A summary of the genome simulated by TEgenomeSimulator

Ting-Hsuan Chen

10 April, 2024

Contents

1 Simulation information

Simulated genome: tair10

CPU	RAM	Time	Timestamp_of_completion
1	426.00 MB	00:05:01	2024-03-28 13:04:34

In this simulation, TEgenomeSimulator did the followings:

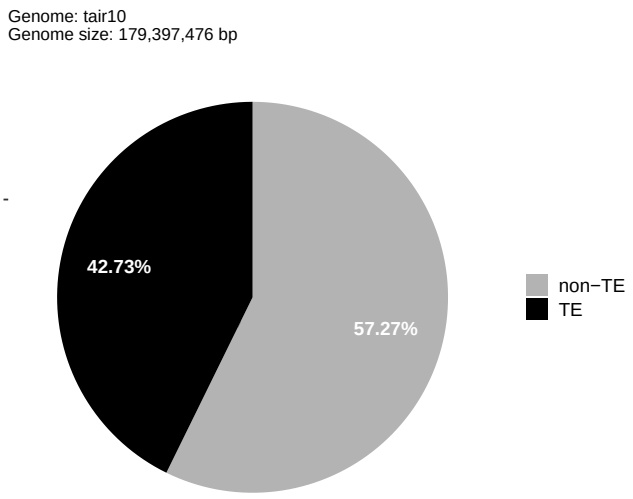
1. It took the TE library file, **athaTEref_ru1.classified.fasta**, which comprises curated TE family sequences from ***A. thalian***, ***Z. maize*** and ***O. sativa***, and simulates multiple TE copies with sequence variations depending on the parameters specified in the table: **TElib_sim_list_tair10.table**. In addition to nucleotide substitution and INDEL, it also simulated fragmentation (as a proportion of TE truncated from 5' end), nested insertion (only for Copia and Gypsy), as well as target site duplication.
2. The simulated TE copies were then randomly inserted into the user-provided TE-depleted genome, **AthaGenome.fa.nonTE**, where TE sequences had been exhaustively detected by multiple TE annotators (e.g. EDTA, RepeatModeler, and EarlGrey) and removed. The final simulated genome can be utilised for benchmarking TE annotators.

2 Loaded files for creating this report

- Genome fasta index file: tair10_genome_sequence_out_nest.fasta.fai
- All TE fasta index file: tair10_repeat_sequence_out_nest.fasta.fai
- All TE gff file: tair10_repeat_annotation_out_nest.gff

3 The proportion of the TE/nonTE sequence

- Total simulated genome size: 179,397,476 bp
- Total simulated TE bases: 76,659,960 bp



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_tair10_genome/report/

File name: 001_sim_genome_size_tair10.png

4 Breaking down the simulated TEs by superfamily

There are total 16 TE superfamilies included in the curated TE library. The following table shows the number of loci, bases and family of each superfamily, as well as the percentage of loci, bases and family.

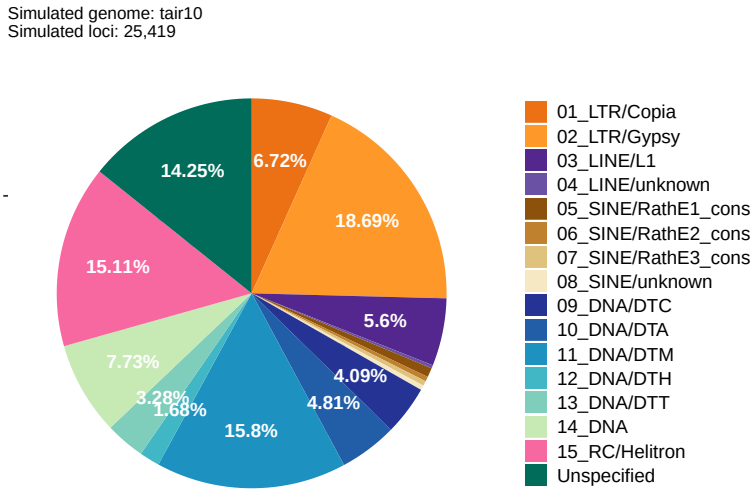
TE_superfamily	loci	bp	family_count	loci_percentage	bp_percentage	family_percentage
01_LTR/Copia	1708	6126854	105	6.72	7.99	32.61
02_LTR/Gypsy	4752	28025168	32	18.69	36.56	9.94
03_LINE/L1	1424	3945391	11	5.60	5.15	3.42
04_LINE/unknown	68	61100	1	0.27	0.08	0.31
05_SINE/RathE1_cons	195	29019	1	0.77	0.04	0.31
06_SINE/RathE2_cons	101	26558	1	0.40	0.03	0.31
07_SINE/RathE3_cons	102	15608	1	0.40	0.02	0.31
08_SINE/unknown	100	36789	2	0.39	0.05	0.62
09_DNA/DTC	1039	5691914	12	4.09	7.42	3.73
10_DNA/DTA	1223	1235984	22	4.81	1.61	6.83
11_DNA/DTM	4015	18278575	70	15.80	23.84	21.74
12_DNA/DTH	428	1179996	3	1.68	1.54	0.93
13_DNA/DTT	833	314640	7	3.28	0.41	2.17
14_DNA	1966	1384591	12	7.73	1.81	3.73
15_RC/Helitron	3842	9458794	34	15.11	12.34	10.56
Unspecified	3623	852838	8	14.25	1.11	2.48
Total	25419	76663819	322	100.00	100.00	100.00

Table saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_tair10_genome/report/

File name: sim_genome_summary_TEsuperfamily_tair10.csv

4.1 Total simulated TE loci categorised by TE superfamily

- Total simulated TE loci: 25,419 loci

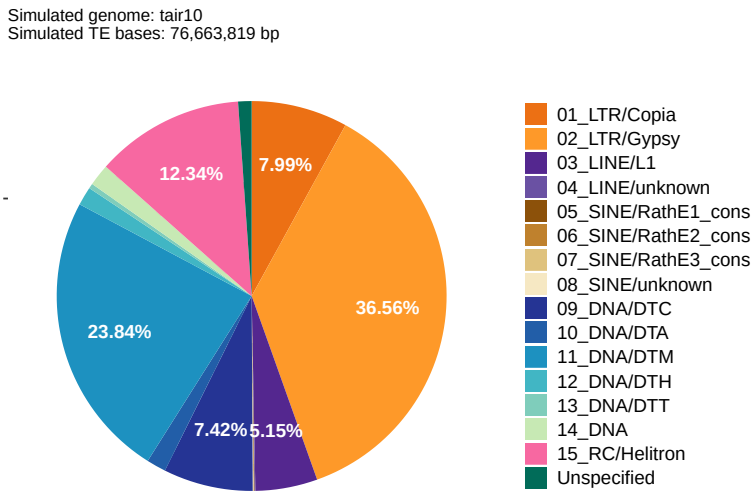


High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_tair10_genome/report/

File name: 002_sim_te_loci_tair10.png

4.2 Total simulated TE bases categorised by TE superfamily

- Total simulated TE bases: 76,663,819 bp



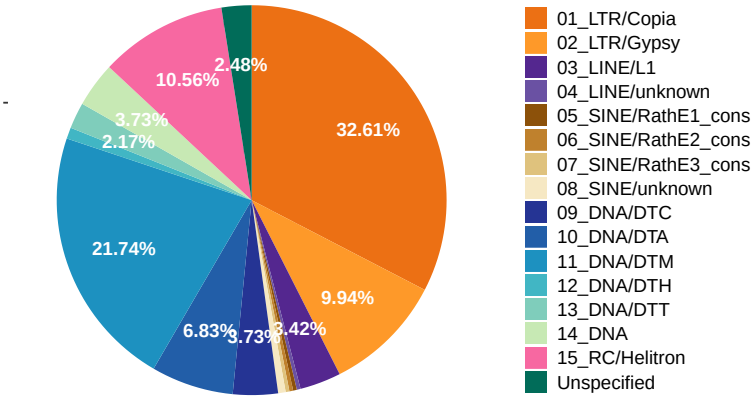
High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_tair10_genome/report/

File name: 003_sim_te_bp_tair10.png

4.3 Simulated TE family per superfamily

This part only depends on the curated TE library. It shouldn't make any difference between simulations.

- Total simulated TE families: 322 families



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_tair10_genome/report/

File name: 004_sim_te_family_tair10.png

5 Extracting full information from TE's gff file

Have a look at the full info extracted from TE's gff file (row 101 to 110):

	Chr	V3	Start	End	strand	TE_rename
101	Chr1	RC	700261	710483	+	15_RC/Helitron
102	Chr1	LINE	716059	720385	+	03_LINE/L1
103	Chr1	SINE	728335	728677	-	08_SINE/unknown
104	Chr1	DNA	731674	746083	+	11_DNA/DTM
105	Chr1	LTR	749545	754584	-	01_LTR/Copia
106	Chr1	DNA	756024	756463	+	14_DNA
107	Chr1	DNA	763365	763731	+	10_DNA/DTA
108	Chr1	DNA	764604	764861	+	13_DNA/DTT
109	Chr1	DNA	776075	780921	+	11_DNA/DTM
110	Chr1	DNA	781715	795968	+	11_DNA/DTM

	ID	TE_id	TE_family
101	HELITRONY2#RC/Helitron_TE0000095	TE0000095	HELITRONY2
102	ATLINE1_1#LINE/L1_TE0000096	TE0000096	ATLINE1_1
103	ATSINE4#SINE_TE0000097	TE0000097	ATSINE4
104	VANDAL4#DNA/MuDR_TE0000098	TE0000098	VANDAL4
105	ATCOPIA33#LTR/Copia_TE0000099	TE0000099	ATCOPIA33
106	ATDNA12T3_2#DNA_TE0000100	TE0000100	ATDNA12T3_2
107	ATHATN4#DNA/HAT_TE0000101	TE0000101	ATHATN4
108	ATHPOGON1#DNA/Pogo_TE0000102	TE0000102	ATHPOGON1
109	VANDAL13#DNA/MuDR_TE0000103	TE0000103	VANDAL13
110	ARNOLD1#DNA/MuDR_TE0000104	TE0000104	ARNOLD1

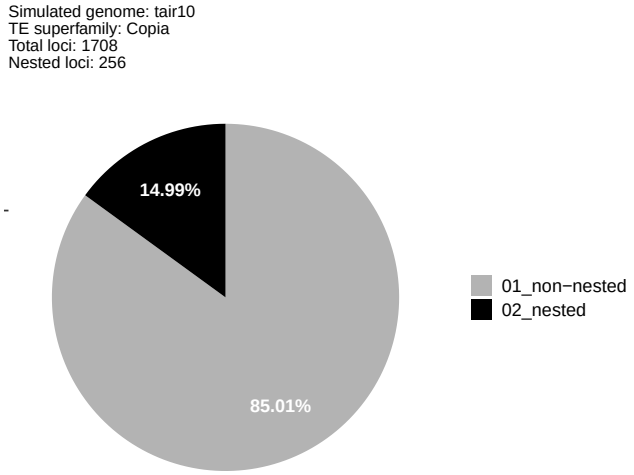
	Subclass	Identity	Integrity	Nested_in	Cut_at	Cut_by
101	RC	0.73	0.92	NA	NA	NA
102	LINE	0.68	0.74	NA	NA	NA
103	SINE	0.79	1.0	NA	NA	NA
104	DNA	0.78	0.78	NA	NA	NA
105	LTR	0.73	1.0	NA	NA	NA
106	DNA	0.82	0.53	NA	NA	NA
107	DNA	0.73	0.83	NA	NA	NA
108	DNA	0.79	0.47	NA	NA	NA
109	DNA	0.84	0.4	NA	NA	NA
110	DNA	0.8	0.9	NA	NA	NA

Full table saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_tair10_genome/report/

File name: sim_genome_TE_insertion_info_tair10.csv

6 Simulation of nested TE insertions

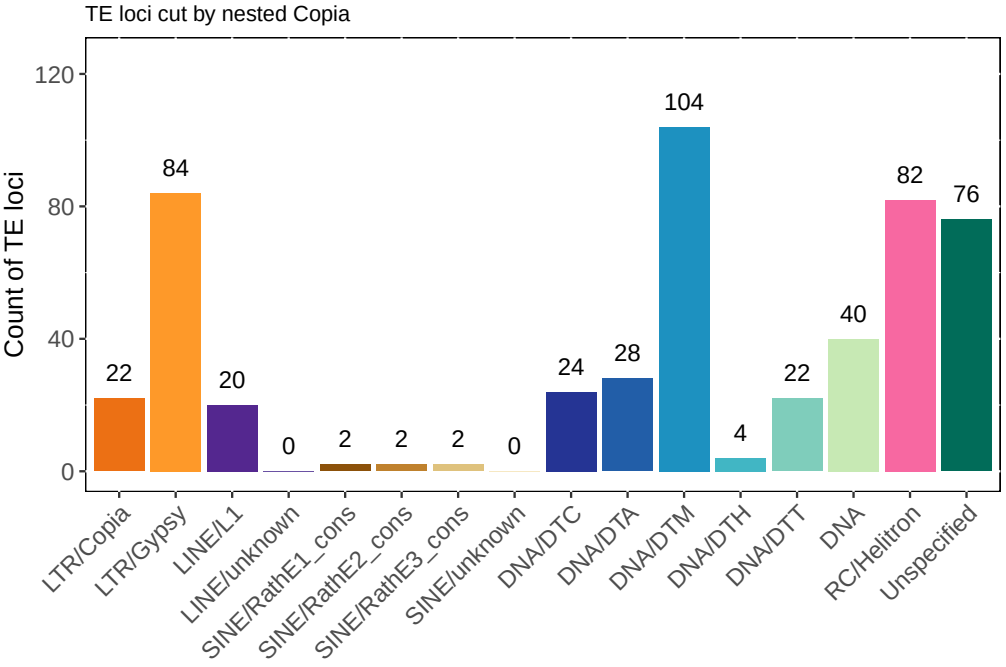
6.1 Nested and non-nested Copia



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_tair10_genome/report/

File name: 005_sim_te_nested_copia_tair10.png

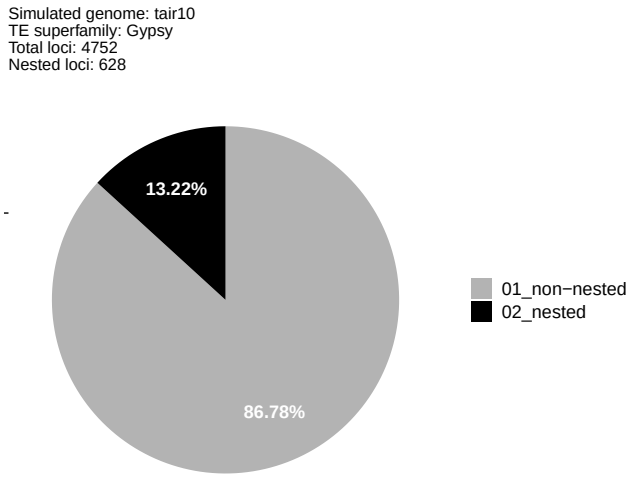
6.2 TE loci cut by nested Copia



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_tair10_genome/report/

File name: 006_sim_te_cutby_nested_copia_tair10.png

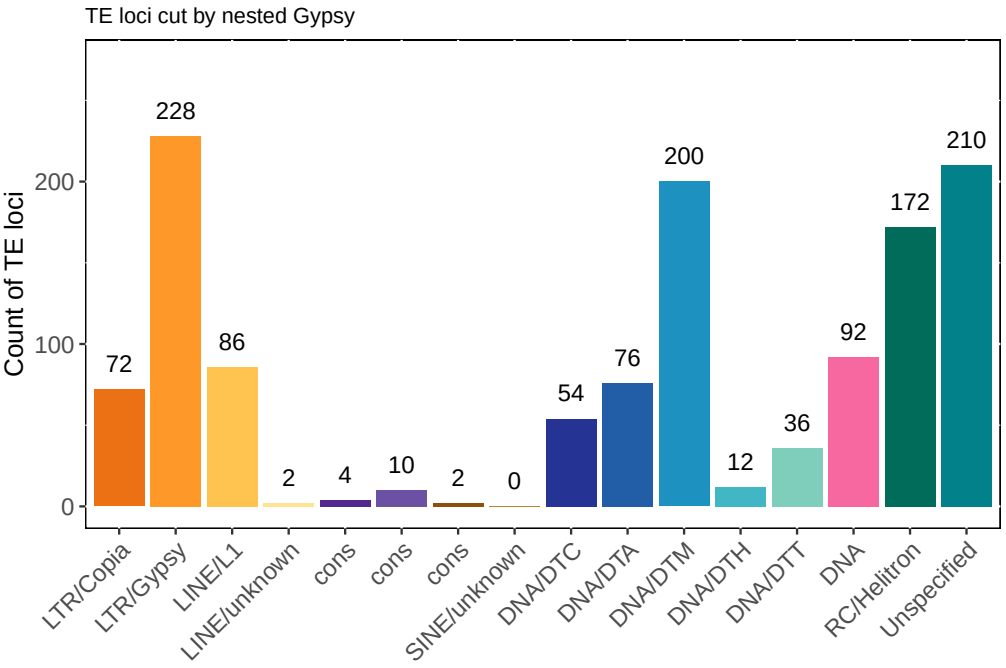
6.3 Nested and non-nested Gypsy



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_tair10_genome/report/

File name: 007_sim_te_nested_gypsy_tair10.png

6.4 TE loci cut by nested Gypsy



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_tair10_genome/report/

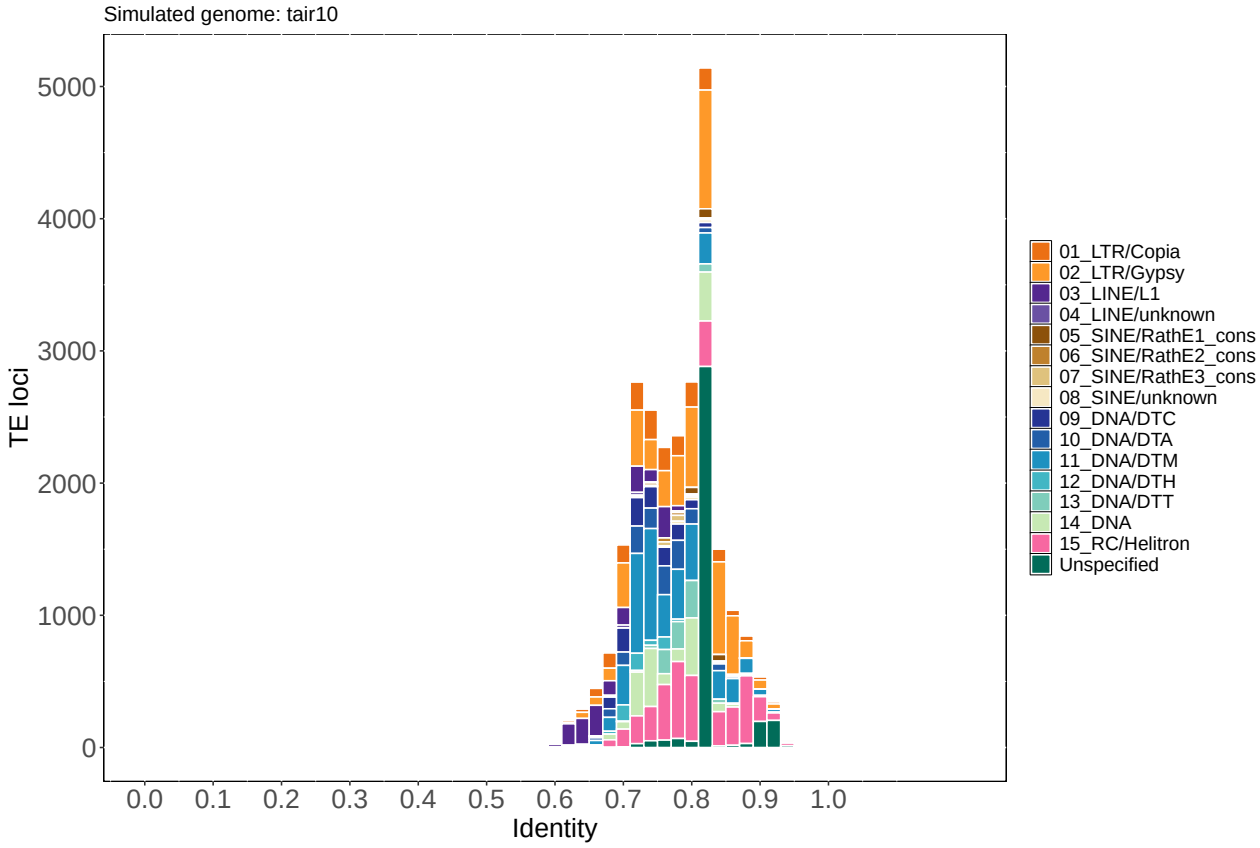
File name: 008_sim_te_cutby_nested_gypsy_tair10.png

7 The distribution of TE loci identity

How did TEgenomeSimulator simulate sequence identity in this simulation?

1. The distribution of sequence identity of each TE family was defined by the **idn** and **sd** values (user provided) stored in **TElib_sim_list_tair10.table**.
2. TEgenomeSimulator took the **idn** as mean identity and **sd** as standard deviation to create a distrubution, from which a value was sampled as the **simulated identity** of a TE member. Therefore, **simulated divergence = 1 - simulated identity**
3. Inherited from it's predecessor, denovoTE-eval, TEgenomSimulator broke down the **simulated divergence** into **substitution** and **INDELs** (i.e. **divergence = substitution % + INDEL %**).
4. The **INDEL %** was defined by the **indels** value from **TElib_sim_list_tair10.table**.

```
## Warning: Removed 32 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```




```
## Warning: Removed 32 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```

High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_tair10_genome/report/

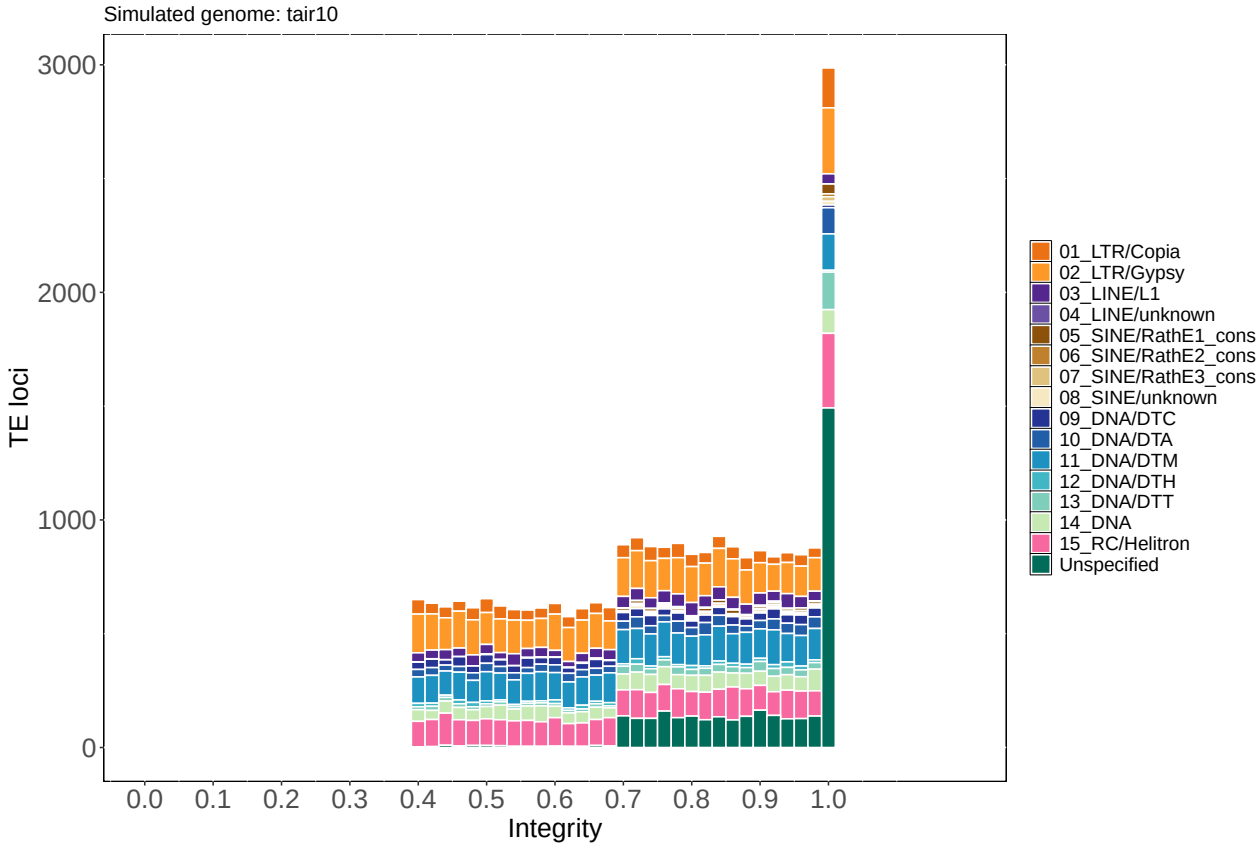
File name: 009_sim_te_loci_identity_tair10.png

8 The distribution of TE loci integrity

How did TEgenomeSimulator simulate sequence integrity in this simulation?

1. TEgenomeSimulator considers sequence integrity as **integrity = (1 - (TE locus length / full length))**
2. The length of a TE locus is decided by INDELs and fragmentation.
3. In the fragmentation step, if a TE length is shorter than 500 bp, TEgenomeSimulator would randomly select a value between 70 and 90 as the fraction of the sequence to be removed from 5' end; otherwise the simulator randomly chooses a value between 40 and 90. (This is the same as in denovoTE-eval)
4. The number of fragmented loci was defined by the value of **frag** in **TElib_sim_list_tair10.table**. This value was taken as a proportion of total TE loci to undergo fragmentation step.

```
## Warning: Removed 32 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```




```
## Warning: Removed 32 rows containing missing values or values outside the scale range
## (`geom_bar()`).
```

High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_tair10_genome/report/

File name: 010_sim_te_loci_integrity_tair10.png