

# Genome Evolution in the 21st Century

James A. Shapiro

University of Chicago

<http://shapiro.bsd.uchicago.edu>

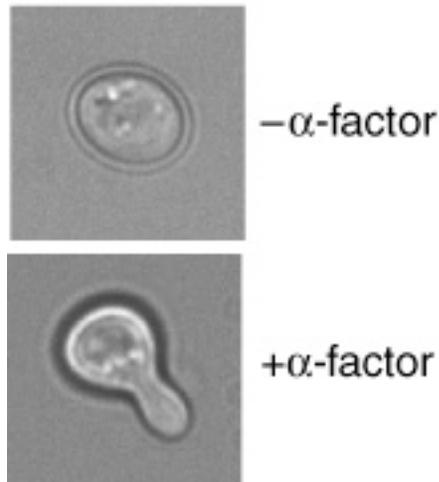
# Basic Principles of Genome Informatics

1. cells are cognitive entities;
2. cells use signal transduction networks to process internal and external information;
3. signal transduction involves reversible covalent bonds, weak interactions, cooperativity, and allostery;
4. DNA does not operate by itself and requires generic formatting to complex accurately with other molecules;
5. DNA stores information in sequence data files and metastable nucleoprotein complexes connected to signal transduction networks;
6. DNA storage has multiple functional requirements (physical & informatic organization, reading, transmission to progeny, repair, writing);
7. Cells have natural genetic engineering tools to restructure DNA molecules in many non-random ways.

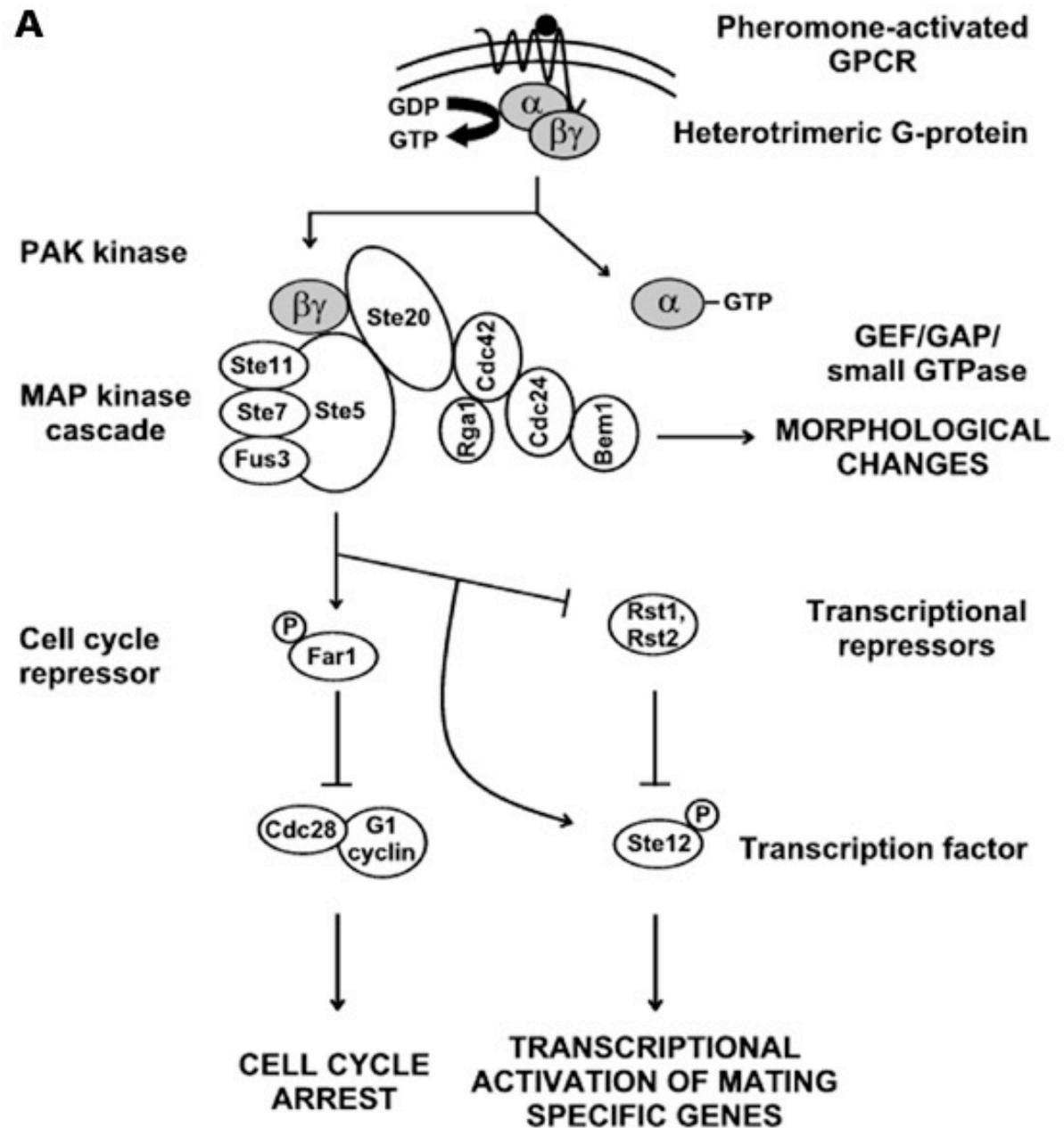
# Examples of cognitive behavior by cells

- Metabolic regulation in fluctuating environments
- Cell-cycle regulation (checkpoints)
- Error correction and damage repair
- Cellular differentiation
- Cell migration (either autonomously or coordinately during multicellular development)
- Wound healing and responses to invasion

# Signal Transduction in Yeast Mating



Pheromone Response  
Frank van Drogen et al. MAP kinase dynamics in response to pheromones in budding yeast. *Nature Cell Biology* 3, 1051 - 1059 (2001).



Mary J. Cismowski et al. Genetic screens in yeast to identify mammalian nonreceptor modulators of G-protein signaling. *Nature Biotechnology* 17, 878 - 883 (1999).

# What must DNA nucleoprotein complexes do in living cells?

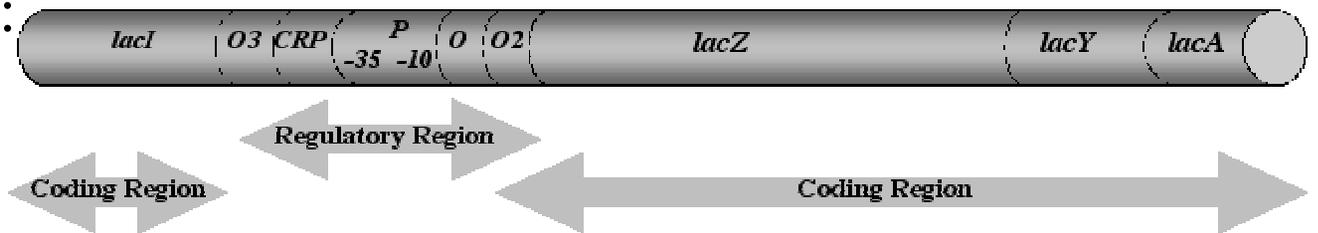
1. Package genome correctly in the cell
2. Access RNA and protein data files (transcription)
3. Control timing and level of data file transcription
4. Replicate information in genome once per cell cycle
5. Proofread replicated DNA
6. Transmit replicated DNA correctly to daughter cells
7. Repair genome damage
8. Restructure genome when necessary

**Without #1-#7, normal cell proliferation cannot occur; without #8, evolution is not possible.**

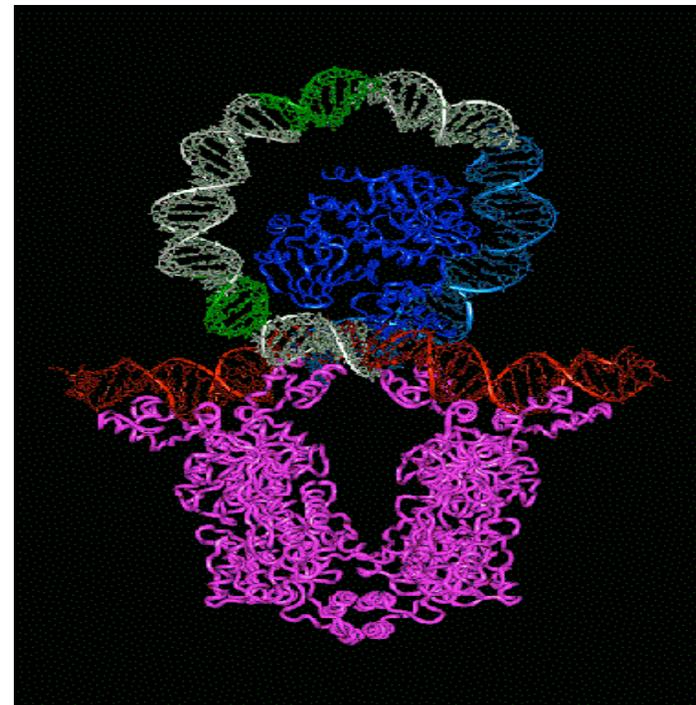
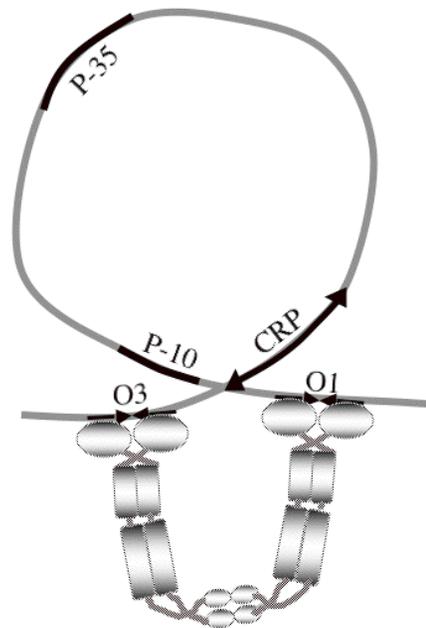
# DNA formatting to execute a cognitive algorithm - *E. coli* discriminates glucose and lactose

1. The algorithm: IF lactose present AND glucose not present AND cell can synthesize active LacZ (beta-galactosidase) and LacY (lactose permease), THEN transcribe *lacZYA* from *lacP*

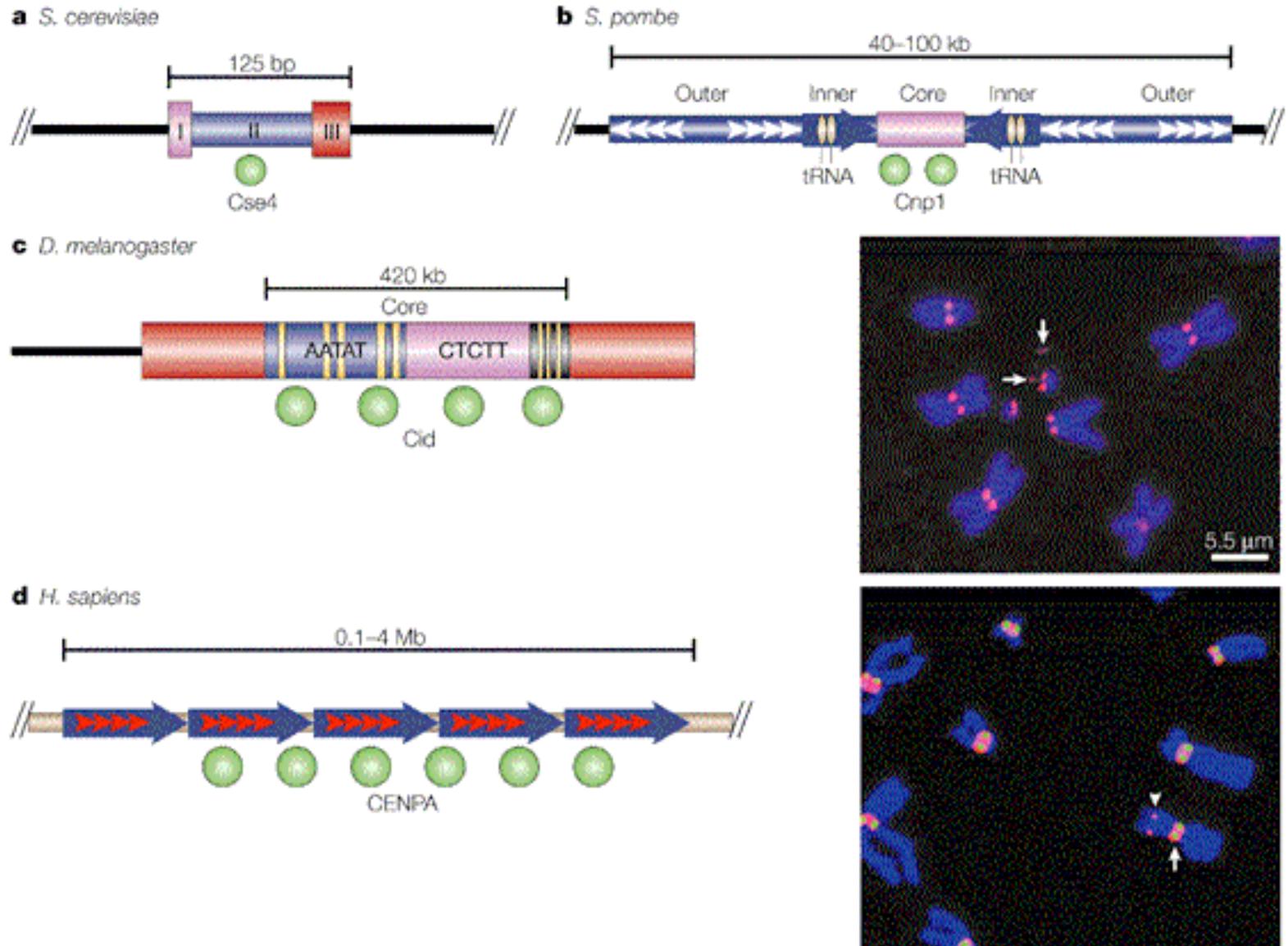
2. The formatted DNA:



3. A nucleoprotein complex (LacI repressor bound to *lacO* operators):



# DNA formatting for chromosome transmission to daughter cells - centromeres



B A Sullivan, M D  
Blower & G H  
Karpen  
DETERMINING  
CENTROMERE  
IDENTITY:  
CYCLICAL  
STORIES AND  
FORKING  
PATHS Nature  
Reviews Genetics  
2; 584-596 (2001)

# Repetitive DNA/Mobile Genetic Elements in the Human Genome

Classes of interspersed repeat in the human genome

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

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860 - 921 (2001)

# Importance of repetitive DNA in several genomes

Species	Genome size <sup>1</sup>	% repetitive DNA	% coding sequences	Reference
<b>Animals</b>				
<i>Caenorhabditis elegans</i>	100 MB	16.5	14	Stein <i>et al.</i> (2003)
<i>Caenorhabditis briggsae</i>	104 MB	22.4	13	Stein <i>et al.</i> (2003)
<i>Drosophila melanogaster</i>	175 MB	33.7 (female) ~57 (male) <sup>2</sup>	<10	Bennett <i>et al.</i> (2003); Celniker <i>et al.</i> (2002)
<i>Ciona intestinalis</i>	157MB	35	9.5	Dehal <i>et al.</i> (2002)
<i>Fugu rubripes</i>	365MB	15	9.5	Aparicio <i>et al.</i> (2002)
<i>Canis domesticus</i>	2.4GB	31	1.45	Kirkness <i>et al.</i> (2003)
<i>Mus musculus</i>	2.5GB	40	1.4	Mouse Genome Sequencing Consortium (2002)
<i>Homo sapiens</i>	2.9 GB	≥50	1.2	International Human Genome Consortium (2001)
<b>Plants</b>				
<i>Arabidopsis thaliana</i>	125-157 MB	13-14	21	<i>Arabidopsis</i> Genome Initiative (2000) Bennett <i>et al.</i> (2003)
<i>Oryza sativa</i> (indica)	466 MB	42	11.8	Yu <i>et al.</i> (2002)
<i>Oryza sativa</i> (Japonica)	420 MB	45	11.9	Goff <i>et al.</i> (2002)
<i>Zea mays</i>	2.5 GB	77	1	Meyers <i>et al.</i> (2001)

1. MB = megabases ( $10^6$  base pairs), GB = gigabases ( $10^9$  base pairs)

2. The *D. melanogaster* Y chromosome is largely heterochromatic repetitive DNA.

Shapiro & Sternberg. 2005. Why repetitive DNA is essential to genome function. *Biol. Revs.* **80**, 227-50

# Natural Genetic Engineering - the cellular toolbox for genome restructuring

- Homologous recombination
- Non-homologous end-joining
- Site-specific recombination
- DNA transposons (large-scale rearrangements)
- Retrotransposons (small-scale rearrangements)
- Homing and retrohoming introns and inteins
- Mutator polymerases

# Some ways cells make use of natural genetic engineering

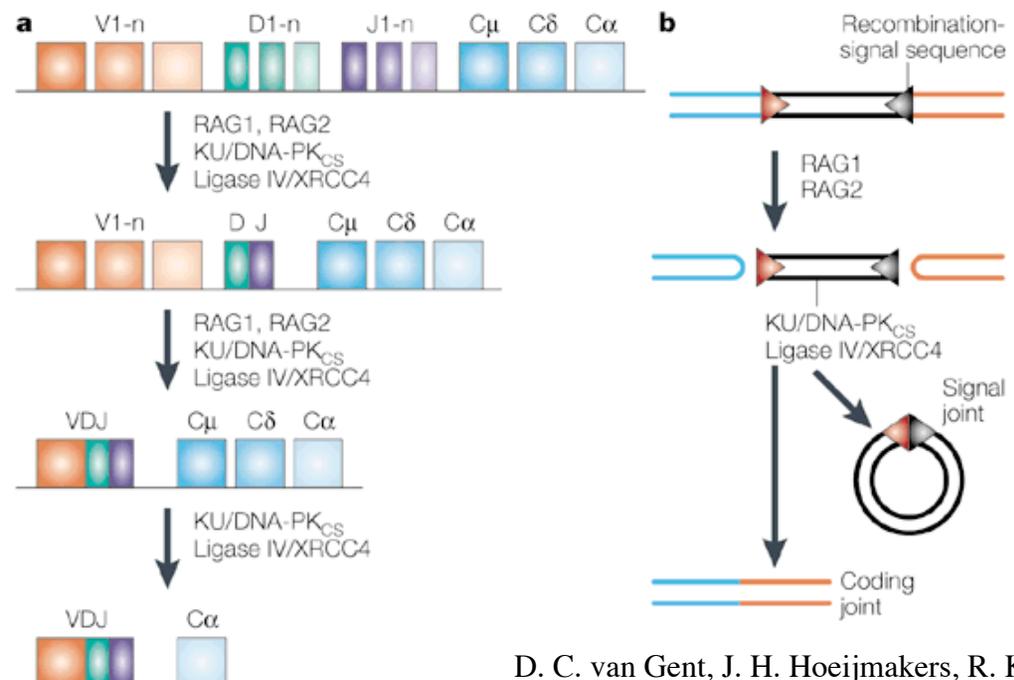
- Regulation of protein synthesis by targeted homologous recombination in yeast;
- Regulation of protein synthesis by site-specific recombination in bacteria;
- Regulation of protein synthesis by gene conversion and chromosome rearrangement in trypanosomes;
- Protein structural engineering by site-specific recombination in bacteria;
- Extend chromosome ends (telomeres) in insects;
- Rapid protein evolution by multiple coordinated DNA rearrangements and directed hypermutation in vertebrate immune systems.

# **The Mammalian Immune System: A rapid protein evolution factory - specificity, flexibility and regulation in functional natural genetic engineering**

- Combinatorial joining of germ-line cassettes (VDJ joining at transposon-like DNA cleavage signals)  $\implies \sim 10^6$  antibodies;
- Internucleotide flexibility in VDJ joining  $\implies \sim 10^3$  greater diversity;
- Incorporation of untemplated polynucleotides at VD and DJ junctions  $\implies > 10^2$  greater diversity;
- Antigenic stimulation of B cells  $\implies$  rapid selection of appropriate VDJ heavy chain & VJ light chain combinations;
- Somatic hypermutation restricted to sequences encoding antigen-binding region of antibody molecules in stimulated B cells  $\implies$  production of antibodies with higher affinity;
- Lymphokine-directed “class switching” of heavy chain constant region sequences (transcription-mediated breakage and rejoining)  $\implies$  effective antibodies targeted to different sites in the body;
- B cell specificity; ordered VD - DJ - VJ joining; allelic exclusion
- Sequential activation of somatic hypermutation and class switching

# The Mammalian Immune System: An Evolved Rapid Evolution System

## VDJ recombination (transposon-like breakage)

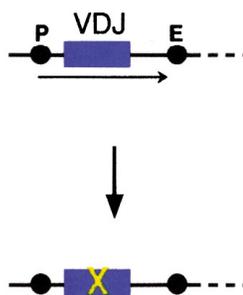


D. C. van Gent, J. H. Hoeijmakers, R. Kanaar, Chromosomal Stability And The DNA Double-Stranded Break Connection 2, 196 (2001)  
Nature Reviews | Genetics

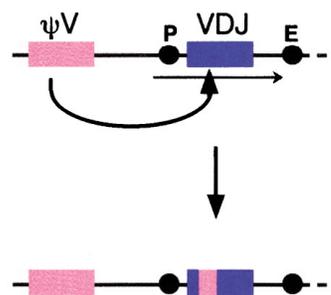
### A. Ig Heavy Chain Locus



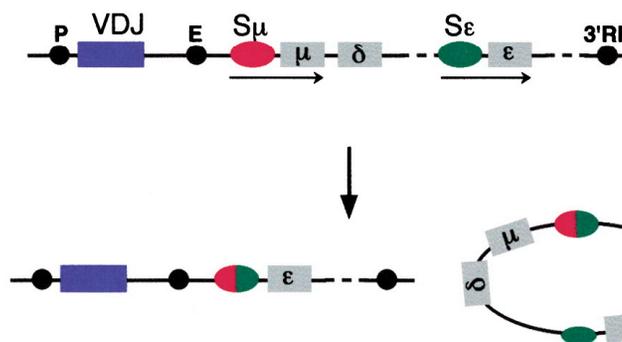
### B. Hypermutation



### C. Gene Conversion



### D. Class Switch Recombination



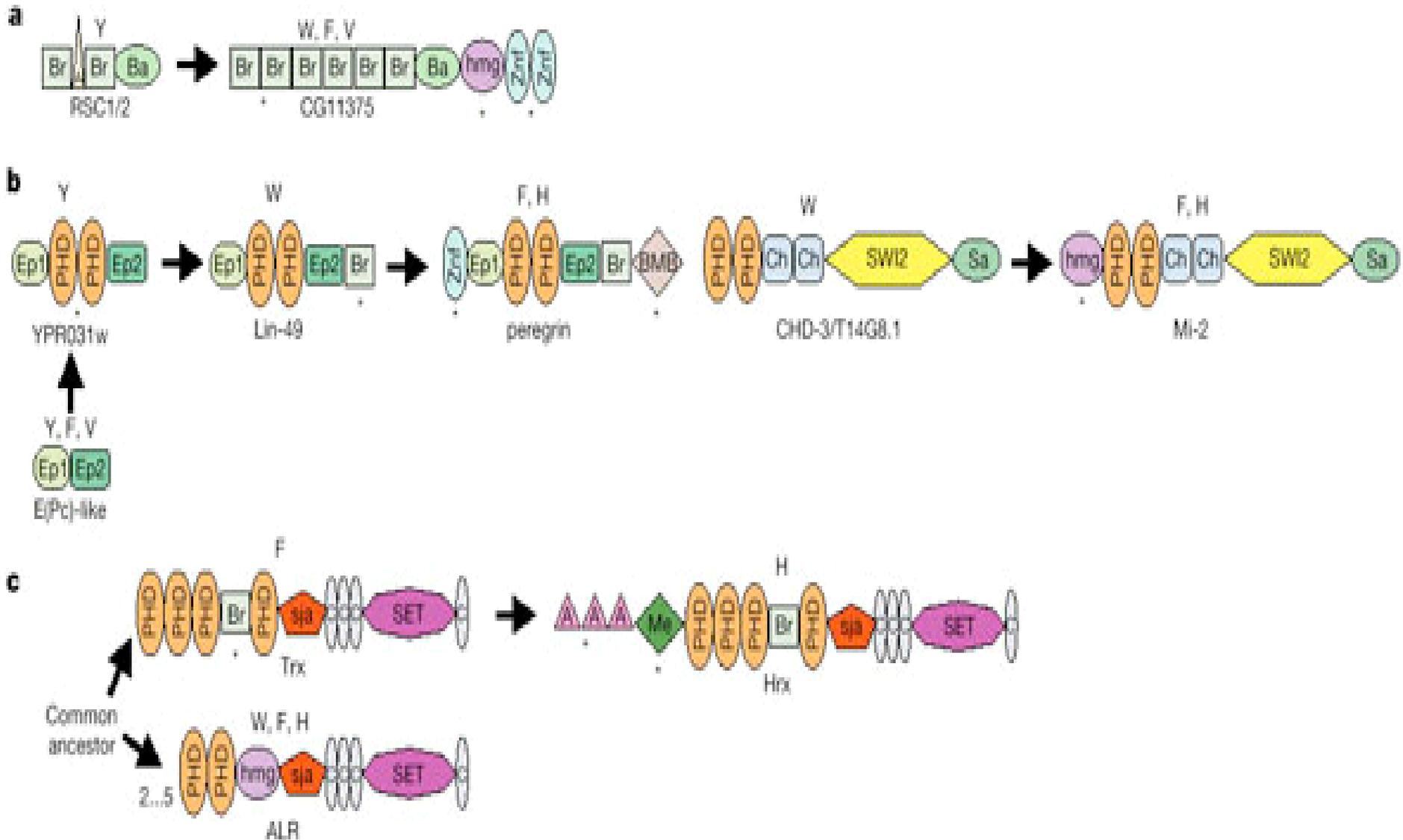
Somatic hypermutation & class switch recombination - transcription directed

Tasuko Honjo, Kazuo Kinoshita, and Masamichi Muramatsu. 2001. Molecular Mechanism of Class Switch Recombination: Linkage with Somatic Hypermutation. Annu. Rev. Immunol.;

# What have whole genome sequences revealed about evolution?

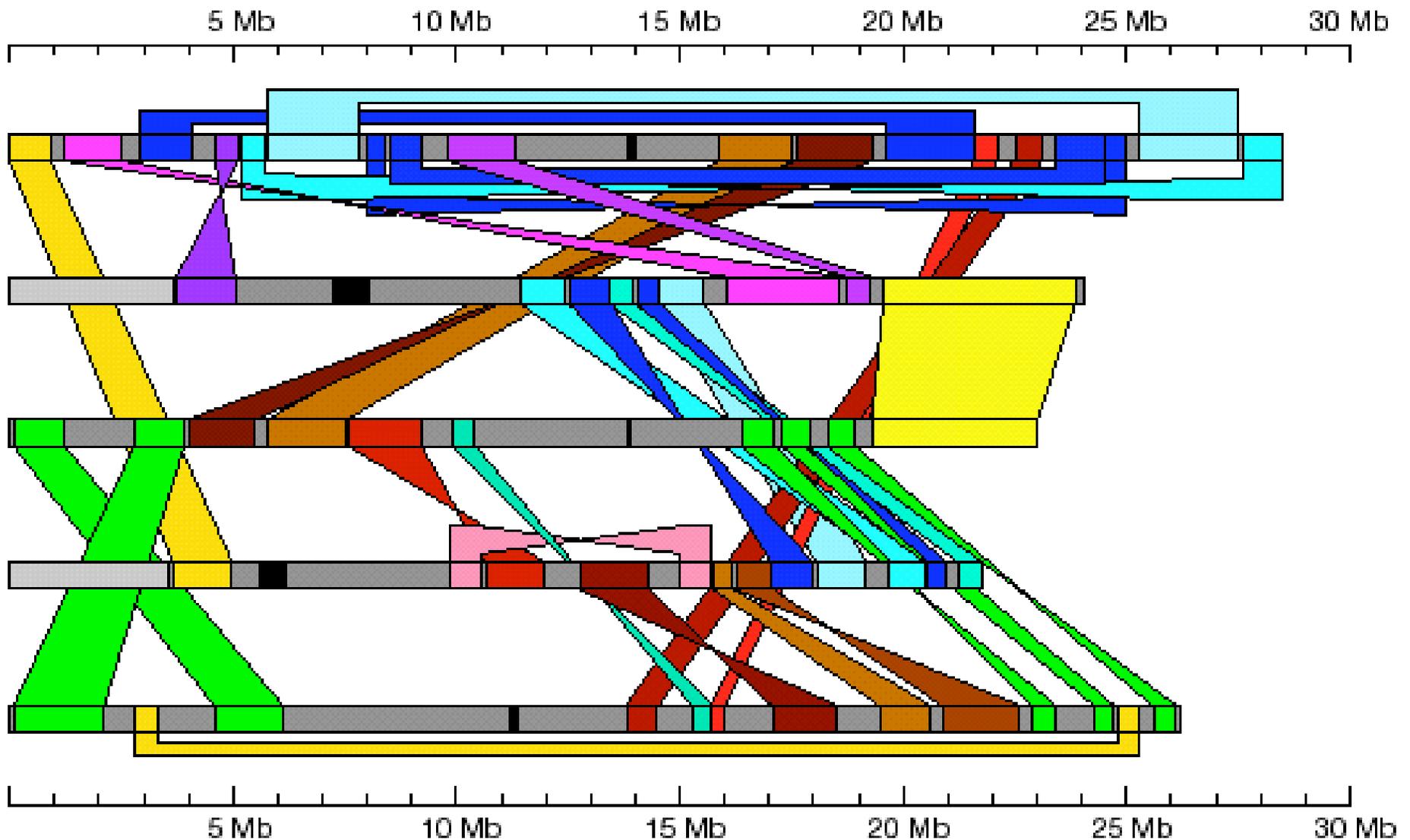
1. Protein evolution by domain shuffling and domain accretion;
2. Protein evolution by retrotransposon insertion in introns;
3. Protein families: taxonomically-specific duplications and transpositions;
4. Regulatory region evolution by transposon and retrotransposon insertions;
5. Formation of chromatin domains by clustered transposon and retrotransposon insertions;
6. Segmental duplications and rearrangements;
7. Whole genome duplications (tetraploidization);
8. Genome-wide rearrangement of “syntenic” blocks;
9. Cell evolution by symbiosis.

# Protein evolution by domain accretion



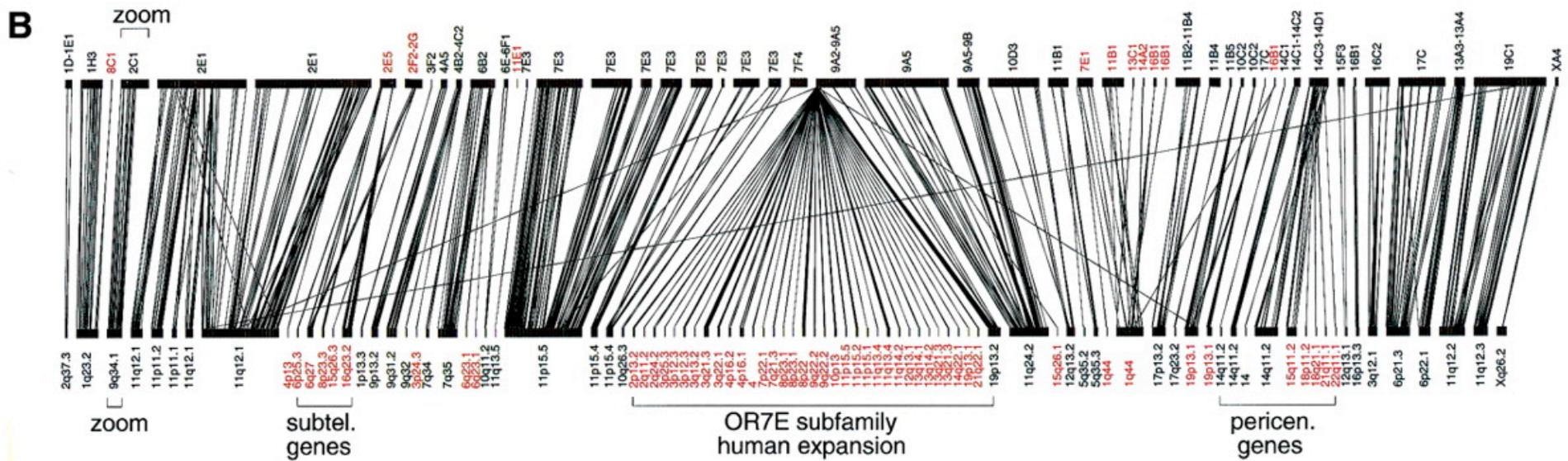
International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 409, 860 - 921 (2001)

# Segmental Duplications in Arabidopsis



Segmentally duplicated regions in the Arabidopsis genome. Individual chromosomes are depicted as horizontal grey bars (with chromosome 1 at the top), centromeres are marked black. Coloured bands connect corresponding duplicated segments. Similarity between the rDNA repeats are excluded. Duplicated segments in reversed orientation are connected with twisted coloured bands. The scale is in megabases. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796 - 815 (2000).

# Protein Amplification: Olfactory Receptors



Most pairs of similar OR gene clusters (labeled in black) fall into established syntenic chromosomal regions (<http://www.ncbi.nlm.nih.gov/Homology/index.html>), but some (labeled in red) do not.  
 Young, J. M. et al. Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Hum. Mol. Genet.* 11, 535-546 (2002)

# Mouse-Human Synteny

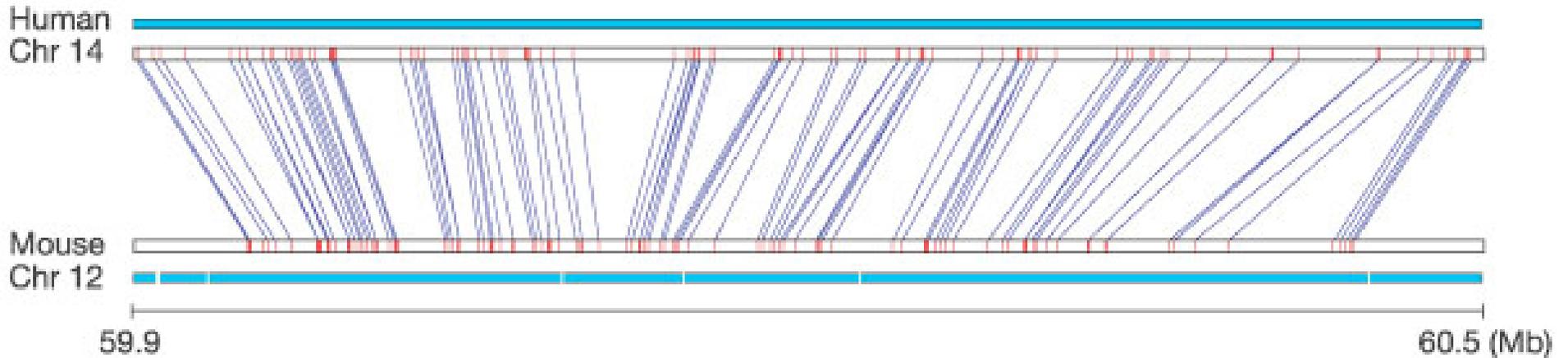


Figure 2 Conservation of synteny between human and mouse. We detected 558,000 highly conserved, reciprocally unique landmarks within the mouse and human genomes, which can be joined into conserved syntenic segments and blocks (defined in text). A typical 510-kb segment of mouse chromosome 12 that shares common ancestry with a 600-kb section of human chromosome 14 is shown. Blue lines connect the reciprocal unique matches in the two genomes. The cyan bars represent sequence coverage in each of the two genomes for the regions. In general, the landmarks in the mouse genome are more closely spaced, reflecting the 14% smaller overall genome size.

Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520 - 562 (2002)

# Mouse-Human: Synteny and Scrambling

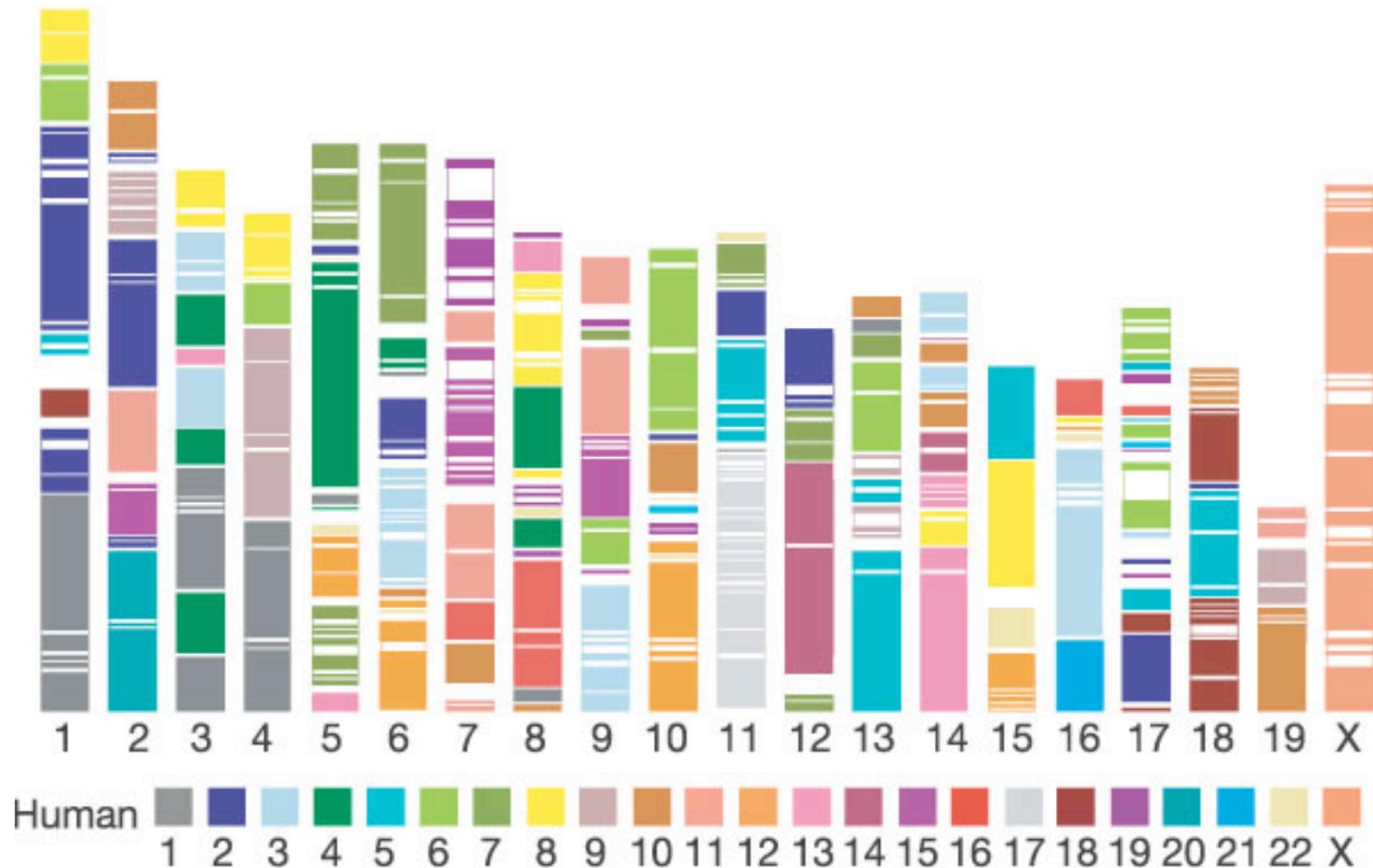
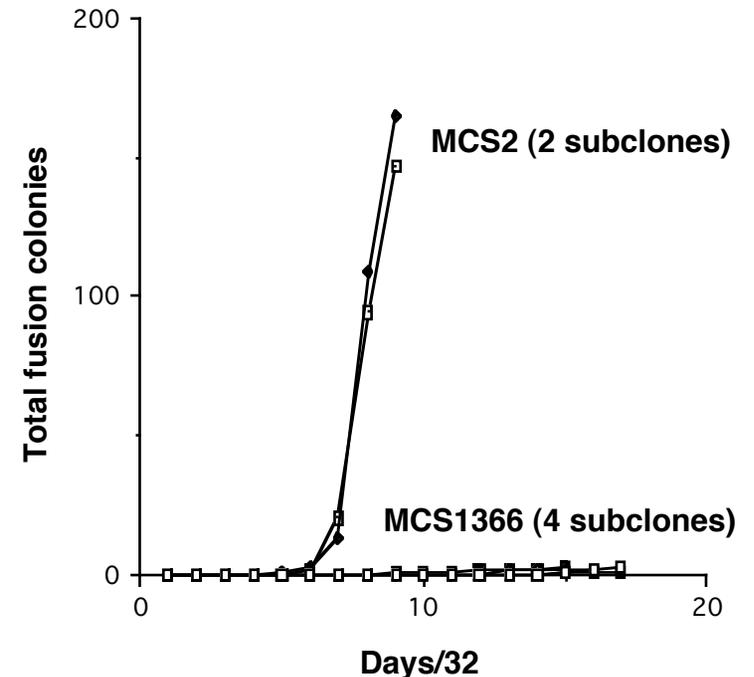
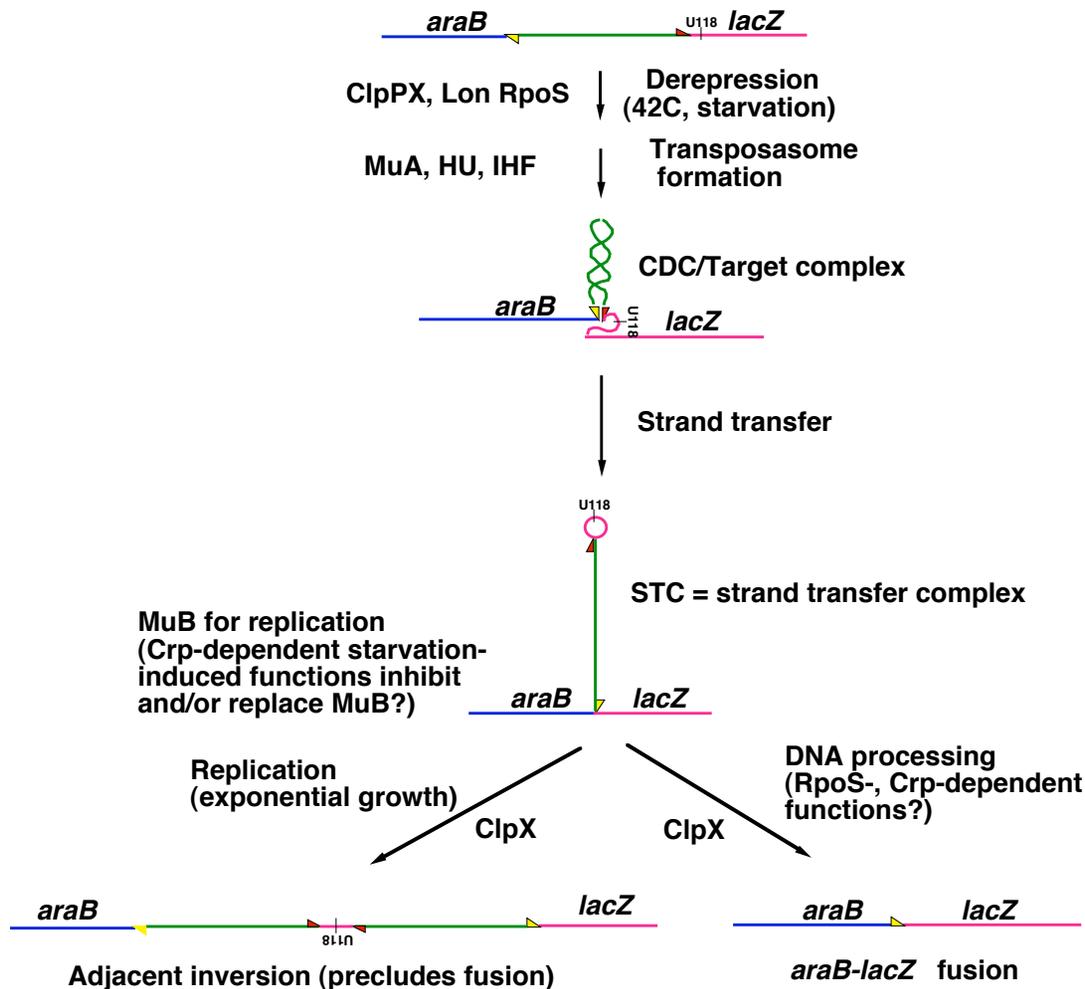


Figure 3 Segments and blocks >300 kb in size with conserved synteny in human are superimposed on the mouse genome. Each colour corresponds to a particular human chromosome. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520 - 562 (2002)

# Temporal & metabolic regulation of natural genetic engineering



# Targeting of natural genetic engineering

## Known molecular mechanisms:

- Sequence recognition by proteins (yeast mating-type switching, ribosomal LINE elements, homing introns, VDJ joining);
- Protein-protein interaction with transcription factors or chromatin proteins (Ty retrotransposon targeting);
- Sequence recognition by RNA (reverse splicing of group II retrohoming introns);
- Transcriptional activation of target DNA (somatic hypermutation; class-switch recombination).

## Unknown mechanisms:

- Telomere targeting of certain LINE elements in insects;
- HIV & MLV targeting upstream of transcribed regions;
- P factor homing directed by transcription, chromatin signals.

# Advantages of non-random searches of genome space at evolutionary crises

- Genome changes occur under stress or other conditions, when they are most likely to prove beneficial;
- Multiple related changes can occur when a particular natural genetic engineering system is activated;
- Rearrangement of proven genomic components increases the chance that novel combinations will be functional;
- Targeting can increase the probability of functional integration and reduce the risk of system damage;
- Rearrangements followed by localized changes provide opportunities for fine tuning once novel function has been achieved.

# A 21st Century view of genomes and evolution

- All genome functions are interactive (no Cartesian dualism, genome always in communication with rest of cell);
- Every genome component operates as part of a complex information-processing system (no “one gene-one trait” correlation);
- Genome systems are organized and integrated into cell networks by repetitive DNA;
- Genome change is a regulated biological function;
- Natural genetic engineering processes are subject to biological feedback at multiple levels.

# Some Recent Examples of 21st Century Thinking

- **” To our knowledge, this is the first example of TEs initiating synchronous, developmentally regulated expression of multiple genes in mammals. We propose that differential TE expression triggers sequential reprogramming of the embryonic genome during the oocyte to embryo transition and in preimplantation embryos.”**

Peaston AE, et al. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev Cell*. 2004 Oct;7(4):597-606.

- **” Combining these results with recent findings about RNAi, we suggest that specific repetitive elements, as well as density, play a role in determining higher-order chromatin packaging.”**

Elizabeth E Slawson, et al. Comparison of dot chromosome sequences from *D. melanogaster* and *D. virilis* reveals an enrichment of DNA transposon sequences in heterochromatic domains.

*Genome Biology* 2006, 7:R15 Subject areas: Genetics, Evolution The electronic version of this article is the complete one and can be found online at: <http://genomebiology.com/2006/7/2/R15>

- **“Together, these data suggest that transposable elements have a profound impact on the *M. oryzae* genome by creating localized segments with increased rates of chromosomal rearrangements, gene duplications and gene evolution.”**

Michael R Thon, et al. The role of transposable element clusters in genome evolution and loss of synteny in the rice blast fungus *Magnaporthe oryzae*. *Genome Biology* 2006, 7:R16 Subject areas: Microbiology and parasitology, Evolution, Genome studies The electronic version of this article is the complete one and can be found online at: <http://genomebiology.com/2006/7/2/R16>



