CENTER OF
PLANT SCIENCES

Sant’Anna
Scuola Universitaria Superiore Pisa

Advanced Genomics

The shifting genomic landscape

- Indeed, repetitive DNA explains much of the **C paradox**
- Bursts of TE expansions may explain different sizes also in close taxa
- A balance exists between new insertions and DNA loss (via, *e.g.* unequal recombination)

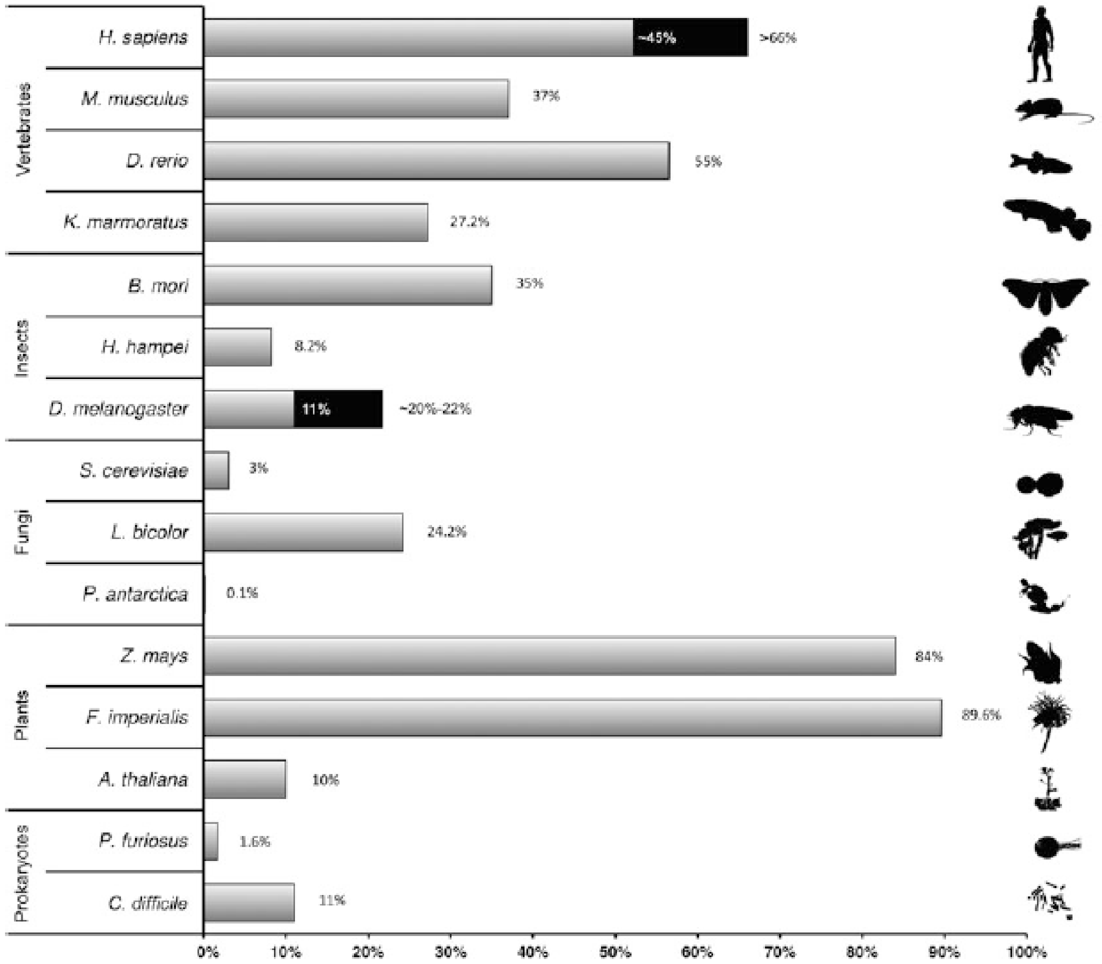
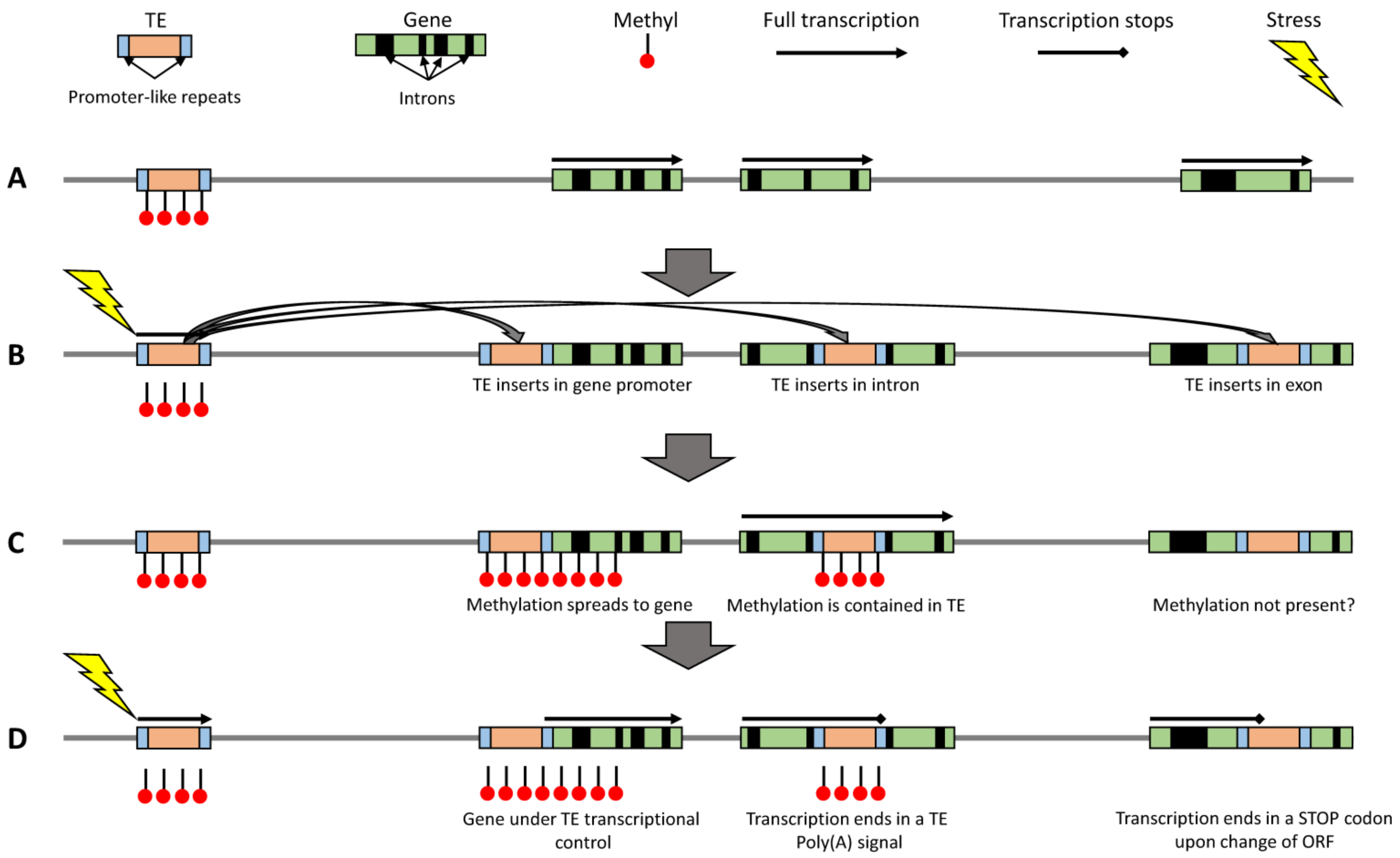


Fig. 1 TE content in the genome of different organisms expressed as percentage of the genome: *Homo sapiens* (~45% [12], >66% [13]) *Mus musculus* [143], *Saccharomyces cerevisiae* [144], *Arabidopsis thaliana* [145], *Pyrococcus furiosus* [146], *Clostridium difficile* [147], *Danio rerio* [133], *Kryptolebias marmoratus* [148], *Bombyx mori* [149], *Hypothenemus hampei* [150], *Drosophila melanogaster* (11%, [68], ~20% [69]), *Pseudozyma antarctica*, and *Laccaria bicolor* [151]. *Zea mays* [152] and *Fritillaria imperialis* [8]. All estimates were obtained with homology-based methods except [13] that uses P-cloud and [69] that uses de novo approaches

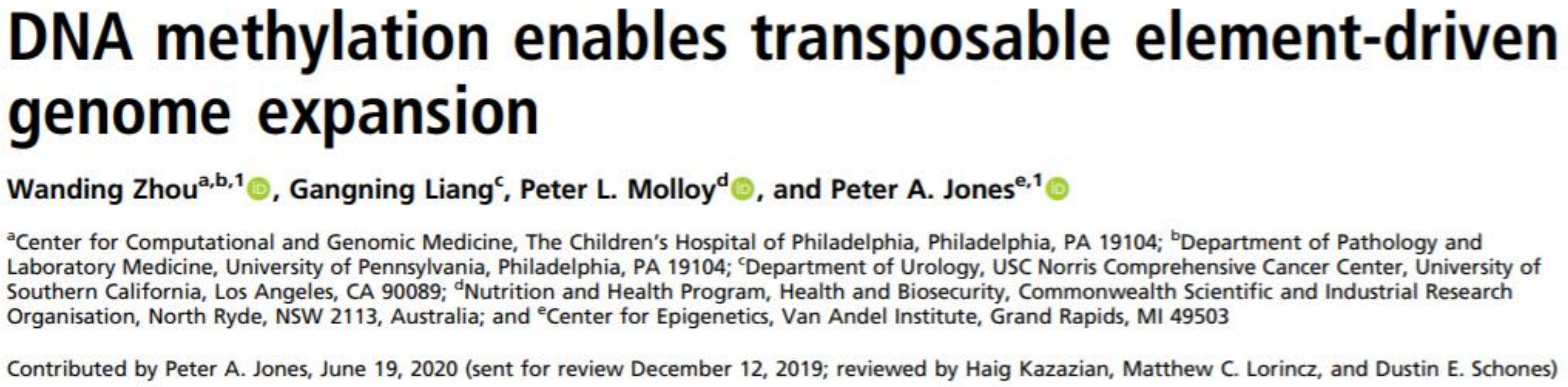
How to keep TEs at bay?

- Methylation (targeted by siRNA) can hinder TE transcription and hence replication and translocation
- TE met can spread to nearby sequences
- Stress can change methylation and unlock TE movement



TE
Promoter-like repeats
Gene
Introns
Methyl
Full transcription
Transcription stops
Stress
A
B
TE inserts in gene promoter
TE inserts in intron
TE inserts in exon
C
Methylation spreads to gene
Methylation is contained in TE
Methylation not present?
D
Gene under TE transcriptional control
Transcription ends in a TE Poly(A) signal
Transcription ends in a STOP codon upon change of ORF

- There’s a delicate equilibrium b/w TE density and genome «evolvability»
- The host genome tries to suppress their movement, but not too much

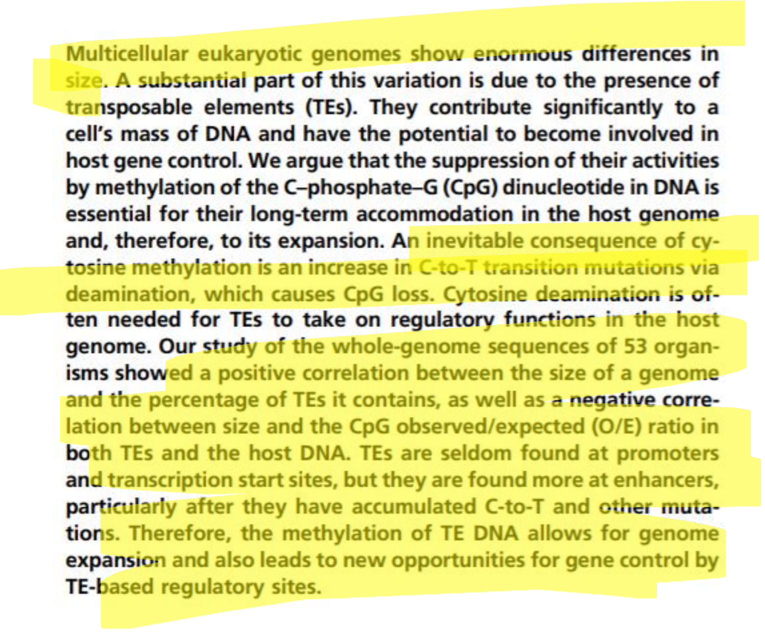


DNA methylation enables transposable element-driven genome expansion

Wanding Zhou[a,b,1], Gangning Liang[c], Peter L. Molloy[d], and Peter A. Jones[e,1]

[a]Center for Computational and Genomic Medicine, The Children’s Hospital of Philadelphia, Philadelphia, PA 19104; [b]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104; [c]Department of Urology, USC Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA 90089; [d]Nutrition and Health Program, Health and Biosecurity, Commonwealth Scientific and Industrial Research Organisation, North Ryde, NSW 2113, Australia; and [e]Center for Epigenetics, Van Andel Institute, Grand Rapids, MI 49503

Contributed by Peter A. Jones, June 19, 2020 (sent for review December 12, 2019; reviewed by Haig Kazazian, Matthew C. Lorincz, and Dustin E. Schones)

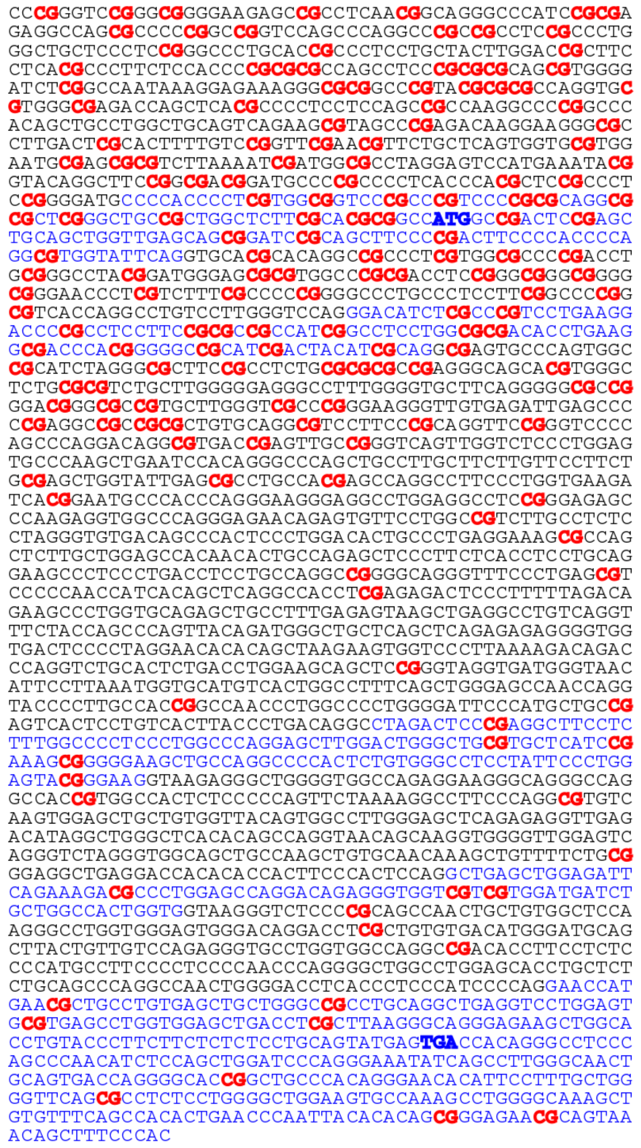


Multicellular eukaryotic genomes show enormous differences in size. A substantial part of this variation is due to the presence of transposable elements (TEs). They contribute significantly to a cell’s mass of DNA and have the potential to become involved in host gene control. We argue that the suppression of their activities by methylation of the C–phosphate–G (CpG) dinucleotide in DNA is essential for their long-term accommodation in the host genome and, therefore, to its expansion. An inevitable consequence of cytosine methylation is an increase in C-to-T transition mutations via deamination, which causes CpG loss. Cytosine deamination is often needed for TEs to take on regulatory functions in the host genome. Our study of the whole-genome sequences of 53 organisms showed a positive correlation between the size of a genome and the percentage of TEs it contains, as well as a negative correlation between size and the CpG observed/expected (O/E) ratio in both TEs and the host DNA. TEs are seldom found at promoters and transcription start sites, but they are found more at enhancers, particularly after they have accumulated C-to-T and other mutations. Therefore, the methylation of TE DNA allows for genome expansion and also leads to new opportunities for gene control by TE-based regulatory sites.

- Eukaryotic genomes display a 64,000-fold variation in their sizes, mostly due to transposable elements
- A long-term outcome of methylation is an increase in C-to-T transition mutations both in the TEs and host DNA; this can be observed as a decreased proportion of CpG dinucleotides over evolutionary time
- Moving TE provide additional DNA for evolvability of the host organism

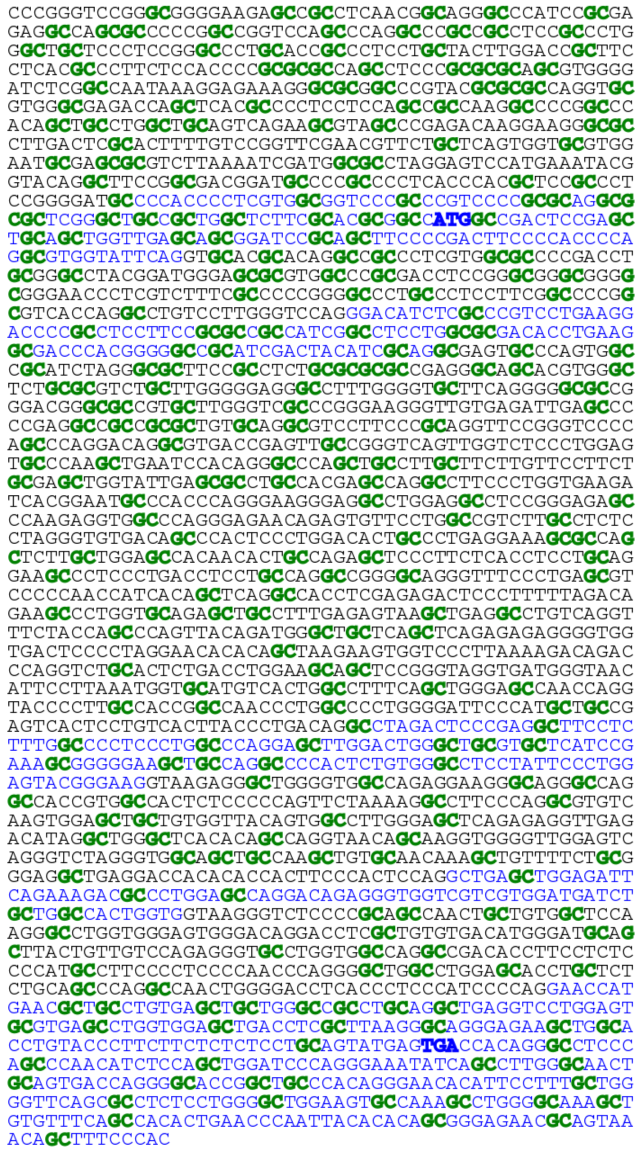
- “CpG islands” are special places in the genome
- CpG dinucleotides are underrepresented due to deamination
- Around 60-70% of human genes have a CpG island in their promoter region
- In evolutionary times, one may look at the unbalance of CpG to understand the age of TE insertion

CpG sites



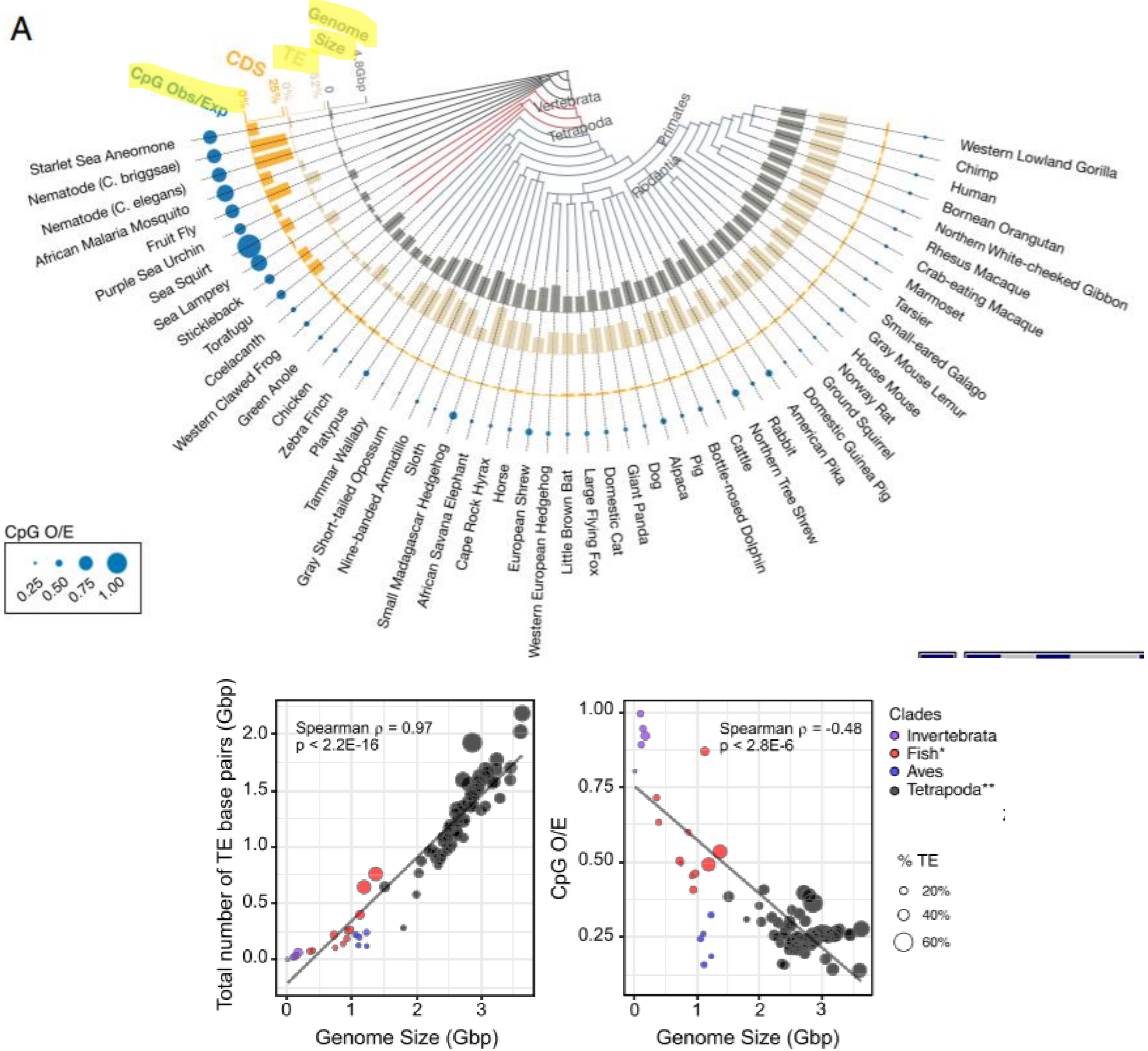
```
CCCGGGTCCGGGCGGGGAAGAGCCGCCTCAACGGCAGGGCCCATCCGCGA
GAGGCCAGCGCCCCCGGCCGGTCCAGCCCAGGCCCGCCGCCTCCGCCCTG
GGCTGCTCCCTCCGGGCCCTGCACCGCCCTCCTGCTACTTGGACCGCTTC
CTCACGCCCTTCTCCACCCCGCGCGCCAGCCTCCCGCGCGCAGCGTGGGG
ATCTCGGCCAATAAAGGAGAAAGGGCGCGGCCCGTACGCGCGCCAGGTGC
GTGGGCGAGACCAGCTCACGCCCCTCCTCCAGCCGCCAAGGCCCCGGCCC
ACAGCTGCCTGGCTGCAGTCAGAAGCGTAGCCCGAGACAAGGAAGGGCGC
CTTGACTCGCACTTTTGTCCGGTTCGAACGTTCTGCTCAGTGGTGCGTGG
AATGCGAGCGCGTCTTAAAATCGATGGCGCCTAGGAGTCCATGAAATACG
GTACAGGCTTCCGGCGACGGATGCCCCGCCCCTCACCCACGCTCCGCCCT
CCGGGGATGCCCCACCCCTCGTGGCGGTCCCGCCCGTCCCCGCGCAGGCG
CGCTCGGGCTGCCGCTGGCTCTTCGCACGCGGCCATGGCCGACTCCGAGC
TGCAGCTGGTTGAGCAGCGGATCCGCAGCTTCCCCGACTTCCCCACCCCA
GGCGTGGTATTCAGGTGCACGCACAGGCCGCCCTCGTGGCGCCCCGACCT
GCGGGCCTACGGATGGGAGCGCGTGGCCCGCGACCTCCGGGCGGGCGGGG
CGGGAACCCTCGTCTTTCGCCCCCGGGGCCCTGCCCTCCTTCGGCCCCGG
CGTCACCAGGCCTGTCCTTGGGTCCAGGGACATCTCGCCCGTCCTGAAGG
ACCCCGCCTCCTTCCGCGCCGCCATCGGCCTCCTGGCGCGACACCTGAAG
GCGACCCACGGGGGCCGCATCGACTACATCGCAGGCGAGTGCCCAGTGGC
CGCATCTAGGGCGCTTCCGCCTCTGCGCGCGCCGAGGGCAGCACGTGGGC
TCTGCGCGTCTGCTTGGGGGAGGGCCTTTGGGGTGCTTCAGGGGGCGCCG
GGACGGGCGCCGTGCTTGGGTCGCCCGGGAAGGGTTGTGAGATTGAGCCC
CCGAGGCCGCCGCGCTGTGCAGGCGTCCTTCCCGCAGGTTCCGGGTCCCC
AGCCCAGGACAGGCGTGACCGAGTTGCCGGGTCAGTTGGTCTCCCTGGAG
TGCCCAAGCTGAATCCACAGGGCCCAGCTGCCTTGCTTCTTGTTCCTTCT
GCGAGCTGGTATTGAGCGCCTGCCACGAGCCAGGCCTTCCCTGGTGAAGA
TCACGGAATGCCCACCCAGGGAAGGGAGGCCTGGAGGCCTCCGGGAGAGC
CCAAGAGGTGGCCCAGGGAGAACAGAGTGTTCCTGGCCGTCTTGCCTCTC
CTAGGGTGTGACAGCCCACTCCCTGGACACTGCCCTGAGGAAAGCGCCAG
CTCTTGCTGGAGCCACAACACTGCCAGAGCTCCCTTCTCACCTCCTGCAG
GAAGCCCTCCCTGACCTCCTGCCAGGCCGGGGCAGGGTTTCCCTGAGCGT
CCCCCAACCATCACAGCTCAGGCCACCTCGAGAGACTCCCTTTTTAGACA
GAAGCCCTGGTGCAGAGCTGCCTTTGAGAGTAAGCTGAGGCCTGTCAGGT
TTCTACCAGCCCAGTTACAGATGGGCTGCTCAGCTCAGAGAGAGGGGTGG
TGACTCCCCTAGGAACACACAGCTAAGAAGTGGTCCCTTAAAAGACAGAC
CCAGGTCTGCACTCTGACCTGGAAGCAGCTCCGGGTAGGTGATGGGTAAC
ATTCCTTAAATGGTGCATGTCACTGGCCTTTCAGCTGGGAGCCAACCAGG
TACCCCTTGCCACCGGCCAACCCTGGCCCCTGGGGATTCCCATGCTGCCG
AGTCACTCCTGTCACTTACCCTGACAGGCCTAGACTCCCGAGGCTTCCTC
TTTGGCCCCTCCCTGGCCCAGGAGCTTGGACTGGGCTGCGTGCTCATCCG
AAAGCGGGGGAAGCTGCCAGGCCCCACTCTGTGGGCCTCCTATTCCCTGG
AGTACGGGAAGGTAAGAGGGCTGGGGTGGCCAGAGGAAGGGCAGGGCCAG
GCCACCGTGGCCACTCTCCCCCAGTTCTAAAAGGCCTTCCCAGGCGTGTC
AAGTGGAGCTGCTGTGGTTACAGTGGCCTTGGGAGCTCAGAGAGGTTGAG
ACATAGGCTGGGCTCACACAGCCAGGTAACAGCAAGGTGGGGTTGGAGTC
AGGGTCTAGGGTGGCAGCTGCCAAGCTGTGCAACAAAGCTGTTTTCTGCG
GGAGGCTGAGGACCACACACCACTTCCCACTCCAGGCTGAGCTGGAGATT
CAGAAAGACGCCCTGGAGCCAGGACAGAGGGTGGTCGTCGTGGATGATCT
GCTGGCCACTGGTGGTAAGGGTCTCCCCGCAGCCAACTGCTGTGGCTCCA
AGGGCCTGGTGGGAGTGGGACAGGACCTCGCTGTGTGACATGGGATGCAG
CTTACTGTTGTCCAGAGGGTGCCTGGTGGCCAGGCCGACACCTTCCTCTC
CCCATGCCTTCCCCTCCCCAACCCAGGGGCTGGCCTGGAGCACCTGCTCT
CTGCAGCCCAGGCCAACTGGGGACCTCACCCTCCCATCCCCAGGAACCAT
GAACGCTGCCTGTGAGCTGCTGGGCCGCCTGCAGGCTGAGGTCCTGGAGT
GCGTGAGCCTGGTGGAGCTGACCTCGCTTAAGGGCAGGGAGAAGCTGGCA
CCTGTACCCTTCTTCTCTCTCCTGCAGTATGAGTGACCACAGGGCCTCCC
AGCCCAACATCTCCAGCTGGATCCCAGGGAAATATCAGCCTTGGGCAACT
GCAGTGACCAGGGGCACCGGCTGCCCACAGGGAACACATTCCTTTGCTGG
GGTTCAGCGCCTCTCCTGGGGCTGGAAGTGCCAAAGCCTGGGGCAAAGCT
GTGTTTCAGCCACACTGAACCCAATTACACACAGCGGGAGAACGCAGTAA
ACAGCTTTCCCAC
```

GpC sites



```
CCCGGGTCCGGGCGGGGAAGAGCCGCCTCAACGGCAGGGCCCATCCGCGA
GAGGCCAGCGCCCCCGGCCGGTCCAGCCCAGGCCCGCCGCCTCCGCCCTG
GGCTGCTCCCTCCGGGCCCTGCACCGCCCTCCTGCTACTTGGACCGCTTC
CTCACGCCCTTCTCCACCCCGCGCGCCAGCCTCCCGCGCGCAGCGTGGGG
ATCTCGGCCAATAAAGGAGAAAGGGCGCGGCCCGTACGCGCGCCAGGTGC
GTGGGCGAGACCAGCTCACGCCCCTCCTCCAGCCGCCAAGGCCCCGGCCC
ACAGCTGCCTGGCTGCAGTCAGAAGCGTAGCCCGAGACAAGGAAGGGCGC
CTTGACTCGCACTTTTGTCCGGTTCGAACGTTCTGCTCAGTGGTGCGTGG
AATGCGAGCGCGTCTTAAAATCGATGGCGCCTAGGAGTCCATGAAATACG
GTACAGGCTTCCGGCGACGGATGCCCCGCCCCTCACCCACGCTCCGCCCT
CCGGGGATGCCCCACCCCTCGTGGCGGTCCCGCCCGTCCCCGCGCAGGCG
CGCTCGGGCTGCCGCTGGCTCTTCGCACGCGGCCATGGCCGACTCCGAGC
TGCAGCTGGTTGAGCAGCGGATCCGCAGCTTCCCCGACTTCCCCACCCCA
GGCGTGGTATTCAGGTGCACGCACAGGCCGCCCTCGTGGCGCCCCGACCT
GCGGGCCTACGGATGGGAGCGCGTGGCCCGCGACCTCCGGGCGGGCGGGG
CGGGAACCCTCGTCTTTCGCCCCCGGGGCCCTGCCCTCCTTCGGCCCCGG
CGTCACCAGGCCTGTCCTTGGGTCCAGGGACATCTCGCCCGTCCTGAAGG
ACCCCGCCTCCTTCCGCGCCGCCATCGGCCTCCTGGCGCGACACCTGAAG
GCGACCCACGGGGGCCGCATCGACTACATCGCAGGCGAGTGCCCAGTGGC
CGCATCTAGGGCGCTTCCGCCTCTGCGCGCGCCGAGGGCAGCACGTGGGC
TCTGCGCGTCTGCTTGGGGGAGGGCCTTTGGGGTGCTTCAGGGGGCGCCG
GGACGGGCGCCGTGCTTGGGTCGCCCGGGAAGGGTTGTGAGATTGAGCCC
CCGAGGCCGCCGCGCTGTGCAGGCGTCCTTCCCGCAGGTTCCGGGTCCCC
AGCCCAGGACAGGCGTGACCGAGTTGCCGGGTCAGTTGGTCTCCCTGGAG
TGCCCAAGCTGAATCCACAGGGCCCAGCTGCCTTGCTTCTTGTTCCTTCT
GCGAGCTGGTATTGAGCGCCTGCCACGAGCCAGGCCTTCCCTGGTGAAGA
TCACGGAATGCCCACCCAGGGAAGGGAGGCCTGGAGGCCTCCGGGAGAGC
CCAAGAGGTGGCCCAGGGAGAACAGAGTGTTCCTGGCCGTCTTGCCTCTC
CTAGGGTGTGACAGCCCACTCCCTGGACACTGCCCTGAGGAAAGCGCCAG
CTCTTGCTGGAGCCACAACACTGCCAGAGCTCCCTTCTCACCTCCTGCAG
GAAGCCCTCCCTGACCTCCTGCCAGGCCGGGGCAGGGTTTCCCTGAGCGT
CCCCCAACCATCACAGCTCAGGCCACCTCGAGAGACTCCCTTTTTAGACA
GAAGCCCTGGTGCAGAGCTGCCTTTGAGAGTAAGCTGAGGCCTGTCAGGT
TTCTACCAGCCCAGTTACAGATGGGCTGCTCAGCTCAGAGAGAGGGGTGG
TGACTCCCCTAGGAACACACAGCTAAGAAGTGGTCCCTTAAAAGACAGAC
CCAGGTCTGCACTCTGACCTGGAAGCAGCTCCGGGTAGGTGATGGGTAAC
ATTCCTTAAATGGTGCATGTCACTGGCCTTTCAGCTGGGAGCCAACCAGG
TACCCCTTGCCACCGGCCAACCCTGGCCCCTGGGGATTCCCATGCTGCCG
AGTCACTCCTGTCACTTACCCTGACAGGCCTAGACTCCCGAGGCTTCCTC
TTTGGCCCCTCCCTGGCCCAGGAGCTTGGACTGGGCTGCGTGCTCATCCG
AAAGCGGGGGAAGCTGCCAGGCCCCACTCTGTGGGCCTCCTATTCCCTGG
AGTACGGGAAGGTAAGAGGGCTGGGGTGGCCAGAGGAAGGGCAGGGCCAG
GCCACCGTGGCCACTCTCCCCCAGTTCTAAAAGGCCTTCCCAGGCGTGTC
AAGTGGAGCTGCTGTGGTTACAGTGGCCTTGGGAGCTCAGAGAGGTTGAG
ACATAGGCTGGGCTCACACAGCCAGGTAACAGCAAGGTGGGGTTGGAGTC
AGGGTCTAGGGTGGCAGCTGCCAAGCTGTGCAACAAAGCTGTTTTCTGCG
GGAGGCTGAGGACCACACACCACTTCCCACTCCAGGCTGAGCTGGAGATT
CAGAAAGACGCCCTGGAGCCAGGACAGAGGGTGGTCGTCGTGGATGATCT
GCTGGCCACTGGTGGTAAGGGTCTCCCCGCAGCCAACTGCTGTGGCTCCA
AGGGCCTGGTGGGAGTGGGACAGGACCTCGCTGTGTGACATGGGATGCAG
CTTACTGTTGTCCAGAGGGTGCCTGGTGGCCAGGCCGACACCTTCCTCTC
CCCATGCCTTCCCCTCCCCAACCCAGGGGCTGGCCTGGAGCACCTGCTCT
CTGCAGCCCAGGCCAACTGGGGACCTCACCCTCCCATCCCCAGGAACCAT
GAACGCTGCCTGTGAGCTGCTGGGCCGCCTGCAGGCTGAGGTCCTGGAGT
GCGTGAGCCTGGTGGAGCTGACCTCGCTTAAGGGCAGGGAGAAGCTGGCA
CCTGTACCCTTCTTCTCTCTCCTGCAGTATGAGTGACCACAGGGCCTCCC
AGCCCAACATCTCCAGCTGGATCCCAGGGAAATATCAGCCTTGGGCAACT
GCAGTGACCAGGGGCACCGGCTGCCCACAGGGAACACATTCCTTTGCTGG
GGTTCAGCGCCTCTCCTGGGGCTGGAAGTGCCAAAGCCTGGGGCAAAGCT
GTGTTTCAGCCACACTGAACCCAATTACACACAGCGGGAGAACGCAGTAA
ACAGCTTTCCCAC
```

APRT gene sequence



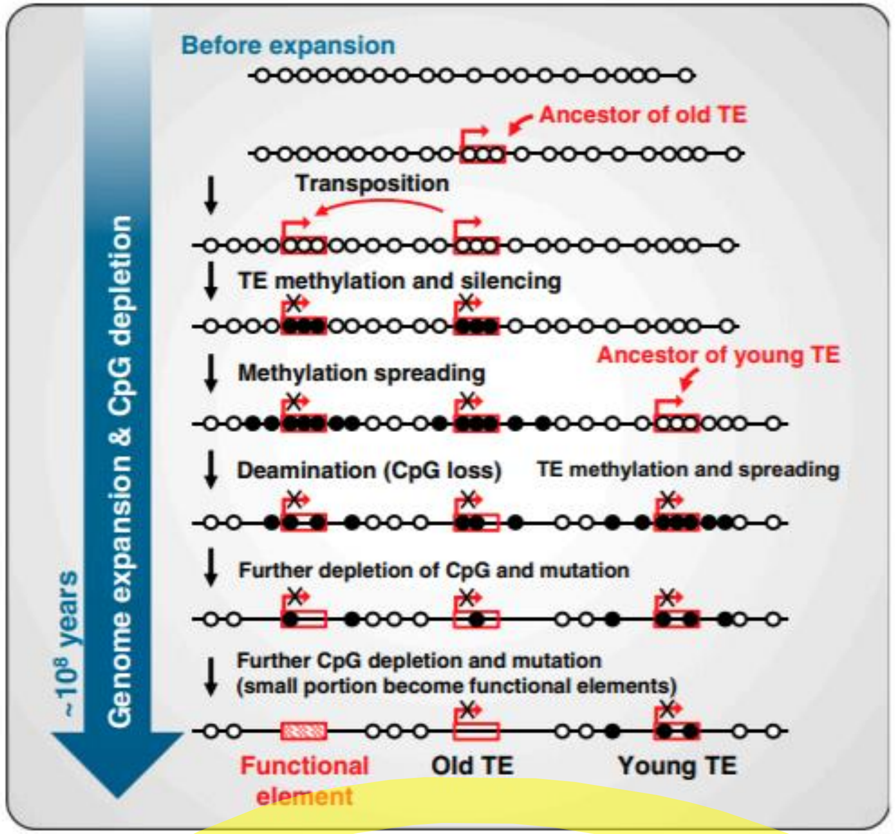
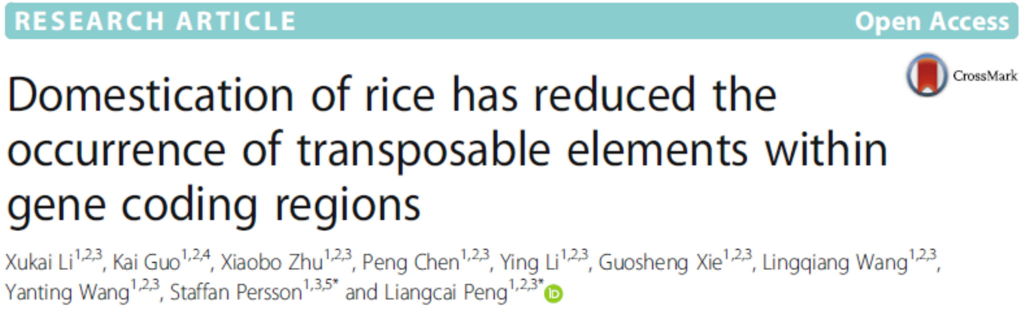

Before expansion
Ancestor of old TE
Transposition
TE methylation and silencing
Methylation spreading
Ancestor of young TE
Deamination (CpG loss)
TE methylation and spreading
Further depletion of CpG and mutation
Further CpG depletion and mutation
(small portion become functional elements)
Functional element
Old TE
Young TE
~10⁸ years
Genome expansion & CpG depletion

Fig. 6. CpG methylation contributes to TE-mediated genome expansion and ultimately to CpG depletion by deamination and neofunctionalization of TEs in the expanded genome. The model depicts an early genome with no TEs and the unmethylated CpG sites shown as open circles and methylated CpGs as solid black circles. At this stage, the CpG O/E ratio is about 1. Insertion and transposition of a TE lead to its de novo methylation (shown as black circles) and silencing of the TE. Methylation can then spread into the flanking host DNA. Methylated CpGs have an enhanced mutation frequency relative to unmethylated CpGs and a half-life of about 35 million y in the primate germline (10). Over evolutionary time, this leads to an overall depletion of CpGs in the entire genome with the exception of CpG islands (11) and ultimately to the creation of new functional elements such as enhancers, depicted by the decreasing number of methylation sites and a decrease in CpG O/E ratio.

Why more TEs near centromeres?

- With time, accumulated TEs can get lost by (illegitimate) recombination and deletion
- In regions with low recombination, TEs are lost with less efficiency
- Selection may also purge TEs close to genes; their methylation will expand to nearby loci and inactivate genes reducing individual fitness

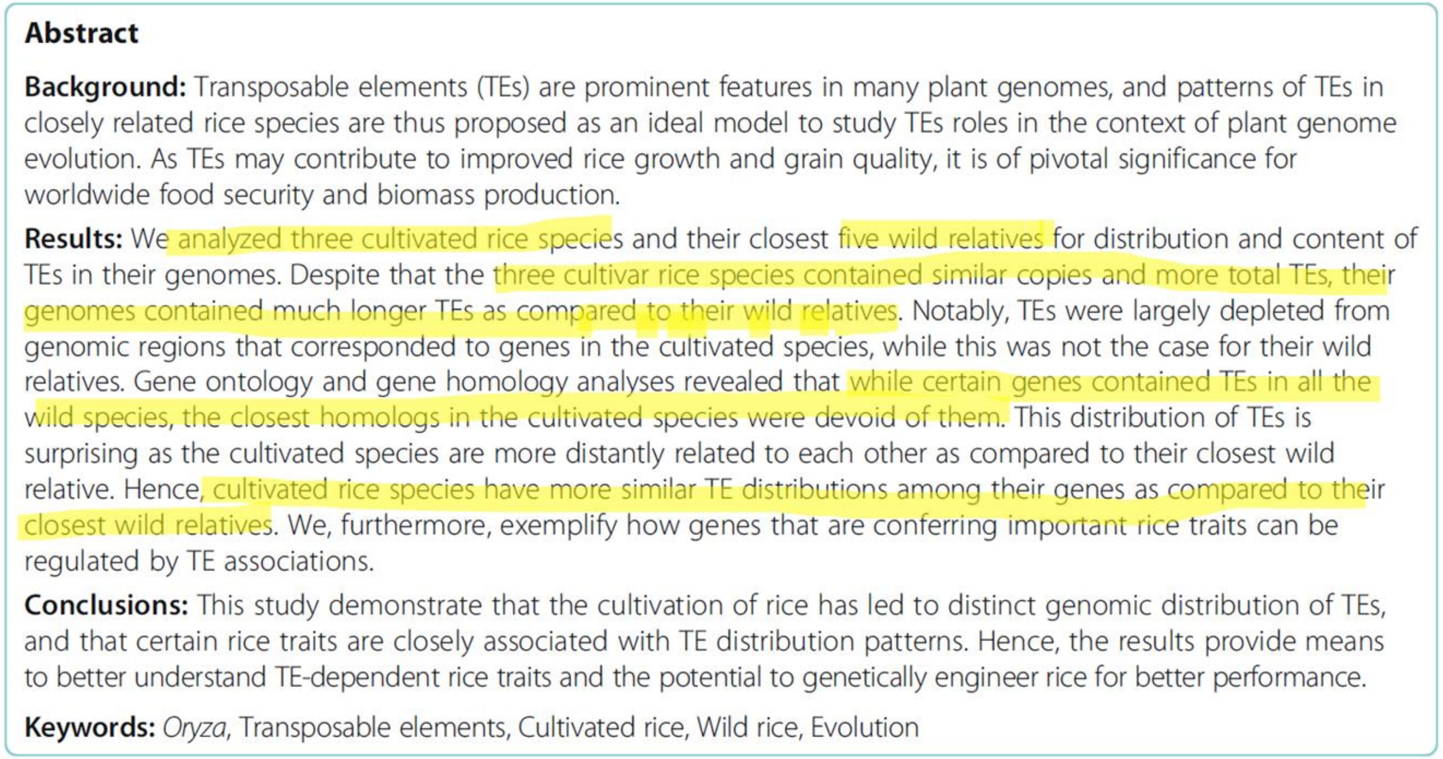
TEs interfering with genes are a major force of evolution



RESEARCH ARTICLE Open Access

Domestication of rice has reduced the occurrence of transposable elements within gene coding regions

Xukai Li[1,2,3], Kai Guo[1,2,4], Xiaobo Zhu[1,2,3], Peng Chen[1,2,3], Ying Li[1,2,3], Guosheng Xie[1,2,3], Lingqiang Wang[1,2,3], Yanting Wang[1,2,3], Staffan Persson[1,3,5*] and Liangcai Peng[1,2,3*]



Abstract

Background: Transposable elements (TEs) are prominent features in many plant genomes, and patterns of TEs in closely related rice species are thus proposed as an ideal model to study TEs roles in the context of plant genome evolution. As TEs may contribute to improved rice growth and grain quality, it is of pivotal significance for worldwide food security and biomass production.

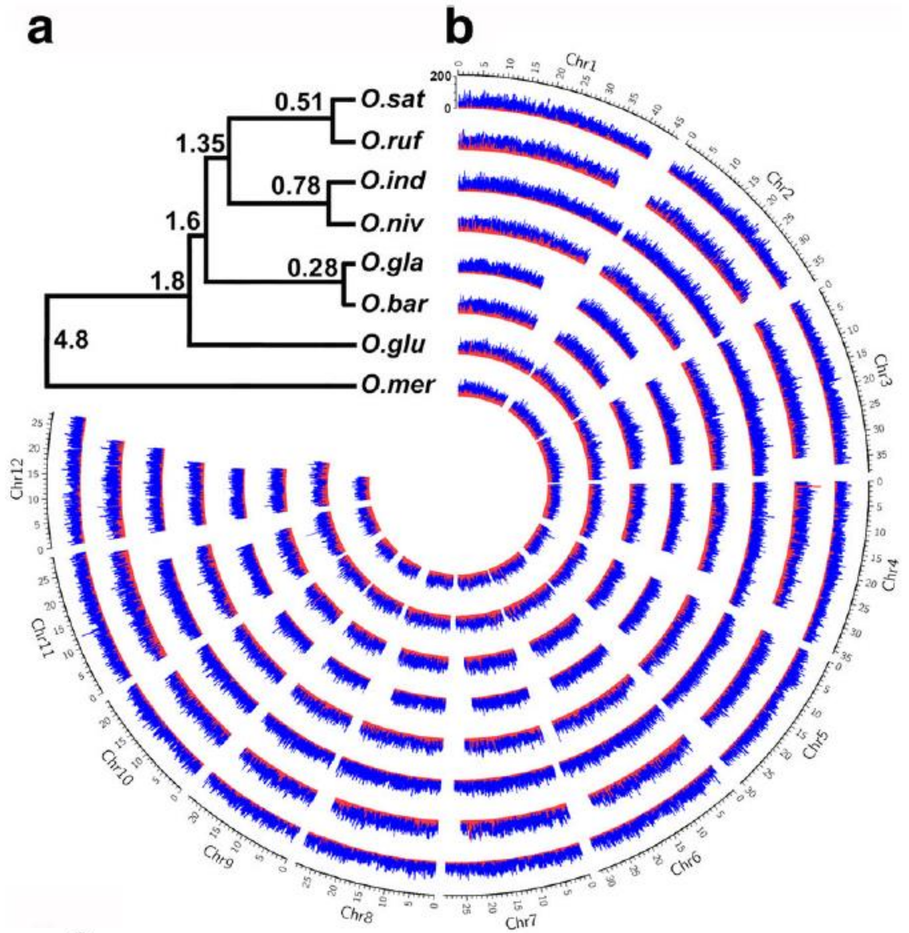
Results: We analyzed three cultivated rice species and their closest five wild relatives for distribution and content of TEs in their genomes. Despite that the three cultivar rice species contained similar copies and more total TEs, their genomes contained much longer TEs as compared to their wild relatives. Notably, TEs were largely depleted from genomic regions that corresponded to genes in the cultivated species, while this was not the case for their wild relatives. Gene ontology and gene homology analyses revealed that while certain genes contained TEs in all the wild species, the closest homologs in the cultivated species were devoid of them. This distribution of TEs is surprising as the cultivated species are more distantly related to each other as compared to their closest wild relative. Hence, cultivated rice species have more similar TE distributions among their genes as compared to their closest wild relatives. We, furthermore, exemplify how genes that are conferring important rice traits can be regulated by TE associations.

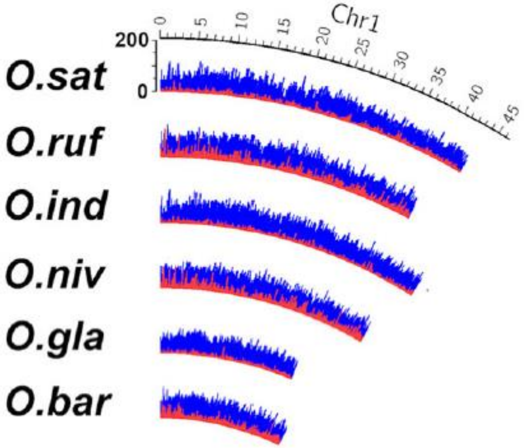
Conclusions: This study demonstrate that the cultivation of rice has led to distinct genomic distribution of TEs, and that certain rice traits are closely associated with TE distribution patterns. Hence, the results provide means to better understand TE-dependent rice traits and the potential to genetically engineer rice for better performance.

Keywords: *Oryza*, Transposable elements, Cultivated rice, Wild rice, Evolution

In the rice species complex, different cultivated rice have different origins (Asia, Africa)

- Authors compare sequence information across cultivated and wild species
- Though distantly related, cultivated rice show convergence in TE distribution (less in genic regions)





- Gene regions in the cultivated species are devoid of TEs, and genes of certain functions tend to be similarly affected by TEs in cultivated species, but different in wild rice species
- E.g. *GIF1* (grain filling) show conserved gene collinearity and structure across cultivated and wild, but TE content and positions clearly differ
- These TEs might affect either alternative splicing or changes in expression of the *GIF1* locus, which may contribute to the grain-filling

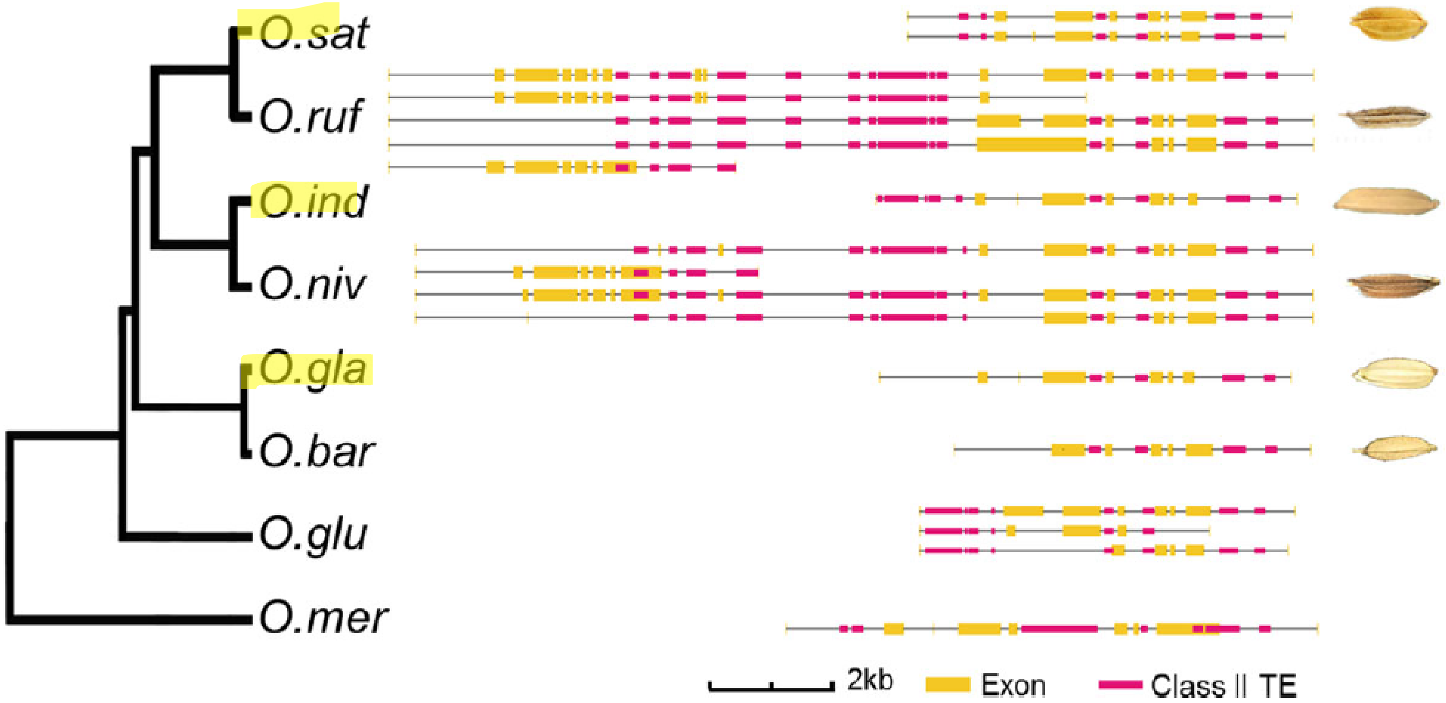
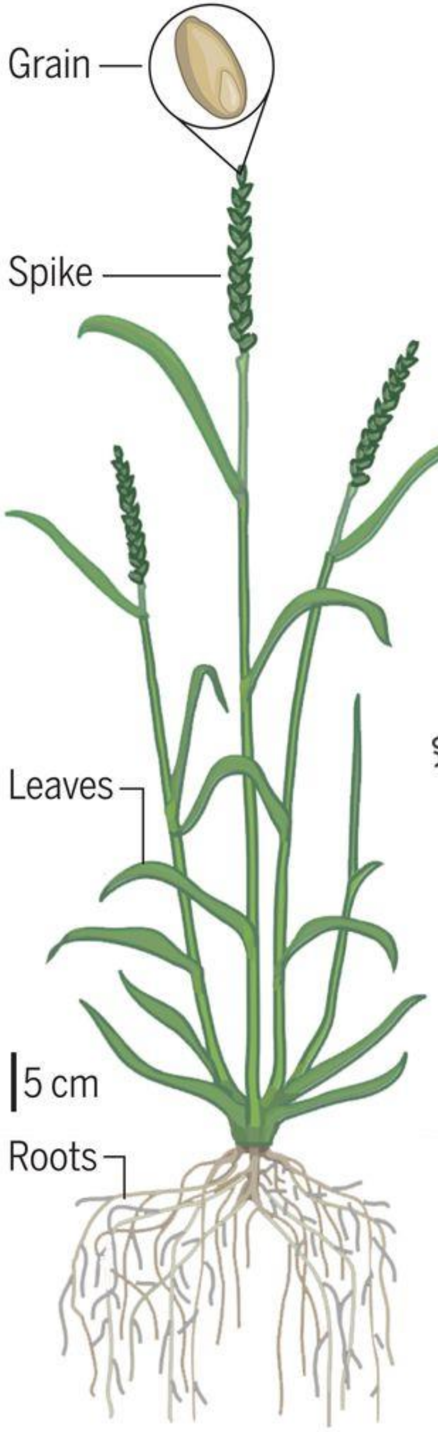
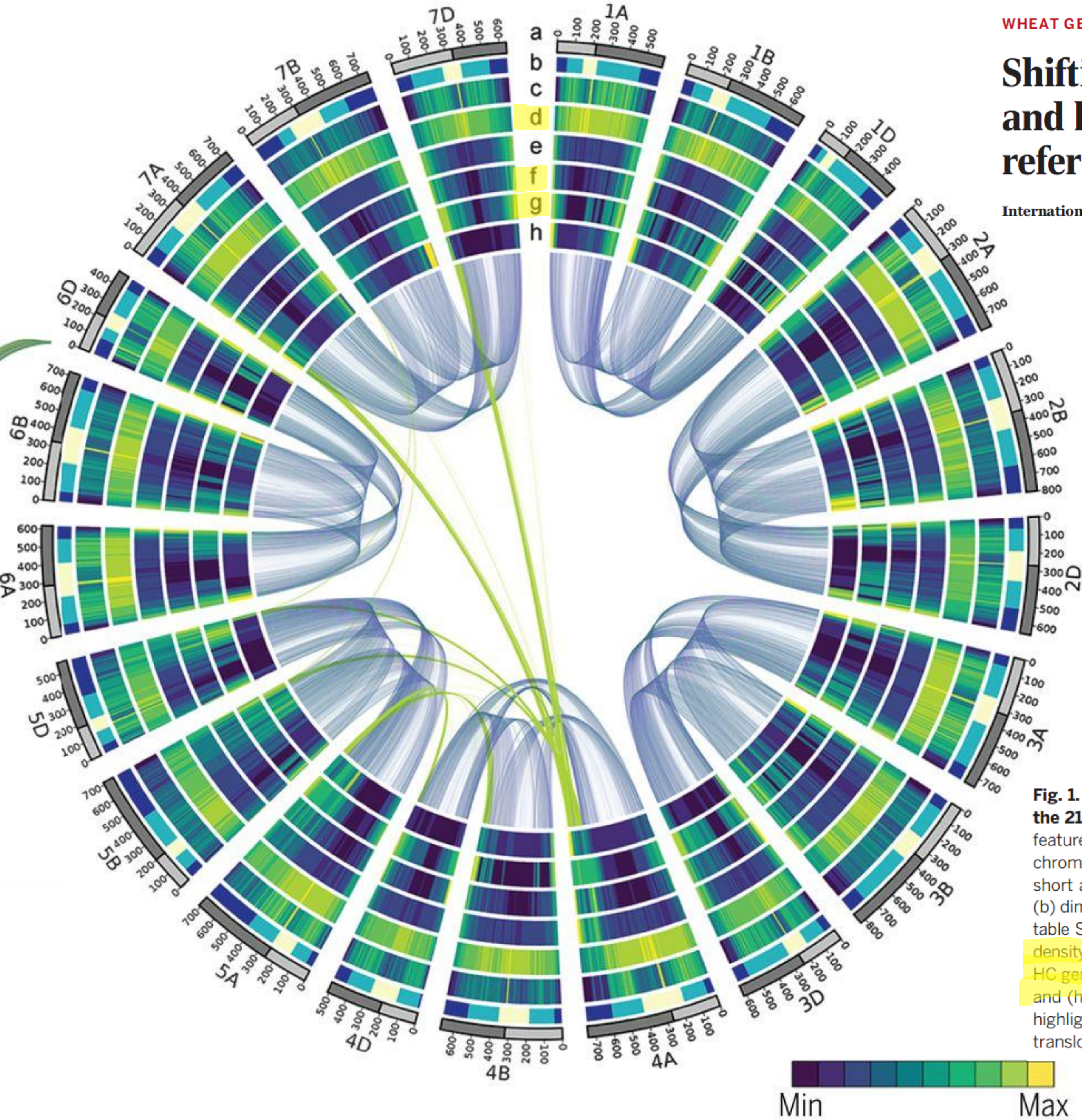


Fig. 6 Comparisons of gene structure and TE locations of *GIF1* gene critical for grain filling in the eight rice species. Organization of exons, introns and TEs of *GIF1* (*GRAIN INCOMPLETE FILLING 1*; LOC_Os04g33740) gene in gene body and 2-kbp flanking sequences of the gene. Seed images of ancestral wild rice and cultivated rice are shown to the *right*





WHEAT GENOME

Shifting the limits in wheat research and breeding using a fully annotated reference genome

International Wheat Genome Sequencing Consortium (IWGSC)*

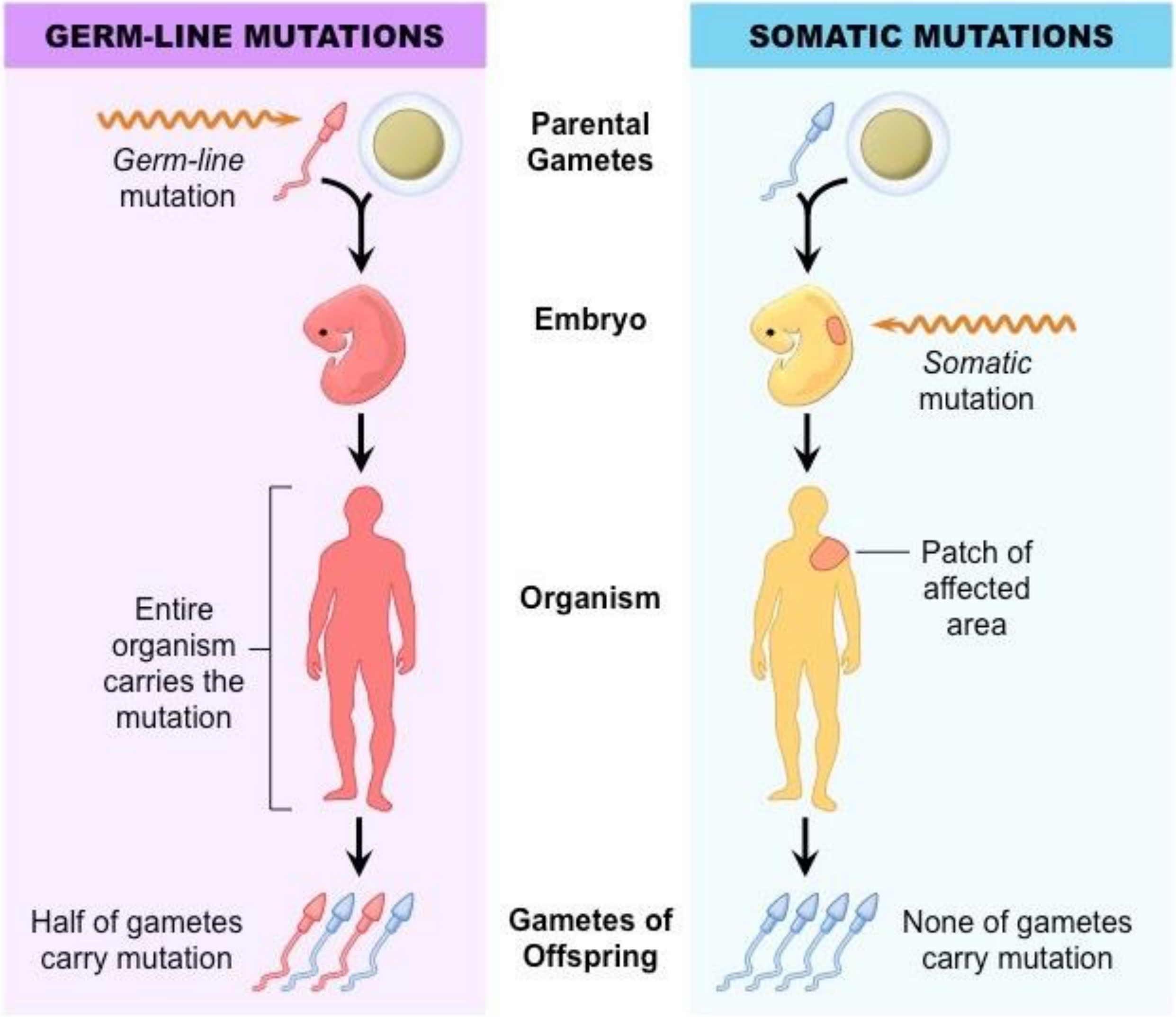
Fig. 1. Structural, functional, and conserved synteny landscape of the 21 wheat chromosomes. (**A**) Circular diagram showing genomic features of wheat. The tracks toward the center of the circle display (a) chromosome name and size (100-Mb tick size; light gray bar indicates the short arm and dark gray indicates the long arm of the chromosome); (b) dimension of chromosomal segments R1, R2a, C, R2b, and R3 [(*18*) and table S29]; (c) *K*-mer 20-frequencies distribution; (d) LTR-retrotransposons density; (e) pseudogenes density (0 to 130 genes per Mb); (f) density of HC gene models (0 to 32 genes per Mb); (g) density of recombination rate; and (h) SNP density. Connecting lines in the center of the diagram highlight homeologous relationships of chromosomes (blue lines) and translocated regions (green lines). (**B**) Distribution of Pfam domain

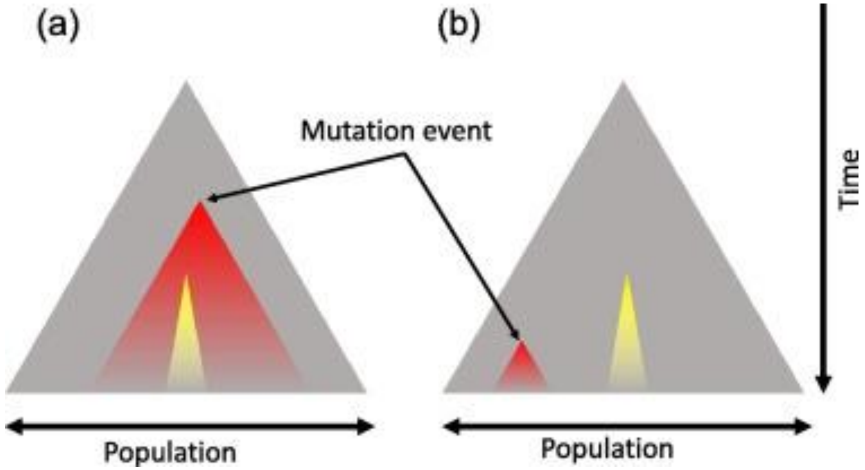
A changing genomic landscape

- Mutations: the raw material on which evolution works
- Three broad classes:
  - Point mutations (and small indels)
  - Copy number variants (CNVs)
  - Chromosomal mutations (inversions, translocations, ploidy)

- The rate of nucleotide substitutions is estimated to be 1 in 10^8 per generation, implying that 30 nucleotide mutations would be expected in each human gamete
- Locus-specific mutation rates for CNV have been observed to be ~100 to 10,000 times higher than those for nucleotide substitution rates (and very diverse)

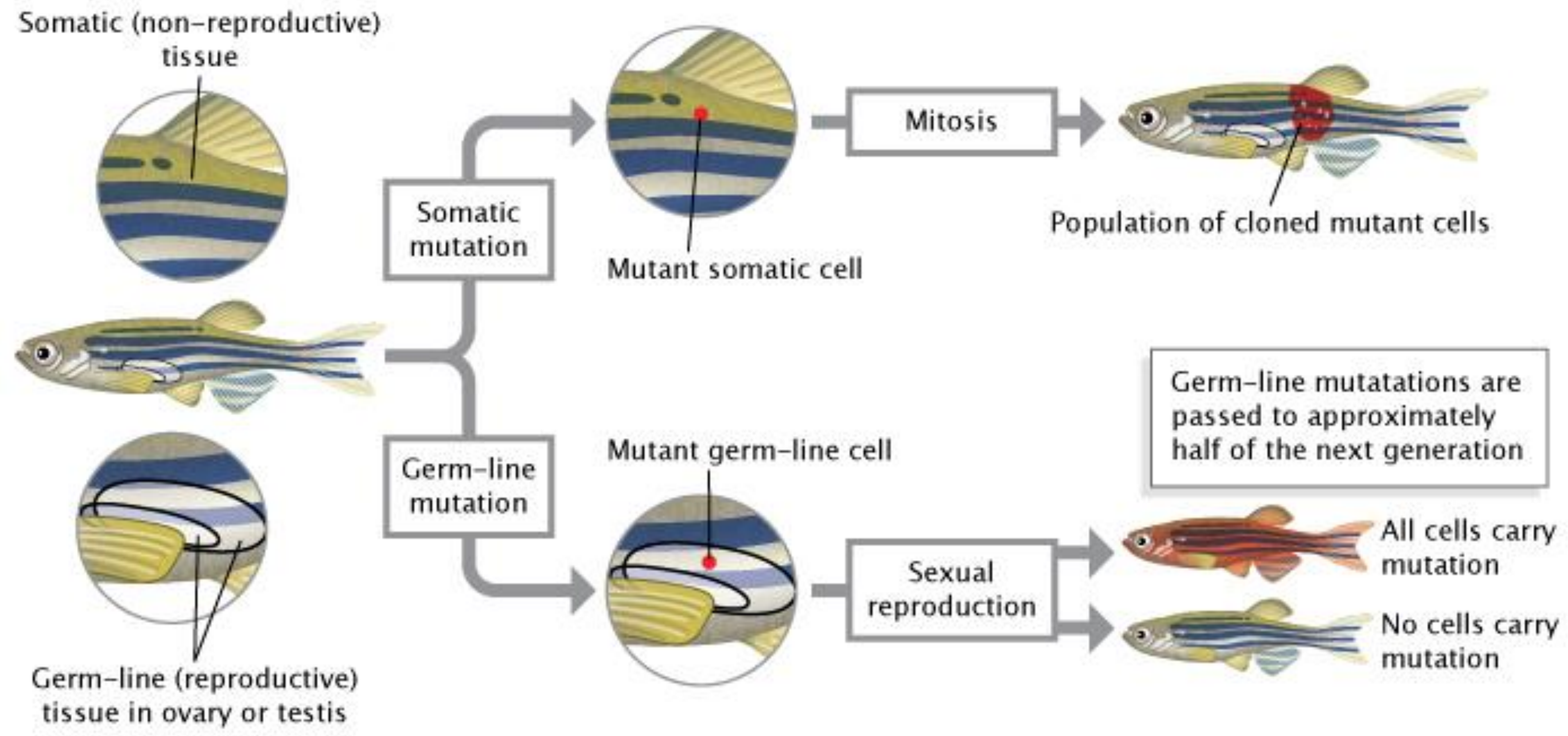
Timing and location of a mutation makes a difference (in evolutionary terms)





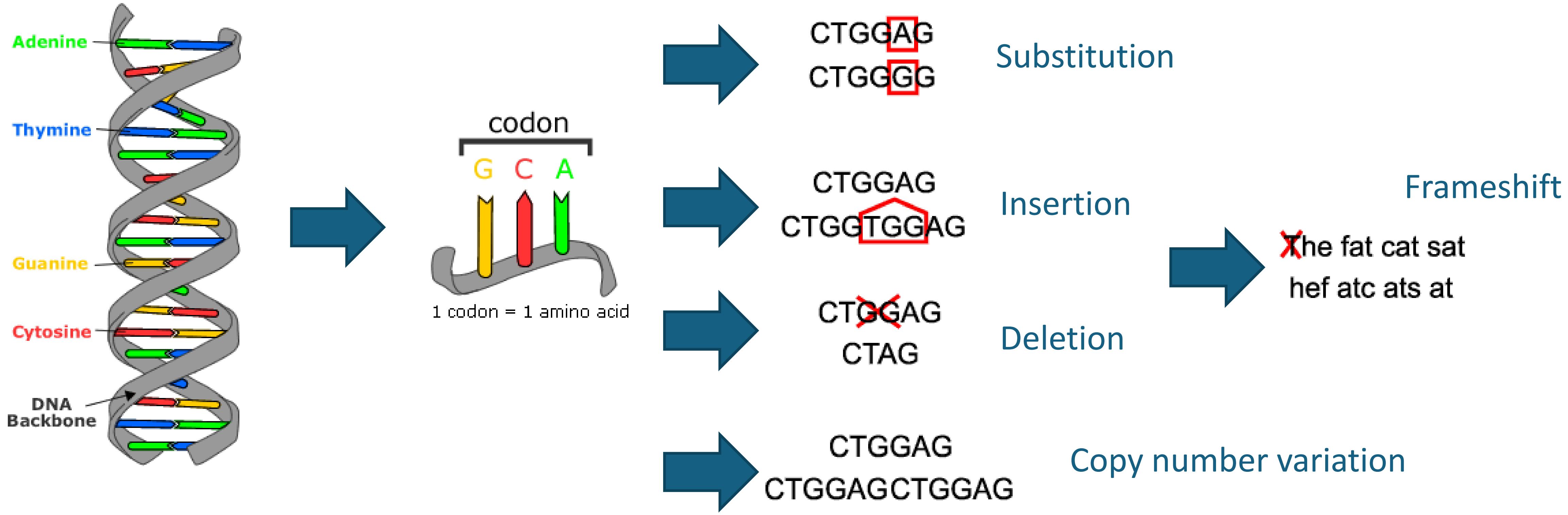
An earlier mutation (a) produces a larger population of mutant cells than a later mutation (b). Grey triangle: whole cell lineages of an individual; red triangle: a mutant cell lineage; yellow triangle: a germline cell lineage. Depending on the timing of a mutation event, the affected cell population size varies, including the germline cells.

- All mutations having a phenotypic outcome will be subjected to natural selection
- Only those mutations taking place in the germ line will be propagated in the population


Somatic (non-reproductive) tissue
Somatic mutation
Mutant somatic cell
Mitosis
Population of cloned mutant cells
Germ-line mutation
Mutant germ-line cell
Sexual reproduction
Germ-line mutatations are passed to approximately half of the next generation
All cells carry mutation
No cells carry mutation
Germ-line (reproductive) tissue in ovary or testis

Point mutations and small insertions / deletions

- Changes affecting only one or few nucleotides
- Impact the DNA sequence and reading frame



CTGGAG
CTGGGG
Substitution

CTGGAG
CTGGTGGAG
Insertion

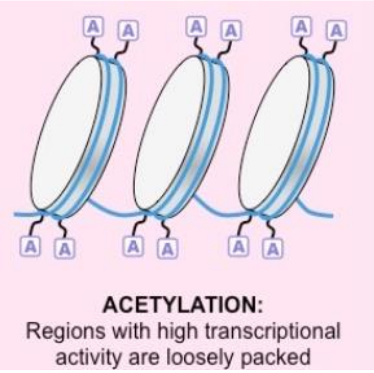
CT~~GG~~AG
CTAG
Deletion

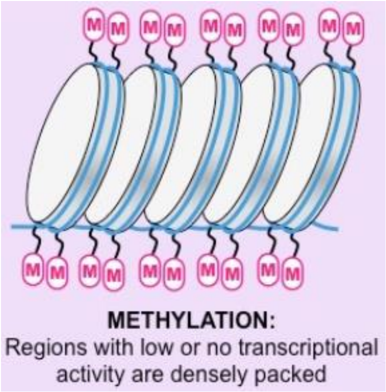
Frameshift
~~T~~he fat cat sat
hef atc ats at

CTGGAG
CTGGAGCTGGAG
Copy number variation

Not all point mutations are equal

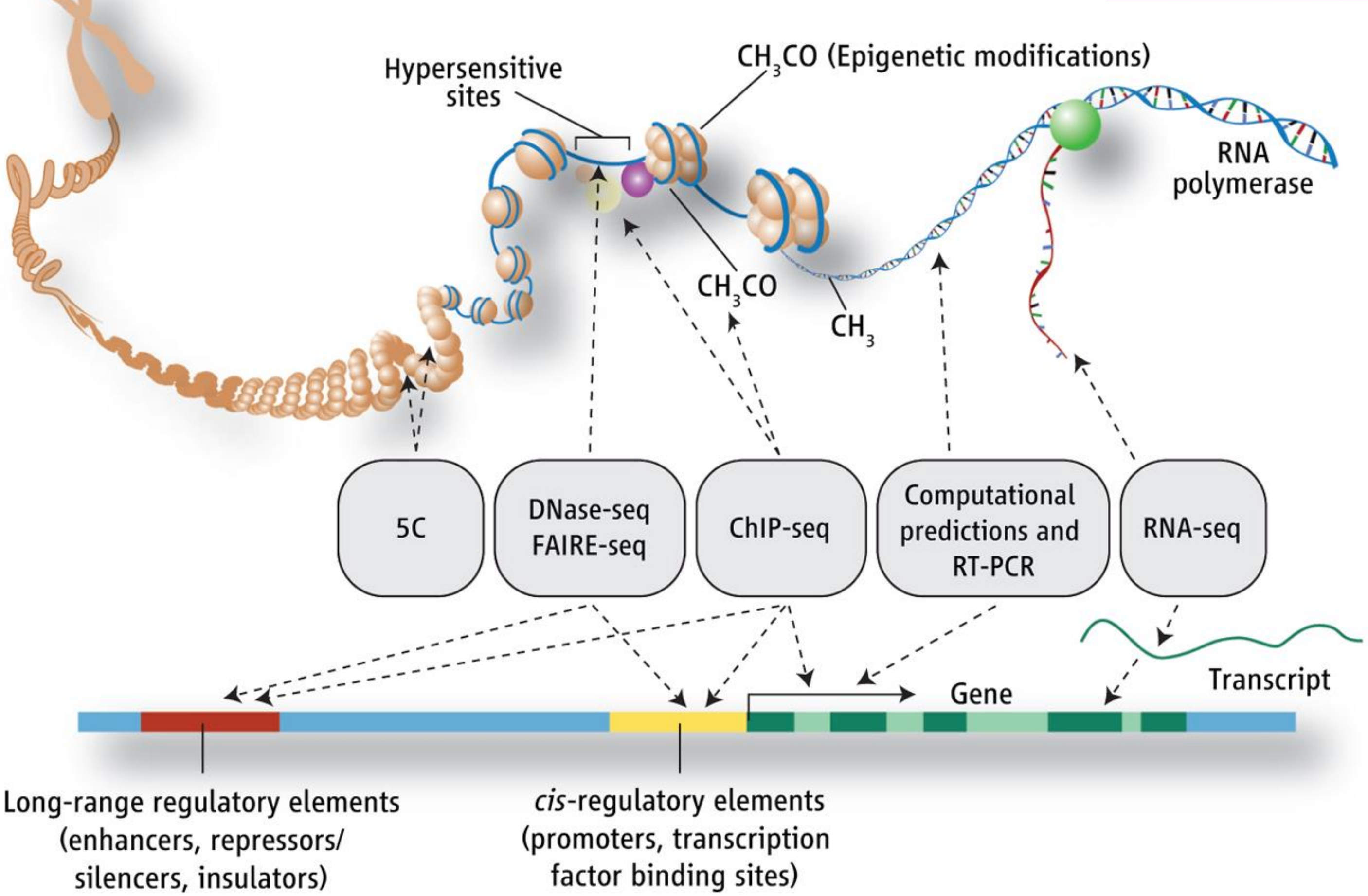
- Euchromatine or heterochromatine
- Genic or intergenic
- Introns or exons
- …


A
ACETYLATION:
Regions with high transcriptional activity are loosely packed


M
METHYLATION:
Regions with low or no transcriptional activity are densely packed

Humans

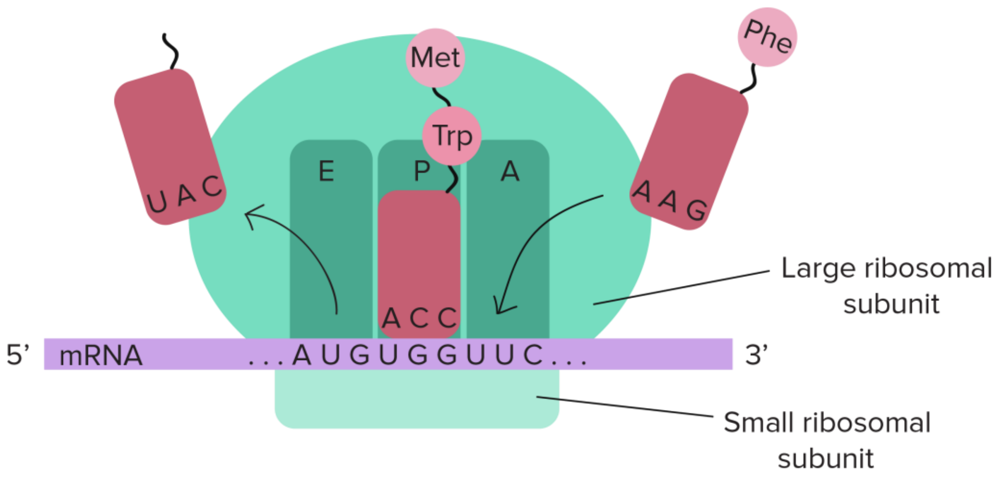
- 3.2 Bbp, only about 21,000 genes (1-3% of the genome)
- Yet 80% of the human genome serves some purpose, biochemically speaking

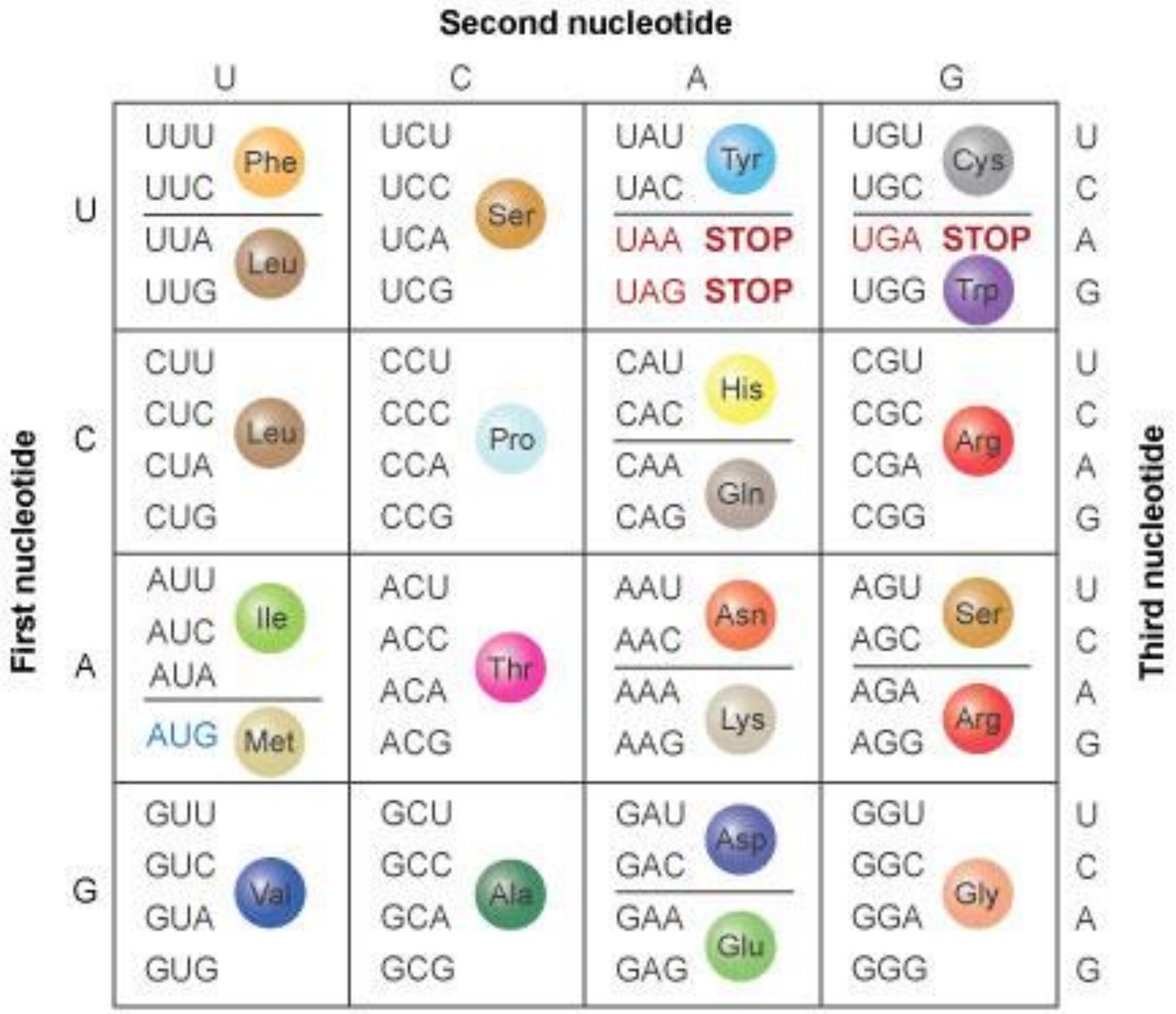

Hypersensitive sites
CH$_3$CO (Epigenetic modifications)
RNA polymerase
CH$_3$CO
CH$_3$
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Transcript
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)

Not all point mutations are equal - 2

- Nonsense, missense, neutral (and frameshift)

Depending on the nucleotide base that is affected by the mutation, the outcome may be different in term of aminoacids incorporated in proteins



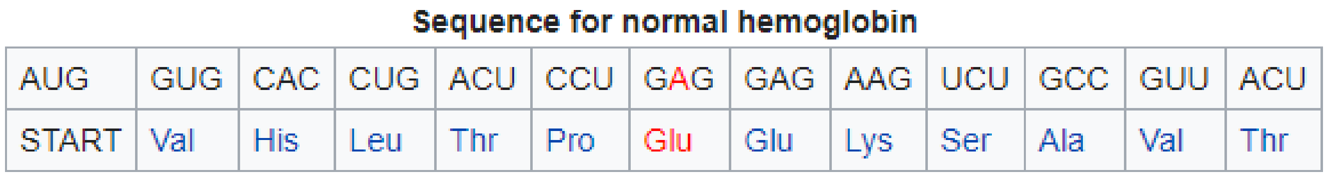


Second nucleotide

First nucleotide	U	C	A	G	Third nucleotide
U	UUU UUC Phe UUA UUG Leu	UCU UCC UCA UCG Ser	UAU UAC Tyr UAA STOP UAG STOP	UGU UGC Cys UGA STOP UGG Trp	U C A G
C	CUU CUC CUA CUG Leu	CCU CCC CCA CCG Pro	CAU CAC His CAA CAG Gln	CGU CGC CGA CGG Arg	U C A G
A	AUU AUC AUA Ile AUG Met	ACU ACC ACA ACG Thr	AAU AAC Asn AAA AAG Lys	AGU AGC Ser AGA AGG Arg	U C A G
G	GUU GUC GUA GUG Val	GCU GCC GCA GCG Ala	GAU GAC Asp GAA GAG Glu	GGU GGC GGA GGG Gly	U C A G

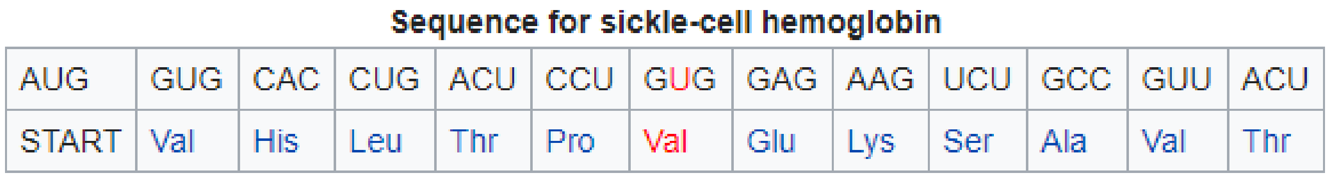
Mutation in the sickle cell anemia

- Autosomal recessive disease
- Just one mutation on the hemoglobin beta gene (chromosome 11), 147 amino acids long
- Glutamic acid to Valine
- Red blood cells deformed, cannot effectively carry around oxigen



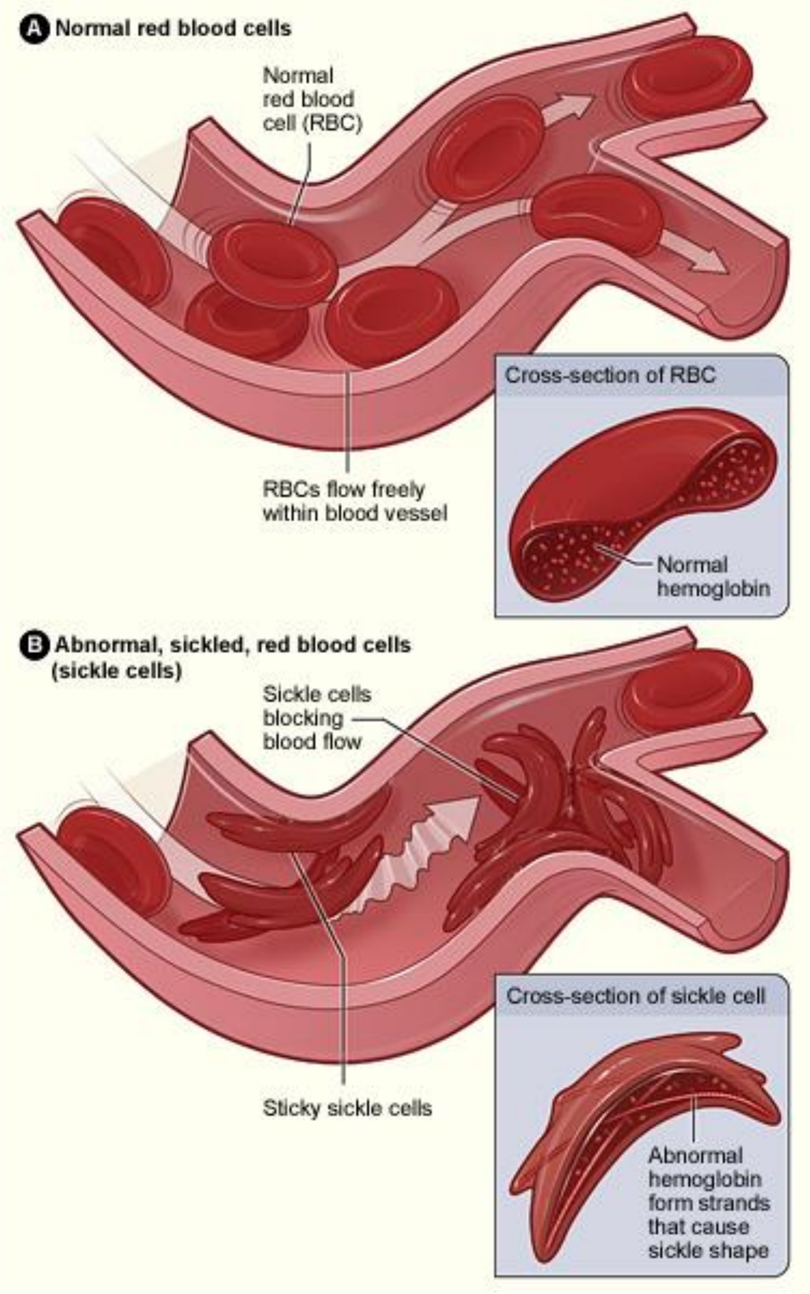
Sequence for normal hemoglobin

AUG	GUG	CAC	CUG	ACU	CCU	GAG	GAG	AAG	UCU	GCC	GUU	ACU
START	Val	His	Leu	Thr	Pro	Glu	Glu	Lys	Ser	Ala	Val	Thr



Sequence for sickle-cell hemoglobin

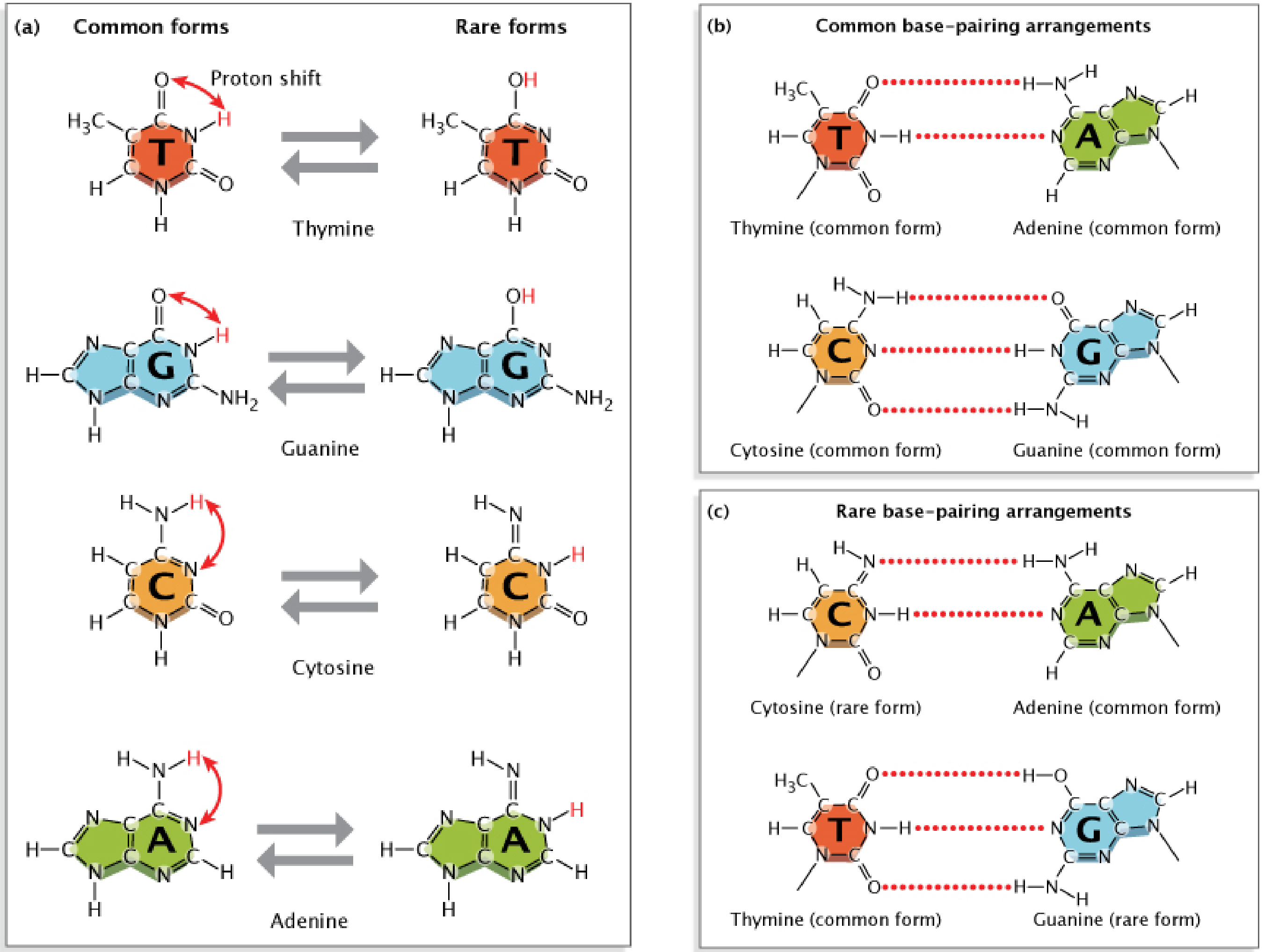
AUG	GUG	CAC	CUG	ACU	CCU	GUG	GAG	AAG	UCU	GCC	GUU	ACU
START	Val	His	Leu	Thr	Pro	Val	Glu	Lys	Ser	Ala	Val	Thr


A Normal red blood cells
Normal red blood cell (RBC)
RBCs flow freely within blood vessel
Cross-section of RBC
Normal hemoglobin
B Abnormal, sickled, red blood cells (sickle cells)
Sickle cells blocking blood flow
Sticky sickle cells
Cross-section of sickle cell
Abnormal hemoglobin form strands that cause sickle shape

Mutation can be due to internal mechanisms or external causes

- Endogenous
- Exogenous

There is (infrequent) chemical instability of Nts



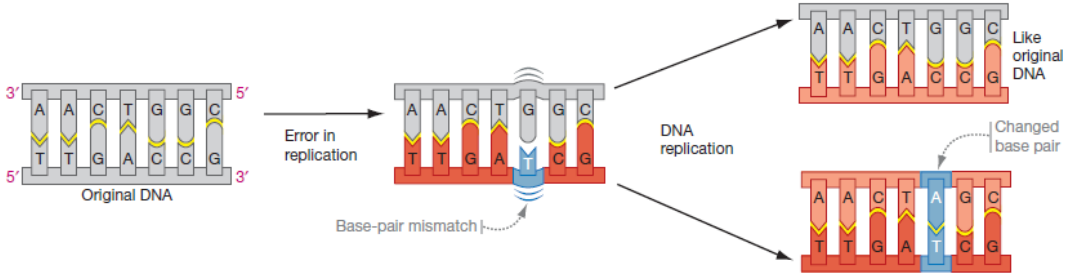


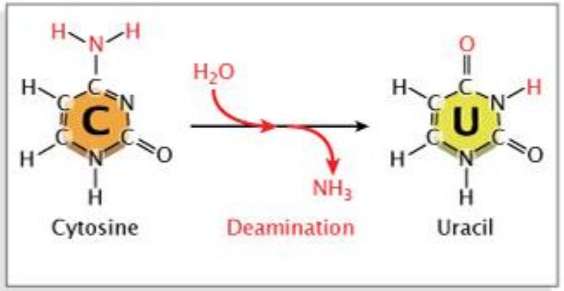
FIGURE 16.8 Unrepaired Mistakes in DNA Synthesis Lead to Point Mutations.

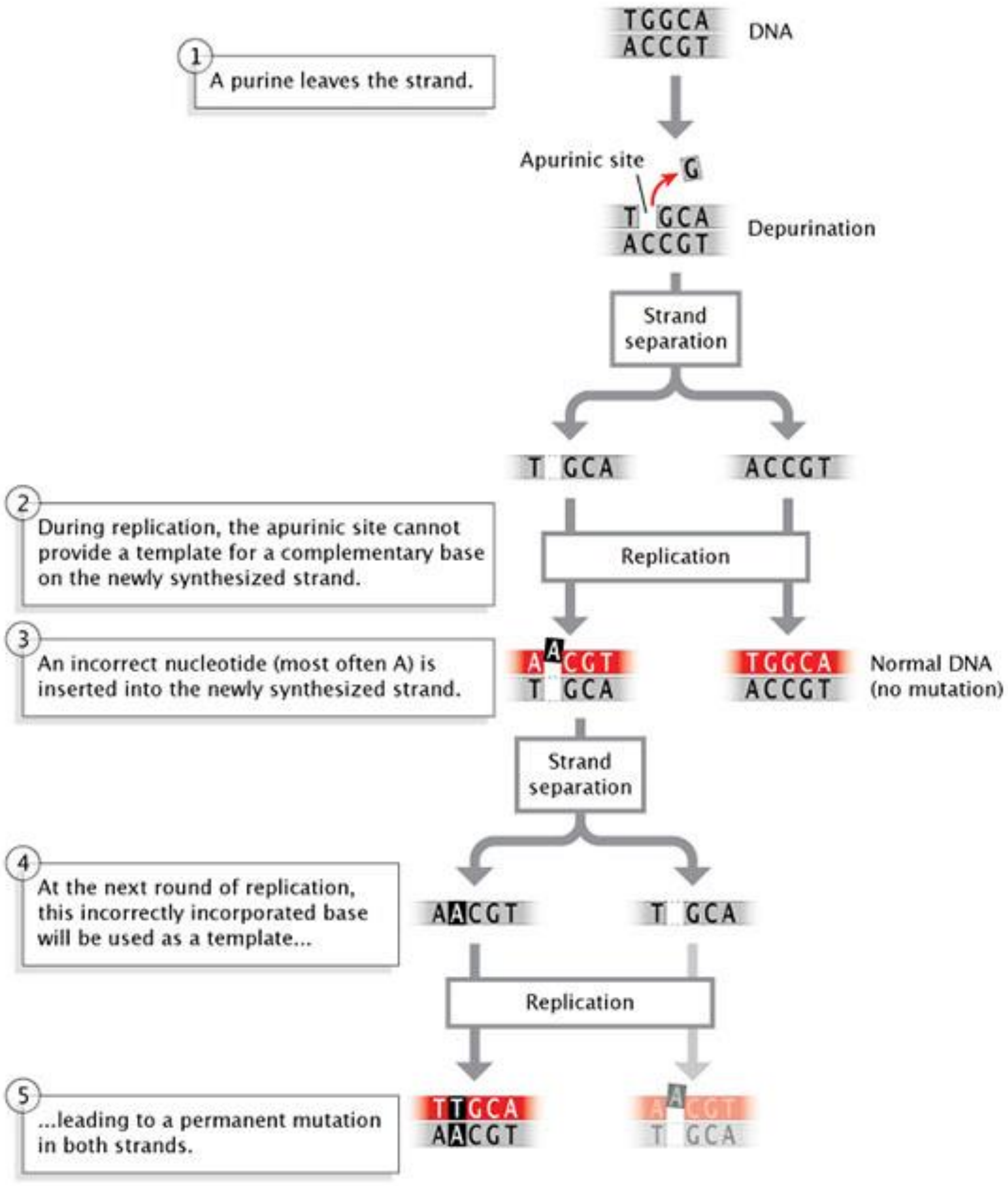
Depurination

- A purine base is lost through hydrolysis (without altering the backbone)
- If not promptly corrected by repair enzymes, the nucleotide change is incorporated in the next DNA replication

Deamination

- Removal of an amine group from a base
- Cytosine becomes Uracil and pairs with Adenine instead of Guanine




1 A purine leaves the strand.

2 During replication, the apurinic site cannot provide a template for a complementary base on the newly synthesized strand.

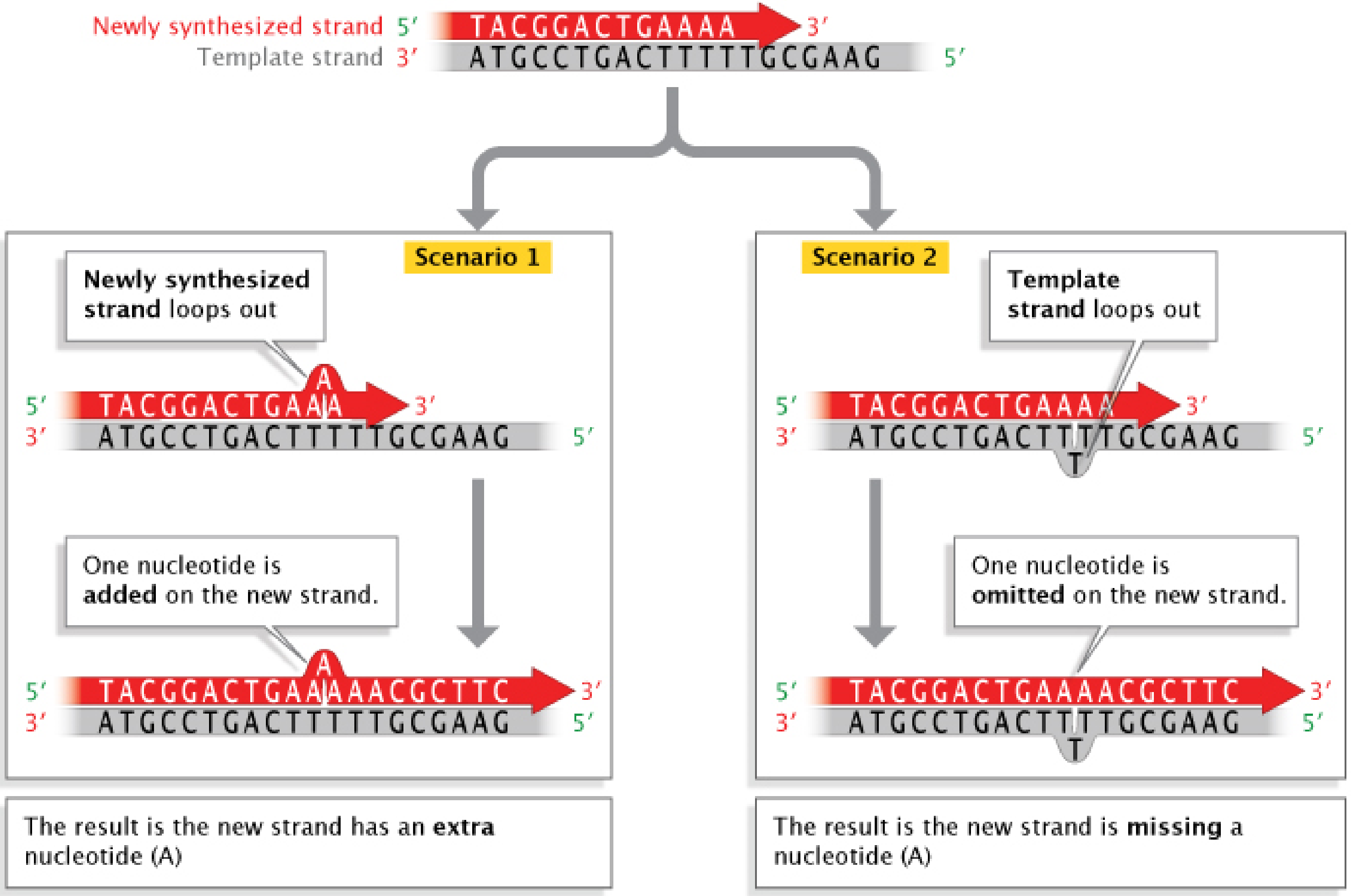
3 An incorrect nucleotide (most often A) is inserted into the newly synthesized strand.

4 At the next round of replication, this incorrectly incorporated base will be used as a template...

5 ...leading to a permanent mutation in both strands.

Errors in the DNA replication fork

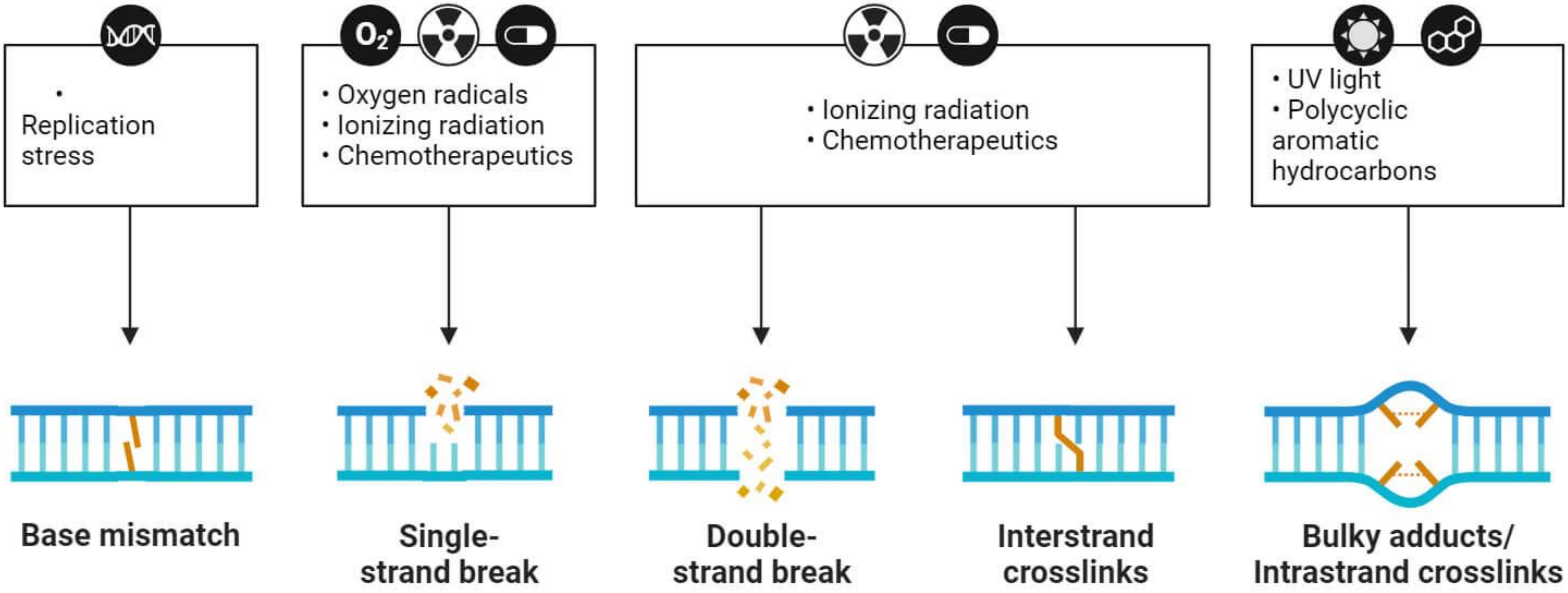
- Slipping of DNA pol
- More common with highly repeated sequences


Newly synthesized strand 5′ TACGGACTGAAAA 3′
Template strand 3′ ATGCCTGACTTTTTGCGAAG 5′
Scenario 1
Newly synthesized strand loops out
A
5′ TACGGACTGAAA 3′
3′ ATGCCTGACTTTTTGCGAAG 5′
One nucleotide is added on the new strand.
A
5′ TACGGACTGAAAAACGCTTC 3′
3′ ATGCCTGACTTTTTGCGAAG 5′
The result is the new strand has an extra nucleotide (A)
Scenario 2
Template strand loops out
5′ TACGGACTGAAAA 3′
3′ ATGCCTGACTTTTGCGAAG 5′
T
One nucleotide is omitted on the new strand.
5′ TACGGACTGAAAACGCTTC 3′
3′ ATGCCTGACTTTTGCGAAG 5′
T
The result is the new strand is missing a nucleotide (A)

Mutation agents

- UV may cause the hydrolysis of a cytosine base to a hydrate form; mispair with adenine, eventually a thymine gets incorporated
- Alternatively, UV light can also cause pyrimidine dimers (covalent bonds between adjacent pyrimidine bases)
- Ionizing radiation (X-ray, etc) induces double-stranded breaks in DNA
- Free radicals and oxidizing agents:
    - E.g. dioxin intercalates between base pairs, disrupting the integrity of the DNA helix
    - Many more….

Common Causes of DNA Damage



- Replication stress

- Oxygen radicals
- Ionizing radiation
- Chemotherapeutics

- Ionizing radiation
- Chemotherapeutics

- UV light
- Polycyclic aromatic hydrocarbons

Base mismatch

Single-strand break

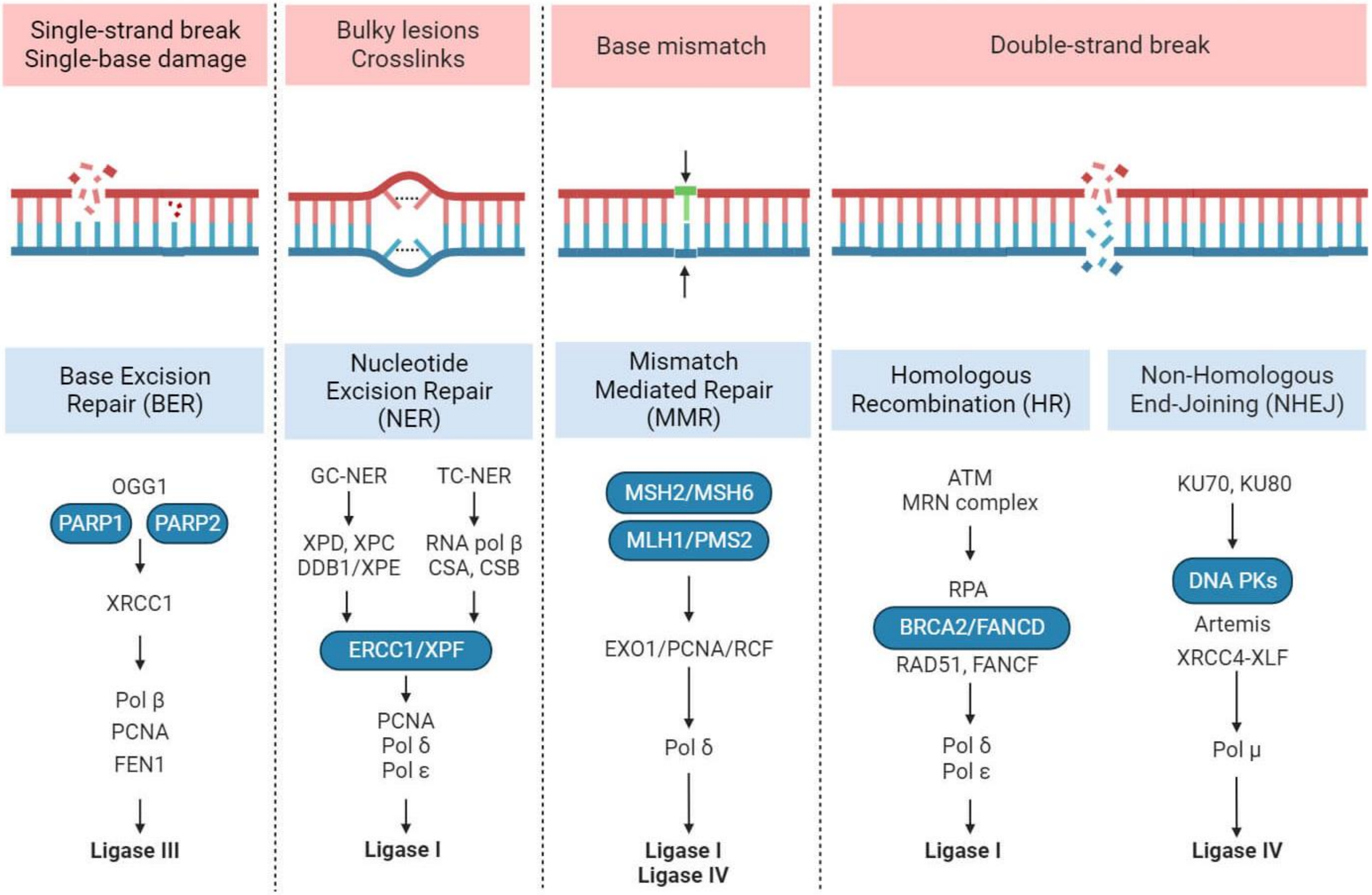
Double-strand break

Interstrand crosslinks

Bulky adducts/ Intrastrand crosslinks

DNA Repair Mechanisms

Microbe Notes



Single-strand break
Single-base damage

Base Excision Repair (BER)

OGG1
PARP1 PARP2
↓
XRCC1
↓
Pol β
PCNA
FEN1
↓
Ligase III

Bulky lesions
Crosslinks

Nucleotide Excision Repair (NER)

GC-NER → XPD, XPC, DDB1/XPE
TC-NER → RNA pol β, CSA, CSB
↓
ERCC1/XPF
↓
PCNA
Pol δ
Pol ε
↓
Ligase I

Base mismatch

Mismatch Mediated Repair (MMR)

MSH2/MSH6
MLH1/PMS2
↓
EXO1/PCNA/RCF
↓
Pol δ
↓
Ligase I
Ligase IV

Double-strand break

Homologous Recombination (HR)

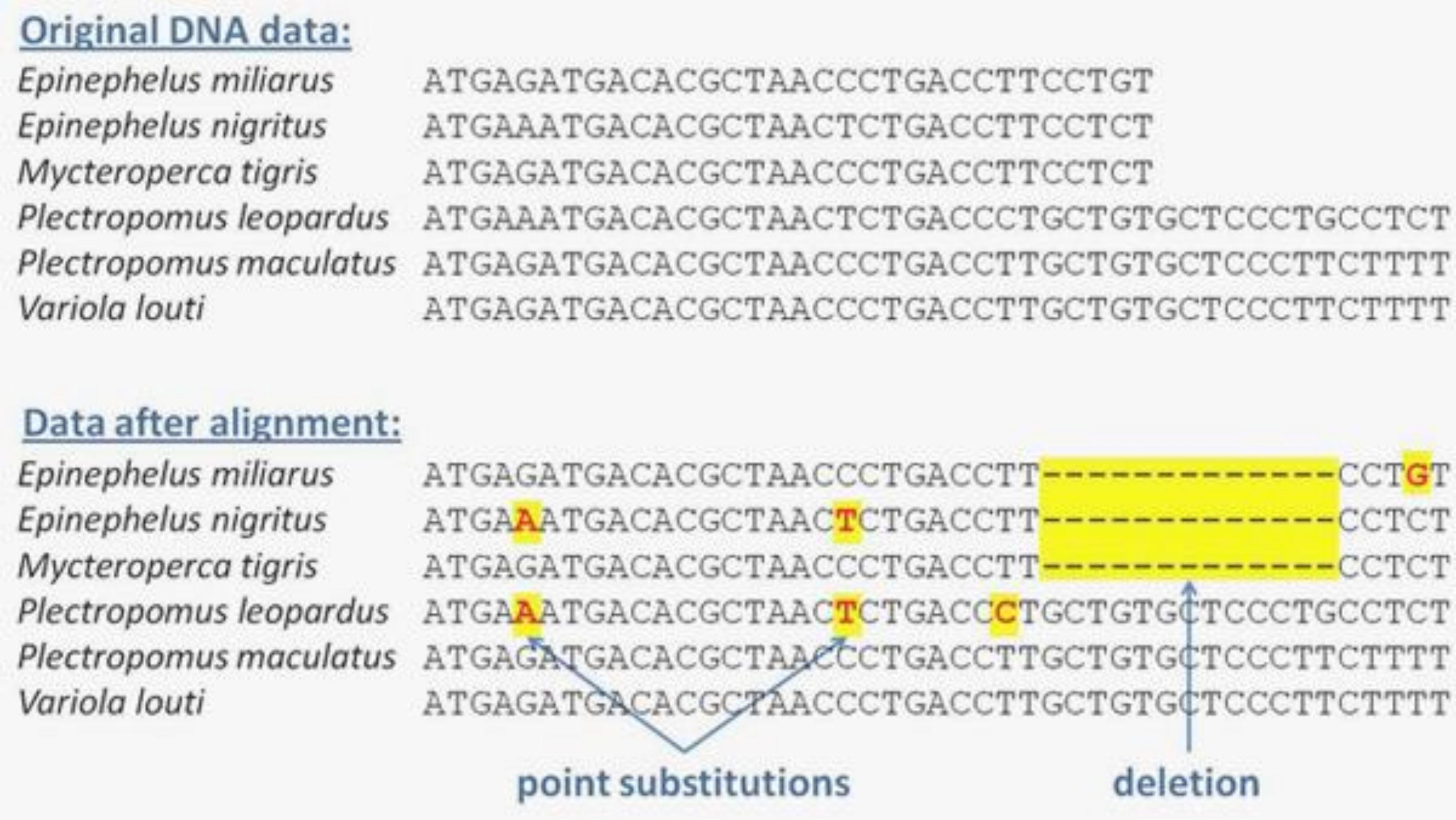
ATM
MRN complex
↓
RPA
BRCA2/FANCD
RAD51, FANCF
↓
Pol δ
Pol ε
↓
Ligase I

Non-Homologous End-Joining (NHEJ)

KU70, KU80
↓
DNA PKs
Artemis
XRCC4-XLF
↓
Pol µ
↓
Ligase IV

Making sense of mutations in genomic sequences

- Sequence alignment is used to detect mutations
- Large mutations are more challenging to identify, depending on sequencing technology used



Original DNA data:

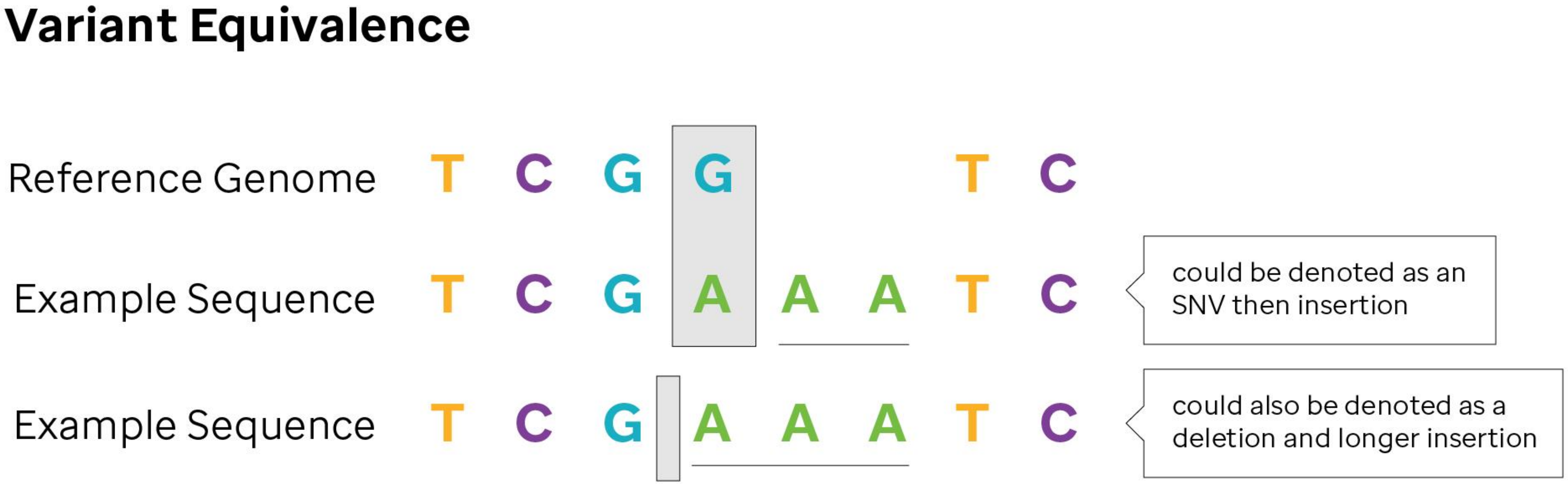
```
Epinephelus miliarus      ATGAGATGACACGCTAACCCTGACCTTCCTGT
Epinephelus nigritus      ATGAAATGACACGCTAACTCTGACCTTCCTCT
Mycteroperca tigris       ATGAGATGACACGCTAACCCTGACCTTCCTCT
Plectropomus leopardus    ATGAAATGACACGCTAACTCTGACCCTGCTGTGCTCCCTGCCTCT
Plectropomus maculatus    ATGAGATGACACGCTAACCCTGACCTTGCTGTGCTCCCTTCTTTT
Variola louti             ATGAGATGACACGCTAACCCTGACCTTGCTGTGCTCCCTTCTTTT
```

Data after alignment:

```
Epinephelus miliarus      ATGAGATGACACGCTAACCCTGACCTT-------------CCTGT
Epinephelus nigritus      ATGAAATGACACGCTAACTCTGACCTT-------------CCTCT
Mycteroperca tigris       ATGAGATGACACGCTAACCCTGACCTT-------------CCTCT
Plectropomus leopardus    ATGAAATGACACGCTAACTCTGACCCTGCTGTGCTCCCTGCCTCT
Plectropomus maculatus    ATGAGATGACACGCTAACCCTGACCTTGCTGTGCTCCCTTCTTTT
Variola louti             ATGAGATGACACGCTAACCCTGACCTTGCTGTGCTCCCTTCTTTT
```

The definition of mutations vary also in relation with bioinformatic tools

- The bonduaries between types of mutations are not always clear cut



Variant Equivalence

Reference Genome T C G G T C

Example Sequence T C G A A A T C

could be denoted as an SNV then insertion

Example Sequence T C G A A A T C

could also be denoted as a deletion and longer insertion