Genomes summary

1. >930 bacterial genomes sequenced.
2. Circular. Genes densely packed.
3. 2-10 Mbases, 470 - 7,000 genes
4. Genomes of >200 eukaryotes (45 "higher") sequenced.
5. Linear chromosomes
6. On average, ~50% of gene functions "known".
7. Human genome: <40,000 genes code for >120,000 proteins.
   Large gene families (e.g. 500 protein kinases)
   98% of human DNA is noncoding.
   ~3% of human DNA = simple repeats (satellites, minisatellites, microsatellites)
   ~50% of DNA = mobile elements (DNA transposons, retrotransposons (LTR and nonLTR) & pseudogenes)

Bacterial genome sizes

Predicted genes in bacterial species	
Mycoplasma genitalium	470
Mycoplasma mycoides	985
E. coli	4,288
B. anthracis	5,508
P. aeruginosa	5,570
Mycobacterium leprae	1,604
Mycobacterium tuberculosis	3,995

+ ~930 sequenced microbial genomes
(http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi)

Small and large

Genome sizes

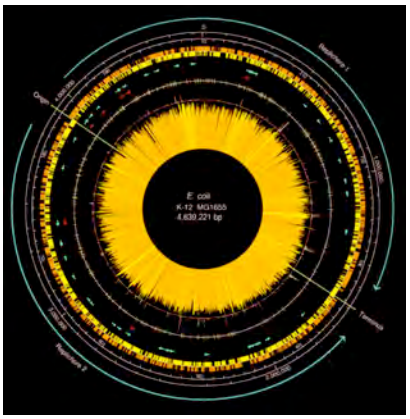
Table 20.1 Genome Sizes and Numbers of Genes

Organism	Genome Size	Estimated Number of Genes	Genes per Mb*
H. influenzae (bacterium)	1.8 Mb*	1,700	950
S. cerevisiae (yeast)	12 Mb	6,000	500
C. elegans (nematode)	97 Mb	19,000	200
A. thaliana (plant)	100 Mb	25,000	200
D. melanogaster (fruit fly)	180 Mb	13,000	100
H. sapiens (human)	3,200 Mb	30,000–40,000	10

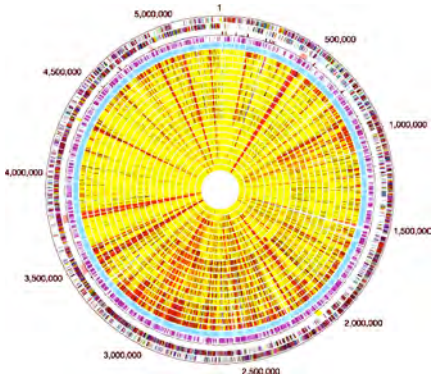
*Mb = million base pairs

Gene density down in mammals

Bacterial genomes are circular and densely packed with genes - 1

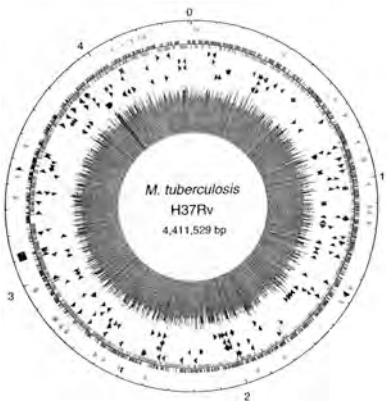


E. coli. Genes (circles 1 & 2).

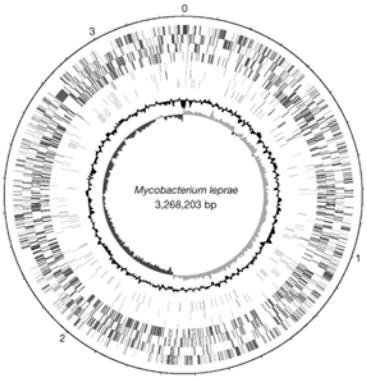


B. anthracis. Genes (circles 1 & 2).

Bacterial genomes are circular and densely packed with genes - 2

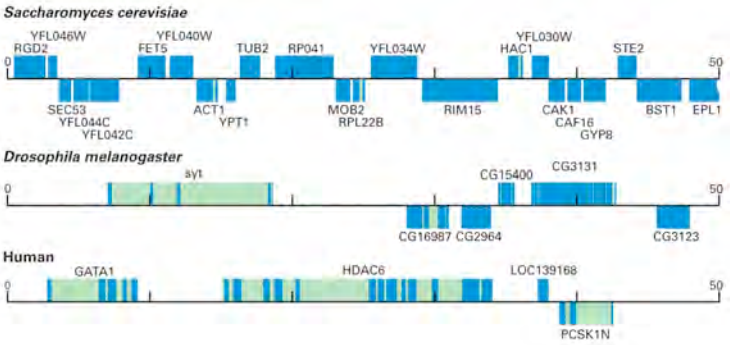


M. tuberculosis (4.41 MB).
Genes (circles 1 & 2).



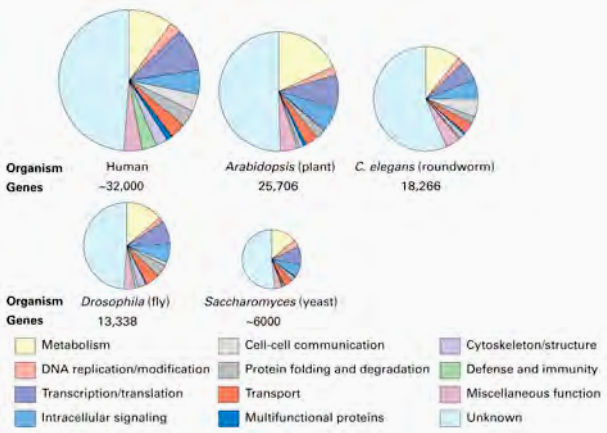
M. leprae (4.41 MB).
Genes (circles 1 & 2),
1116 pseudogenes (circles 3 & 4).

Representative gene arrangements in 50 kb segments of yeast, fly and human DNA.



Few yeast genes contain introns (exons are blue). Genes above and below the line are transcribed in opposite directions.

Numbers and types of genes in different eukaryotes



About half the genes encode proteins of unknown function.

Human genome: <2% ORFs & 48% repeats

Human genome:
<40,000 genes
Average ~3 proteins/gene
98% of DNA is noncoding
Individuals 99.9% identical
(1 difference/1000 bp means many markers for mapping).
Large families of repeats.
481 sequences >200 bp that are absolutely conserved in mouse.
Large gene families (E.g. ~500 Ser/Thr protein kinases many Zn^{2+} fingers, etc.)

TABLE 10-1 Major Classes of Eukaryotic DNA and Their Representation in the Human Genome

Class	Length	Copy Number in Human Genome	Fraction of Human Genome, %
Protein-coding genes			
Solitary genes	Variable	1	~1.5* (0.8)†
Duplicated or diverged genes in gene families	Variable	2–~1000	~1.5* (0.8)†
Tandemly repeated genes encoding rRNAs, tRNAs, snRNAs, and histones	Variable	20–300	0.3
Repetitious DNA			
Simple-sequence DNA	1–500 bp	Variable	3
Interspersed repeats‡			
DNA transposons	2–3 kb	300,000	3
LTR retrotransposons	6–11 kb	440,000	8
Non-LTR retrotransposons			
LINEs	6–8 kb	860,000	21
SINEs	100–300 bp	1,600,000	13
Processed pseudogenes‡	Variable	1–~100	~0.4
Unclassified spacer DNA	Variable	n.a.§	~2.5

*Complete transcription units, including introns.
†Protein-coding exons. The total number of human protein-coding genes is estimated to be 30,000–35,000, but this number is based on current methods for identifying genes in the human genome sequence and may be an underestimate.
‡Not applicable.

SOURCE: E. S. Lander et al., 2001, *Nature* 409:860.

Human genome: individuals 99.9% identical

For every 1000 people . . .

Sequencing revealed one major allele for most genes in populations

Human populations have not been genetically isolated for very long (~2-3 M years)

Many variations have not had time to spread throughout populations.

Human genome: individuals 0.1% different

For every person . . .

Lots of variation!

3.2×10^9 bp/genome x 0.001 changes/bp =

Human genome: individuals 0.1% different

For every person . . .

Lots of variation!

3.2×10^9 bp/genome x 0.001 changes/bp =

3.2×10^6 changes/genome

Human genome: individuals 0.1% different

For every person . . .

Lots of variation!

3.2×10^9 bp/genome x 0.001 changes/bp =

3.2×10^6 changes/genome

Two major types of variation

- SNPs
- Repeated DNA - short to long repeats

Variations produce RFLPs (Restriction Fragment Length Polymorphisms)!

SNPs

Single Nucleotide Polymorphisms (Changes of a single base)

- Some are neutral
- Some alter gene function

Identifying SNPs

- Phenotype (disease), e.g Sickle cell anemia
- Sequencing genes/cDNAs
- Restriction digest

RFLPs

Restriciton Fragment Length Polymorphisms (Changes of restriction enzyme sites)

RFLPs

Restriciton Fragment Length Polymorphisms (Changes of restriction enzyme sites)

For every random 3 x 10^6 SNPs:
~1/256 will be in 4-base restriction sites
--> ~10^4 RFLPs for EACH four-base cutter!
~1/4096 will be in 6-base restriction sites
--> ~ 7.5 x 10^2 RFLPs for EACH six-base cutter!

Lots of markers (RFLPs) to map genes by linkage to RFLPs

Human genome: 48% repeats

Human genome:
<40,000 genes
Average ~3 proteins/gene
95% of DNA is noncoding
Individuals 99.9% identical
(1 difference/1000 bp means many markers for mapping).
Large families of repeats.
Satellites (micro, mini and conventional)
Transposons
Retrotransposons

TABLE 10-1 Major Classes of Eukaryotic DNA and Their Representation in the Human Genome

Class	Length	Copy Number in Human Genome	Fraction of Human Genome, %
Protein-coding genes			
Solitary genes	Variable	1	~1.5* (0.8)†
Duplicated or diverged genes in gene families	Variable	2–~1000	~1.5* (0.8)†
Tandemly repeated genes encoding rRNAs, tRNAs, snRNAs, and histones	Variable	20–300	0.3
Repetitious DNA			
Simple-sequence DNA	1–500 bp	Variable	3
Interspersed repeats			
DNA transposons	2–3 kb	300,000	3
LTR retrotransposons	6–11 kb	440,000	8
Non-LTR retrotransposons			
LINEs	6–8 kb	860,000	21
SINEs	100–300 bp	1,600,000	13
Processed pseudogenes	Variable	1–~100	~0.4
Unclassified spacer DNA	Variable	n.a.‡	~2.5

*Complete transcription units, including introns.
†Protein-coding exons. The total number of human protein-coding genes is estimated to be 30,000–35,000, but this number is based on current methods for identifying genes in the human genome sequence and may be an underestimate.
‡Not applicable.
SOURCE: E. S. Lander et al., 2001, *Nature* 409:860.

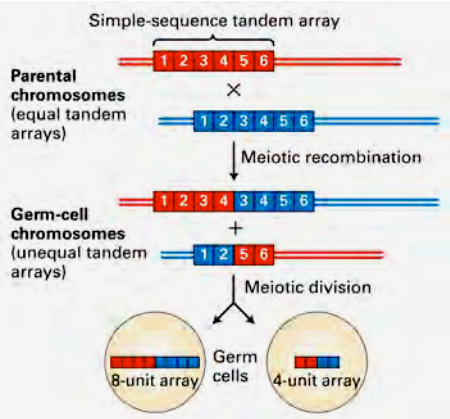
Satellites

Microsatellites: 1 - 13 bps in ~150 bp arrays

Minisatellites: 15-100 bps in 1-5 kb arrays

Satellites: 14 - 500 bps in 20-100 kb arrays

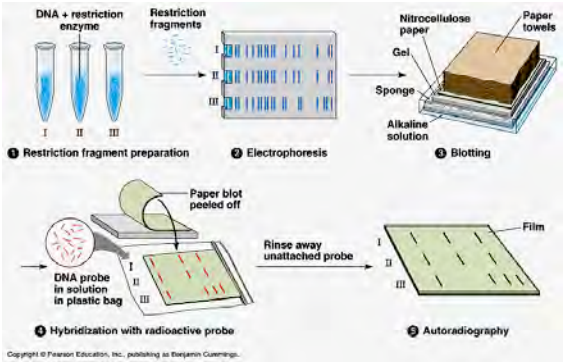
Origins of length polymorphisms in simple-sequence repeats.



Generation of length differences by unequal crossing over in meiosis

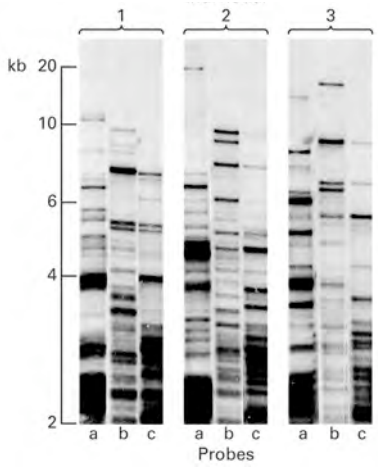
"Southern" blotting detects DNA sequences by hybridization

1. Digest DNA using restriction enzyme(s)
2. Run gel
3. Transfer DNA from gel to (nitrocellulose) paper.
4. Denature DNA, hybridize probe DNA, and wash off excess probe.
5. Detect the probe on the paper. E.g. by autoradiography.



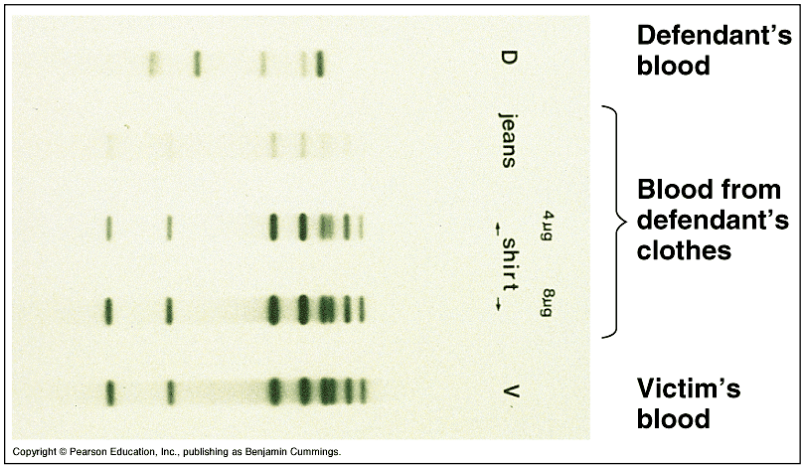
Different distributions of minisatellites

Three repeats (a, b, c) in 3 people (1, 2, 3)



Southern blot of HinfI-digested DNA

RFLPs -- DNA "finger print" in a murder case



Southern blot of DNA samples digested with a restriction enzyme

Human genome: 48% repeats

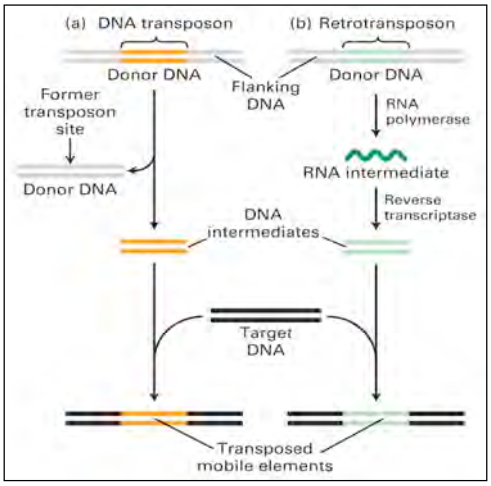
Human genome:
<40,000 genes
Average ~3 proteins/gene
95% of DNA is noncoding
Individuals 99.9% identical
(1 difference/1000 bp means many markers for mapping).
Large families of repeats.
Satellites (micro, mini and conventional)
Transposons
Retrotransposons

TABLE 10-1 Major Classes of Eukaryotic DNA and Their Representation in the Human Genome

Class	Length	Copy Number in Human Genome	Fraction of Human Genome, %
Protein-coding genes			
Solitary genes	Variable	1	~1.5* (0.8)†
Duplicated or diverged genes in gene families	Variable	2–~1000	~1.5* (0.8)†
Tandemly repeated genes encoding rRNAs, tRNAs, snRNAs, and histones	Variable	20–300	0.3
Repetitious DNA			
Simple-sequence DNA	1–500 bp	Variable	3
Interspersed repeats			
DNA transposons	2–3 kb	300,000	3
LTR retrotransposons	6–11 kb	440,000	8
Non-LTR retrotransposons			
LINEs	6–8 kb	860,000	21
SINEs	100–300 bp	1,600,000	13
Processed pseudogenes	Variable	1–~100	~0.4
Unclassified spacer DNA	Variable	n.a.‡	~2.5

*Complete transcription units, including introns.
†Protein-coding exons. The total number of human protein-coding genes is estimated to be 30,000–35,000, but this number is based on current methods for identifying genes in the human genome sequence and may be an underestimate.
‡Not applicable.
SOURCE: E. S. Lander et al., 2001, *Nature* 409:860.

Two major classes of mobile elements



Proks and euks
DNA intermediate

Eukaryotes
RNA intermediate

Some consequences of repeat sequences in eukaryotes

Genomic diversity in individuals and species. The most common retrotransposon sequences in the human genome are derived from endogenous retroviruses (ERVs). Most of these >440,000 sequences consist only of isolated LTRs, which arise from recombination between the ends.

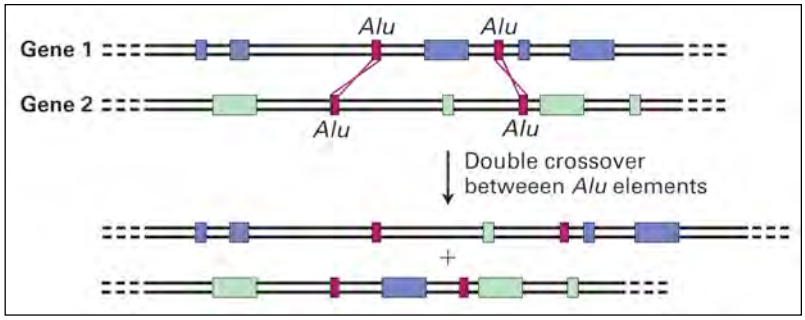
Gene families arise by duplication and divergence.

"Pseudogenes" arise from RT acting on mRNAs.

New genes arise by "exon shuffling".

Exon shuffling may create new proteins in eukaryotes

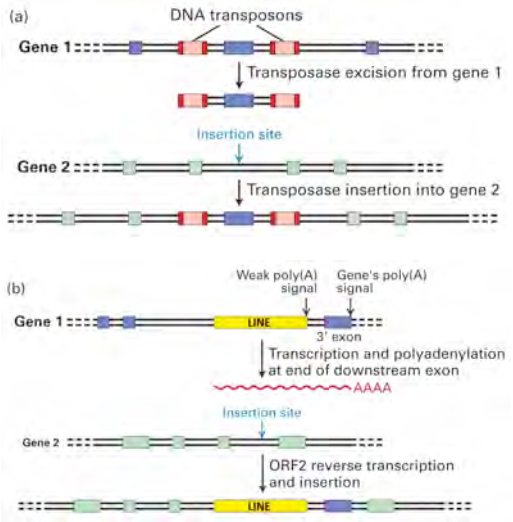
Mechanism 1: **Recombination between homologous interspersed repeats** in the introns of separate genes would produce a new combination of exons.



Exon shuffling may create new proteins in eukaryotes

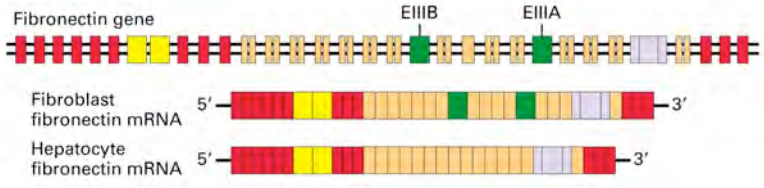
Mechanism 2:
Transposition of an exon
(a) DNA hopping of flanking transposons

(b) Reverse transcription of a LINE RNA extending into the 3′ exon of gene 1 can produce a DNA that gives gene 2 a new 3′ exon upon integration.



Possible results of exon shuffling

1. Modular proteins (with alternate splicing patterns). E.g. Fibronectin gene and mRNA.

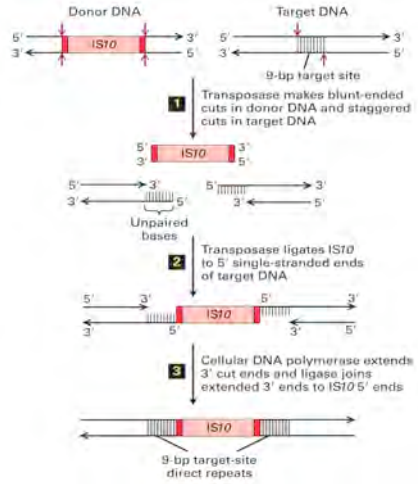


2. Separate proteins that form a complex in one organism are sometimes fused into a single polypeptide chain in another organism.

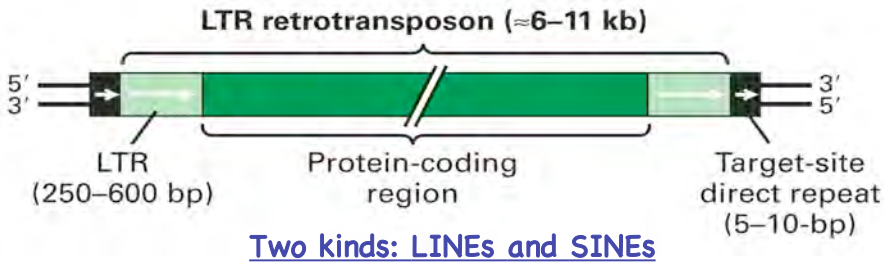
Genomes summary

1. >930 bacterial genomes sequenced.
2. Circular. Genes densely packed.
3. 2-10 Mbases, 470 - 7,000 genes
4. Genomes of >200 eukaryotes (45 "higher") sequenced.
5. Linear chromosomes
6. On average, ~50% of gene functions "known".
7. Human genome: <40,000 genes code for >120,000 proteins.
   Large gene families (e.g. 500 protein kinases)
   98% of human DNA is noncoding.
   ~3% of human DNA = simple repeats (satellites, minisatellites, microsatellites)
   ~50% of DNA = mobile elements (DNA transposons, retrotransposons (LTR and nonLTR) & pseudogenes)

Model for DNA transposition in bacteria



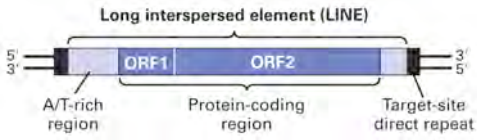
Structure of a eukaryotic LTR retrotransposon



Two kinds: LINEs and SINEs

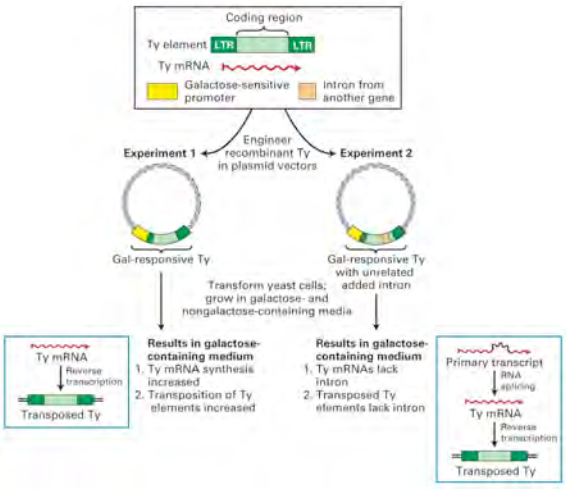
Long Interspersed Elements: encode proteins including RT

Short Interspersed Elements: deletion of protein-coding region



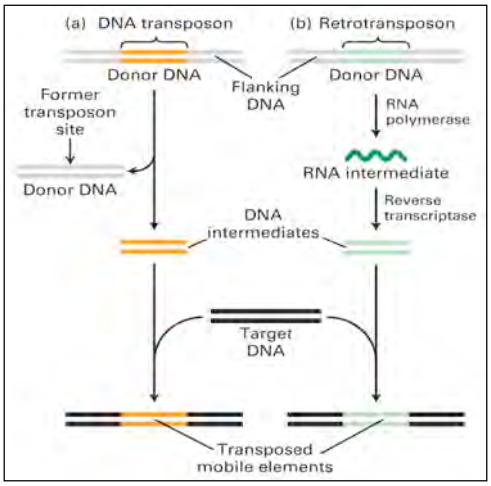
ORF1=RNA binding protein; ORF2=RT and endonuclease.

Experiments with yeast Ty elements demonstrated an RNA intermediate



Introns lost in transposed Tys!

Summary: Two major classes of mobile elements



Proks and euks
DNA intermediate

Eukaryotes
RNA intermediate
LINEs and SINEs