Contextualization and Classification of Natural Genetic Engineering Operators

James A. Shapiro
University of Chicago
jsha@uchicago.edu
http://shapiro.bsd.uchicago.edu

Cells as sophisticated, powerful cognitive and computing entities

- Need for reliability/predictability in cellular reproduction and behavior
- Complexity of cellular functions, cell cycle
- Regulation to coordinate innumerable biochemical and biomechanical events
- Execution of life-cycle and developmental programs
- Responses to changing conditions
- Responses to errors and damage

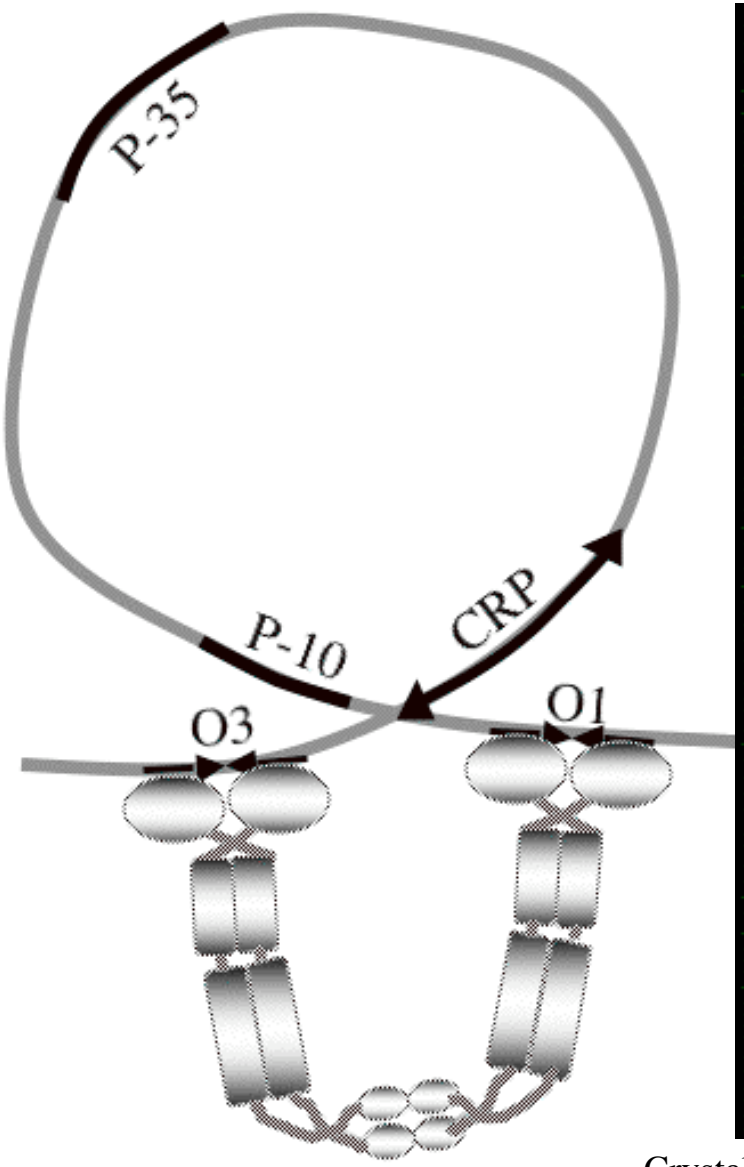
Role of DNA in cellular information processing - I. Information storage at various time scales

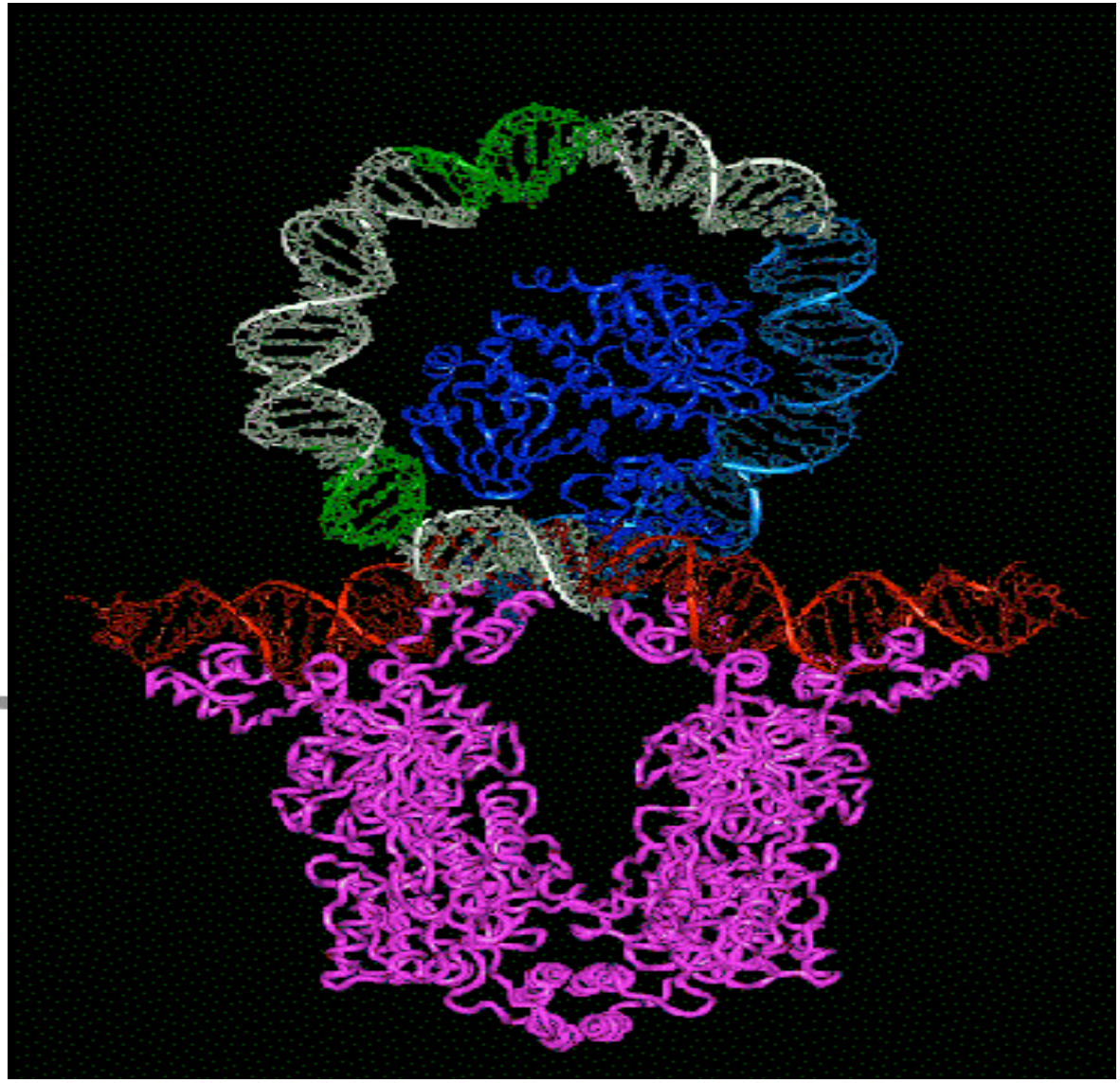
- Long-term/DNA sequence structure – changes by point mutation, DNA rearrangement
- Multiple cell generations/Epigenetic modification and complex formation – erasable and resettable
- Within cell cycle/nucleoprotein complex formation – quickly modified in response to transient signals

Role of DNA in cellular information processing - II. DNA organized as a storage medium

- Necessary interactions with other molecules to carry out all important functions ==> combinatorial nature of information processing (genetic loci as microprocessors with multiple molecular connections)
- Distinction between genome formatting and data files; hierarchical nature of genome organization
- Multi-step nature of reading genomic data files with regulation/control at each step

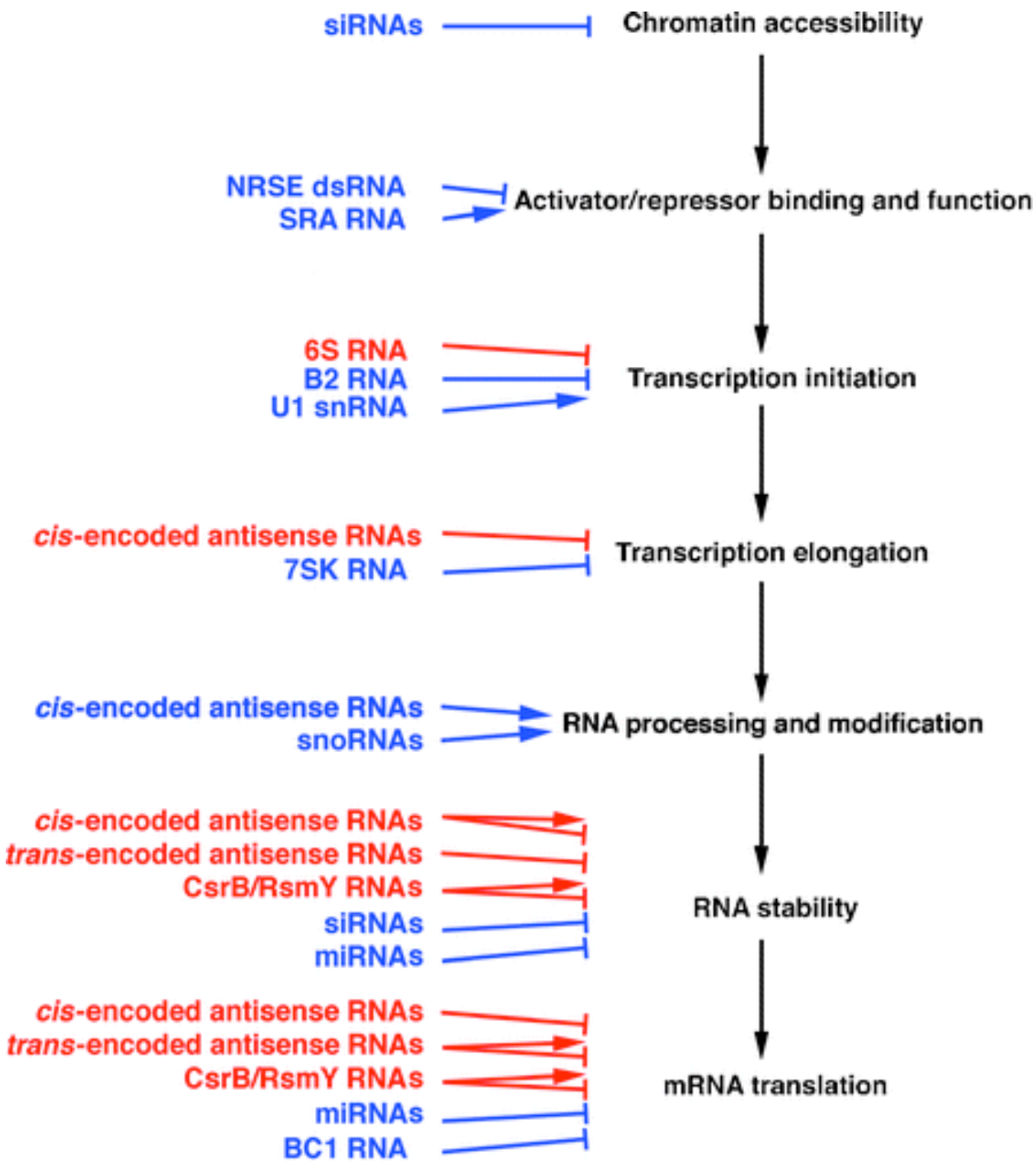
LacI Repressor Binding to *lacO* Sites and DNA Looping - Cooperativity Plus





Crystal Structure of the Lactose Operon Repressor and Its Complexes with DNA and Inducer. Mitchell Lewis, et al. Science, Vol. 271(Mar. 1, 1996), pp. 1247-1254.

Small RNA regulation at steps in expression of genomic data files



Storz G, Altuvia S, Wassarman KM. An abundance of RNA regulators. Annu Rev Biochem. 2005 74: 199-217.

Molecular basis of DNA sequence change

- Templated polymerization
  - DNA template
  - RNA template
- Untemplated polymerization
  - Terminal transferases
  - Mutator/trans-lesion polymerases
- Phosphodiester bond cleavage and resealing in novel combinations
  - Single-strand (SS)
  - Double-strand (DS)
  - DNA-DNA
  - RNA-DNA

Point mutation (local changes)

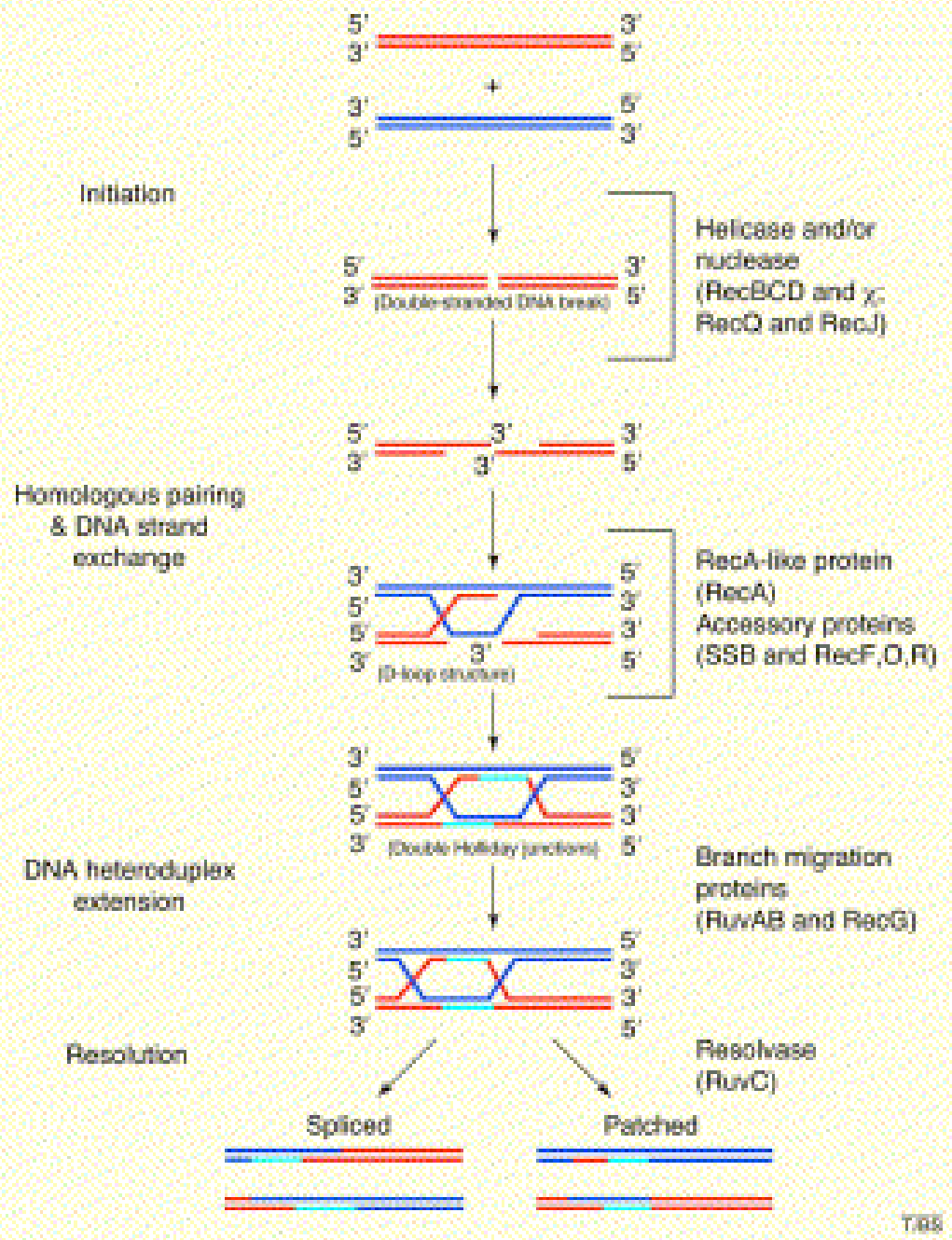
- Base substitutions
- Frameshifts (insertion or deletion of a few nucleotides)
- Microdeletions

- Mutator polymerases
- Topoisomerases
- Nucleotide modifying enzymes

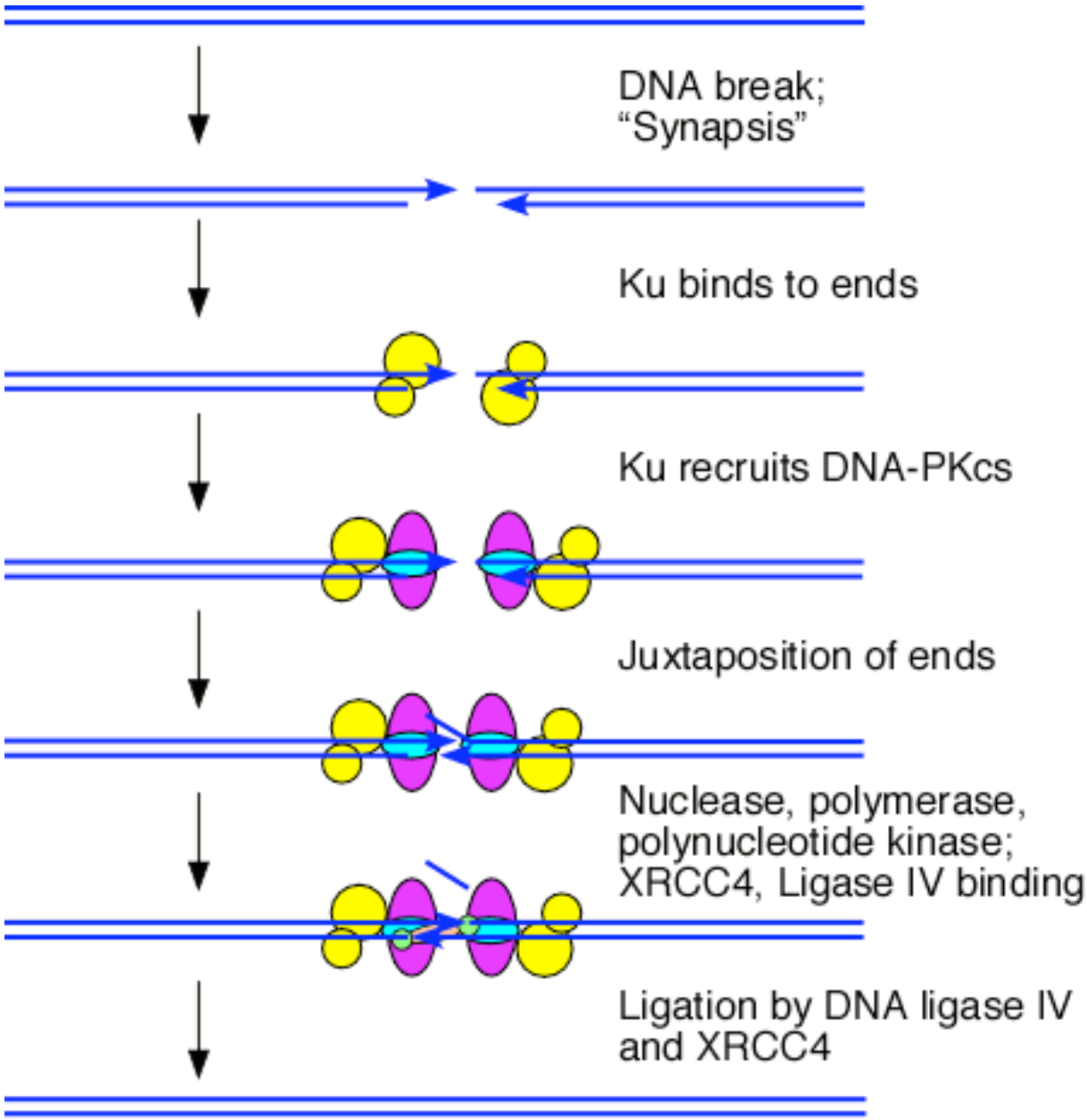
- Protein engineering by somatic hypermutation (transcription, deamination)

Homologous exchange

- Initiated by DNA breaks (can be targeted)
- Multiple biochemical functions required (nucleases, helicases, polymerases, ligases, strand-pairing proteins)
- stimulatory sequences, such as Chi
- Alternative single-strand annealing mechanism (SSA)

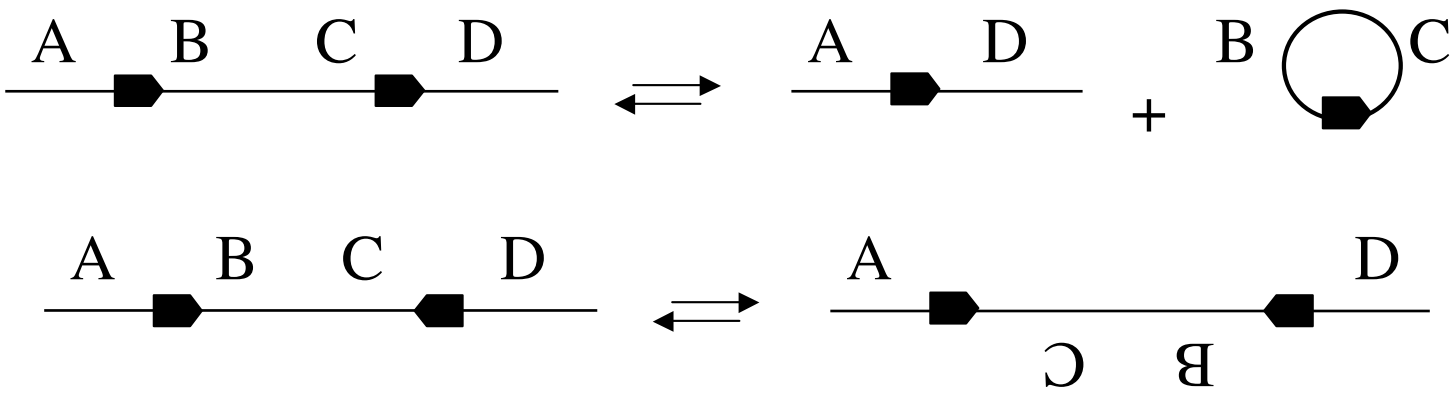


Non-homologous end-joining (NHEJ)



http://asajj.roswellpark.org/huberman/DNA_Repair/dsbreak.html

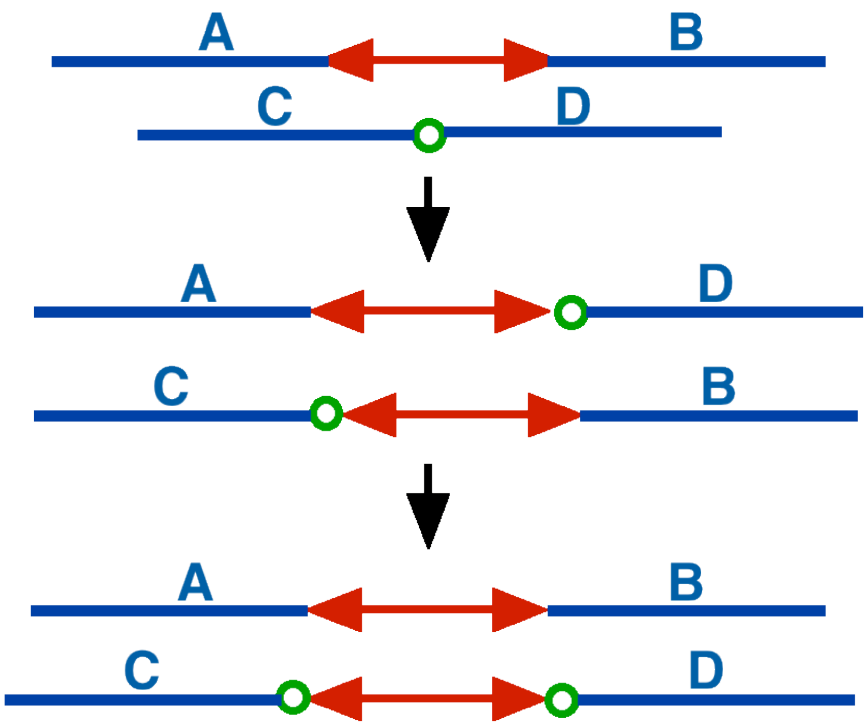
Site-specific exchange



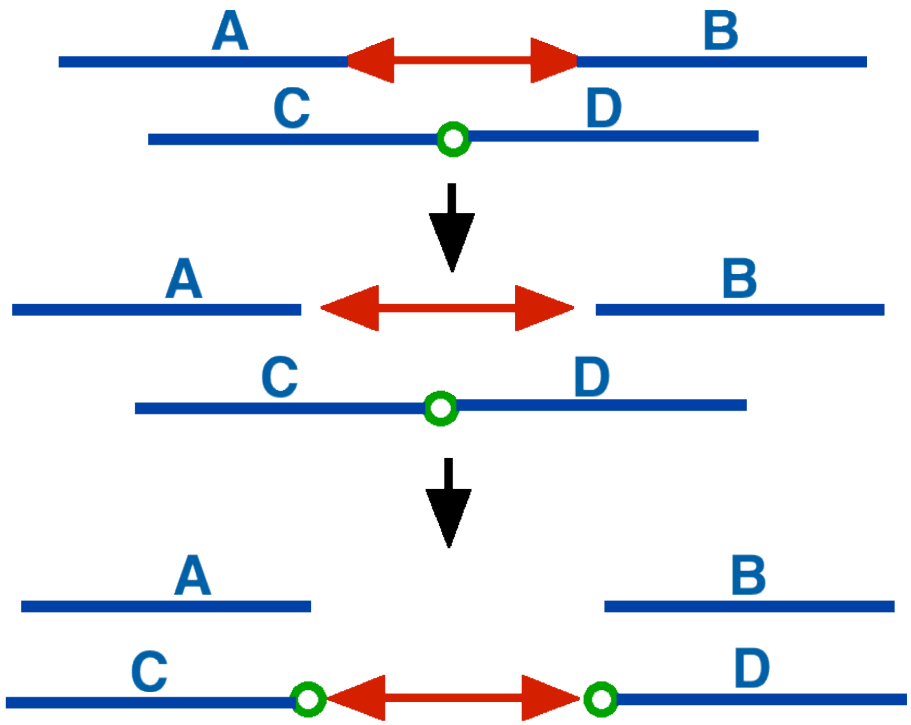
- Insertion and excision (viral genomes, antibiotic resistance and other coding sequence cassettes, genomic “islands”)
- Separation of double DNA circles into single circles for distribution to daughter cells
- Control of expression by inversion of transcription signals
- Control of expression by inversion of coding sequences
- Protein engineering by inversions within coding sequences
- Direction of exchange regulated by recombinase proteins
- Protein-DNA intermediates

DNA transposition

Replicative

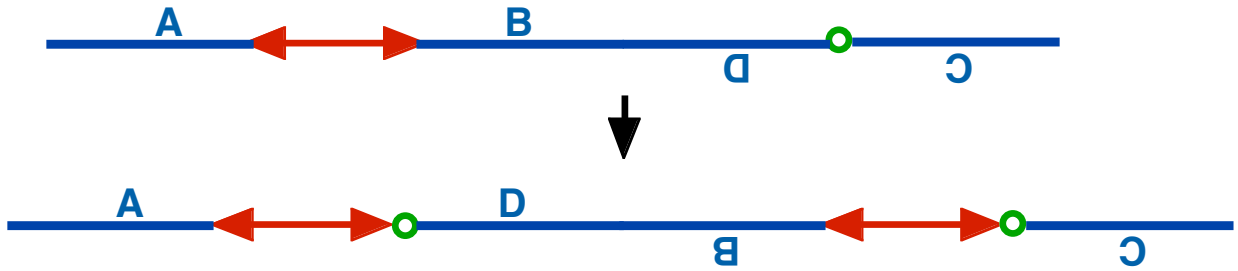


Cut-and-paste

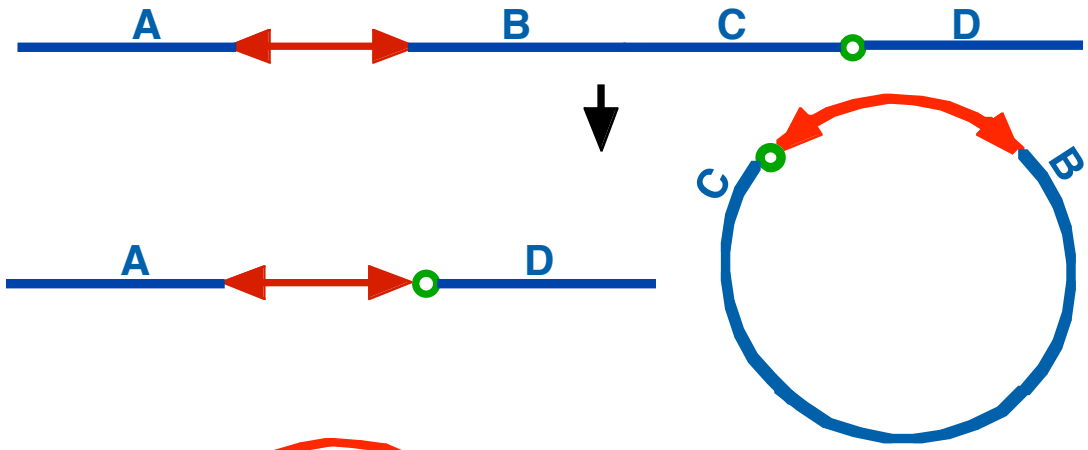


Replicative transposon rearrangements

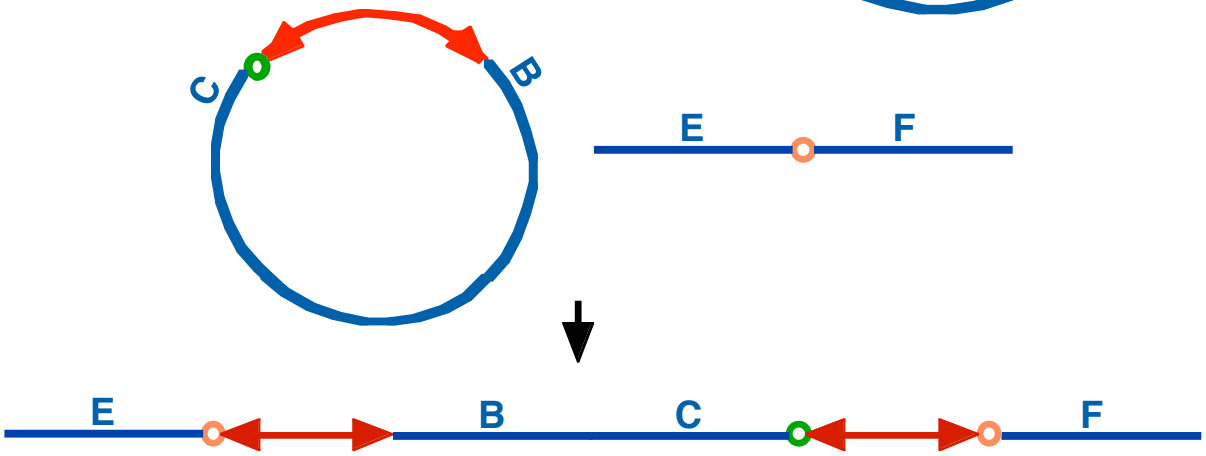
1. Inversion



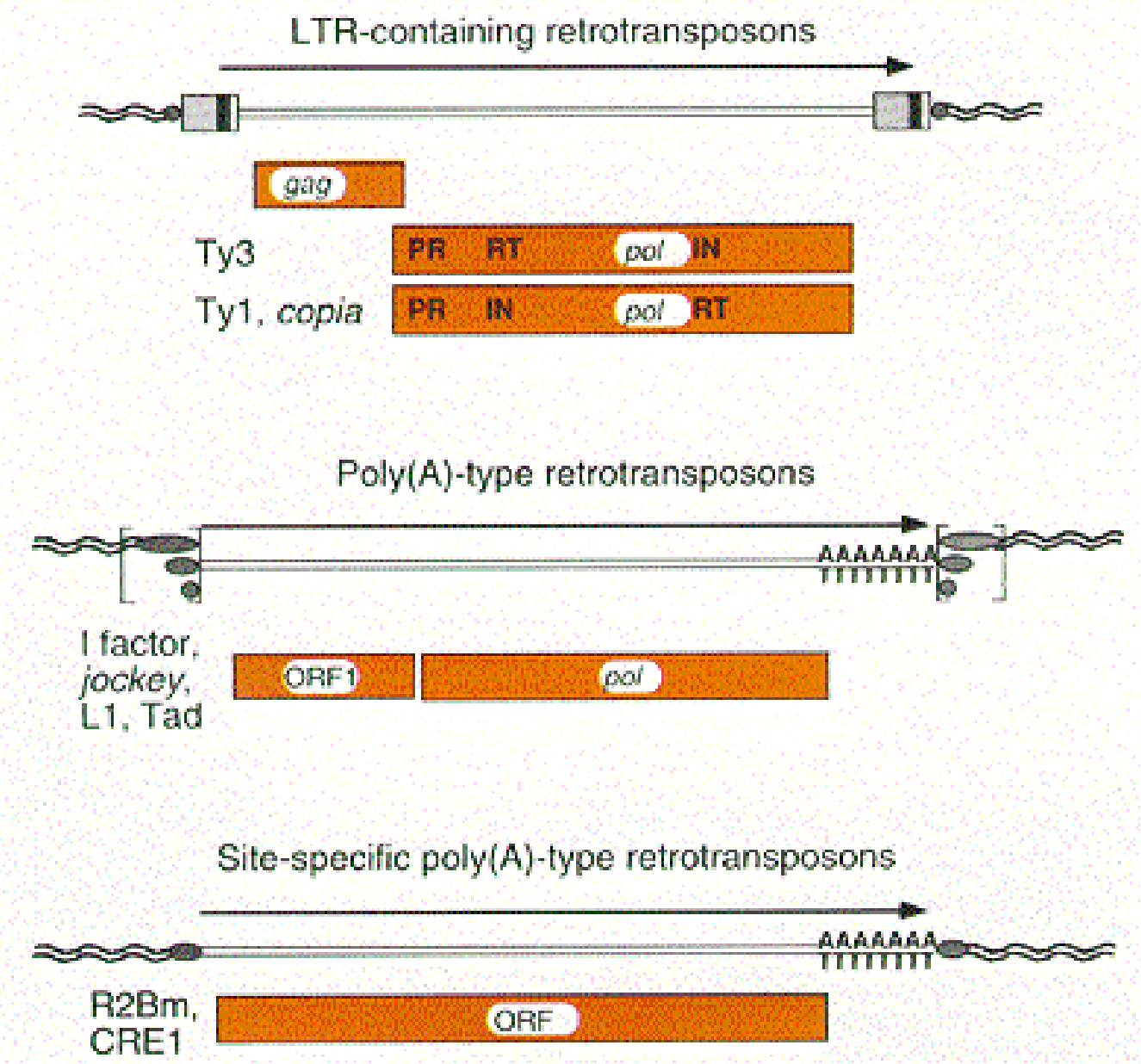
2. Deletion /excision



3. Fusion /insertion



Retro-transposons



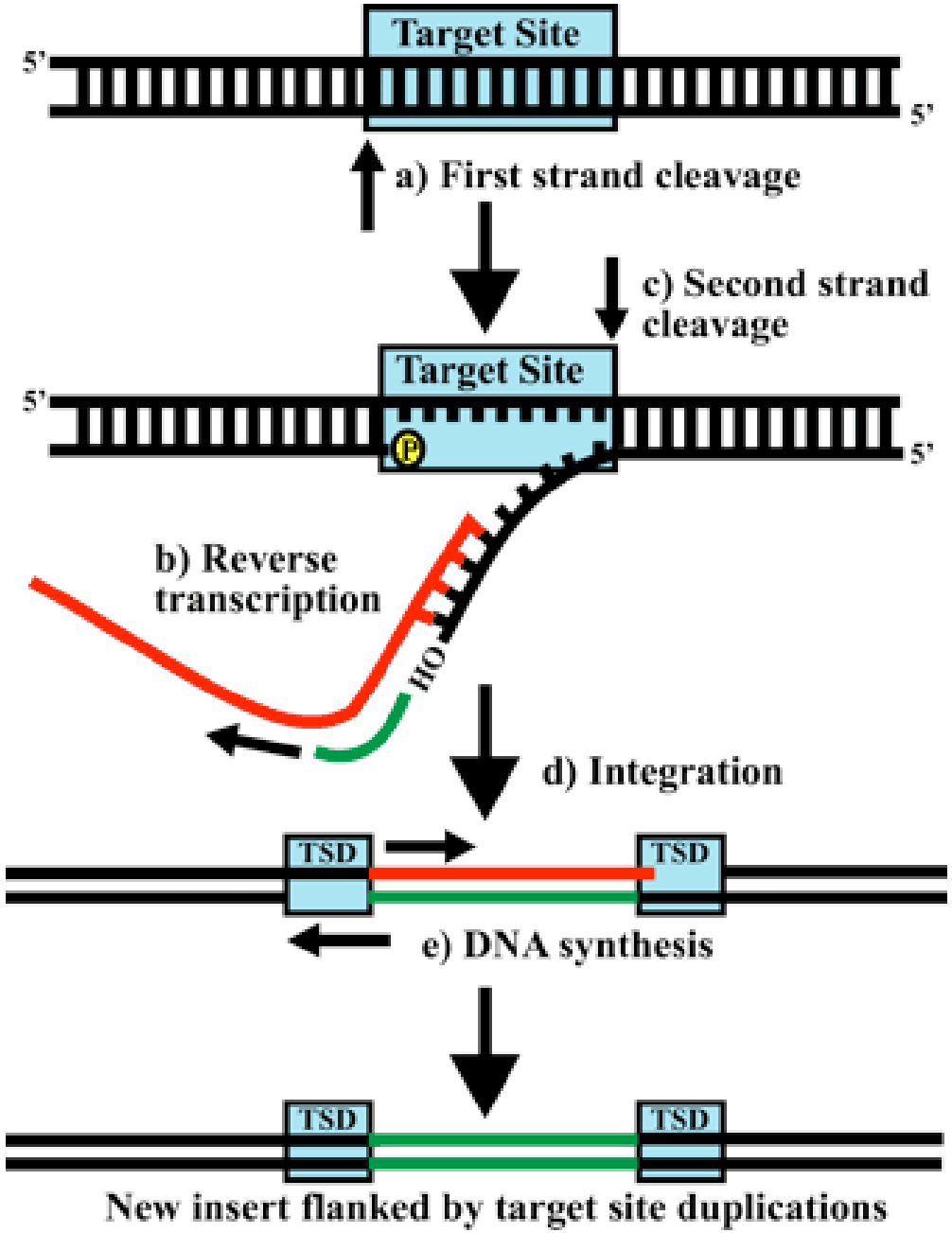
Retroviruses. Coffin, John M.; Hughes, Stephen H.; Varmus, Harold E. Cold Spring Harbor Laboratory Press; c1997.

Transposable Elements in the Human Genome

Classes of interspersed repeat in the human genome

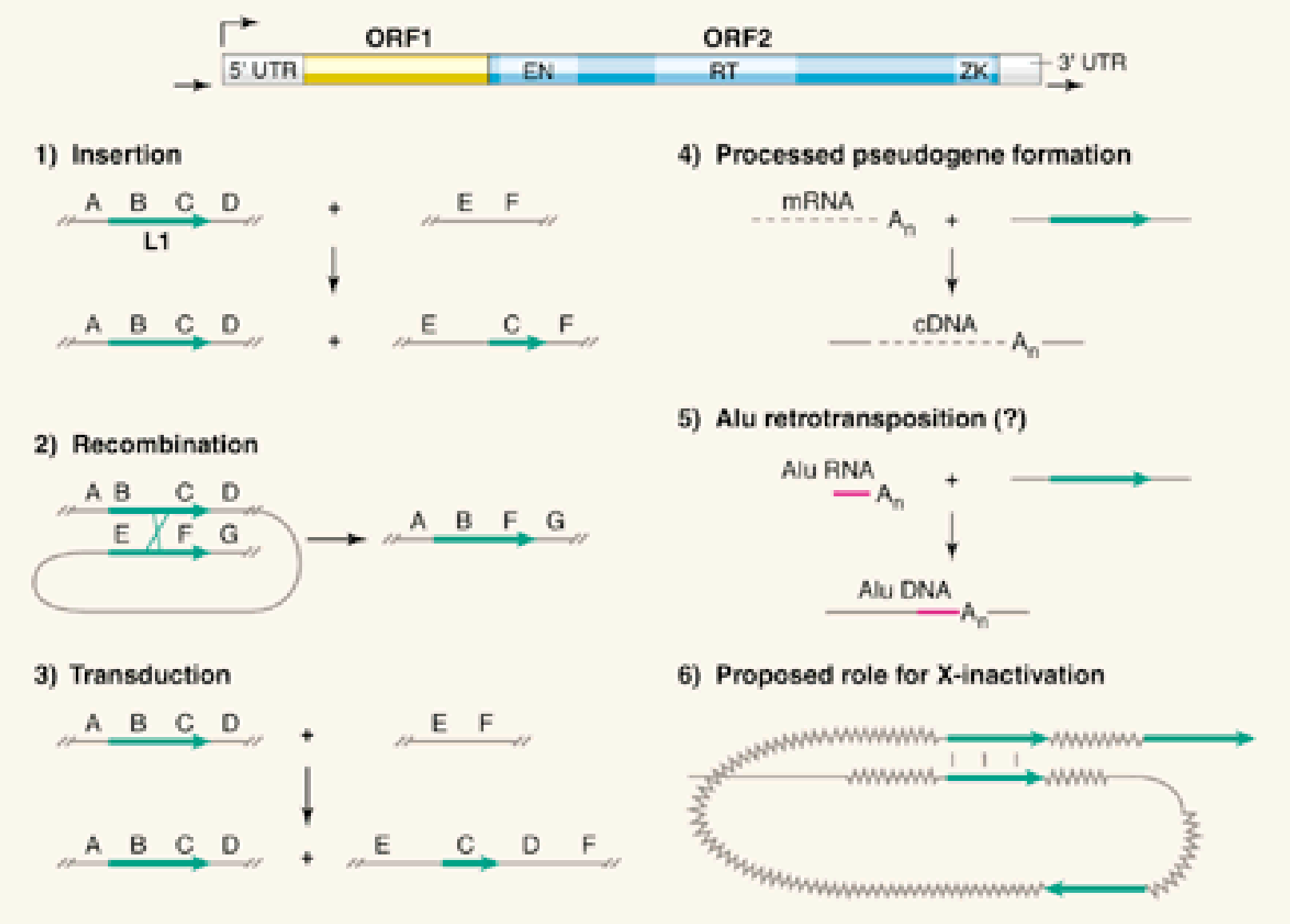
			Length	Copy number	Fraction of genome
LINEs	Autonomous	ORF1 ORF2 (pol) AAA	6–8 kb	850,000	21%
SINEs	Non-autonomous	A B AAA	100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous	gag pol (env)	6–11 kb	450,000	8%
	Non-autonomous	(gag)	1.5–3 kb		
DNA transposon fossils	Autonomous	transposase	2–3 kb	300,000	3%
	Non-autonomous		80–3,000 bp		

LINE Target-Primed Reverse Transcription



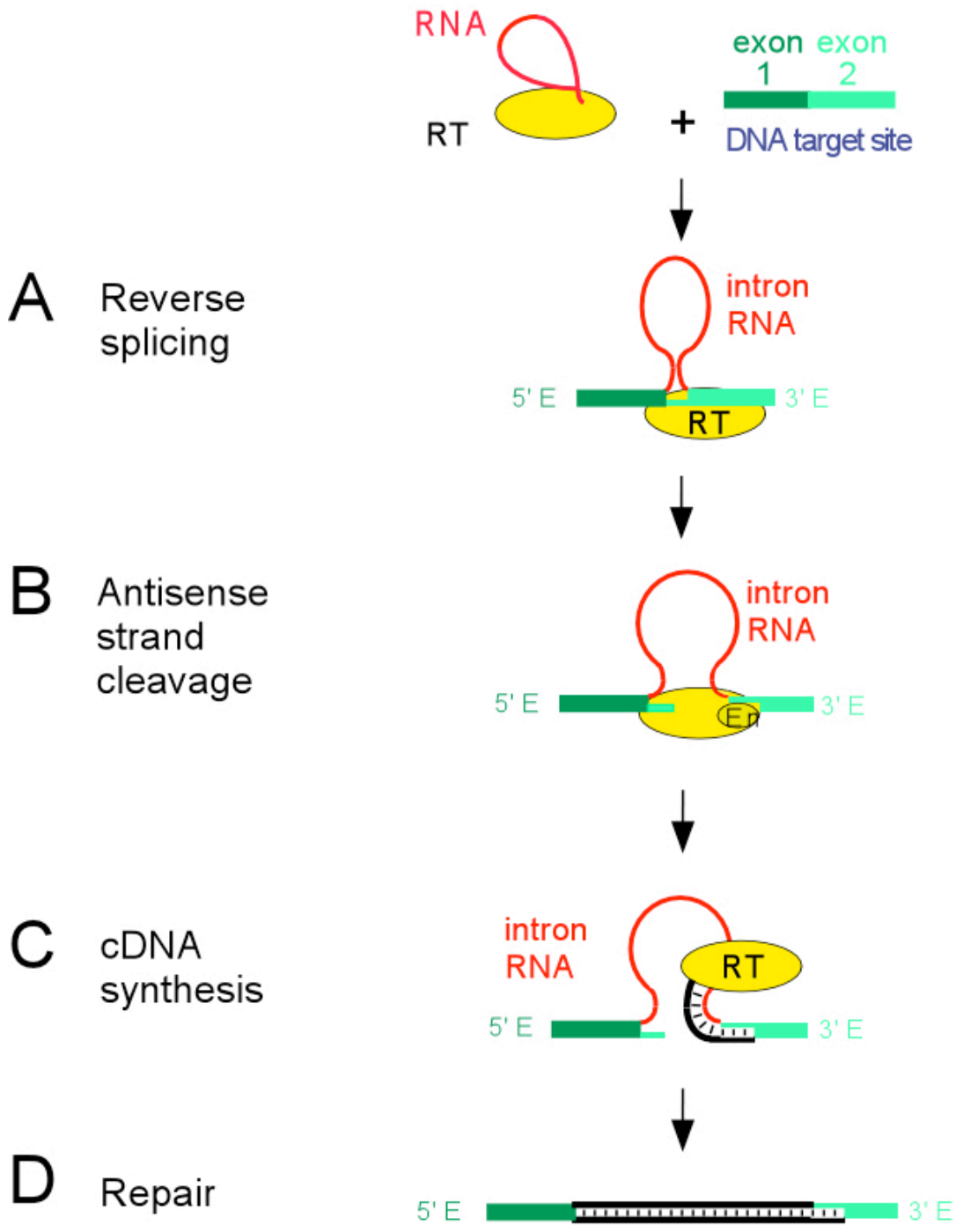
Sandra L. Martin and Frederic D. Bushman. Nucleic Acid Chaperone Activity of the ORF1 Protein from the Mouse LINE-1 Retrotransposon. Molecular and Cellular Biology, 2001, p. 467-475, Vol. 21

LINE element-mediated rearrangements



HH Kazazian. 2000. L1 retrotransposons shape the mammalian genome. Science 289, 1152-53

Reverse splicing



www.fp.ucalgary.ca/group2introns/mobility.htm

DNA transfer and uptake systems

- Viruses and virus-like particles
- Naked DNA extrusion and uptake in bacteria
- Conjugative transfer systems in bacteria
- DNA transfer into animal cells and chromosomal integration
- Interspecific and interkingdom transfer (Agrobacterium tumor formation in plants)

Control of Natural Genetic Engineering: Suppression and Activation

- Repressors, inhibitors, methylation control of
- Competence induction (Vibrio/chitin; B. subtilis peptides)
- Plant wound signals in Agrobacterium T-DNA transfer, retrotransposon activation
- “Stress” induction (UV, DS breaks, starvation, etc.)
- Hybrid dysgenesis in plants and animals
- RNAi, piwi RNAs

Control of Natural Genetic Engineering: Targeting within the genome

- protein sequence recognition (nucleases, recombinases, transposases)
- RNA base-pairing to DNA guide sequences (reverse splicing)
- Coupling to transcription
  - retrotransposon integration (protein-protein interaction)
  - transcription-dependent DS breaks)
- Coupling to chromatin (retrotransposon integration)
- P-element “homing”

Consequences of natural genetic engineering

- Protein engineering by domain/exon swapping and accretion
  - Within life cycle (multiple targeted rearrangements)
  - Evolutionary time scale (PackMuLEs, helitrons, LINEs)
- Accretive evolution of complex signals for transcription and other steps in data file expression
- chromatin formatting of extensive genomic segments encompassing multiple genetic loci (roles of TEs, repeats)
- Accretive evolution and amplification of higher-order structures (integrons, TEs, Hox complexes)
- Evolution of synteny blocks (segments with same genetic loci)
- Repeat accumulation and establishment of genome system architecture

Some computational parallels

- Addressable functions
- Routines and subroutines
- Combinatorial circuit building from established components with known functional properties
- Amplification and reuse of circuit components
- Maintaining syntactical correctness -- targeting