

## Supplementary Text

# TE-AID gallery

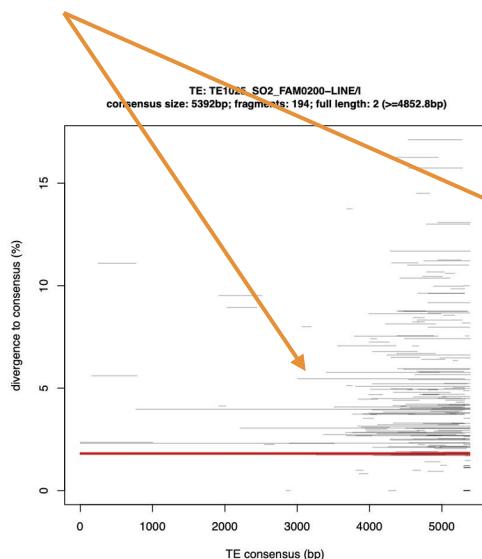
This document provides examples  
of the most common outputs from  
running the TE-AID  
(<https://github.com/clemgoub/TE-Aid>)

# Part I – “Canonical” examples

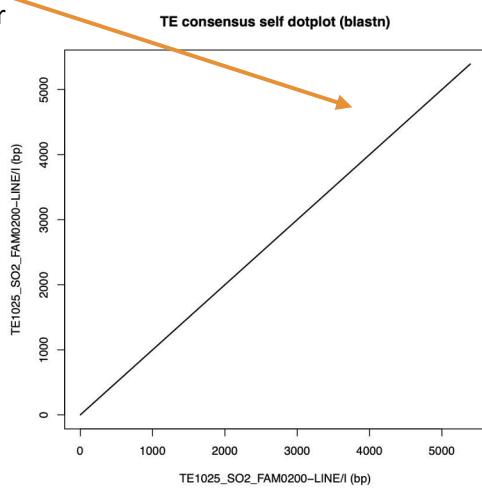
Canonical consensus: LINE (LINE/I)

**Species:** *Sitophilus oryzae*  
(Insecta: Curculionidae)

Typical LINE  
5' truncation pattern



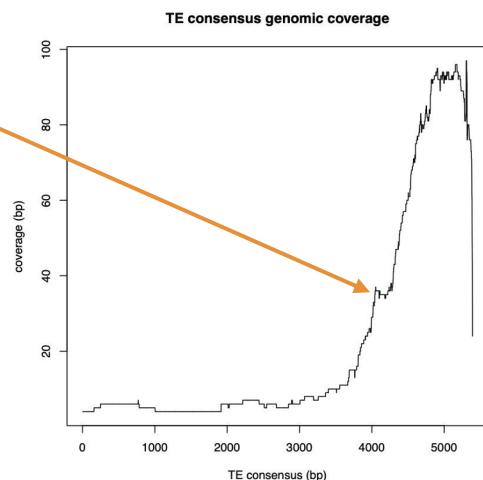
No TIR or  
LTR, as it is  
expected for  
LINEs



*orf1*      *orf2*



LINE elements do not have any termini structural feature. However, they are usually truncated at 5' end which could be appreciated by analysing either the alignment or the genomic coverage plot of a full-length element. Autonomous elements have an ORF with at least an EN and a RT proteins, and some types also possess another ORF (*orf1* in the figure) whose function remains unclear. EN, endonuclease.



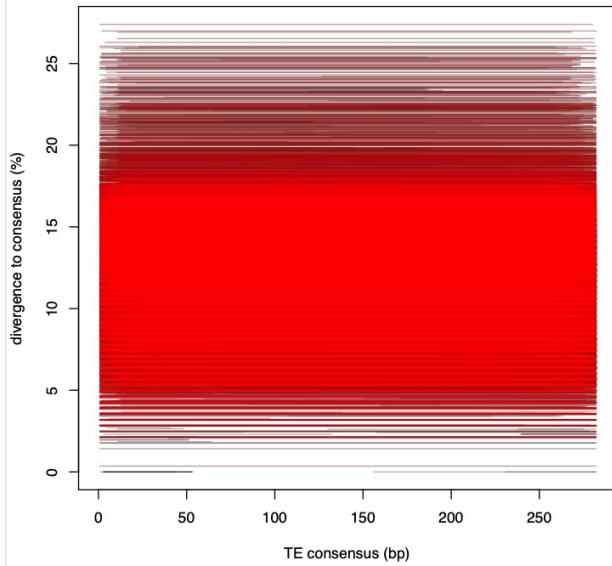
Strong hits against  
LINE gag and pol  
proteins in the  
correct  
orientation  
(compatible with  
5' deletion  
pattern)

Canonical consensus: SINE (*A*/*uY*)

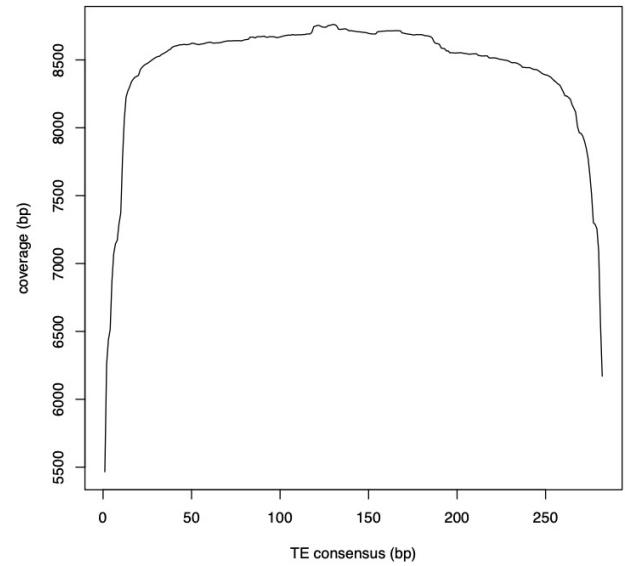
7SL/tRNA

Species: *Homo sapiens*

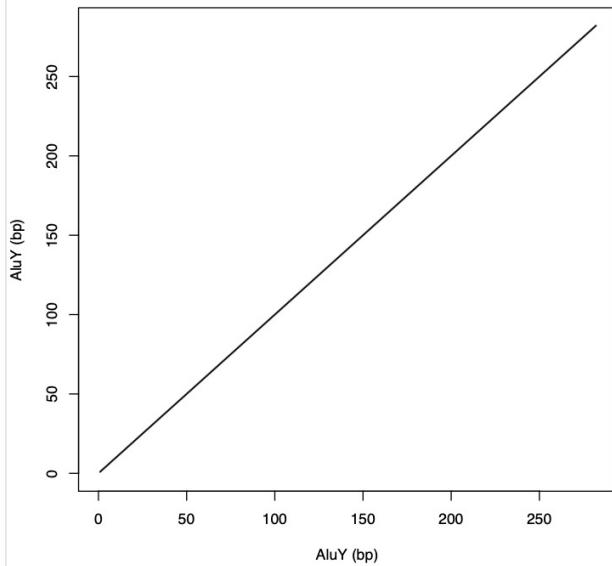
TE: AluY  
consensus size: 282bp; fragments: 9055; full length: 7742 ( $\geq 253.8$ bp)



TE consensus genomic coverage

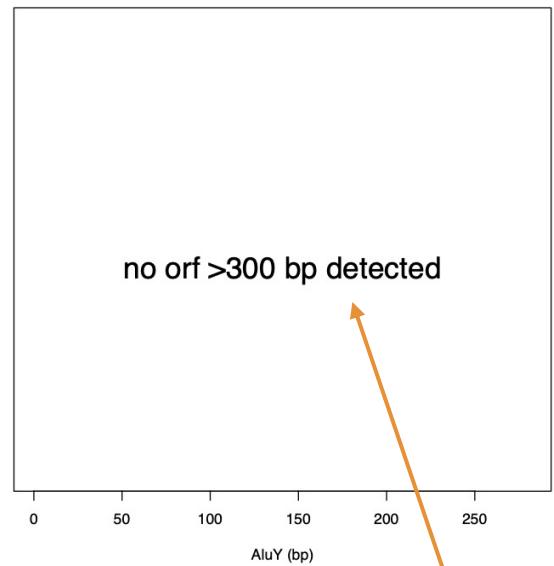


TE consensus self dotplot (blastn)



Short (< 500bp)  
consensus

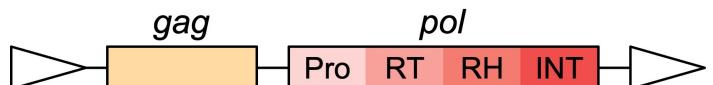
TE consensus structure and protein hits



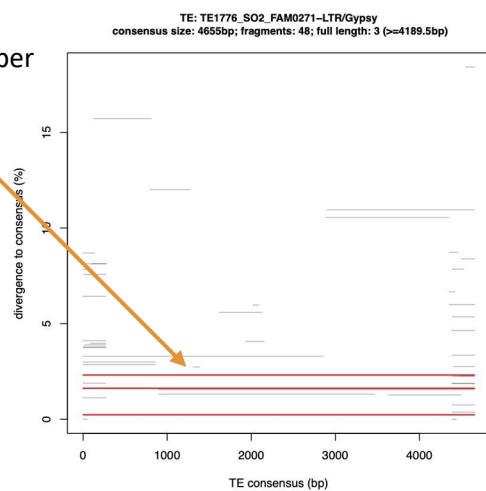
no orf >300 bp detected

## Canonical consensus: LTR (LTR/Gypsy)

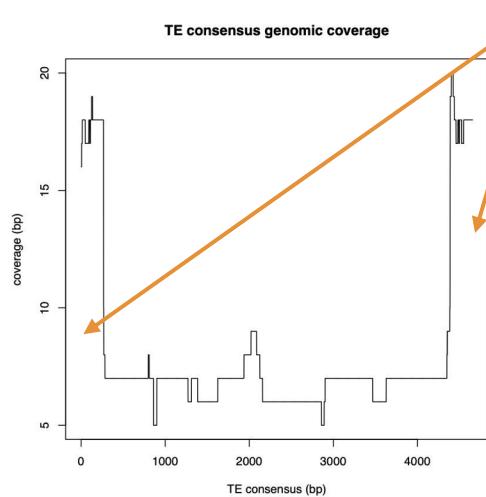
**Species:** *Sitophilus oryzae*  
(Insecta: Curculionidae)



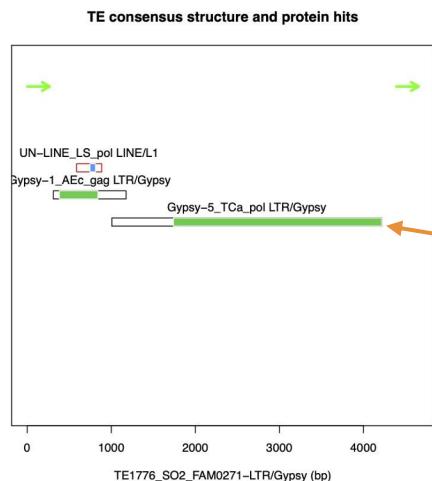
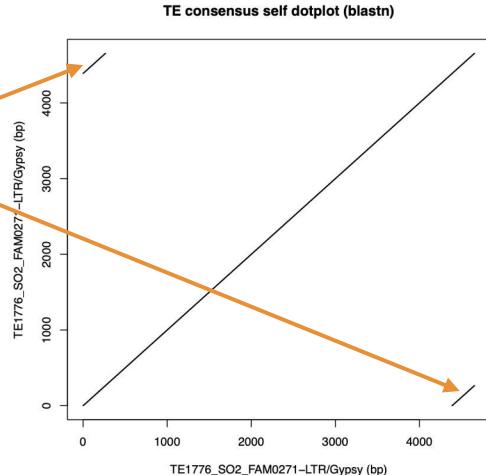
Low copy number  
3 full length



More “solo”-LTR than full-length copies



LTRs



In this example, only a few full-length copies of the element are found. This is not necessarily the case for all LTRs. Depending on the TE family and the species, LTR elements can be found in hundreds or thousands of copies.

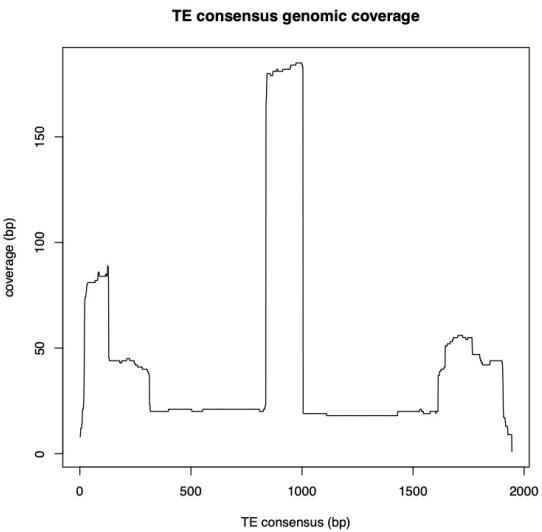
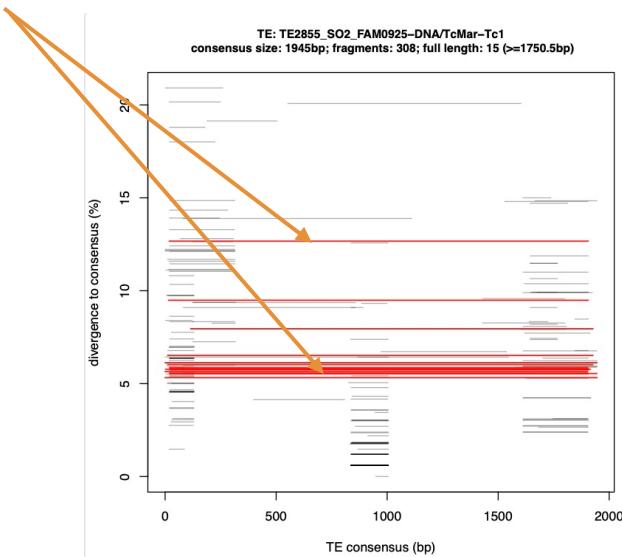
Canonical consensus: DNA/TIR

*orf 1*

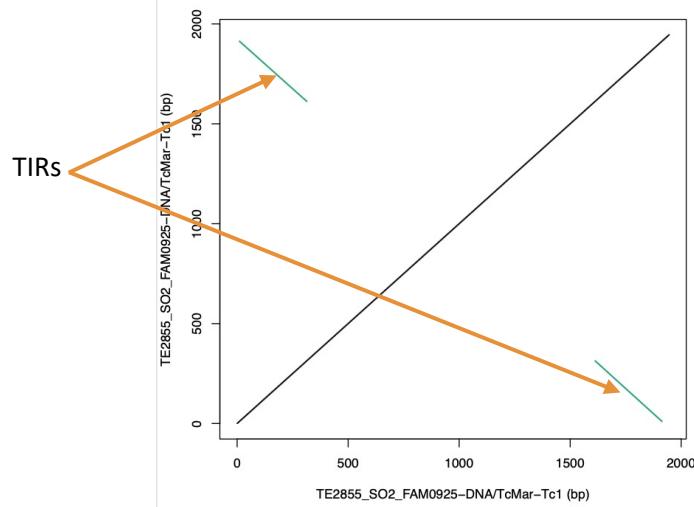
Species: *Sitophilus oryzae*  
(Insecta: Curculionidae)

TR

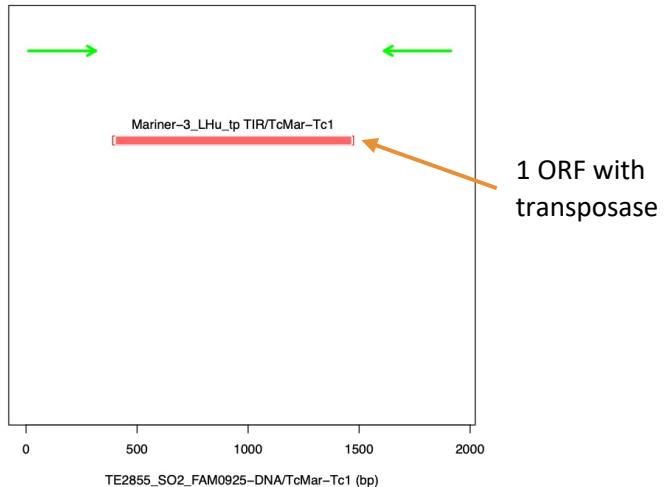
Full-length (autonomous)



TE consensus self dotplot (blastn)



TE consensus structure and protein hits



In this example, we can identify the hallmarks of DNA/TIR transposons: two terminal inverted repeats (TIRs) flanking a single transposase ORF. The full length consensus is of expected size, ~2000bp. In this particular example, we can notice that a portion of the transposase is highly conserved, suggesting that this sequence may have a functional role either for the transposon or the host.

# **Part II – “Not so canonical” examples**

Chimerical consensus: LINE + Unknown

**Species:** *Sitophilus oryzae* (Insecta: Curculionidae)

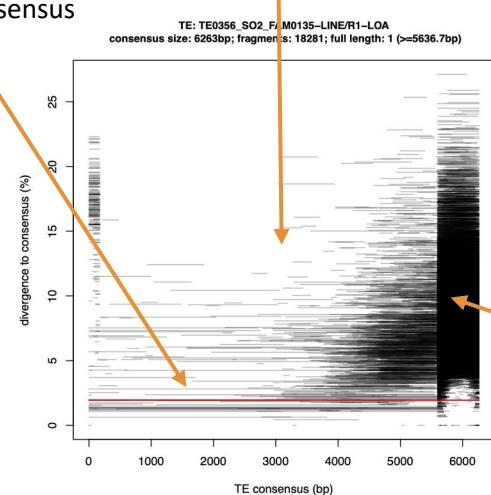
*orf1*      *orf2*



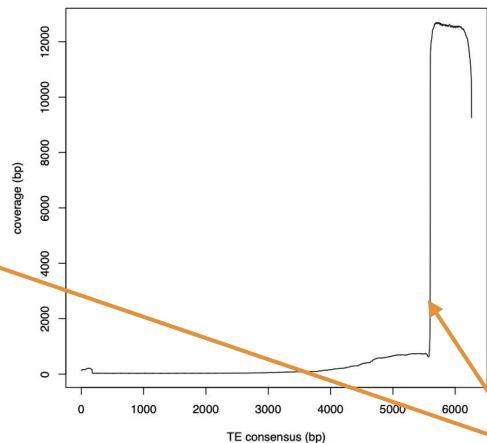
LINE elements do not have any termini structural feature. However, they are usually truncated at 5' end which could be appreciated by analysing either the alignment or the genomic coverage plot of a full-length element. Autonomous elements have an ORF with at least an EN and a RT proteins, and some types also possess another ORF (*orf1* in the figure) whose function remains unclear. EN, endonuclease.

Typical LINE  
5' truncation pattern

Only 1 hit support the complete consensus sequence

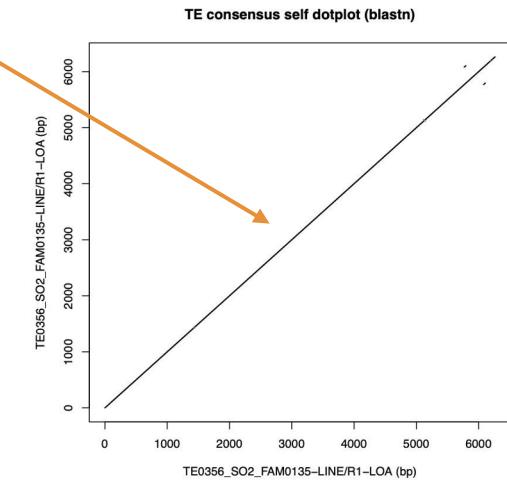


TE consensus genomic coverage

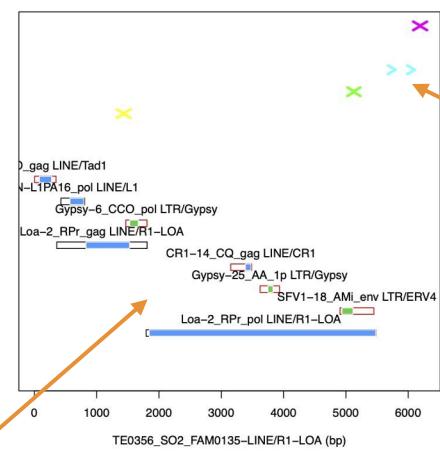


Unknown repeat ~1kb  
High coverage

No TIR or LTR, consistent with a LINE



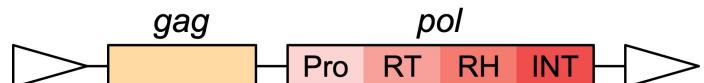
TE consensus structure and protein hits



These repeats are too small and too close (< 1kb) to each other to be LTRs. There is also no LTR protein hit

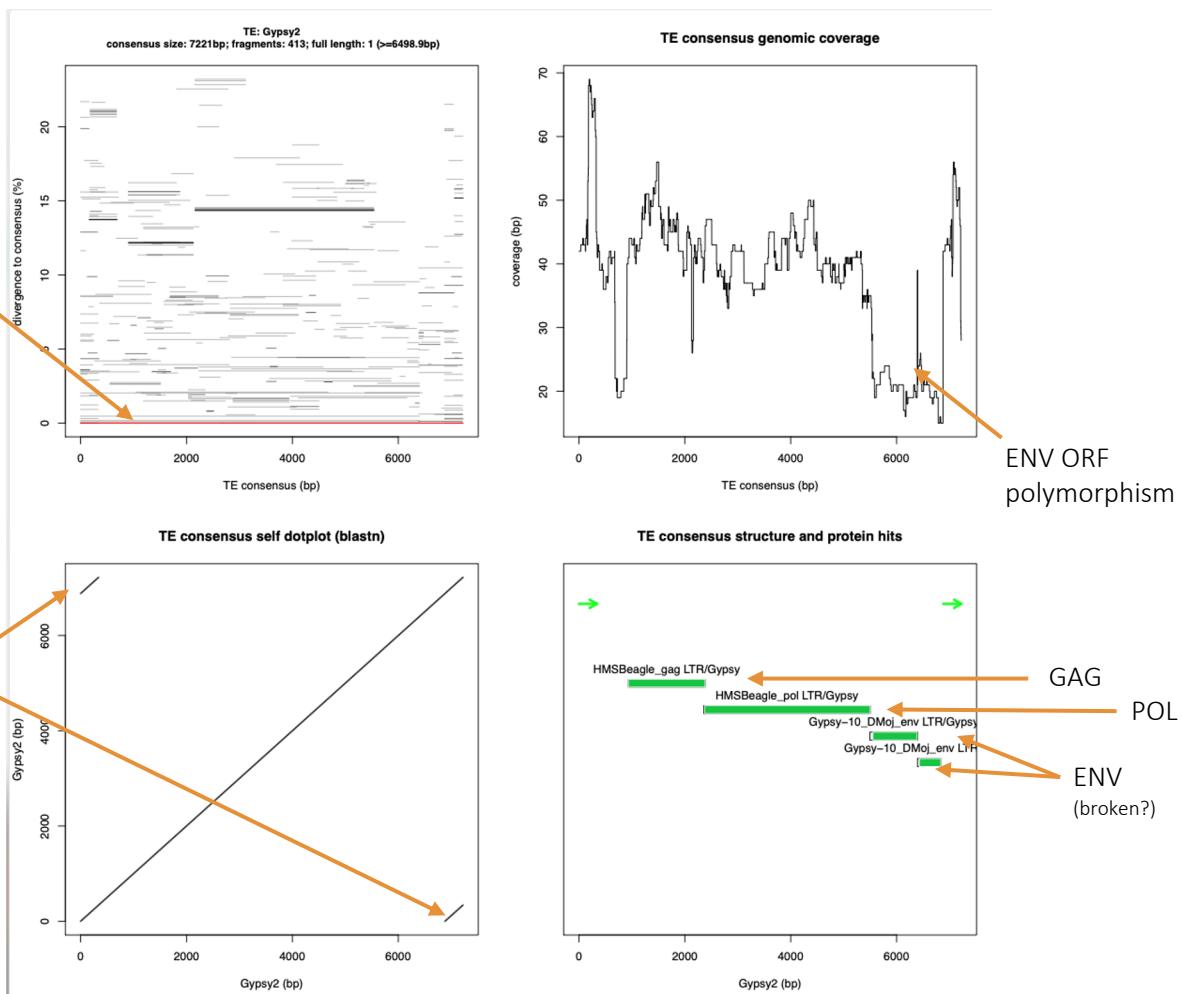
Strong hits against LINE gag and pol proteins in the correct orientation (compatible with 5' deletion pattern)

Structural Variation: LTR/Gypsy



**Species:** *Drosophila melanogaster*

1 full-length copy



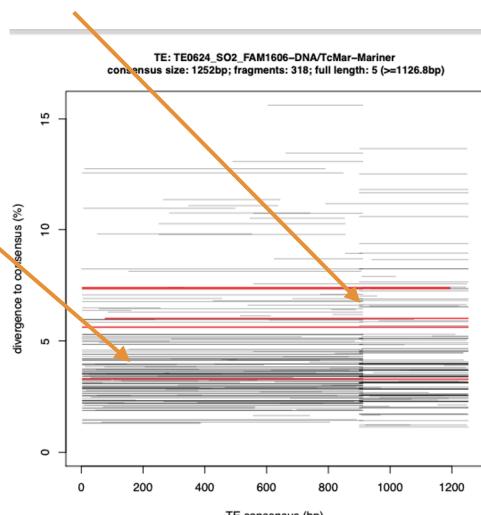
Broken ORF: TIR/Tc1-Mariner

**Species:** *Sitophilus oryzae*  
(Insecta: Curculionidae)



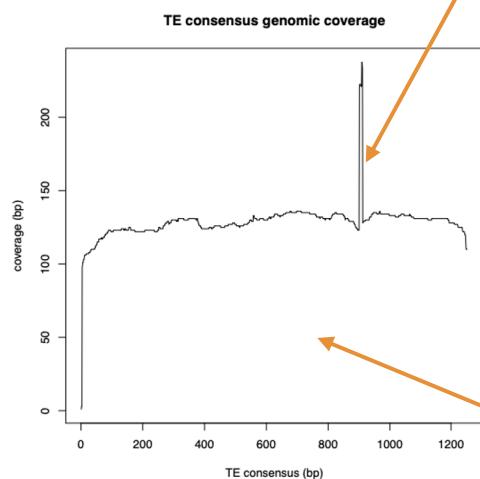
Short overlap suggest that  
a few residues at this  
location may be incorrect  
(deletion? IT stops the  
blast elongation)

Coverage x2 at overlap  
between blastn hits

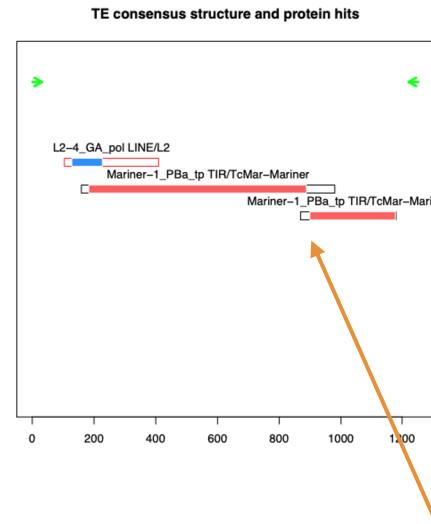
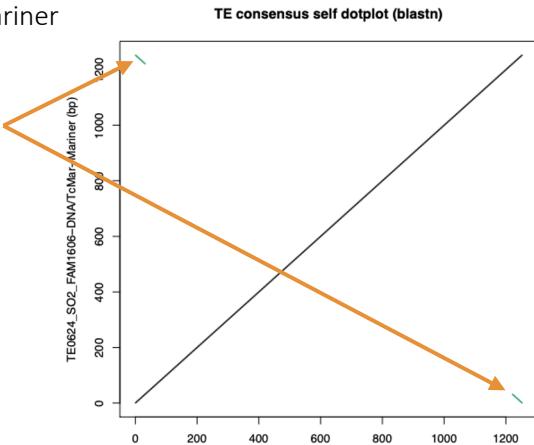


There are 5  
“long” copies  
(red lines)  
supporting  
the  
consensus

Short TIRs as seen  
in Tc1/Mariner  
elements



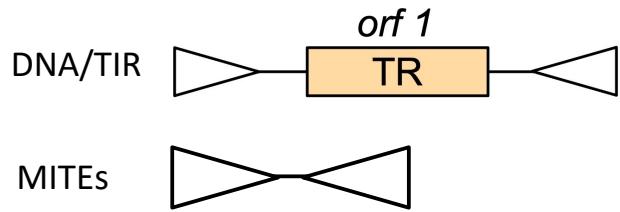
Most copies seem  
close to full length  
if not considering  
the break in the  
overlap



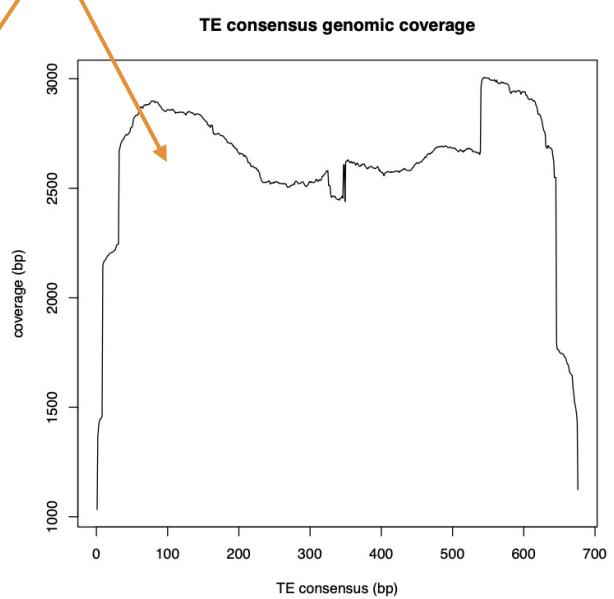
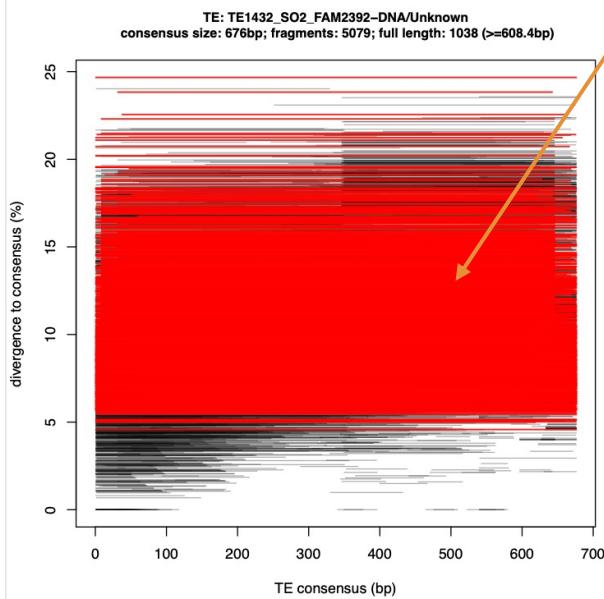
Ideally there should be only  
one transposase ORF. We can  
see that the bastp hit is almost  
continuous over the 2 ORFs

Canonical consensus: MITE (non-autonomous DNA/TIR)

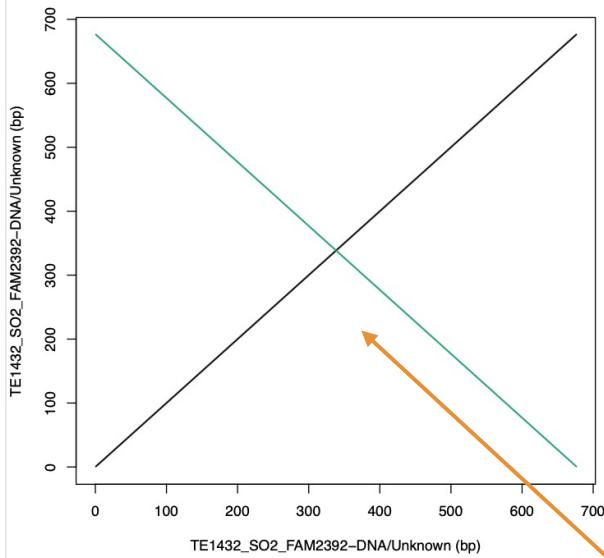
**Species:** *Sitophilus oryzae*  
(Insecta: Curculionidae)



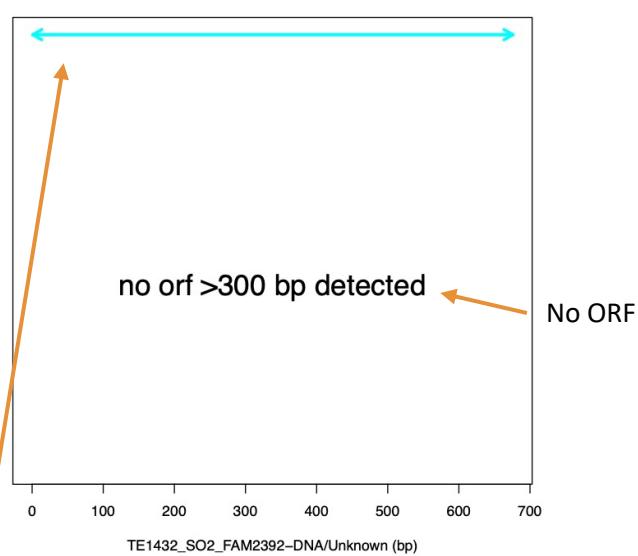
High copy number



TE consensus self dotplot (blastn)



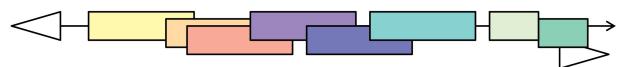
TE consensus structure and protein hits



TIRs create a palindrome structures

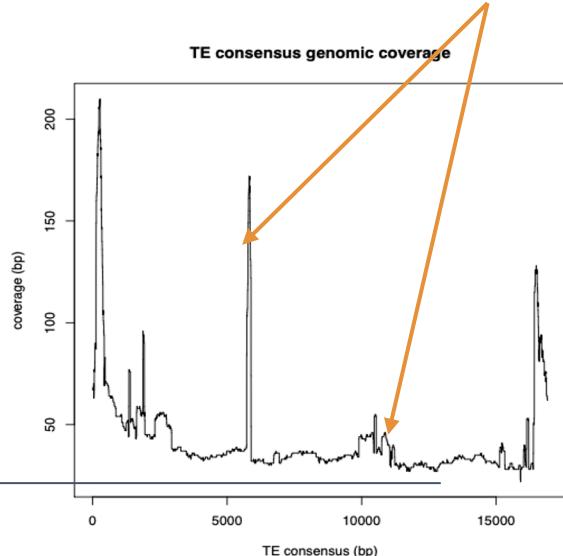
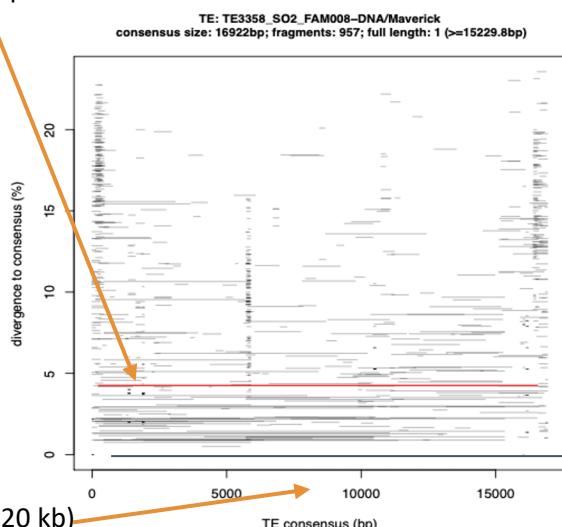
## Part III – Other examples (less common TEs or genomic repeats)

Other examples: Maverick/Polinton (Class II)

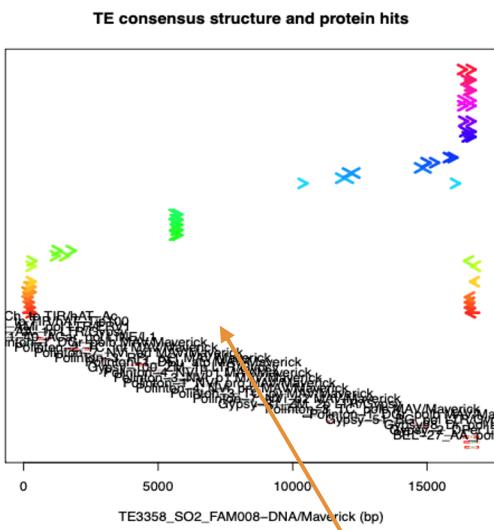
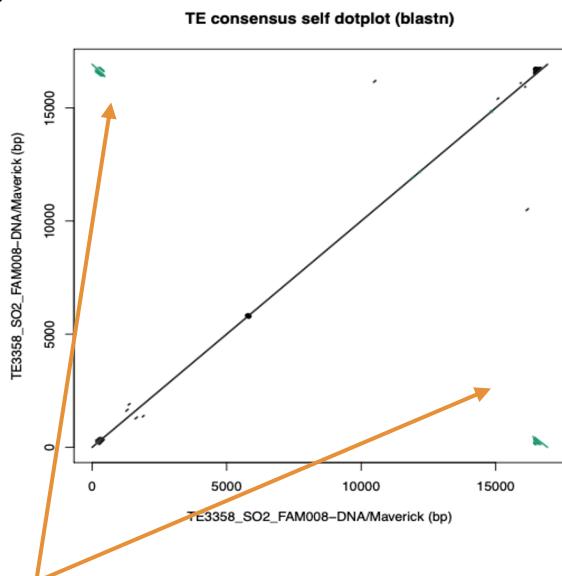


**Species:** *Sitophilus oryzae*  
(Insecta: Curculionidae)

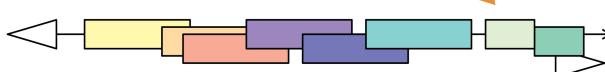
One full-length copy is present in this example



Large (10-20 kb)  
consensus



Flanking TIRs  
are present



The bases pileup of the hits show supports all the way through the consensus with some regions over-represented (includes hit overlaps)

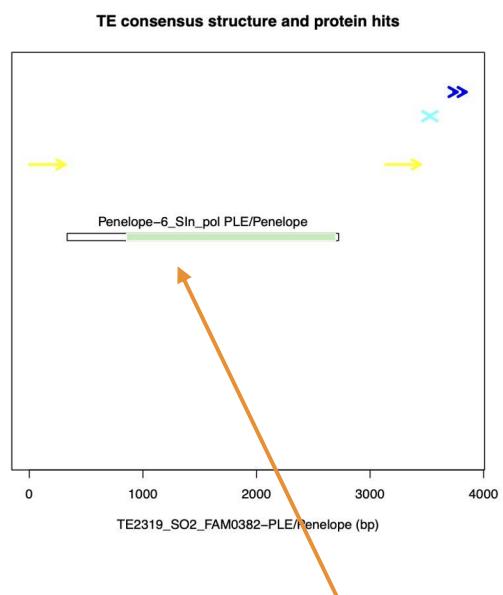
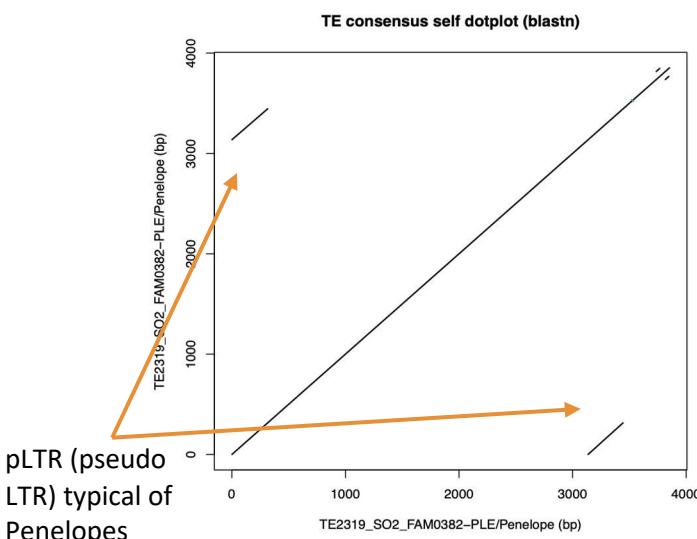
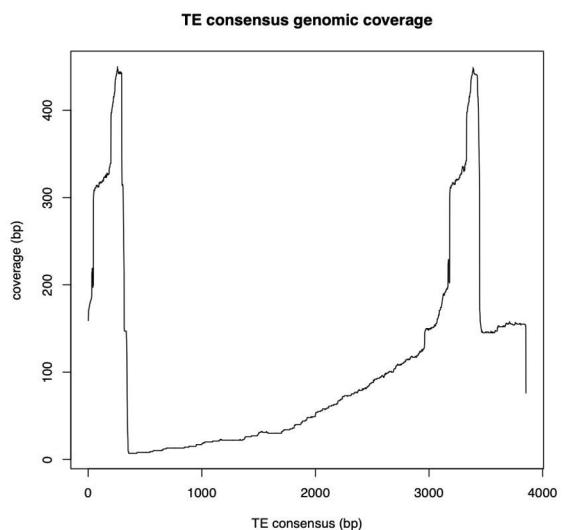
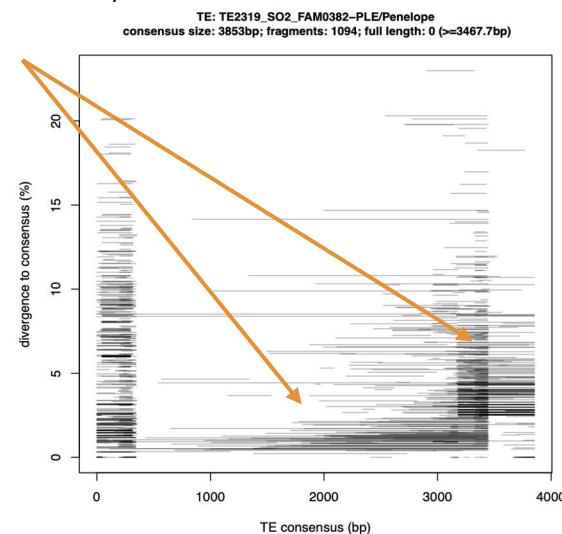
Multiple ORFs with Maverick/Polinton protein hits overlapping each other.

Other examples: Penelope (Class I)

**Species:** *Sitophilus oryzae*  
(Insecta: Curculionidae)

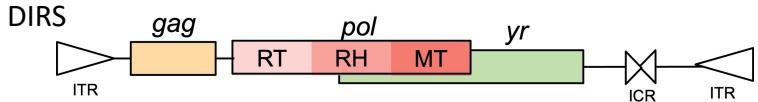


Typical 5'  
truncation  
(similar to LINEs)



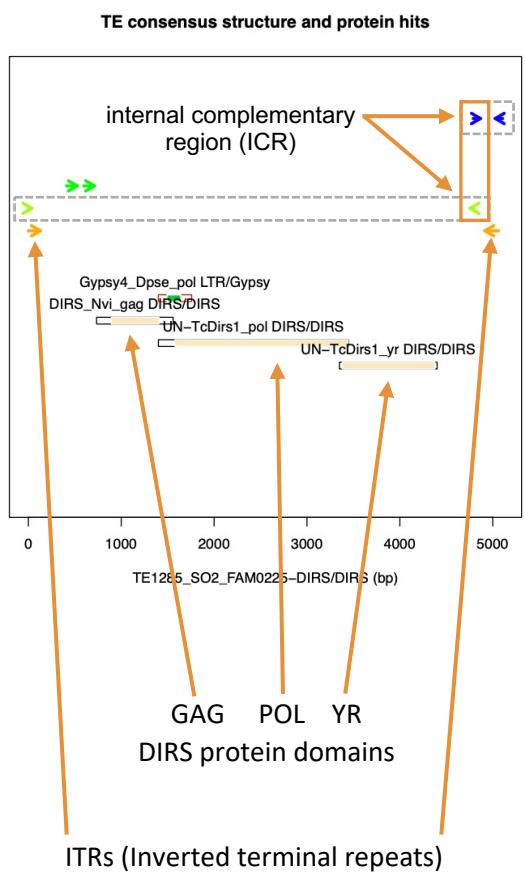
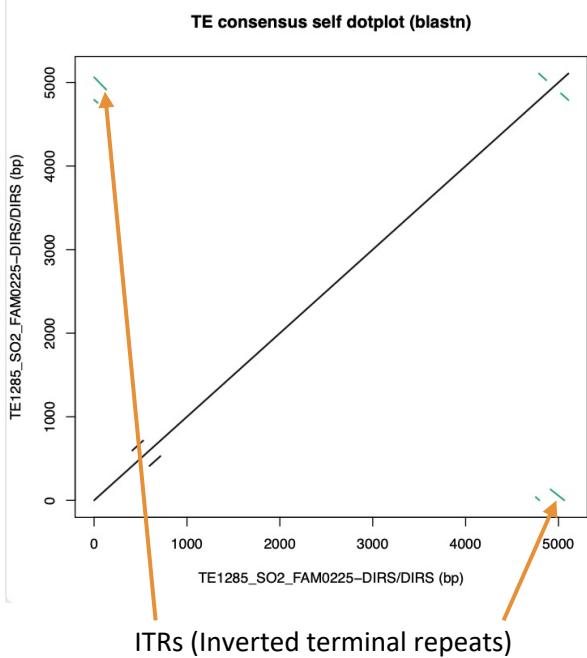
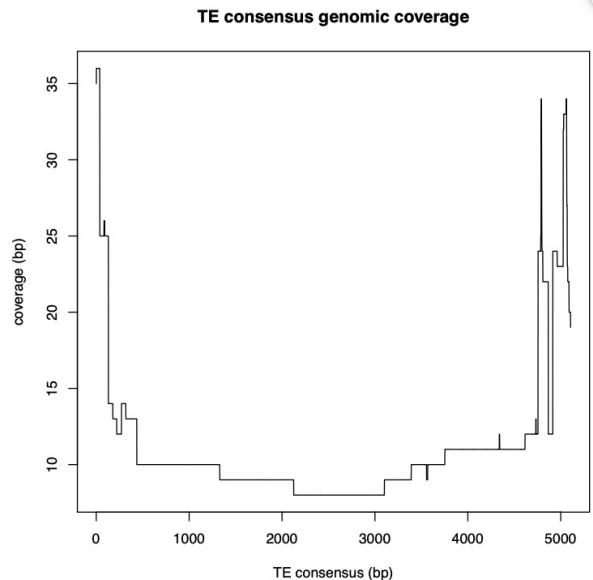
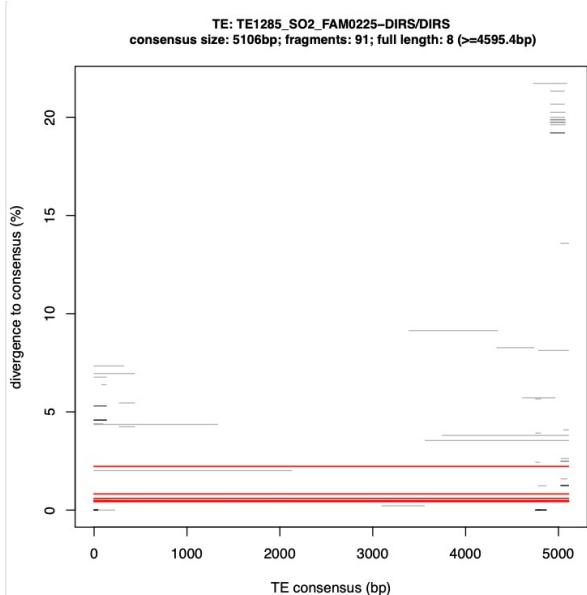
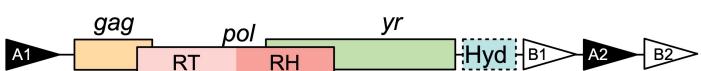
Single penelope ORF  
starting at the 5' pLTR

Other examples: DIRS (Class I)



**Species:** *Sitophilus oryzae* (Insecta: Curculionidae)

Ngaro

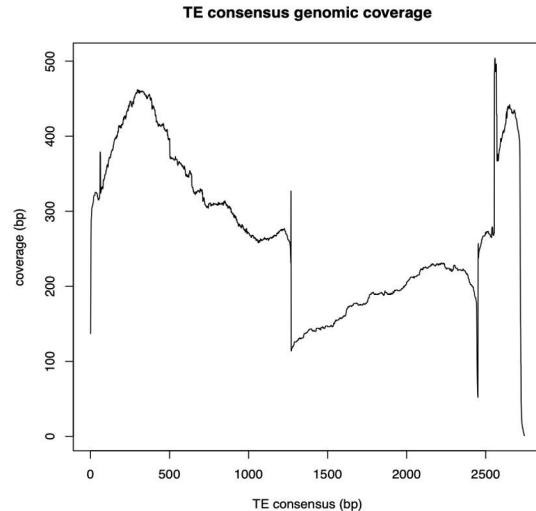
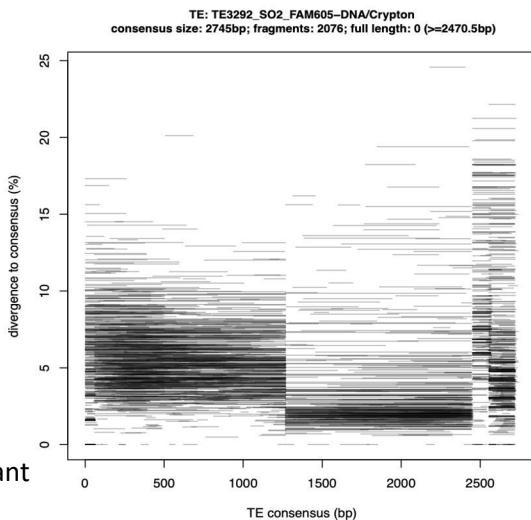


Other examples: Crypton

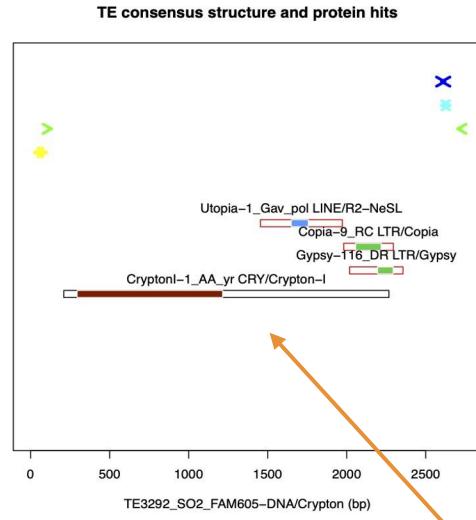
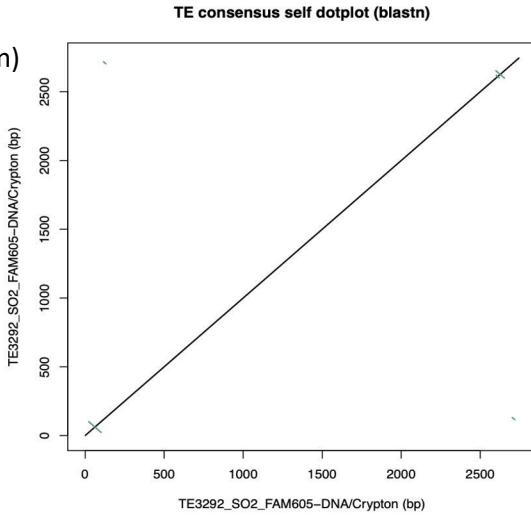
*orf 1*

YR

Species: *Sitophilus oryzae*  
(Insecta: Curculionidae)



No significant structural features, or very short TIRs (shown)

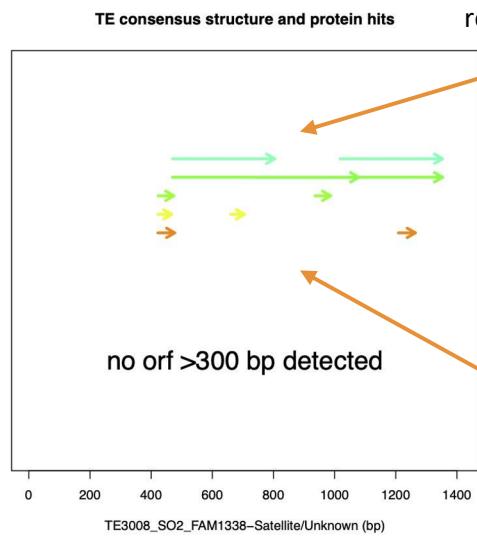
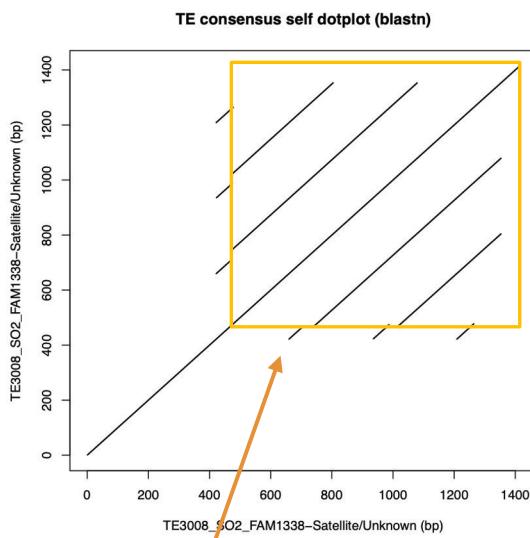
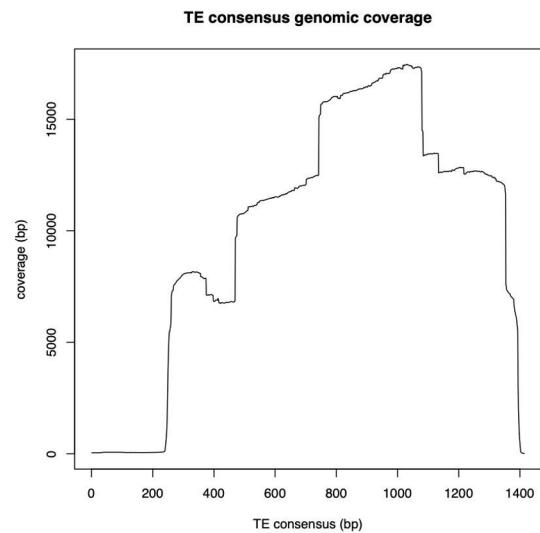
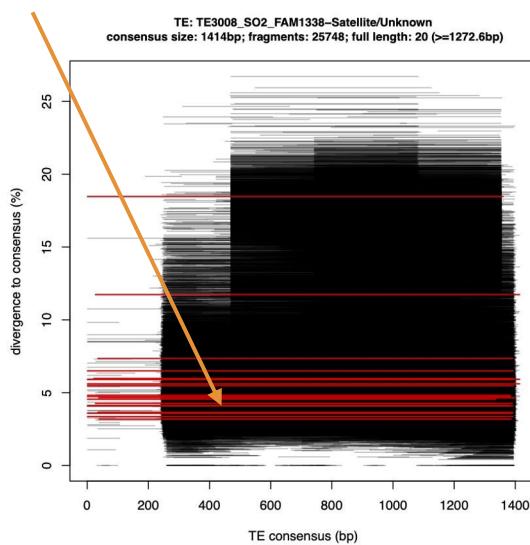


Single Crypton ORF as expected

## Other examples: Satellites

**Species:** *Sitophilus oryzae*  
 (Insecta: Curculionidae)

Strong support for full  
 Length consensus



Overlapping  
 repeats arrays

No ORFs

Characteristic satellite  
 dot-plot pattern (tandem  
 repeats arrays)

Tandem Repeat Finder output:

Period size: 275 Copynumber: 3.2 Consensus size: 270  
 Consensus pattern (270 bp):  
 AATCCGGTTGCCAAAAAATCAGCTTTTACGTCAGCGAGTTAACGAGAATCGGGATTAAATCAATG  
 ACCAATTTTTGAAACACCTGTATAATTAAAATCTTAAATATGTAAAAAAACAAAAATAAA  
 AATATAACGGTTTAAGCGTTTCCGAAAACAACCGTTTTAAATAAAAAAATATGTAAATT  
 TGAGGGGGTGGTTCACCCCTCTTGGAAATTGCAAAATCCCTTATCGAACCTCTACA  
 CCCTAAAGTC