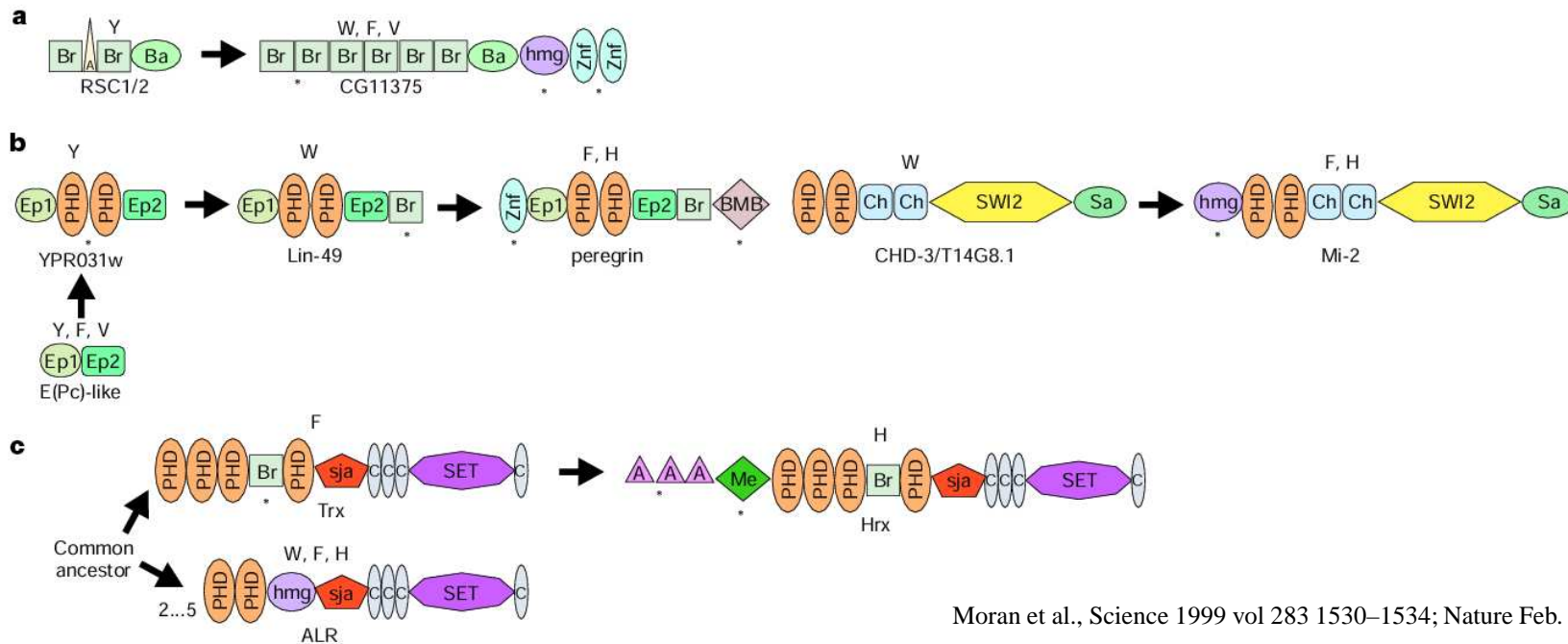
(**Right:** mouse chromosomes. **Left:** human chromosomes, colored according to which mouse chromosome region correspond to)

Transposable elements: Effect on genome

LINE elements may **duplicate** parts of genome. May change genome by:

- Moving **promoters and enhancers** to existing genes, changing expression patterns
- Moving **exons** into existing genes, adding new protein domains

Latter is called **exon shuffling** or **domain accretion**, major mode of protein evolution:



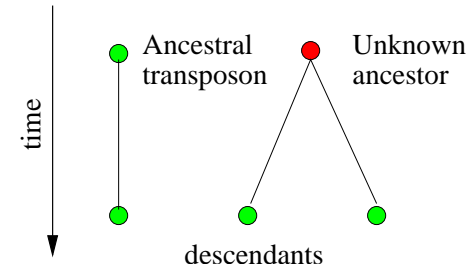
(Y=yeast, V=vertebrate, W=worm, F=fly, H=human.)

Domains are added (indicated by *) mostly at **either end** of protein.

Transposable elements: Looking into the past

Because of their high copy number, the **ancestors** of transposable elements can be recovered.

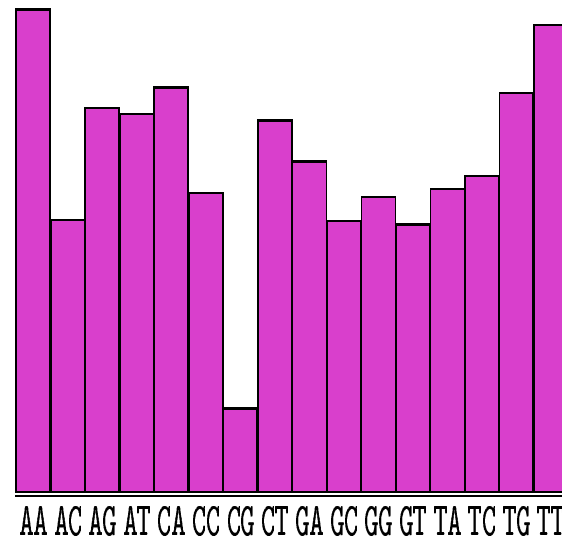
- Useful to measure ‘**forward**’ evolution, instead of ‘**forward+backward**’ in the usual case of comparing two descendant sequences. Gives a direction to time:



- **Indel process**: Small **deletions more frequent** than small insertions.
- **Root placement**: about **twice** as many substitutions occurred on mouse lineage compared to human lineage, since human-mouse split.
- Knowledge of ancestors gives method for measuring **transposon activity over time** (see before).
- Age of many individual TEs can be estimated, and so gives a way to see if **mutation process changed over time**.

3. CpG methylation

Relative frequencies of various **dinucleotides** in human DNA:



(This is for noncoding DNA on human chromosome 21; qualitatively same result holds for coding DNA, and on other chromosomes)

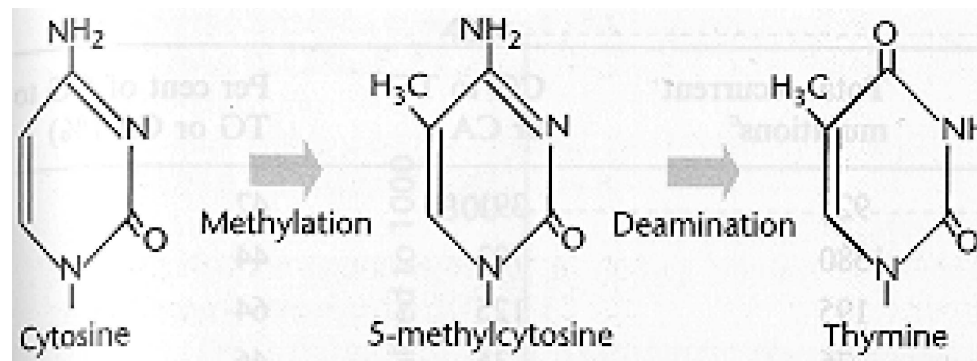
Clearly, there is a **deficiency of CG pairs**, by about a factor 5.

Notation: To distinguish from C-G pairing bases, these CG pairs along the sequence are denoted by **CpG**, where p = **phosphate** connecting C and G nucleotides.

CpG methylation

Explanation:

- Accidental deamination of **cytosine** results in **uracil**, recognized and repaired.
- Cytosine in **CpG** pairs are often *methyated*.
- Deamination of **methyl-cytosine** results in **thymine**, illegal pairing of two legal nucleotides. Unclear whether T or pairing G is wrong, sometimes incorrectly repaired.



Result:

EHG D 120

- Rate of **CpG** → **TpG** mutation increased (factor 10-20)
- **CpG** → **CpA** rate also increased (same factor), due to process on **reverse strand**.
(Process is strand-symmetric, because **CpG** is a **palindrome**: identical to its reverse complement)

In **neutrally evolving DNA**: $\approx 10\%$ of mutations occur on **1% CpGs**!

CpG islands

- Human genome:
 - C+G content ranges from 0.34–0.55 (average 0.42),
 - Expected CpG frequency about $0.21 \times 0.21 = 0.044$
 - Actual CpG frequency about factor 5 lower than expected ($\approx 1\%$)
- Special regions (“CpG islands”)
 - Higher C+G content (≈ 0.65),
 - Actual CpG frequency about as expected ($0.32 \times 0.32 \approx 10\%$).
 - ≈ 1000 bp long. Constitute about 1% of genome.

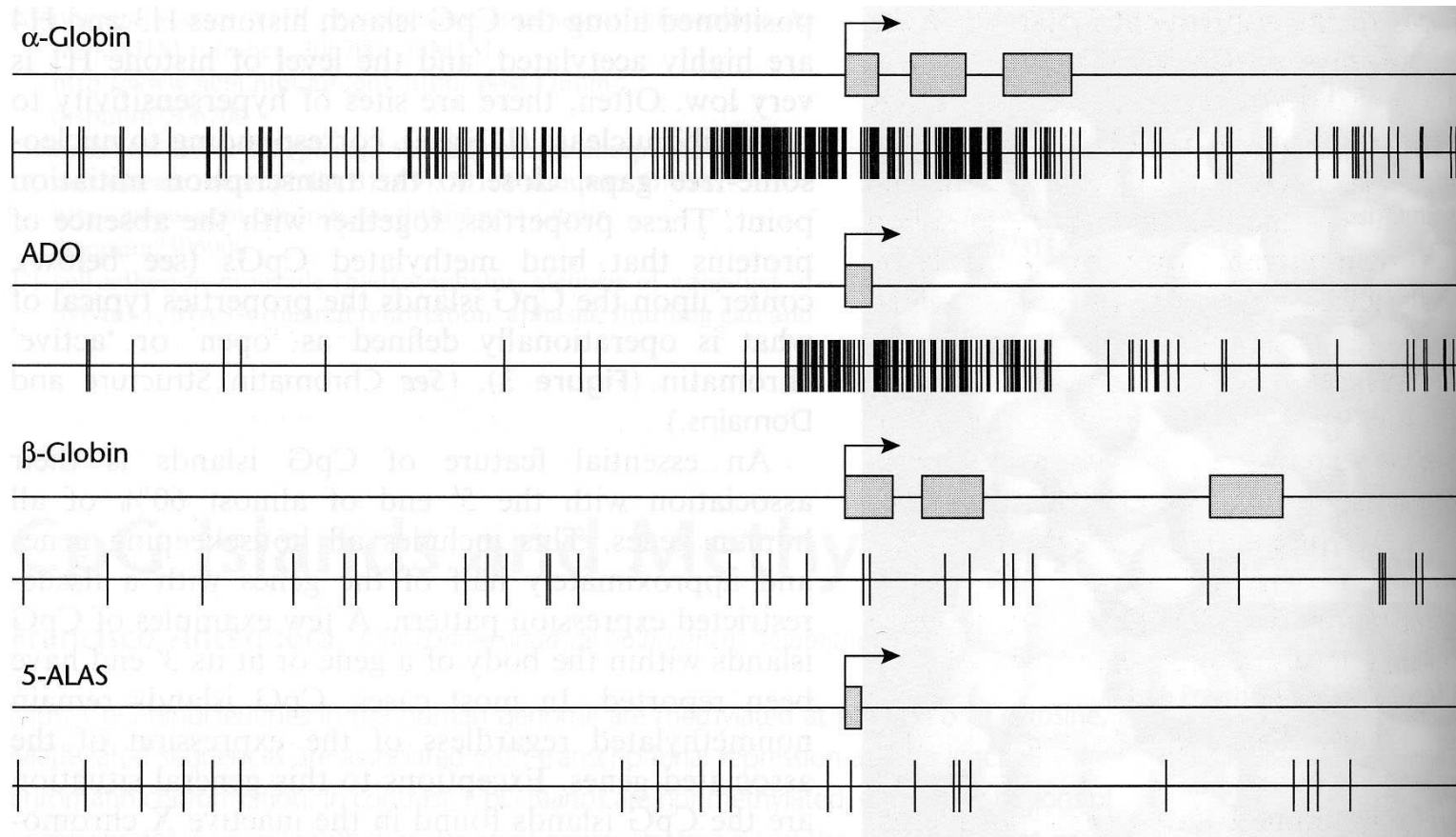
Explanation: CpG islands are unmethylated throughout development. No mutation pressure on CpGs, also leading to higher C+G levels.

CpG islands are found in promotor region of all housekeeping genes, and in about half of the genes with tissue-specific expression pattern.

Suggests that methylation is related to gene expression

CpG islands - examples

Some examples of CpG islands:



Vertical lines: CpG dinucleotide; **grey boxes:** exons, **arrow:** translation start site.

Function of CpG methylation

Methylation of DNA **suppresses transcription**. This is used in various ways:

- **Genetic imprinting**

Of about 45 mouse genes, either paternal or maternal gene is methylated and not expressed.

- **X silencing**

Females have two X chromosomes; to correct for double dosage, one X is methylated.

- **Protection against transposons?**

Most transposons are methylated and not expressed. However in some groups of organisms methylation and amount of transposons in genome seem unrelated.

EHG C 960; Simmen et al., Science 1999 vol 283 1164; Walsh et al., Nat Gen 1998 vol 20 116

- **Gene regulation / cell specialization?**

There seems to be no consensus whether methylation is involved in this.

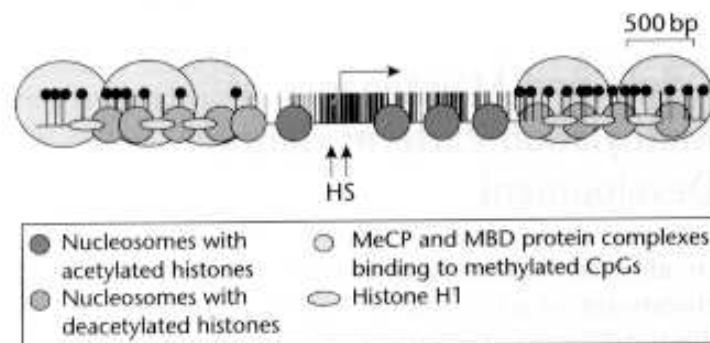
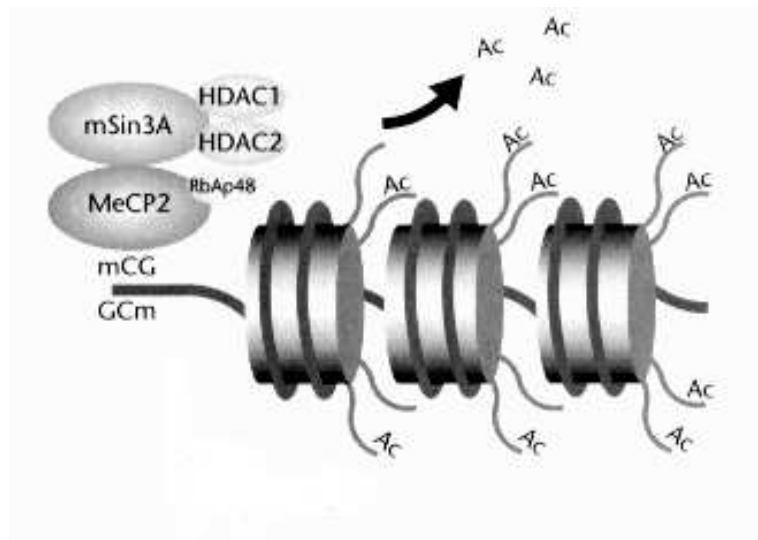
Yes: Regev et al., Mol Biol Evol 15(7)880, 1998; EHG D 115; **No:** EHG C 959

Methylation and gene silencing - mechanism

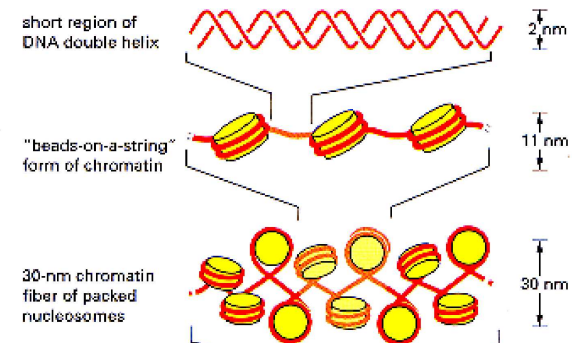
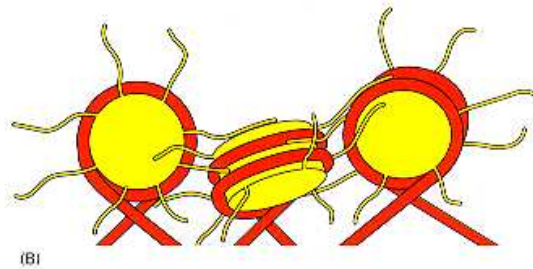
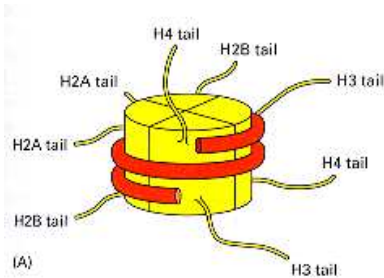
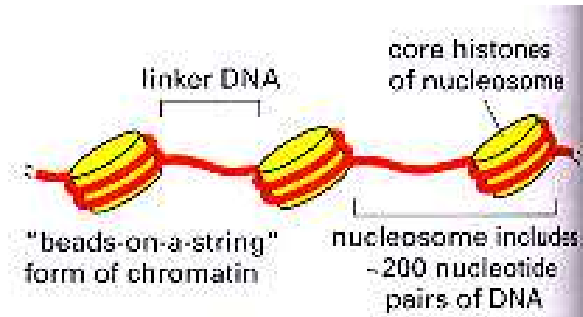
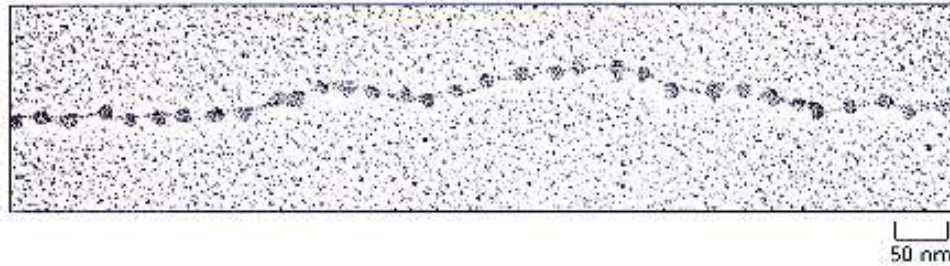
Proposed mechanism for gene silencing by methylation:

- Methyl-CpG binding protein (MeCP2) binds to DNA
- Forms complex with histone de-acetylase (HDAC1,2)
- Acetyl groups get removed from histone tails
- De-acetylated histones form tighter chromatin structure, preventing transcription

Summary: Methylation → change in chromatin packing → suppression of transcription



Methylation and gene silencing - pictures

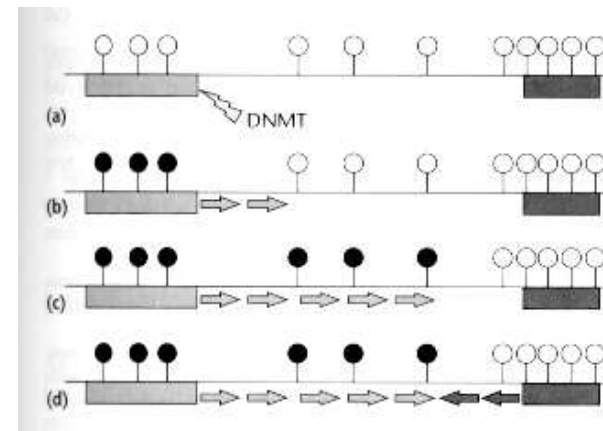


How does methylation pattern come about?

- Normally, methylation patterns are **copied when cell divides**
- Sperm and egg cells have (almost) normal methylation patterns
- After fertilization:
 - Parental genome is (almost) **demethylated** within 4 hours
 - Maternal genome more slowly (passively?) **demethylated**
 - From 120-cell stage, DNA is slowly **re-methylated** until birth

Suggested mechanism for formation of methylation pattern:

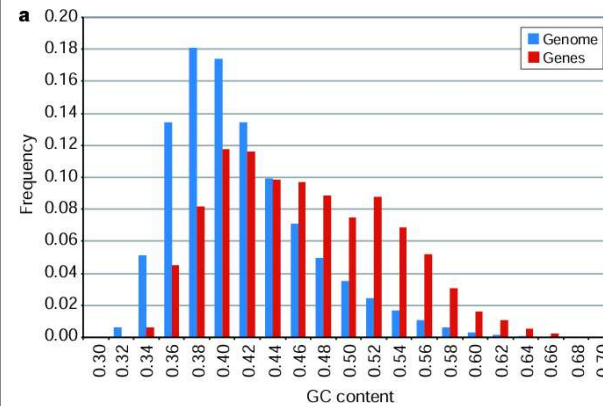
- Methylation starts at **specific target sites**
(See picture, DNMT = DNA methyltransferase)
- Spreads slowly, over many cell divisions.
Is actively stopped at promotor regions.



In mice, the SINE **B1** is **methylation target**.

Methylation and C+G content variation?

- Overall C+G content varies significantly in genome (0.32–0.60)



- C+G content in CpG islands is 0.65 on average

Could C+G content variation be result of variation of methylation levels?

- C+G variation positively correlated to:
 - Gene density
 - Transposon density
- Alus have high C+G content (63%)
- Alus occupy 10% of genome, on average

Could C+G content variation be direct result of Alu accumulation in gene-rich DNA?

4. Ongoing effects

The two mechanisms discussed before continue to have an effect on our genome. In particular, they cause **disease**.

Diseases caused by transposable elements (TEs):

- 0.1% of human disease is caused by Alu insertions
- 0.1% of human disease is caused by L1 insertions
- 0.3% caused by unequal homologous recombination of TEs

Currently **11 cases** of human L1 disease-causing insertion known.

Most insertions occurred **before fertilization** or early in embryogenesis.

One cancer-causing insertion was present in cancer tissue, not in healthy tissue.

See www.med.upenn.edu/genetics/labs/kazazian/human.html for list of diseases caused by retrotransposon insertions.

Ongoing effects - transposable elements

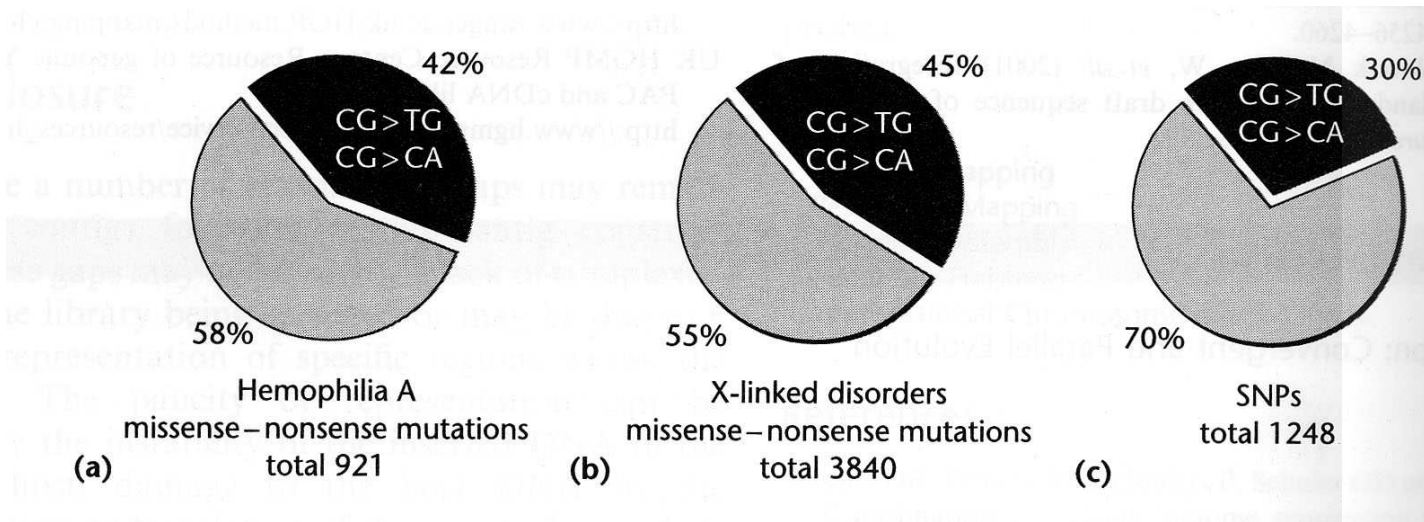
In mice, transposable elements are much more active: **10%** of spontaneous mutations causing a noticable effect are due to transposable elements.



Say cheese: just one piece of 'junk' DNA can produce several coat colours in genetically identical mice.

Ongoing effects - CpG methylation

About 30 – 50% of disease-causing mutations occur in CpG dinucleotides.



Some numbers:

EHG C 950

- 42% of acid-changing mutations in F8 gene (hemophilia A) are CpG-related.
- 30% of a set of 1248 single-nucleotide polymorphisms are CpG-related.
- Mutation rates (per site, per generation, in humans):
 - CpG \rightarrow TpG and CpG \rightarrow CpA: $\approx 9.7 \cdot 10^{-8}$
 - Other mutations (either at CpG sites or not): $\approx 6.5 \cdot 10^{-9}$

CpG-related mutations and disease

Eleven genetic disorders, and proportion of spontaneous amino-acid-changing mutations related to methylated CpG dinucleotides.

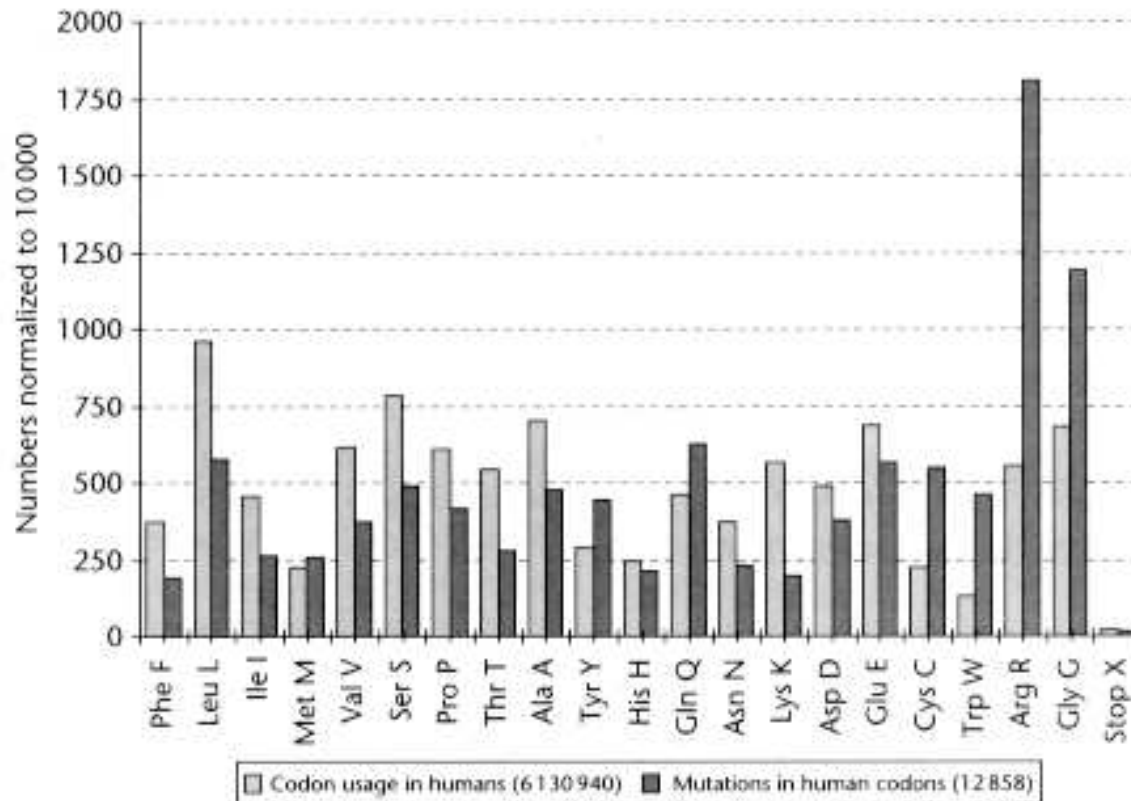
Table 1 X-linked disorders and mutations in CG dinucleotides

Disorder ^a	Gene	OMIM no. ^b	Total recurrent mutations ^c	CG to TG or CA	Per cent of CG to TG or CA (%)
Hemophilia A	<i>F8</i>	306700	921	391	42
Hemophilia B	<i>F9</i>	306900	1580	703	44
Rett syndrome	<i>MECP2</i>	300005	195	125	64
OTC deficiency	<i>OTC</i>	311250	76	35	46
Adrenoleukodystrophy	<i>ABCD1</i>	300100	235	146	62
X-linked agammaglobulinemia	<i>BTK</i>	300300	309	93	30
Lowe syndrome	<i>OCRL</i>	309000	53	28	53
Pyruvate dehydrogenase deficiency	<i>PDHA1</i>	312170	50	25	50
X-linked myotubular myopathy	<i>MTM1</i>	310400	95	49	51
Androgen receptor insensitivity	<i>AR</i>	313700	250	85	34
X-linked hypophosphatemia	<i>PHEX</i>	307800	76	32	42
Total			3840	1712	44

CpG-related mutations are responsible for 40 – 50% of genetic disorders.

Ongoing effects - CpG methylation

Arginine and glycine mutate more often than other amino acids.



Explanation: Codons for Arg are: AGA, AGG, CGU, CGC, CGA, CGG.

Explanation for hypermutability of Gly less clear.

5. Is Alu functional?

Our genome: >1M copies of Alu, occupying >10%.

Does not need to imply function. Abundance may be due to successful copying strategy (truly selfish genes).

But, Alu is often clearly deleterious (generally mutagenic, and involved in specific diseases), so a symbiotic role would help to explain abundance.

Distribution of Alu in genome is puzzling (more in gene-rich regions with high G+C content, in contrast to L1) and suggests positive selection pressure → biological function.

More information:

AFA Smit, Interspersed repeats and other mementos of transposable elements in mammalian genomes, Curr Opin Gen Dev 1999, 9:657–663.

W.M. Chu et al., Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR, Mol. Cell. Biol., Jan. 1998, 58–68.

C.W. Schmid, Does SINE evolution preclude Alu function?, Nuc. Acids Res. 1998 26(20), 4541–4550.

Alu function

Cell stress

- **Cell stress** (heat, viral infection, toxins) rapidly increases amount of Alu RNA transcripts
- Alu RNAs **bind and inhibit** an enzyme (PKR, protein kinase)
- This (after additional steps) increases protein translation

Cell stress increases SINE transcription in **insects** as well.

AFA Smit, Curr Opin Gen Dev 1999 9:657

Role in CpG methylation / nucleotide organisation?

Observations:

- Mouse B1 SINE is **target** for methylation in embryonic cells
- Specific type of young Alu is **undermethylated** in human sperm

Do **Alus** direct **methylation** and thereby DNA packing, C+G and CpG content?

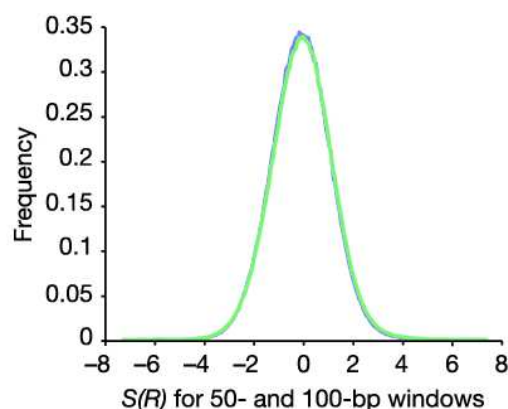
EHG D 135; EHG D 114; EHG C 956

6. Amount of DNA under selection

Align human/mouse genomes. If no selection, we expect to see certain average number of mutations in a window. Summarize this by **conservation score**:

- μ = fraction of **identical** sites in **large window** (μ depends on local C+G content)
- p = fraction of **identical** sites in window R of size $n = 50, 100$
- $S = S(R) = \frac{p - \mu}{\sqrt{\mu(1 - \mu)/n}}$

If mutations are independent and identically distributed, then S has (approximately) **normal** distribution with mean 0 and standard deviation 1.



Hypothesis: Ancestral repeats are **neutrally evolving**, and **representative** of n.e. DNA.

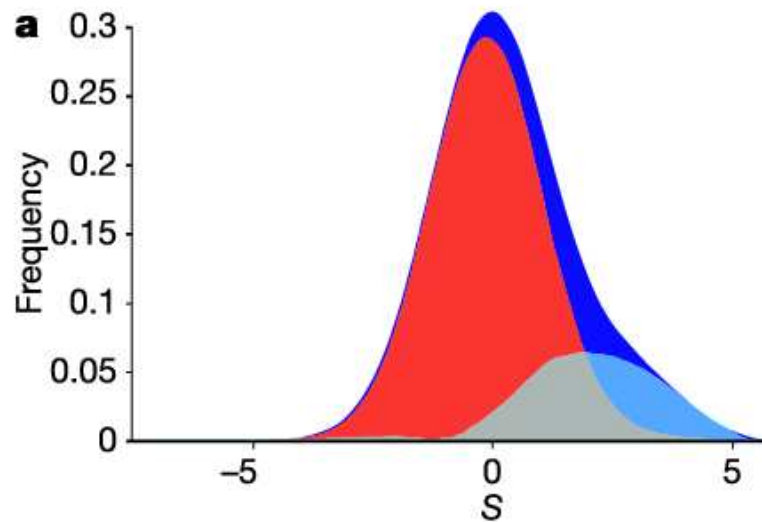
Left: Distribution of S , on 165Mb of ancestral repeats; window sizes 50 and 100.

Indeed bell-shaped, but **larger standard deviation** than normal distribution: $\sigma = 1.19$ for $n = 50$; $\sigma = 1.23$ for $n = 100$.

Amount of DNA under selection

Now we know distribution of $S(R)$ for DNA **not under selection**.

Distribution of $S(R)$ on **all** DNA gives us **proportion of DNA under selection**:



Nature vol. 420, 5 dec. 2002, 520–562

- Compute $S(R)$ for windows R covering entire (alignable, 40%) genome **(blue)**.
(It has a bulge at large $S \Rightarrow$ DNA under purifying selection.)
- Subtract shape of distribution of neutrally evolving DNA **(red)**.
- What remains is the proportion under selection (grey).

This is **5%** of total genome. Coding regions = **1.5%**.

What is remaining 3.5%?

“Genomic dark matter”

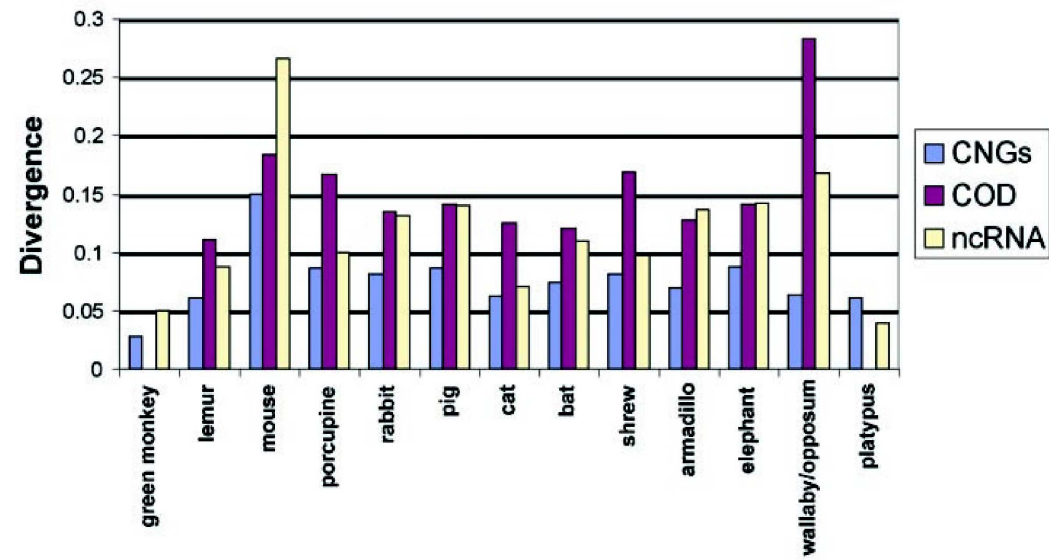
Is there a **substantial amount** of **conserved DNA** with **unknown function** in our genome?

The answer seems to be **Yes:**

By direct comparison of human chromosome 21 DNA against 13 other mammals:

(mouse, monkey, lemur, porcupine, rabbit, pig, cat, bat, shrew, armadillo, elephant, wallaby, platypus)

- **220 “Conserved Non-Genic”** (CNG) sequences of unknown function were found.
- These CNGs are **significantly more conserved** than protein-coding DNA.
- By extrapolation, CNGs **cover 0.3 to 1.0 %** of our genome.



Dermitzakis et al., Mol Biol. Evol. 19, 1114 (2002); Dermitzakis et al., Science vol 302 1033, (7 nov 2003)

Example of CNG sequence

CNG

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

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

COD

Human	GCTCAGTCAC	TOCAGAATCC	CTGAGAAAAG	CAATAGAGGC	TGTATCACCG
LemurG..	T.....A.....
MouseATGG.....	CA.T.....
Porcupine	..T...C..A.AT.T..A
Rabbit	..T.....A.TG.....A.....A.....
Pig	..T.T.....A.TA.....	CA.....A
CatA..	..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—GC....
Elephant	...A...C.	T.....	T.....T.....C...G
Opossum	...AG...AC.	TG.C.....	GAG.C.T—G	..C.G.C.T..

Is 5% of genome really under selection? Some thoughts

Are ancient repeats representative for neutrally evolving DNA?

Facts:

1. Few main classes (e.g. **Alu/B1** and **L1**) account for large proportion of ancient repeats.
 2. Alus (especially young Alus) have high **C+G** and **CpG** content: **9×** more CpGs than in human DNA, and **33%** of all genomic CpGs are in Alus.
 3. LINEs have low **C+G** content.
-

- (1:) Nucleotide distribution (and therefore substitution rate) of ancient repeats (in the past!) may **differ** from neutral DNA,
- (1+2+3:) Mutation rate for Alus is **higher** than average. Mutation rate on LINEs is lower than on Alus. This may explain observed **lack of mutational independence** ($\sigma \approx 1.2$), since type of repeat is clearly not independent along sequence.
- (1+2:) High mutation rate on Alus suggests that estimated average number of conserved sites for neutrally evolving DNA (μ) is **underestimate**. This would make conservation score S on entire genome come out high, and make genome look more conserved than it is.

Conclusions

- **Transposable elements**, or “selfish genes”, are probably responsible for over 50% of our genome.
- Not all are purely selfish though: **Alu may be a symbiont**.
Probably involved in **cell stress** and **DNA methylation / chromatin structure**.
- Methylation of **CpG dinucleotides** accounts for many mutations. Incorporating nearest-neighbour effects will be substantial improvement for DNA substitution models.
- Somewhere between 0.3% and 3.5% of our genome is **evolutionary conserved**, up to 1.0% even highly conserved, and has **unknown function**.

Recommended literature

- *Initial sequencing and comparative analysis of the mouse genome*, Nature vol. 420, Dec. 2002, 520–562.
- *Initial sequencing and analysis of the human genome*, Nature vol. 409, Feb. 2001, 860– 921
- *The origin of interspersed repeats in the human genome*, AFA Smit, Curr Opin Gen Dev 1996 6:743–748, and
Interspersed repeats and other mementos of transposable elements in mammalian genomes, AFA Smit, Curr Opin Gen Dev 1999 9:657–663.
Overview of transposable elements, lots of references.
- *Does SINE evolution preclude Alu function*, CW Schmid, Nuc Acids Res 1998 26(20) 4541–4550.
What Alus do for us.
- *Nature Encyclopedia of the Human Genome*. (Referred to as EHG in these slides.)
Five volumes with short up-to-date articles by experts.