Construction of custom repeat libraries for genome annotation

Ning Jiang
Dept. of Horticulture
Michigan State University
March 7, 2018

Outline

- Classification of transposable elements (TEs)
- Abundance and insertion preference
- The relationship between TEs and genes
- TE detection methods
- Construction of repeat library

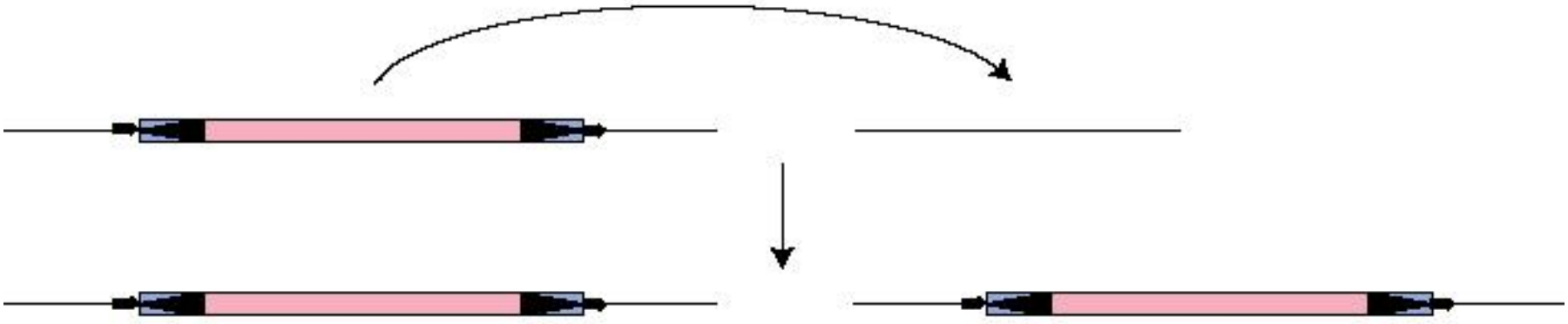
What are in the genome

- Structural component, centromere and telomere.
- Genes – protein genes (coding gene) and non-coding RNA genes
- Intergenic sequences

Genomes contain both unique and repetitive sequences (ATGC)

- Unique sequences
  - genes and regulatory regions
- Repetitive sequences
  - gene families
  - tandem repeats, centromeric repeats, telomeric repeats
  - transposable elements (TEs)

Genomes contain large amount of transposable elements (TEs) “Jumping genes”



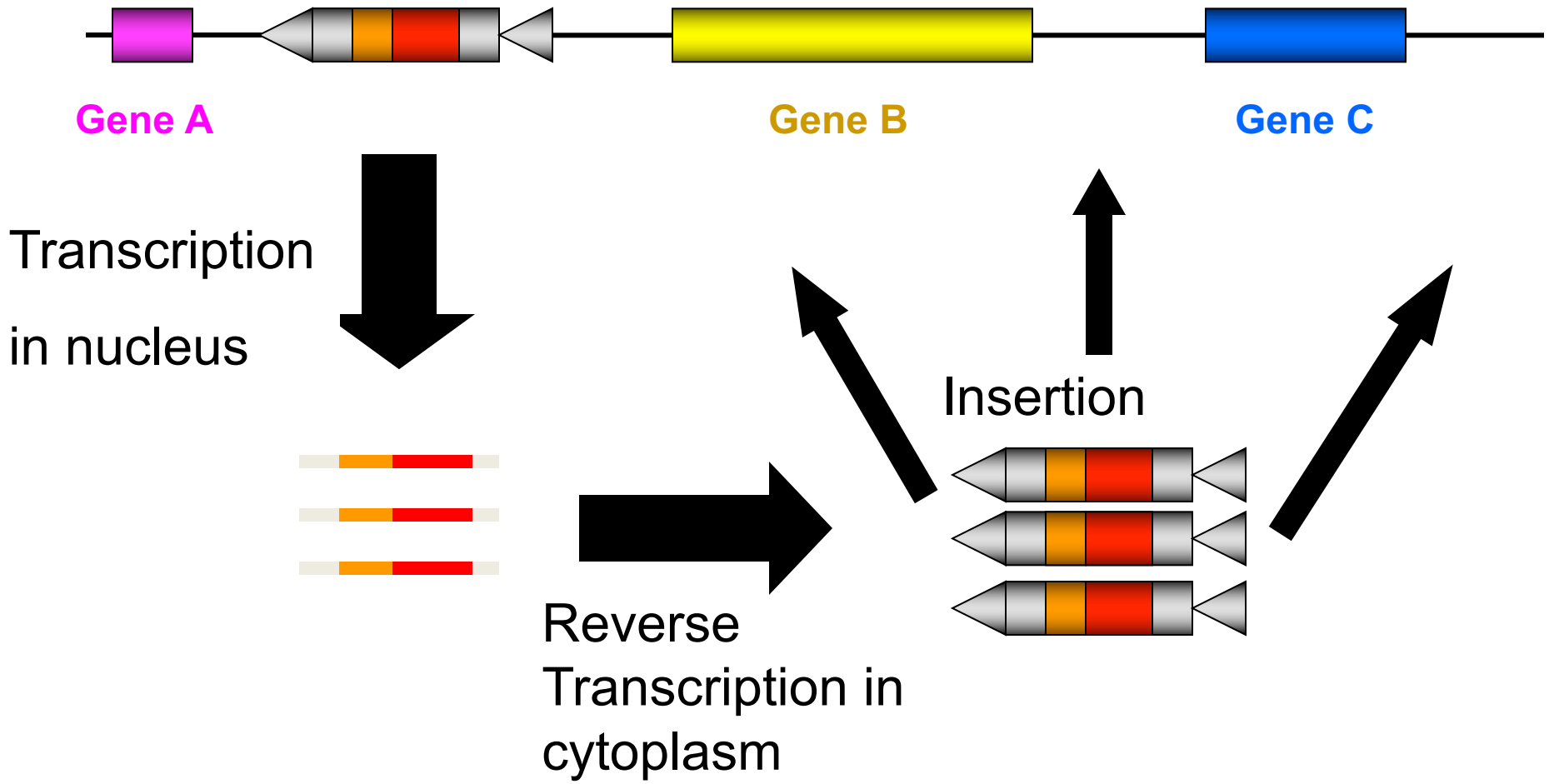
Barbara McClintock

Classification of plant transposable elements (class I)

- Class I (retrotransposon, RNA transposon, copy and paste mechanism)
  - LTR (Long Terminal Repeat)elements
    - Copia like
    - Gypsy like
    - Endogenous retrovirus
  - Non LTR elements
    - LINE
    - SINE

Long Terminal Repeat (LTR) elements

Transposition of LTR elements “Copy and paste”



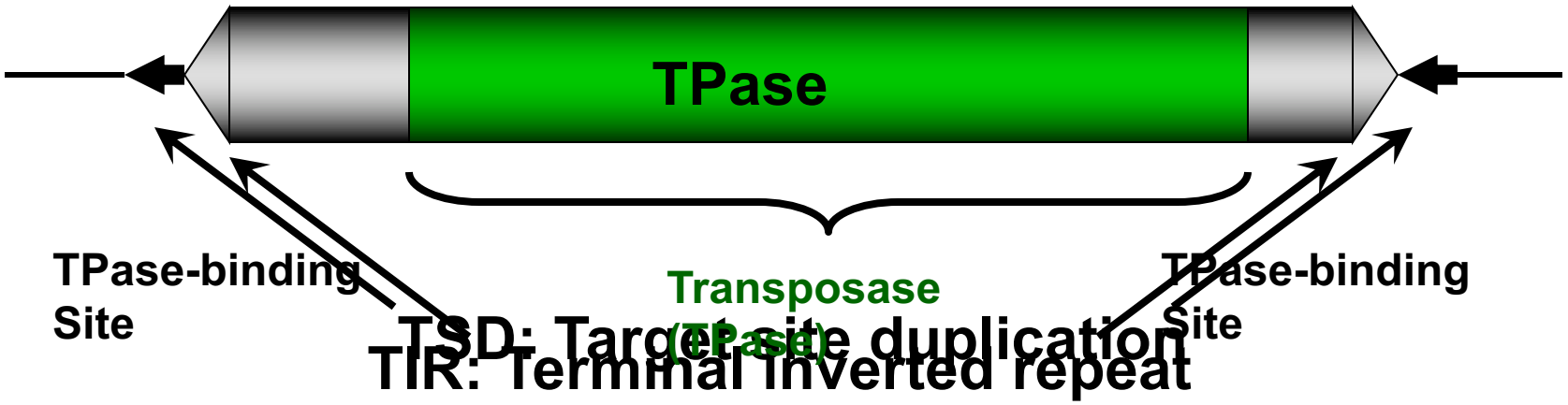
Classification of plant transposable elements (class II)

- Class II (DNA transposon)
  - Subclass I (cut and paste mechanism)
    - Ac/Ds (hAT), En/Spm (CACTA), Mutator (MULEs), PIF/Harbinger, TC1/Mariner
  - Subclass II (replicative mechanism)
    - Helitron

Wicker et al. 2007. Nature Reviews Genetics

Class II elements

Terminal Inverted repeat (TIR) elements

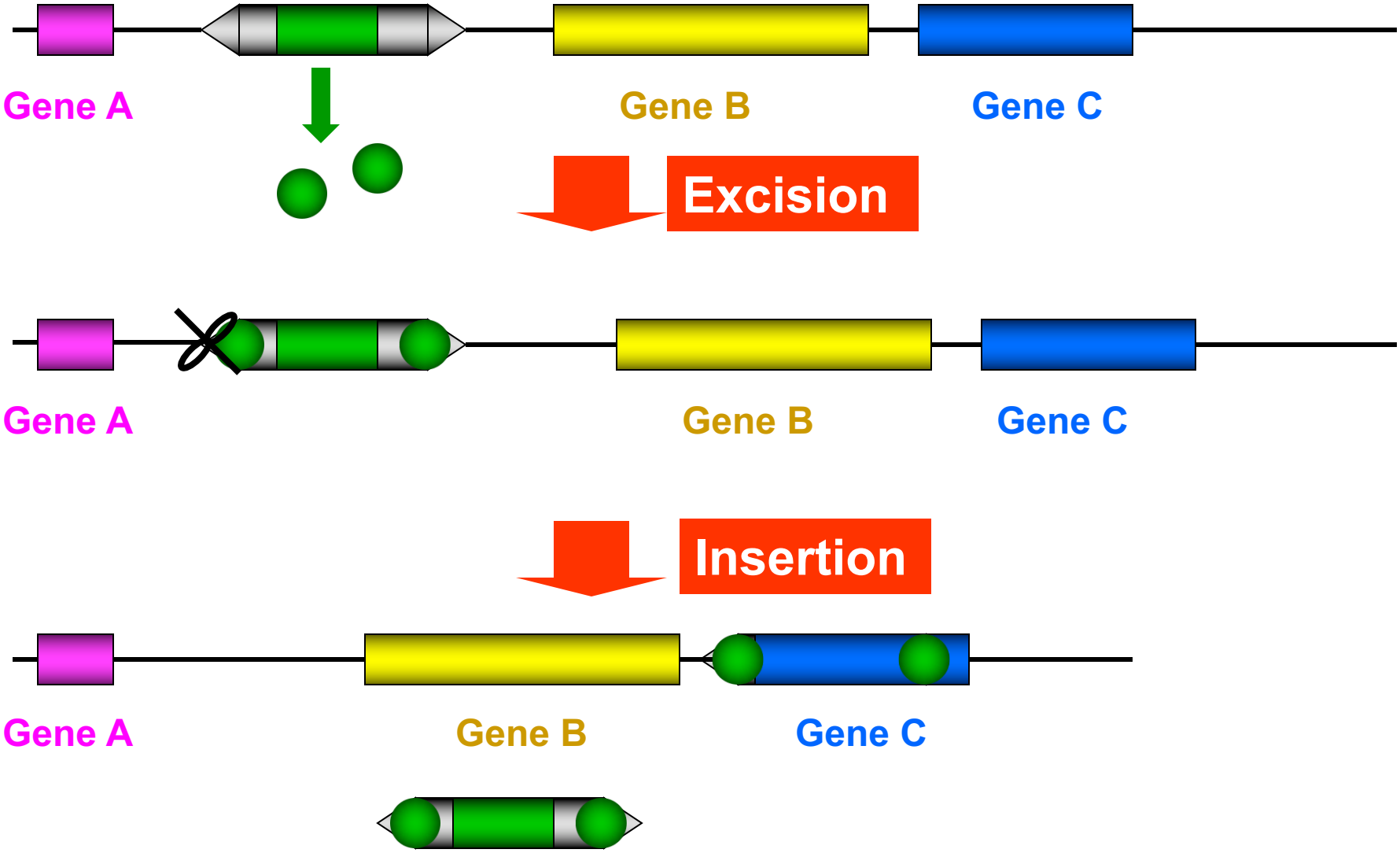


Hierarchy of TE classification

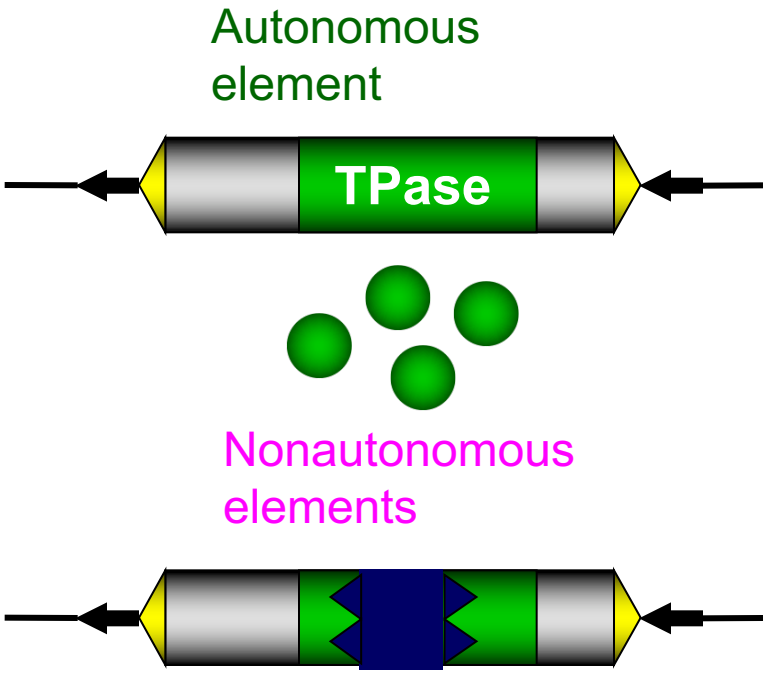
- Class – Class II or DNA transposon
  - Subclass 1 – “cut and paste”
    - Order – TIR (terminal inverted repeat)
      - Superfamily - “Mutator-like element” (MULE)
        - » Family – Mutator
          - Subfamily – Mu1, Mu2, Mu3
            - Individual elements

The definition of family or subfamily is more or less arbitrary. Wick et al. proposed that if two elements share 80% identity in 80% of the element sequence, they belong to the same family.

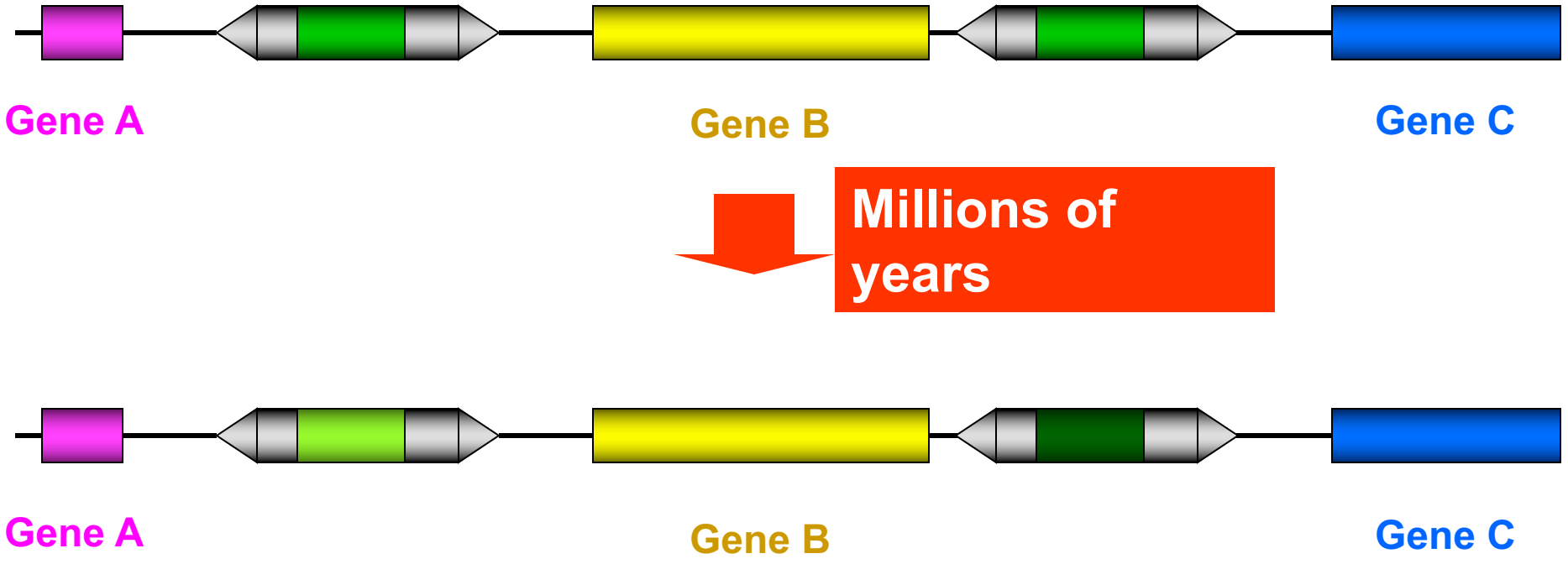
Transposition of DNA elements “cut and paste”



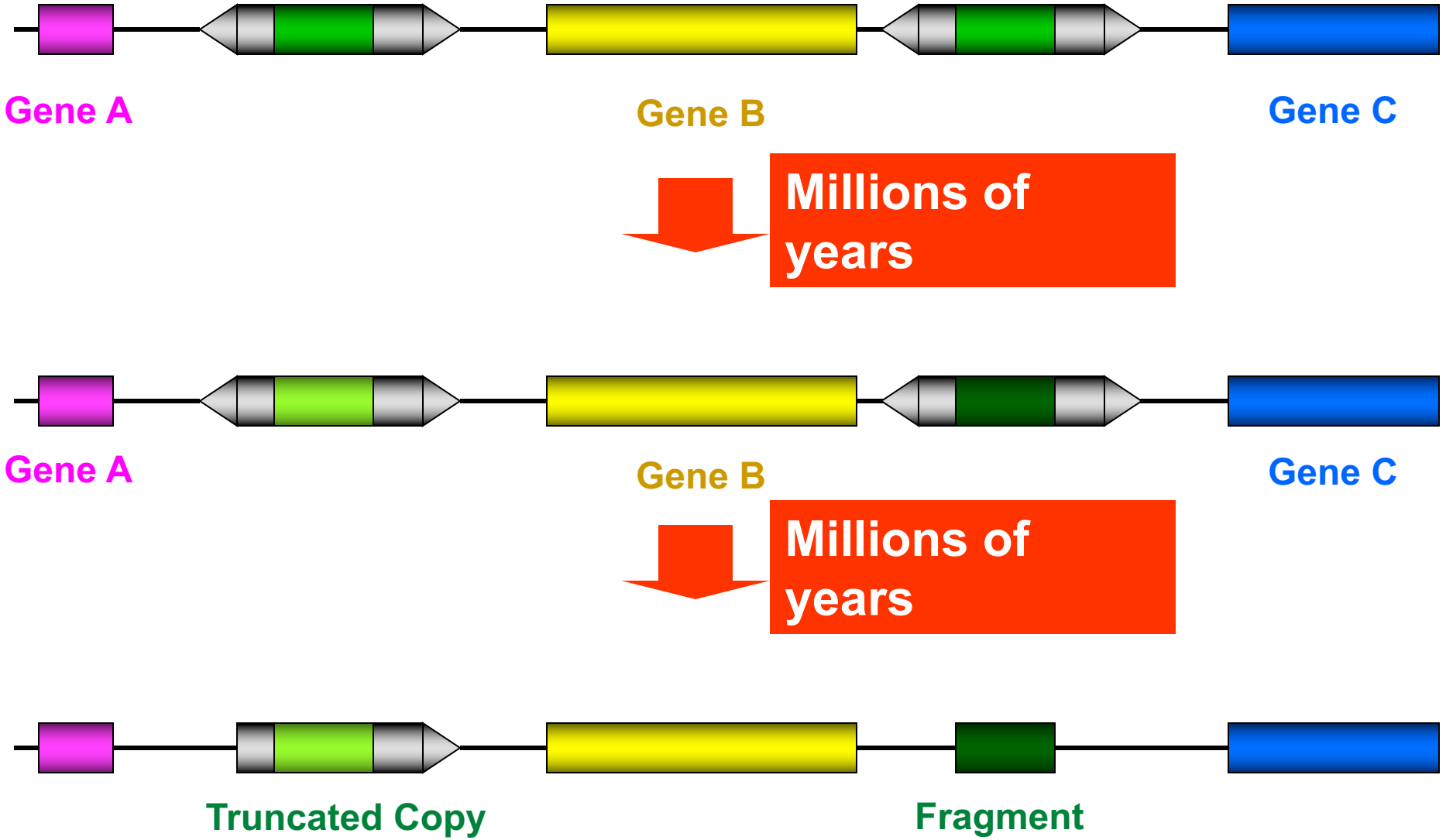
Autonomous and non-autonomous elements (DNA transposons)



Newly duplicated elements are identical



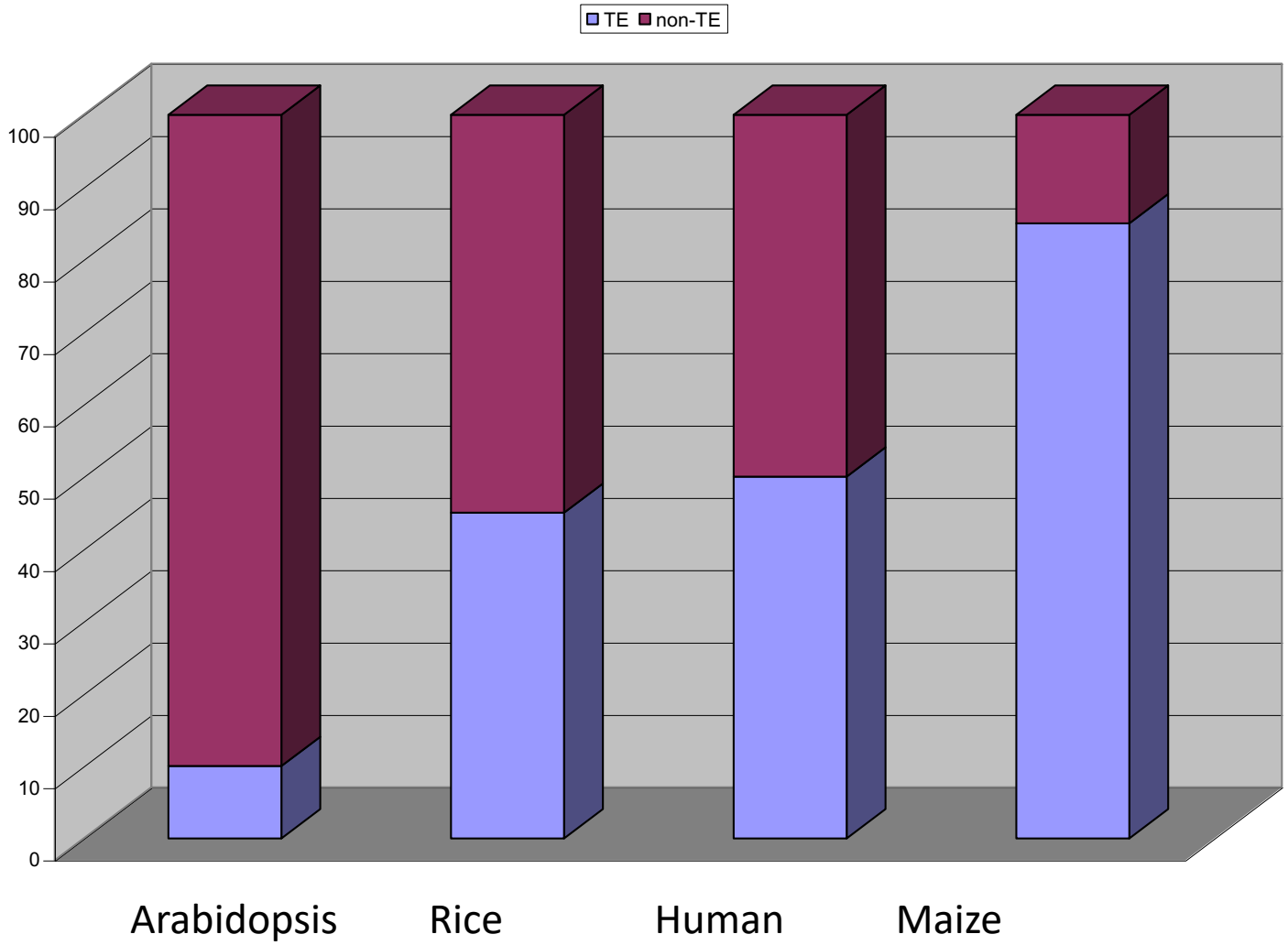
Transposons are only recognizable for a few million years



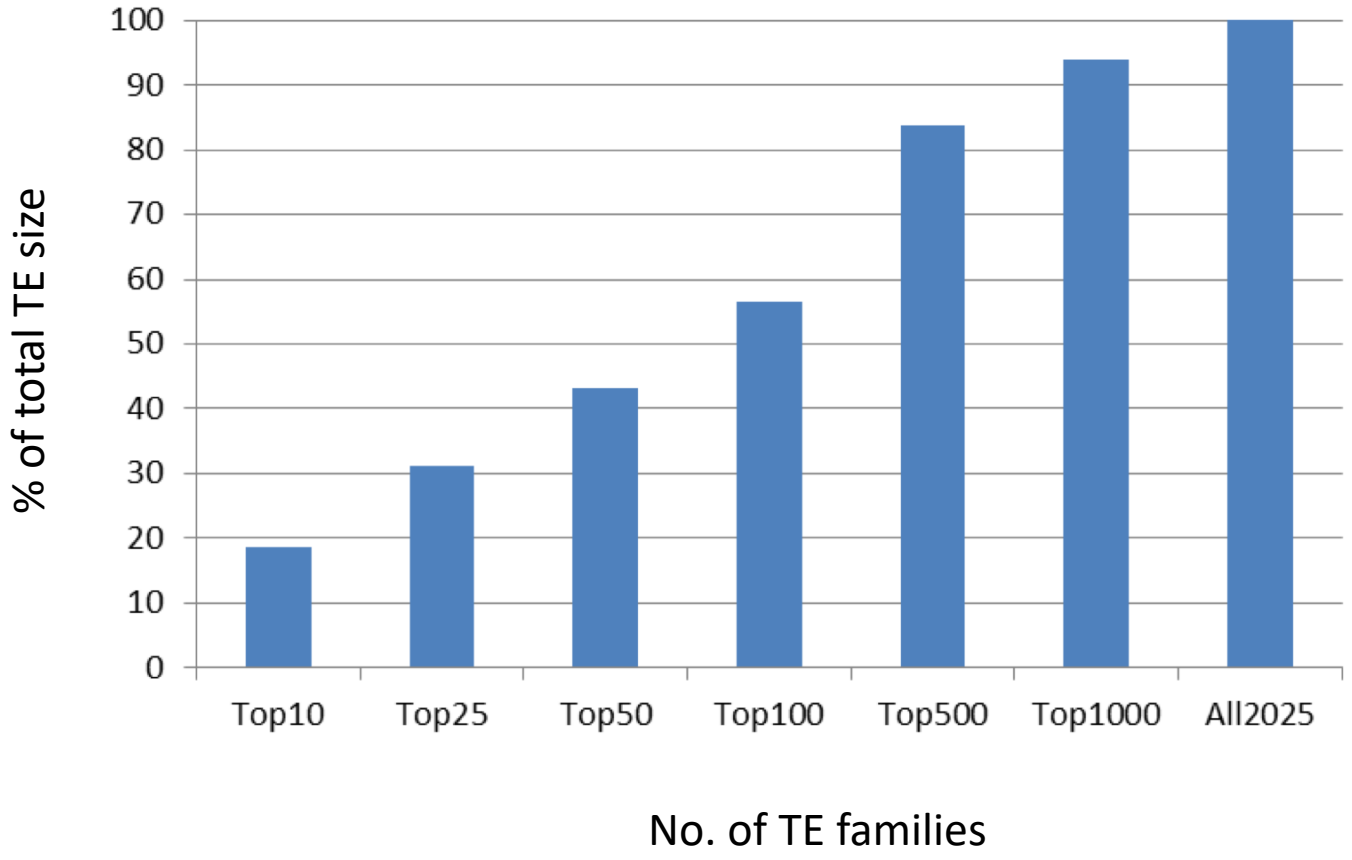
Outline

- Classification of transposable elements (TEs)
- Abundance and insertion preference
- The relationship between TEs and genes
- TE detection methods
- Construction of repeat library

Transposable elements are major components of eukaryotic genomes



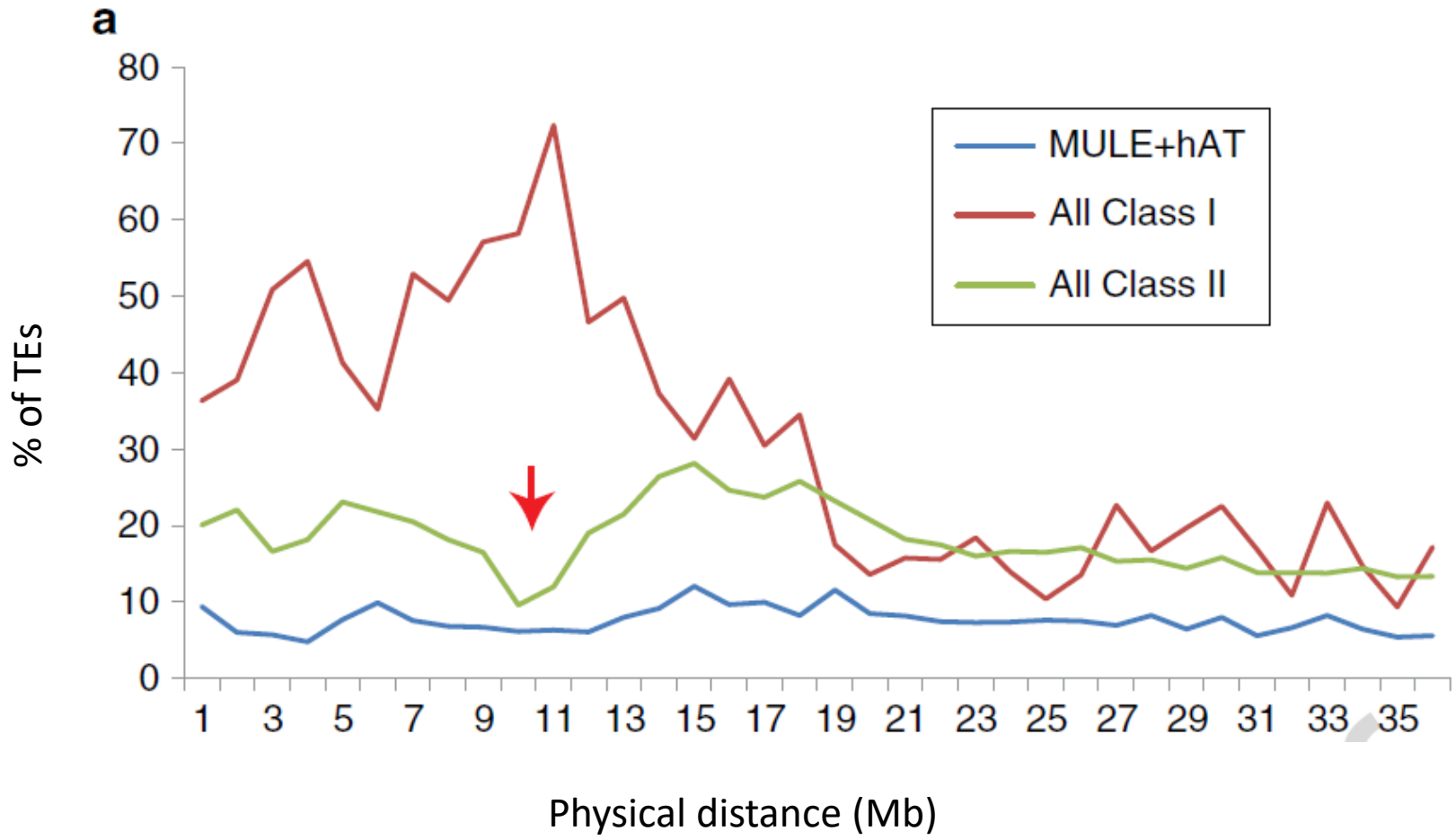
Few most abundant TE families contribute to a large portion of TE size



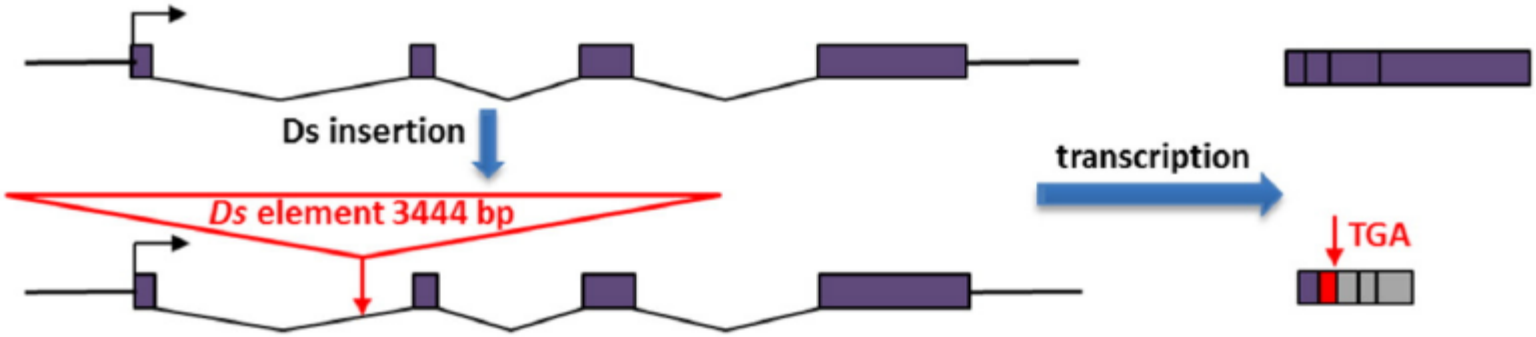
Should we care about the less abundant elements at all?

- Most active TEs are low copy number elements
- It depends on your research purpose

Different transposons have different niches in the genome



DNA transposons are frequently associated with genes



Chen et al. Plant Mol. Biol. 2012

Transposons nested with other transposons

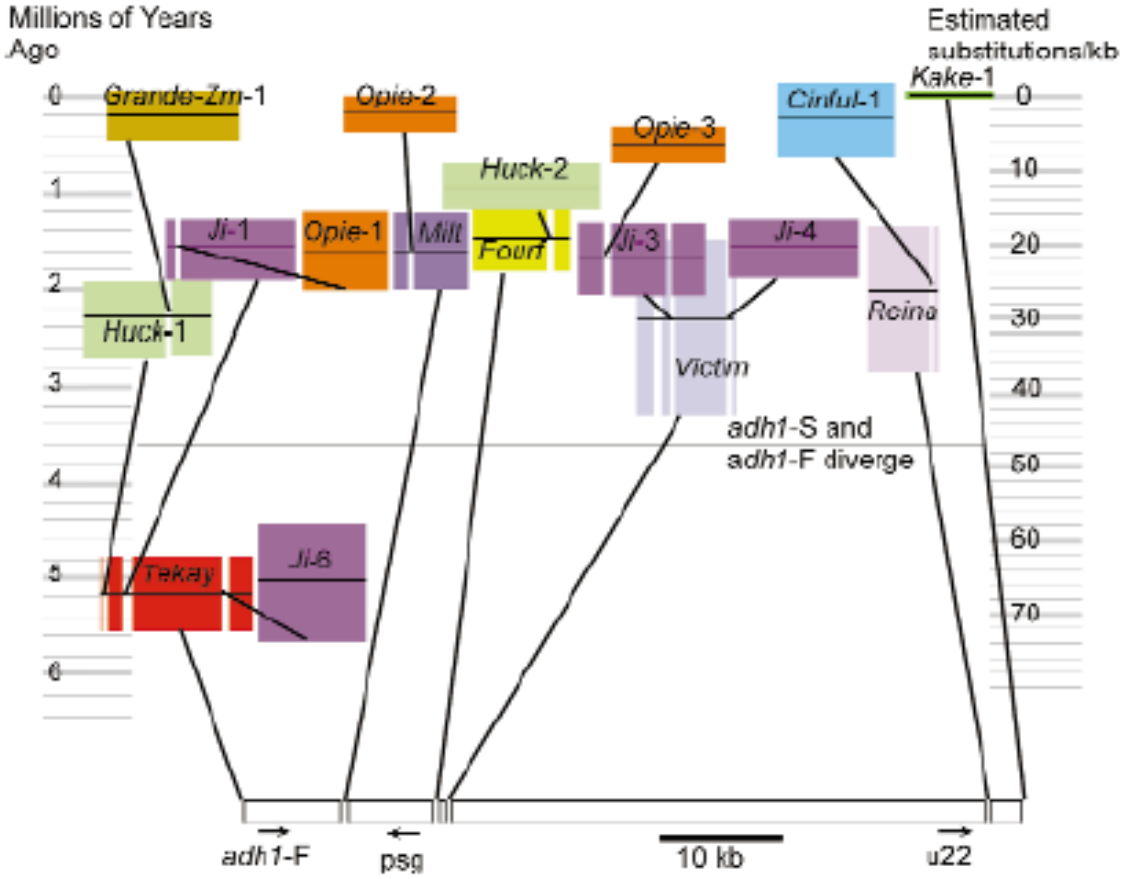


Fig. 2 Temporal arrangement of maize *adh1*-F region retrotransposons. Coloured boxes, retrotransposons; breaks in boxes, insertion sites; horizontal lines through box center, estimated insertion date of a retrotransposon; box height, standard deviation above and below the estimate.

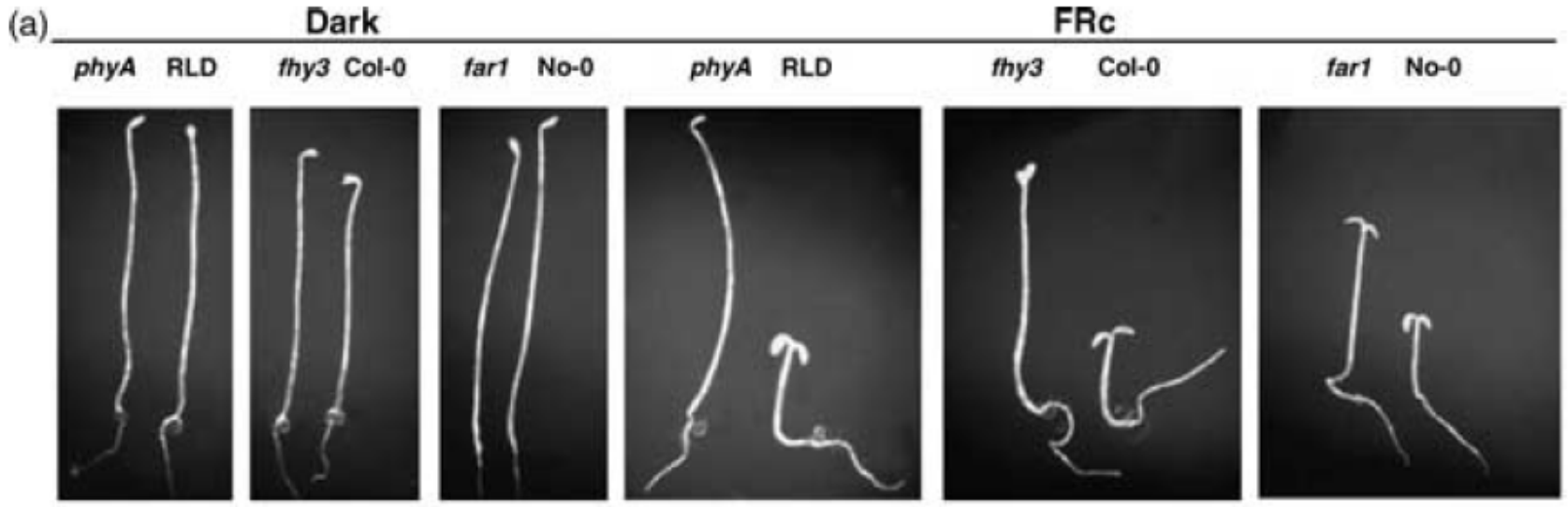
SanMiguel et al. Nature Genetics 1998

Outline

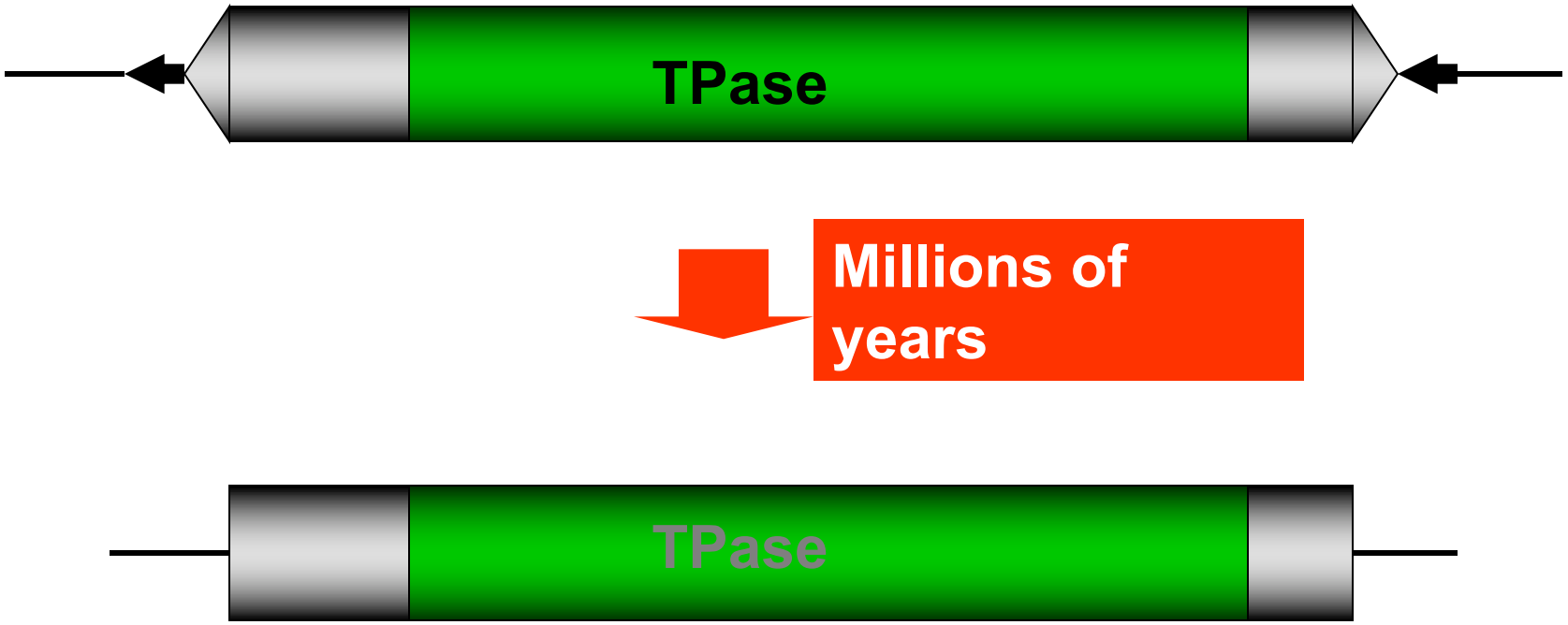
- Classification of transposable elements (TEs)
- Abundance and insertion preference
- The relationship between TEs and genes
- TE detection methods
- Construction of repeat library

Two genes in phytochrome pathway are derived from MULE transposons

- The FHY3 and FAR1 genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway Matthew et al. The Plant Journal (2003)

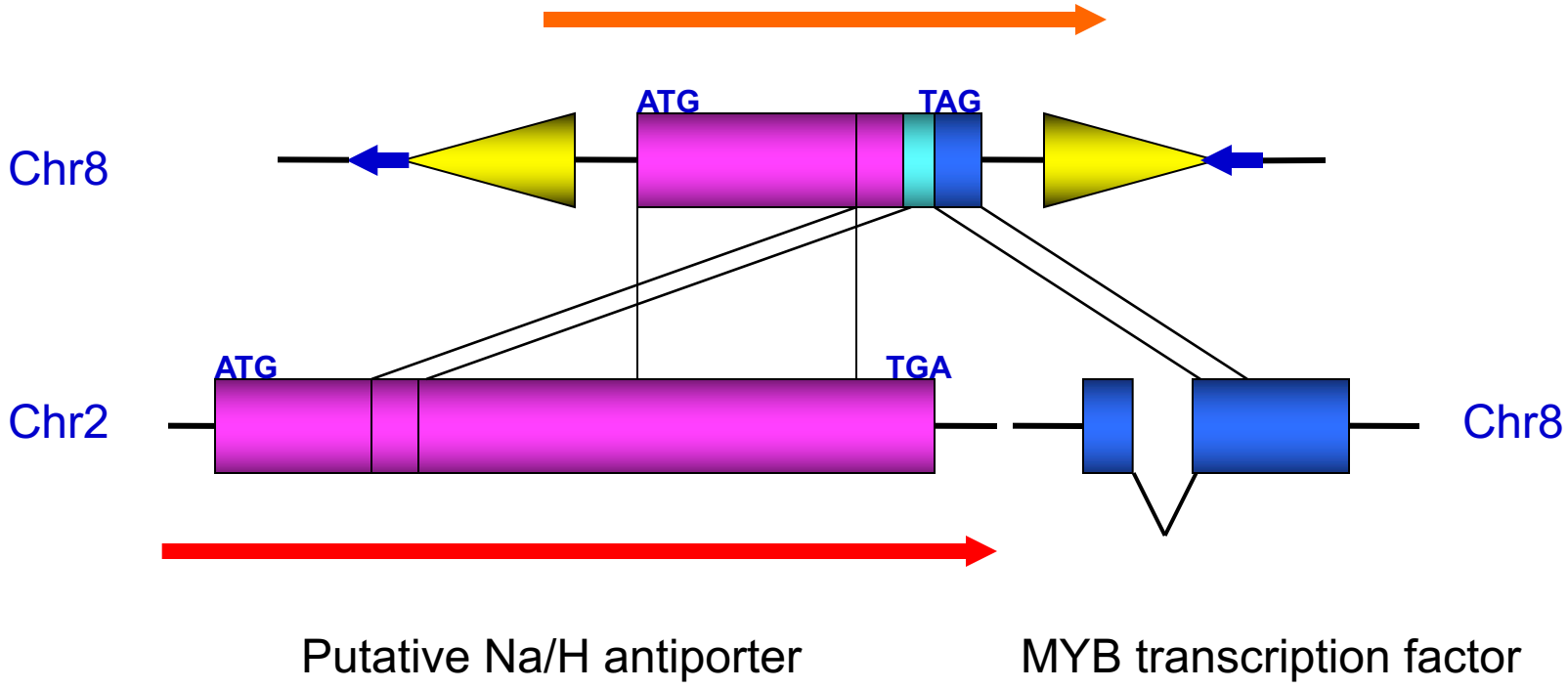


Domestication of transposons – autonomous transposons become normal genes



Transposases are DNA binding proteins!

Transposons can duplicate and recombine gene sequences



Outline

- Classification of transposable elements (TEs)
- Abundance and insertion preference
- The relationship between TEs and genes
- TE detection methods
- Construction of repeat library

Why do we remove repeats prior to annotation of genes?

- Reduce the use of computational power, particularly for large genomes
- Minimize the interference of TEs Improve the accuracy the gene prediction
- Construct a repeat library and mask them out

Need for custom repeat libraries

- Most TE sequences are not conserved at nucleotide level except among closely related species (divergence for a few million years)
- Custom libraries are usually small in size
- Sensitivity vs. specificity
- Specificity is more important then sensitivity in this case

TE detection methods

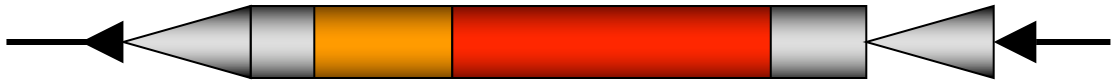
- Homology based
  - Homology to known TEs, such as RepeatMasker
  - De novo methods
- Structure based
  - Using structural features of TEs for identification. Methods developed for MITEs, LTR elements, SINEs, Helitron, etc.

De novo identification methods

- No requirement for knowledge about the genome
- Any repetitive sequences will be recovered, good or bad
- Cannot identify low copy number TEs
- Most de novo methods do not classify elements

Structure based methods

- Terminal repeat, terminal sequence, target site duplication can all be used as structural features for research
- Identify both high copy and low copy number TEs
- Cannot identify old, degenerated copies
- Cannot identify novel elements



Outline

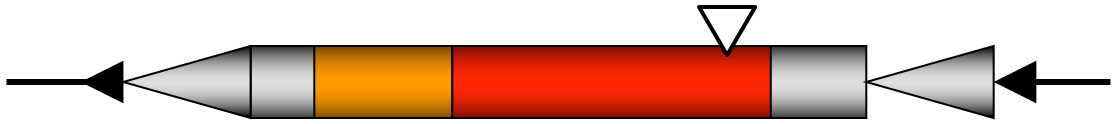
- Classification of transposable elements (TEs)
- Abundance and insertion preference
- The relationship between TEs and genes
- TE detection methods
- Construction of repeat library

Repeat Library Construction-Advanced

- http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-Advanced
- Optimized for plant genomes, but also applicable for other genomes
- Will have another update later 2018

MITEs (miniature inverted repeat transposable elements)

- MITEs (< 600 bp) are numerically most abundant TEs in plant genomes
- Frequently associated with genes
- Identify small TEs first to minimize misclassification of elements



MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences

Yujun Han and Susan R. Wessler*

Department of Plant Biology, University of Georgia, Athens, GA 30602, USA

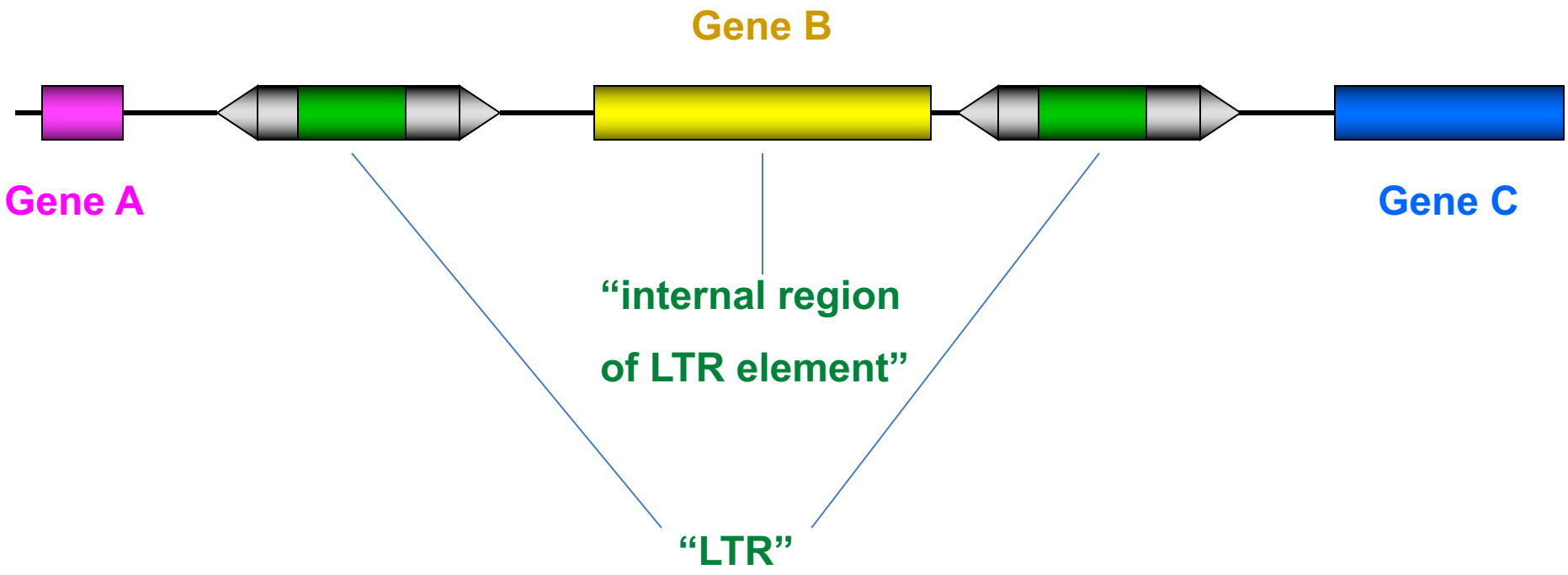
Received July 15, 2010; Revised September 8, 2010; Accepted September 13, 2010

- Low false positive rate
- Reasonable computation time
- For large genomes (> 500 Mb), use partial genomic sequences

LTR retrotransposons

- Largest component of plant genomes
- Most elements are large in size but there are small elements called terminal-repeat retrotransposon in miniature (TRIM)
- Many programs developed with high sensitivity, but false positives have been an issue

False positive LTR elements could be very toxic



Breakthrough Technologies

Check for updates

LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons[1][OPEN]

Shujun Ou and Ning Jiang[2]

Department of Horticulture, Michigan State University, East Lansing, Michigan 48824

ORCID IDs: 0000-0001-5938-7180 (S.O.); 0000-0002-2776-6669 (N.J.)

- Using output from LTRharvest, MGEScan-LTR and LTR_finder to maximize sensitivity
- Filtering false positives to improve specificity
- Identifying elements with non-canonical terminal sequences (seven types)

LTR_retriever

- Provides gff for all intact LTR elements in the genome
- Estimate insertion time of intact elements
- Building non-redundant libraries of LTR elements to reduce downstream requirement for computation
- Applicable to corrected long-reads
- Multithreading, good for big genomes

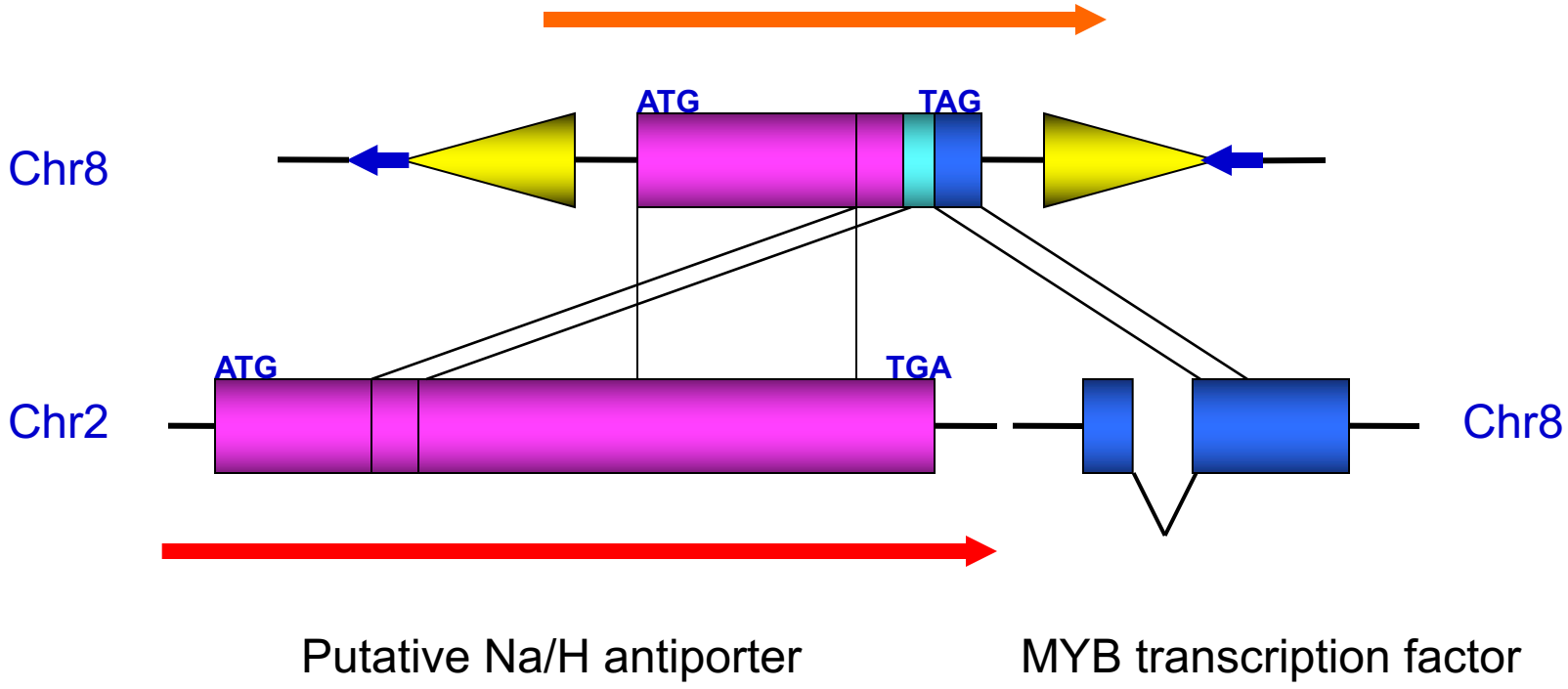
Identification of the remainder of repeats - RepeatModeler

- Combining output from two de novo programs: RECON and RepeatScout
- Provides classification for some repeats, but not all classifications are correct
- Single-threading, slow, run it after masked with outputs (libraries) from MITE-hunter and LTR_retriever
- For large genomes, proceed step-wisely

Step-wise implementation of RepeatModeler

- For a large genome, use a small portion of genomic sequence first
- Use the output to mask a larger portion of the genome, then run RepeatModeler on the masked sequences, or exclude the masked sequence to reduce the physical size of sequences
- Repeat this process on the remainder of the sequences

Real transposons (not false positives) could contain gene sequences



Excluding gene sequences from repeat libraries

- Blast against a plant protein database
- Using ProtExcluder to remove the gene sequences
- The default is to remove the matched portion as well as 50 bp flanking sequences but it can be customized

Final repeat libraries

- MITEs, LTR elements, classified TEs from RepeatModeler, most likely true TEs
- Unknown repeats from RepeatModeler, most of them are ancient TEs but could contain non-TE sequences or even novel gene families, so use it with caution

Thanks to

- National Science Foundation

- Michigan State University

- All users, especially those who provided feedbacks