Suplementary Material 2: Curation of the TE databases

Using the REPET pipeline v2.5 (Flutre et al., 2011), transposable elements (TEs) were predicted and annotated independently on both ESM015 (Japanese isolate) and EP155 (first reference isolate) genomes. From the prediction on the Japanese isolate curated assembly, 29 consensus TEs sequences were detected and used to annotate 3,670 partial and complete copies accounting for 2.49Mb (5.8 % of the genome). These consensus were manualy curated according to the autors specifications (see materials and methods). We suppressed five consensus with no full length copy annotated in the genome and eight redundant or incomplete consensus (see sup. Data. 2 for details). We cut a part of sequence in two consensus since it was only supported by one hsp. Last, we replaced two incomplete consensus of a gypsy element by only one, complete and detected in another *C. parasitica* genome not presented in this study. We finally kept 15 TE consensus sequences. This new TEs data base (TEs_db1) have been used as input TEs sequences for a second annotation of the ESM015 genome. 1,669 partial and complete copies accounting for 3.28Mb (7.6% of the genome) were annotated (Table 2). After the curation of a TEs database, we would expect to lose annotation power because we deleted some consensus. However, our annotation of the ESM015 genome using the curated TEs_db1 had higher TEs recovery rate than the first one, which confirms that this database is consistent. This increase is mainly due to the fact that we have replaced the two parts of a gypsy element (actually both ends) with the complete element detected in another genome (which carries the central part of the element).

The TE prediction performed independently on the EP155 genome lead to 28 TEs consensus (TEs_db2) used to annotate 2,510 partial and complete copies accounting for 4,98 Mb (11,34 % of the genome). The EP155 genome was also annotated using the TEs_db1 from ESM015, leading to 1,767 copies accounting for 3.72 Mb (8.5% of the genome). Intersecting both annotations allowed us to recover the copies annotated on the basis of consensus specific to TEs_db2. Three LARD (Large retrotransposon derivatives, class I) corresponded to the main part of the additionnal copies detected using TEs_db2 (1.21Mb out of 1.26Mb). The sequences of these three LARD TEs were added to TEs_db1 to create a final TEs data base (TEs_db3, available here : XXXXXXX), counting for 18 TEs consensus representative of both ESM015 and EP155 genomes.