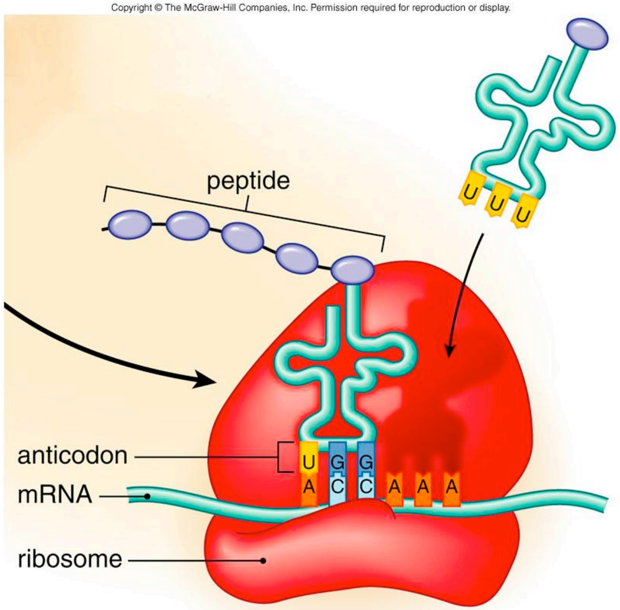
Non-Coding DNA and RNA

Agenda

- Most of the genome (in eukaryotes) doesn't code for proteins, though some of it may still be functional, structurally important, mutagenic, or biologically interesting
- Overview of types of non-coding DNA/RNA
- Small RNAs
- TEs

Non-coding & repetitive DNA may be non-coding, but it is/may be still important!

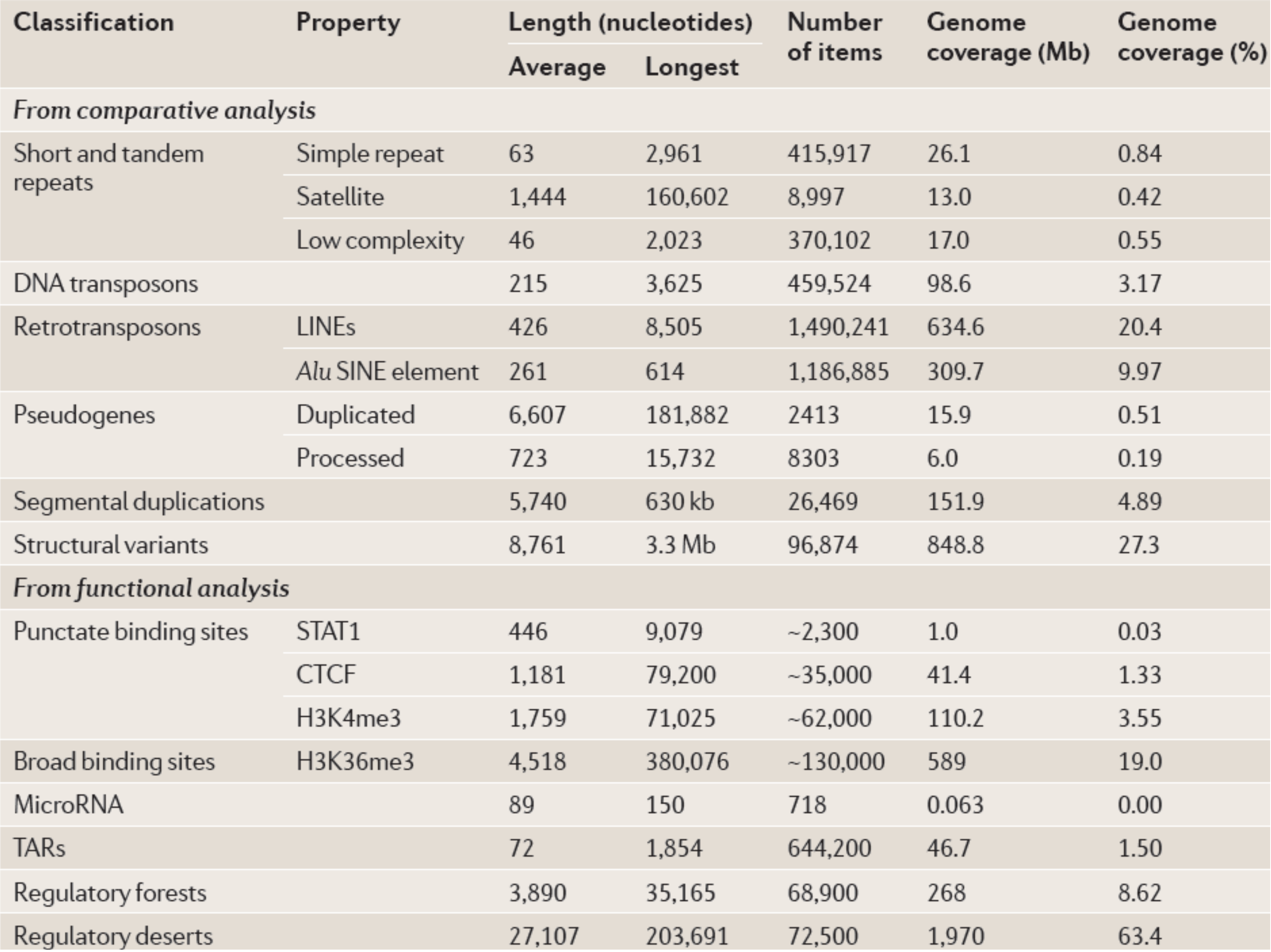


tRNA–amino acid at ribosome

- Introns, self-splicing introns
- Pseudogenes
- Telomeres, centromeres
- Cis- and trans-regulatory elements
- Binding sites
- Transposable elements (TEs)
- Short tandem repeats (1-5 bp)
- Noncoding functional RNAs (big & small RNAs, many kinds)
  - e.g., rRNAs and tRNAs
  - e.g., miRNAs
- Noncoding “elements” (NCEs; often conserved, functional?)
  - e.g., lincRNAs

Very prevalent!

An example from the human genome



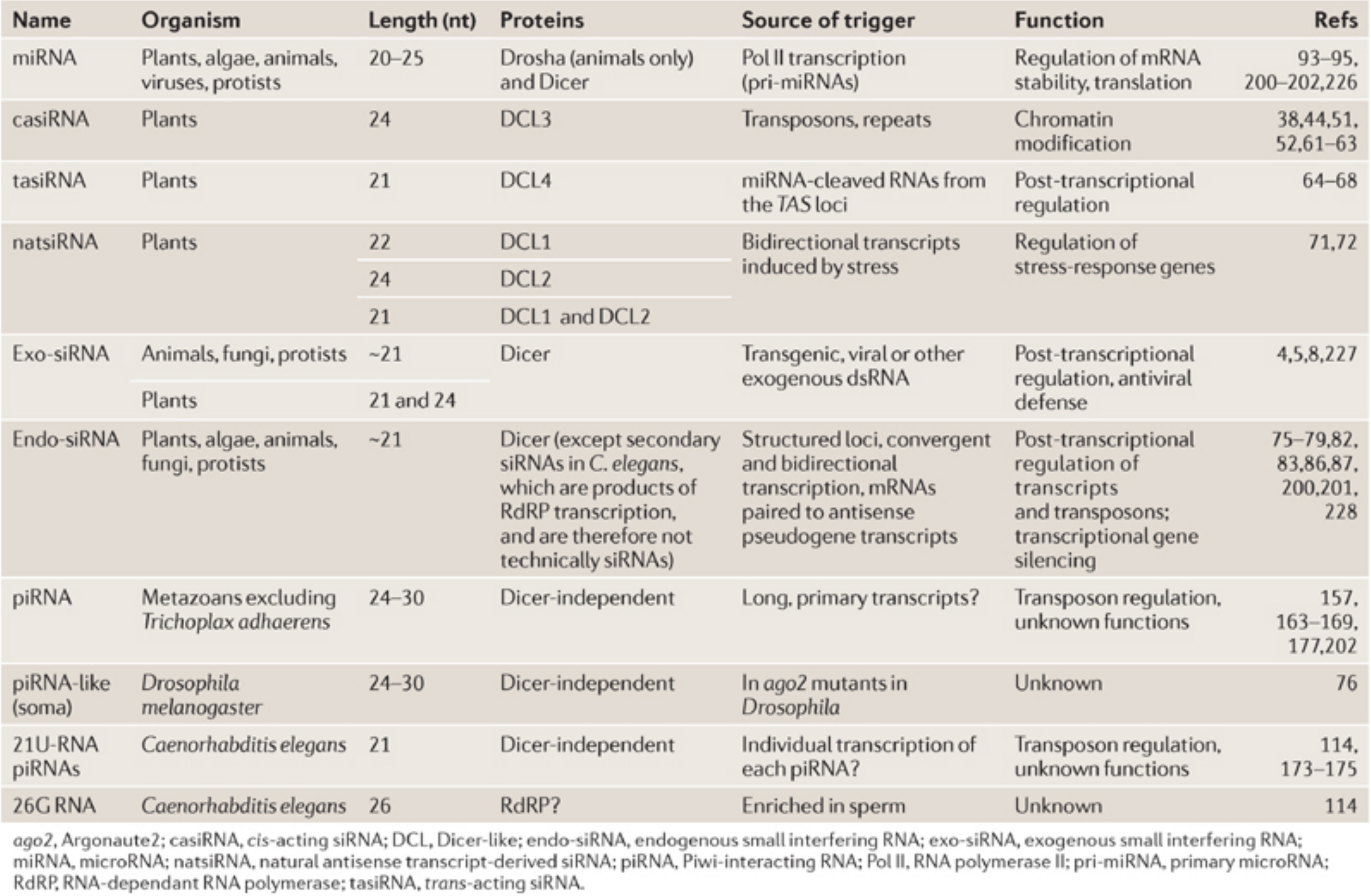
Classification	Property	Length (nucleotides)		Number of items	Genome coverage (Mb)	Genome coverage (%)
		Average	Longest			
From comparative analysis						
Short and tandem repeats	Simple repeat	63	2,961	415,917	26.1	0.84
	Satellite	1,444	160,602	8,997	13.0	0.42
	Low complexity	46	2,023	370,102	17.0	0.55
DNA transposons		215	3,625	459,524	98.6	3.17
Retrotransposons	LINEs	426	8,505	1,490,241	634.6	20.4
	Alu SINE element	261	614	1,186,885	309.7	9.97
Pseudogenes	Duplicated	6,607	181,882	2413	15.9	0.51
	Processed	723	15,732	8303	6.0	0.19
Segmental duplications		5,740	630 kb	26,469	151.9	4.89
Structural variants		8,761	3.3 Mb	96,874	848.8	27.3
From functional analysis						
Punctate binding sites	STAT1	446	9,079	~2,300	1.0	0.03
	CTCF	1,181	79,200	~35,000	41.4	1.33
	H3K4me3	1,759	71,025	~62,000	110.2	3.55
Broad binding sites	H3K36me3	4,518	380,076	~130,000	589	19.0
MicroRNA		89	150	718	0.063	0.00
TARs		72	1,854	644,200	46.7	1.50
Regulatory forests		3,890	35,165	68,900	268	8.62
Regulatory deserts		27,107	203,691	72,500	1,970	63.4

Alexander et al. 2010

Many types of small RNAs

More discovered all the time…

- snRNA
- snoRNA
- gRNA
- miRNA
- piRNA
- siRNA
- casiRNA
- tasiRNA
- rasiRNA



Name	Organism	Length (nt)	Proteins	Source of trigger	Function	Refs
miRNA	Plants, algae, animals, viruses, protists	20–25	Drosha (animals only) and Dicer	Pol II transcription (pri-miRNAs)	Regulation of mRNA stability, translation	93–95, 200–202,226
casiRNA	Plants	24	DCL3	Transposons, repeats	Chromatin modification	38,44,51, 52,61–63
tasiRNA	Plants	21	DCL4	miRNA-cleaved RNAs from the *TAS* loci	Post-transcriptional regulation	64–68
natsiRNA	Plants	22	DCL1	Bidirectional transcripts induced by stress	Regulation of stress-response genes	71,72
		24	DCL2			
		21	DCL1 and DCL2			
Exo-siRNA	Animals, fungi, protists	~21	Dicer	Transgenic, viral or other exogenous dsRNA	Post-transcriptional regulation, antiviral defense	4,5,8,227
	Plants	21 and 24				
Endo-siRNA	Plants, algae, animals, fungi, protists	~21	Dicer (except secondary siRNAs in *C. elegans*, which are products of RdRP transcription, and are therefore not technically siRNAs)	Structured loci, convergent and bidirectional transcription, mRNAs paired to antisense pseudogene transcripts	Post-transcriptional regulation of transcripts and transposons; transcriptional gene silencing	75–79,82, 83,86,87, 200,201, 228
piRNA	Metazoans excluding *Trichoplax adhaerens*	24–30	Dicer-independent	Long, primary transcripts?	Transposon regulation, unknown functions	157, 163–169, 177,202
piRNA-like (soma)	*Drosophila melanogaster*	24–30	Dicer-independent	In *ago2* mutants in *Drosophila*	Unknown	76
21U-RNA piRNAs	*Caenorhabditis elegans*	21	Dicer-independent	Individual transcription of each piRNA?	Transposon regulation, unknown functions	114, 173–175
26G RNA	*Caenorhabditis elegans*	26	RdRP?	Enriched in sperm	Unknown	114

ago2, Argonaute2; casiRNA, *cis*-acting siRNA; DCL, Dicer-like; endo-siRNA, endogenous small interfering RNA; exo-siRNA, exogenous small interfering RNA; miRNA, microRNA; natsiRNA, natural antisense transcript-derived siRNA; piRNA, Piwi-interacting RNA; Pol II, RNA polymerase II; pri-miRNA, primary microRNA; RdRP, RNA-dependant RNA polymerase; tasiRNA, *trans*-acting siRNA.

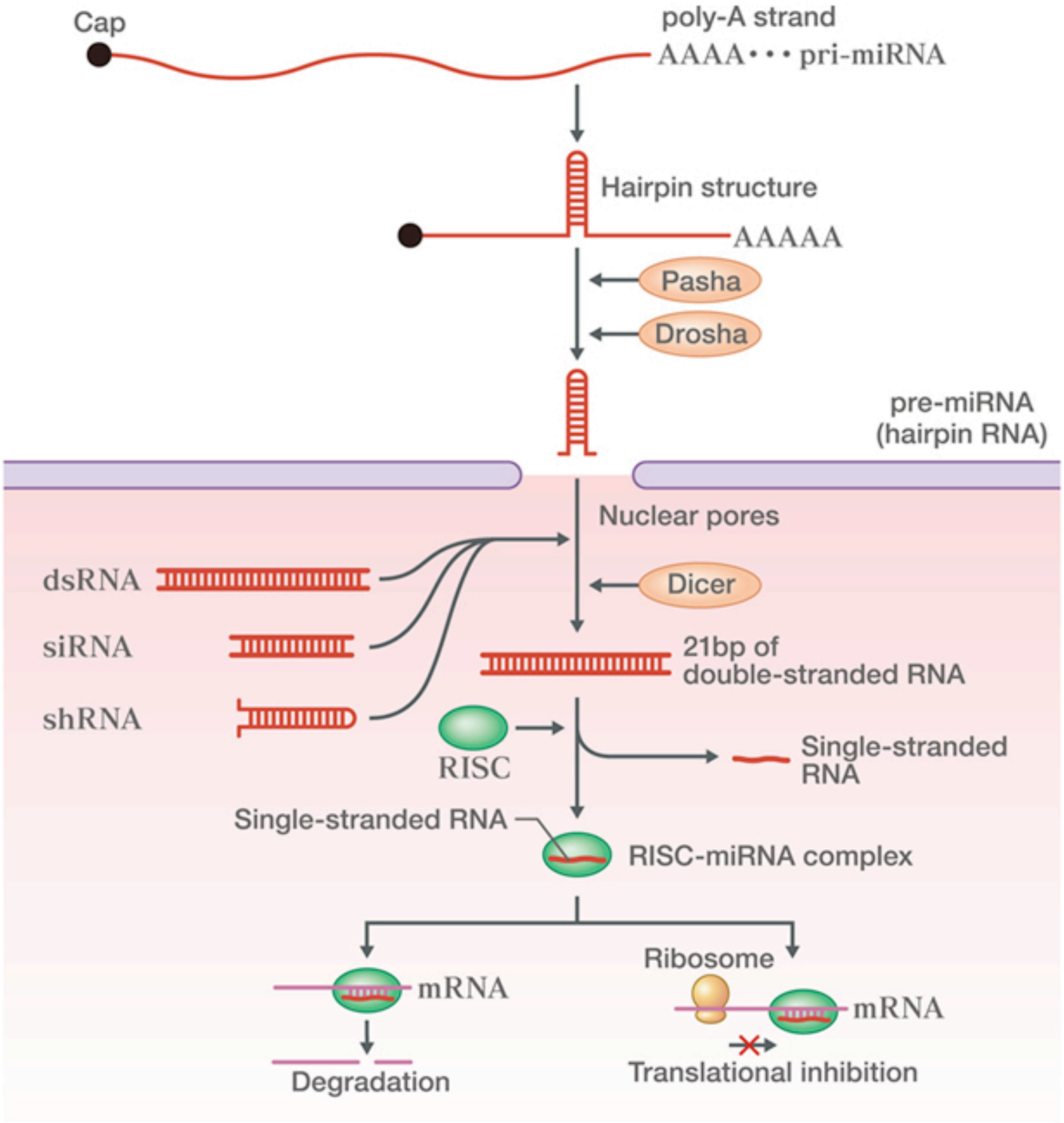
Ghildiyal and Zamore 2009

Functions of small RNAs

- Gene regulation
- Antiviral defense
- Regulation of host functions by viruses
- Immune system regulation
- Maintenance of stem cells
- Chromatin remodeling

- “Knockdowns”
- Antiviral therapy
- Anticancer therapy

- Genetic diseases



Using Small RNAs in the Lab: RNAi

– *C. elegans unc22* encodes muscle protein twitchin

- Mutants show uncoordinated “twitching” movement

Wild type worm

+ double-stranded *unc-22* RNA ⇨ twitching!

– RNAi can rapidly and efficiently silence a gene

– Specific

– Results from dsRNA

– Only small amounts required

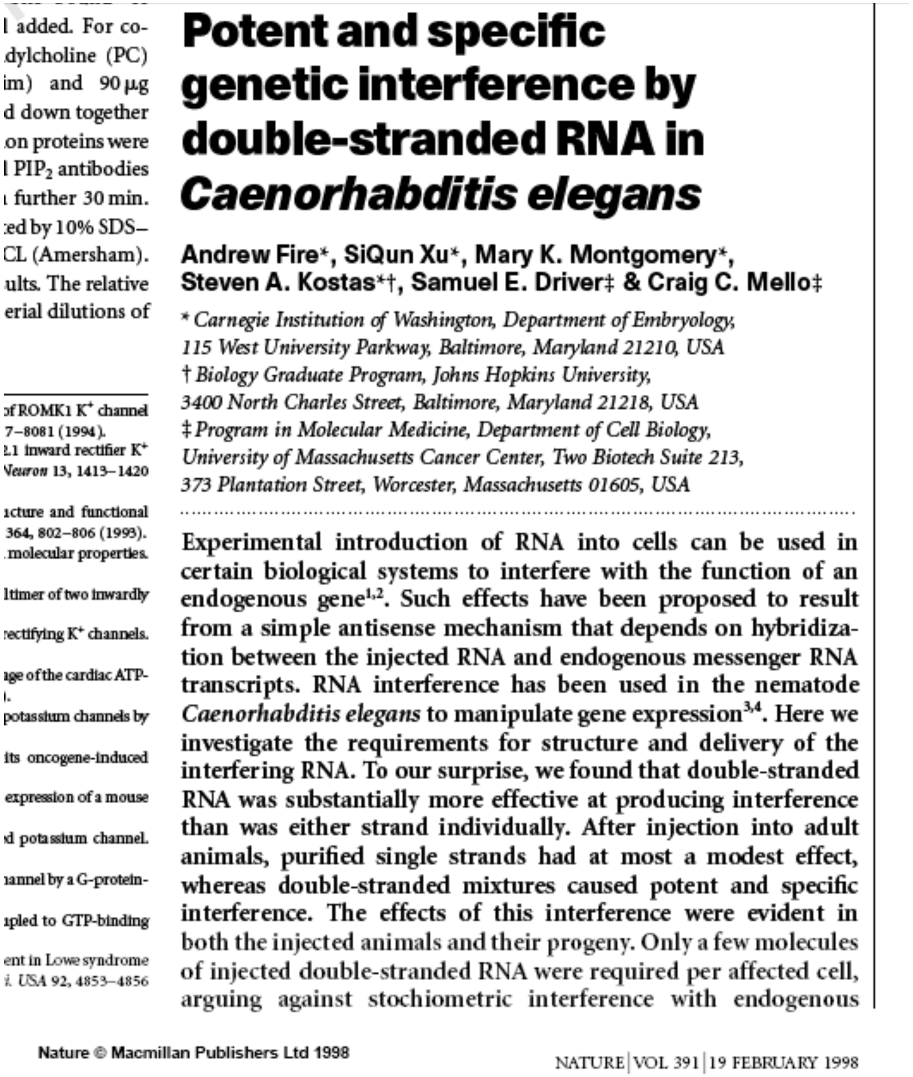
– Can inject or even feed the dsRNA

Don’t ignore weird results when you get them!

Nobel Prize in 2006: RNA interference (RNAi)

Andrew Fire and Craig Mello

"for their discovery of RNA interference –
gene silencing by double-stranded RNA"



l added. For co-
dylcholine (PC)
im) and 90 µg
d down together
on proteins were
l PIP_2 antibodies
ı further 30 min.
ed by 10% SDS–
CL (Amersham).
ults. The relative
erial dilutions of

of ROMK1 K^+ channel
7–8081 (1994).
2.1 inward rectifier K^+
Neuron 13, 1413–1420

ıcture and functional
364, 802–806 (1993).
molecular properties.

ltimer of two inwardly

rectifying K^+ channels.

ıge of the cardiac ATP-
).

potassium channels by

its oncogene-induced

expression of a mouse

d potassium channel.

ıannel by a G-protein-

ıpled to GTP-binding

ent in Lowe syndrome
i. USA 92, 4853–4856

Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*

Andrew Fire*, SiQun Xu*, Mary K. Montgomery*, Steven A. Kostas*†, Samuel E. Driver‡ & Craig C. Mello‡

* *Carnegie Institution of Washington, Department of Embryology, 115 West University Parkway, Baltimore, Maryland 21210, USA*

† *Biology Graduate Program, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218, USA*

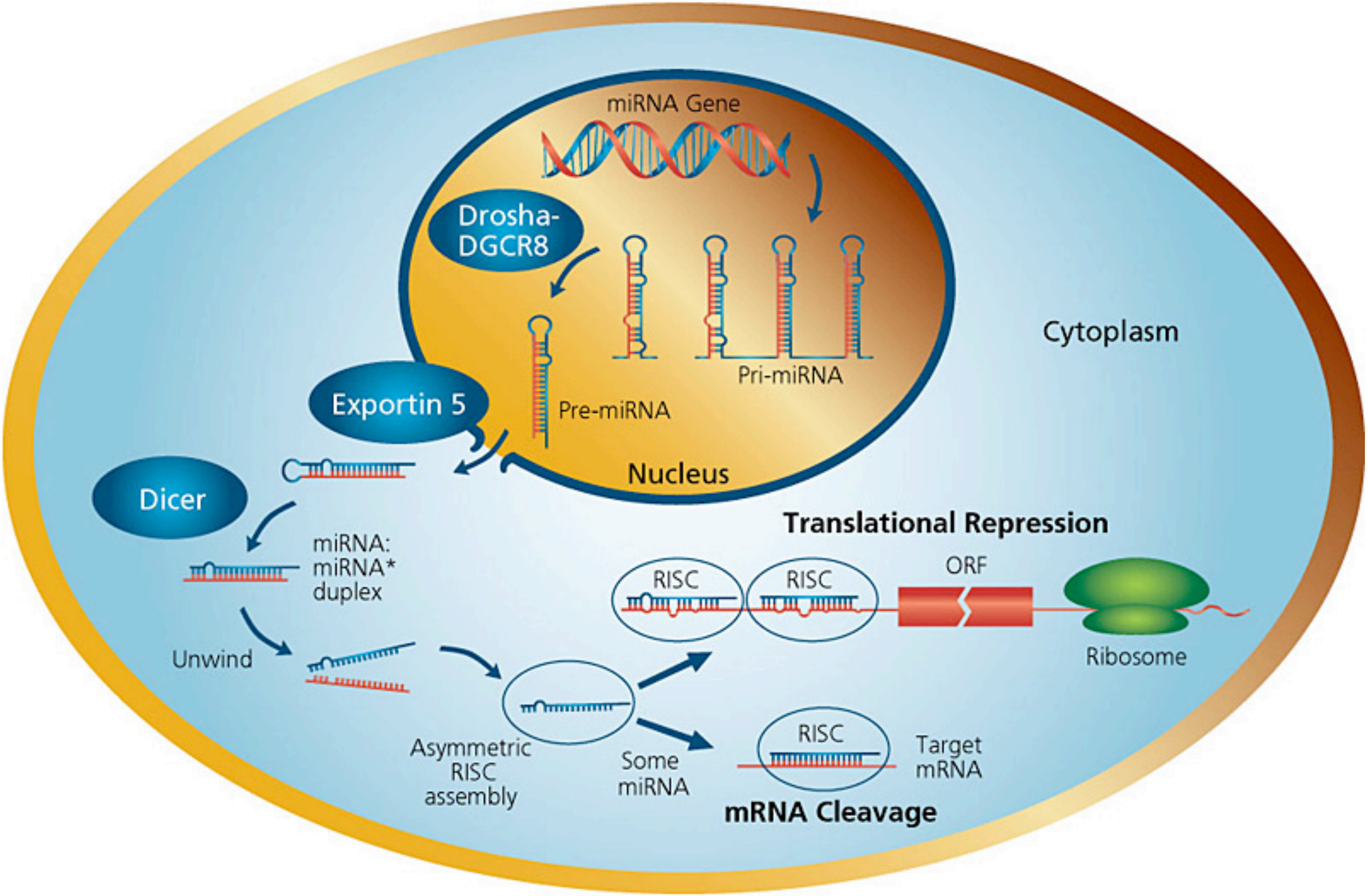
‡ *Program in Molecular Medicine, Department of Cell Biology, University of Massachusetts Cancer Center, Two Biotech Suite 213, 373 Plantation Street, Worcester, Massachusetts 01605, USA*

Experimental introduction of RNA into cells can be used in certain biological systems to interfere with the function of an endogenous gene[1,2]. Such effects have been proposed to result from a simple antisense mechanism that depends on hybridization between the injected RNA and endogenous messenger RNA transcripts. RNA interference has been used in the nematode *Caenorhabditis elegans* to manipulate gene expression[3,4]. Here we investigate the requirements for structure and delivery of the interfering RNA. To our surprise, we found that double-stranded RNA was substantially more effective at producing interference than was either strand individually. After injection into adult animals, purified single strands had at most a modest effect, whereas double-stranded mixtures caused potent and specific interference. The effects of this interference were evident in both the injected animals and their progeny. Only a few molecules of injected double-stranded RNA were required per affected cell, arguing against stochiometric interference with endogenous

3 of the Categories of Small RNAs

	Micro RNAs	Small interfering RNAs	Piwi-interacting RNAs
Description	miRNAs	siRNAs	piRNAs
	~22 nt	20-25 nt	26-31 nt
	ssRNA precursor, hairpin	dsRNA precursor, cut up	ssRNA precursor
	DICER-dependent		DICER-independent
Distribution	Proks, Euks, & Viruses	Euks	Animals only
	Translation Inhibition		
	Cleave Complimentary RNAs		
		Chromatin Modification	

We can use what we know about miRNA biogenesis and function to search for them using bioinformatic tools



Small RNA Prediction

- miRNA secondary structure
  - “hairpin” structure and stability
  - >15 nt paired region, no internal hairpins
- search regions flanking known small RNAs
  - miRNAs are often found in clusters
  - cleavage sites for processing
- comparative genomics
  - many miRNAs are evolutionarily conserved
  - some miRNAs found in gene families
  - Databases: NONCODE, miRBase
- target sequences
  - Identify potential genes that may be silenced by the candidate miRNA

Small RNA prediction: Challenges

- Small!
- Untranslated
- Generated from a larger transcript
- May be encoded in introns or other “junk” sequences
- Lack consensus sequence clues because recently discovered

Part II

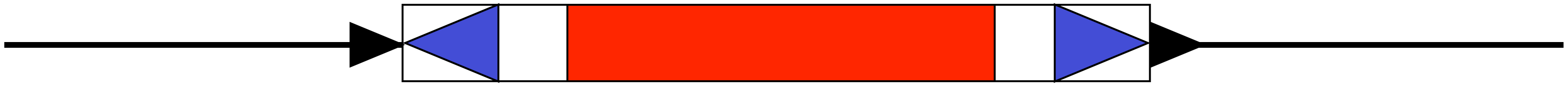
Transposable Elements

A TE is a piece of DNA that is, or once was, capable of moving or replicating and reinserting in the genome.

Other names: Mobile elements, selfish DNA, genomic parasites.

Features: Common, mobile, potentially replicative.

Types: Many

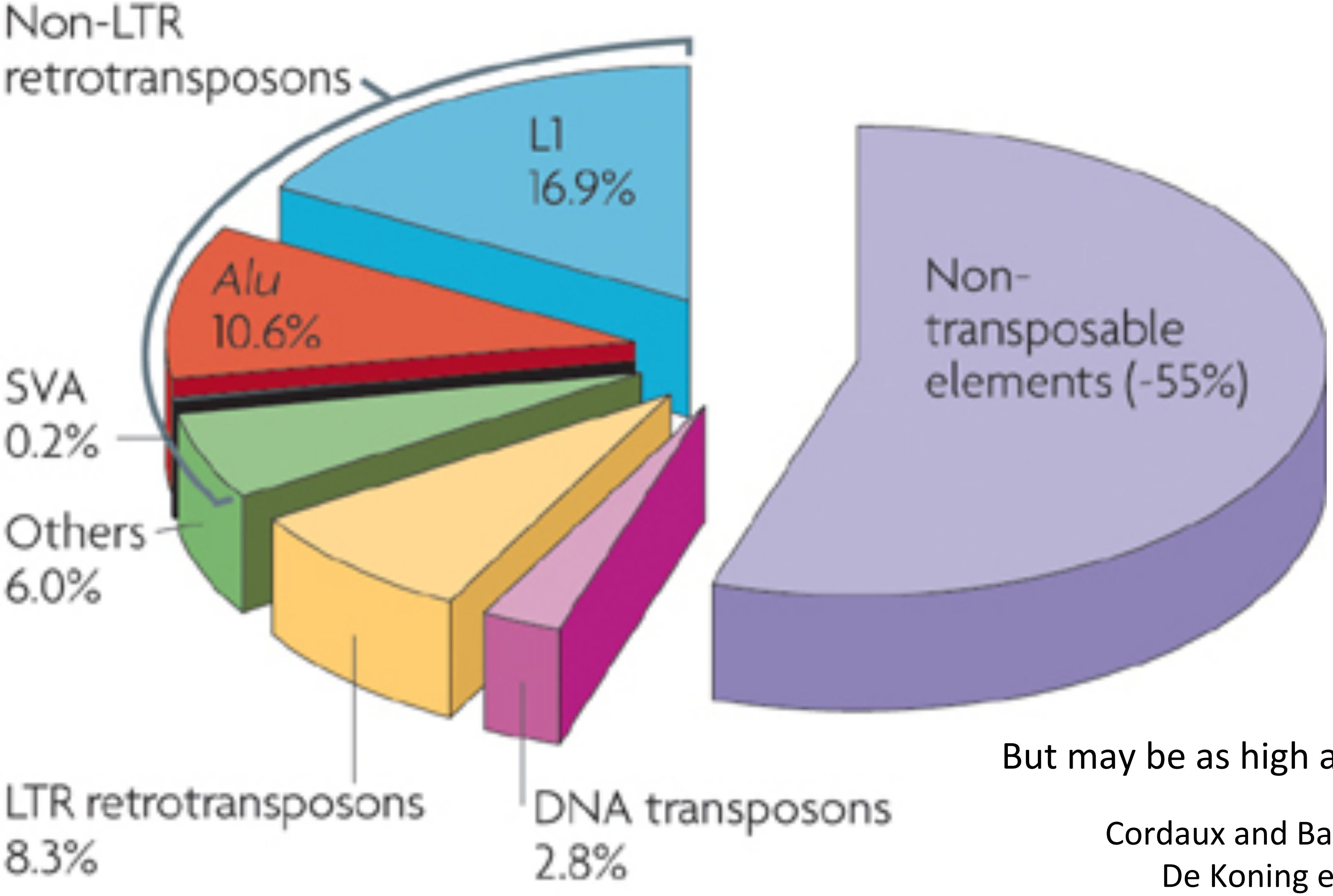


Can resemble genes (ORFs, sometimes introns)

But often include many unique motifs (inverted repeats, direct repeats)

Sometimes found in clusters, often more abundant in heterchromatin

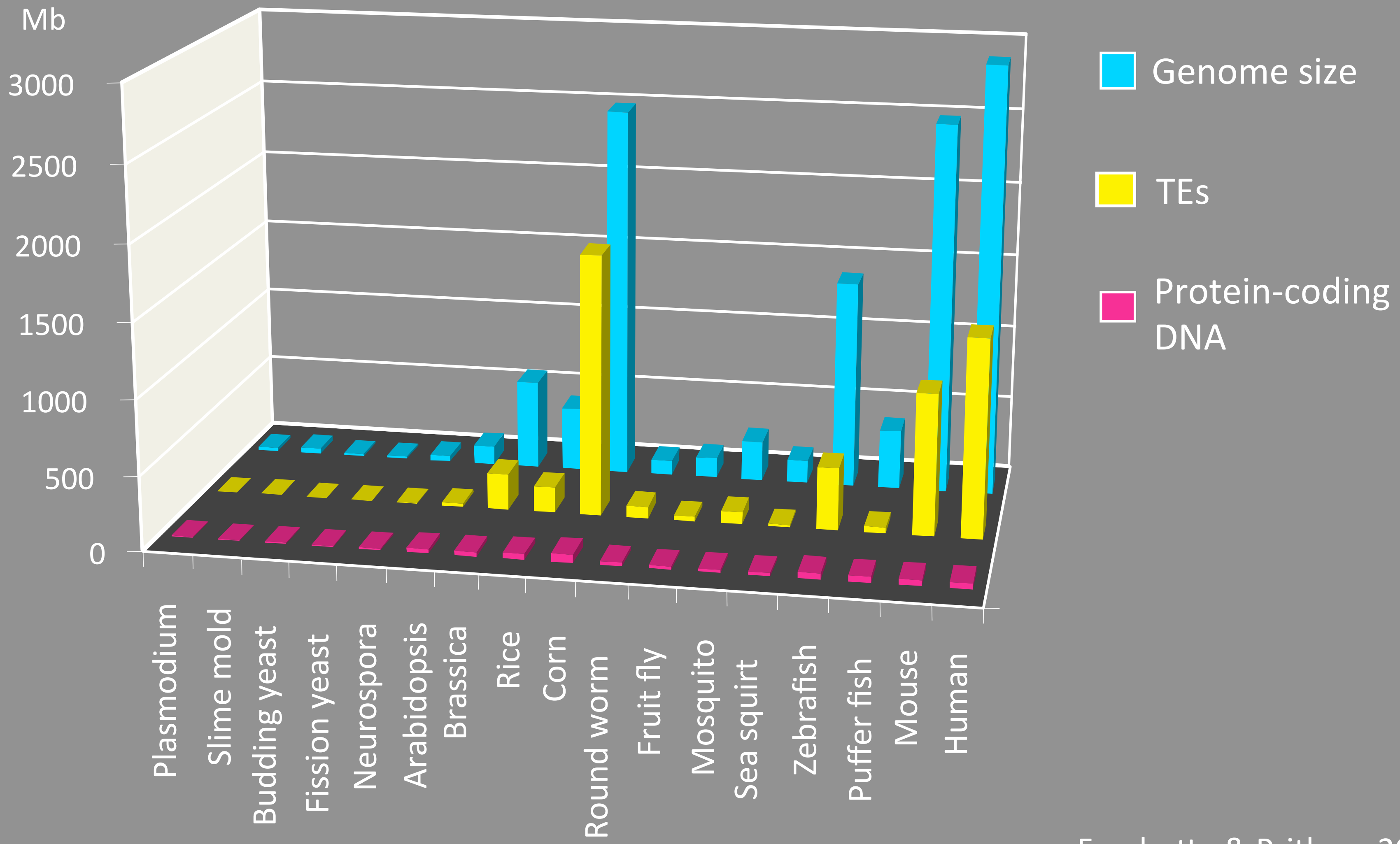
How common are they?



But may be as high as 66%!

Cordaux and Batzer 2009
De Koning et al. 2012

TEs, not genes, explain genome size differences across species



Feschotte & Pritham 2006

Understanding the quantity and distribution of TEs is critical to understanding both their positive and negative impact on the genome

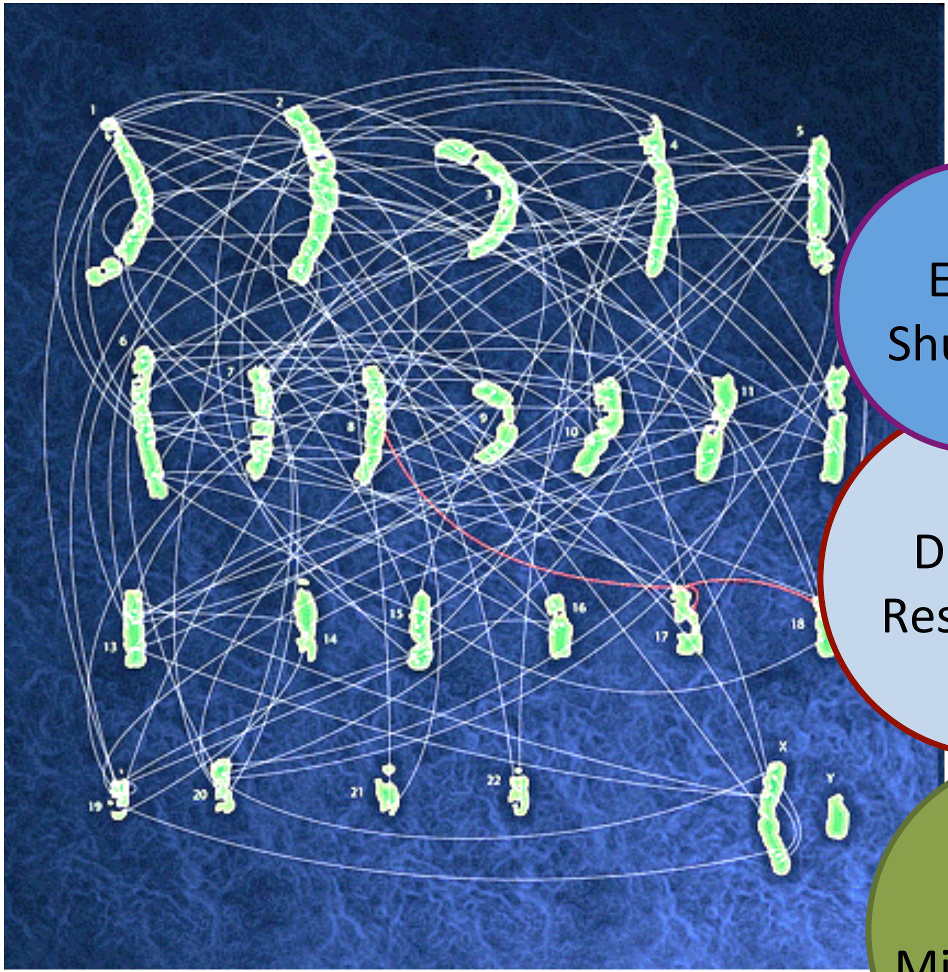


Image courtesy of Bang Wong, Jinchuan Xing, and Mark A. Batzer

Gene Regulation

Chromosome Trafficking

Exon Shuffling

Alternative Splicing

Development

Disease Resistance

Telomere Elongation

Genome Enlargement

Interrupting Genes

Gene Misregulation

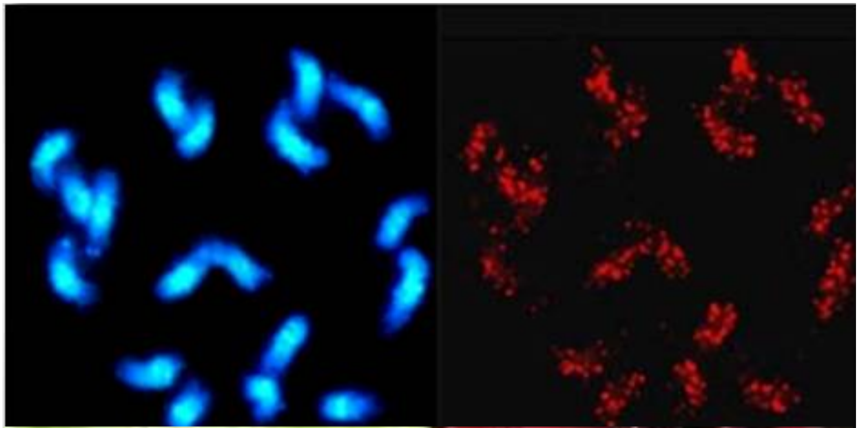
Nonhomologous recombination

So, there are many reasons we want to find TEs

(and there are many programs out there for finding them!)

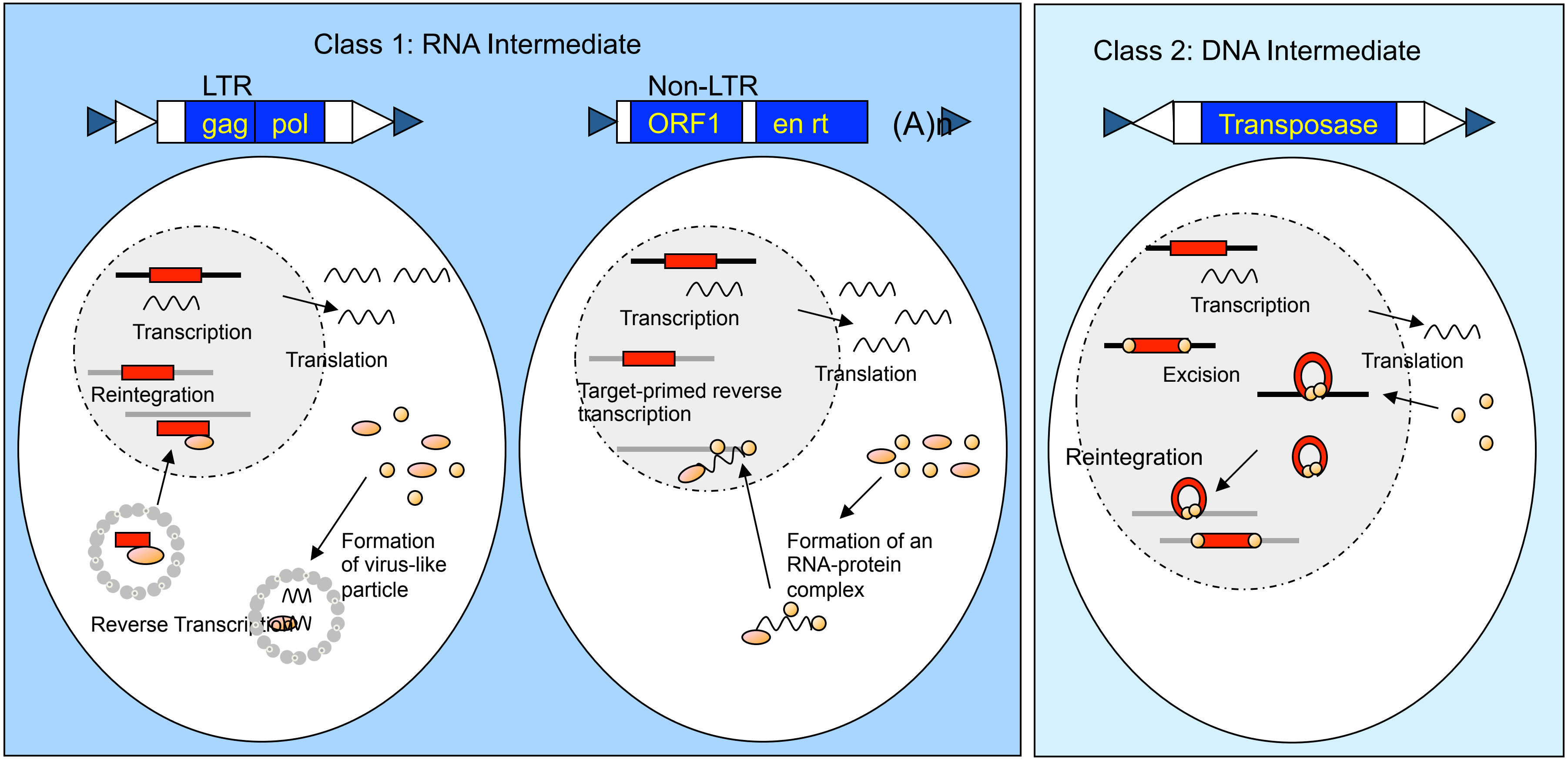
- Cytogenetic techniques
  - Staining
  - FISH

    (difficult to quantify!)
- Bioinformatic techniques
  - RepeatMasker
  - RepeatScout
  - RepeatExplorer
  - CENSOR
  - MGE-Scan (LTR and non-LTR)
  - BLAST
  - Many others

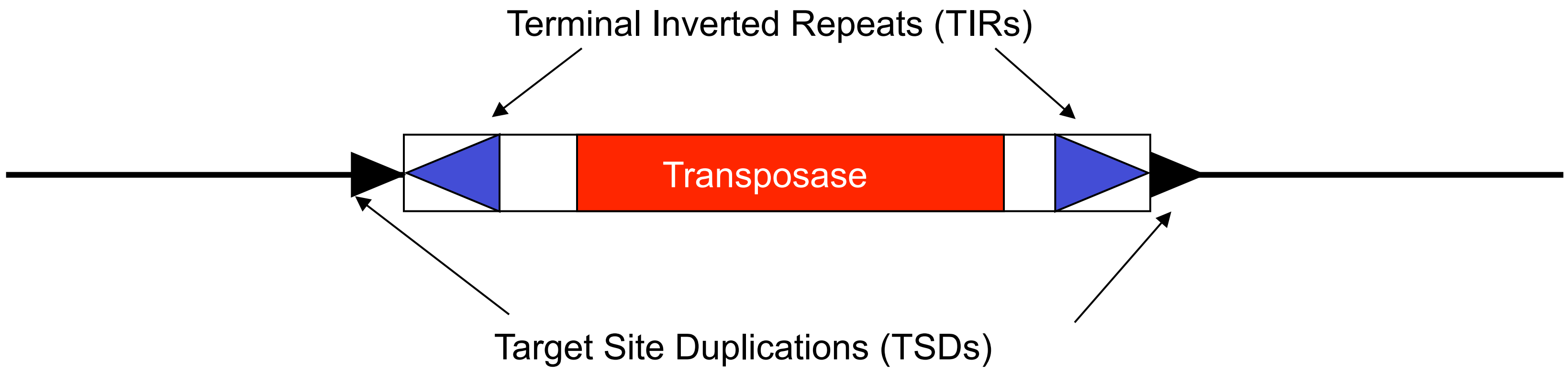


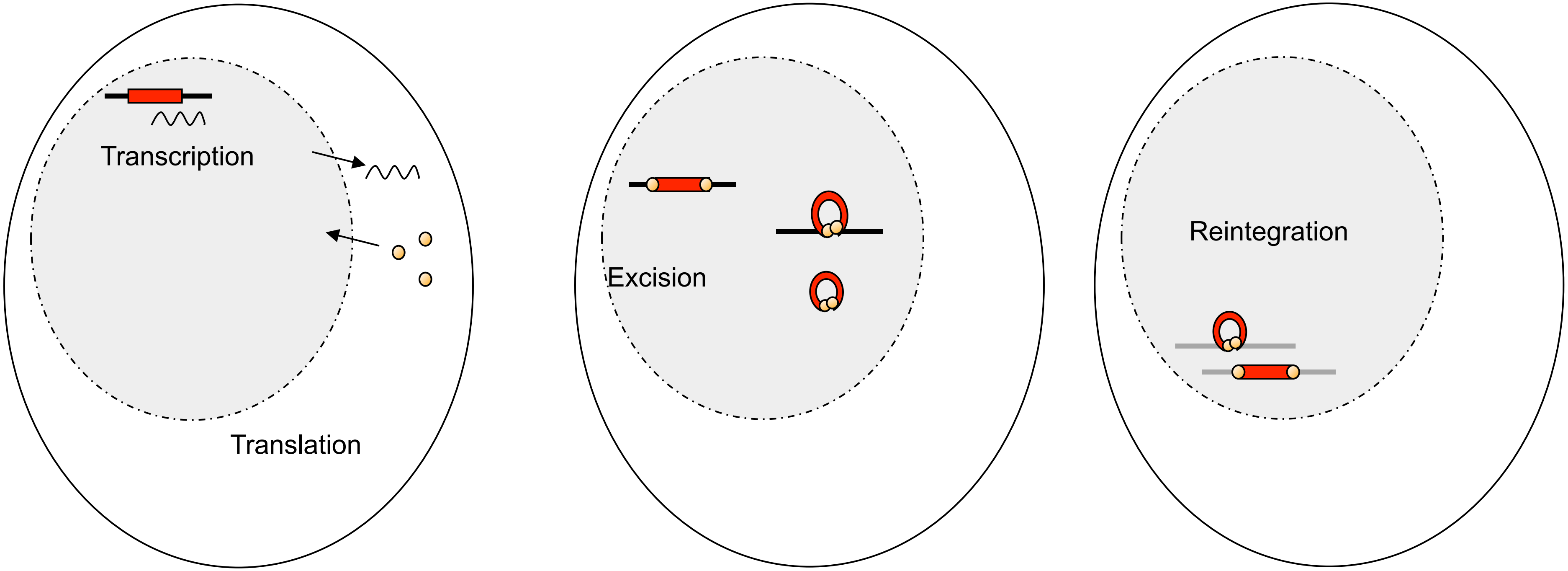
Menzel et al. 2006

Homology-based Searching:
2 Main Classes of TEs Characterized by Different Proteins

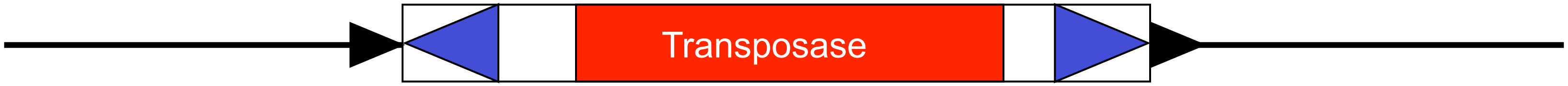


An Example of How A TE Moves


Terminal Inverted Repeats (TIRs)
Transposase
Target Site Duplications (TSDs)

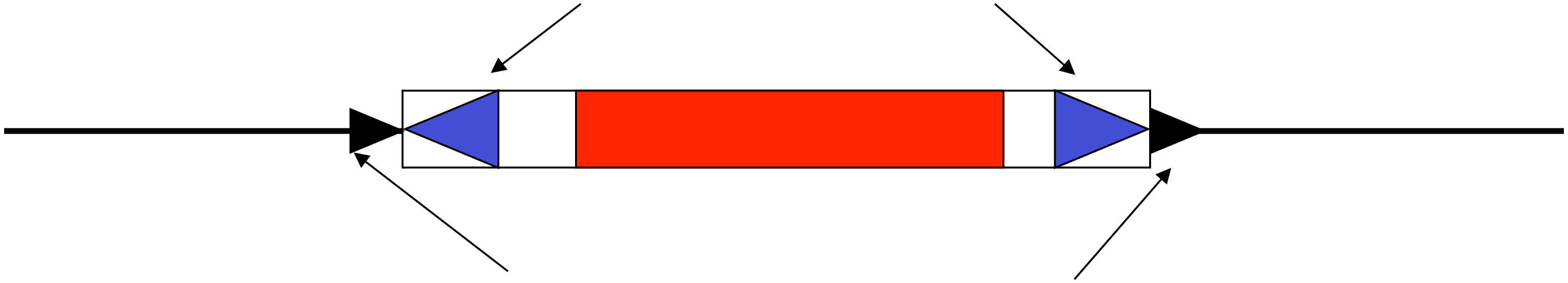


Homology-based Searching



Davies et al. 2000

Motif-Based Searching



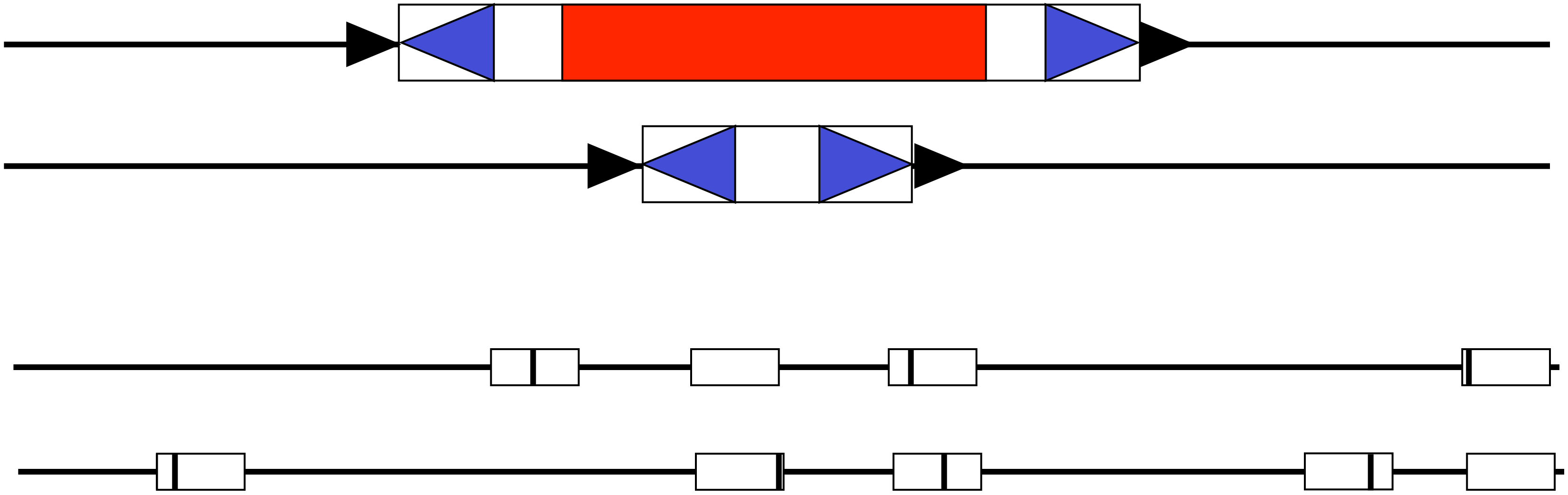
Terminal Inverted Repeats (TIRs)

Target Site Duplications (TSDs)

Binding Sites
Integration Sites
Length
Copy Number

Davies et al. 2000

When TEs Replicate → Family of TEs



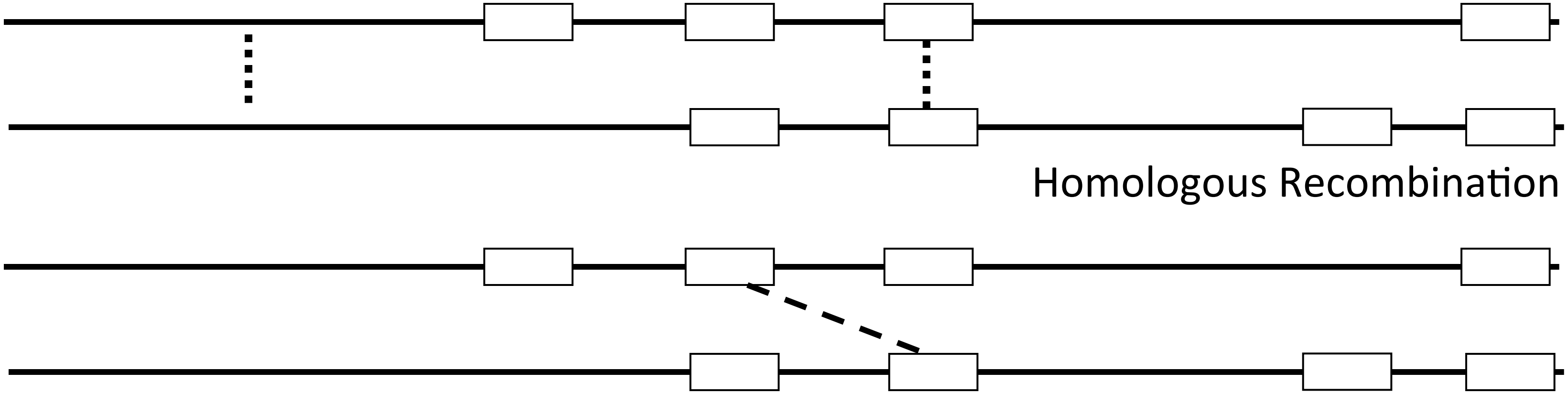
TEs accumulate mutations over time.

Copies can be used to estimate the founder (mobile) element.

All the copies of a particular TE type in a genome are referred to as a “TE family”

```
          AGTTAGATCA
          AGCTAGATCT
          ACTTAGATCT
          AGTTTGAGCT
          AGTTAGATCT
          AGTGAGATCT
          CGTTAGATCT
          AGTTAGATGT
Consensus AGTTAGATCT
```

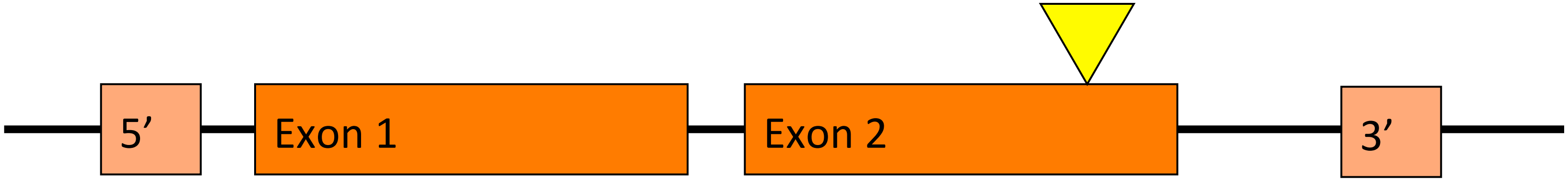
Repetitive Sequence Throughout the Genome Can lead to Non-Homologous Recombination



Homologous Recombination

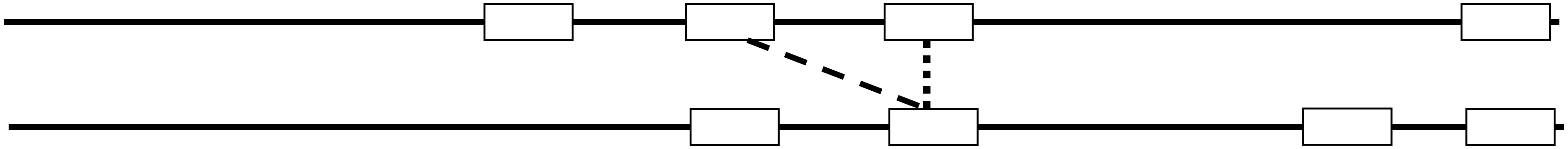
Non- Homologous Recombination

Indirect Costs: Increased risk of non-homologous recombination and therefore indels



Direct Costs: Increased risk of interrupting genes

If Indirect costs are important….



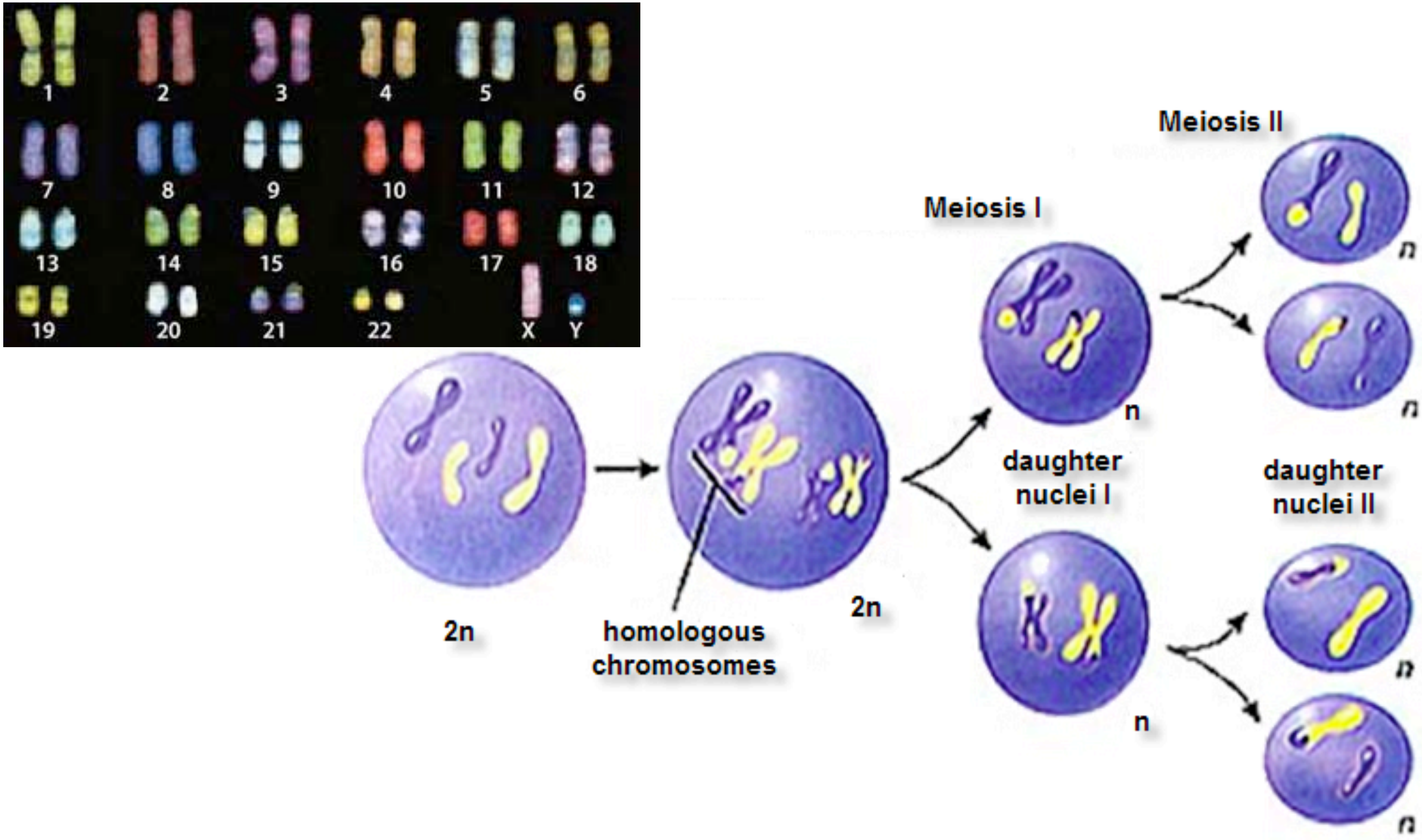
….we would expect TEs to accumulate
in regions of low recombination
because the risk of non-homologous recombination
would also be lower.

How can we test this by looking for TEs using bioinformatic methods?

Compare TE levels in recombining and non-recombining regions!

Which region of the human genome does not recombine during meiosis?

(or recombines the least?)



So, let's use CENSOR to compare how many TEs there are on the Y versus on the X versus versus on a randomly selected autosome and see if low recombination areas accumulate more TEs than expected.

Chromosome Type	Accession #	Length of BAC clone (bp)	Number of element fragments	Length of repetitive DNA in basepairs	*Percent of the BAC composed of repetitive sequence*
Autosome	**AC005690.8**				
X	**AC233302.2**				
Y	**AC244170.3**				