Repeated Elements in Plant Genomes

1. Telomeres
2. Centromeric Repeats
3. Retrotransposons (Class I Transposons)
4. DNA transposons (Class II Transposons)
5. MITEs (Miniature Inverted Terminal Repeat Elements)

Telomeres

Telomeres are the physical ends of linear chromosomes

Consists of nucleic acid/protein complexes in the vast majority of cases

Present in eukaryotic organisms

Molecular clock to monitor replicative history of the cell

Telomeres

Telomeres are maintained using:

1. RNA template (TER locus)
2. Reverse Transcriptase activity

Telomerase activity maintains the terminal DNA repeats

Telomerase binding proteins (TRFs) bind single and double stranded telomerase repeats

TRFs

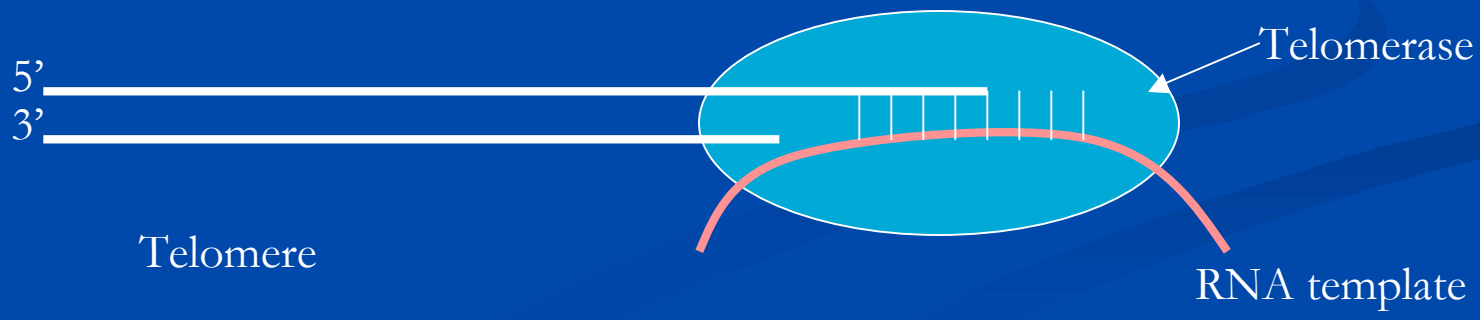
1. Protect against DNA repair
2. End-joining of chromosomes
3. Spurious exonuclease activity

Telomeres

Initial sequencing of end fragments of DNA from chromosomes showed they possessed tandem arrays of simple repeats

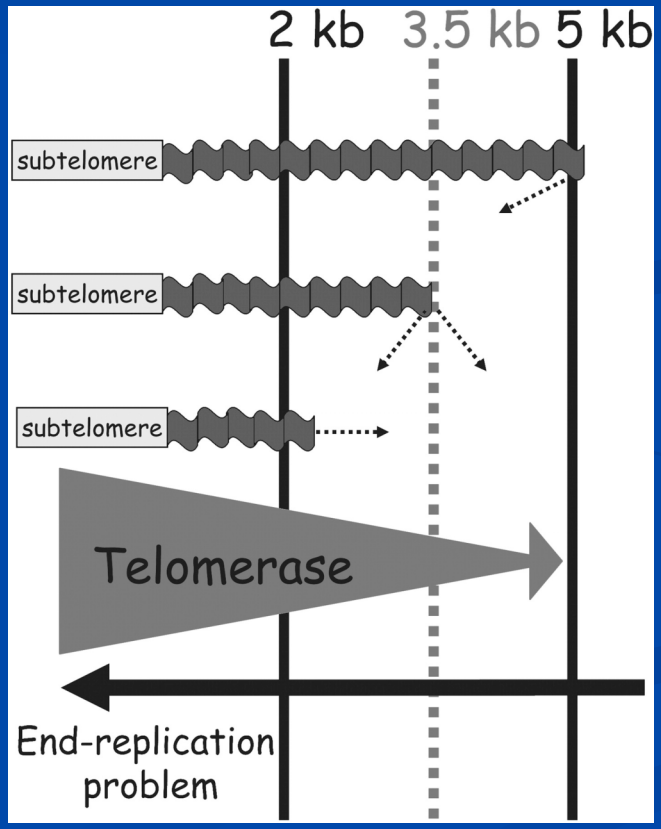
Humans	(TTAGG)n
Arabidopsis	(TTAGGG)n
Rice	(TTTAGGG)n

The RNA template from the TER locus is a complement to the repeat and is used to extend the telomere



Telomeres

This coordinated activity solves the end-replication problems for the chromosome and ensures the telomeres maintain their length



McKnight and Shippen, 2004

Telomeres

Loss of telomerase activity will yield severe phenotypes after several mitotic cycles

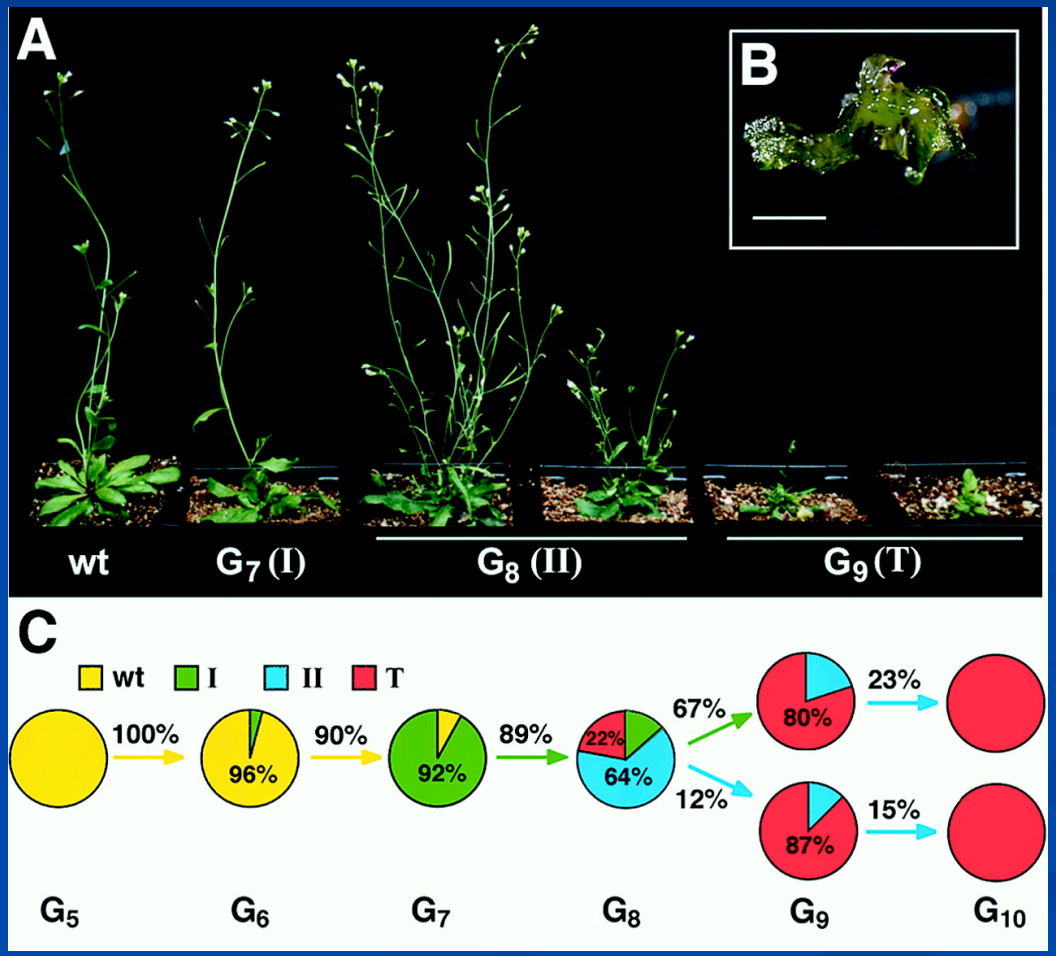
Arabidopsis plants lacking telomerase will begin showing pleiotropic effects in the 6^{th} and 7^{th} generations

By the 9^{th} generation, these plants have entered a terminal stage of sterility and dwarfism

By the 10^{th} generation, the effects are lethal.

Telomeres

Visualization of the phenotypic progression in successive generations resulting from a loss of telomerase

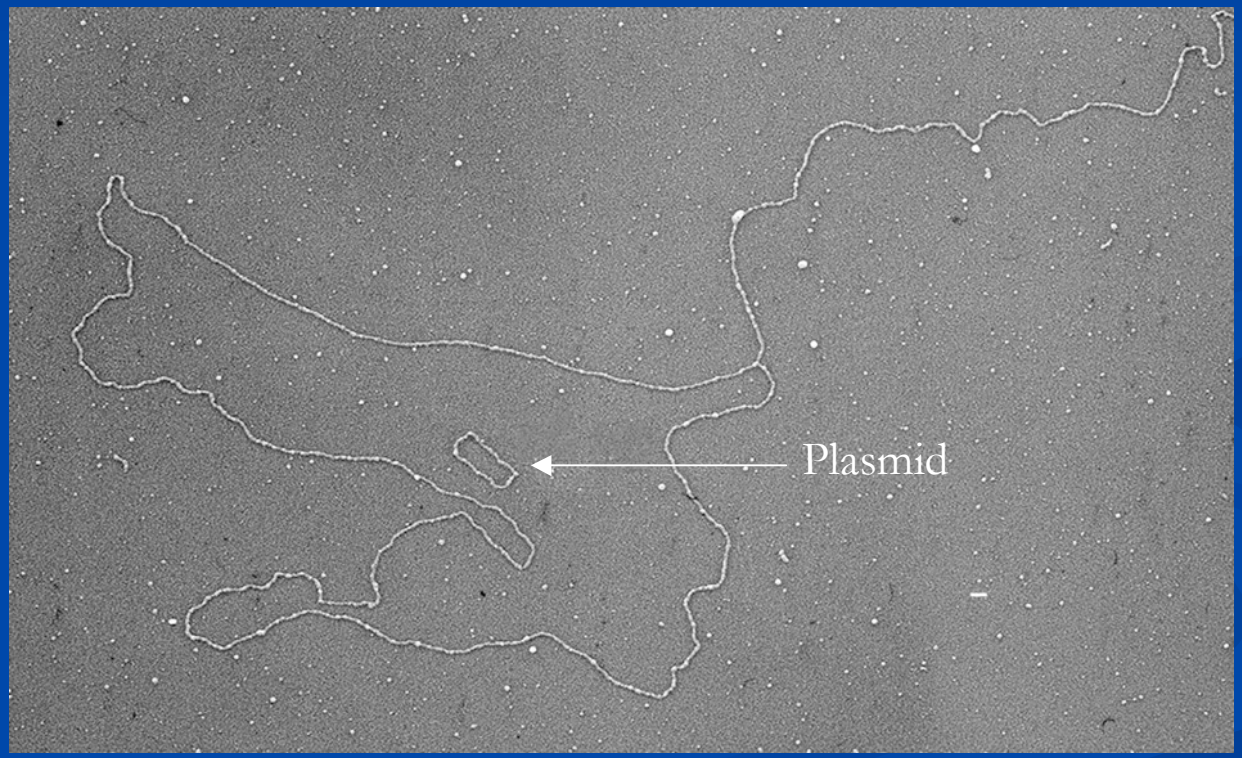


Riha et al 2001

Telomeres

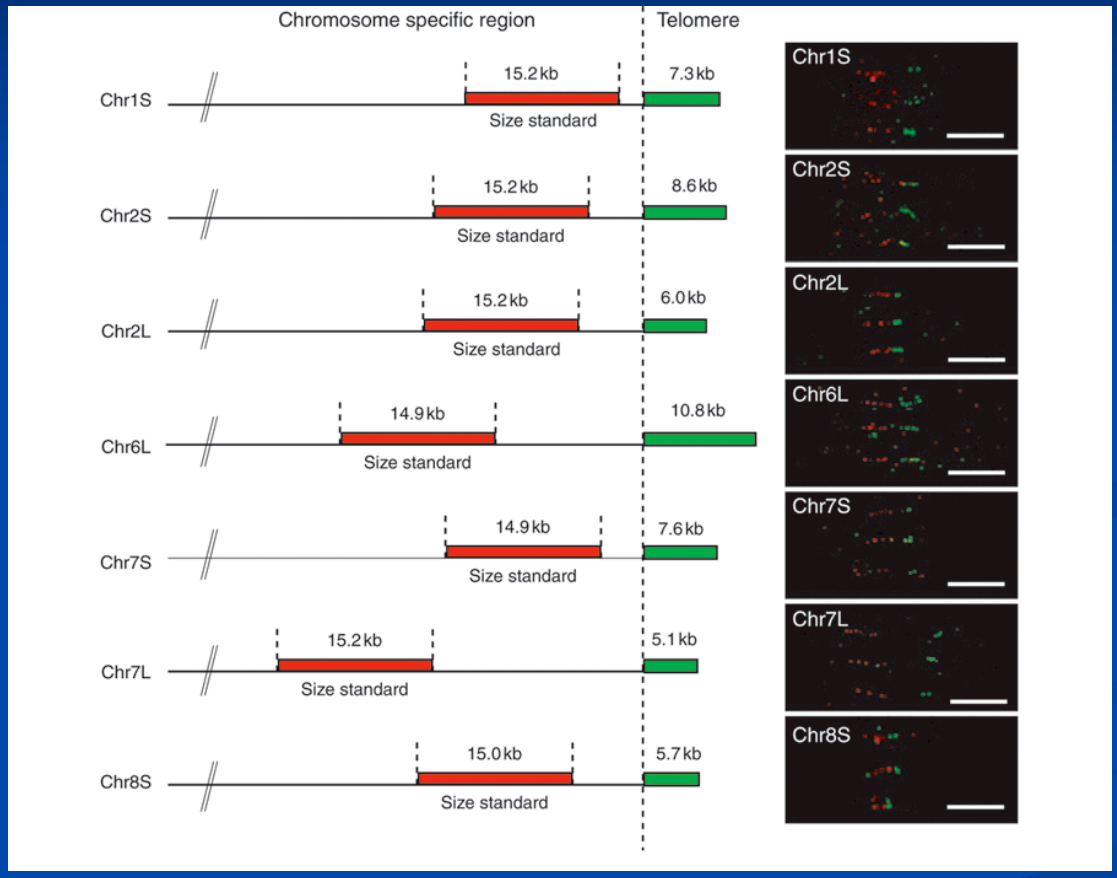
SEM of a telomere loop from Pea

Note circular plasmid ~3kbp in length inside telomere loop



McKnight and Shippen, 2004

Telomeres in rice have been characterized



Mizuno et al., 2006

Subtelomere/telomere junctions have polymorphic telomere repeats

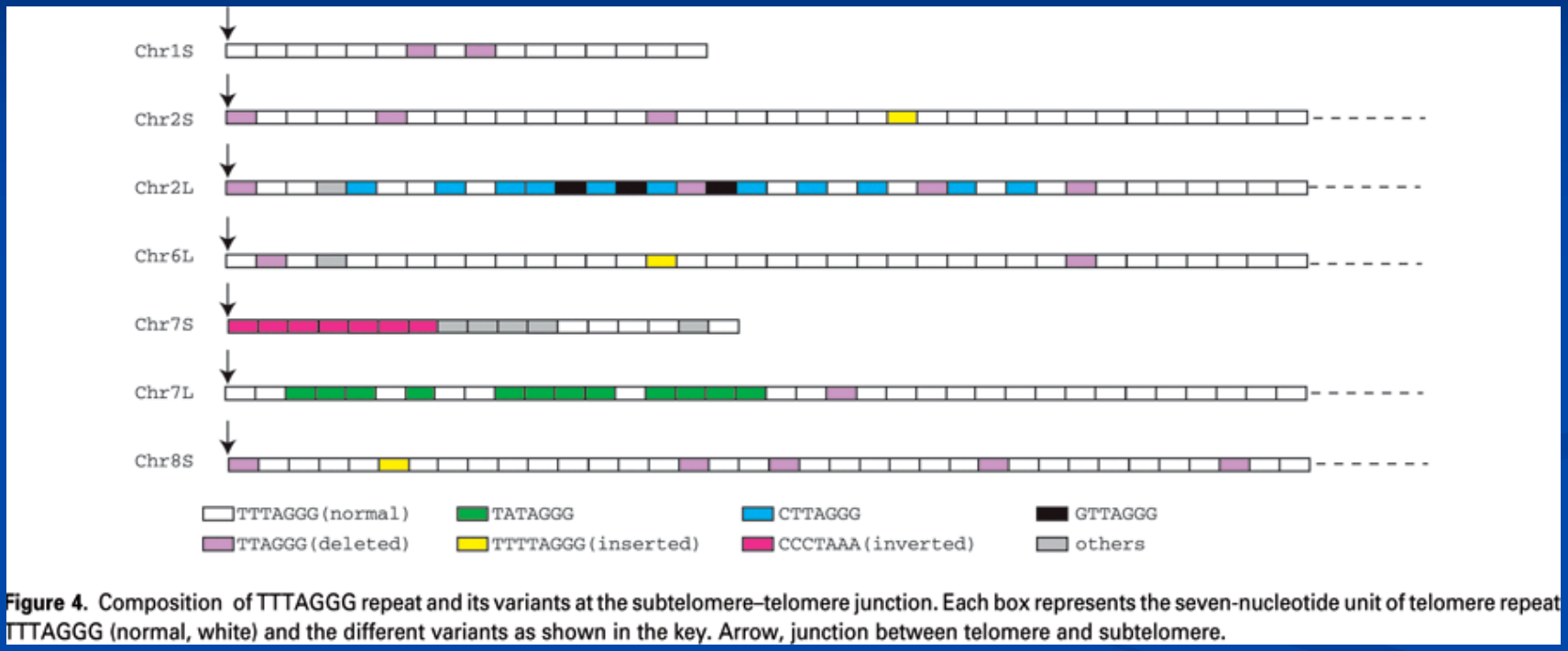


Figure 4. Composition of TTTAGGG repeat and its variants at the subtelomere–telomere junction. Each box represents the seven-nucleotide unit of telomere repeat TTTAGGG (normal, white) and the different variants as shown in the key. Arrow, junction between telomere and subtelomere.

Mizuno et al., 2006

Centromeres

Centromeres are heterochromatic components of the genome with vital roles

Centromeres serve as the assembly point for the kinetochore for post-replicative chromosome division

Centromeres are:

1. Relatively “gene” poor
2. Dense with various types of repeats

These repeats consist of satellite DNA and transposable elements

Centromeres

Estimated sizes range from 125 bp (yeast) to several megabases (maize)

Varying structural arrangements:

An ordered arrangement of repeats (fission yeast)

Tandem arrays of repeated sequence studded with transposable elements (plants, humans)

The core centromere binds the protein CENH3

CENH3 is a variant of the histone H3 but is associates specifically with the centromere

CENH3 among species has conserved histone domain but a divergent N terminal domain

Centromeres

In rice, the centromeric satellite repeats are 155 bp in length

These satellite repeats are called CentO in rice

Centromeric repeats are species specific and widely divergent among eukaryotes

Satellite repeat organization can vary widely among the chromosomes of a species

The centromeric and pericentromeric regions also have significant content of retrotransposons

The combined size and repetitive nature of centromeres make them difficult to sequence completely

Centromeres

Centromeres from rice chromosomes 4 and 8 have been sequenced completely

Chromosome 4:

18 separate tracts of CentO repeats clustered in 124 kb of sequence

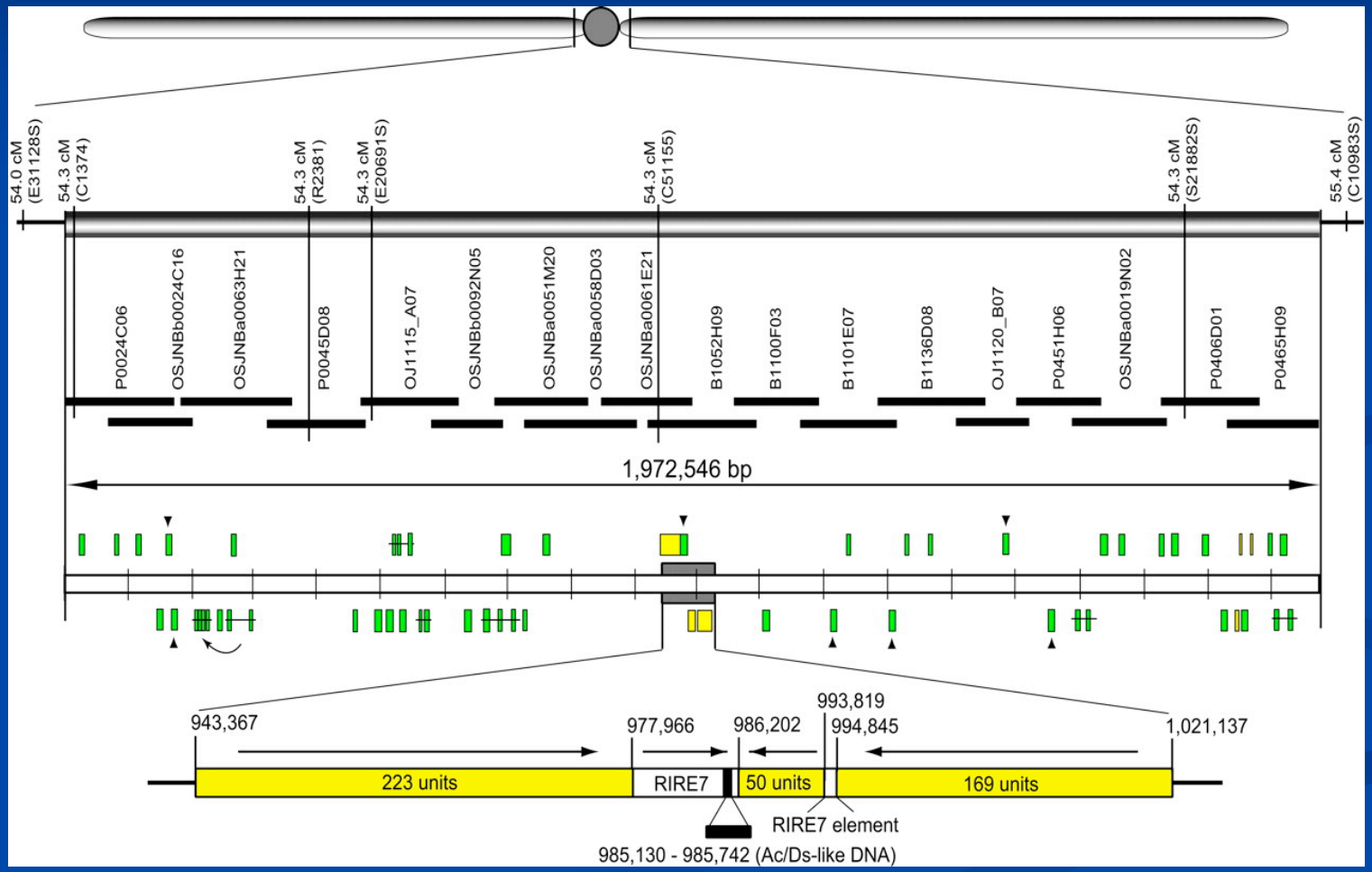
Chromosome 8:

3 separate tracts of CentO repeats clustered in 78kb of sequence

Note that the arrangement of the CentO repeats is very distinct between the two different centromeres in rice

Centromeres

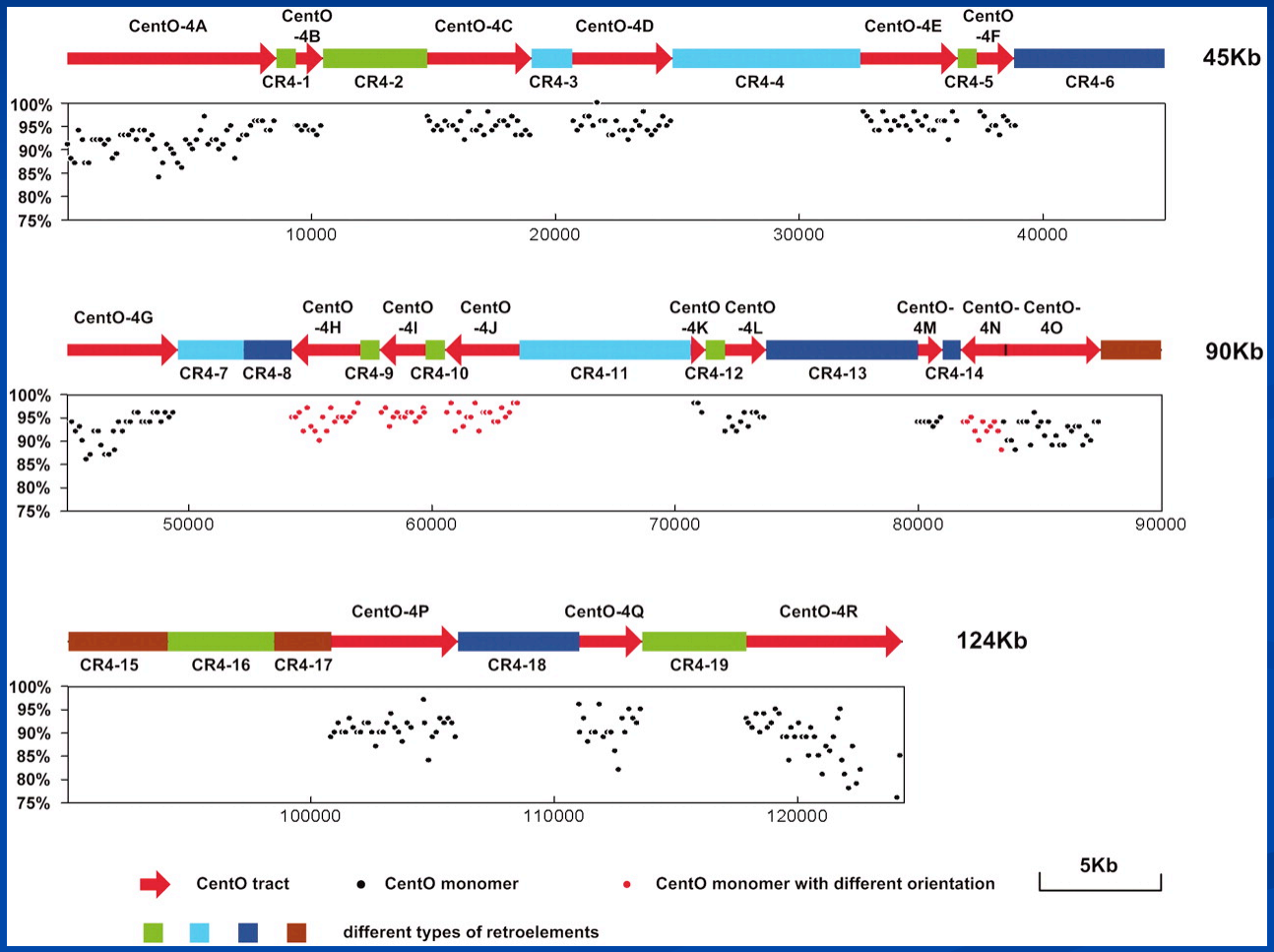
Schematic of the centromere of chromosome 8 of rice



Wu et al 2004

Centromeres

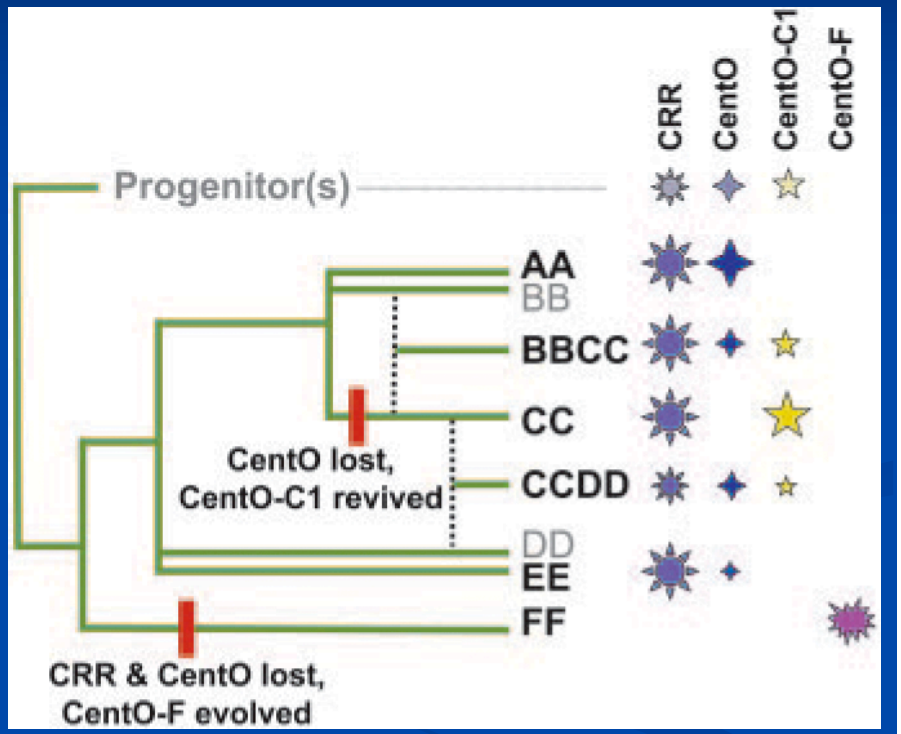
Schematic of the centromere of chromosome 4 of rice



Zhang et al 2004

Divergence in Centromere repeats in the Oryza genus

CRR – retroelement specific to centromeres
CentO – Oryza centromeric repeat
CentO-C1 – Centromeric repeat that shares homology with maize and rice
CentO-F – No homology to CentO or CentO-C1 Novel centromeric repeat
AA, BB, CC, DD, EE, FF are the names of genome types in Oryza genus
AABB is a tetraploid alloploidy event



Dawe, 2005

Note that the CC genomes have replaced the CentO with a divergent repeat

Note the FF genome has a novel centromeric repeat (AA diverged from FF ~7-9 million years ago

Retrotransposons

Retrotransposons are Class I transposable elements

Ubiquitous in the plant kingdom, well studied in monocots

A major constituent of many plant genomes

Mobilization via an RNA intermediate that leads to accumulation within the genome

Significant structural relationships to retroviruses

Can create mutations and affect transcription of neighboring genes

Retrotransposons

- Retrotransposons are Class I transposable elements

Features common to these elements:

LTR – Long terminal repeats

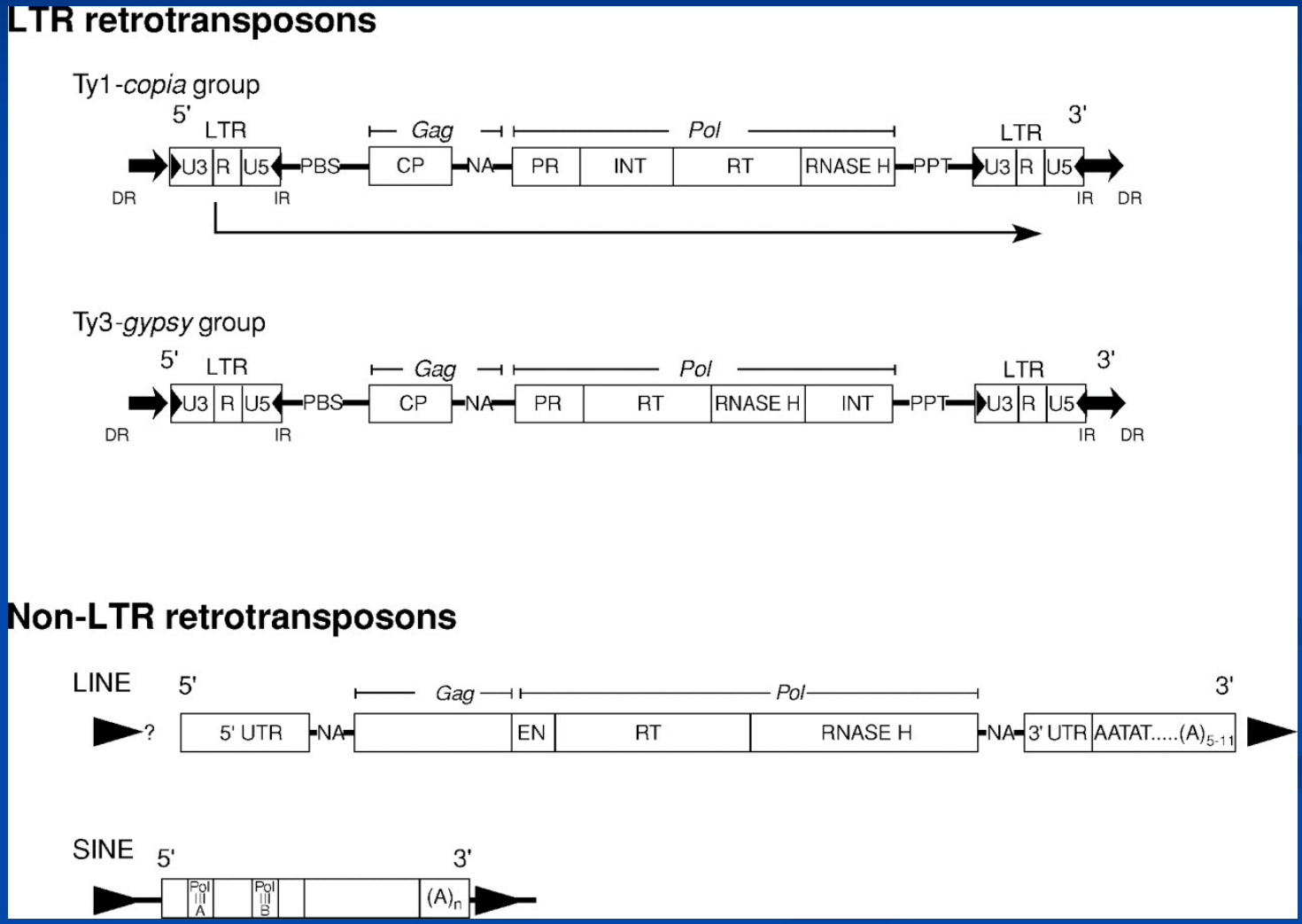
PBS – Primer binding site

Coding sequence – *gag, pol, int* genes

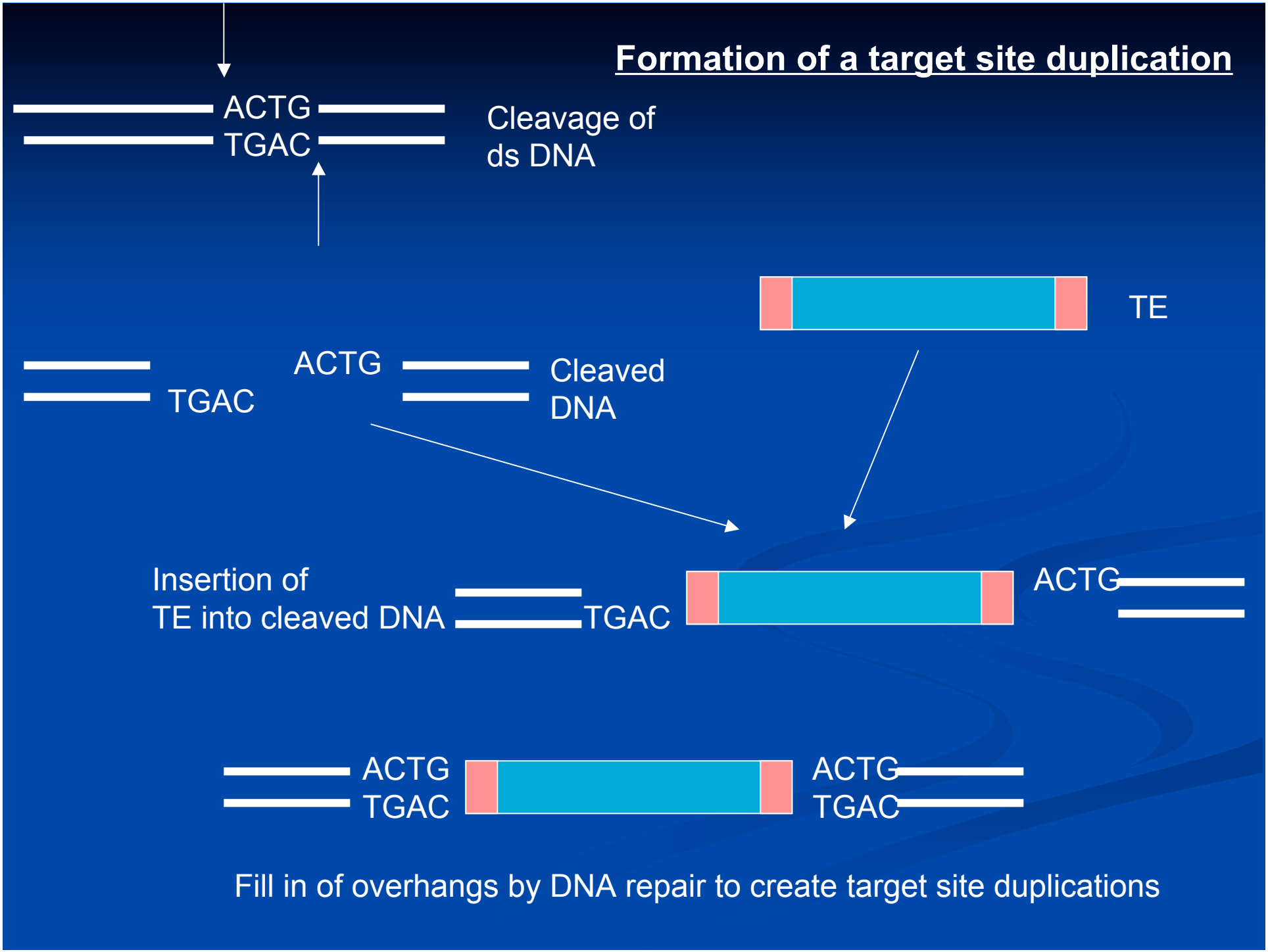
PPT – Polypurine tract

TSD – Target site duplication

Retrotransposons



Kumar and Bennetzen 1999


Formation of a target site duplication
ACTG
TGAC
Cleavage of
ds DNA
TE
TGAC
ACTG
Cleaved
DNA
Insertion of
TE into cleaved DNA
TGAC
ACTG
ACTG
TGAC
ACTG
TGAC
Fill in of overhangs by DNA repair to create target site duplications

Retrotransposons

LINE – Long Interspersed Repetitive Elements

LINEs are related to LTR transposons, but distinct in their structure

Differences between LINEs and retrotransposons:

- LINEs lack LTRs
- gag protein encodes a endonuclease activity (cleave DNA)
- pol has RT and RNaseH motifs but lacks an integrase
- Has internal RNA pol II and pol III promoters

Retrotransposons

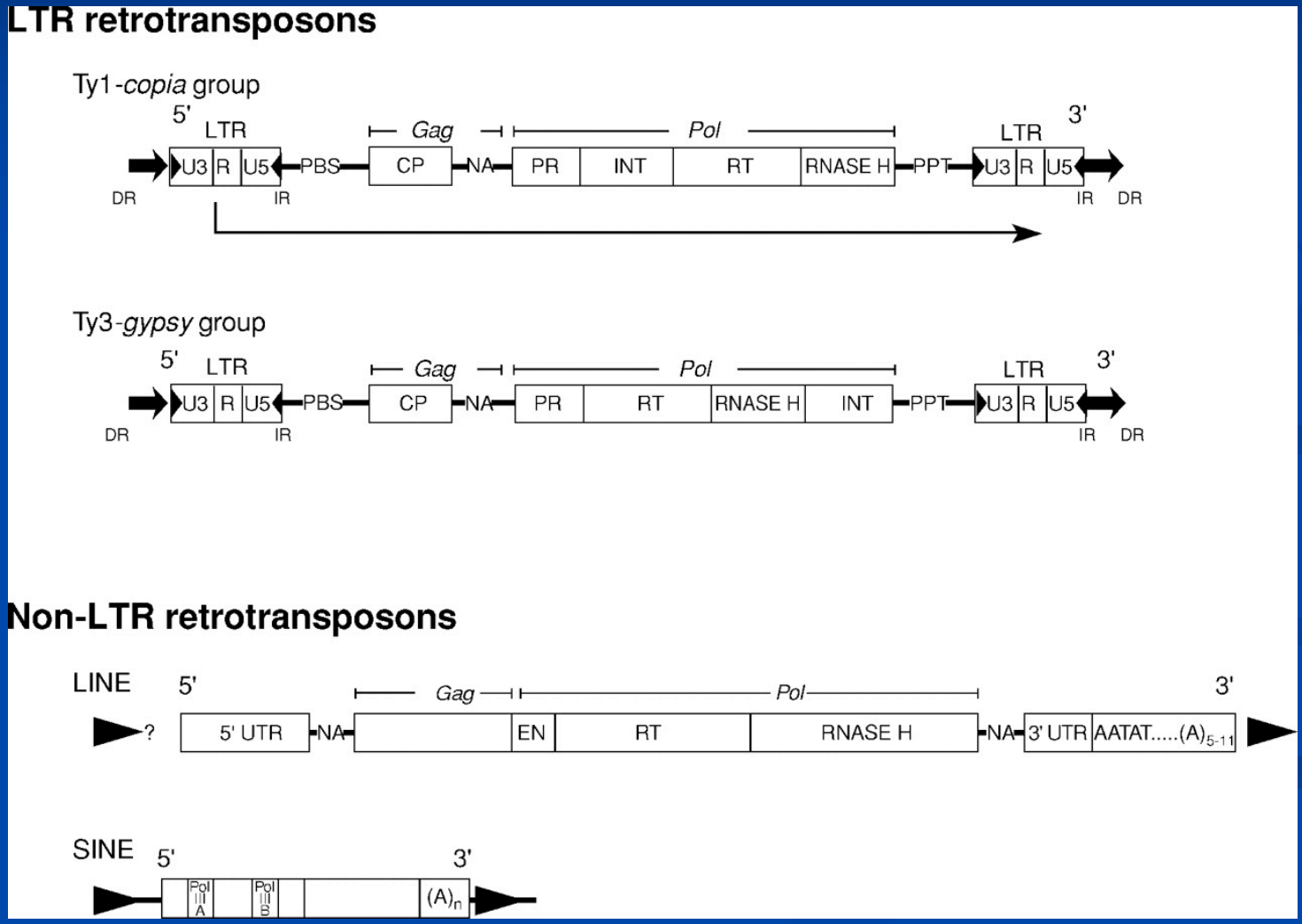
SINE – Short Interspersed Nuclear Element

SINEs are originally derived from tRNA sequences

SINEs are distinct from retrotransposons

- Short (<500bp) nonautonomous elements
- These elements lack LTRs and introns
- Possess an encoded polyA tail
- Cross-mobilization would need to be the method for transposition

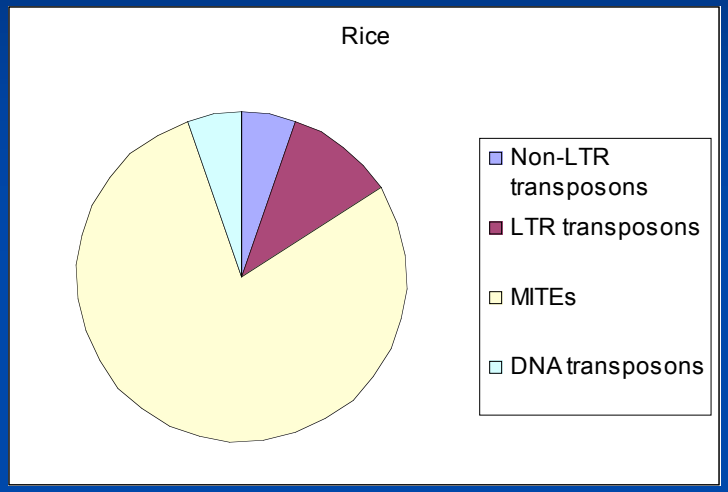
Retrotransposons

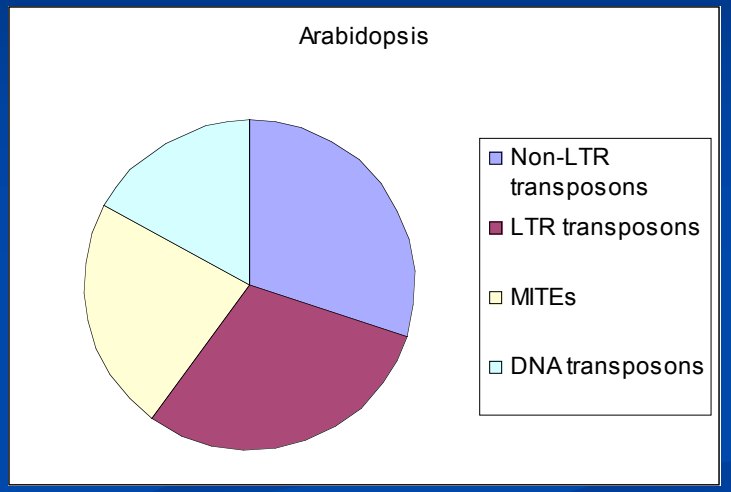


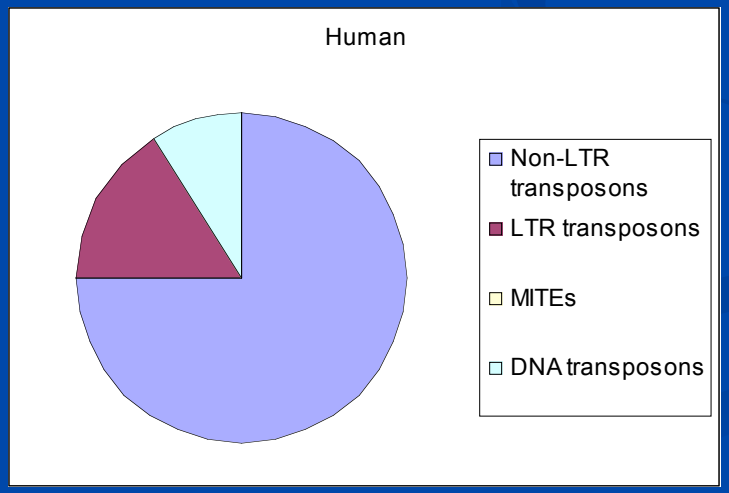
Kumar and Bennetzen 1999

Retrotransposons

Copy number for classes of elements varies among genomes







Retrotransposons

Retrotransposons are ubiquitous in higher eukaryotes:

Maize genome is ~3000 Mbp
- >50% genome is comprised of retrotransposons

Rice genome is ~375 Mbp
- ~20% genome is retrotransposons

Arabidopsis genome is ~130 Mbp
- < 10% genome is retransposons

Retroelement copy number is a major determinant of genome size variation in higher plants

Retrotransposons

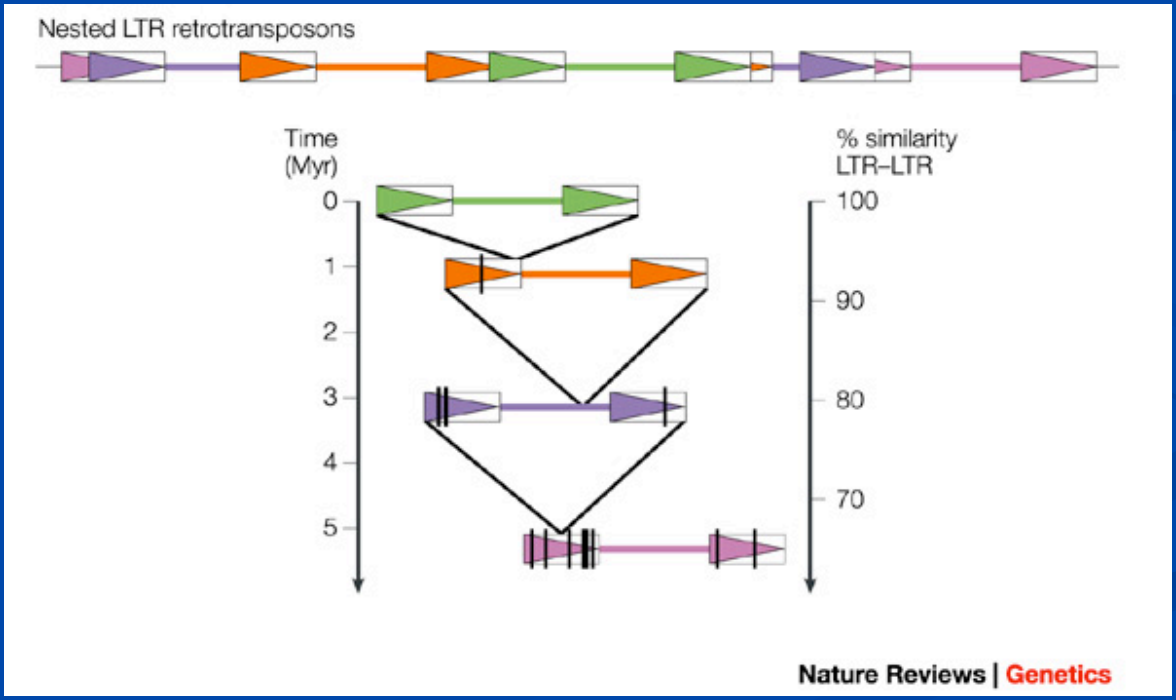
Maize and sorghum comparison as an illustration:

- Diverged an estimated ~15 mya from one another
- Both have 10 chromosomes
- Excellent conservation of gene order (synteny)
- Maize genome is >4x larger than the sorghum genome
- Sequence analysis indicates that maize genome expansion is due to retrotransposon expansion

Retrotransposons

In maize, retroelements are often found as “nested” insertions.
(Nested means that one element is inserted into another which is inserted into another)

Using the tandemly repeated LTRs, you can estimate the age of the retrotransposon by looking at rate of mutation



Feschotte et al 2002

Tos17 mediated gene tagging

The *Tos* family of retrotransposons have been characterized in rice

Three of the *Tos* family (*Tos10, Tos17, Tos19*) have been shown to be active under tissue culture conditions

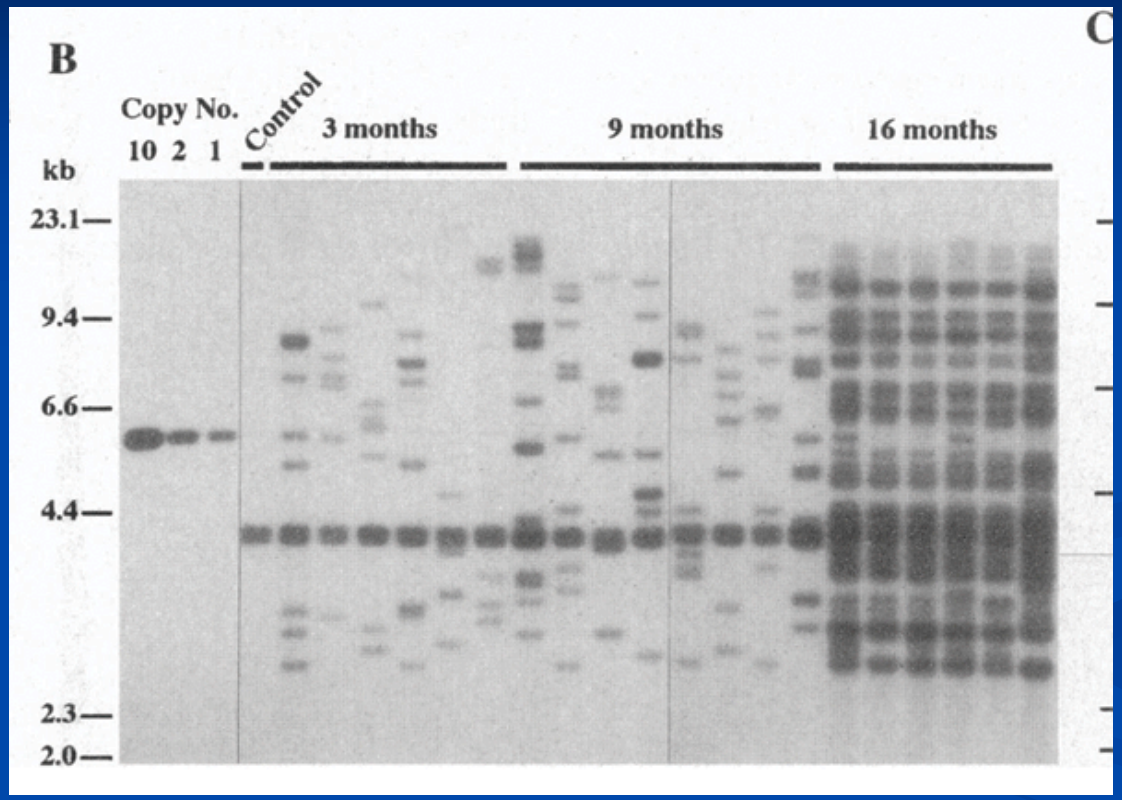
Tos17 was found to only have two copies in the Nipponbare genome

Tos17, when activated, has a preference for insertion into low copy sequences in the rice genome

Tos17 activation leads to a gradual accumulation of Tos17 elements in the genome

Tos17 is being used as a functional genomics tool in rice for tagging genes

Tos17 mediated gene tagging



The Southern Blot shows the accumulation of *Tos17* elements in plants that were regenerated from calli that had been in tissue culture for 3, 9, and 16 months.

Hirochika et al.., 1996

DNA Transposons

DNA Transposons are Class II transposable elements

Ubiquitous in the plant kingdom

May be autonomous or non-autonomous elements

Mobilization via a cut and paste mechanism

Low copy number per genome (<100 per genome per family)

Can create mutations and affect transcription of neighboring genes

DNA Transposons

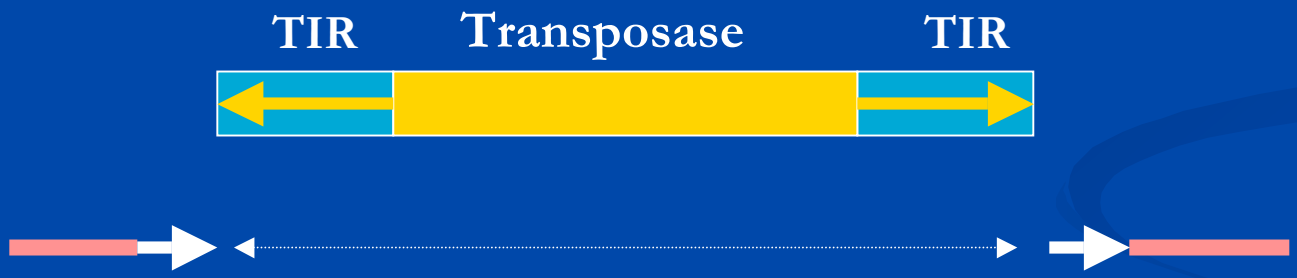
DNA transposons (Class II) have several key features

1. Target site duplications produced upon insertion
2. An ORF containing the catalytic domain for transposase
3. TIR (Terminal Inverted Repeats) that can form a hairpin
4. Subterminal regions that may possess binding motifs for transposase

DNA Transposons

DNA Transposons mobilize via a cut and paste mechanism

Cutting the transposon from genomic DNA



End-joining genomic DNA (pink)

Transposon inserts at a new location

DNA Transposons

Autonomous elements encode (minimally) a full-length transposase and TIRs

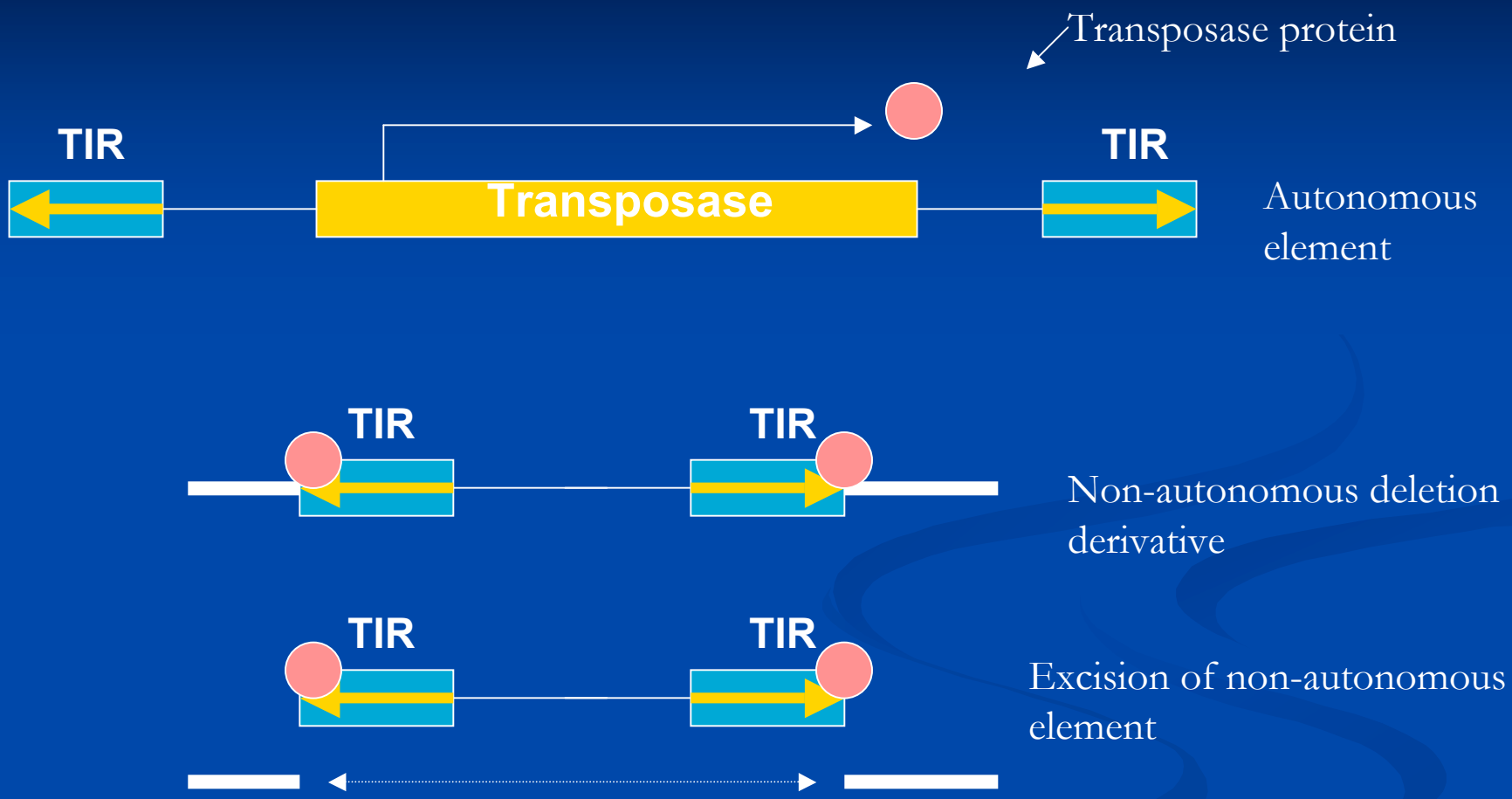
Non-autonomous elements are truncation of the parent (autonomous) elements

Autonomous

Non-autonomous

Non-autonomous

DNA Transposons



Non-autonomous elements can be moved in trans by a transposase encoded by the autonomous element

DNA Transposons

DNA Transposons (autonomous and non-autonomous) are used for functional genomics

In rice: Use of *Activator* and *Ds* from maize by transformation

These elements can insert into a gene leading to a non-functional allele and phenotype

Example: The promoter of *frizzy panicle* locus was tagged with *Ds*

These mutations are now called “transposon-tagged” and can be cloned

Example: Screen for *Ds* using PCR to obtain flanking sequence

MITEs

MITEs are Miniature Inverted Terminal Repeat Elements

Ubiquitous in the plant kingdom

Commonly associated with genic regions

Can attain high copy number (>10,000 per genome/family)

Derived from DNA class II transposons in many cases

Rapid expansion (burst) in genomes

MITEs

Generalized features of MITEs

1. Small relative size (<600 bp)
2. TIRs that are similar in size with DNA transposons
3. 3bp TSD
4. Share TIR sequence motifs with DNA tranposons
5. Mobilization via transposases produced from autonomous DNA transposon *in trans*
6. Extremely high copy numbers
7. Phylogenies are indicative of rapid expansion

MITEs

MITEs were originally found in a computer search of maize genomic DNA

The original element *Tourist* was found in the waxy locus of maize

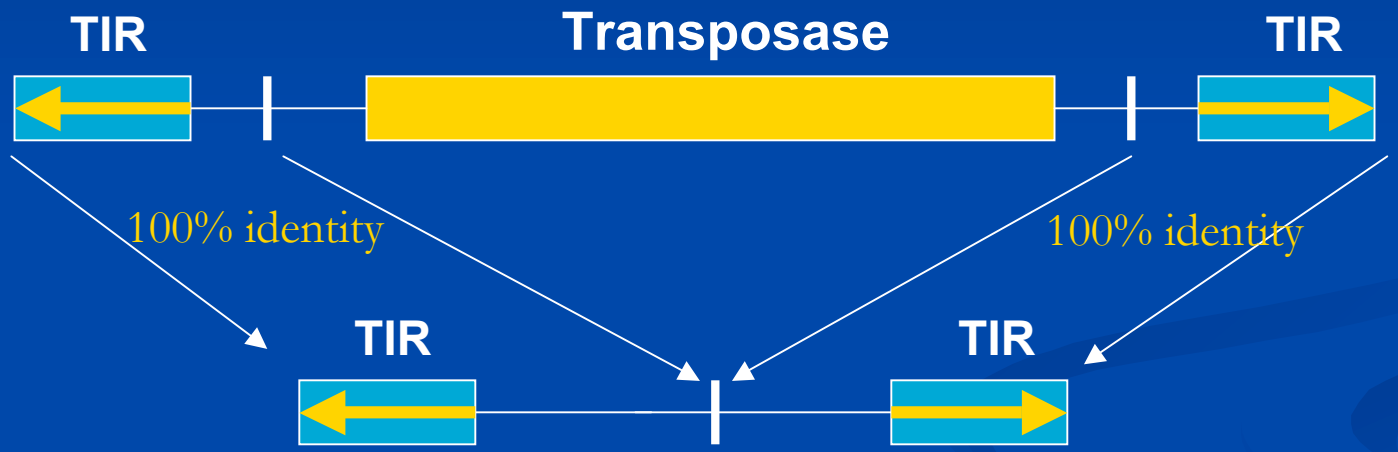
Stowaway was found in sorghum genomic DNA

MITEs are found throughout the plant kingdom

**MITEs are viewed as derivatives of autonomous elements which may be recent or ancient

MITEs

MITEs can be the product of a direct deletion:

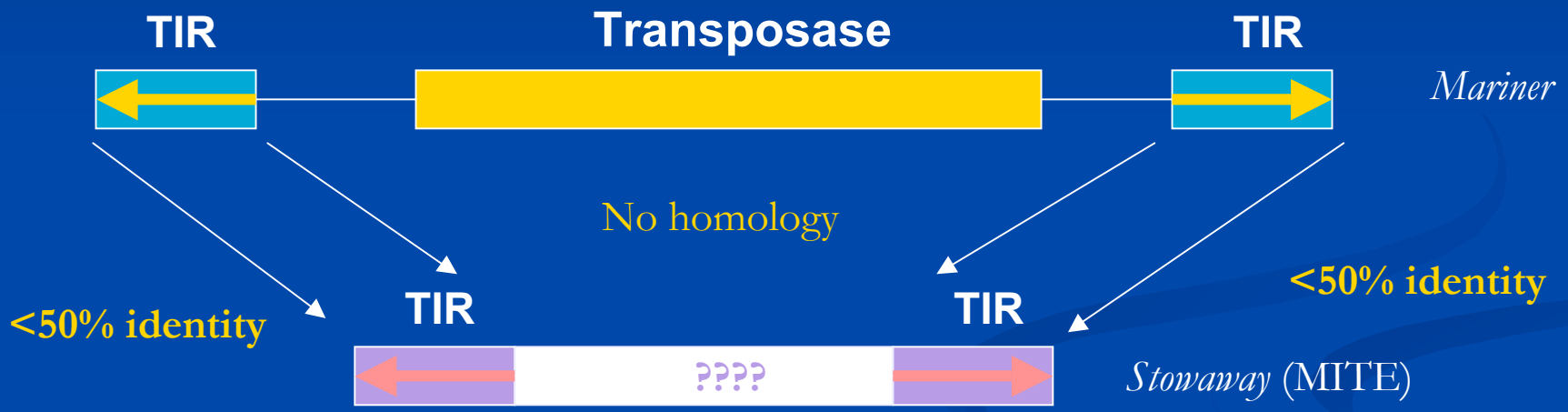


Example: mPING is a direct deletion of the autonomous element Ping
mPIF is a direct deletion of the autonomous element PIF

Copy number: 72 copies *mPING* and 1 copy *PING* in rice genome

MITEs

MITEs can be highly diverged from a presumptive autonomous element:



Example: *Stowaway* has extremely limited homology in its TIRs with its autonomous parent *mariner*
Stowaway has no central homology with *mariner*

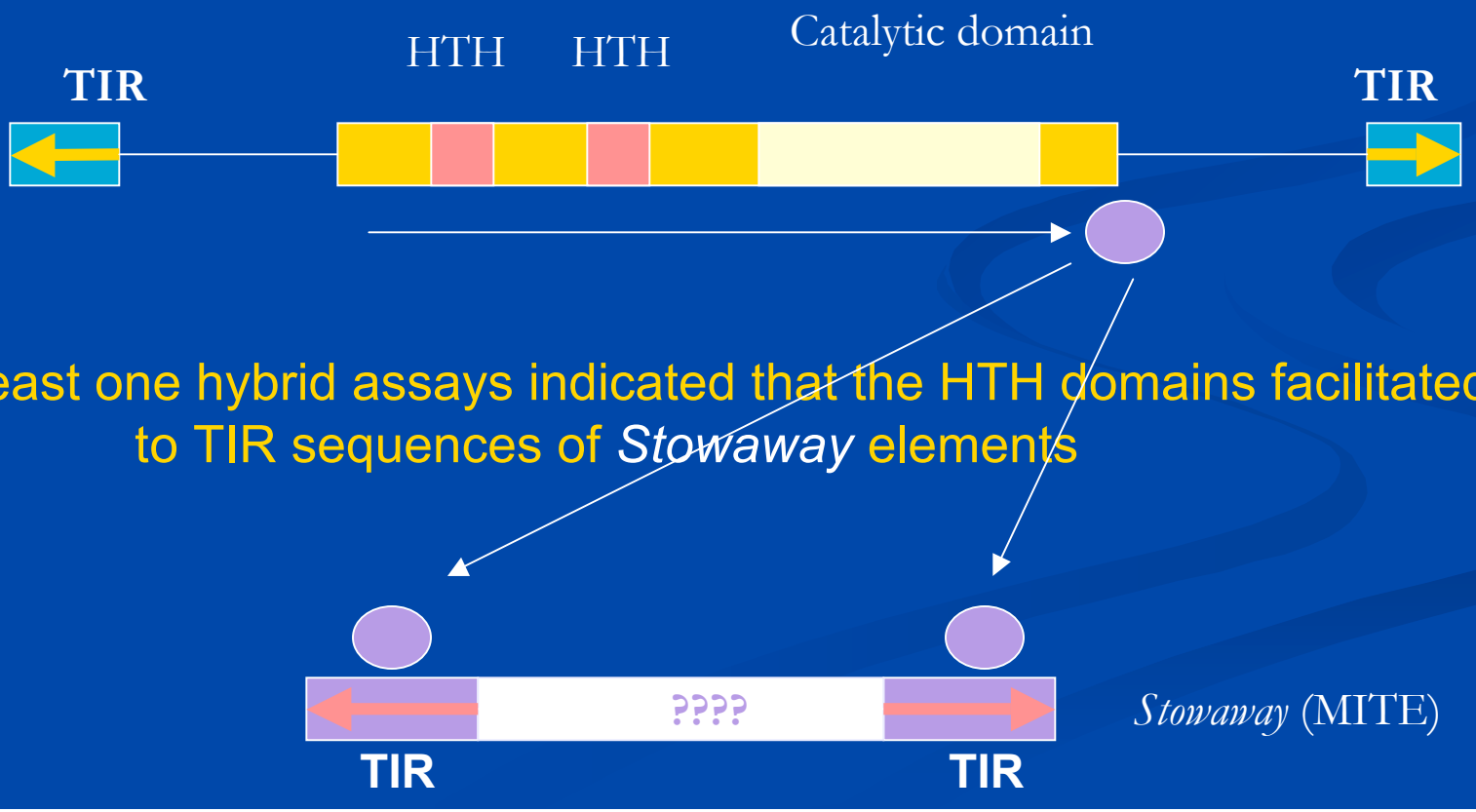
Copy number: 34 copies of *mariner* and 22,000 copies of *Stowaway*

DNA Transposons

***OsMar5* (Mariner family of transposable elements)**

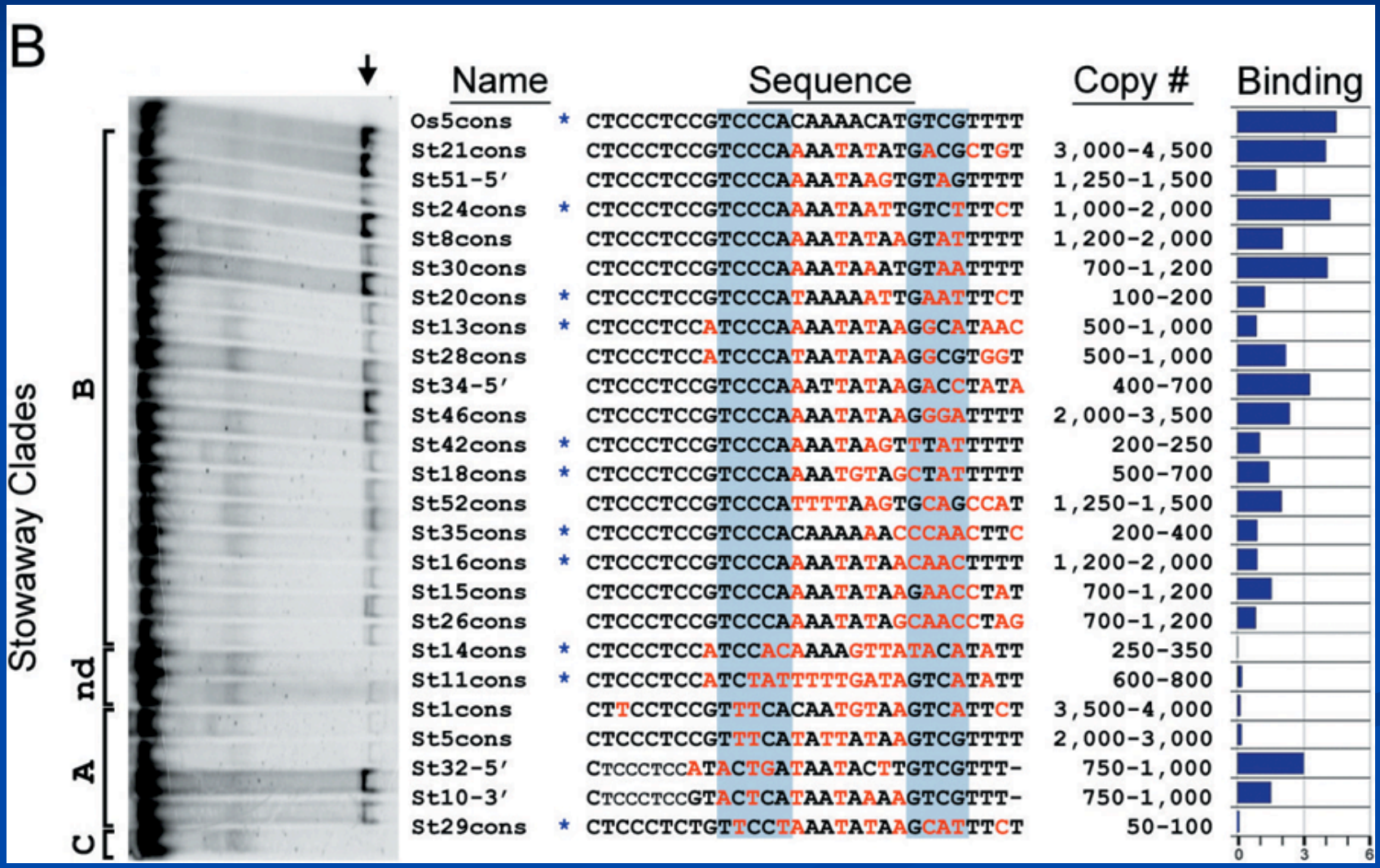
- HTH- Helix-turn-helix domain involved in DNA binding

- Catalytic domain is responsible for transposition



Yeast one hybrid assays indicated that the HTH domains facilitated binding to TIR sequences of *Stowaway* elements

DNA Transposons



Feschotte et al., 2005

Pack-MULEs

Pack MULEs are an interesting twist where gene amplification, exon swapping and transposons meet

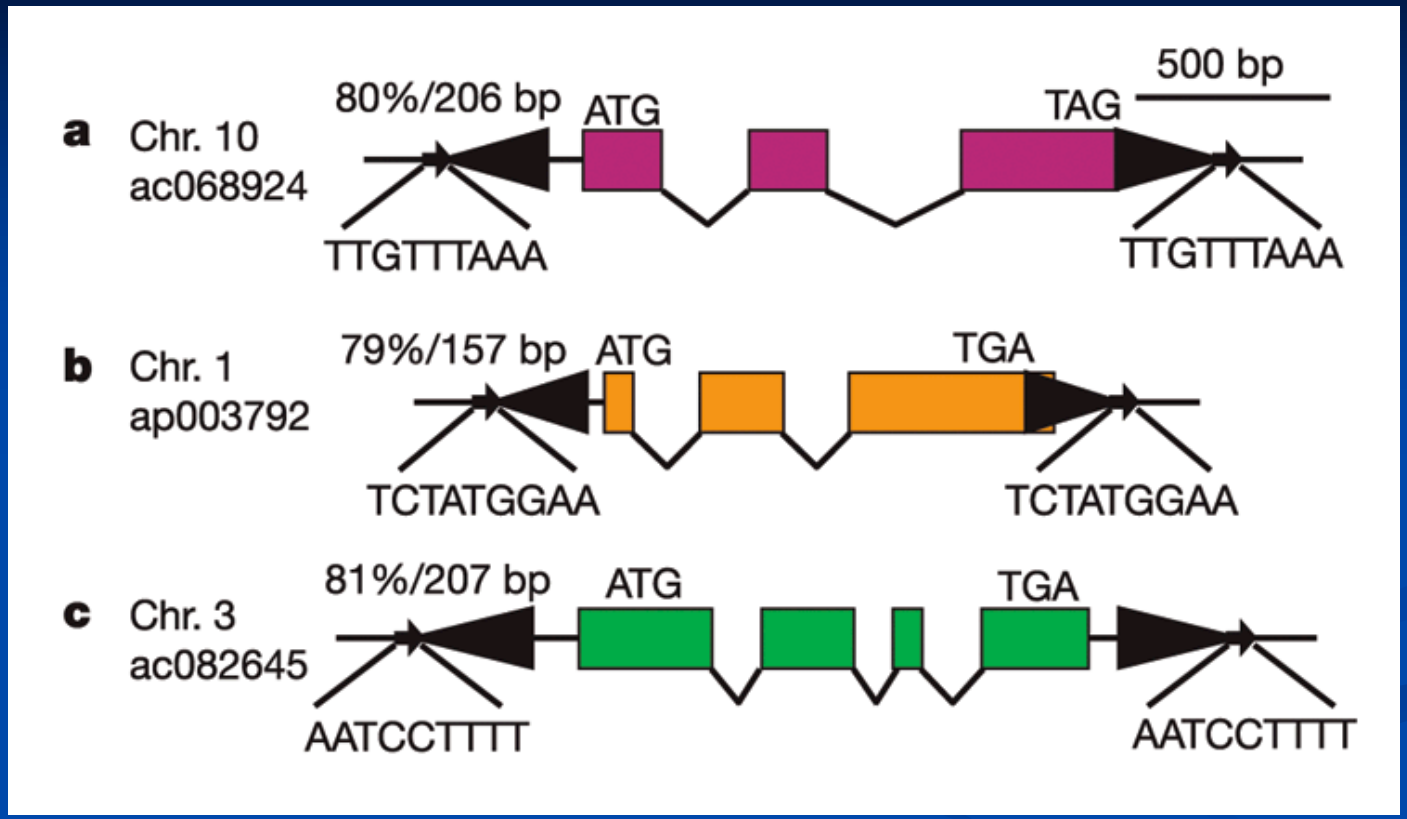
MULEs are *Mutator*-like elements

Mutator (*Mu*) is an element that was originally identified in maize
- Maize lines were grown in radioactive conditions and *Mu* became active

Mu –like elements have been identified in other grass species

Mu is a bit different than other DNA transposons, it has a long tandem site duplication (8-10bp) and has very long TIRs (hundreds of bp)

Pack-MULEs



Pack MULEs are *Mu*-like elements in rice that have captured genes/exons between the TIRs

Note in the figure above the TSDs are the small arrows, the TIRs are the larger arrows and the contained gene is shown in color (with ATG and TGA shown)

These capture genes can be mobilized by the *Mutator* element AND they can amplify their copy number

Jiang et al., 2004

Pack-MULEs

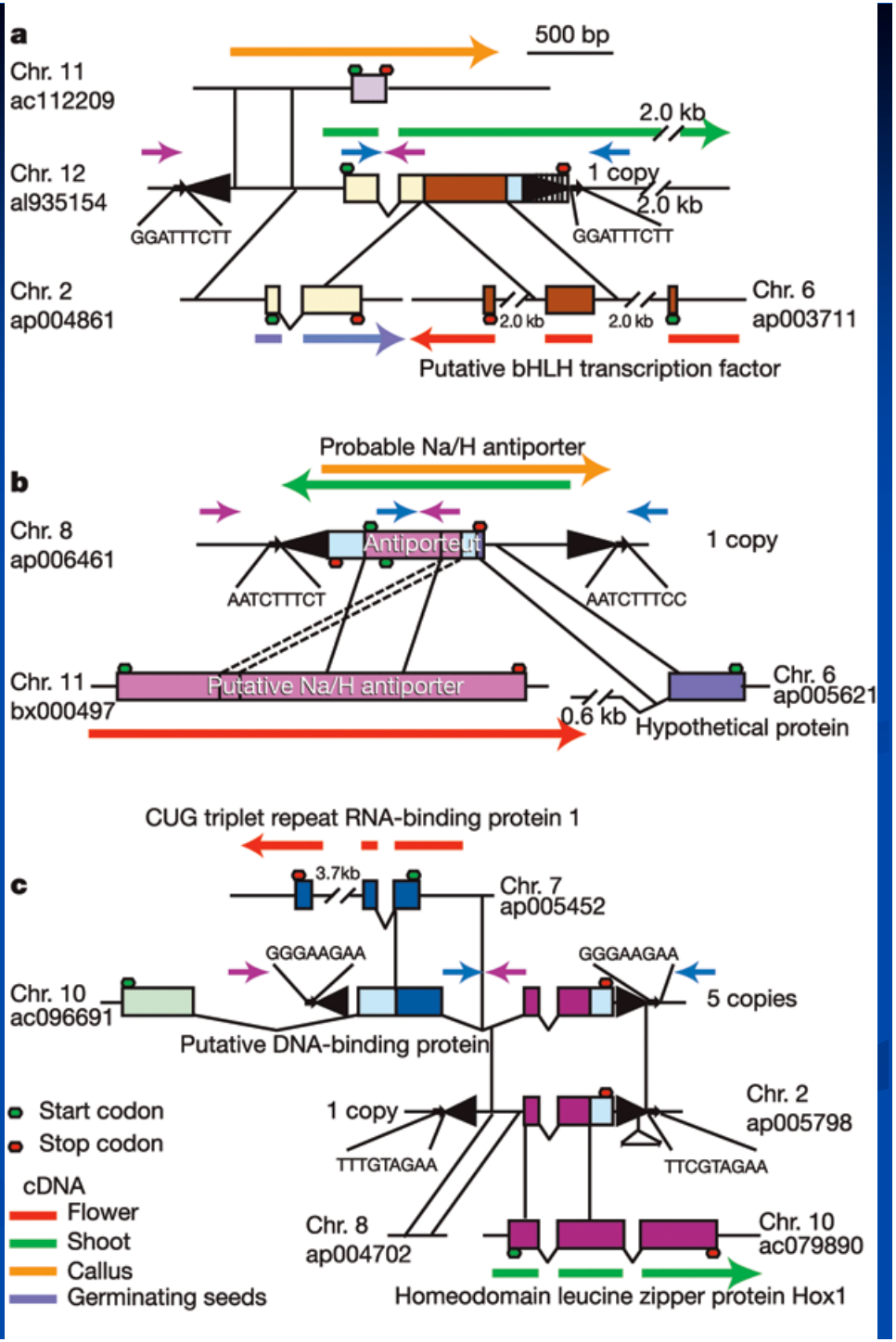
PACK-Mules can also contain more than one gene

In fact, the Pack-MULEs can merge together exons from genes that are genetically unlinked

The figure to the right is busy but shows how the origins of the genes/exons Pack-MULEs

This offers an interesting mechanism whereby novel gene combinations can be generated by *Mu* elements and amplified

Jiang et al., 2004



How to Identify Repeats?

- Sequence similarity search using preexisting databases of known repeat sequences
- Algorithms locating repeats exclusively relying on sequence composition

Programs for Repeat Searches

- **CENSOR** (Jurka et al., 1996)
  early program, slow
- **RepeatMasker** (Smit et al., 1996)
  most popular, sensitive, good functionality, uses cross_match, slow
- **MaskerAid** (Bedell et al., 2000)
  uses WU-BLAST, an enhancement of RepeatMasker in speed (~ 30 times), not as sensitive as RepeatMasker
- **BLAST, flast** …
  basically any similarity search program can identify repeats using a library

Major drawback of similarity searches:

requires a repeat library (e.g. Repbase), which is available only for the well-studied organisms.

Programs for *de-novo* Repeat Identification

- Miropeats (printrepeats, Parsons, 1995)

  uses ICAass, graphically display repeats, can only handle several hundred thousand bp

- REPuter and REPfind (Kurtz et al., 2001)

  first applied suffix trees in repeat mining. REPfind is a newer version that can identify degenerate repeats. Applies statistical significance

- RepeatFinder (Volfovsky et al., 2001)

  merges repeats where a merged repeat exists elsewhere in the genome at lease once. Boundaries not well defined. Group members may not share similarity at all

Programs for *de-novo* Repeat Identification, cont'd

- RECON (Bao and Eddy, 2002)

  WU-BLAST for pair-wise alignment, multiple alignment used to define boundaries of repeat elements. Boundaries of repeat families not available.

- PILER (Edgar and Myers, 2005)a suite of tools. uses its own PALS for pair-wise alignment

  PILER-DF: to detect Dispersed Families of transposable elements

  PILER-PS: to detect Pseudo-Satellites – repeats clustered locally

  PILER-TA: to detect Tandem Arrays

  PILER-TR: to detect repeat families of members with Terminal Repeats

- RepeatScout (Price and Pevzner, 2005) no pair-wise alignment needed.

  Genome is first scanned for "word" of fixed length. Starting from the most frequently found word,

  RepeatScout will extend the word in both directions, terminating at the most appropriate points (determined by score) for boundaries.

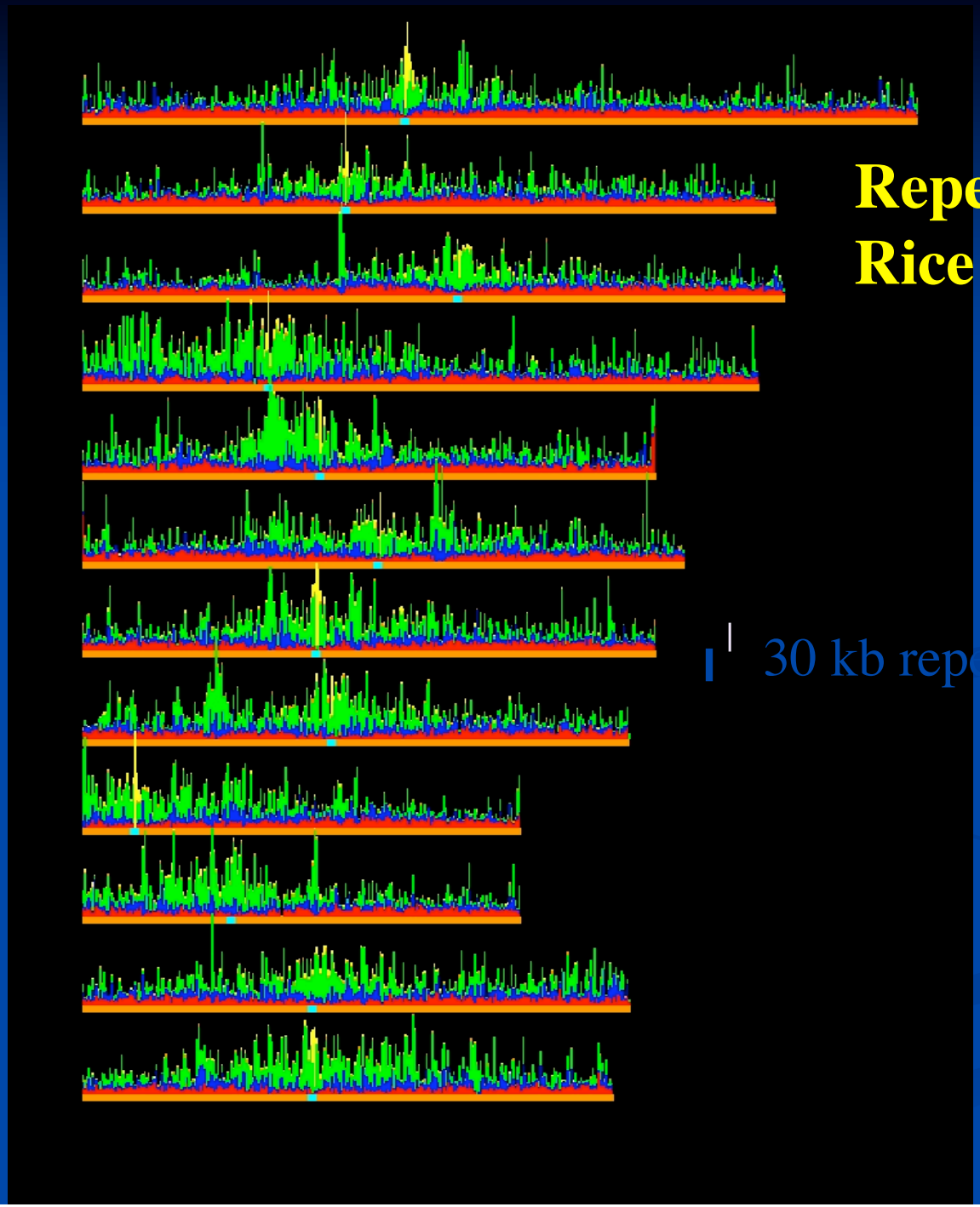
  Consensus sequence for families is generated.

Major drawback of these programs: large gene families will be included as "repeats".

Construction of TIGR Plant Repeat Database -- Methods

- Collecting repetitive sequences from public database: GenBank, TREP, individual projects, etc
- Evaluate the sequences, remove erroneous entries
- Classification and coding
  Repeat database for a family (e.g. TIGR Gramineae Repeat Database)
- Search the family repeats against available genomic sequences of a genus. Matches are extracted and coded, and then combined with repeats obtained previously from public databases, to create the TIGR Repeat Database for that genus.
- The TIGR Plant Repeat Databases (Nucleic Acids Res. 2004 Jan)

Repeat Distribution of Rice Genome



Citations

McKnight and Shippen (2004) Plant Telomere Biology. Plant Cell 16:794.

Riha, McKnight, Griffing, and Shippen (2001) Living with Genome Instability: Plant repsonses to Telomere Dysfunction. Science 291:1797.

Henikoff, Ahmad, and Malik (2001) The centromere paradox:Stable inheritance with rapidly evolving DNA. Science 293:1098.

Wu et al (2004) Composition and Structure of the Centromeric Rgion of Rice Chromosome 8. Plant Cell. 16:967.

Zhang et al (2004) Structural Features of the Rice Chromosome 4 centromere. Nucleic Acids Research. 32:2023.

Kumar and Bennetzen. (1999) Plant Retrotransposons. Annual Review of Genetics 33:479.

Feschotte, Jiang, and Buell (2002) Plant Transposable Elements: Where Genetics meets Genomics. Nature Genetics Reviews. 3:329.

Citations

Mizuno et al. (2006) Sequencing and characterization of telomere and subtelomere regions on rice chromosomes 1S, 2S, 2L, 6L, 7S, 7L, and 8S. Plant Journal. 46:206-217.

Dawe, K. (2005) Centromere renewal and replacement in the plant kingdom. PNAS. 102(33):11573-4.

Lee et al. (2005) Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in Oryza species. PNAS. 102(33):11793-8.

Jiang et al. (2004) Pack-MULE transposable elements mediate gene evolution in plants. Nature 431:569

Hirochika et al. (1996) Retrotransposons of rice involved in mutations induced by tissue culture. PNAS. 93:7783-7788.

Feschotte, et al. (2005) DNA-binding specificity of rice mariner-like transposases and interactions with Stowaway MITEs. NAR. 33:2153.