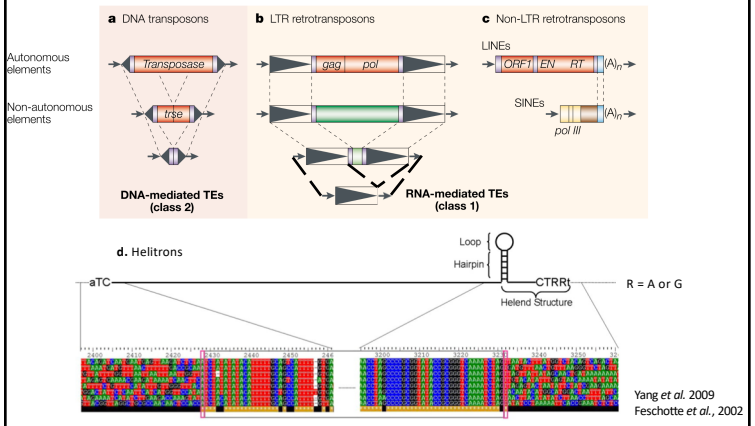
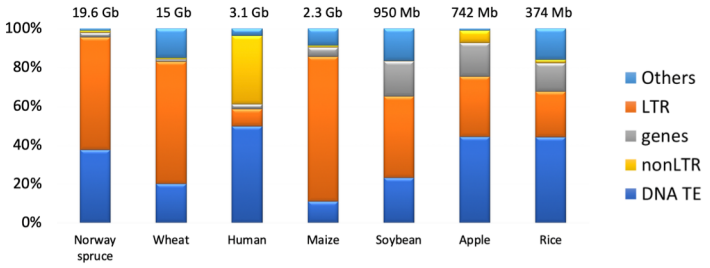
Transposable Element Annotation

Annotating Transposable Elements



Yang *et al.* 2009
Feschotte *et al.*, 2002

TEs are prevalent in eukaryotic genomes



Approaches to TE Annotation

- Copy-number based
  - 👍: sensitively identify repetitive sequences
  - 👎: not knowing what are the sequences, could be duplicated genes
- Homology based
  - 👍: reuse prior knowledge; quick
  - 👎: limited by prior knowledge; many TEs sequences are not conserved
- Structural based
  - 👍: codable; independent of database
  - 👎: limited by knowledge of sequence structures; high FDR

Structural Features of TEs

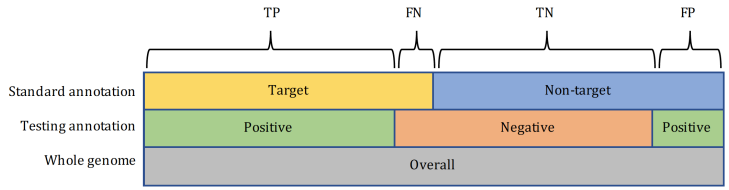
Class	Superfamily		TSD	Element size		TIR or LTR size	Terminal sequence (5'...3')
	Traditional name	Systematic classification		Autonomous	Non-autonomous		
Class I	LTR-*Copia*	RLC	5 bp	4–11 kb	292 bp to 4 kb	85 bp to 3 kb	TG...C/G/
	LTR-*Gypsy*	RLG	5 bp	5–20 kb	2–5 kb	120 bp to 6 kb	TG...CA
	LINE	RIL/RII	Variable	Up to 9 kb	NA	NA	...AAA
	SINE	RST/RSL	Variable	NA	80–500 bp	NA	...AAA
Class II	*Ac/Ds/hAT*	DTA	8 bp	3–6 kb	110 bp–3 kb	5–22 bp	C/TA...TA
	En/Spm/dSpm/CACTA	DTC	3 bp	6–21 kb	200 bp–6 kb	12–28 bp	CACTA/G. C/TAGTG
	MuDR/Mutator/Mu/MULE	DTM	7–11 bp, mostly 9	4–16 kb	120 bp–3 kb	0–800 bp	G/C...G/C
	PIF/Harbinger/Tourist	DTH	TNA	3–7 kb	80 bp–3 kb	14–60 bp	GGG/CC... GG/CCC GG/AGCA TGCT/CC
	Tc1/Mariner/Stowaway	DTT	TA	3–7 kb	80 bp–3 kb	11–120 bp	CTCCCTC GGAGGG
	Helitron	DHH	None	5–17 kb	150 bp–20 kb	None	TC...CTR/

(Zhao, D., A. Ferguson, N. Jiang, BBA Gene Regulatory Mechanisms 2016)

Many Different TE Annotation Software

- General TE annotator
  - RepeatModeler, RepeatScout, RepeatMasker, RepBase, Red, RECON, CENSOR, PILER, REPET, GenericRepeatFinder (GRF)
- LTR retrotransposons
  - LTR_STRUC, LTR_MINER, LTR_FINDER, LTRharvest, MGEScan_LTR/LTRrho, LTR Annotator, LTR_retriever, LtrDetector, GRF
- Non-LTR retrotransposons
  - SINE-Finder, SINE_scan, SINEBase, MGEScan_nonLTR, TSDFinder
- MITE
  - MITE-Hunter, detectMITE, IRF, TIRvish, GRF, miteFinderII, P-MITE, TIR-Learner, MITE-Tracker, MUSTv2, MITE Digger, RSPB
- Helitron
  - HelSearch2, HelitronFinder, HelitronScanner

How to Pick a Software?



$$\text{Sensitivity} = P(\text{positive}|\text{target}) = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{Specificity} = P(\text{negative}|\text{non\_target}) = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

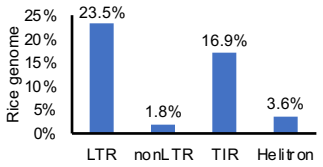
$$\text{Accuracy} = P(\text{true\_classification}) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad \text{Precision} = P(\text{target}|\text{positive}) = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$F_1 = \frac{2 * \text{Sensitivity} * \text{Precision}}{\text{Sensitivity} + \text{Precision}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \qquad \text{FDR} = P(\text{non\_target}|\text{positive}) = \frac{\text{FP}}{\text{TP} + \text{FP}}$$

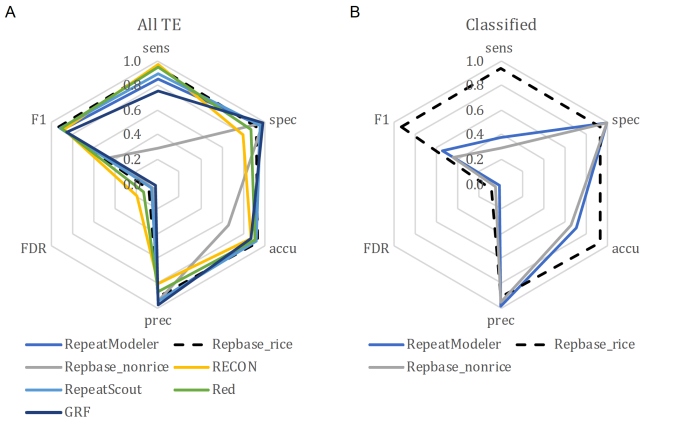
Ou et al. (2019) Genome Biol.

Subject genome – *Oryza sativa* (rice)

- High-quality TE space assembly (LAI = 20, "gold" quality)
- Harbors wide range of TEs in reasonable abundance (46%)
- Small genome (375 Mb) allows quick analyses
- Has a manually curated TE library.

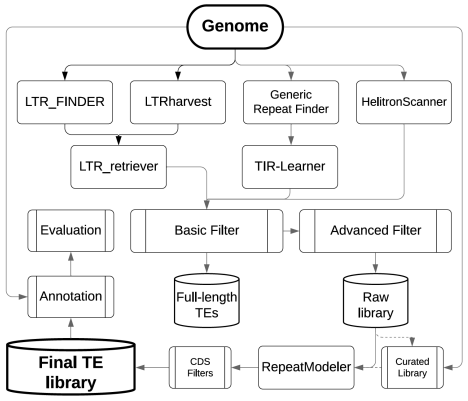


General TE Identifiers



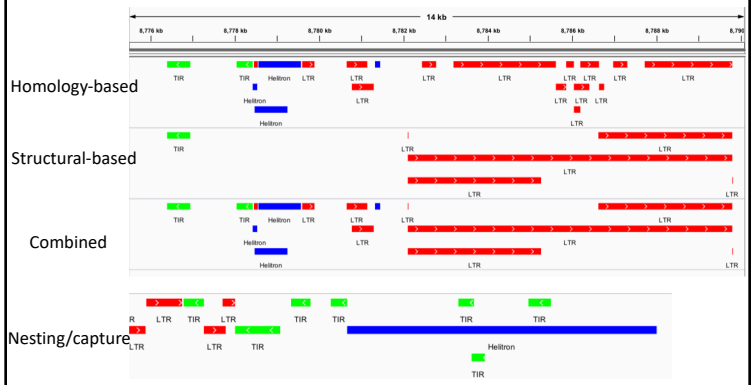
Ou et al. (2019) Genome Biol.

Example of a Full TE Annotation Pipeline

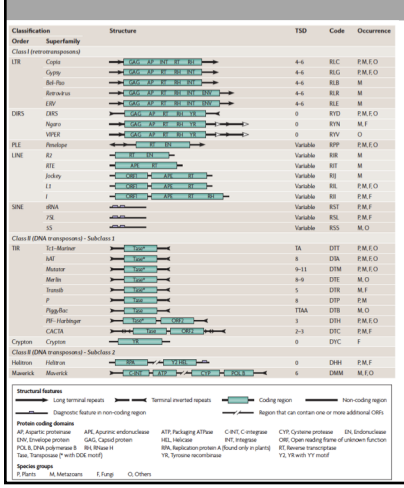


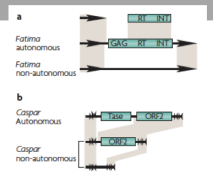
Ou et al. (2019) Genome Biol.

Example of a Full TE Annotation Pipeline



Autonomous vs. Non-Autonomous





Autonomous and non-autonomous TEs

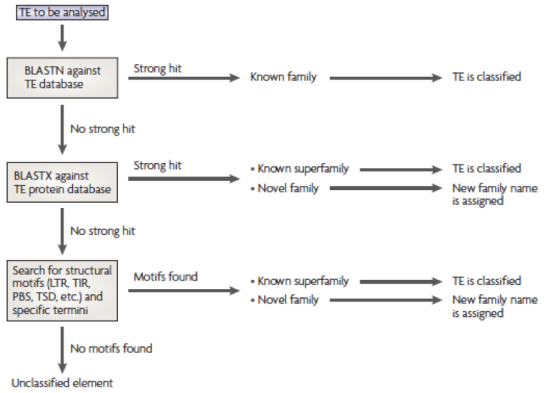
Our definition of family also addresses another problem: many autonomous TEs give rise to non-autonomous deletion derivatives [FIG. 2]. Here we define an element as autonomous simply if it appears to encode all the domains that are typically necessary for its transposition, without implying that the element is either functional or active. A family can contain various elements that have been rendered defective from point mutations or small insertions or deletions (indels), but that nevertheless retain not only all the coding regions of an active TE, but also sufficiently high DNA identity for family membership. Such elements that are defective owing to their limited mutations were probably functional in the recent past[18].

Autonomous but defective elements stand in contrast to non-autonomous TEs, which we define as any group of elements that lacks some (or all) of the domains found in autonomous elements[18]. Usually, non-autonomous

80-80-80 Rule

- Used to define family and subfamily
- TE Family – group of TEs that have high DNA sequence similarity in their coding region/internal domain, or in their terminal repeat region
- 80% sequence similarity in at least 80% of aligned sequence when considering segments of 80bp or longer

Superfamily and Family Classification



TE Annotation Takeaways

- Like gene annotation – there are rules…but they aren't as defined as they are for genes
- Annotations are done using structure and homology
- Different programs used to annotate different TE orders each with different accuracy levels
- Family classification important for downstream biological interpretation