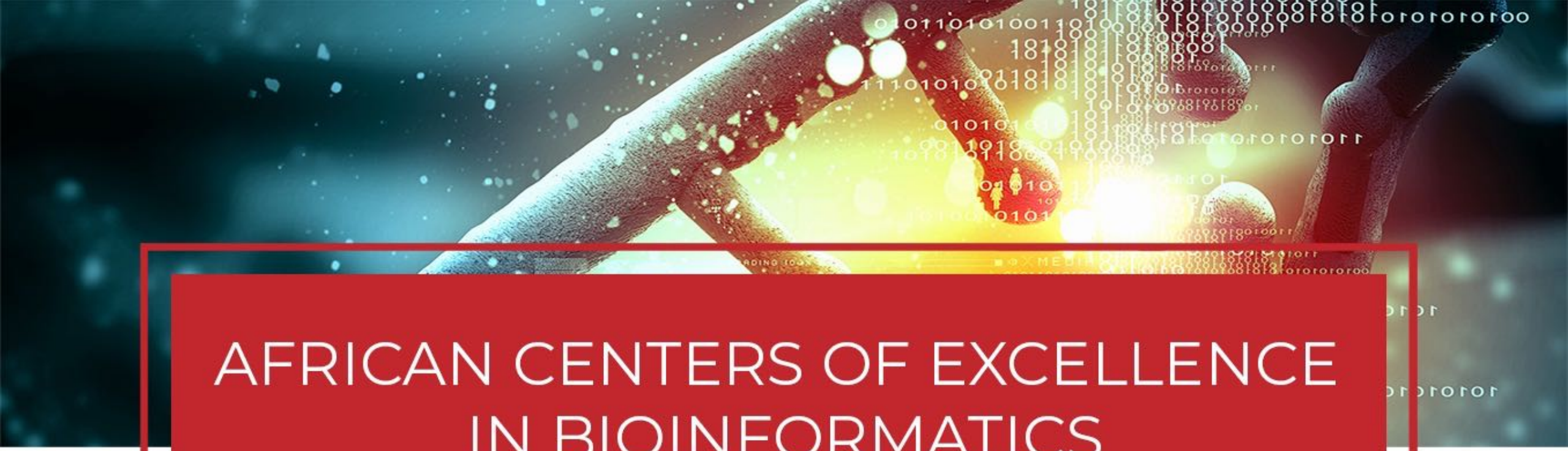


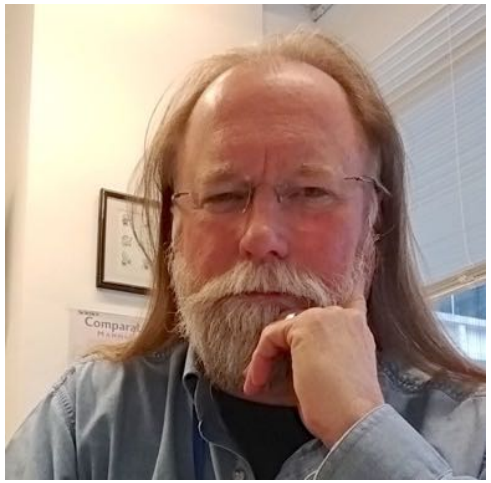
AFRICAN CENTERS OF EXCELLENCE IN BIOINFORMATICS

KAMPALA, UGANDA

GENETIC EVOLUTION/DIVERSITY MECHANISMS

MSB 7101

Today's Instructor



Dr. Kurt Wollenberg, Ph.D. in Genetics

Ongoing Computational Biology projects:

- Molecular evolution of drug resistance in *M. tuberculosis*
- CLAG protein family evolution

- Bioinformatics and Computational Biosciences Branch (BCBB), NIAID
- National Institutes of Health, Bethesda, MD USA.
- Contact our team via email:
  - Email: bioinformatics@niaid.nih.gov
  - Instructor: kurt.wollenberg@nih.gov

Class Materials

- NIAID Box folder:
  - https://nih.app.box.com/folder/132214849477
- NIAID github repository:
  - https://github.com/niaid/ACE-2021

GENETIC DIVERSITY

Where does it come from?

- Review of relevant cell biology topics
- Somatic/germ-line cells
- Diploid vs haploid
- Mitosis, meiosis, and DNA replication
- Clonal reproduction

GENETIC DIVERSITY

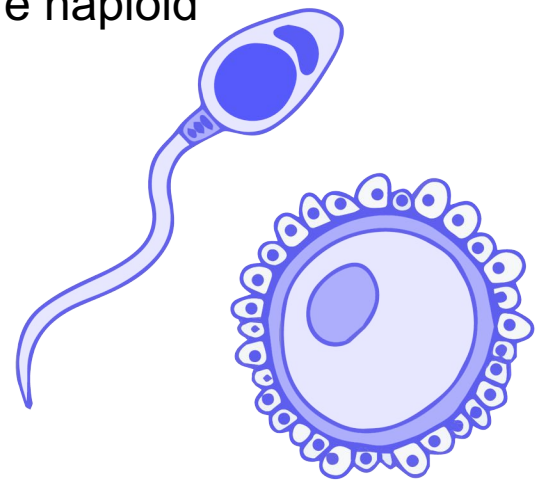
Somatic and germ-line cells

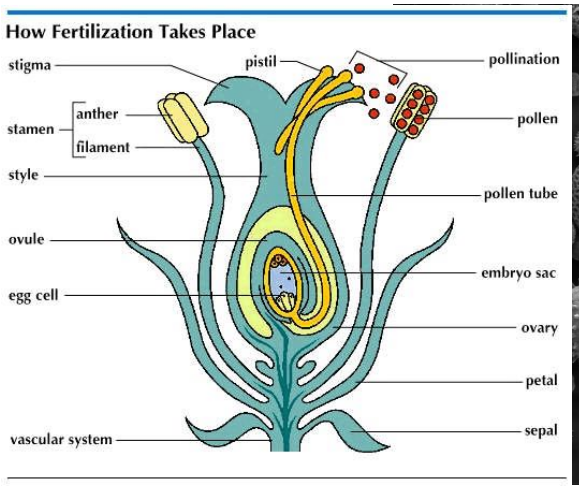
Somatic cells

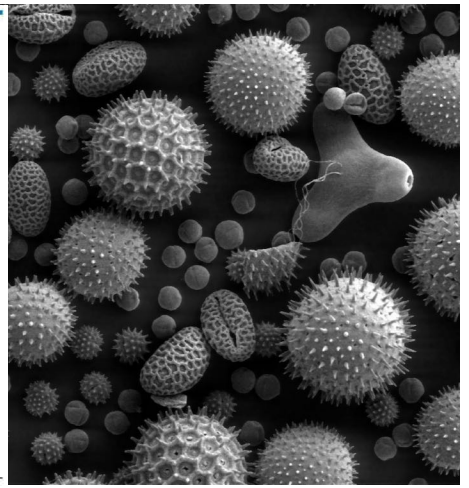
- Make up the majority of tissues
- Do not pass on to the next generation
- Typically diploid, though in plants this is complicated

Germ line cells

- Specialized cells that pass on genetic information to the next generation ➡ Gametes
- These cells are haploid







GENETIC DIVERSITY

Diploid vs haploid

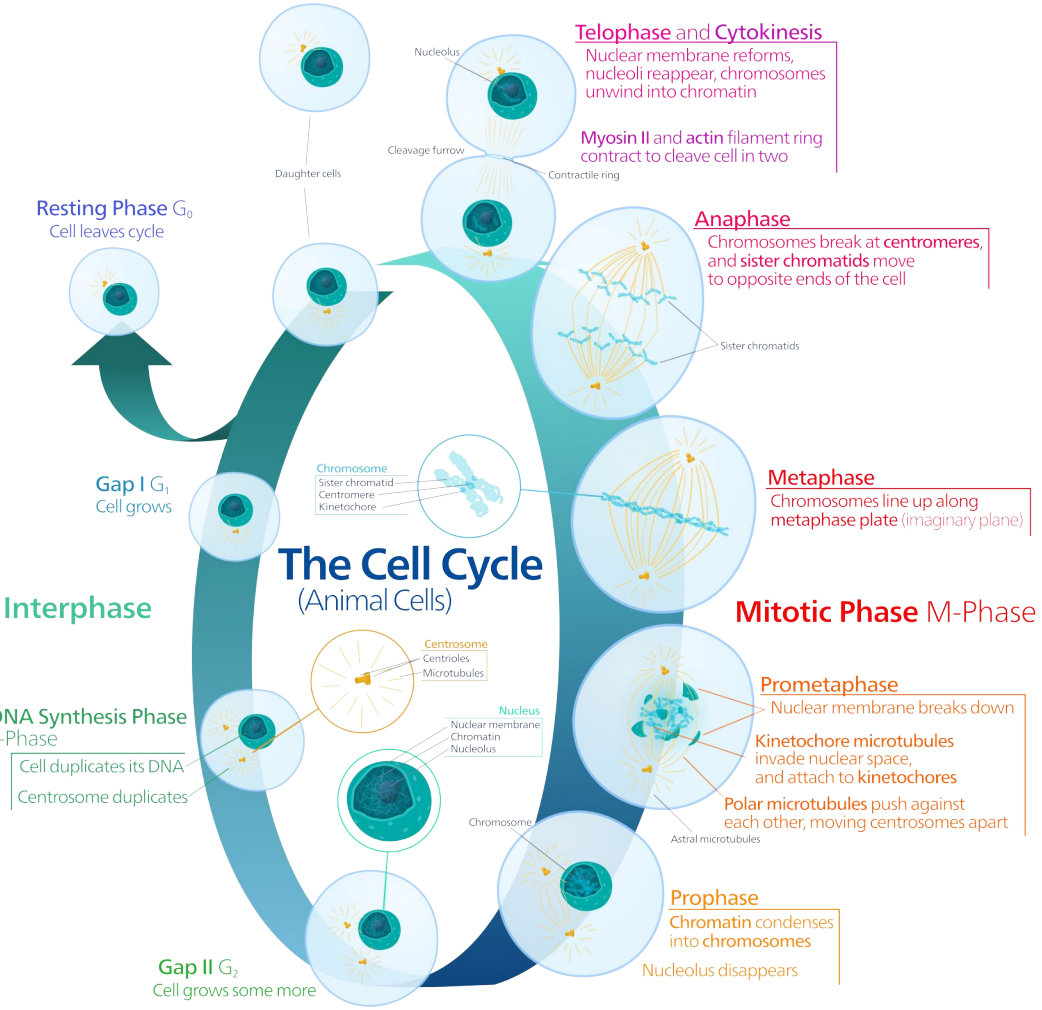
Diploid: The cell nucleus contains two copies of each chromosome

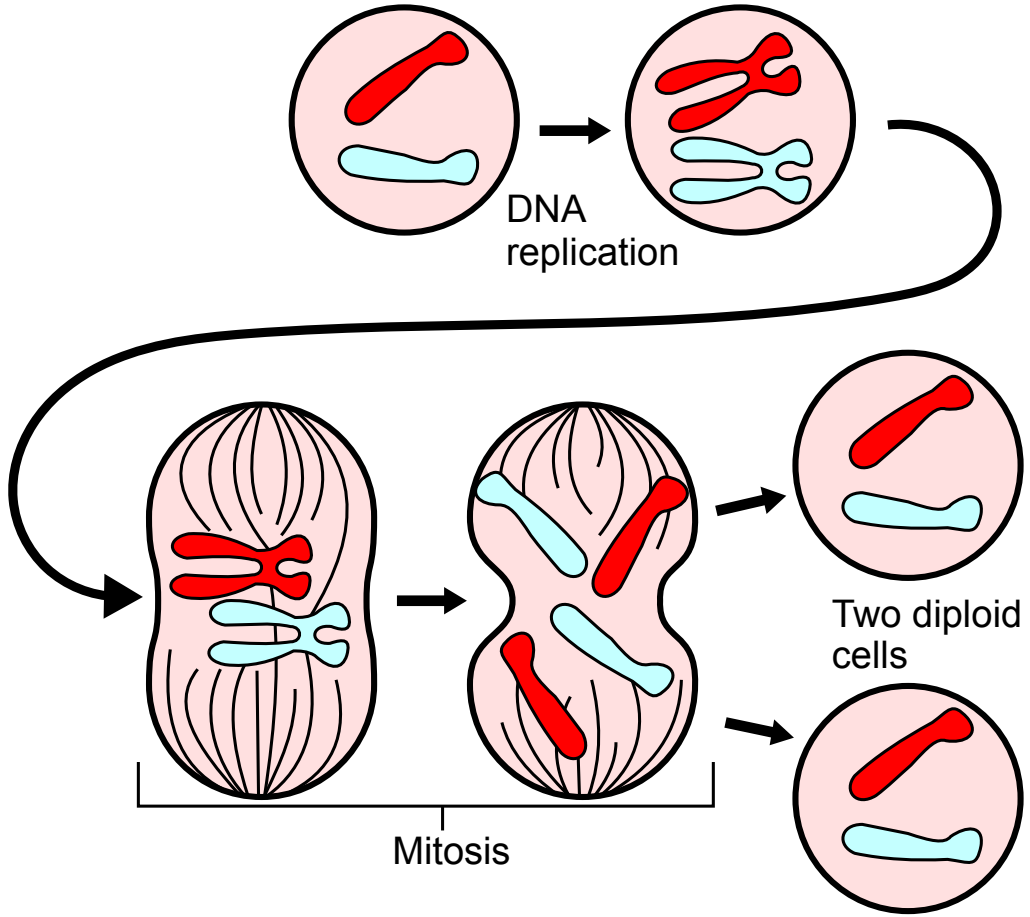
Polyploid: The cell nucleus contains more than two copies of each chromosome

Haploid: The cell contains only one copy of each chromosome. The chromosome may or may not be contained in a cell nucleus.

GENETIC DIVERSITY

Mitosis and Meiosis – Mitosis: somatic cell duplication

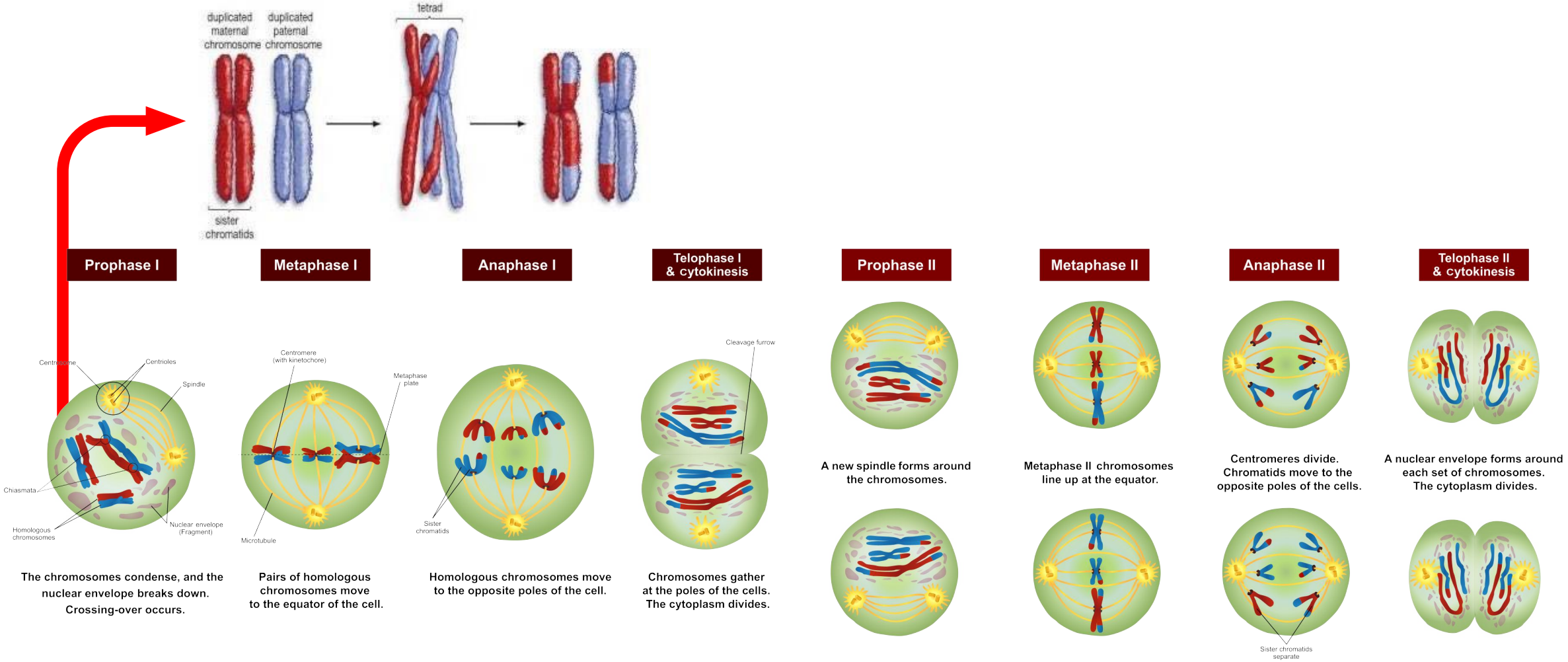




By Kelvinsong - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=22965076

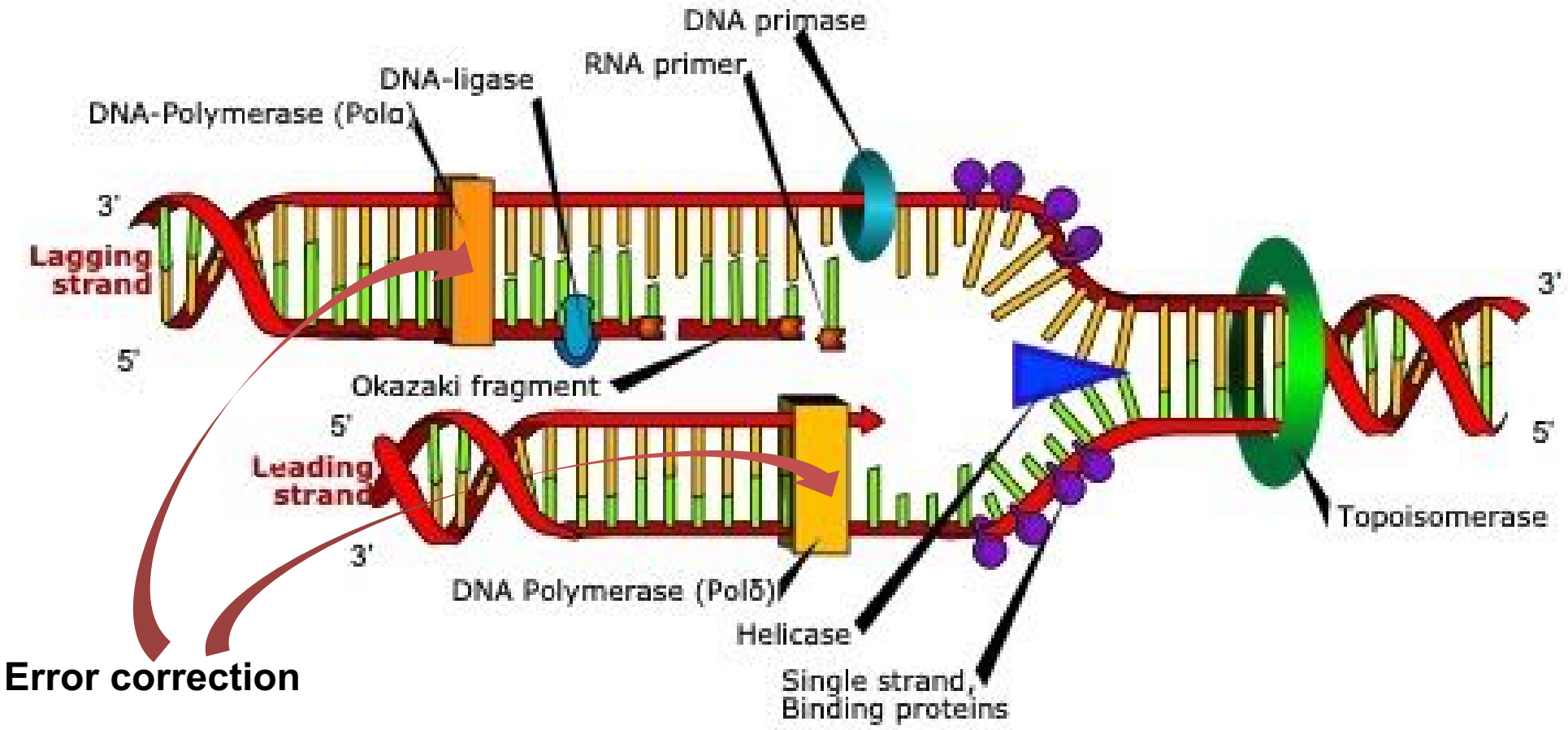
GENETIC DIVERSITY

Mitosis and Meiosis – Meiosis: Duplication and reduction to produce gametes



GENETIC DIVERSITY

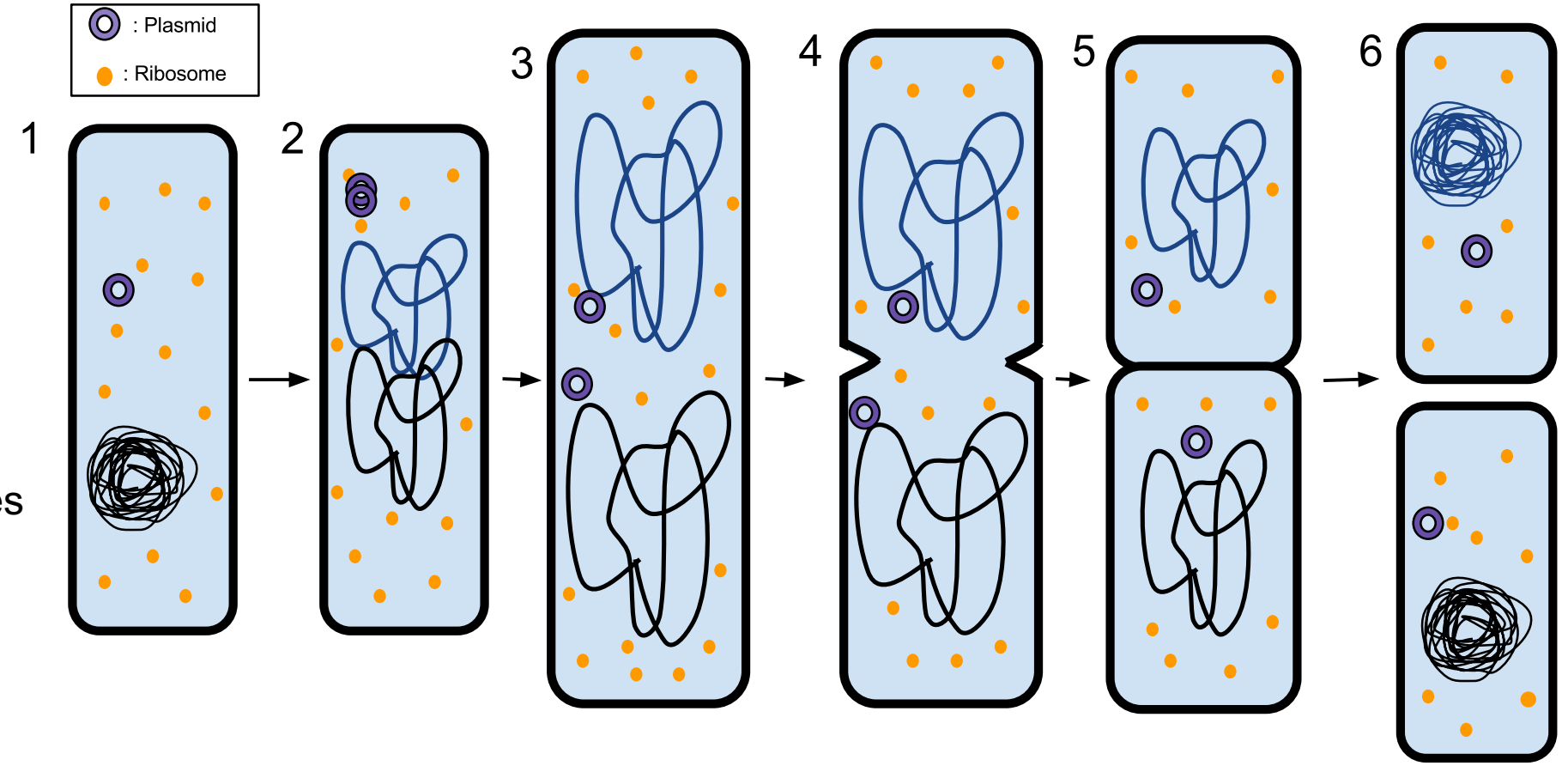
DNA replication



GENETIC DIVERSITY

Clonal reproduction: Bacteria (also Archaea, some plants, animals, and fungi)

1. Parent cell
2. Genetic material duplication
3. Cell elongates and genetic material migrates
4. Cell wall invaginates
5. Invaginations meet
6. Daughter cells separate

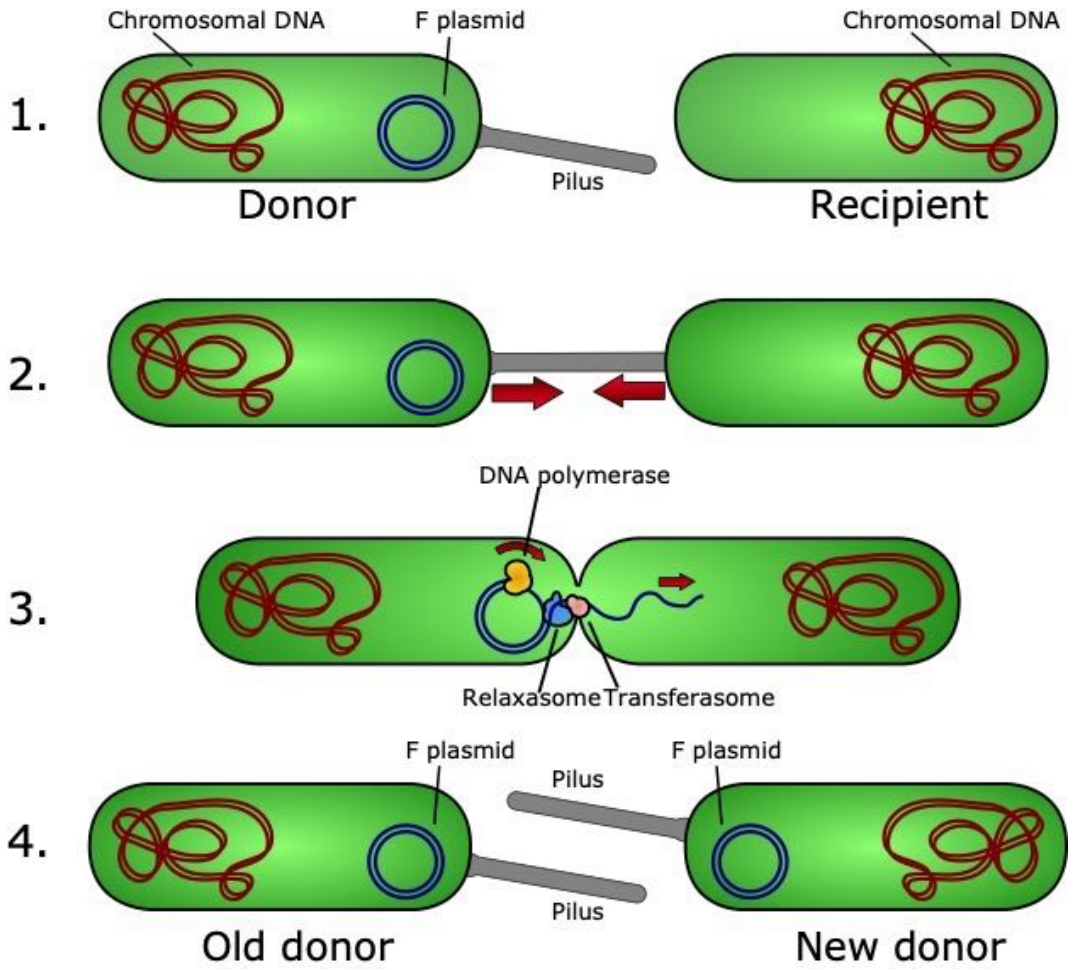


GENETIC DIVERSITY

Not quite clonal reproduction: lateral gene transfer

Bacterial Conjugation

Plasmid transfer between cells

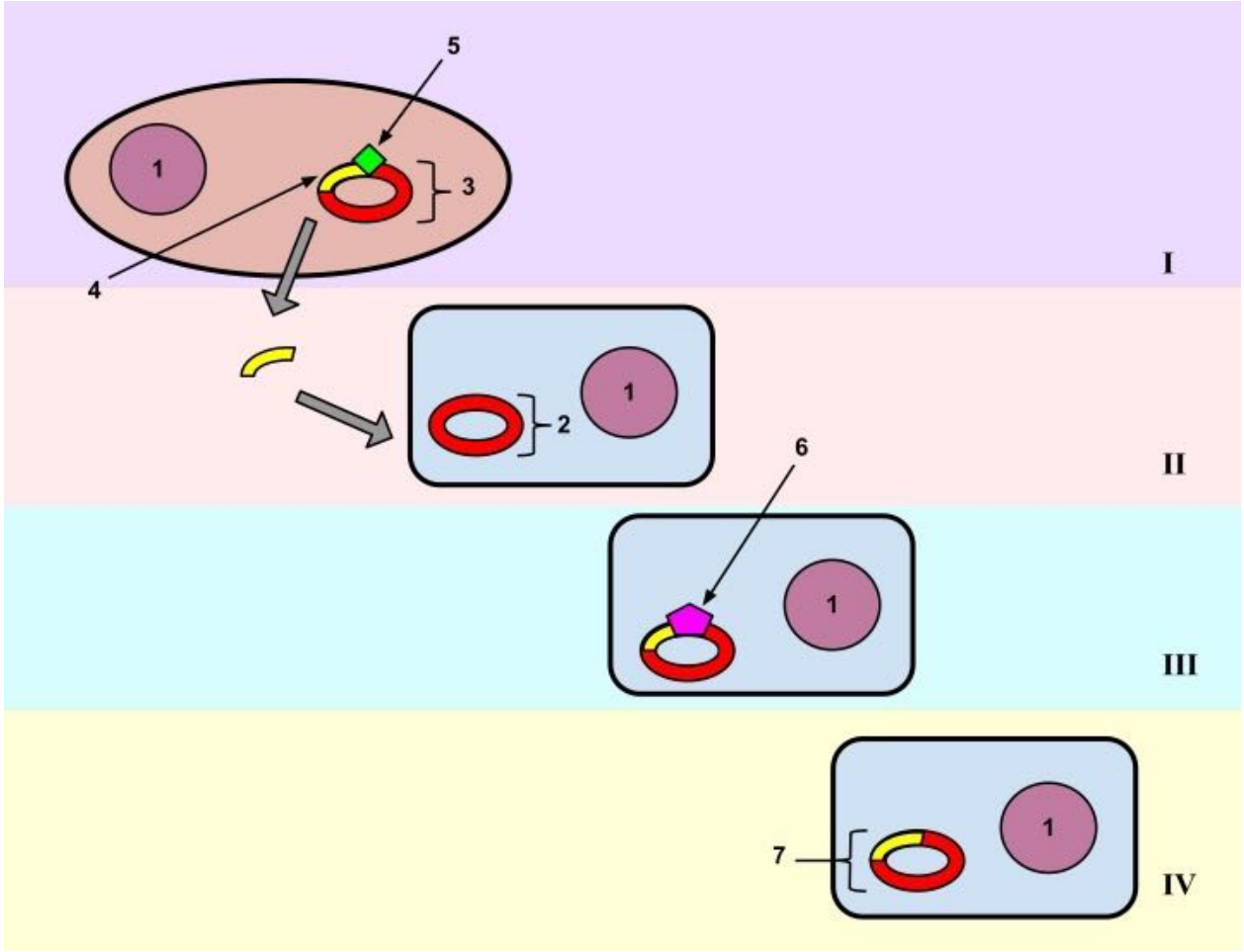


GENETIC DIVERSITY

Not quite clonal reproduction: lateral gene transfer

Bacterial Transformation

- Receptor cell must be competent
- Competence usually induced by starvation or crowding
- Incorporation into chromosome easier if same species (homologous)

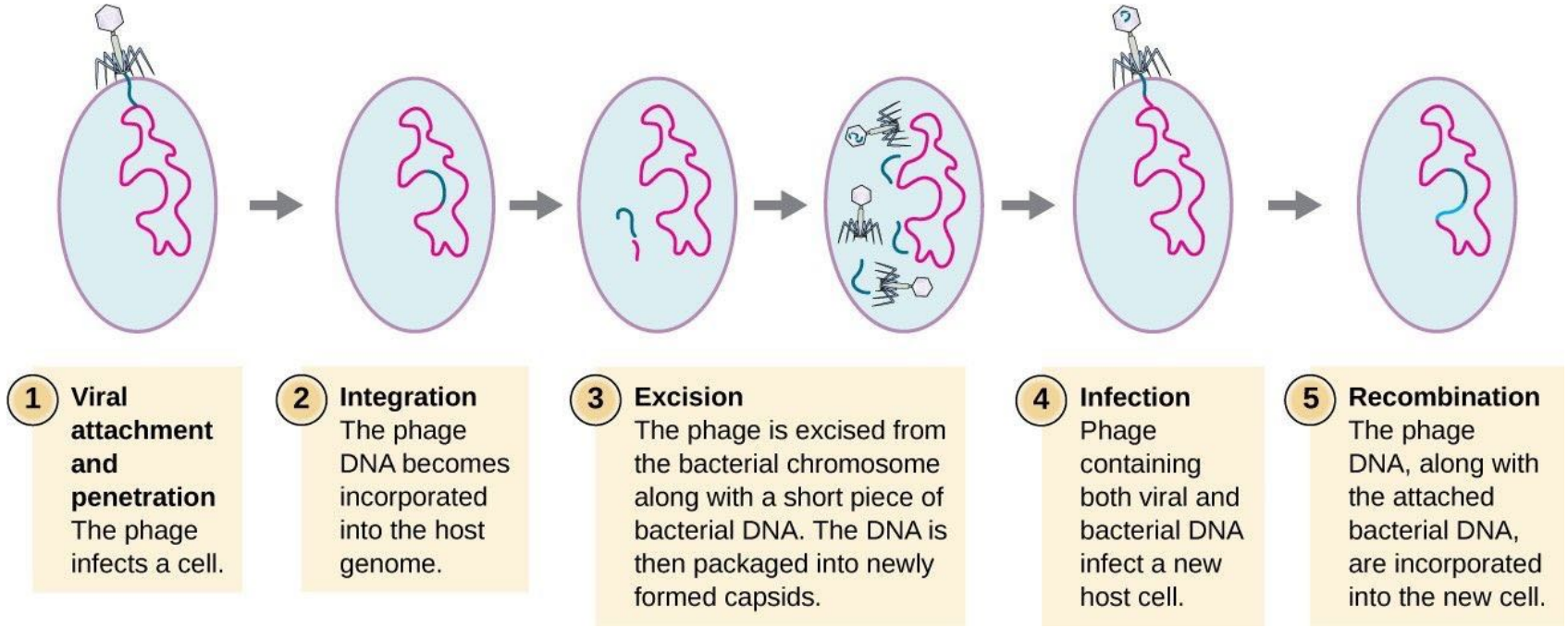


GENETIC DIVERSITY

Not quite clonal reproduction: lateral gene transfer

Bacterial Transduction

- Viral transfer of host bacterial DNA to other bacteria
- Can be general (random host DNA) or specialized (specific host DNA).



GENETIC DIVERSITY

BREAK

DIVERSITY MECHANISMS

How does genomic diversity spread?

- Mutations – changes to individual nucleotides
- Substitutions – mutations that persist over time
- Substitution rate – the speed at which substitutions accumulate over time in a lineage

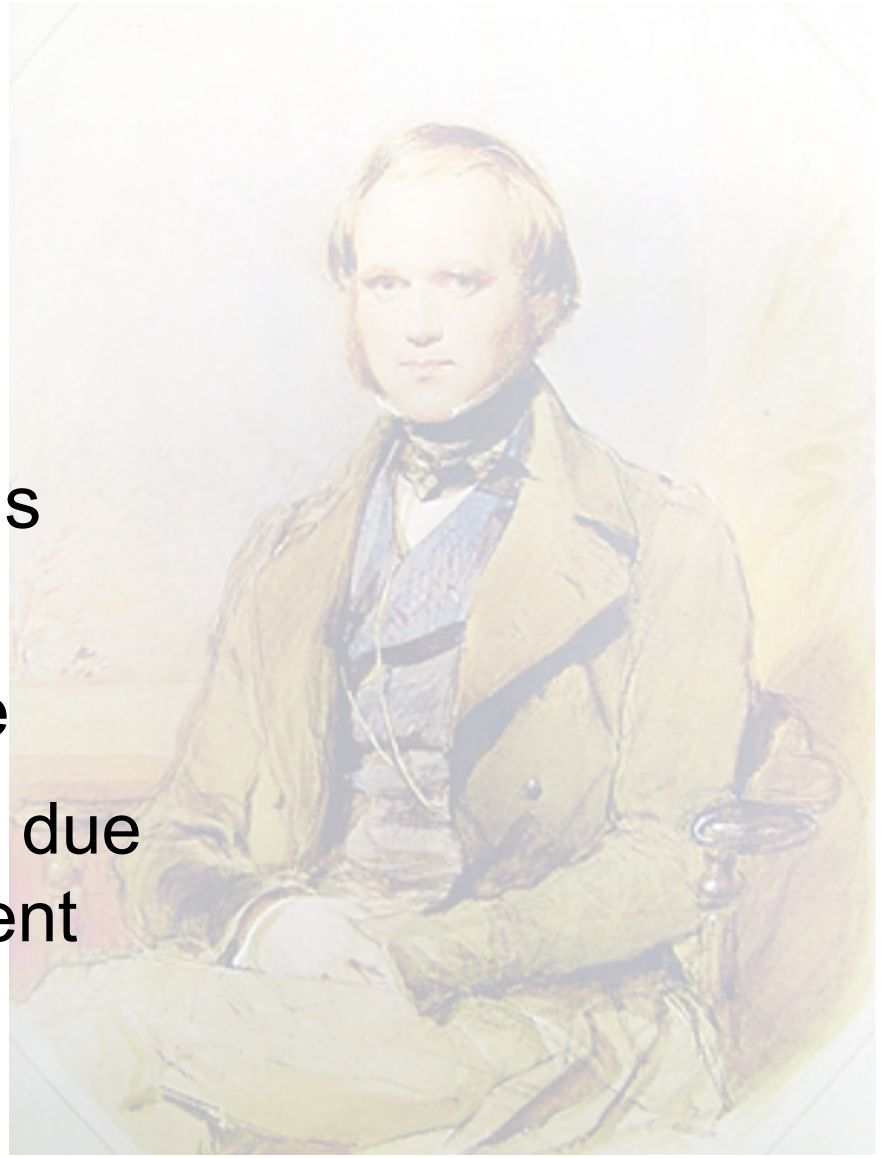
DIVERSITY MECHANISMS

How does genomic diversity spread?

- Chromosomal variation – changes to chromosome structure
- Genomic variation – changes in the chromosomal content of genomes

EVOLUTIONARY DIVERSITY

- Charles Darwin (1809 - 1882)
  - Artificial selection: crops and livestock
  - Fossil record: change over time
  - Biogeography: similarities across regions and differences within locales
- Descent with modification ⇨ continuity of life
- Natural selection: change in population traits due to individual interactions with their environment

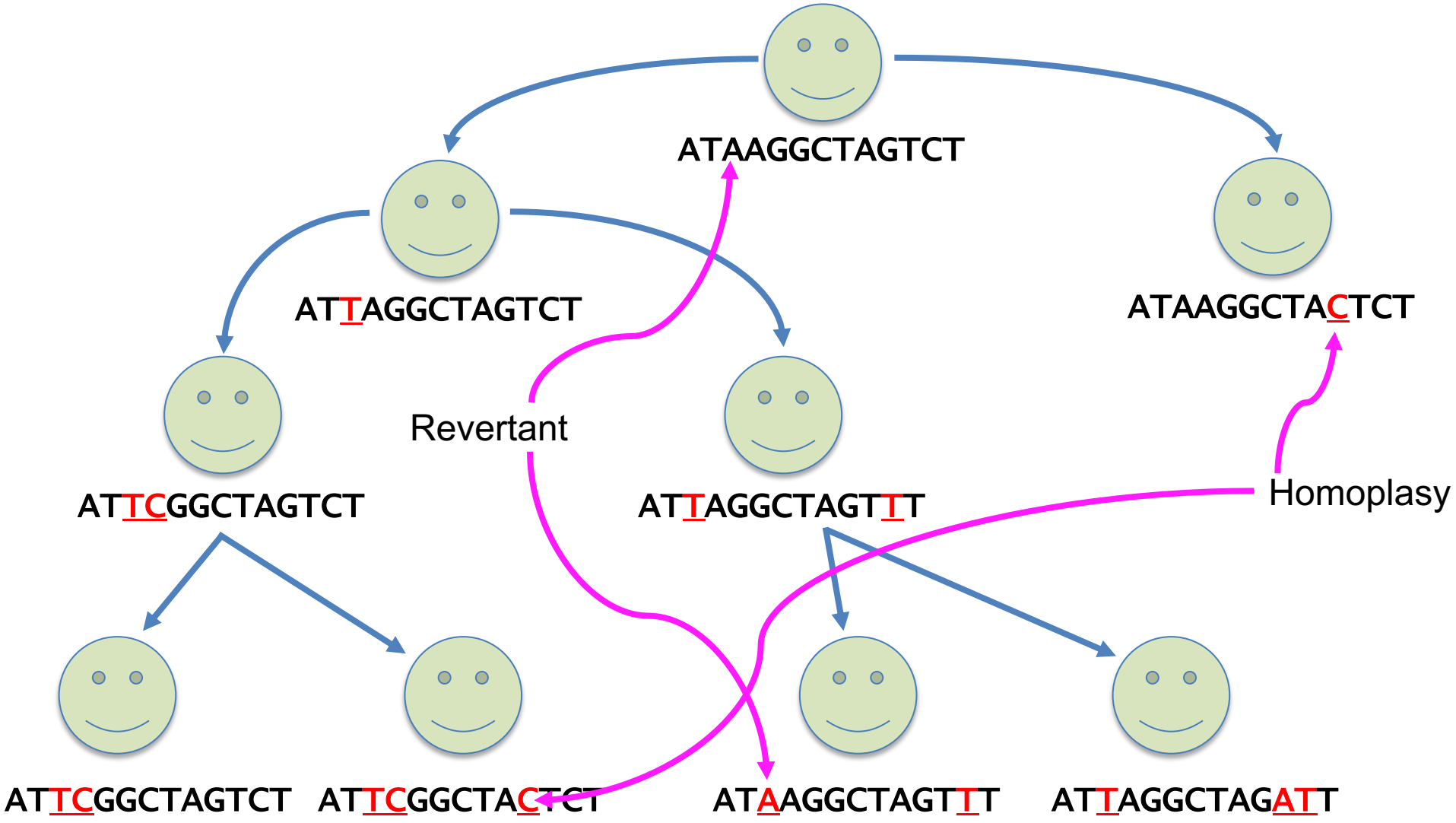


EVOLUTIONARY DIVERSITY

"Naturalists try to arrange the species, genera, and families in each class, on what is called the Natural System. But what is meant by this system?" p.413

"… I believe this element of descent is the hidden bond of connexion which naturalists have sought under the term of the Natural System" p. 433

EVOLUTIONARY DIVERSITY

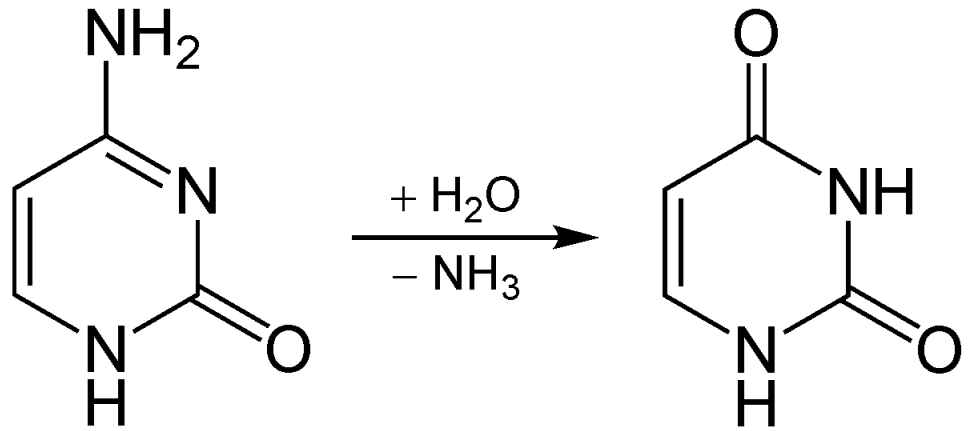


DIVERSITY MECHANISMS

- Mutation
  - Substitution
    - Deamination of cytosine to uracil
    - Transitions and transversions
- Deletion
- Insertion
- Recombination

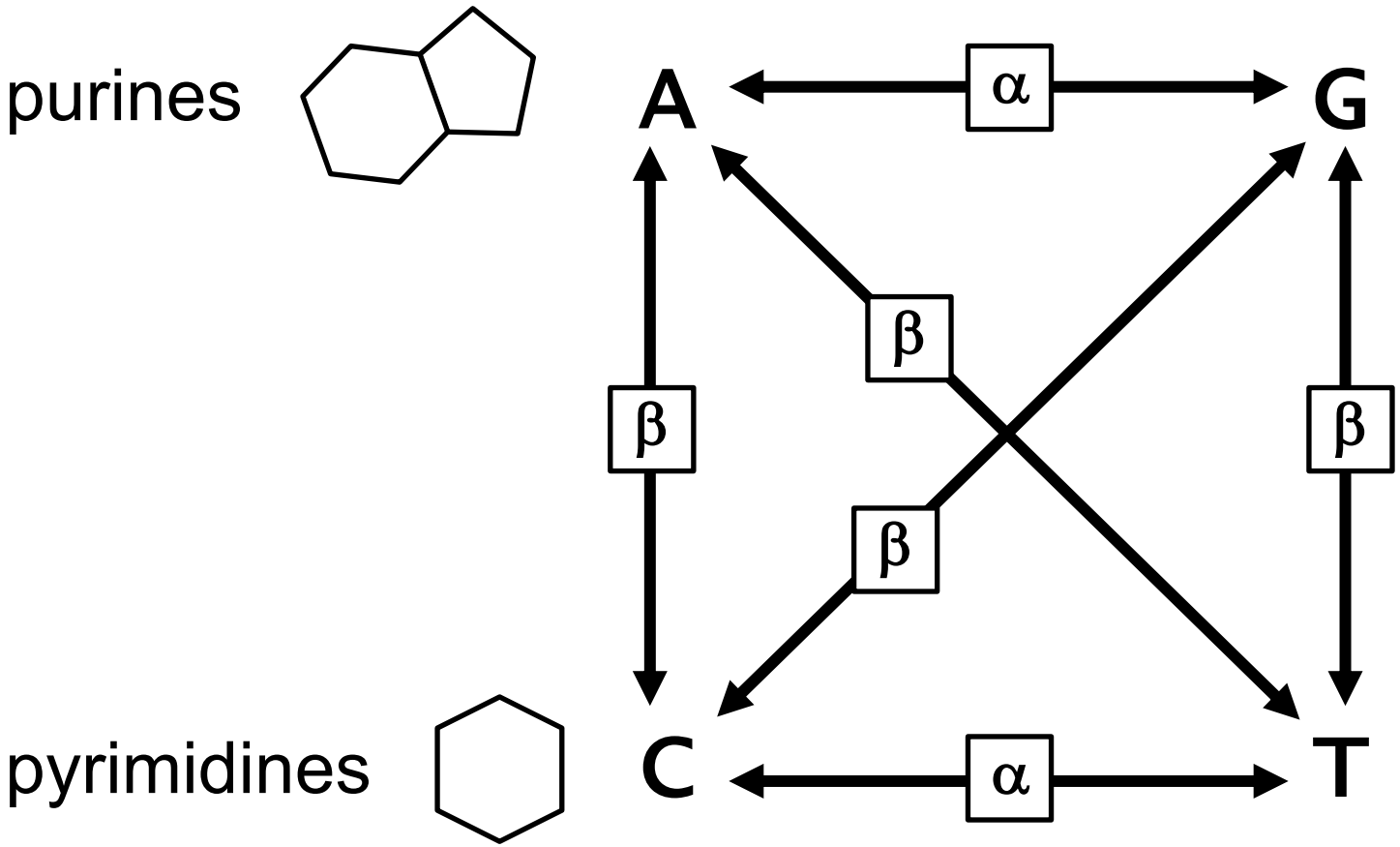
DIVERSITY MECHANISMS

Deamination of cytosine to uracil



DIVERSITY MECHANISMS

Transitions and transversions



Two substitution rates:

transition = α

transversion = β

DIVERSITY MECHANISMS

Insertion/Deletion

- Can affect single nucleotides and long sequences of nucleotides
- Single-base indels associated with single-nucleotide repeats should be investigated due to the potential to be a sequencing artifact
  - AAAAAAT vs AAAAA-T Is the 6th “A” real or an artifact?
- Example: The *Mycobacterium tuberculosis* H37Rv reference genome, when compared against the CDC1551 genome, was found to have several deletions of thousands of nucleotides containing many functional loci

DIVERSITY MECHANISMS

Insertion/Deletion

CDC1551 coordinates	Length	Gene name or product
150887-151067	180	PE_PGRS
624668-624758	90	PE_PGRS
744075-744608	533	Alpha-mannosidase
1121754-1121769	15	Hypothetical
1191505-1191697	192	PE_PGRS
1213846-1213891	45	PE_PGRS
1480513-1482187	1674	Adenylate cyclase
1612509-1612530	21	Hypothetical
1632424-1632451	27	PE_PGRS
1633446-1634201	755	PE_PGRS
1885204-1885214	10	ABC transporter
1974051-1974211	160	PPE
1978715-1985523	6808	Phospholipase C, Glycosyl transferase, Oxidoreductase, Membrane protein
1993920-1994873	953	Hypothetical
2130695-2130710	15	Hypothetical
2134757-2134767	10	Hypothetical
2143342-2143387	45	Conserved hypothetical
2160664-2160941	277	PPE
2266057-2271057	5000	Conserved hypothetical, Hypothetical, Conserved hypothetical, Helicase
2629977-2630917	940	Hypothetical, Hypothetical, PPE
2633463-2634259	796	PPE
2701714-2701735	21	Aryl sulfatase
2862694-2863350	656	Lipoprotein
3524545-3526695	2150	PPE
3685803-3685859	56	Hypothetical
3705263-3709688	4425	MoaB, MoaA, Hypothetical, Transcription regulator, Hypothetical, Transposase
3730852-3730870	18	PE_PGRS
3733433-3733511	78	PE_PGRS
3922614-3922632	18	PE_PGRS
3924305-3924313	8	PE_PGRS
3926618-3926693	75	PE_PGRS
3935210-3935555	345	PE_PGRS
3940711-3940747	36	PE_PGRS
3941109-3941184	75	PE_PGRS
4086588-4086606	18	PE_PGRS

Table 2. R. D. Fleischmann et al. J. Bacteriol. 2002; doi:10.1128/JB.184.19.5479-5490.2002

DIVERSITY MECHANISMS

The CDC1551 genome compared to the H37Rv genome

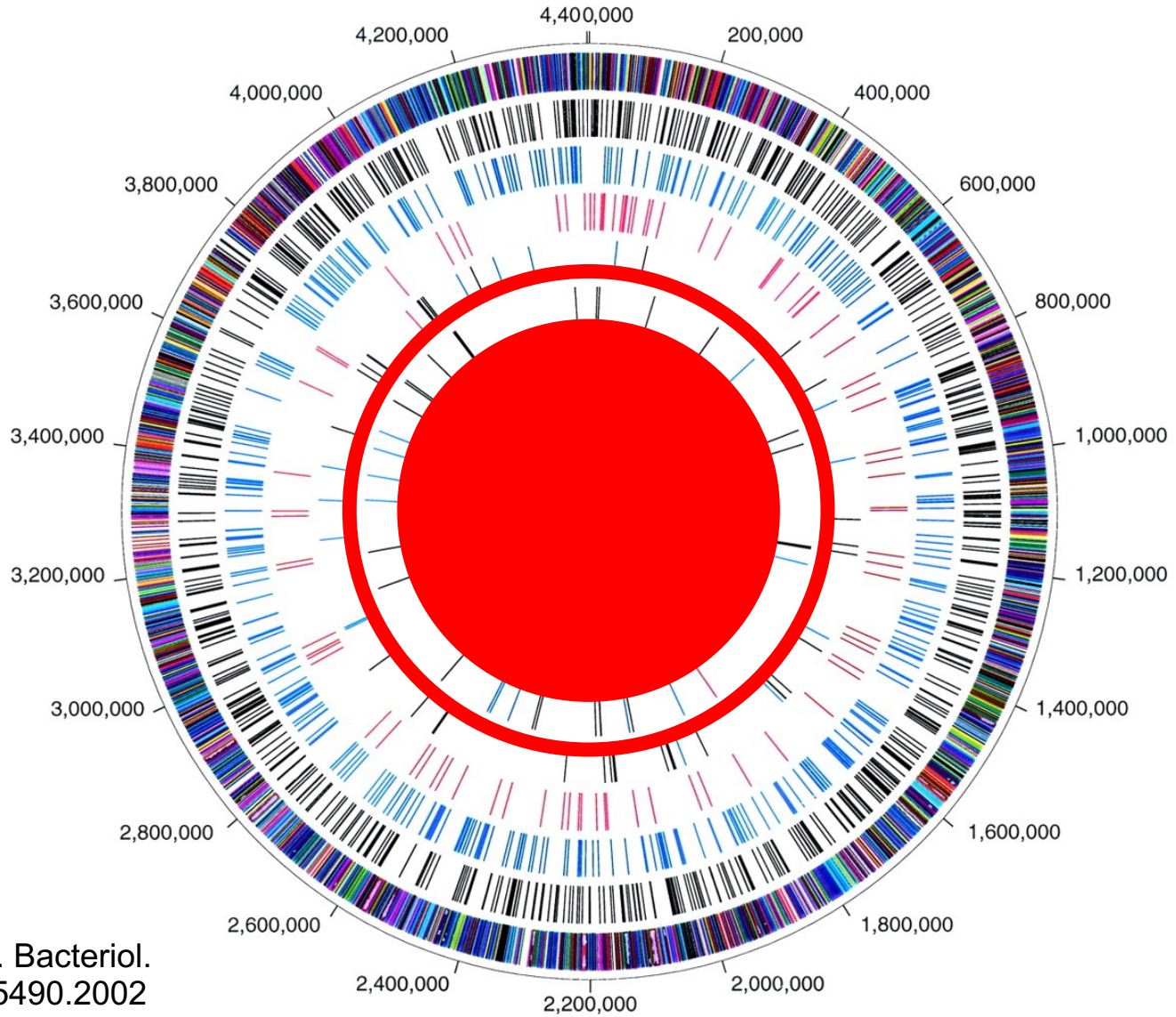
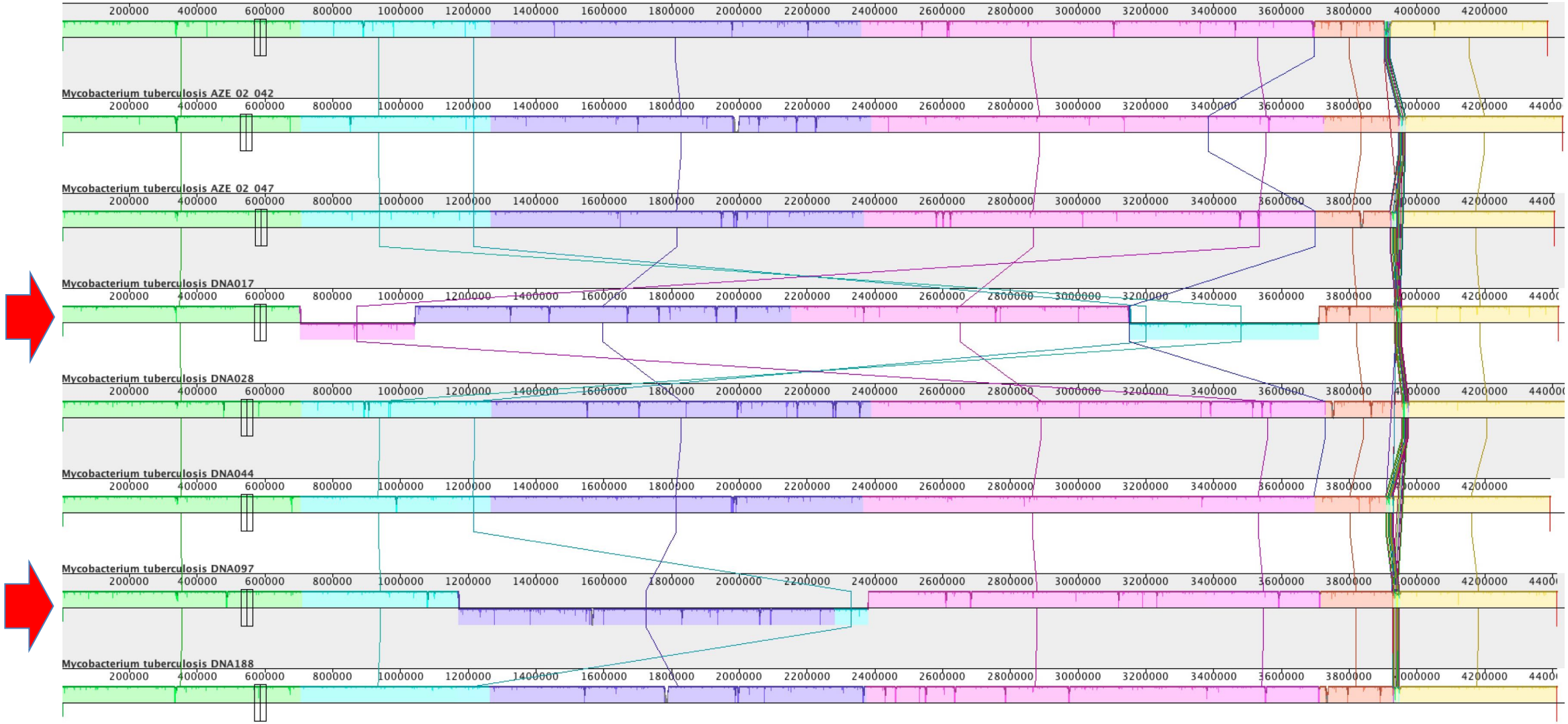


Figure 1. R. D. Fleischmann et al. J. Bacteriol. 2002; doi:10.1128/JB.184.19.5479-5490.2002

DIVERSITY MECHANISMS



Eight complete *M. tuberculosis* genomes showing rearrangements

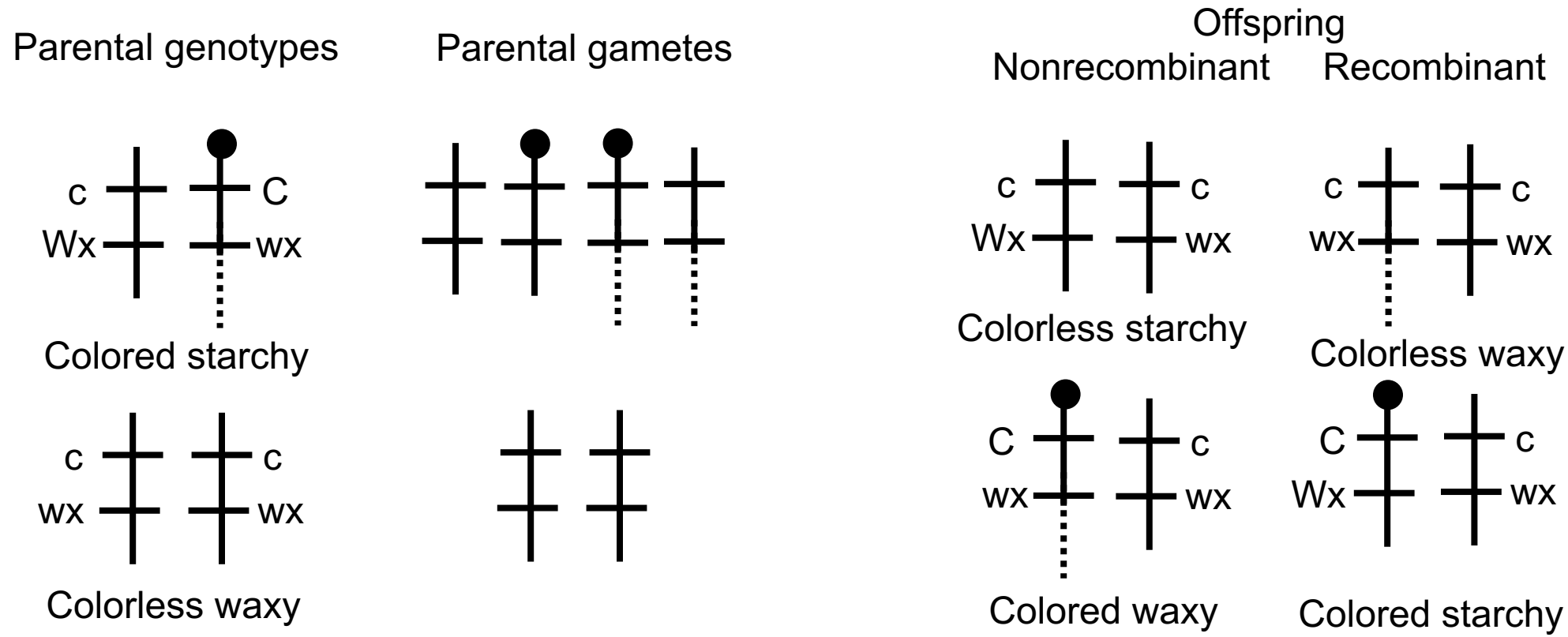
DIVERSITY MECHANISMS

Recombination

- Mostly occurs during prophase I of meiosis
- Paired homologous chromosomes (maternal and paternal) swap genetic material – “crossing over”
- First described as changes in pairing of phenotypic characteristics
- Appears as linked nucleotide variants across a region
- Analytically, subregions of sequences have divergent evolutionary histories: gene genealogies for regions do not match

DIVERSITY MECHANISMS

Recombination



DIVERSITY MECHANISMS

Repeat expansion

- Repetitive genetic elements
  - Microsatellites – 2-6 nucleotide motifs
  - Minisatellites – 10-100 nucleotide motifs
  - VNTRs – any type of repetitive element where the number of repeats varies from individual to individual

DIVERSITY MECHANISMS

Repeat expansion

- Mechanisms of change in repeat numbers
  - Replication slippage/slipped-strand mispairing
    - Occurs during DNA replication
  - Unequal sister-chromatid exchange
    - Occurs during mitosis (typically) and meiosis (rarely)
  - Unequal crossing-over
    - Occurs typically during meiosis

DIVERSITY MECHANISMS

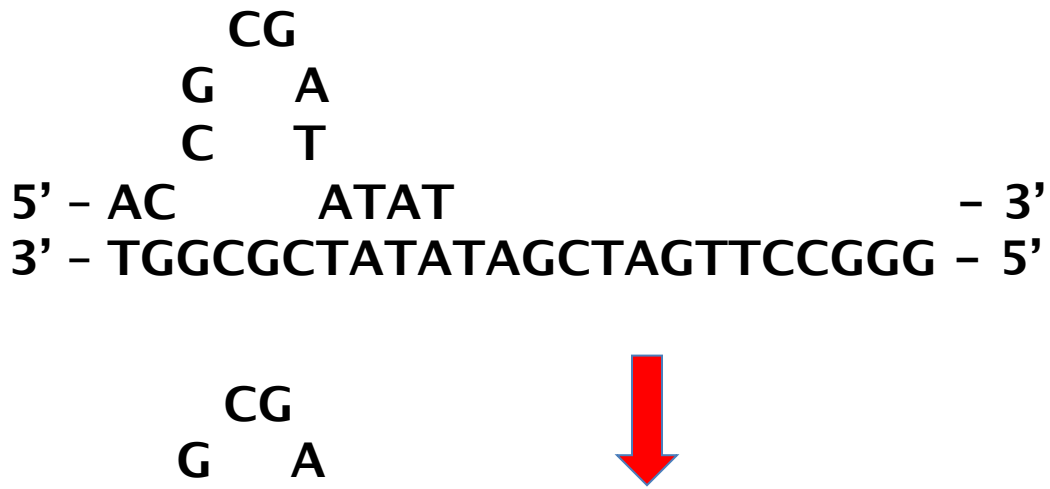
Repeat expansion

- Mechanisms of change in repeat numbers
  - Replication slippage/slipped-strand mispairing
    - Bulge in replicated strand – increased number of repeats
    - Bulge in template strand – reduced number of repeats

DIVERSITY MECHANISMS

Replication slippage/slipped-strand mispairing

5’ – ACCGCGATATAT – 3’
3’ – TGGCGCTATATAGCTAGTTCCGGG – 5’



DIVERSITY MECHANISMS

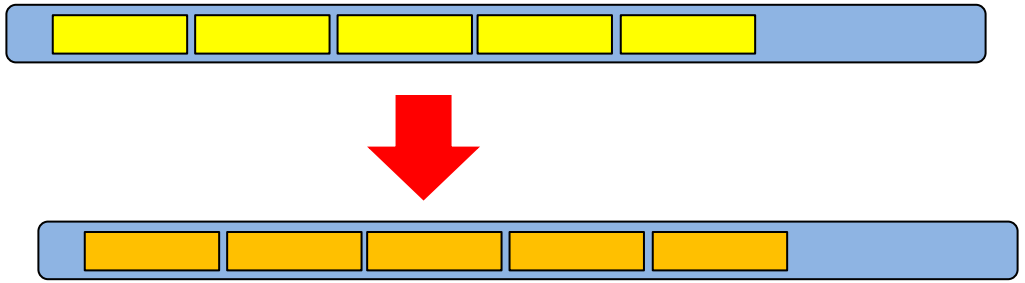
Repeat expansion

- Mechanisms of change in repeat numbers
  - Unequal sister-chromatid exchange
    - Occurs during mitosis (typically) and meiosis (rarely)
    - Occurs between the original chromosome and the replicated chromosome (sister chromatids)
    - Results in reduced number of repeats in one chromatid and an increased number of repeats in the other chromatid

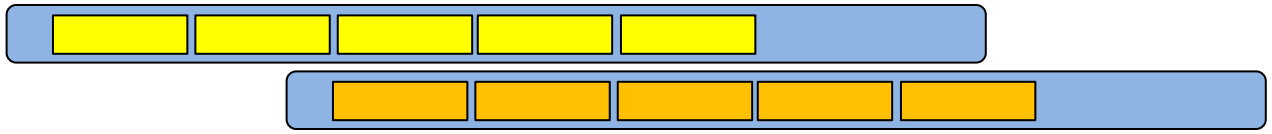
DIVERSITY MECHANISMS

Unequal sister-chromatid exchange

1: S-phase chromatid duplication



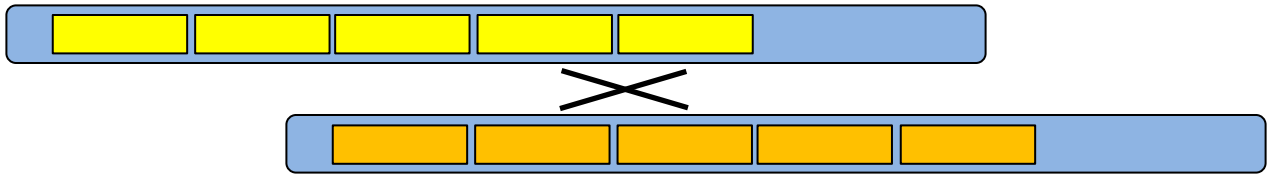
2: Mismatching of repeated loci



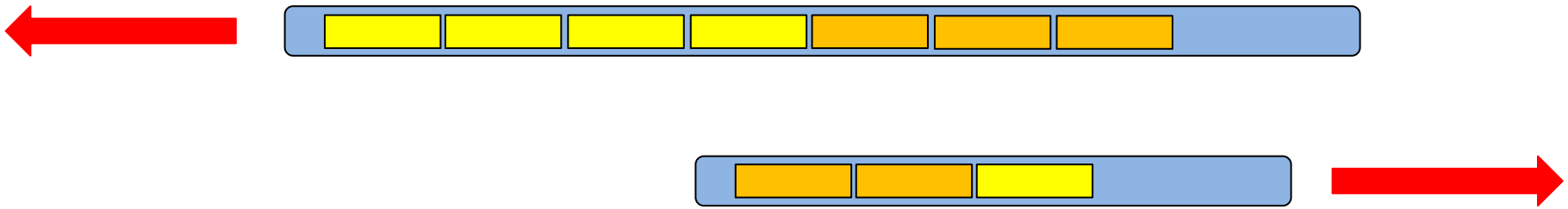
DIVERSITY MECHANISMS

Unequal sister-chromatid exchange

3: Homologous recombination of mismatched chromatids



4: Unequal sister chromatids segregate to different daughter cells



DIVERSITY MECHANISMS

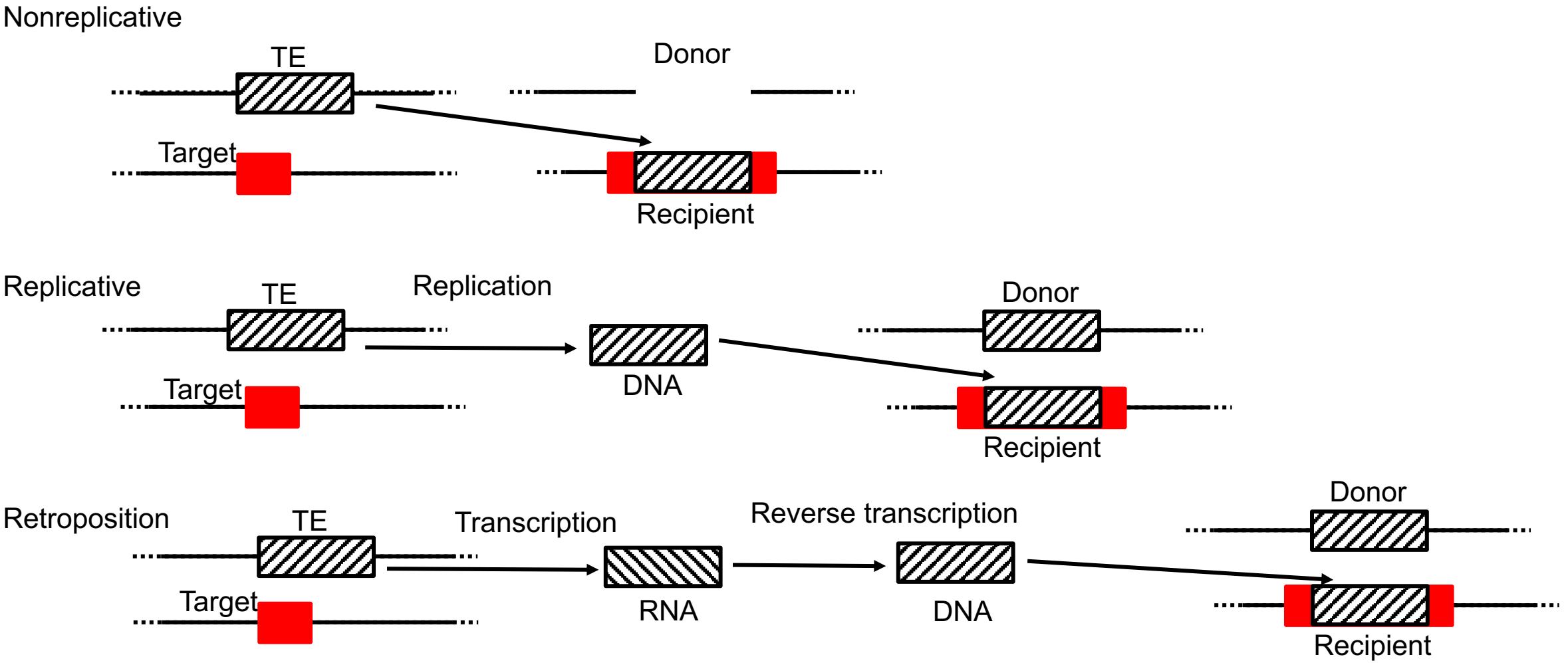
- Genetic Duplication Events
  - Transposition
  - Whole gene – gene families
  - A part of a chromosome (multiple loci)
  - Whole chromosome
  - Whole genome

DIVERSITY MECHANISMS

- Genetic element duplication – Transposition
- Discovered by Barbara McClintock in maize (1940s)
  - Nonreplicative transposition – the element itself transposes
  - Replicative transposition – element copies are transposed
    - Direct DNA transposition
    - Transposition using an RNA intermediary - Retroposition

DIVERSITY MECHANISMS

Genetic element duplication – Transposition



DIVERSITY MECHANISMS

Transposable Elements (TE)

- TE insertion results in short (4-12 bp) repeats flanking the element
  - If repeats have the same orientation they are called direct repeats
- Classes of TEs
  - Insertion Sequences
  - Transposons
  - Retroelements

DIVERSITY MECHANISMS

Transposable Elements (TE)

- Insertion Sequences
  - Found in bacteria, bacteriophages, and plasmids (but also McClintock's Maize Controlling Elements)
  - Usually small, 700-2500 bp
  - Carry genes necessary for transposition (transposases) at a minimum

DIVERSITY MECHANISMS

Transposable Elements (TE)

- Transposons
  - Found in prokaryotes and eukaryotes
  - Larger (2500-7000 bp) and more complex
  - Quite variable genomic structure
    - *P* elements in *Drosophila* contain introns
  - Bacterial transposons often carry genes for antibiotic, heavy metal, and heat resistance
  - Some bacteriophages (*Mu*) are transposons, and contain genes for the proteins needed to package the virus

DIVERSITY MECHANISMS

Transposable Elements (TE)

- Retroelements
  - Retroviruses are retroelements containing the *gag*, *pol*, and *env* genes
    - *gag*: polyprotein processed into nucleocapsid and virion matrix proteins
    - *pol*: polyprotein processed into reverse transcriptase, Rnase H, integrase, and aspartate proteinase
    - *env*: the envelope protein
    - Many retroviruses contain additional genes
  - Retrotransposons are retroelements with long terminal repeats (LTRs)
    - Similar to retroviruses but lack *env* so cannot infect other cells
    - Extremely variable in structure and composition

DIVERSITY MECHANISMS

Transposable Elements (TE)

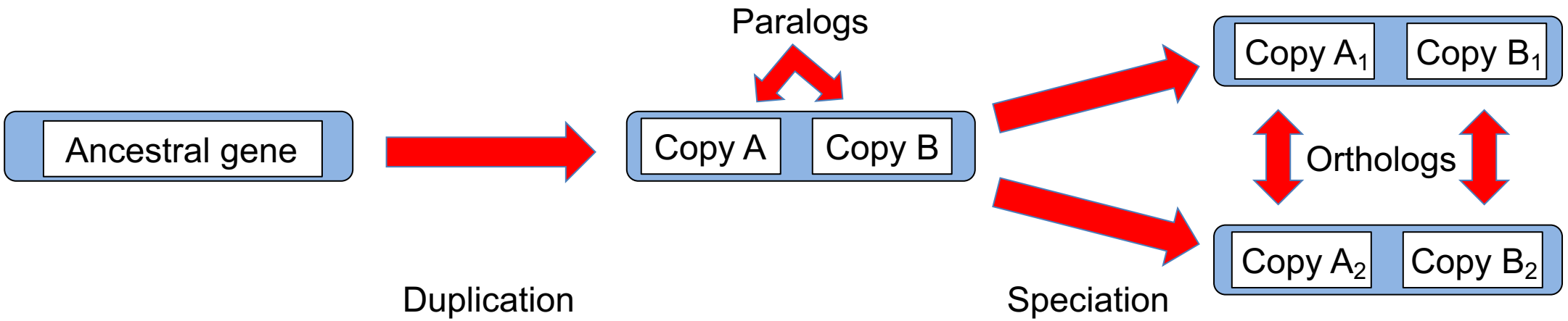
- Retroelements
  - Retroposons – similar to retrotransposons but lack LTRs
    - Extremely variable group
    - Use reverse transcriptase and can transpose natively
  - Pararetroviruses – a family of DNA viruses
    - Use reverse transcription for replication but cannot transpose natively
    - Appear to share a common origin with retroviruses
    - Contain LTRs
    - Hepatitis B and cauliflower mosaic virus are pararetroviruses

DIVERSITY MECHANISMS

- Gene Duplication
  - Gene families: orthologs and paralogs
  - Generation of pseudogenes
  - Subfunctionalization
  - Ex: Recruitment and modification of trypsinogen in Antarctic notothenoid fishes to produce antifreeze glycoprotein
- Inversion
- Translocation
- Domain Shuffling

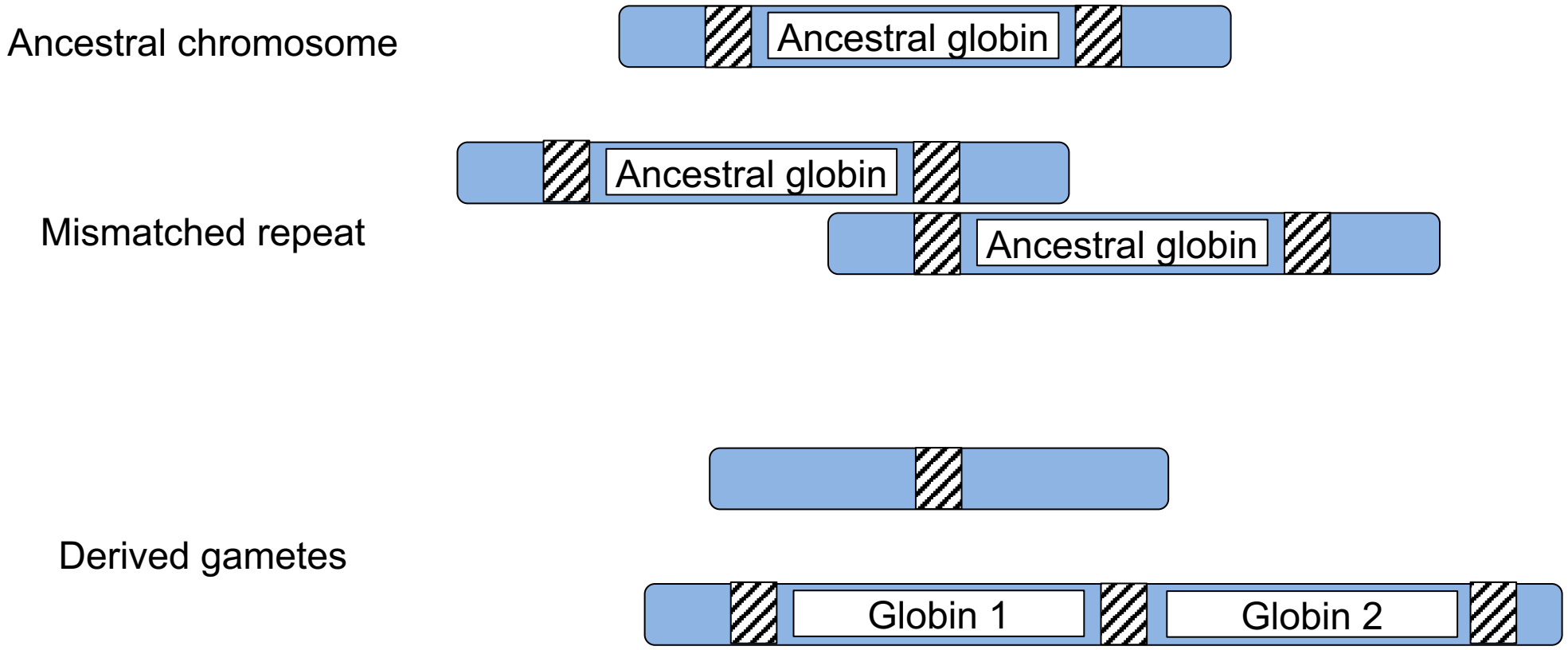
DIVERSITY MECHANISMS

- Gene Duplication
  - Gene families: orthologs and paralogs
    - Orthologs: genes derived from speciation
    - Paralogs: genes derived from gene duplication



DIVERSITY MECHANISMS

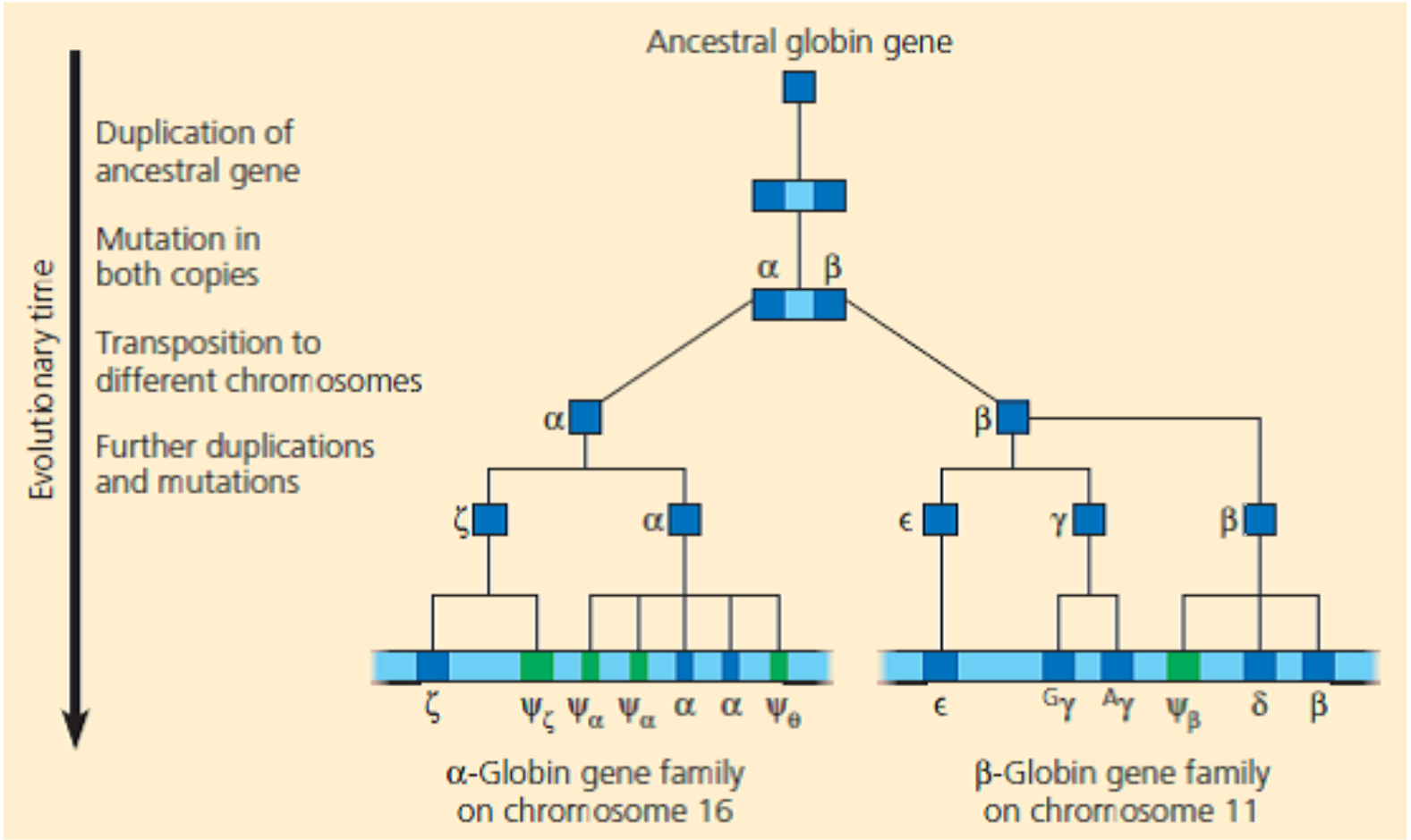
Gene Duplication – The globin genes



Ridley, Evolution, 1993

DIVERSITY MECHANISMS

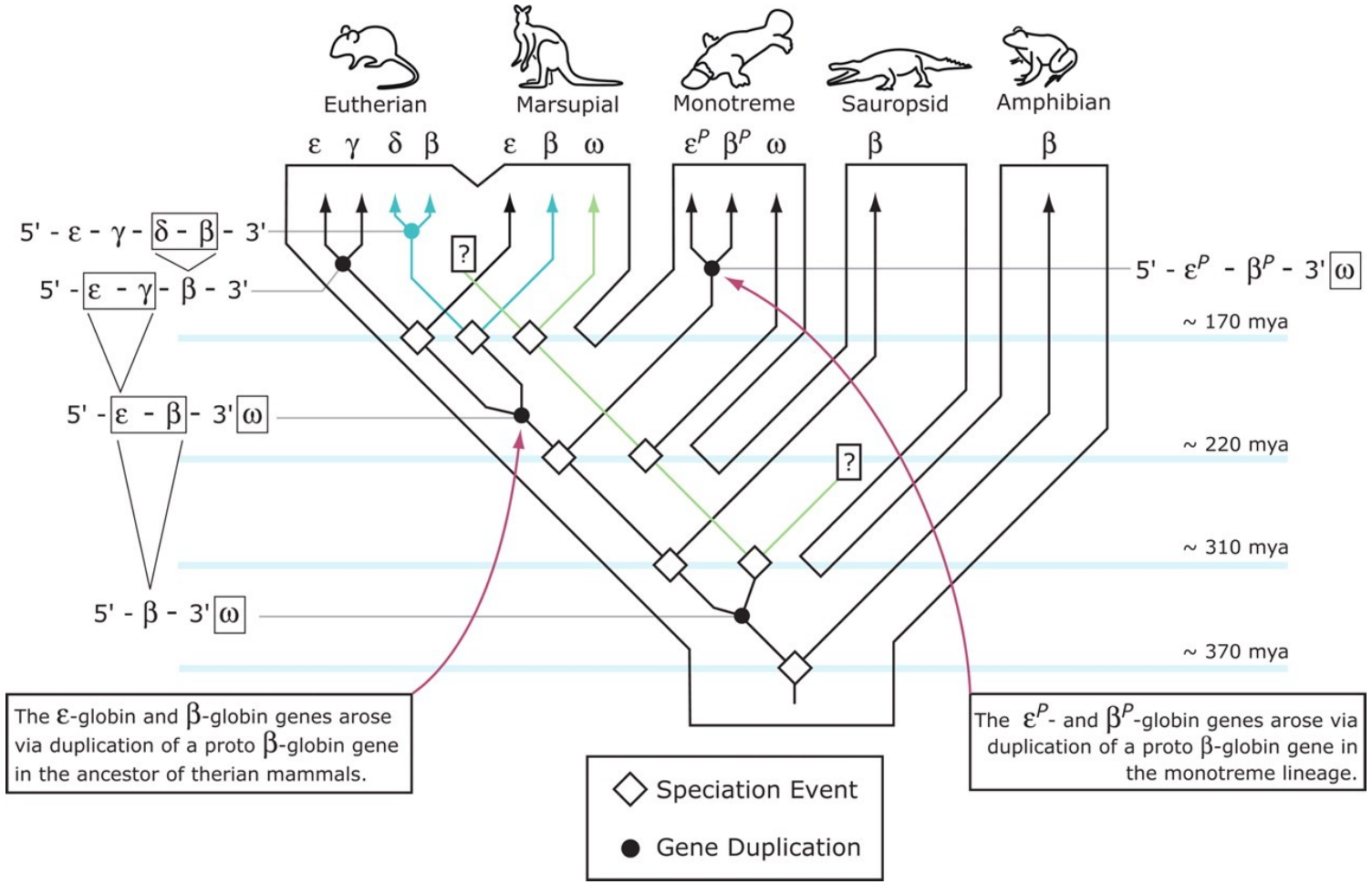
Gene Duplication – The globin genes



Campbell Biology, 11th edition

DIVERSITY MECHANISMS

Gene Duplication – The globin genes



JC Opazo, et al. PNAS 2008, doi.org/10.1073/pnas.0710531105

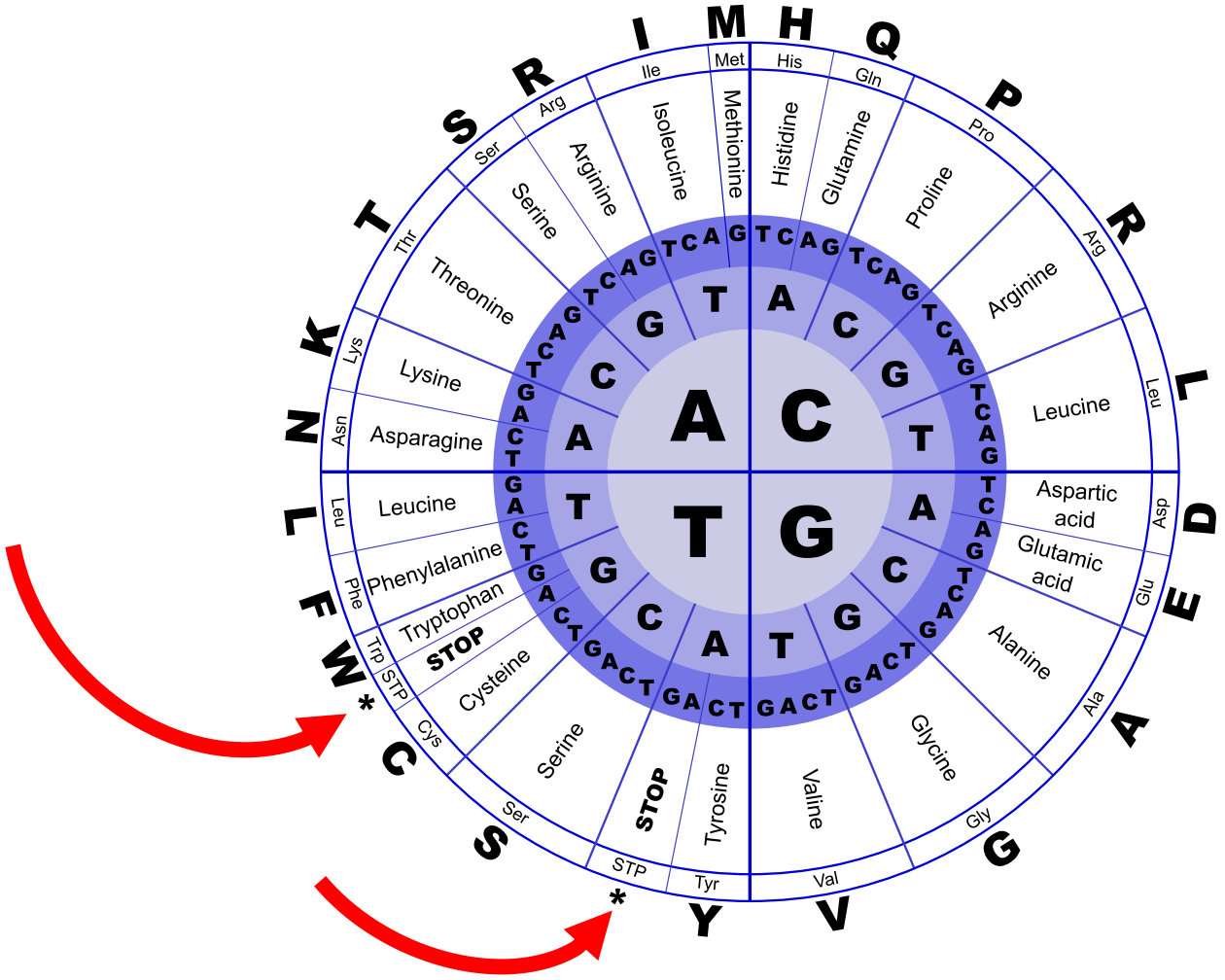
DIVERSITY MECHANISMS

Gene Duplication – The fate of duplicated genes

- Functional divergence – independent accumulation of substitutions leading to changes in protein structure and function
- Loss of function - pseudogenes
  - Generation of premature stop codon
    - Substitution
    - Insertion/deletion → frame shift
  - Loss of “ATG” start codon
  - Missense mutation
  - Frameshift mutation
  - Regulatory sequence mutation

DIVERSITY MECHANISMS

Substitution to a STOP codon



DIVERSITY MECHANISMS

Substitution to a STOP codon

Original codon - AA	**A**GA - R	**A**A(G/A) - L
	CGA - R	**C**A(G/A) - Q
	GGA - G	**G**A(G/A) - E
	~~T**A**A - Stop~~	T**C**(G/A) - S
	T**C**A - S	~~T**G**A - Stop~~
	T**T**A - L	T**G**G - W
	TG**C** - C	T**T**(G/A) - L
	TG**G** - W	TA**C** - Y
	TG**T** - C	TA**T** - Y
Resulting STOP	TGA - Stop	TA(G/A) - Stop

DIVERSITY MECHANISMS

- Gene Duplication
  - Subfunctionalization: complementary, degenerative mutations in gene regulatory regions

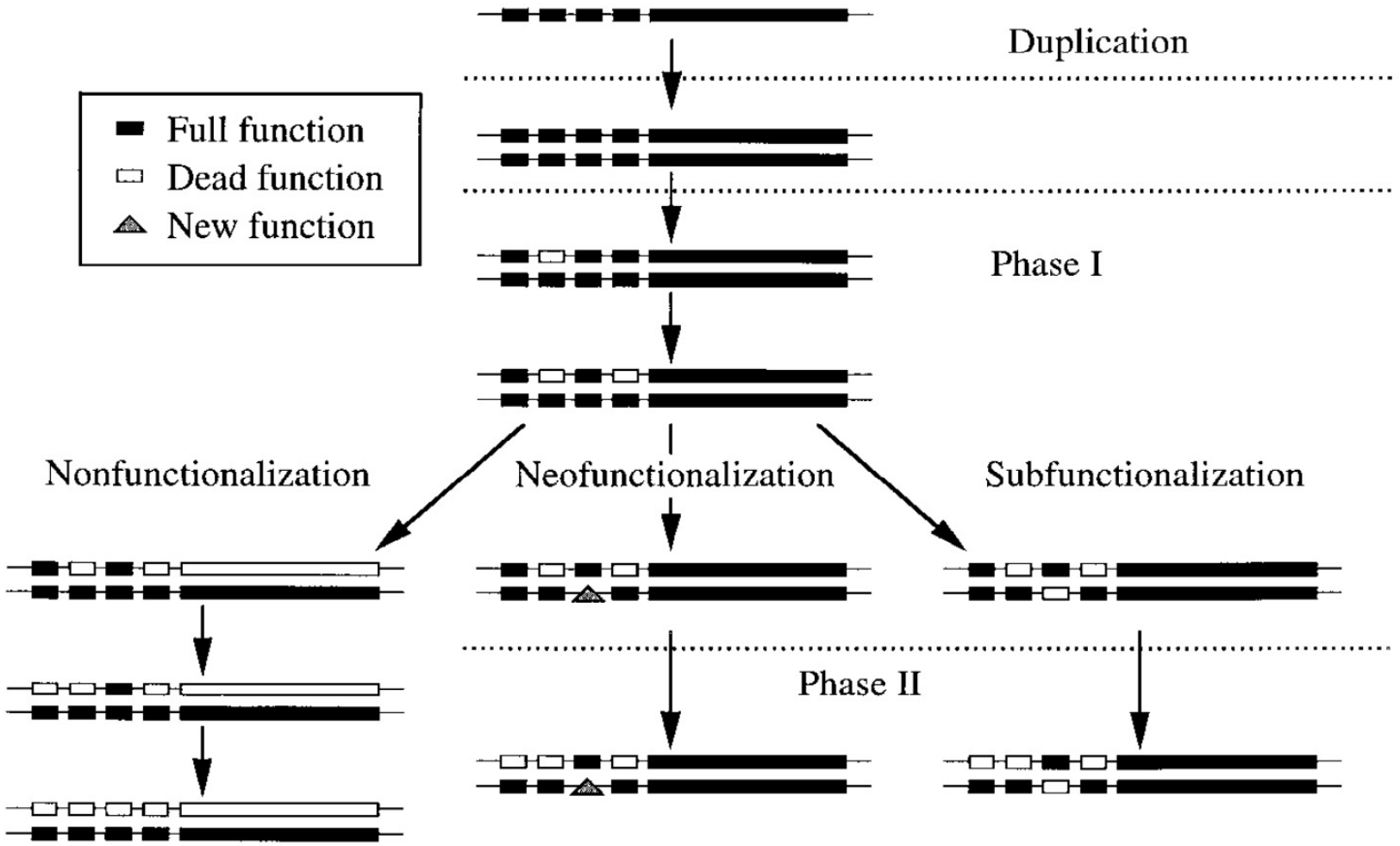


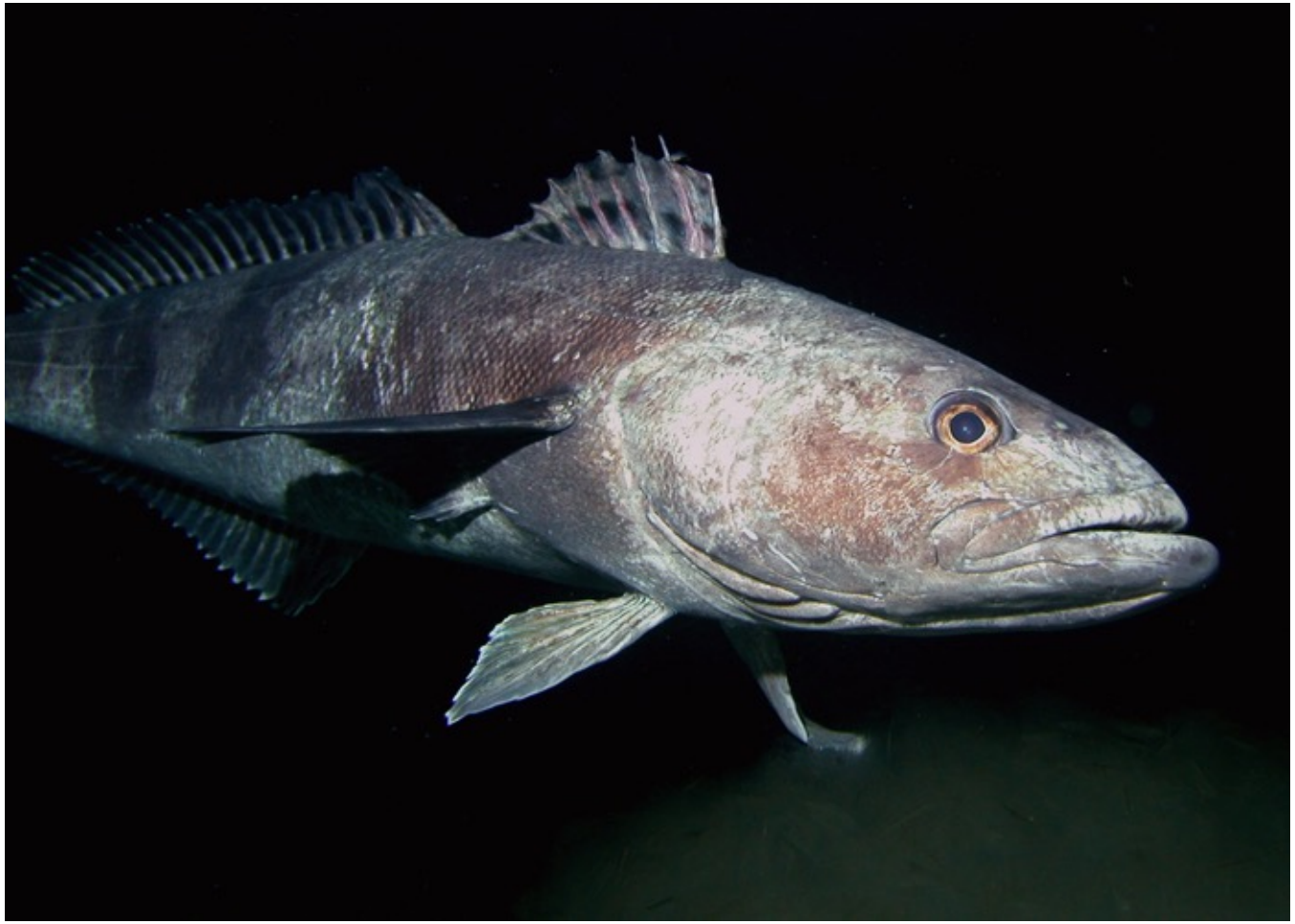
Figure 1. Force, et al. Genetics 1999

DIVERSITY MECHANISMS

Recruitment of a gene for a novel function: Antarctic fish antifreeze protein

Antarctic toothfish
Dissostichus mawsoni

Lives in subfreezing waters around Antarctica, yet its blood does not freeze



DIVERSITY MECHANISMS

Recruitment of a gene for a novel function: Antarctic fish antifreeze protein

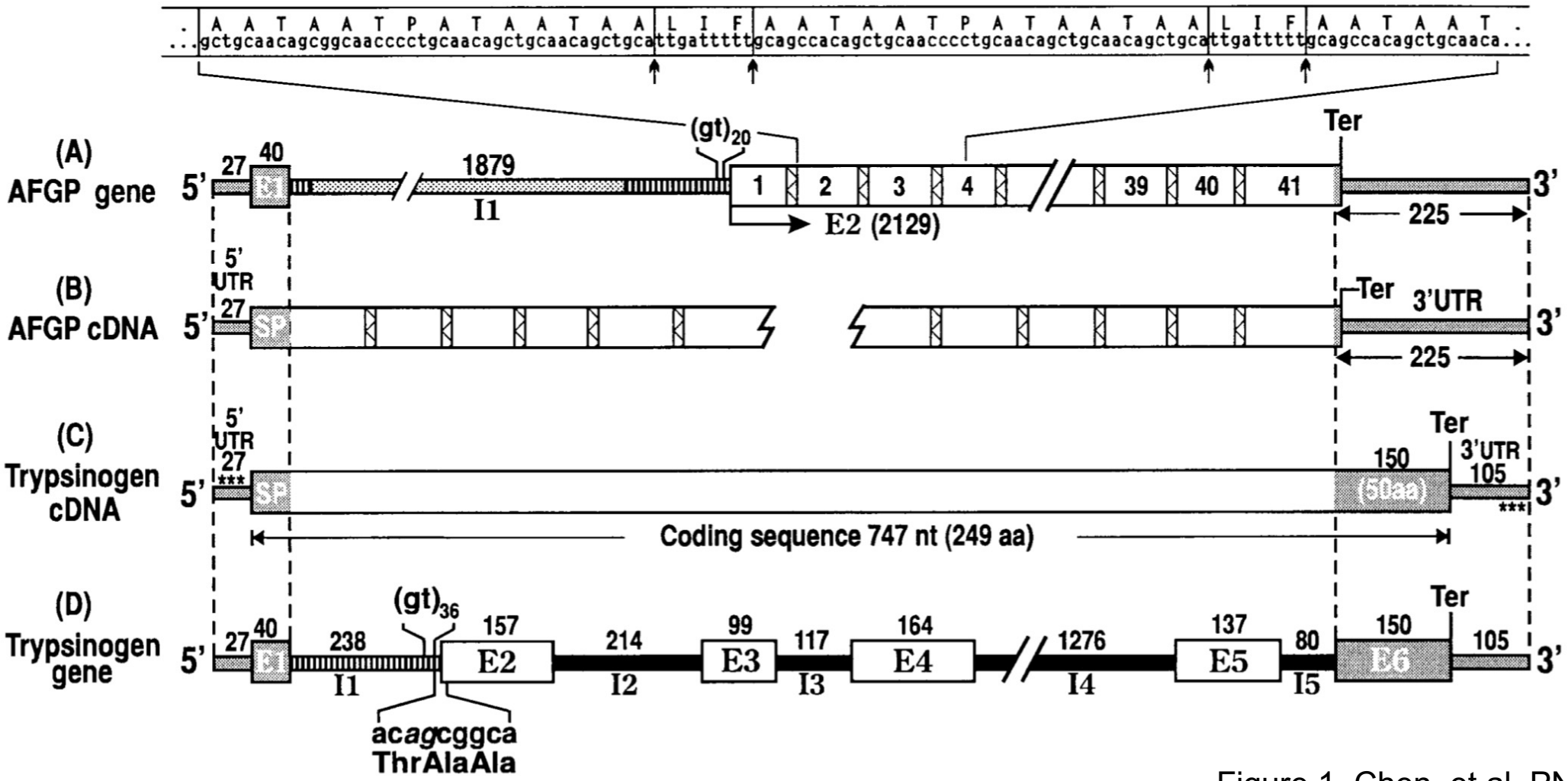
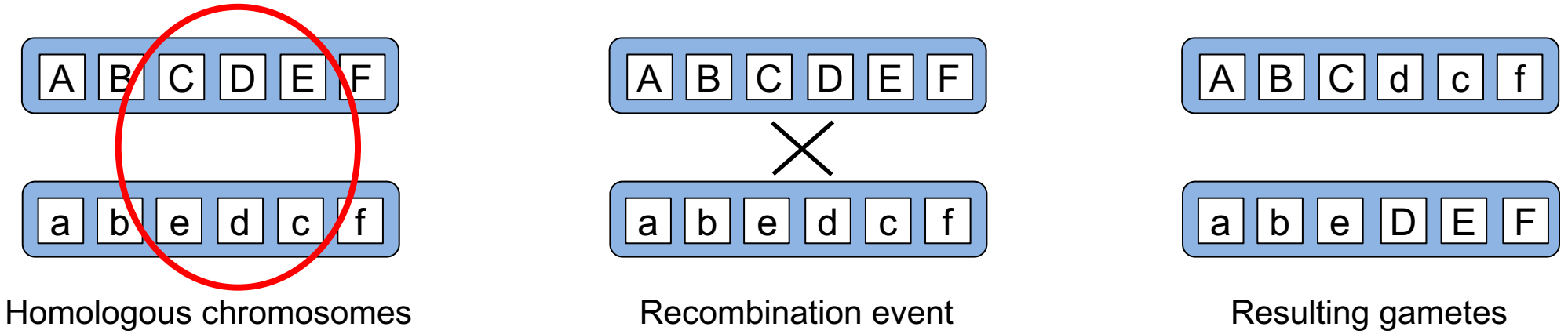


Figure 1. Chen, et al. PNAS 1997

DIVERSITY MECHANISMS

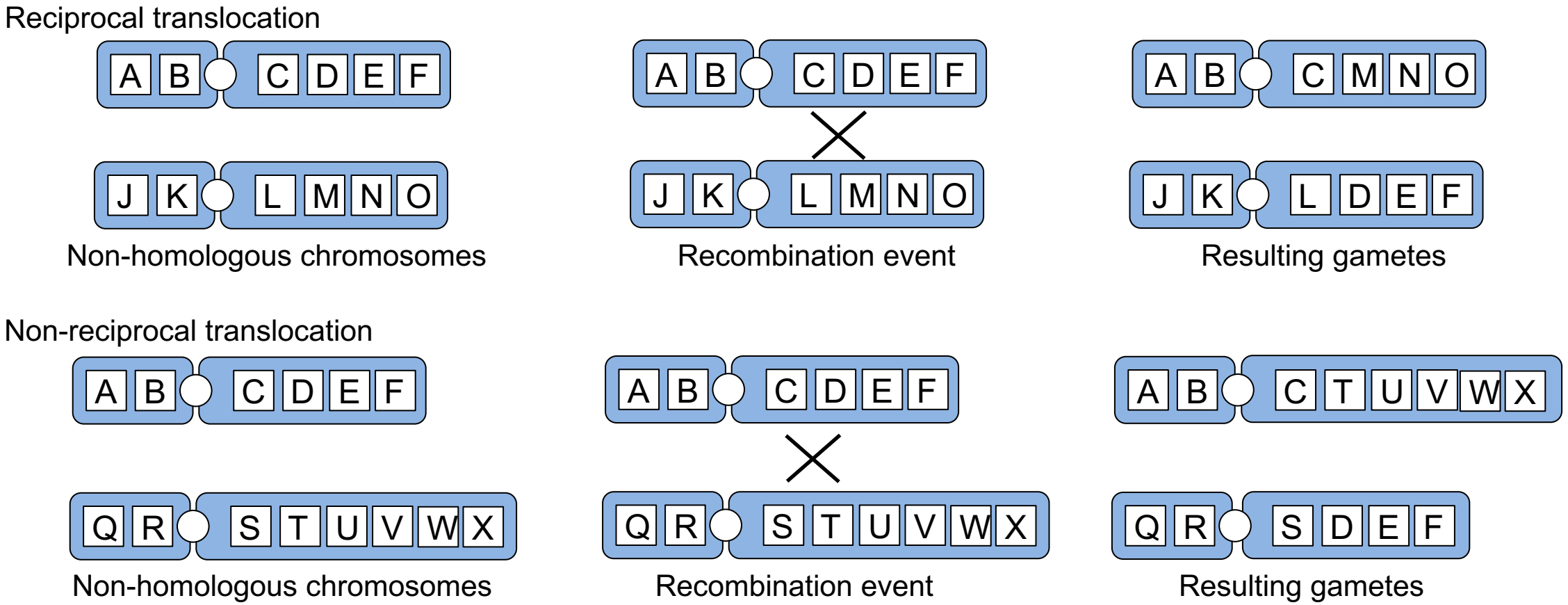
- Chromosome inversions
  - Chromosome region with multiple genes in reverse order with respect to the homologous chromosome region
  - Recombination within an inversion is suppressed because both recombinants will be missing entire genes



Homologous chromosomes | Recombination event | Resulting gametes

DIVERSITY MECHANISMS

- Chromosomal translocations
  - Mispairing of chromosomes during recombination



DIVERSITY MECHANISMS

- Domain shuffling
  - Protein functional domains
  - Protein structural domains

DIVERSITY MECHANISMS

Protein functional domains – the bHLH domain

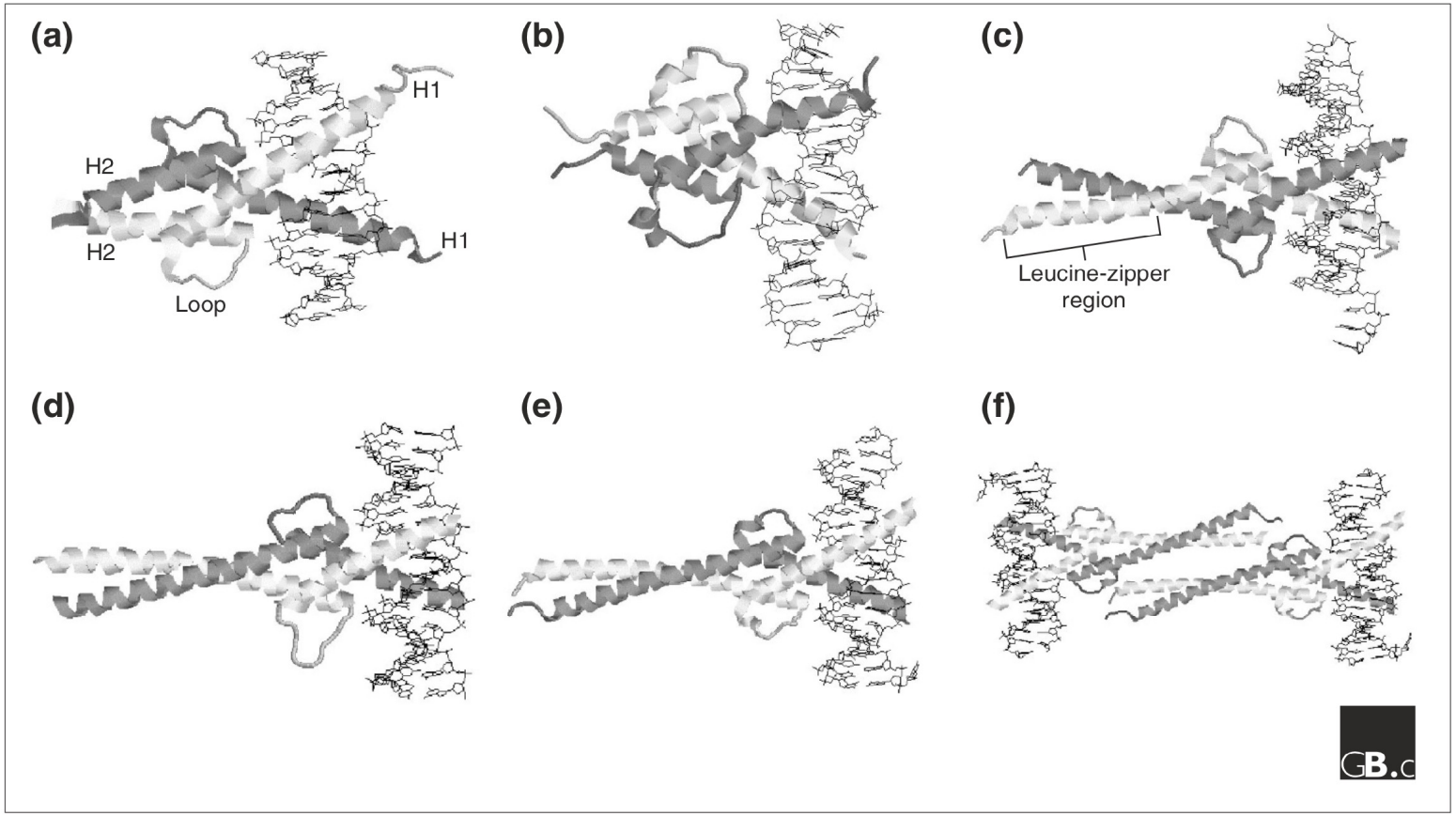
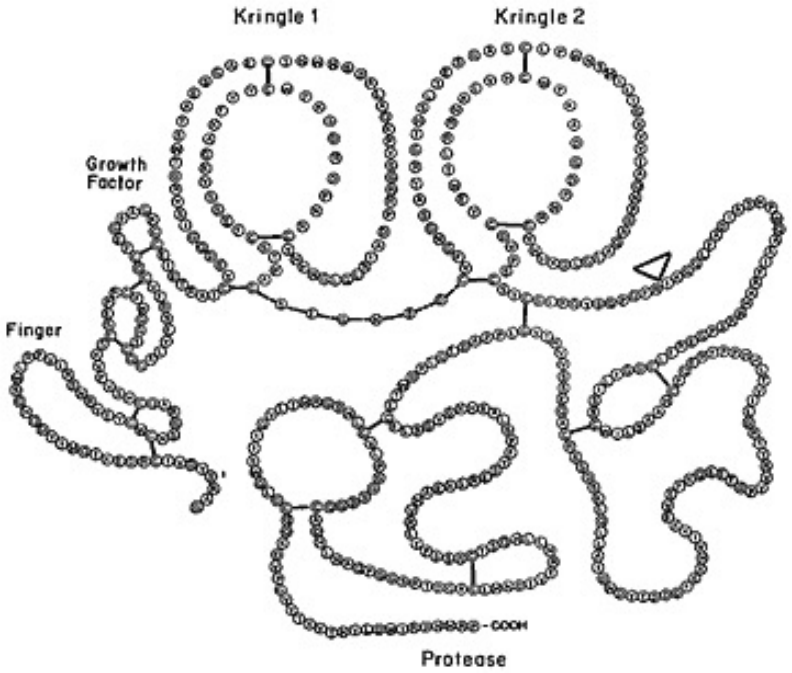


Figure 1
Representative structures of bHLH proteins from the Protein Data Bank [22]. In each diagram, the protein is shown as a secondary-structure cartoon and the DNA double helix is shown in stick representation. **(a)** MyoD bHLH-domain homodimer (PDB code 1mdy). **(b)** Pho4 bHLH-domain homodimer (1am9). **(c)** SREBP-1a bHLH-domain homodimer (1aoaC). **(d)** Max-Mad heterodimer (1nlw). **(e)** Max-Myc heterodimer (1nkp). **(f)** Max-Myc heterotetramer (1nkp). In (d-f) the Max HLH monomer is shown in dark gray. The scales are not comparable between different structures.

Jones, Genome Biology 2004

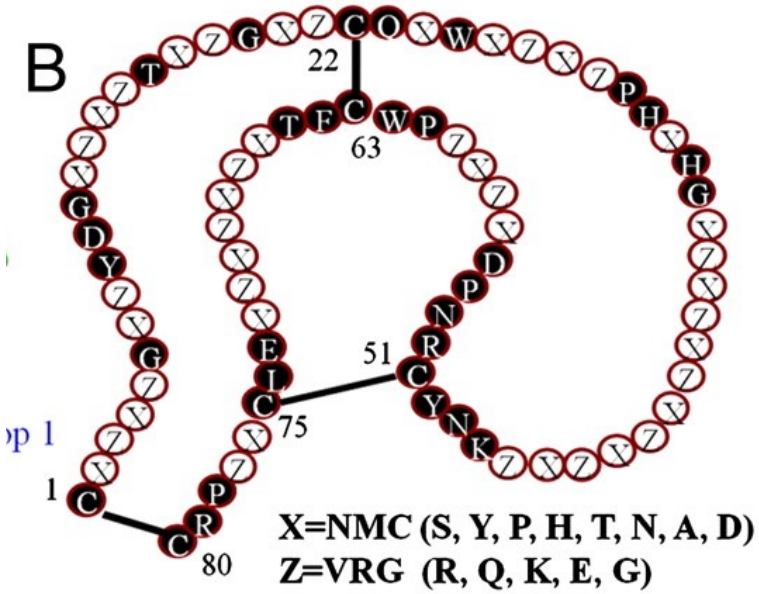
DIVERSITY MECHANISMS

Protein domain shuffling – tissue plasminogen activator



Byeon and Llinás J Mol Biol 1991

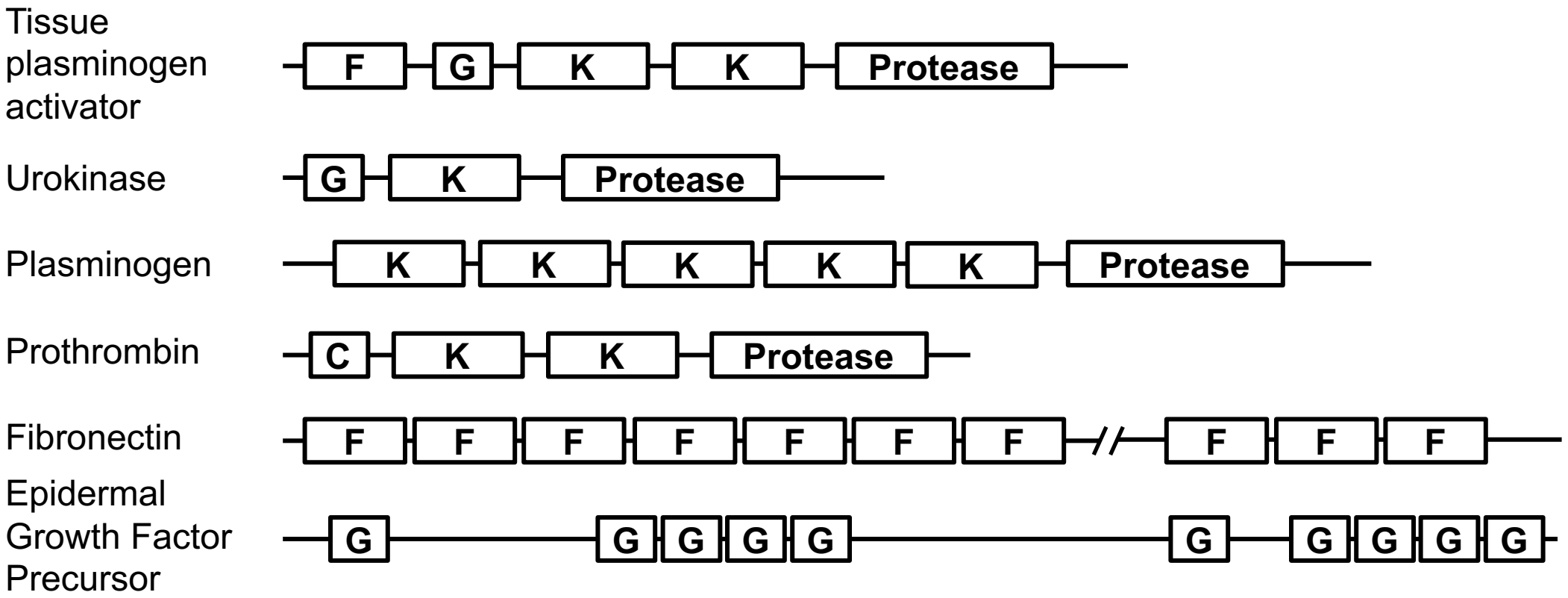
The Kringle Domain



Lee, et al.PNAS 2010 https://doi.org/10.1073/pnas.1001541107

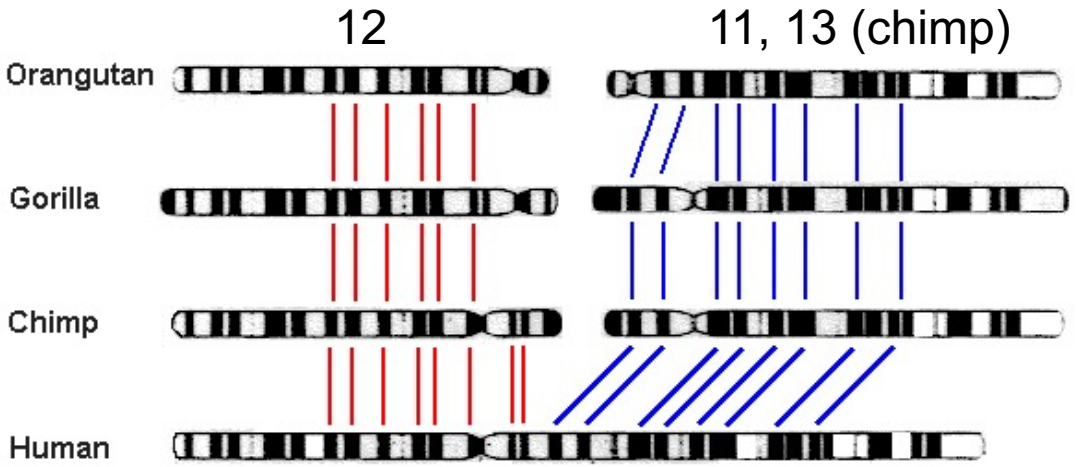
DIVERSITY MECHANISMS

Protein domain shuffling – mosaic proteins



DIVERSITY MECHANISMS

- Chromosome breaking and merging
  - Humans have 46 chromosomes (32 pairs)
  - Chimpanzees, gorillas, and orangutans have 48
  - Human chromosome 2 corresponds to two ape chromosomes

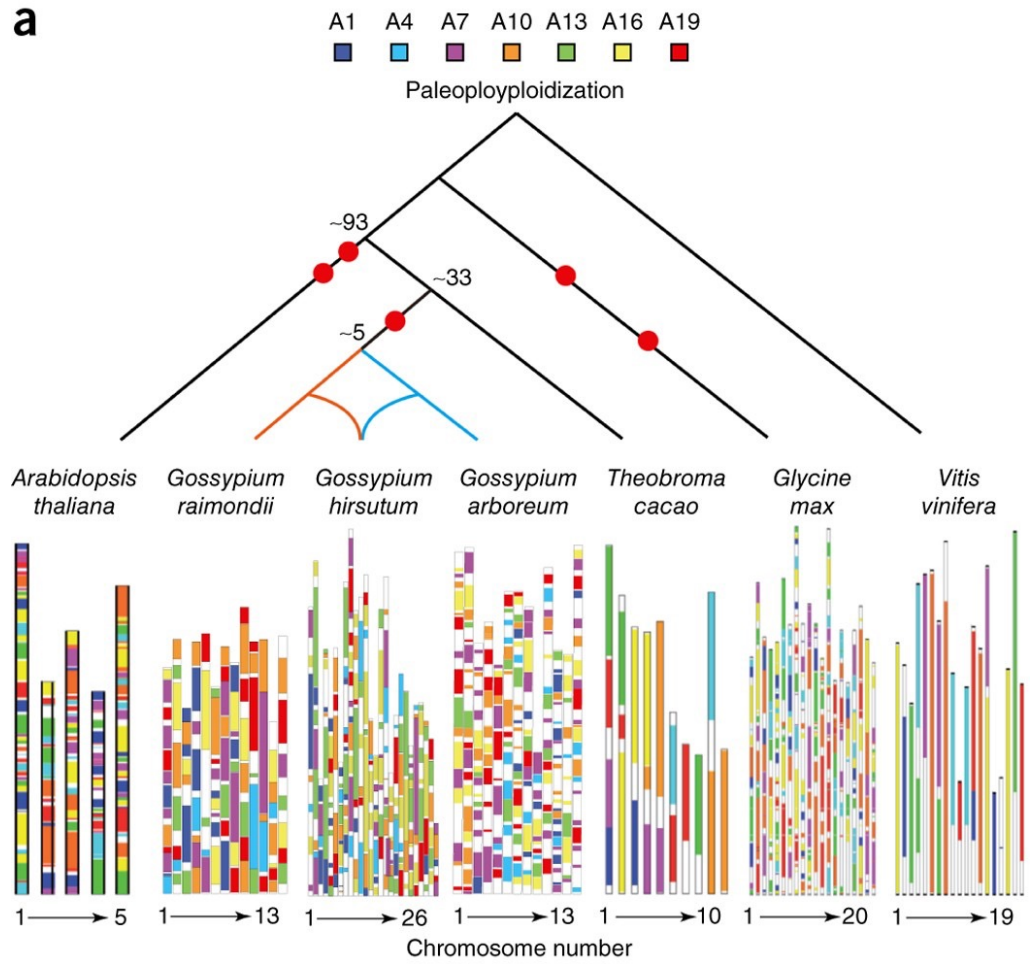


DIVERSITY MECHANISMS

- Genome duplication – complete doubling of all chromosomes
  - Polyploidy
    - Many species of plants → cotton
  - Salmonid fishes
  - Vertebrate homeobox genes

DIVERSITY MECHANISMS

Genome duplication – cotton polyloidy

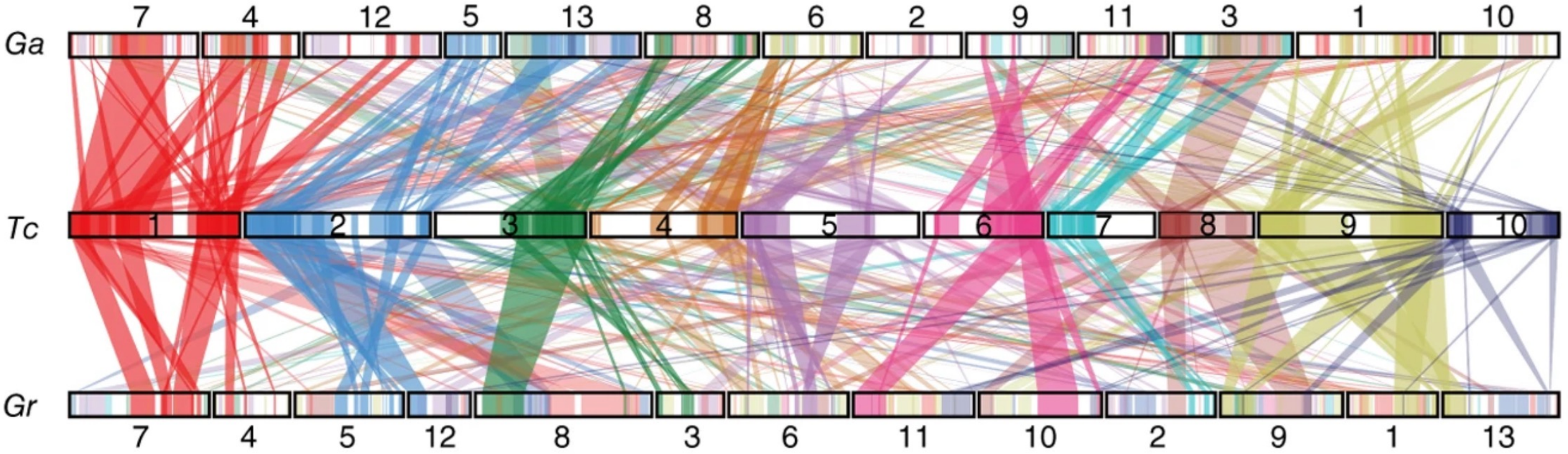


Comparison of syntenic blocks between the *G. raimondii*, *G. arboreum*, and hybrid *G. hirsutum* genomes with the genomes of four other eudicot genomes. The breaking up of syntenic blocks to different chromosomes is evidence of genome duplication events in these lineages.

Li et al. Nat Biotech 2015 doi:10.1038/nbt.3208

DIVERSITY MECHANISMS

Genome duplication – cotton polyloidy



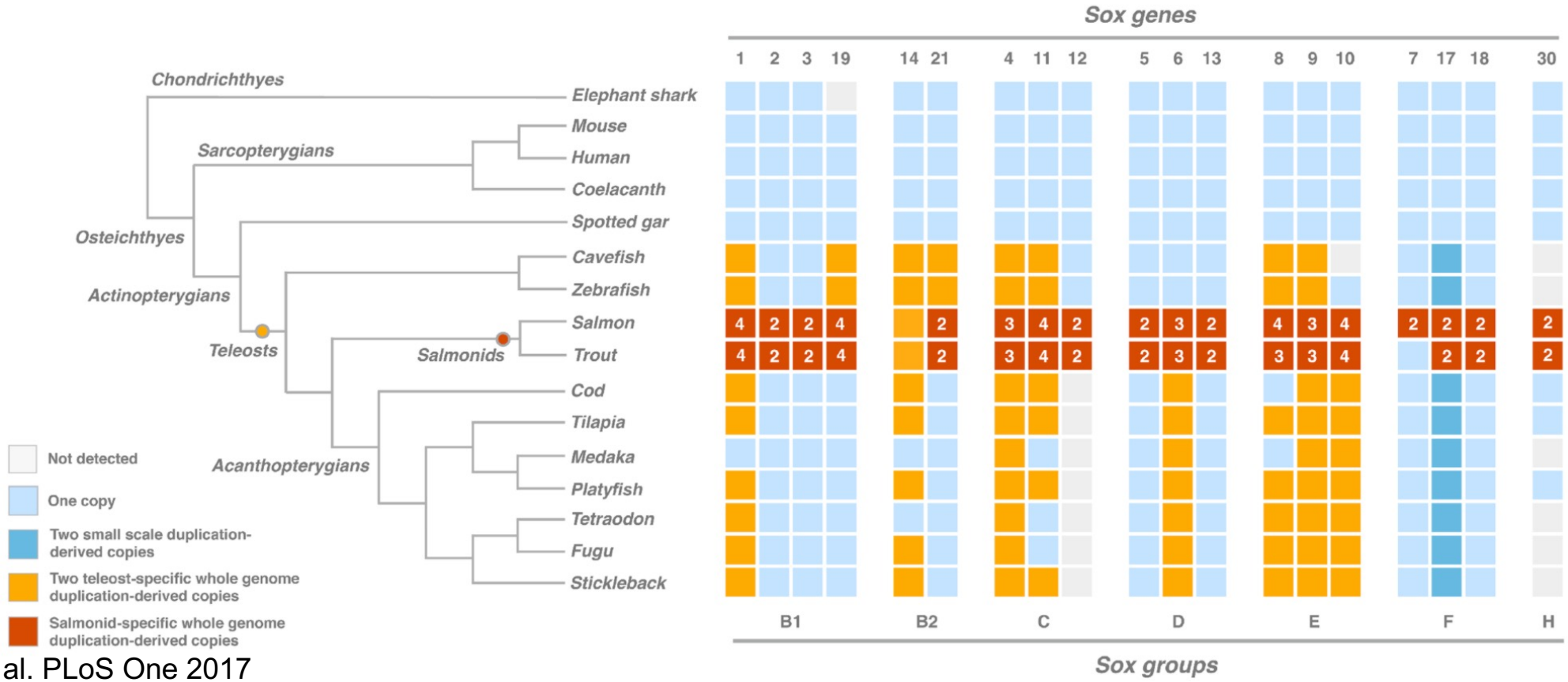
Comparison of syntenic blocks between the *G. raimondii* (*Gr*) and *G. arboreum* (*Ga*) genomes with the genome of the closest ancestor *T. cacao* (*Tc*). Multiple syntenic blocks on different chromosomes is evidence of at least one genome duplication event in the *Gossypium* lineage.

Li et al. Nat Genet 2014 doi.org/10.1038/ng.2987

DIVERSITY MECHANISMS

Landscape of *sox* genes in representative vertebrate genomes.

The *sox* genes (top) are divided into the 7 groups: B1, B2, C, D, E, F and H. Phylogenetic relationships of the different analyzed vertebrate species are indicated on the left. The orange and red circles on the phylogeny represent the teleost-specific WGD and the salmonid-specific WGD, respectively. Light blue squares indicate gene singletons. Orange and dark blue squares indicate duplicates produced either by the teleost-specific WGD or by small-scale duplications (SSDs) respectively. Red squares correspond to genes detected in multiple copies (two, three or four as indicated by the number in the square) in salmonids. White squares are used when no copy was detected. The mammal-specific *SoxA* group is not represented on the figure.



Voldoire, et al. PLoS One 2017

DIVERSITY MECHANISMS

- Genome duplication – Vertebrate homeotic (*Hox*) genes
  - Identified by presence of 180 bp homeobox element
  - Master regulators of transcription
    - Specify body plan
    - Regulate development
  - *Antennapedia* class homeobox genes
    - Specify body segments from anterior to posterior
    - Genes are colinear in location and expression, anterior first
  - Vertebrates have four clusters on separate chromosomes
    - Consistent with two whole-genome duplications in vertebrate evolution

DIVERSITY MECHANISMS

- Genome duplication – Vertebrate homeotic (*Hox*) genes

Wilkins The Evolution of Developmental Pathways 2001

DIVERSITY MECHANISMS

Synopsis

- Mutations – changes to individual nucleotides
  - Substitution, insertion, deletion
- Chromosomal variation – changes to chromosome structure
  - Recombination, repeat expansion, transposition, gene families, inversion, translocation, domain shuffling
- Genomic variation – changes in the chromosomal content of genomes
  - Chromosome merging, polyploidization, genome duplication

DIVERSITY MECHANISMS

Demonstration – Multiple Sequence Alignment

- Data file: IL1A_Mammal120DNA.fasta
  - 120 interleukin-1 alpha DNA sequences from a wide selection of mammals
  - FASTA data format
- Alignment program: MAFFT
  - URL: https://mafft.cbrc.jp/alignment/server/
  - The MAFFT software can also be downloaded and installed locally
- Alignment visualization
  - “VIEW” option on MAFFT results page
  - Aliview software – MSA visualization and editing program

GENETIC EVOLUTION/DIVERSITY MECHANISMS

Thank you