Characterization of the human genome: noncoding DNA, transposable elements, Hox genes, chromosome rearrangements, and gene families

Biosciences 741: Genomics

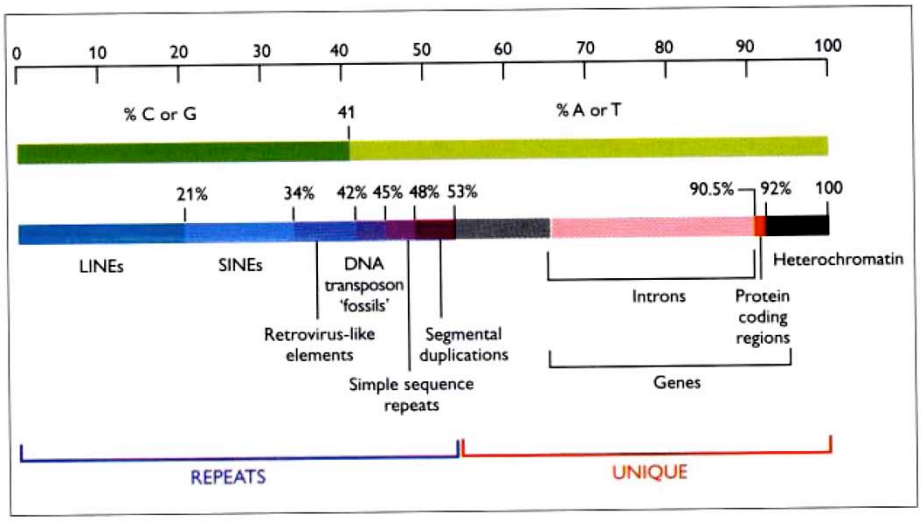
Fall, 2013

Week 10

Outline

- Noncoding DNA – human vs. mouse
- Transposon structure and function
- Transposon distributions (age, GC content, species)
- Hox gene clusters vs. transposons
- Chromosomal rearrangements

Content of the human genome



The functions of noncoding DNA

- The putative functions of noncoding DNA (i.e., DNA that does not encode a protein) include the control of gene expression, chromosome segregation, and recombination rates (~not alternative splicing).
- This sort of "functional noncoding DNA" can encode DNA binding sites (for protein ligands), or can encode RNA gene products (*Xist*, snRNPs, small interfering RNAs, *etc*).
- Although examples of all of these are known, their prevalence across the genome is difficult to estimate.
- The best available estimates come from comparisons with the mouse and rat genomes, which indicate that ~5% of the mammalian genome is conserved, of which ~1.5% codes for protein. The remaining ~3.5% may consist of other functional sequences such as enhancers, silencers, matrix attachment sites, noncoding RNAs, and so on.
- Why use the mouse for this comparison?
- What is the rest of the DNA doing?

Mouse genome sequencing strategy

- Produced a physical map of the mouse genome by fingerprinting and sequencing the ends of clones from a BAC library.
- Whole-genome shotgun sequencing to approximately seven-fold coverage to efficiently generate a sequence database with small contigs (this limited their ability to find segmental duplications - why?)
- Hierarchical shotgun sequencing of BAC clones *(i.e.,* sequencing selected BAC clones, one at a time, by the shotgun method) to build larger contigs, refine the sequence, and confirm the framework assembly.
- Production of a finished sequence by using selected BAC clones directly as templates for directed finishing (*i.e.,* with specific oligo primers).
- The Y chromosome was sequenced by a purely hierarchical strategy because of its relatively high content of repetitive DNA.
- The sequence was oriented and assigned to chromosome by reference to the pre-existing classical genetic map of the mouse genome.

Mouse genome summary

- Mouse genome is ~14% smaller than human genome, probably because of a higher deletion rate.
- More than 90% of both mouse and human genomes consist of regions of generally conserved synteny.
- At the nucleotide level, about 40% of the mouse genome can be aligned with the human genome, with the rest likely to have been INSERTED or deleted in one genome or the other.
- The neutral substitution rate has been roughly 50% since the last common ancestor of mice and humans (which was about 75 ± 10 million years ago).
- The proportion of small (50-100 bp) segments that are under purifying selection is about 5%, of which 1.5% are protein-coding sequences.
- Certain classes of secreted proteins involved in reproduction, host defense and immune response seem to be under positive selection in mammals.

Methods used to detect natural selection in the mouse genome

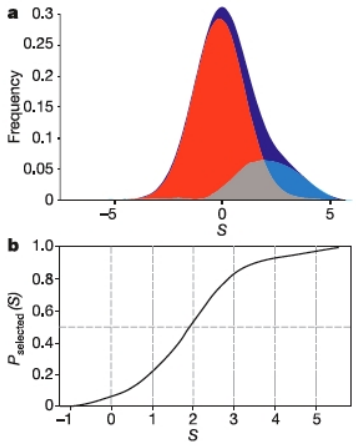


Figure 28 Proportion of the human genome under selection and the probability of a genomic window to be under selection on the basis of conservation score. **a**, The genome-wide density of conservation scores, S_{genome} (dark blue), was decomposed into a mixture of two component densities: $S_{neutral}$ (red) and $S_{selected}$ (light blue and grey). S_{genome} is derived from the conservation scores $S(R)$ for all windows of 50 bp in the human genome with at least 45 bases aligning to mouse. $S_{neutral}$ is a scaled version of the $S_{neutral}$ density from the blue curve in Fig. 23 for the 50-bp windows in ancestral repeats, representing neutrally evolving DNA. $S_{selected}$ is the difference between the blue density and the red component, and thus represents a scaled version of $S_{selected}$, the predicted density for conservation scores of 50-bp windows in the human genome that are evolving under selection. The scaling factors are the estimated mixture coefficients, which are $p_0 = 0.792$ for $S_{neutral}$, and $1 - p_0 = 0.208$ for $S_{selected}$. The coefficient p_0 is calculated as the minimum of the ratio between $S_{genome}(S)$ and $S_{neutral}(S)$ for all values of S, giving a conservative estimate that maximizes the share of the mixture attributed to $S_{neutral}$. **b**, The probability, $P_{selected}(S)$, that a 50-bp window is under selection as a function of its conservation score $S = S(R)$. This function is derived from the mixture decomposition by setting $P_{selected}(S) = 1 - p_0 S_{neutral}(S)/S_{genome}(S)$.

The *C*-value paradox

- The "*C*-value paradox" originally referred to the paradoxical observation that eukaryotes vary by ~100,000-fold in their nuclear DNA content (*C*-value), but this appears to be uncorrelated with their apparent biological complexity, and much more variable than prokaryotes, who vary only ~10-fold in their genome sizes.
- Insects and amphibians of apparently similar complexity vary by ~100-fold in genome size. The world's record largest genome belongs to a protozoan.
- In the genomics era, the C-value paradox grew to encompass related questions - why are eukaryotic genomes so much larger than needed to encode their proteomes? And why are bacterial genomes generally quite close to the size needed to encode their proteomes?
- One view is that the ideal genome size is determined by correlations between genome size and cell size, cellular growth rates, and organismal generation times. The best evidence in favor of this is that intracellular parasites' genomes shrink noticeably and their cells get smaller (but many other explanations of this phenomena are also possible).

The C-value paradox (continued)

- An alternative view is that genomes grow to variable extents because of the continual insertion of transposable elements.
- This is supported by a fairly strong correlation between genome size and the percent of the genome encoded by transposable elements.
- It also helps to explain the fairly strong correlation between genome size and average intron size.
- According to this view, organisms with large genomes either happen to have active transposons, or are not successful in selecting against transposon insertions, or perhaps have lower spontaneous deletion rates used to "take out the trash".
- The most frequent classes of indels are strongly biased in all organisms, with deletions being more frequent than insertions.

Outline

- Noncoding DNA – human vs. mouse
- Transposon structure and function
- Transposon distributions (age, GC content, species)
- Hox gene clusters vs. transposons
- Chromosomal rearrangements

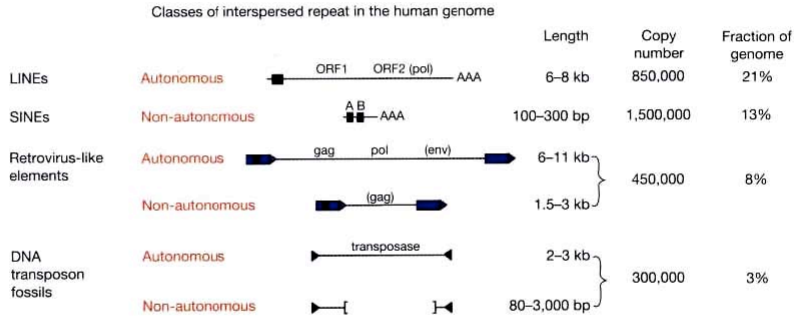


Figure 17 Almost all transposable elements in mammals fall into one of four classes. See text for details.

Table 11 Number of copies and fraction of genome for classes of interspersed repeat

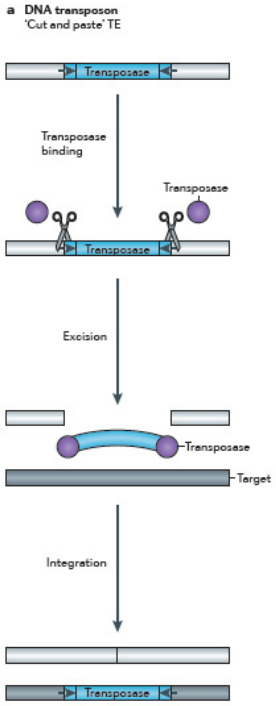
	Number of copies (× 1,000)	Total number of bases in the draft genome sequence (Mb)	Fraction of the draft genome sequence (%)	Number of families (subfamilies)
SINEs	1,558	359.6	13.14	
Alu	1,090	290.1	10.60	1 (~20)
MIR	393	60.1	2.20	1 (1)
MIR3	75	9.3	0.34	1 (1)
LINEs	868	558.8	20.42	
LINE1	516	462.1	16.89	1 (~55)
LINE2	315	88.2	3.22	1 (2)
LINE3	37	8.4	0.31	1 (2)
LTR elements	443	227.0	8.29	
ERV-class I	112	79.2	2.89	72 (132)
ERV(K)-class II	8	8.5	0.31	10 (20)
ERV (L)-class III	83	39.5	1.44	21 (42)
MaLR	240	99.8	3.65	1 (31)
DNA elements	294	77.6	2.84	
hAT group				
MER1-Charlie	182	38.1	1.39	25 (50)
Zaphod	13	4.3	0.16	4 (10)
Tc-1 group				
MER2-Tigger	57	28.0	1.02	12 (28)
Tc2	4	0.9	0.03	1 (5)
Mariner	14	2.6	0.10	4 (5)
PiggyBac-like	2	0.5	0.02	10 (20)
Unclassified	22	3.2	0.12	7 (7)
Unclassified	3	3.8	0.14	3 (4)
Total interspersed repeats		1,226.8	44.83	

The number of copies and base pair contributions of the major classes and subclasses of transposable elements in the human genome. Data extracted from a RepeatMasker analysis of the draft genome sequence (RepeatMasker version 09092000, sensitive settings, using RepBase Update 5.08). In calculating percentages, RepeatMasker excluded the runs of Ns linking the contigs in the draft genome sequence. In the last column, separate consensus sequences in the repeat databases are considered subfamilies, rather than families, when the sequences are closely related or related through intermediate subfamilies.

DNA transposons generally have inverted repeats,

require a single site-specific recombinase (transposase),

and produce direct (not inverted) target site duplications upon insertion.



Levin and Moran (2011)
Nat. Rev. Genet. 12, 615-627.

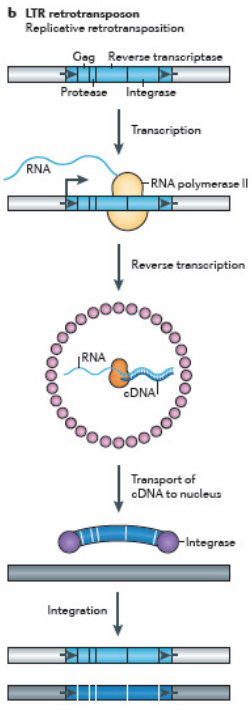
LTR retrotransposons form viral-like particles

Integrase produces a staggered cut in the host chromosome, rather like a restriction endonuclease.

Therefore, the length of the target site duplication is fixed (for any given LTR retrotransposon).

Daughter elements are generally full length, due to precise recognition of the LTR by integrase enzyme.

However, deleted copies (and single LTRs) can be produced by later genomic deletions (or unequal crossing over).



Levin and Moran (2011)
Nat. Rev. Genet. 12, 615-627.

LINE (non-LTR) retrotransposons

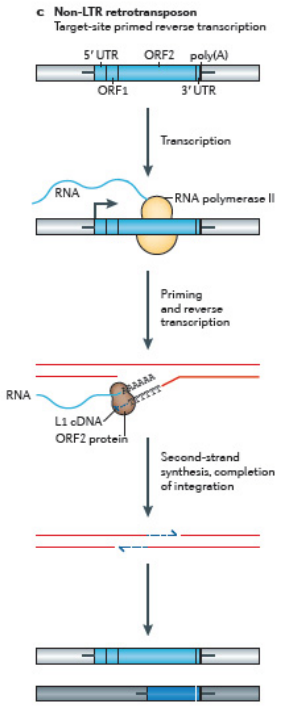
Promoter sequence in 5'UTR; transcript is polyadenylated.

ORF1 Is an RNA-binding protein.
An ORF1 trimer covers 50 bp of RNA
(Basame et al. 2006 J. Mol. Biol. 357, 351-357)

ORF2 is an endonuclease and reverse transcriptase.

Size of target site duplication is variable, because that depends on location of random nick in second chromosomal DNA strand.

Most daughter elements have 5' deletions, due to premature strand-switching by reverse transcriptase => w/o 5'UTR, they will be replication-defective.



Levin and Moran (2011)
Nat. Rev. Genet. 12, 615-627.

Summary of transposon properties

Table 1 | **Classes of transposable elements and their mobility mechanisms**

Class of TE	Structural features	Replication mechanism	Variant forms	Active examples
DNA transposons	• TIRs • Transposase	Transposase-mediated excision of donor dsDNA followed by insertion into the target site	• Some DNA transposons also mobilize via replicative mechanisms • ssDNA transposons lack TIRs: donor ssDNA is inserted into target-site ssDNA, such as for IS*608* of *Helicobacter pylori*	• Tn*7* in *Escherichia coli* • *P* elements in *Drosophila melanogaster* • *Tc1* elements in *Caenorhabditis elegans*
LTR retrotransposons	• LTRs • Gag, protease, reverse transcriptase and integrase	Within virus-like particles, reverse transcriptase copies the mRNA of the TE into a full-length cDNA; integrase inserts the cDNA into target sites	• Solo LTRs are commonly found in genomes and are a result of LTR–LTR recombination	• *Ty1*, *Ty3* and *Ty5* in *Saccharomyces cerevisiae* • *Tf1* and *Tf2* in *Schizosaccharomyces pombe* • *Tnt1* in tobacco
Non-LTR retrotransposons	• One or two ORFs • 5′ truncations and inversion/deletion (for mammalian L1 elements) • Some end in poly(A) tails (for example, L1s); others do not (for example, R2)	An element-encoded endonuclease mediates TPRT. The endonuclease nicks the DNA at the target site and uses the 3′ nicked end for the primer as it reverse transcribes TE mRNA	• Non-autonomous, non-LTR retrotransposons (for example, *Alu* and SVA elements, as well as other eukaryotic SINEs) rely on the endonuclease and reverse transcriptase of an autonomous non-LTR retrotransposon to mediate retrotransposition • The L1 retrotransposition machinery can also mobilize mRNAs (to generate processed pseudogenes) and certain non-coding RNAs (for example, the U6 snRNA)	• L1 in human, mouse, and other mammals • *I* factor in *D. melanogaster* • Zorro3 in *Candida albicans* • R1 and R2 in insects

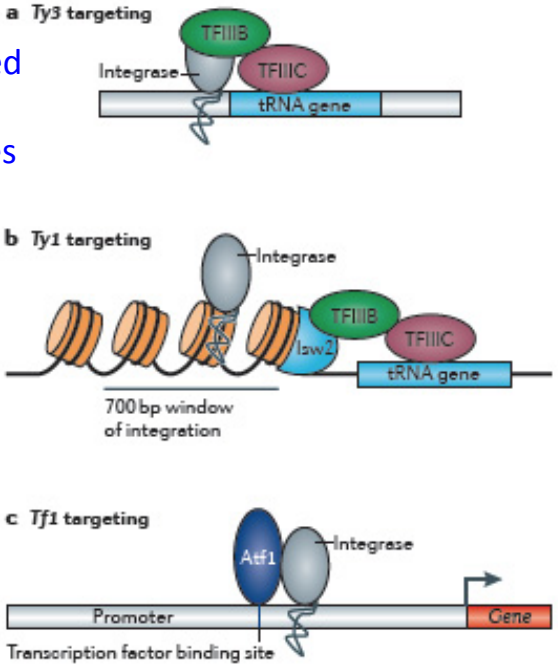
L1, long interspersed element 1; LTR, long terminal repeat; SINE, short interspersed element; snRNA, small nuclear RNA; SVA, SINE-R–VNTR–*Alu*; TE, transposable element; TIR, terminal inverted repeat; TPRT, target-site-primed reverse transcription.

Levin and Moran (2011)
Nat. Rev. Genet. 12, 615-627.

Some yeast retroposons are preferentially targeted near the promoters of particular classes of genes

This is also true of Drosophila P elements (DNA cut and paste transposon) and human HIV (retrovirus), Among others.

It is used for "enhancer trap" screens in mice and flies.

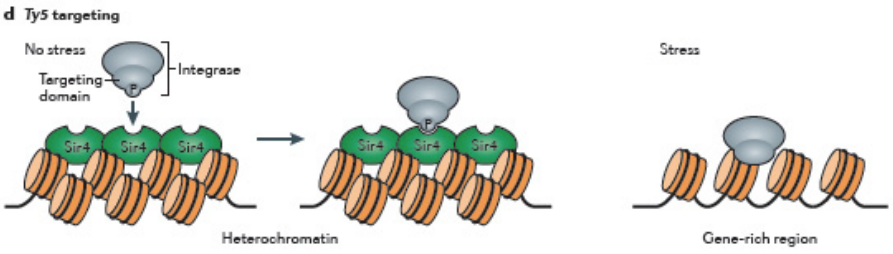


Levin and Moran (2011)
Nat. Rev. Genet. 12, 615-627.

Some retrotransposons are specifically targeted to heterochromatin

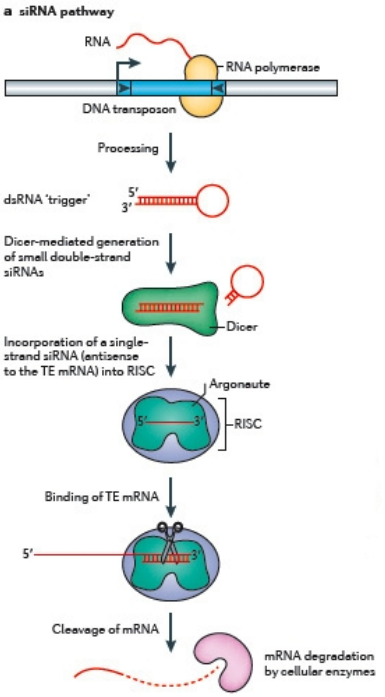
When cells are subjected to stress, the integrase enzyme is dephosphorylated, and integrase becomes able to direct integration in euchromatin.

Similar elements (non-LTR retrotransposons HeTA, TART, TAHRE) are responsible For maintaining the telomeres of Drosophila.



Levin and Moran (2011)
Nat. Rev. Genet. 12, 615-627.

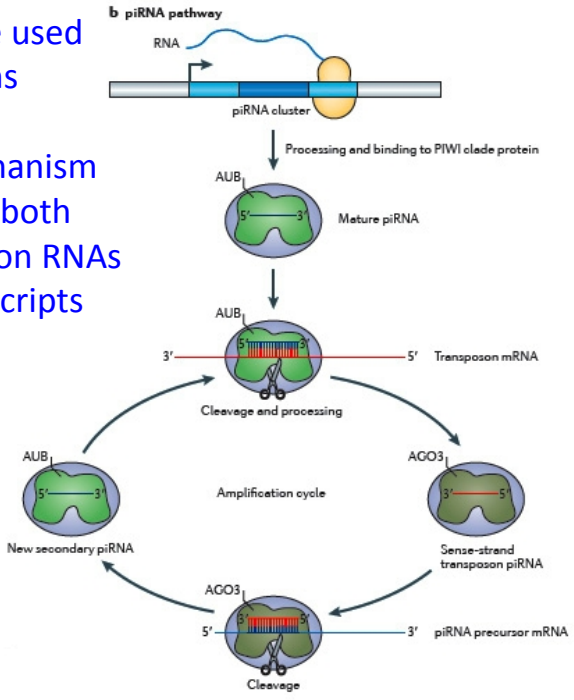
siRNA pathways are used to degrade retrotransposons RNAs



Levin and Moran (2011)
Nat. Rev. Genet. 12, 615-627.

PIWI RNAs (piRNA) are used to degrade transposons

by a "ping-pong" mechanism of mutual cleavage by both sense-strand transposon RNAs vs. piRNA cluster transcripts



Levin and Moran (2011)
Nat. Rev. Genet. 12, 615-627.

Transposition events occur at specific stages of the life cycle

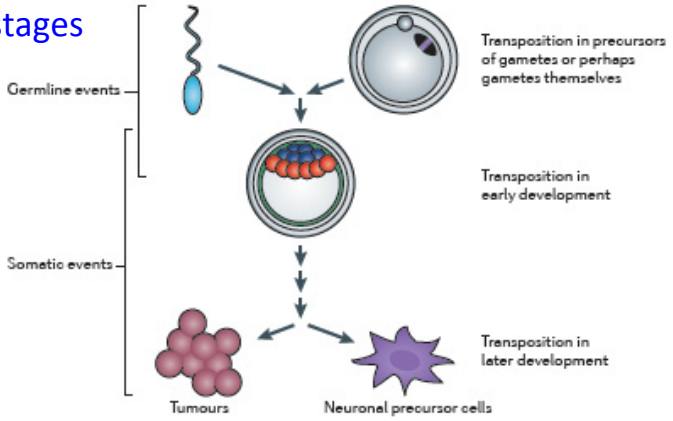
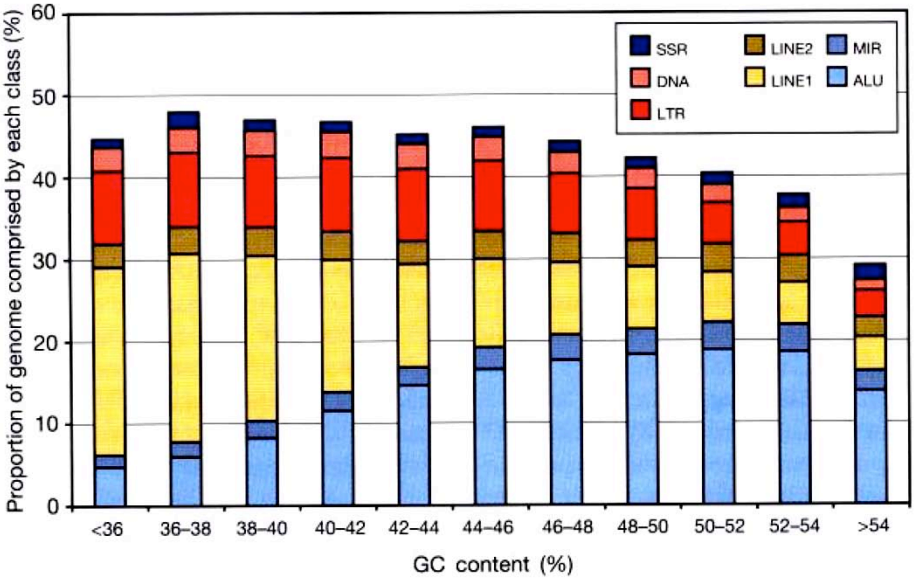


Figure 4 | **Timing of transposition.** Germline transposable element (TE) integration events can result from TE mobility in cells that give rise to gametes or from TE mobility post-fertilization during early development. Embryonic TE mobility in cells that do not contribute to the germ line or mobility at later developmental stages can, in principle, lead to somatic TE integration events. The overlapping brackets signify that some TE insertions in early development can contribute to the germ line, whereas others may not. Endogenous long interspersed element 1 (L1) retrotransposition events can occur in certain tumours, and experiments using engineered human L1s suggest that L1 retrotransposition may also occur during mammalian neurogenesis. Examples of the developmental timing of TE integration events are described in the main text.

Levin and Moran (2011) *Nat. Rev. Genet.* 12, 615-627.

Outline

- Noncoding DNA – human vs. mouse
- Transposon structure and function
- Transposon distributions (age, GC content, species)
- Hox gene clusters vs. transposons
- Chromosomal rearrangements



Mouse-specific and human-specific transposon sub-families were clearly identified

Table 5 **Composition of interspersed repeats in the mouse genome**

	Mouse				Human	
	Thousands of copies	Length occupied (Mb)	Fraction of genome (%)	Lineage specific (%)	Fraction of genome (%)	Lineage specific (%)
LINEs	660	475.3	19.20	16.46	20.99	7.94
LINE1	599	464.8	18.78	16.46	17.37	7.94
LINE2	53	9.4	0.38	–	3.30	–
L3/CR1	8	1.2	0.05	–	0.32	–
SINEs	1,498	202.9	8.22	7.63	13.64	10.74
B1 (Alu)	564	67.3	2.66	2.66	10.74	10.74
B2	348	59.6	2.39	2.39	–	–
B4/RSINE	391	57.1	2.36	2.36	–	–
ID	79	5.3	0.25	0.25	–	–
MIR/MIR3	115	14.1	0.57	–	2.90	–
LTR elements	631	244.3	9.87	8.72	8.55	4.09
ERV_classI	34	16.8	0.68	0.58	2.92	2.02
ERV_classII	127	79.1	3.14	3.14	0.30	0.30
ERV_classIII	37	14.0	0.58	0.32	1.55	0.19
MaLRs (III)	388	112.2	4.82	4.02	3.78	1.58
DNA elements	112	21.8	0.88	0.36	3.03	1.00
Charlie	82	15.2	0.62	0.35	1.41	0.14
Other hATs	8	1.6	0.06	–	0.31	–
Tigger	24	4.4	0.17	–	1.06	0.76
Mariner	1	0.2	0.01	0.01	0.10	0.07
Unclassified	26	9.2	0.38	0.37	0.15	0.14
Total	2,926	953.6	38.55	33.53	46.36	24.05
Small RNAs	19	1.5	0.06	0.04	0.04	0.02
Satellites	7	0.7	0.30	NA	0.34	NA
Simple repeats	960	56.1	2.27	NA	0.87	NA

The two right columns show the fractions in the human genome (excluding chromosome Y) for comparison. These and all other interspersed repeat-related data are based on RepeatMasker analysis (version July 2002, sensitive settings, RepBase release 5.3) of the February 2002 mouse and June 2002 human draft assemblies. Each repeat family in the RepeatMasker library was determined to be either order-specific or 'ancestral repeats' present at orthologous sites, usually on the basis of the average divergence level of the interspersed repeat family copies. For elements with an average divergence of 15–19% in human, copies were checked to be present or absent at mouse orthologous sites, to have inserted in known primate-specific repeats, or to have inserts of known mammalian-wide elements. No mammalian-wide repeats were found in the mouse genome that were not already known from the human genome. NA, not applicable.

Transposon subfamilies shared between human and mouse were equally old in both species,
and were older than all human- and mouse-specific,
and showed the same molecular clock (2x faster in mouse).

Table 6 **Divergence levels of interspersed repeats predating the human–mouse speciation**

Interspersed repeat		Mouse				Human					
Family	Class	kb	Divergence	Range	JC	kb	Divergence	Range	JC	Substitution ratio	Adjusted ratio
L1MA6	LINE1	1,795	0.28	0.04	0.35	2,738	0.16	0.05	0.184	1.92	1.98
L1MA7	LINE1	789	0.28	0.04	0.35	3,502	0.16	0.04	0.181	1.92	1.96
L1MA8	LINE1	951	0.27	0.04	0.34	4,488	0.15	0.04	0.172	1.96	1.96
L1MA9	LINE1	1,032	0.28	0.04	0.35	6,468	0.18	0.05	0.201	1.74	1.86
L1MA10	LINE1	160	0.29	0.04	0.36	1,492	0.19	0.05	0.224	1.61	1.80
L1MB1	LINE1	627	0.29	0.04	0.36	2,947	0.18	0.05	0.211	1.71	1.87
L1MB2	LINE1	725	0.28	0.04	0.35	3,309	0.18	0.06	0.201	1.75	1.87
L1MC1	LINE1	1,389	0.28	0.04	0.36	7,221	0.17	0.05	0.198	1.80	1.92
MLT1A	MaLR	984	0.31	0.04	0.39	2,203	0.21	0.04	0.242	1.62	1.73
MLT1A0	MaLR	1,794	0.30	0.04	0.38	5,424	0.19	0.04	0.219	1.74	1.80
MLT1A1	MaLR	539	0.29	0.04	0.37	1,705	0.19	0.04	0.214	1.74	1.78
MLT1B	MaLR	73	0.28	0.03	0.35	4,482	0.18	0.04	0.203	1.73	1.73
MLT1C	MaLR	2,071	0.30	0.04	0.37	5,511	0.21	0.04	0.245	1.53	1.64
Looper	DNA	33	0.28	0.04	0.34	48	0.18	0.03	0.211	1.62	1.69
MER20	DNA	435	0.29	0.05	0.37	2,205	0.19	0.05	0.222	1.65	1.76
MER33	DNA	232	0.27	0.05	0.33	1,207	0.18	0.04	0.211	1.57	1.63
MER53	DNA	82	0.26	0.05	0.31	524	0.17	0.05	0.191	1.63	1.63
Tigger6a	DNA	97	0.29	0.03	0.37	190	0.18	0.06	0.211	1.77	1.85

Shown are the number of kilobases matched by each subfamily (kb), the median divergence (mismatch) level of all copies from the consensus sequence, the interquartile range of these mismatch levels (range), and a Jukes–Cantor estimate of the substitution level to which the median divergence level corresponds (JC). Notice that RepeatMasker found, on average, four- to fivefold more copies in the human than in the mouse genome, as a result of the higher DNA loss in the rodent lineage as well a failure to identify many highly diverged copies. The two right columns contain the ratio of the JC substitution level in mouse over human, and an adjusted ratio (AR) of the mouse and human substitution level after subtraction from both of the approximate fraction accumulated in the common human–mouse ancestor. For this fraction we have taken the difference between the ancestral repeat average substitution level and least diverged ancestral repeat family (L1MA8). See the Supplementary Information for a discussion of the origin of the variance in the human and mouse ratios.

Transposable elements undergo "bursts" of transposition

- Canonical (or consensus) transposons undergo replicative transposition at a high rate, and also produce near-canonical daughter elements that transpose at a high rate.
- The three most active *L1* subfamilies in the human genome (currently "bursting" are represented by ~10 elements each whose activity can be readily measured in a cell culture assay.
- Families in this state rapidly increase in numbers, until the 10 elements have all mutated into the inactive state.
- Subsequently, each subfamily member diverges separately from the subfamily consensus, leading to a "star-like" phylogeny (complicated by gene conversion *etc.*)
- The age of the subfamily is then inferred from the branch lengths in the star. According to this assay, the youngest transposons in the human genome are <1 million years old (still polymorphic), while the oldest are >> 200 million years old.

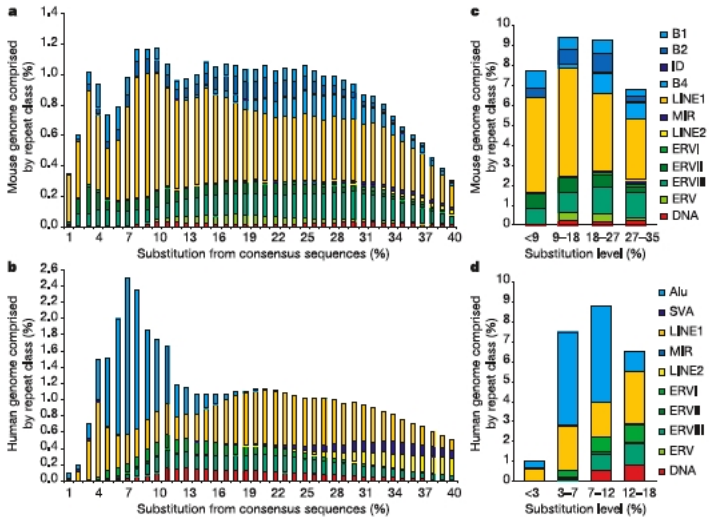


Figure 10 Age distribution of interspersed repeats in the mouse and human genomes. This is an update of Fig. 18 in the IHGC human genome paper[1]. **a**, **b**, Distribution for mouse and human of copies of each repeat class in bins corresponding to 1% increments in substitution level calculated using Jukes–Cantor formula ($K = -3/4 \ln(1 - D^{*}_{\text{mut}}4/3)$) (see Supplementary Information for definition). The first bin for mouse is artificially low because the WGS assembly used for mouse excludes a larger percentage of very recent repeats. **c**, **d**, Interspersed repeats grouped into bins of approximately equal time periods after adjusting for the different rates of substitution in the two genomes. On average, the substitution level has been twofold higher in the mouse than in the human lineage (Table 6), but the difference was initially less and has increased over time. The present rates may differ over fourfold. The activity of transposable elements in the mouse lineage has been quite uniform compared with the human lineage, where an overall decline was interrupted temporarily by a burst of Alu activity. The apparent absence of <2% diverged interspersed repeats in mouse is primarily due to the shotgun sequencing strategy; long, closely similar interspersed repeats very often were not assembled. This is supported by an up to tenfold higher concentration of young L1 and ERV elements at the edges of gaps. The gradually decreasing density of repeats beyond a 30% substitution level reflects in part the limits of the detection method.

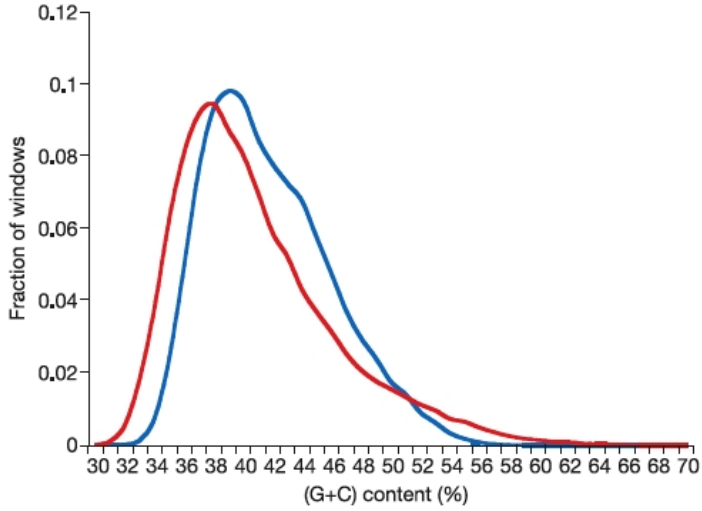


Figure 7 Distribution of (G+C) content in the mouse (blue) and human (red) genomes. Mouse has a higher mean (G+C) content than human (42% compared with 41%), but human has a larger fraction of windows with either high or low (G+C) content. The distribution was determined using the unmasked genomes in 20-kb non-overlapping windows, with the fraction of windows (*y* axis) in each percentage bin (*x* axis) plotted for both human and mouse.

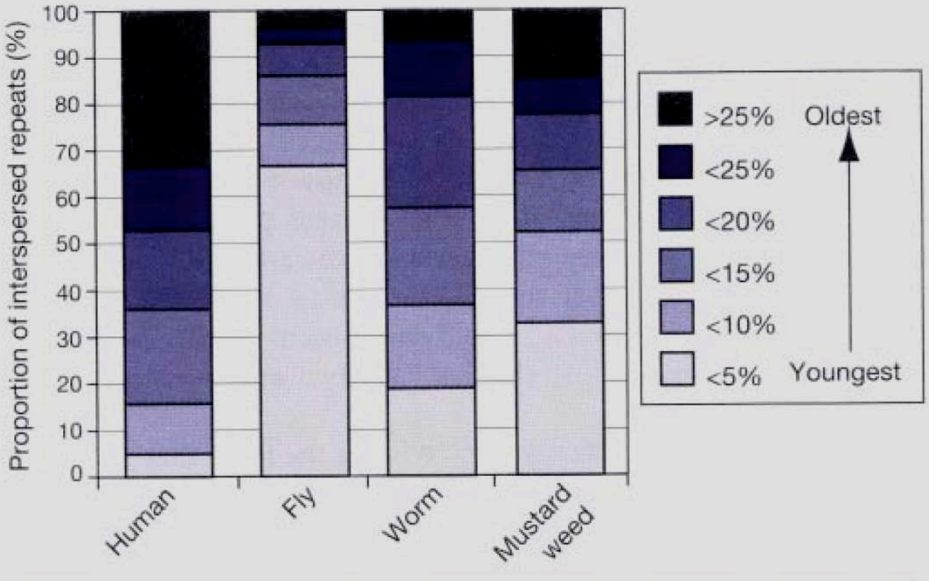


Figure 20 Comparison of the age of interspersed repeats in eukaryotic genomes. The copies of repeats were pooled by their nucleotide substitution level from the consensus.

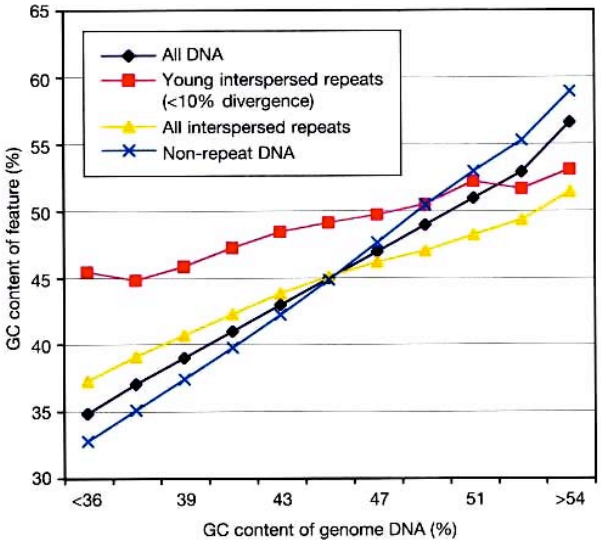


Figure 28 Interspersed repeats tend to diminish the differences between GC bins, despite the fact that GC-rich transposable elements (specifically Alu) accumulate in GC-rich DNA, and AT-rich elements (LINE1) in AT-rich DNA. The GC content of particular components of the sequence (repeats, young repeats and non-repeat sequence) was calculated as a function of overall GC content.

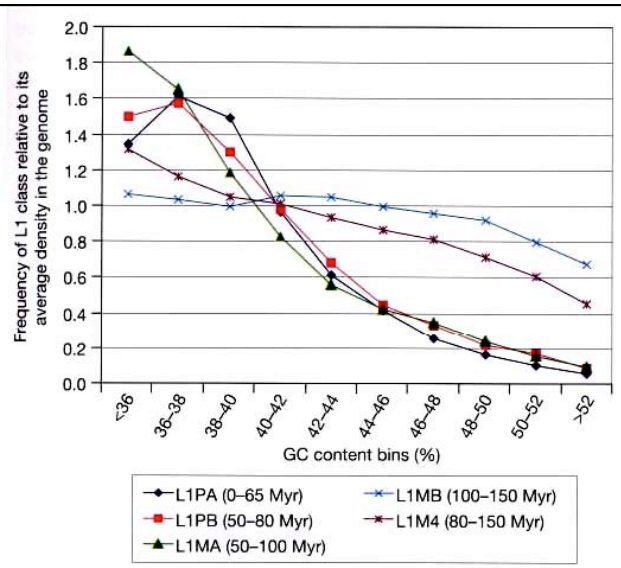


Figure 25 Distribution of various LINE cohorts as a function of local GC content. The divergence levels and ages of the cohorts are shown in the key. (The divergence levels were measured for the 3' UTR of the LINE1 element only, which is best characterized evolutionarily. This region contains almost no CpG sites, and thus 1% divergence level corresponds to a much longer time than for CpG-rich Alu copies).

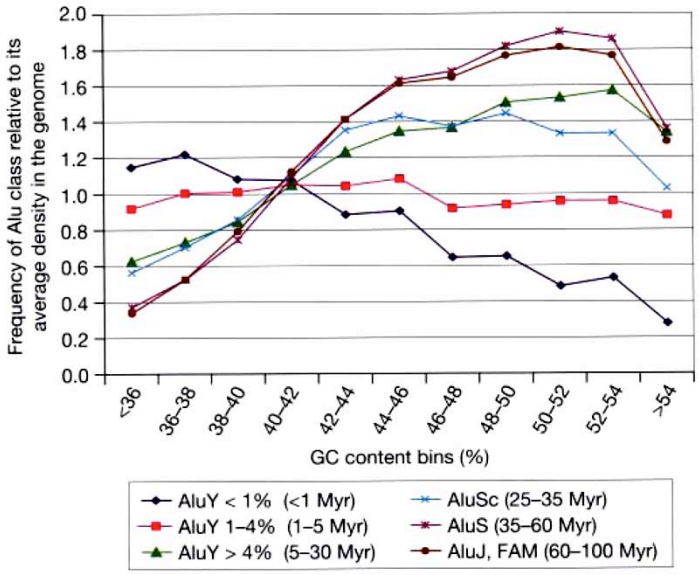


Figure 23 Alu elements target AT-rich DNA, but accumulate in GC-rich DNA. This graph shows the relative distribution of various Alu cohorts as a function of local GC content. The divergence levels (including CpG sites) and ages of the cohorts are shown in the key.

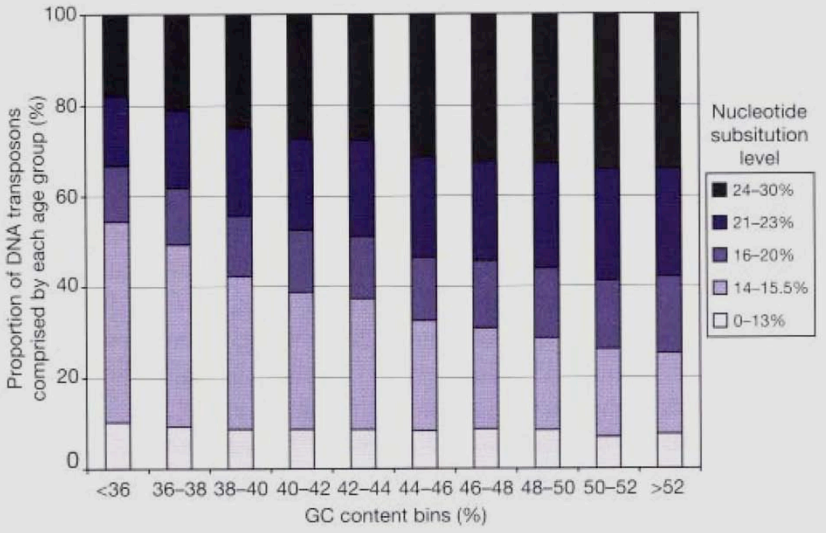


Figure 24 DNA transposon copies in AT-rich DNA tend to be younger than those in more GC-rich DNA. DNA transposon families were grouped into five age categories by their median substitution level (see Fig. 19). The proportion attributed to each age class is shown as a function of GC content. Similar patterns are seen for LINE1 and LTR elements.

Outline

- Noncoding DNA – human vs. mouse
- Transposon structure and function
- Transposon distributions (age, GC content, species)
- Hox gene clusters vs. transposons
- Chromosomal rearrangements

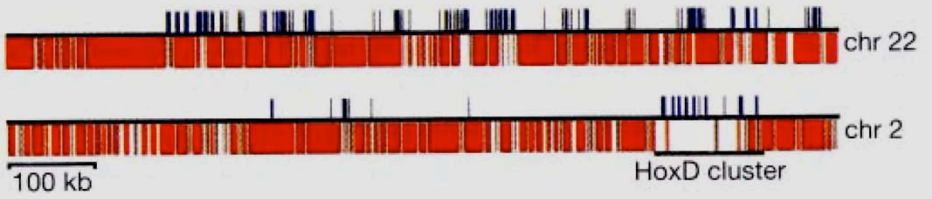


Figure 21 Two regions of about 1 Mb on chromosomes 2 and 22. Red bars, interspersed repeats; blue bars, exons of known genes. Note the deficit of repeats in the HoxD cluster, which contains a collection of genes with complex, interrelated regulation.

Segmental duplications in the human genome

- Sequences were considered to be putative segmental duplications only if <99.5% identity (to avoid contig assembly errors) and >90% identity (to avoid random similarities due to transposons etc.).
- Given an average molecular clock rate in monkeys and apes of ~2 x 10^{-9} bp^{-1} yr^{-1}, a sequence subject to neutral evolution would be expected to retain 90% sequence identity after ~50 million years.
- In other words, the criteria of >90% identity is likely to detect only duplications that occurred relatively recently (since the divergence of Old World monkeys from New World monkeys).
- These recent segmental duplications were found to represent about 3.5% of the human genome sequence. There are undoubtedly many more older duplications that are not included in this total.
- The majority of segmental duplications found in the human genome project were intrachromosomal (2.0%), but many were interchromosomal (1.5%).

Segmental duplications within and between chromosomes

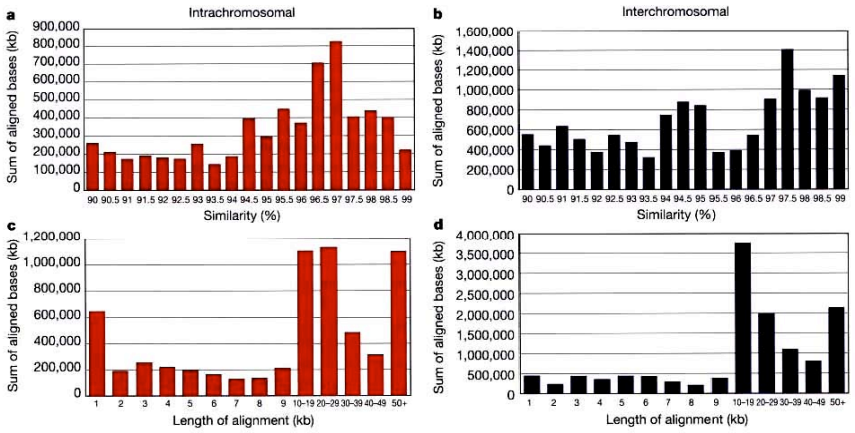


Figure 33 a–d, Sequence properties of segmental duplications. Distributions of length and per cent nucleotide identity for segmental duplications are shown as a function of the number of aligned bp, for the subset of finished genome sequence. Intrachromosomal, red; interchromosomal, blue.

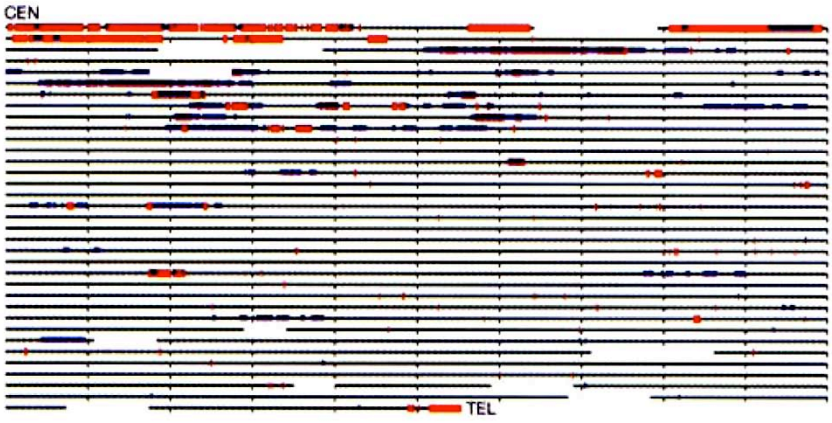


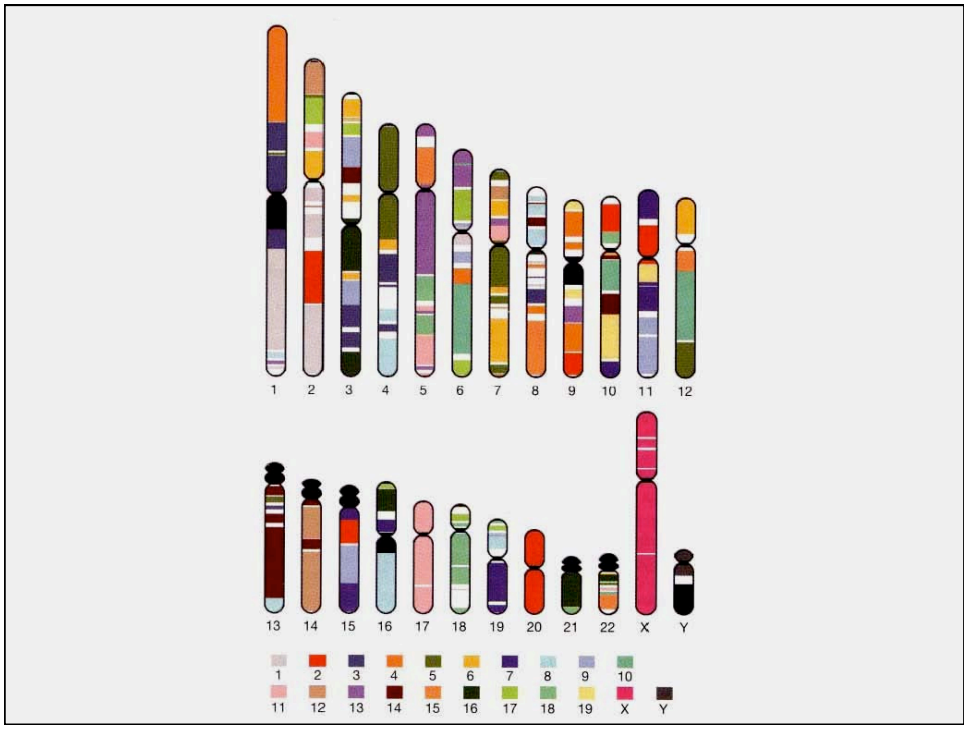
Figure 30 Duplication landscape of chromosome 22. The size and location of intrachromosomal (blue) and interchromosomal (red) duplications are depicted for chromosome 22q, using the PARASIGHT computer program (Bailey and Eichler, unpublished). Each horizontal line represents 1 Mb (ticks, 100-kb intervals). The chromosome sequence is oriented from centromere (top left) to telomere (bottom right). Pairwise alignments with $>$ 90% nucleotide identity and $>$ 1 kb long are shown. Gaps within the chromosomal sequence are of known size and shown as empty space.

Outline

- Noncoding DNA – human vs. mouse
- Transposon structure and function
- Transposon distributions (age, GC content, species)
- Hox gene clusters vs. transposons
- Chromosomal rearrangements

Synteny groups and chromosomal rearrangements

- Synteny groups are conserved chromosomal segments - that is, portions of chromosomes within which the same genes are present in (nearly) the same order in both of the species being analyzed.
- The number of synteny groups depends on the species being compared - from human to chimp, there are several dozen synteny groups, almost the same as the number of chromosomes.
- From human to mouse, there are several hundred synteny groups. From human to fish, there are > 1,000 synteny groups. But synteny groups are still a useful concept when comparing fish to human, because there is still significant conservation of genetic map order.
- The "average" chromosomal rearrangement rate in mammals is roughly 1 chromosomal rearrangement (that is retained and becomes fixed in the population) per 5 million years.
- But no mammal is average - rodents have experienced chromosomal rearrangements several-fold faster than this, while carnivores and most higher primates have undergone rearrangements several-fold slower than average.



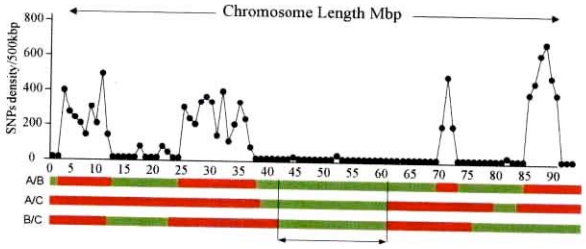


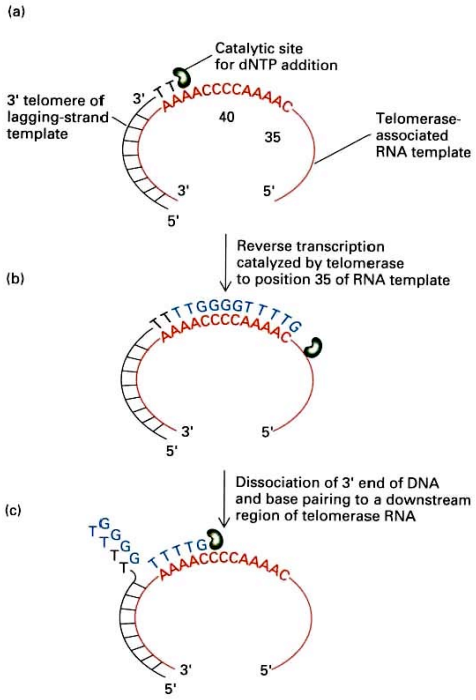
Figure 2. The SNP density between any two inbred strains of mice varies according to the chromosomal region concerned and changes abruptly when passing from one region to the next. This observation is in agreement with historical data on the origin of laboratory inbred strains indicating that they are derived from a small pool of wild ancestors belonging to different subspecies of the genus *Mus*. Because of this polyphyletic origin, the mouse genome can be regarded as a mosaic of chromosomal segments of various sizes. When the SNP density is low, the segments in question share the same ancestral origin and the few observed SNPs are those resulting from recent mutations. When the SNP density is high, on the contrary, the chromosomal segments have a different origin. When three strains are compared, as on the diagram represented here, one can perfectly observe that three homologous segments have a high SNP density on pairwise comparisons, if all three of them have an independent origin stemming, for example, in three different subspecies of the genus *Mus*. When a particular region with low SNP density cosegregates with a particular phenotype, the region in question may harbor the genetic determinants for the phenotype in question. Redrawn with permission from *PNAS* © 2003, Wiltshire et al. (2003).

Telomere structure & function

- Telomeres end in a hairpin loop (that is, the two strands are connected to each other by a continuous sugar-phosphate backbone).
- A chromosome end without a telomere is genetically unstable.
- This is due, in part, to the fact that nucleases can attack a blunt end.
- It is also due, in part, to the fact that blunt ends ("double-stranded chromosome breaks") are recombinogenic (are a rate-limiting step in crossovers between chromosomes).
- Creating a blunt end stimulates crossing over even at places where it is not supposed to occur (such as chromosome fusion).
- Sequences near telomeres (subtelomeric sequences) tend to be GC-rich.

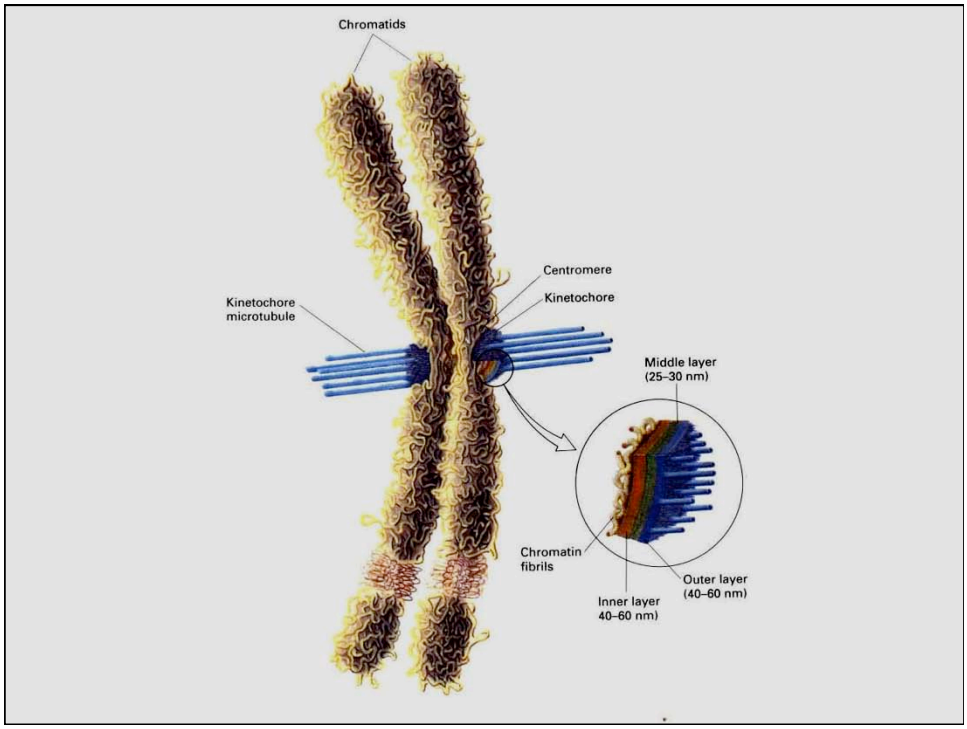
Telomere replication preserves chromosome ends in eukaryotes

- Telomeres contain many copies of a ~6 base repeat (in humans, this is TTAGGG).
- Additional copies of the telomere repeat are synthesized in some cells (for example, germ cells), by a specialized enzyme called telomerase. Other cells (most somatic cells) lack this enzyme.
- The synthesis of additional telomere repeats solves the problem of random RNA priming on the lagging strand, which otherwise would cause the chromosome to get a few hundred bp shorter after every cell division.
- Loss of chromosome ends is believed to be a major cause of aging in humans, due to the death of cells that have lost essential genes due to chromosome shortening.
- Conversely, mutations that cause the inappropriate expression of telomerase is a key step in cancer.



Centromere function

- Centromere binding proteins exclude nucleosomes from the region of their binding. They bind cooperatively (each one helps the next).
- Some centromere binding proteins are evolutionarily related to histones (the proteins that make up the nucleosome particle).
- The centromere binding complex can function as a motor. That is, it can use the energy in ATP to translocate towards the (-) ends of microtubules.
- Directed depolymerization of the (+) end of microtubules also helps in this process. Sort of like tearing up the road behind your car, it prevents you from slipping into reverse.

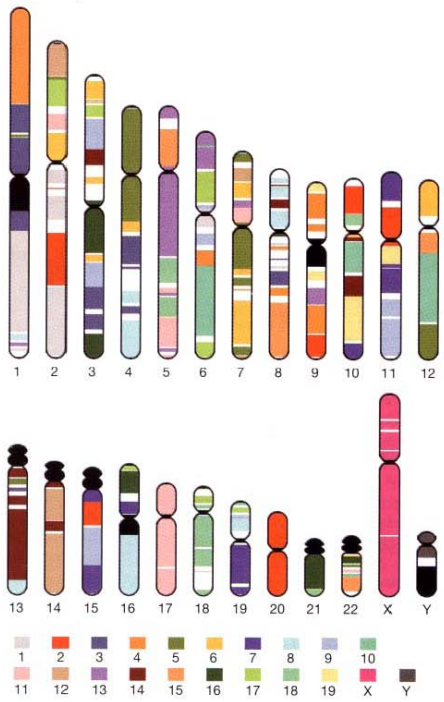


Dosage compensation on the X and Y chromosomes

- The X chromosome is "dosage compensated" in female mammals, by random inactivation of one X chromosome (Barr body) in the early embryo.
- Dosage compensation is essential to provide physiological levels of gene expression in both sexes. Mutations that disrupt dosage compensation are usually lethal.
- Dosage compensation has evolved independently at least three times: in mammals, birds, and insects. Each group uses a different mechanism.
- Dosage compensation in mammals is reasonably well understood, and is regulated by the interactions of noncoding RNA (*Xist*), DNA methylation, and histone modifications. The X chromosome also has an unusually high concentration of L1 LINE elements.
- The Y chromosome is not dosage compensated. PARs are not dosage compensated.

Linkage conservation on the X chromosome

- According to "Ohno's Law", chromosomal rearrangements of the mammalian X chromosome were selected against, during the evolutionary history of the mammals.
- This was because X chromosome rearrangements (particularly translocations to another chromosome) could disrupt dosage compensation.
- Although exceptions are known, this hypothesis has been largely confirmed by chromosome painting and the mammalian genome projects.
- In contrast, most inversions on the Y chromosome are likely to be even less deleterious than inversions on autosomes.
- This is strongly supported by the finding that "normal" male humans are polymorphic for a 3.5 Mb inversion in the Y chromosome.



Discussion questions

- Discuss how sequence comparisons are used to date the ages of retroposon families. What are the assumptions / background information that this method of analysis is based upon (i.e., mechanisms of transposition)? What conclusions can be drawn from this work?
- Why and how do transposons target specific stages of the life cycle for transposition? How do they target specific portions of the genome? How is this utilized in "enhancer trap" screens?
- Why and how do transposons maintain the telomeres of Drosophila chromosomes? How is this situation likely to have arisen?
- Why does the X chromosome represent the largest synteny group in the human genome? What is a synteny group?
- Why and how were transposons excluded from Hox gene clusters? How useful is this observation vs. whole mouse genome comparison in analyzing transcriptional regulatory regions?