A summary of the genome simulated by TEgenomeSimulator

Ting-Hsuan Chen

05 March, 2024

Contents

1 Simulation information

Simulated genome: donghong
Range of simulated TE loci per TE family: 5-500

CPU	RAM	Time	Timestamp_of_completion
1	2.81 GB	00:56:30	2024-02-21 14:58:41

In this simulation, TEgenomeSimulator did the followings:

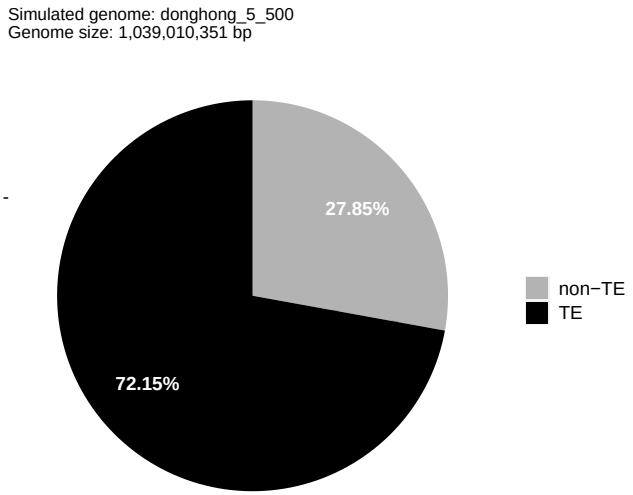
1. It took the TE library file, **combined_curated_TE_lib_ATOSZM_selected.fasta**, which comprises curated TE family sequences from ***A. thalian***, ***Z. maize*** and ***O. sativa***, and simulates multiple TE copies with sequence variations depending on the parameters specified in the table: **TElib_sim_list_5_500.table**. 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, **Donghong.chromosomes.only.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: donghong_5_500_genome_sequence_out_nest.fasta.fai
- All TE fasta index file: donghong_5_500_repeat_sequence_out_nest.fasta.fai
- All TE gff file: donghong_5_500_repeat_annotation_out_nest.gff

3 The proportion of the TE/nonTE sequence

- Total simulated genome size: 1,039,010,351 bp
- Total simulated TE bases: 749,611,011 bp



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_500_genome/report/

File name: 001_sim_genome_size_donghong_5_500.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.

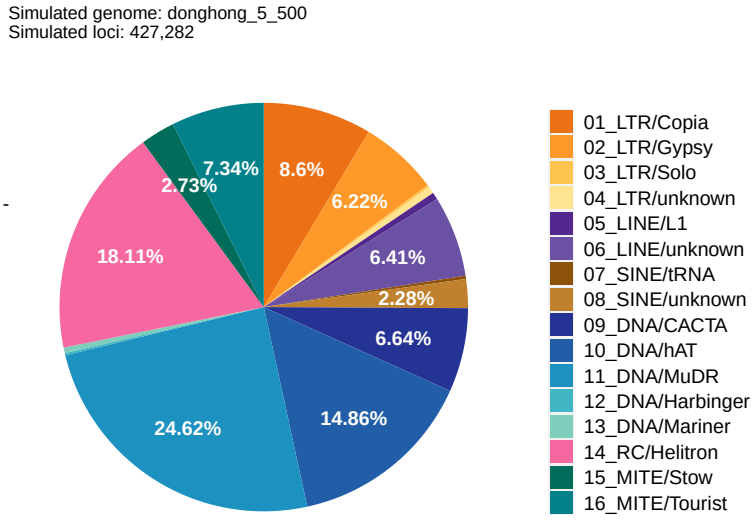
TE_superfamily	loci	bp	family_count	loci_percentage	bp_percentage	family_percentage
01_LTR/Copia	36734	106558750	139	8.60	14.21	8.44
02_LTR/Gypsy	26583	87755088	109	6.22	11.71	6.62
03_LTR/Solo	787	828294	3	0.18	0.11	0.18
04_LTR/unknown	2380	4391552	11	0.56	0.59	0.67
05_LINE/L1	2441	8256516	12	0.57	1.10	0.73
06_LINE/unknown	27386	55213067	94	6.41	7.37	5.71
07_SINE/tRNA	1121	229470	5	0.26	0.03	0.30
08_SINE/unknown	9754	4389678	43	2.28	0.59	2.61
09_DNA/CACTA	28385	78582460	110	6.64	10.48	6.68
10_DNA/hAT	63474	52608504	249	14.86	7.02	15.12
11_DNA/MuDR	105190	193827411	393	24.62	25.86	23.86
12_DNA/Harbinger	776	1738595	3	0.18	0.23	0.18
13_DNA/Mariner	1868	1018444	7	0.44	0.14	0.43
14_RC/Helitron	77399	144441702	314	18.11	19.27	19.06
15_MITE/Stow	11651	2314509	40	2.73	0.31	2.43
16_MITE/Tourist	31353	7496326	115	7.34	1.00	6.98
Total	427282	749650366	1647	100.00	100.00	100.00

Table saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_500_genome/report/

File name: sim_genome_summary_TEsuperfamily_donghong_5_500.csv

4.1 Total simulated TE loci categorised by TE superfamily

- Total simulated TE loci: 427,282 loci

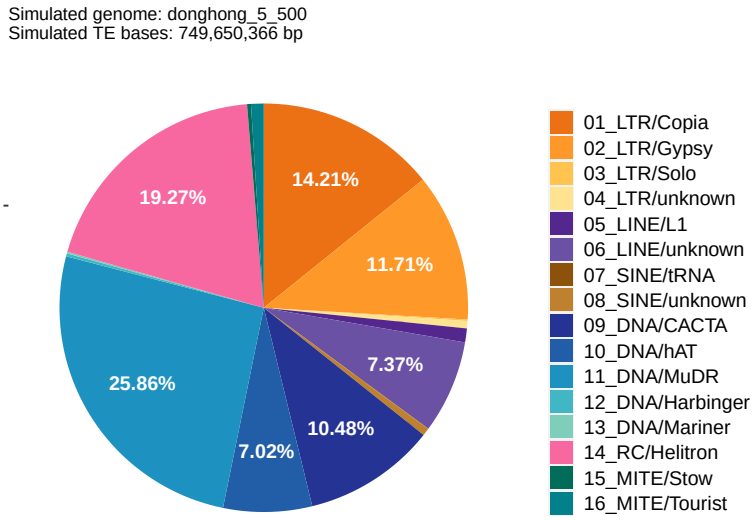


High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_500_genome/report/

File name: 002_sim_te_loci_donghong_5_500.png

4.2 Total simulated TE bases categorised by TE superfamily

- Total simulated TE bases: 749,650,366 bp



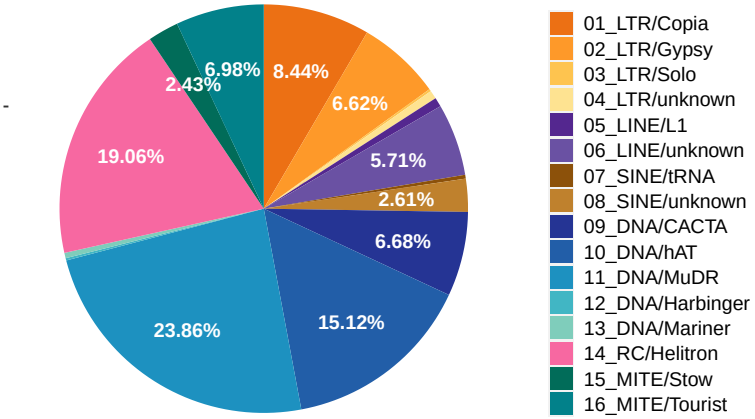
High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_500_genome/report/

File name: 003_sim_te_bp_donghong_5_500.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: 1647 families



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_500_genome/report/

File name: 004_sim_te_family_donghong_5_500.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	Start	End	strand	TE_rename	ID
101	chr1	225186	228549	-	01_LTR/Copia	ATCOPIA14#LTR/Copia_TE0000101
102	chr1	229323	230135	+	14_RC/Helitron	Os1361#DNAnona/Helitron_TE0000102
103	chr1	232050	232201	+	11_DNA/MuDR	Os0536#DNAnona/MULE_TE0000103
104	chr1	232626	232738	-	16_MITE/Tourist	Os0594#MITE/Tourist_TE0000104
105	chr1	234102	234549	+	11_DNA/MuDR	Os0610#DNAnona/MULE_TE0000105
106	chr1	235662	239413	-	14_RC/Helitron	Os1839#DNAauto/Helitron_TE0000106
107	chr1	239574	240366	-	10_DNA/hAT	Os2127#DNAnona/hAT_TE0000107
108	chr1	243840	249477	+	09_DNA/CACTA	ATENSPM9#DNA/CACTA_TE0000108
109	chr1	250269	250536	-	15_MITE/Stow	Os0112#MITE/Stow_TE0000109
110	chr1	250560	250975	-	10_DNA/hAT	ATHATN3#DNA/hAT_TE0000110

	TE_id	TE_family	Subclass
101	TE0000101	ATCOPIA14	LTR_retrotransposon
102	TE0000102	Os1361	Helitron
103	TE0000103	Os0536	TIR_transposon
104	TE0000104	Os0594	MITE
105	TE0000105	Os0610	TIR_transposon
106	TE0000106	Os1839	Helitron
107	TE0000107	Os2127	TIR_transposon
108	TE0000108	ATENSPM9	TIR_transposon
109	TE0000109	Os0112	MITE
110	TE0000110	ATHATN3	TIR_transposon

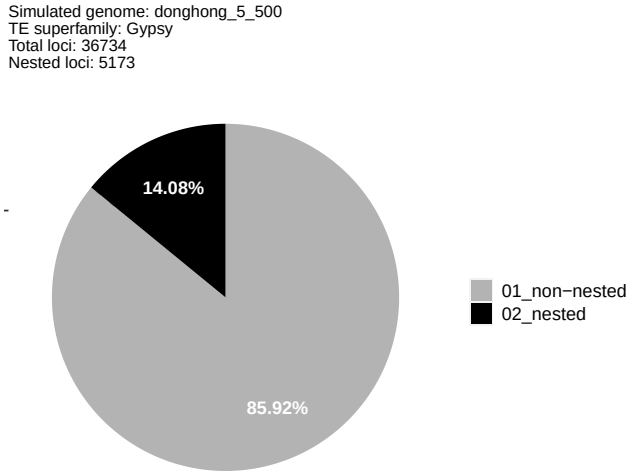
	Identity	Integrity	Nested_in	Cut_at	Cut_by	chrid
101	0.65	0.68	NA	NA	NA	1
102	0.81	0.89	NA	NA	NA	1
103	0.54	0.97	NA	NA	NA	1
104	0.9	0.94	NA	NA	NA	1
105	0.8	0.66	NA	NA	NA	1
106	0.76	0.43	NA	NA	NA	1
107	0.89	0.8	NA	NA	NA	1
108	0.73	0.61	NA	NA	NA	1
109	1.0	1.0	NA	NA	NA	1
110	0.97	0.75	NA	NA	NA	1

Full table saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_500_genome/report/

File name: sim_genome_TE_insertion_info_donghong_5_500.csv

6 Simulation of nested TE insertions

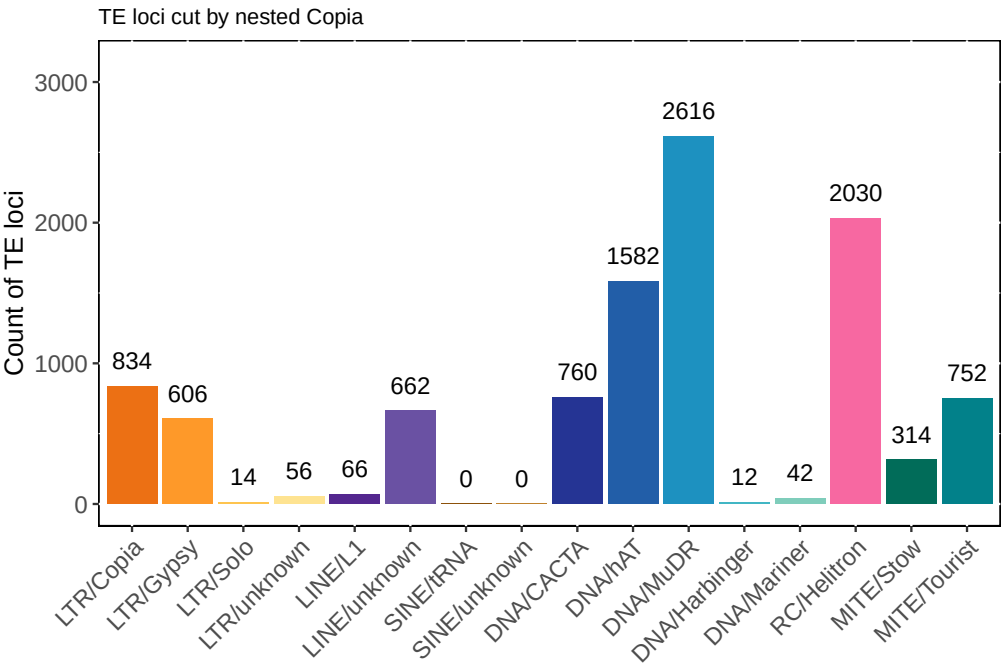
6.1 Nested and non-nested Copia



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_500_genome/report/

File name: 005_sim_te_nested_copia_donghong_5_500.png

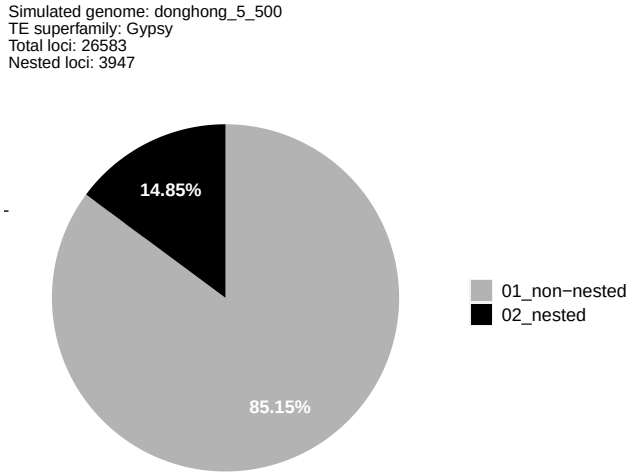
6.2 TE loci cut by nested Copia



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_500_genome/report/

File name: 006_sim_te_cutby_nested_copia_donghong_5_500.png

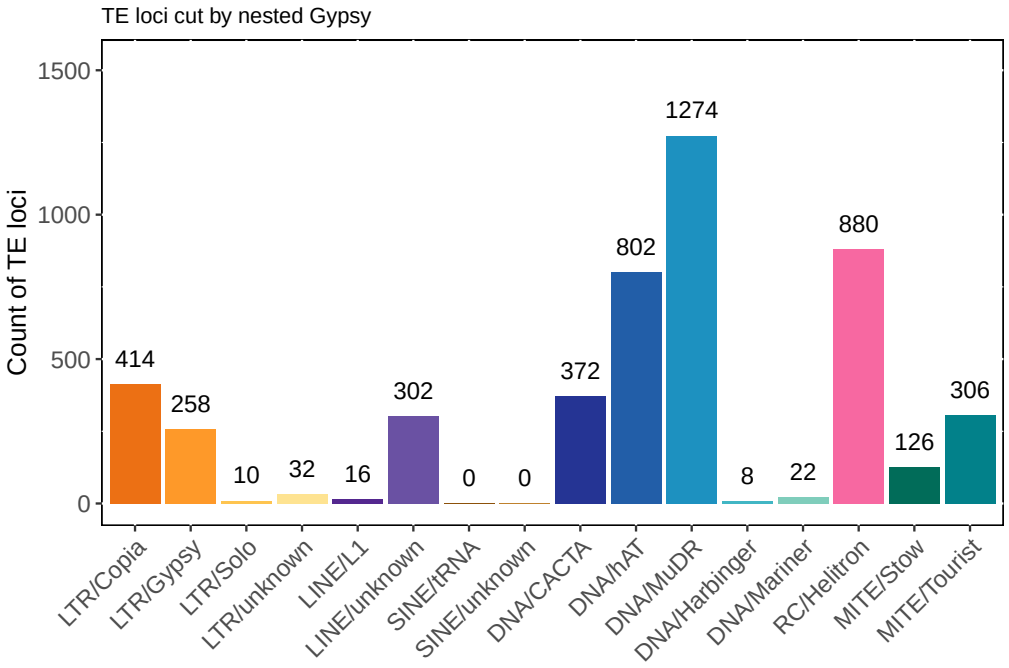
6.3 Nested and non-nested Gypsy



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_500_genome/report/

File name: 007_sim_te_nested_gypsy_donghong_5_500.png

6.4 TE loci cut by nested Gypsy



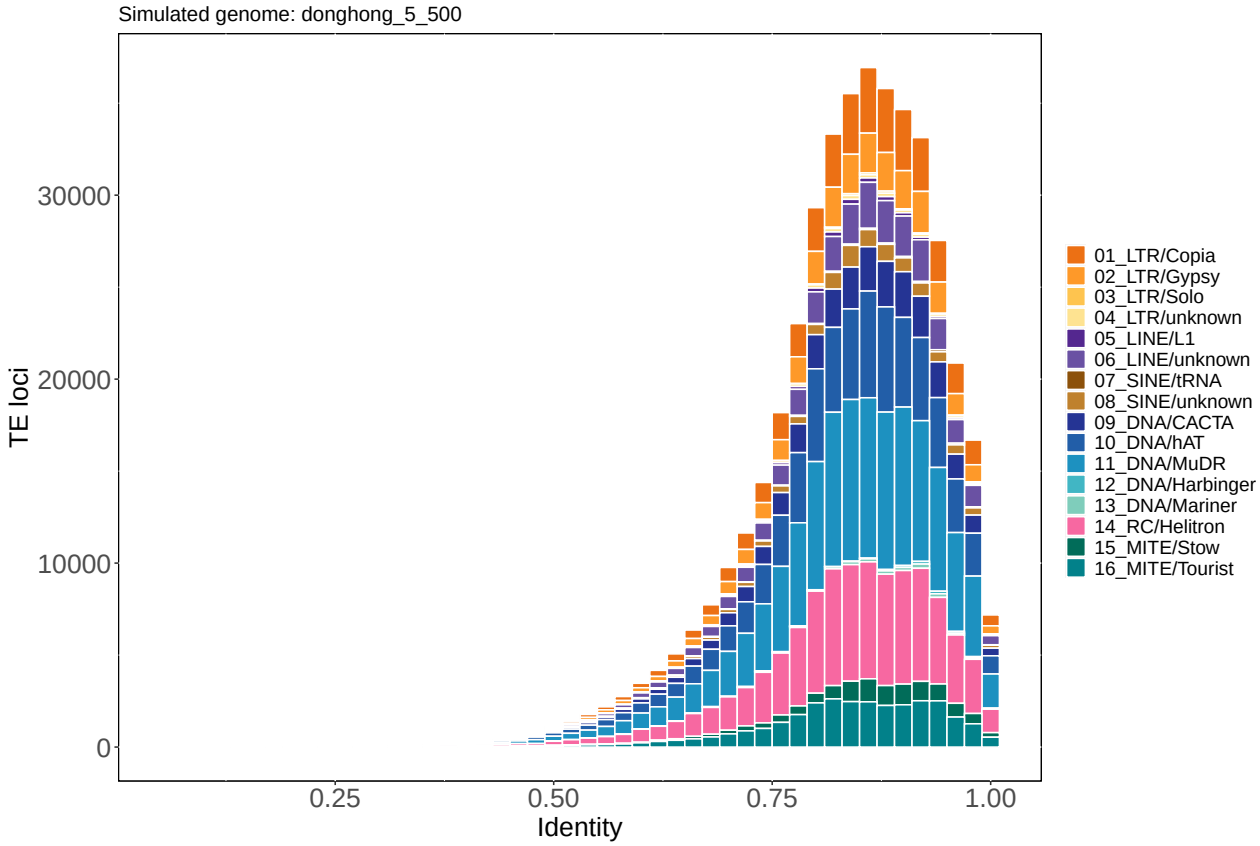
High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_500_genome/report/

File name: 008_sim_te_cutby_nested_gypsy_donghong_5_500.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_5_500.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_5_500.table**.



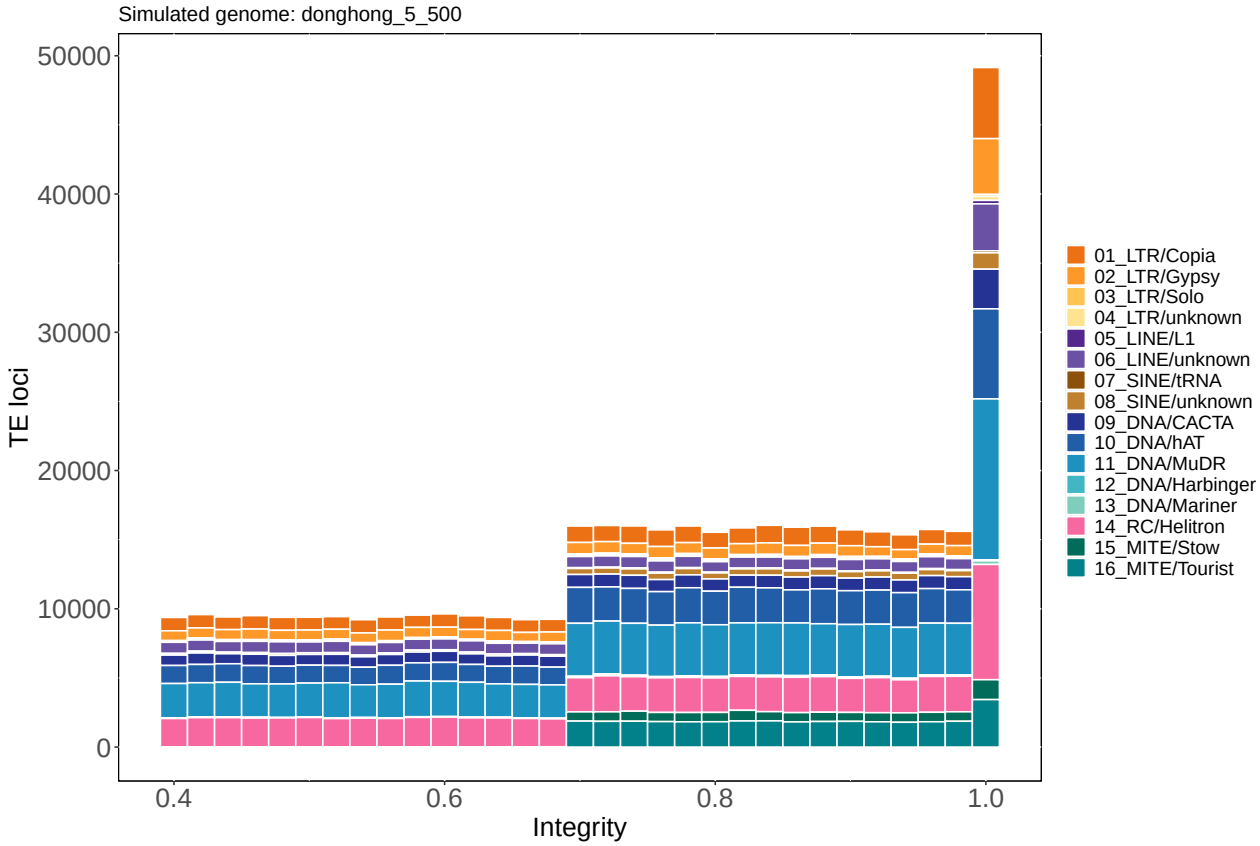
High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_500_genome/report/

File name: 009_sim_te_loci_identity_donghong_5_500.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_5_500.table**. This value was taken as a proportion of total TE loci to undergo fragmentation step.



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_500_genome/report/

File name: 010_sim_te_loci_integrity_donghong_5_500.png