Transposable element (TE) annotation

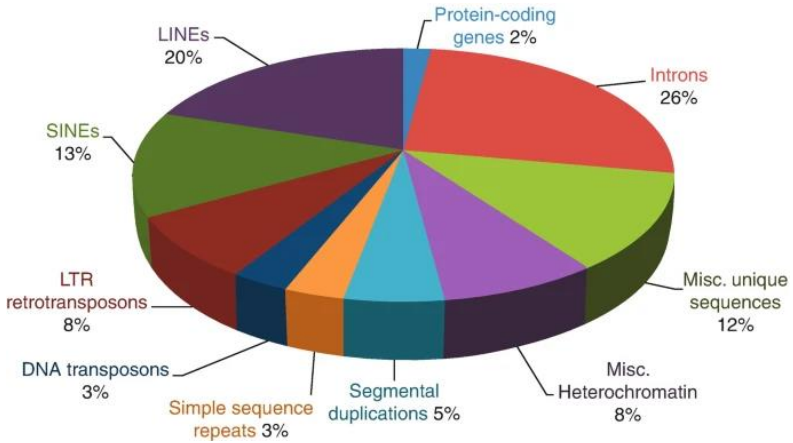
2024 Bioplatforms Australia (BPA) workshop

Schwessinger Lab

Zhenyan Luo

Australian National University

Repeats are a major component of eukaryotic genomes



Composition of the human genome

Repetitive DNA sequences (repeats) are patterns of DNA sequences that occur in multiple copies throughout the genome

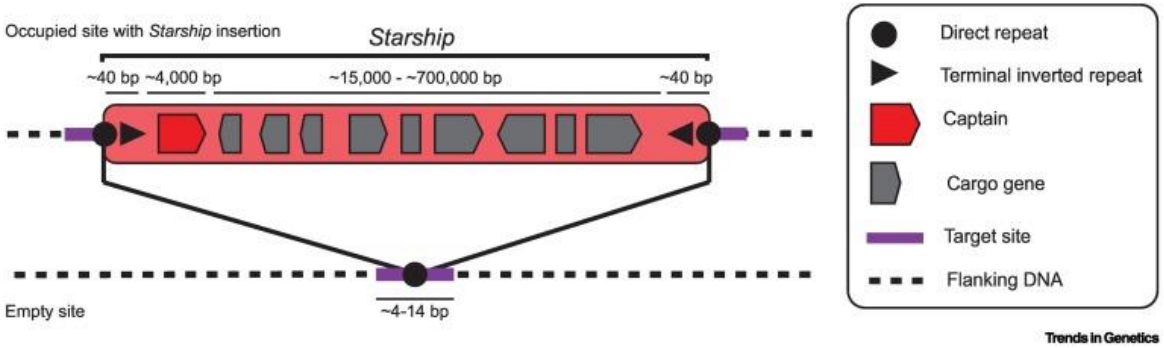
About 50% of the human genome consists of repeats

Roughly 4% of human genes harbor transposable elements in their protein-coding regions

Pena 2021; Bernabe et al. 2023; Nekrutenko & Li 2001

Why transposons VERY interesting?

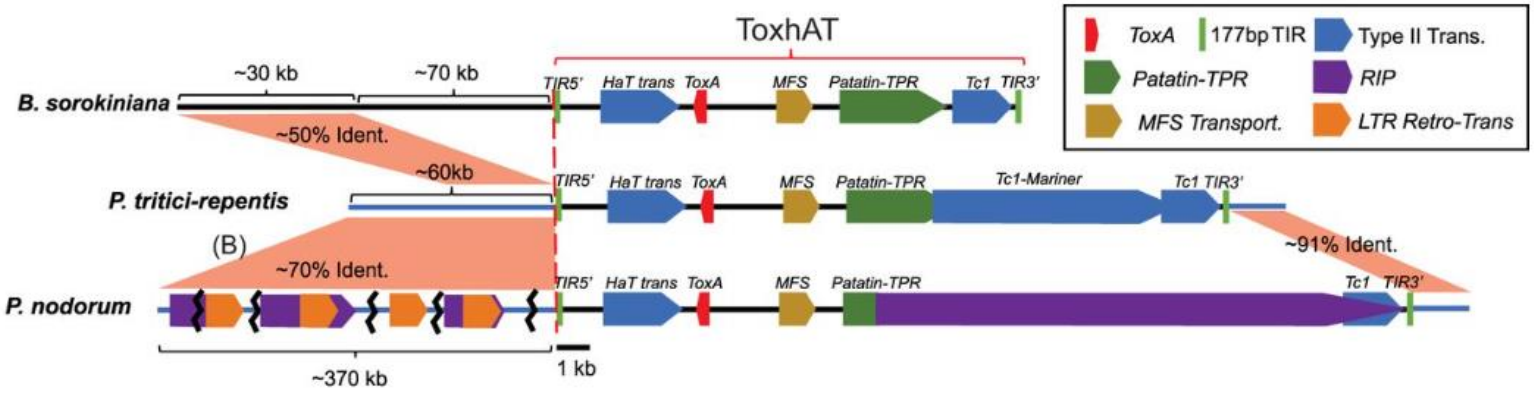
Giant transposon named *Starships* driving gene transfer



Starships are a newly discovered superfamily of gigantic transposons found across hundreds of species of *Pezizomycotina* fungi.

Urquhart et al. 2024

Giant transposon named *Starships* driving gene transfer



ToxhAT is moving as cargo within a *Starship* transposon in *B. sorokiniana*

ToxA enables fungal pathogenicity on susceptible wheat cultivars

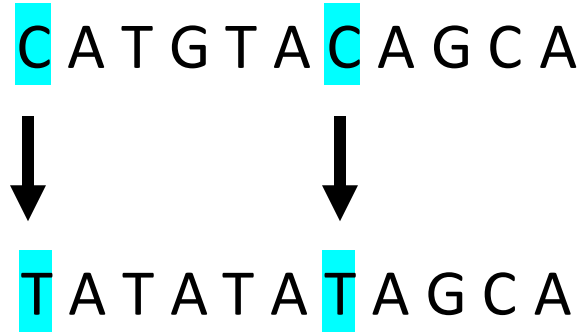
ToxA has been horizontally transferred between three fungal wheat pathogens

McDonald et al. 2019; Bucknell et al.2024

Repeat-induced point mutation (RIP) limit the accumulation of transposons

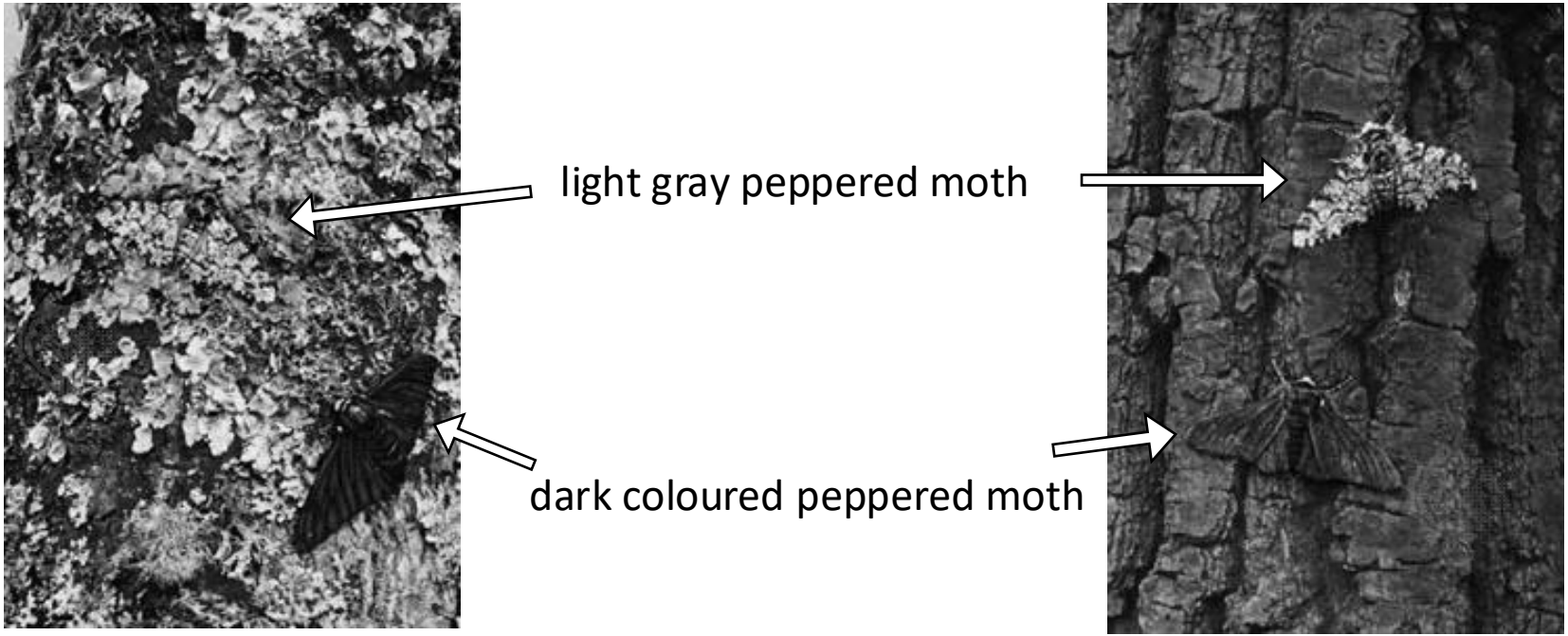
Repeat-induced point (RIP) mutation degrades transposable elements by targeting repeats with C→T mutations

Can occasionally spreads from duplicated sequences into neighboring nonrepetitive regions



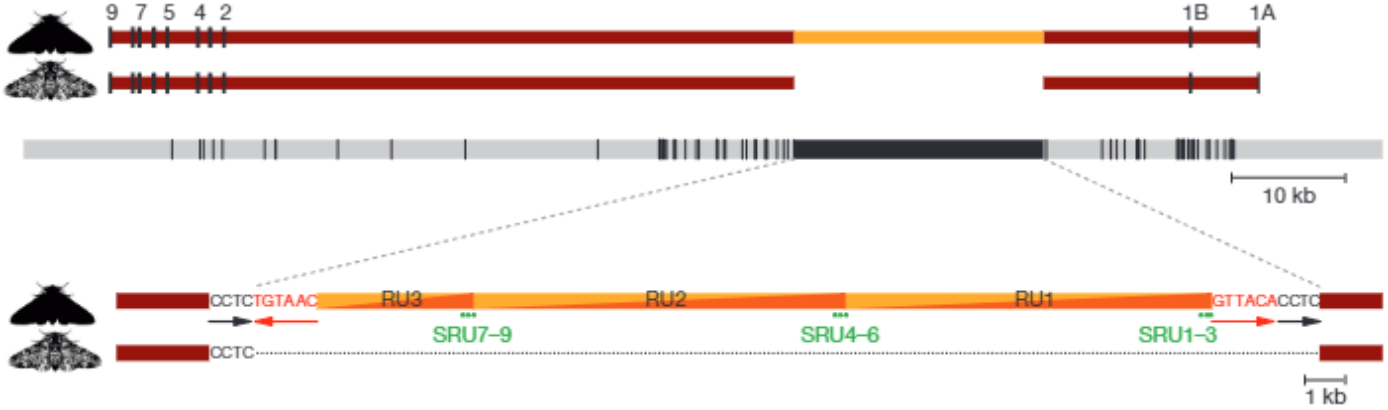
Wang et al. 2020

The insertion of a DNA transposon changed the colour of British peppered moths



https://www.britannica.com/science/industrial-melanism

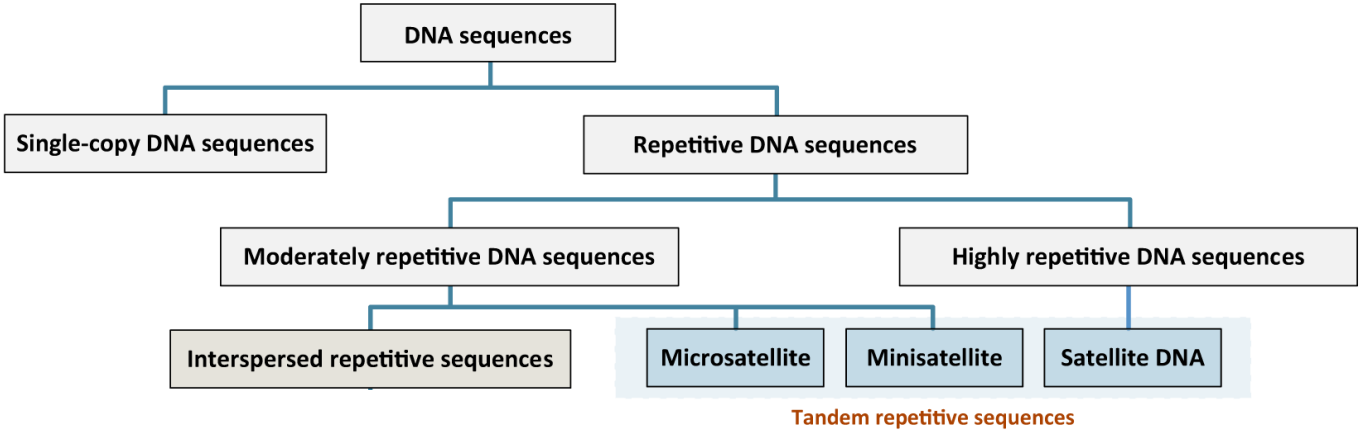
The insertion of a DNA transposon changed the colour of British peppered moths



Industrial melanism of peppered moths in Britain was the insertion of a large, tandemly repeated, transposable element into the first intron of the gene *cortex*

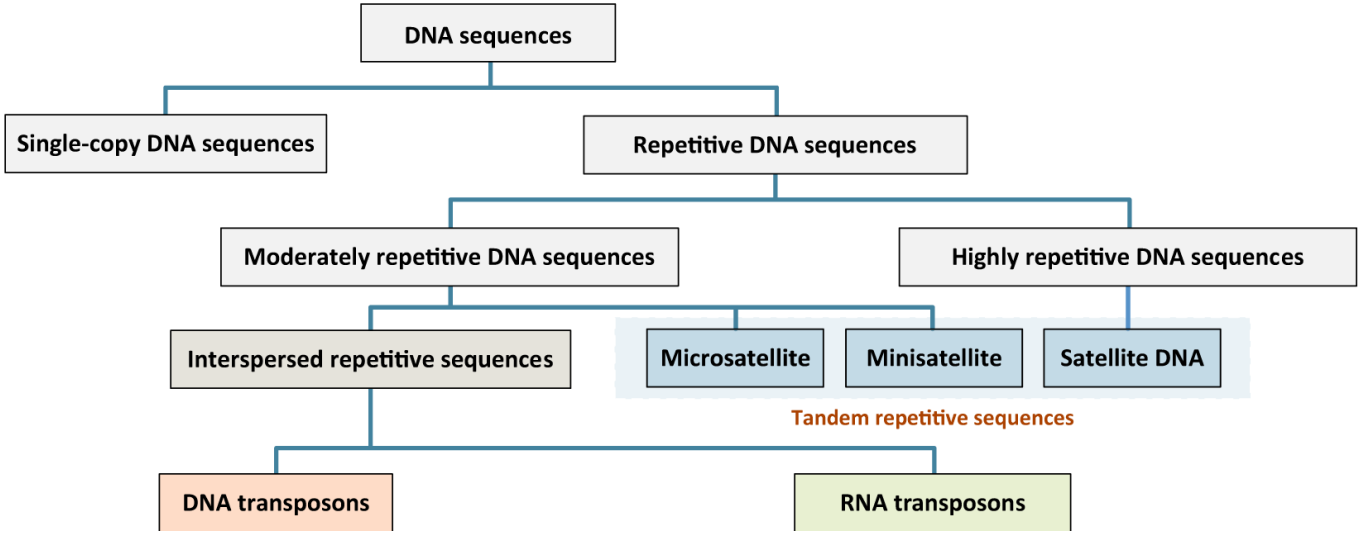
Hof et al. 2016

Classification of transposons



Liao et al. 2023

Classification of transposons



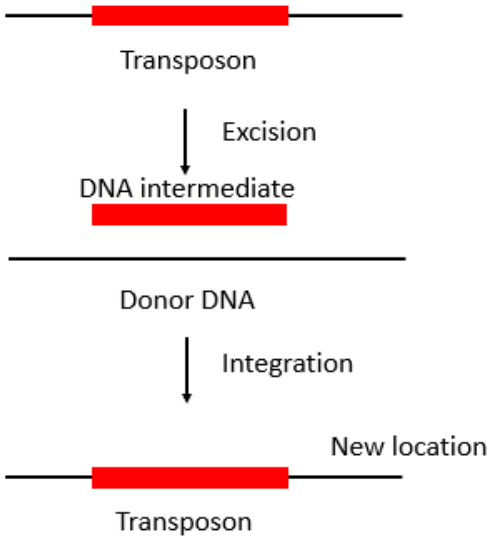
Class

Liao et al. 2023

Transposons can be classified based on transposition mechanism

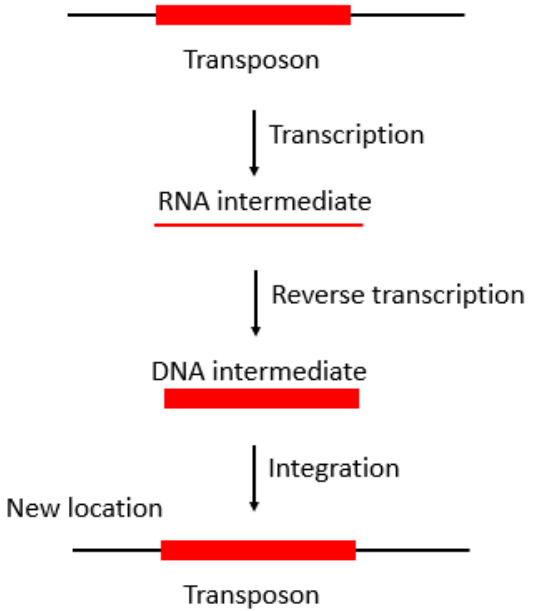
Cut-and-paste

Class II
DNA transposon

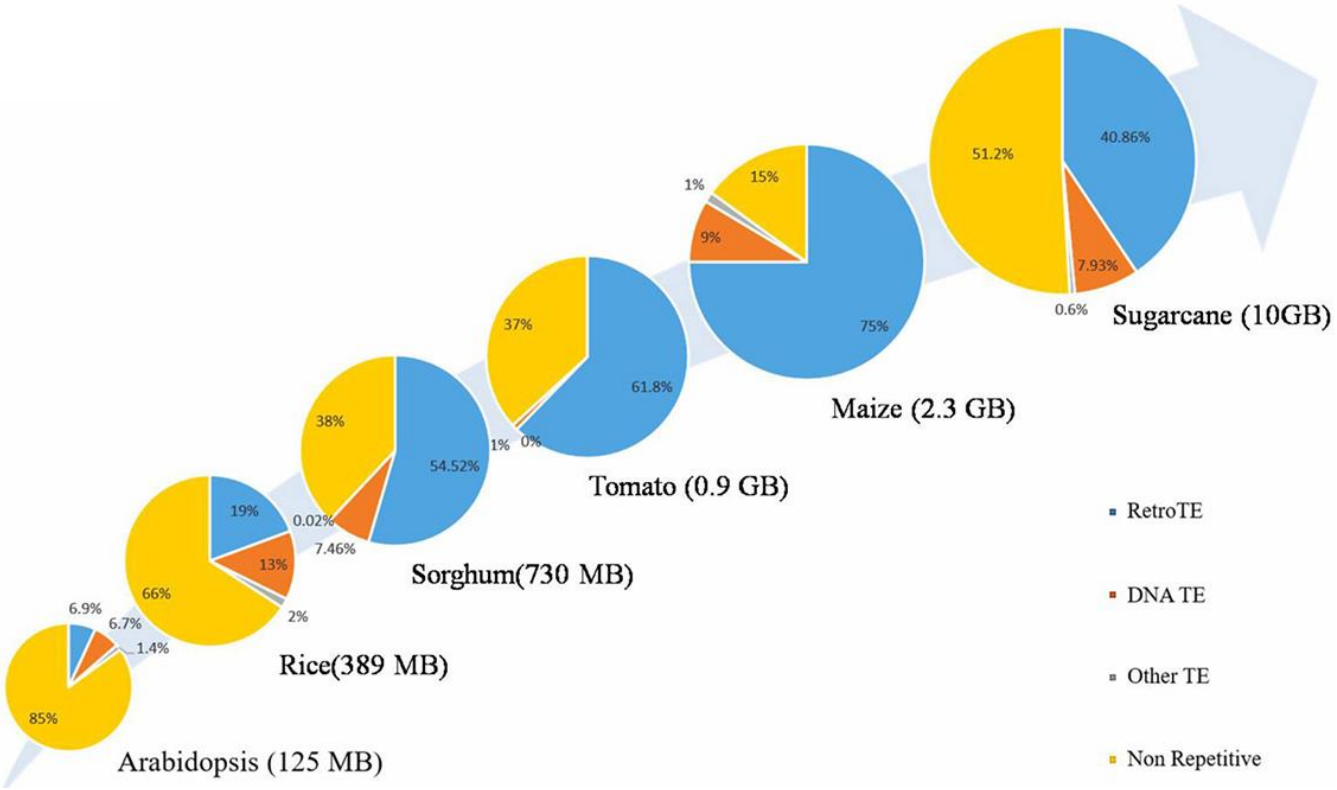


Copy-and-paste

Class I
Retrotransposon

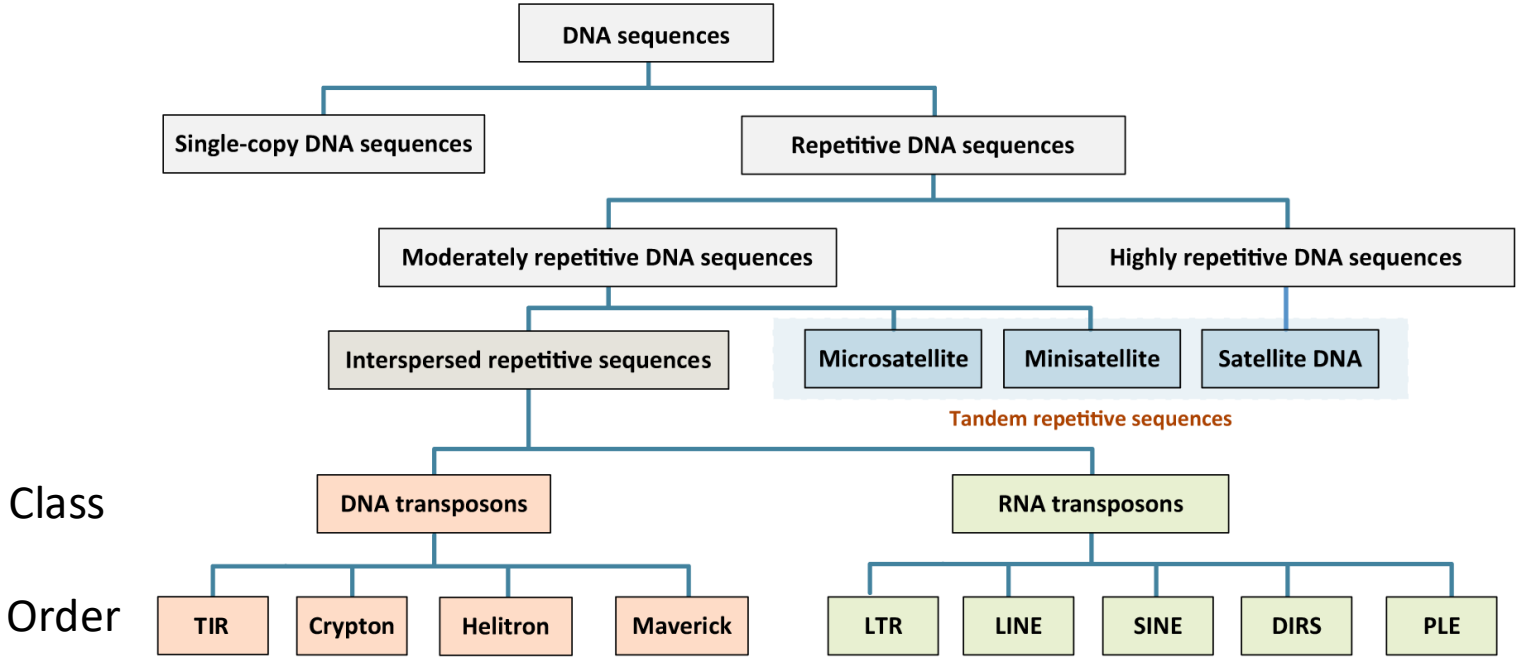


Transposon content varies between species



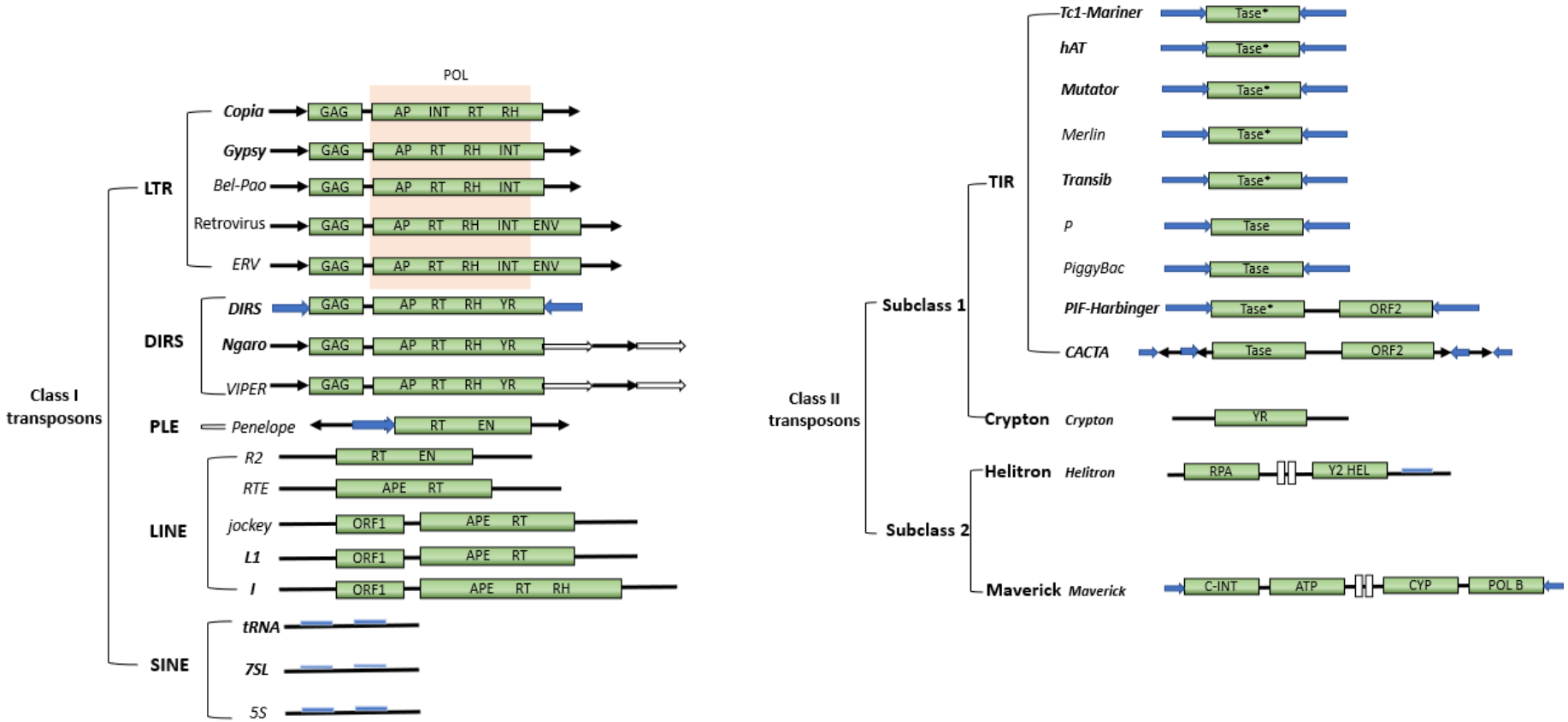
Nagi et al. 2016

Transposons can be further classified based on structure



Liao et al. 2023

Wicker et al. (2007) further classified TEs into superfamily level



Adapted from Wicker et al. 2007

Why proper TE annotation important?

Bad TE annotation can induce errors

- Influence the accuracy of gene annotation
  (Many TEs contain open reading frames that could be incorrectly annotated as host genes)

- Directly affect analysis related to TEs
  - Affect comparing TE copies/coverage within their subfamily

How to annotate TE?

Using exist database

Pros

- Time efficiency
- Enables comparison with other annotation using the same database
- Often peer-reviewed and tested

Cons

- May not work for no-model species
- Not work for species-specific TEs
- Potential bias in the annotation results

Using custom database

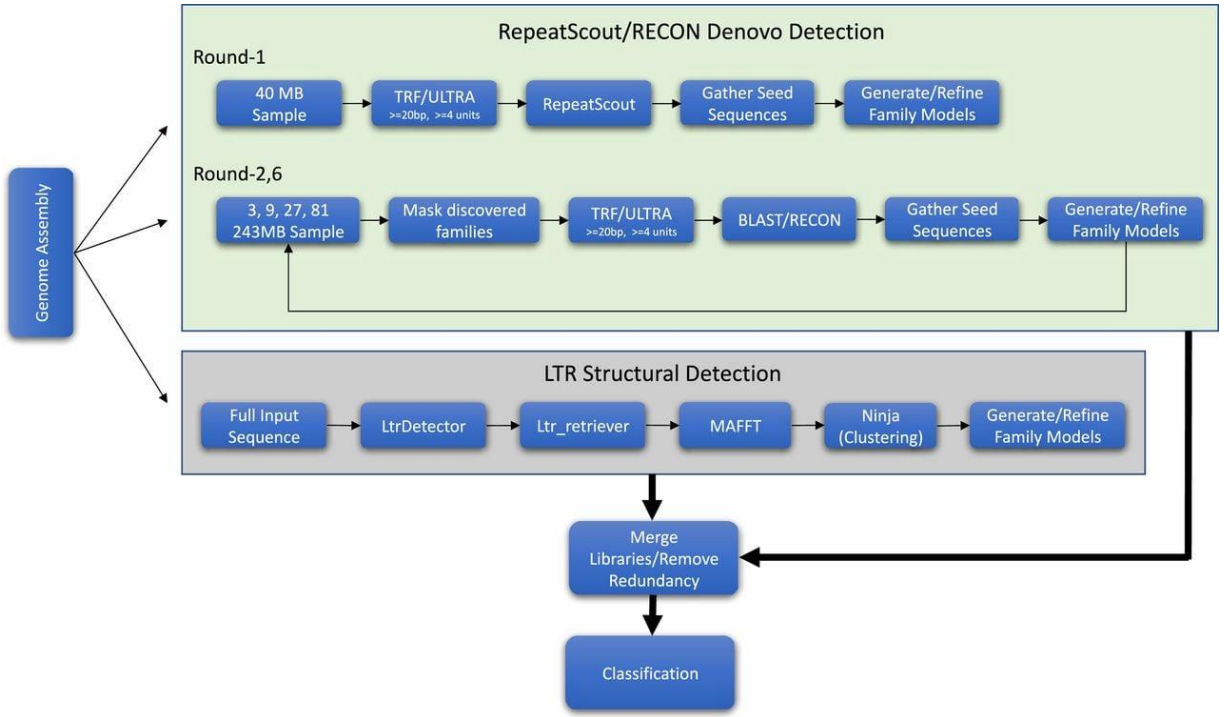
Pros

- Flexible
- Might be able to identify new TE families

Cons

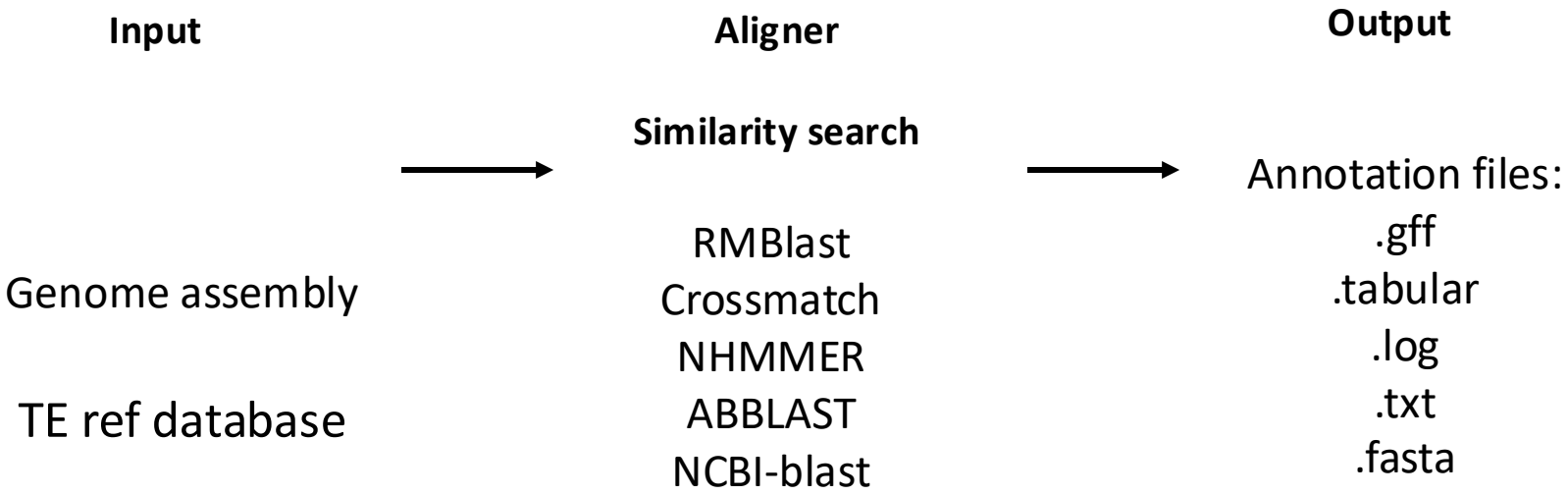
- Fail to detect low-copy number TEs
- Erroneous identification/classification
- Time consuming and labour intensive
- Consistency issues between annotators

RepeatModeler builds custom TE databases



Jullien et al. 2020

RepeatMasker annotates genome assembly using exist/custom TE databases



Jullien et al. 2020

What does RepeatMasker output look like?

Output fasta file contains masked genome

Soft-masked genome

```
CTCAATACCGTCTGAAACGGATGCGGACGGCGGCGGTGTTTTTTGTATGA
AAATATCGGTTTTTAACCGATATTTTCATTCTTTGTCAAACGACGCGCTG
CCGTTTTTCGCGGGCGGATGTTTTTATATTTGTTTCAATTCAATGGATTG
TATTTTAGAGGACGTGTTCCGATACGGCGCGGGTAAATCCTTTTCTGTCA
ATGGCTTATCCGATAGGGCGGTTTTTACTTAGATGGAAAATCACTTCCTT
ATATATCCGCCACGCCGAAGGGCGCATtatggtggattaactttaaaccg
gtacggcgttgccccgccccggctcaaagggaacgattccctaaggcgcc
aagcaccgggcgaaccgattccgtaccatttgtactgcctgcggcccgcc
gccttgtcctgatttttgttaatccgctataTTTTTCTTAACCATCCCTT
CCAACAGCCGTGCGAAGGTTTTTTTATATCAGTGGGAAATGCAATATTCT
ATTGTTTTTATTGTAGAATTTTAAAAACAGATTGTTGTGTTCCGCGTTTT
TGCCGGTTTGGAAGCGGTGGGGCGCATCAGCCTTTTATAAAGGCATATCG
GAAGCCTGTATAAGGTTTTTGAACATATCGATCCTGTTCCCTTGCAGGCG
```

Nucleotides annotated as repeats have been converted to lowercase

Hard-masked genome (Default)

```
CTCAATACCGTCTGAAACGGATGCGGACGGCGGCGGTGTTTTTTGTATGA
AAATATCGGTTTTTAACCGATATTTTCATTCTTTGTCAAACGACGCGCTG
CCGTTTTTCGCGGGCGGATGTTTTTATATTTGTTTCAATTCAATGGATTG
TATTTTAGAGGACGTGTTCCGATACGGCGCGGGTAAATCCTTTTCTGTCA
ATGGCTTATCCGATAGGGCGGTTTTTACTTAGATGGAAAATCACTTCCTT
ATATATCCGCCACGCCGAAGGGCGCATNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTTTTCTTAACCATCCCTT
CCAACAGCCGTGCGAAGGTTTTTTTATATCAGTGGGAAATGCAATATTCT
ATTGTTTTTATTGTAGAATTTTAAAAACAGATTGTTGTGTTCCGCGTTTT
TGCCGGTTTGGAAGCGGTGGGGCGCATCAGCCTTTTATAAAGGCATATCG
GAAGCCTGTATAAGGTTTTTGAACATATCGATCCTGTTCCCTTGCAGGCG
```

Nucleotides annotated as repeats have been converted to N or X

Summary table contains summary of annotation result

```
==================================================
file name: rm_input.fasta
sequences:             2
total length:    2183488 bp  (2056736 bp excl N/X-runs)
GC level:         52.52 %
bases masked:       2328 bp ( 0.11 %)
==================================================
                   number of      length   percentage
                   elements*    occupied  of sequence
--------------------------------------------------
SINEs:                 5          303 bp    0.01 %
      ALUs             0            0 bp    0.00 %
      MIRs             1           76 bp    0.00 %

LINEs:                10          780 bp    0.04 %
      LINE1            1           81 bp    0.00 %
      LINE2            0            0 bp    0.00 %
      L3/CR1           4          221 bp    0.01 %

LTR elements:          0            0 bp    0.00 %
      ERVL             0            0 bp    0.00 %
      ERVL-MaLRs       0            0 bp    0.00 %
      ERV_classI       0            0 bp    0.00 %
      ERV_classII      0            0 bp    0.00 %

DNA elements:          4          333 bp    0.02 %
     hAT-Charlie       0            0 bp    0.00 %
     TcMar-Tigger      2          180 bp    0.01 %

Unclassified:          0            0 bp    0.00 %

Total interspersed repeats:     1416 bp    0.07 %


Small RNA:            15          963 bp    0.05 %

Satellites:            0            0 bp    0.00 %
Simple repeats:        0            0 bp    0.00 %
Low complexity:        0            0 bp    0.00 %
==================================================
```

Copy number

TE classification

Overview information

Total coverage of each category

Total TEs annotated

TR annotated

Annotation (.gff|.tabular) file contains loci of all annotated repeats

Seq id		Feature	Start	End	Divergence(%)	strand		ID of TEs	TE family	Start|End of TE consensus sequence
8416_circular_np2	RepeatMasker	dispersed_repeat	48161	48237	31.6	+	.	ID=1;Target "Motif:tRNA-Arg-AGG" 1 76		
8416_circular_np2	RepeatMasker	dispersed_repeat	48261	48336	26.7	+	.	ID=2;Target "Motif:MIR1_Amn" 1 79		
8416_circular_np2	RepeatMasker	dispersed_repeat	86783	86833	20.0	-	.	ID=3;Target "Motif:tRNA-Met_v" 12 61		
8416_circular_np2	RepeatMasker	dispersed_repeat	385752	385891	38.1	-	.	ID=4;Target "Motif:Tigger19b" 322 458		
8416_circular_np2	RepeatMasker	dispersed_repeat	464771	464828	32.8	+	.	ID=5;Target "Motif:Penelope1_Vert" 814 871		

Score	Deletion(%)	Seq id	Start	End		TE family	Start|End of TE consensus sequence				
	Divergence(%)	Insertion(%)					Classification (Order)				
284	31.6	0.0	1.3	8416_circular_np2	48161	48237	(2131098)	tRNA-Arg-AGG	tRNA	1	76	(0)	1	
184	26.7	5.3	1.3	8416_circular_np2	48261	48336	(2130999)	MIR1_Amn	SINE/MIR	1	79	(151)	2	
229	20.0	0.0	2.0	8416_circular_np2	86783	86833	(2092502)	C	tRNA-Met_v	tRNA	(13)	61	12	3
194	38.1	2.1	4.4	8416_circular_np2	385752	385891	(1793444)	C	Tigger19b	DNA/TcMar-Tigger	(92)	458	322	4
185	32.8	0.0	0.0	8416_circular_np2	464771	464828	(1714507)	Penelope1_Vert	LINE/Penelope	814	871	(208)	5	

An alignment file contains alignments of all annotated repeats against reference

Score	Deletion(%)	Seq id	Start	End	TE family	Start|End of TE consensus sequence	
	Divergence(%) Insertion(%)				Classification		
184	26.7 5.3 1.3	8416_circular_np2	48261	48336	(2130999) MIR1_Amn SINE/MIR	1 79	(151) 2

```
184 26.66 5.26 1.27 8416_circular_np2 48261 48336 (2130999) MIR1_Amn#SINE/MIR 1 79 (151) m_b3s601i0

  8416_circular      48261 TATGGTGGC--TGTAGCTCAGTTGGTAGAGCCCCGGATTGTGATTCCGGT 48308
                             v  v   --   i iii   v i      v i   i  v  i -   v
  MIR1_Amn#SINE          1 TAGGGAGGCAGTGTGGTCTAGTGGATAGAGCACTGGACTGGGACT-CGGG 49

  8416_circular      48309 TGTCGTGGGTTCGAGCCCCA--TCAGCCAC 48336
                           v v v       ? ?i   i--  v
  MIR1_Amn#SINE         50 AGACCTGGGTTCNANTCCCGGCTCTGCCAC 79

Matrix = 25p53g.matrix
Kimura (with divCpGMod) = 32.42
CpG sites = 10, Kimura (unadjusted) = 34.47
Transitions / transversions = 1.00 (10/10)
Gap_init rate = 0.04 (3 / 75), avg. gap size = 1.67 (5 / 3)
```

First row identical to tabular file

Followed by aligning query genome and reference TE sequence

Recommended readings

- Van't Hof, A.E., Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., Hall, N. and Darby, A.C., IJ Saccheri The industrial melanism mutation in British peppered moths is a transposable element., 2016, 534.
- McDonald, M.C., Taranto, A.P., Hill, E., Schwessinger, B., Liu, Z., Simpfendorfer, S., Milgate, A. and Solomon, P.S., 2019. Transposon-mediated horizontal transfer of the host-specific virulence protein ToxA between three fungal wheat pathogens. MBio, 10(5), pp.10-1128.
- Urquhart, A., Vogan, A.A. and Gluck-Thaler, E., 2024. Starships: a new frontier for fungal biology. Trends in Genetics.
- Gladyshev, E., 2017. Repeat-induced point mutation and other genome defense mechanisms in fungi. *The fungal kingdom*, pp.687-699.
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O. and Paux, E., 2007. A unified classification system for eukaryotic transposable elements. Nature reviews genetics, 8(12), pp.973-982.

Australian National University