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: 5-10

CPU	RAM	Time	Timestamp_of_completion
1	842.30 MB	00:00:51	2024-02-21 14:14:22

In this simulation, TEgenomeSimulator did the followings:

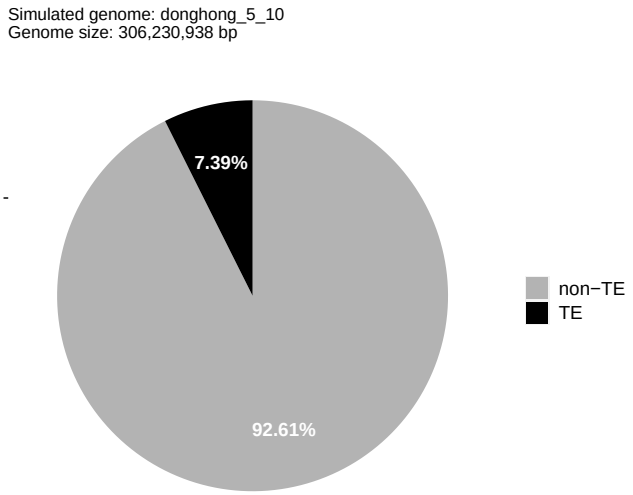
1. It took the TE library file, **combined_curated_TE_lib_ATOSZM_selected.fasta**, which comprises curated TE family sequences from ***A. thalian***, ***Z. maize*** and ***O. sativa***, and simulates multiple TE copies with sequence variations depending on the parameters specified in the table: **TElib_sim_list_5_10.table**. 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_10_genome_sequence_out_nest.fasta.fai
- All TE fasta index file: donghong_5_10_repeat_sequence_out_nest.fasta.fai
- All TE gff file: donghong_5_10_repeat_annotation_out_nest.gff

3 The proportion of the TE/nonTE sequence

- Total simulated genome size: 306,230,938 bp
- Total simulated TE bases: 22,638,471 bp



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_10_genome/report/

File name: 001_sim_genome_size_donghong_5_10.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.

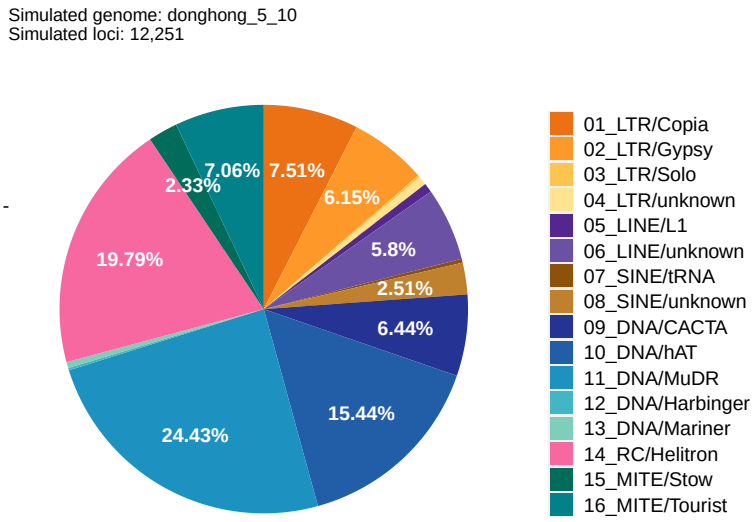
TE_superfamily	loci	bp	family_count	loci_percentage	bp_percentage	family_percentage
01_LTR/Copia	920	2907611	139	7.51	12.84	8.44
02_LTR/Gypsy	753	2476491	109	6.15	10.94	6.62
03_LTR/Solo	21	28080	3	0.17	0.12	0.18
04_LTR/unknown	84	110541	11	0.69	0.49	0.67
05_LINE/L1	89	312729	12	0.73	1.38	0.73
06_LINE/unknown	710	1470449	94	5.80	6.50	5.71
07_SINE/tRNA	38	8237	5	0.31	0.04	0.30
08_SINE/unknown	307	153493	43	2.51	0.68	2.61
09_DNA/CACTA	789	2109253	110	6.44	9.32	6.68
10_DNA/hAT	1891	1627563	249	15.44	7.19	15.12
11_DNA/MuDR	2993	6228941	393	24.43	27.51	23.86
12_DNA/Harbinger	23	68688	3	0.19	0.30	0.18
13_DNA/Mariner	58	32961	7	0.47	0.15	0.43
14_RC/Helitron	2424	4835648	314	19.79	21.36	19.06
15_MITE/Stow	286	60556	40	2.33	0.27	2.43
16_MITE/Tourist	865	207940	115	7.06	0.92	6.98
Total	12251	22639181	1647	100.00	100.00	100.00

Table saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_10_genome/report/

File name: sim_genome_summary_TEsuperfamily_donghong_5_10.csv

4.1 Total simulated TE loci categorised by TE superfamily

- Total simulated TE loci: 12,251 loci

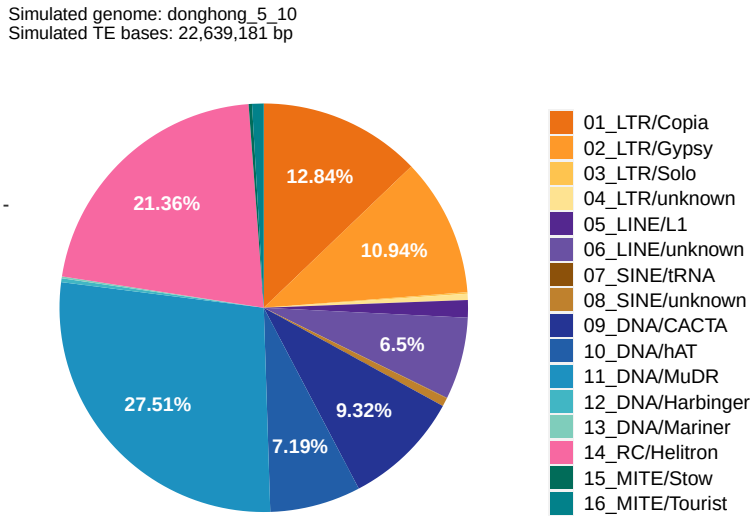


High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_10_genome/report/

File name: 002_sim_te_loci_donghong_5_10.png

4.2 Total simulated TE bases categorised by TE superfamily

- Total simulated TE bases: 22,639,181 bp



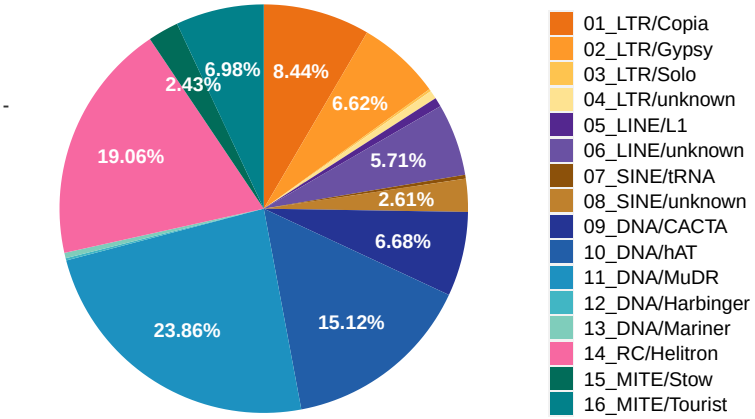
High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_10_genome/report/

File name: 003_sim_te_bp_donghong_5_10.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_10_genome/report/

File name: 004_sim_te_family_donghong_5_10.png

5 Extracting full information from TE's gff file

Have a look at the full info extracted from TE's gff file (row 100 to 109):

	Chr	Start	End	strand	TE_rename	ID
100	chr1	2573750	2574278	-	16_MITE/Tourist	Os0702#MITE/Tourist_TE0000096
101	chr1	2595879	2596849	-	14_RC/Helitron	Os2409#DNAnona/Helitron_TE0000097
102	chr1	2608489	2608619	+	11_DNA/MuDR	Os0372#DNAnona/MULE_TE0000098_1
103	chr1	2608625	2613534	+	01_LTR/Copia	ATCOPIA13#LTR/Copia_TEn000003
104	chr1	2613539	2613734	+	11_DNA/MuDR	Os0372#DNAnona/MULE_TE0000098_2
105	chr1	2625733	2629620	-	09_DNA/CACTA	Os1544#DNAnona/CACTA_TE0000099
106	chr1	2633990	2639706	-	01_LTR/Copia	ATCOPIA76#LTR/Copia_TE0000100
107	chr1	2705427	2705731	+	16_MITE/Tourist	Os1928#MITE/Tourist_TE0000101
108	chr1	2740223	2741792	+	14_RC/Helitron	Os0980#DNAauto/Helitron_TE0000102
109	chr1	2766526	2770040	+	11_DNA/MuDR	Os0896#DNAauto/MULE_TE0000103

	TE_id	TE_family	Subclass
100	TE0000096	Os0702	MITE
101	TE0000097	Os2409	Helitron
102	TE0000098_1	Os0372	TIR_transposon
103	TEn000003	ATCOPIA13	LTR_retrotransposon
104	TE0000098_2	Os0372	TIR_transposon
105	TE0000099	Os1544	TIR_transposon
106	TE0000100	ATCOPIA76	LTR_retrotransposon
107	TE0000101	Os1928	MITE
108	TE0000102	Os0980	Helitron
109	TE0000103	Os0896	TIR_transposon

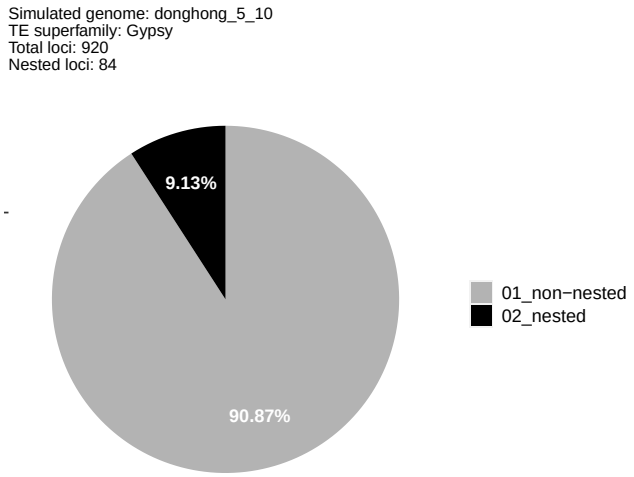
	Identity	Integrity	Nested_in	Cut_at	Cut_by	chrid
100	0.86	1.0	NA	NA	NA	1
101	0.87	0.58	NA	NA	NA	1
102	0.94	0.9	NA	0.4	ATCOPIA13#LTR/Copia_TEn000003	1
103	0.94	1.0	TE0000098	NA	NA	1
104	0.94	0.9	NA	0.4	ATCOPIA13#LTR/Copia_TEn000003	1
105	0.71	0.84	NA	NA	NA	1
106	0.91	1.0	NA	NA	NA	1
107	0.66	0.98	NA	NA	NA	1
108	0.94	0.58	NA	NA	NA	1
109	0.76	1.0	NA	NA	NA	1

Full table saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_10_genome/report/

File name: sim_genome_TE_insertion_info_donghong_5_10.csv

6 Simulation of nested TE insertions

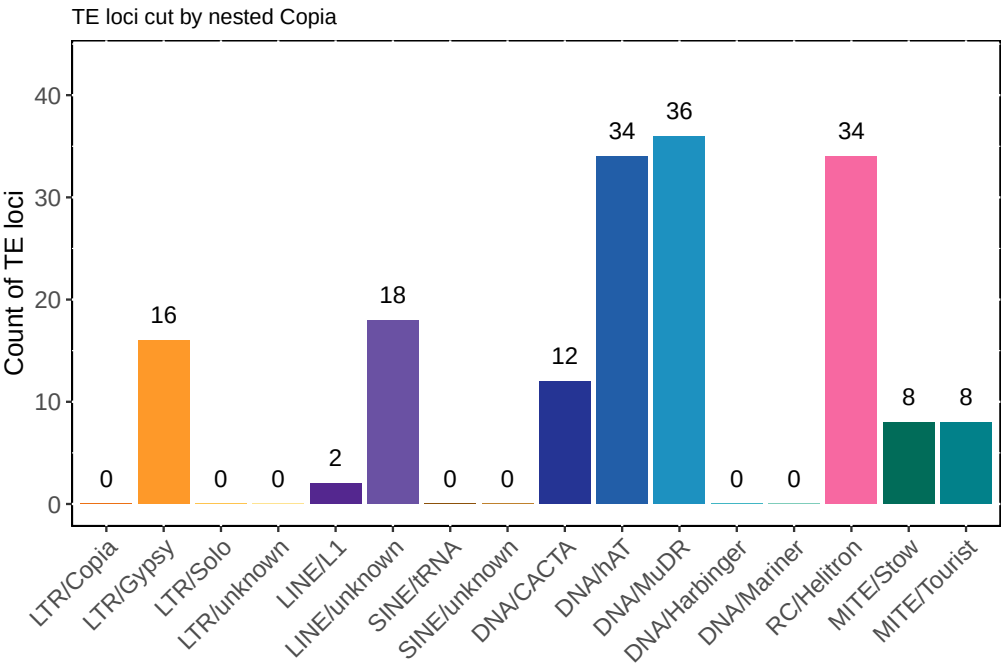
6.1 Nested and non-nested Copia



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_10_genome/report/

File name: 005_sim_te_nested_copia_donghong_5_10.png

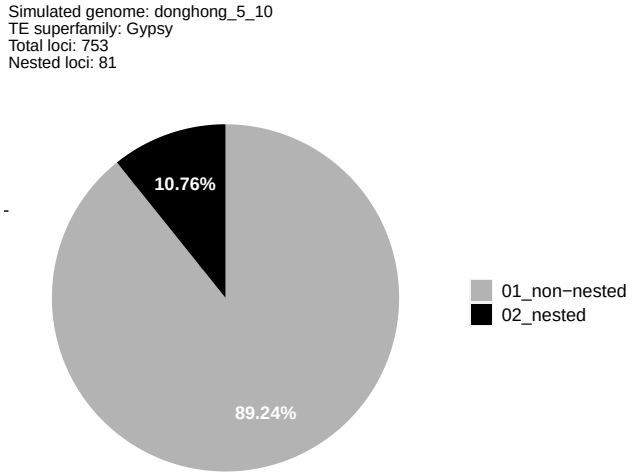
6.2 TE loci cut by nested Copia



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_10_genome/report/

File name: 006_sim_te_cutby_nested_copia_donghong_5_10.png

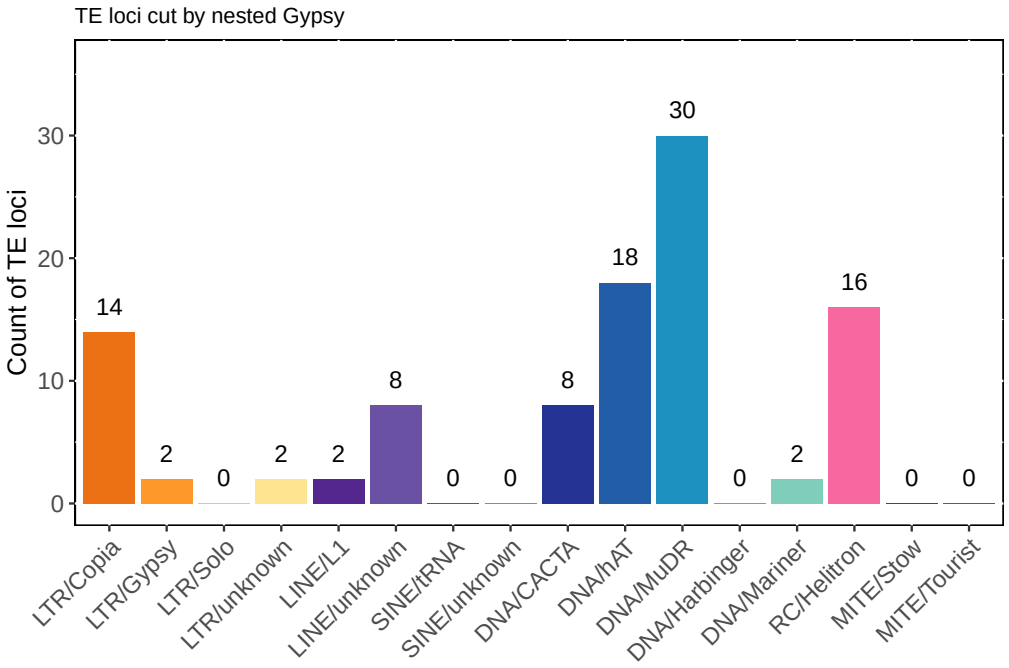
6.3 Nested and non-nested Gypsy



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_10_genome/report/

File name: 007_sim_te_nested_gypsy_donghong_5_10.png

6.4 TE loci cut by nested Gypsy



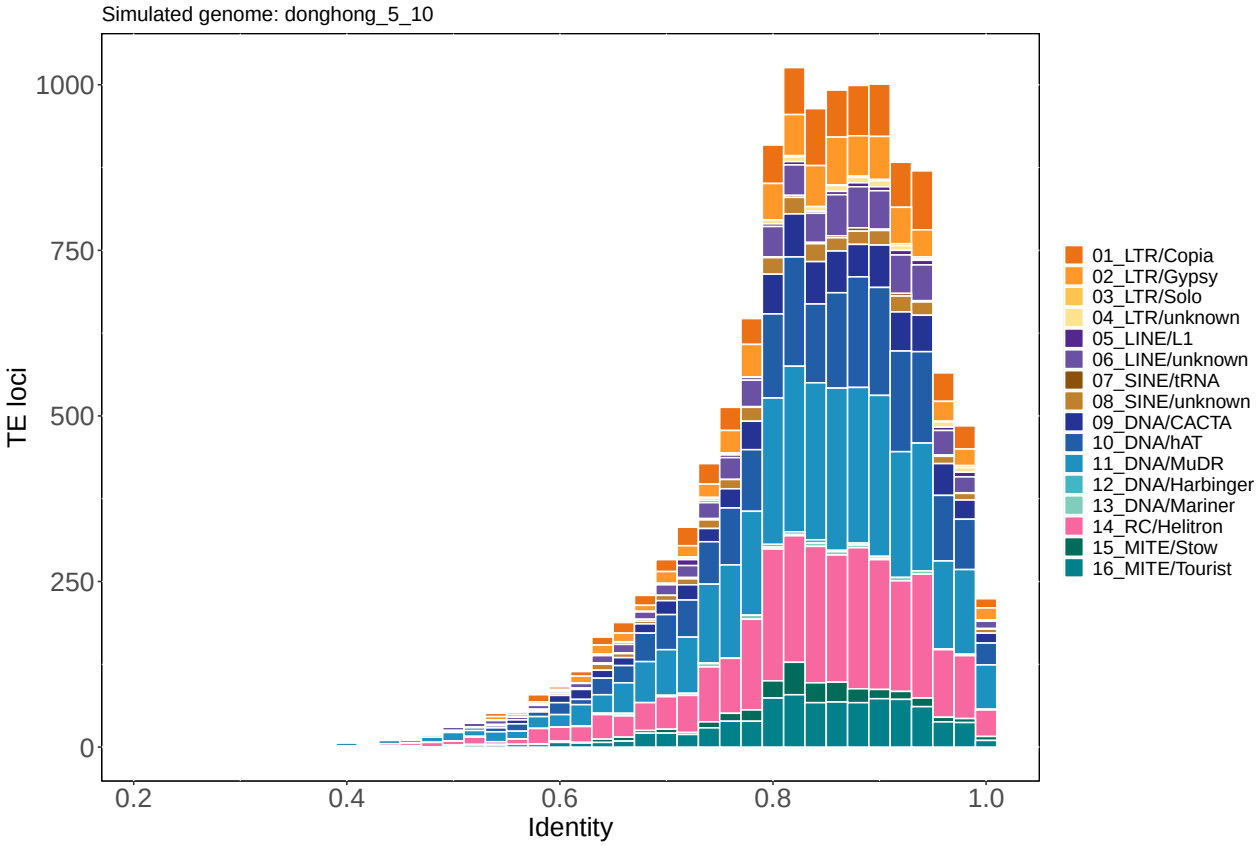
High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_10_genome/report/

File name: 008_sim_te_cutby_nested_gypsy_donghong_5_10.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_10.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_10.table**.



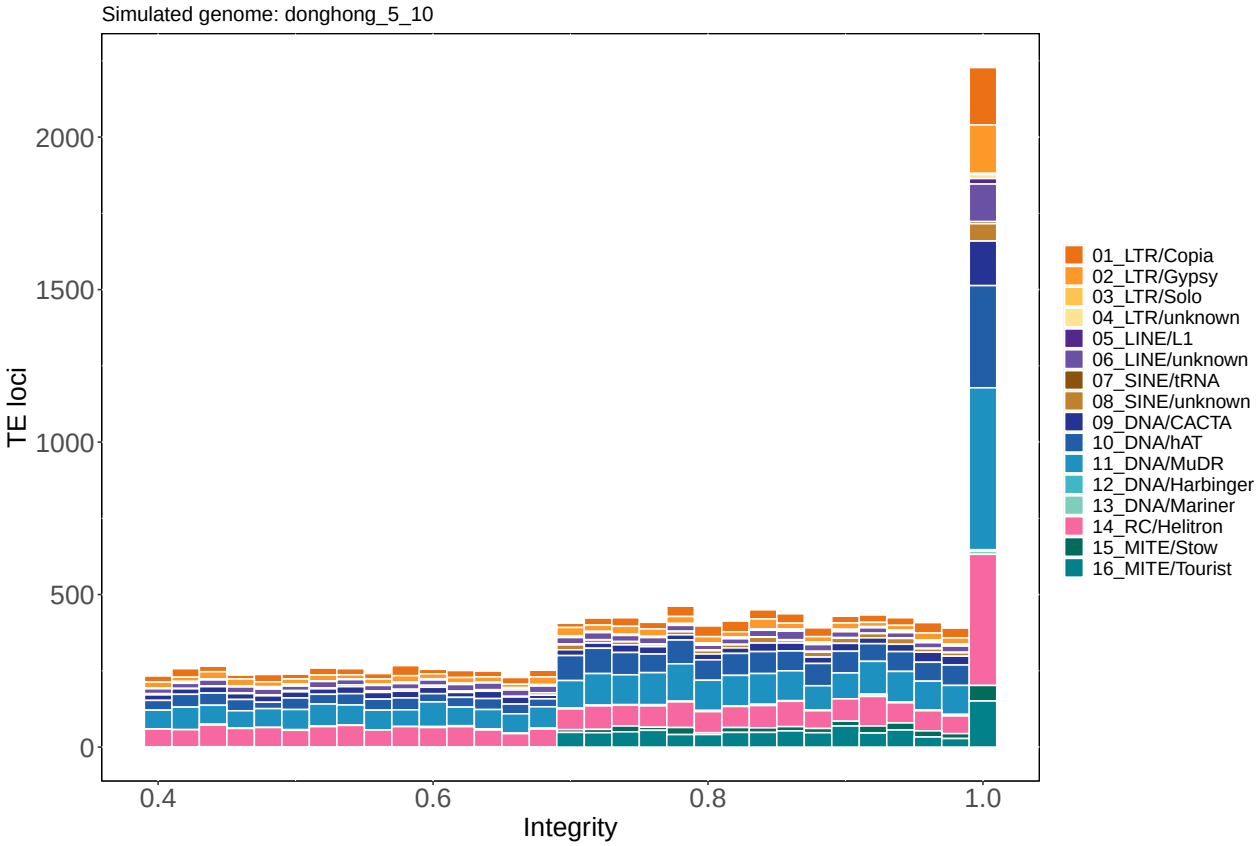
High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_10_genome/report/

File name: 009_sim_te_loci_identity_donghong_5_10.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_10.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_10_genome/report/

File name: 010_sim_te_loci_integrity_donghong_5_10.png