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: 5-1000

CPU	RAM	Time	Timestamp_of_completion
1	5.10 GB	04:38:58	2024-02-21 18:09:13

In this simulation, TEgenomeSimulator did the followings:

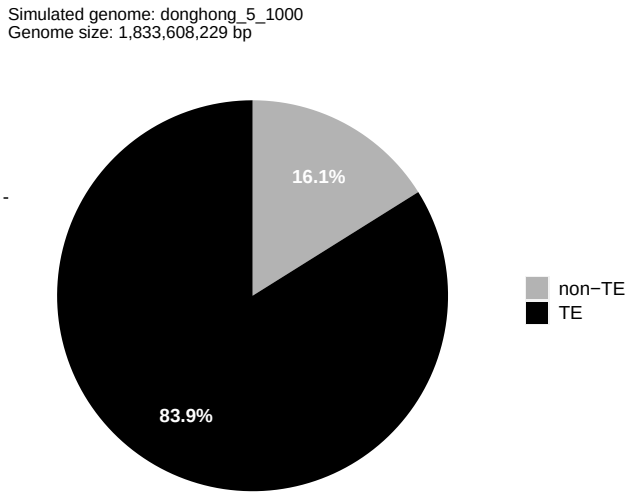
1. It took the TE library file, **combined_curated_TE_lib_ATOSZM_selected.fasta**, which comprises curated TE family sequences from ***A. thalian***, ***Z. maize*** and ***O. sativa***, and simulates multiple TE copies with sequence variations depending on the parameters specified in the table: **TElib_sim_list_5_1000.table**. 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_1000_genome_sequence_out_nest.fasta.fai
- All TE fasta index file: donghong_5_1000_repeat_sequence_out_nest.fasta.fai
- All TE gff file: donghong_5_1000_repeat_annotation_out_nest.gff

3 The proportion of the TE/nonTE sequence

- Total simulated genome size: 1,833,608,229 bp
- Total simulated TE bases: 1,538,396,978 bp



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_1000_genome/report/

File name: 001_sim_genome_size_donghong_5_1000.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.

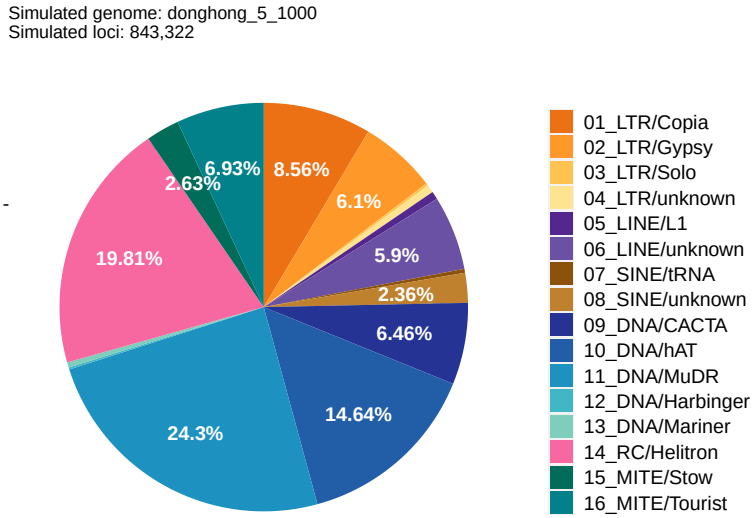
TE_superfamily	loci	bp	family_count	loci_percentage	bp_percentage	family_percentage
01_LTR/Copia	72148	216975884	139	8.56	14.10	8.44
02_LTR/Gypsy	51457	164926232	109	6.10	10.72	6.62
03_LTR/Solo	1661	2083502	3	0.20	0.14	0.18
04_LTR/unknown	5247	10521646	11	0.62	0.68	0.67
05_LINE/L1	5262	16993827	12	0.62	1.10	0.73
06_LINE/unknown	49733	96697544	94	5.90	6.29	5.71
07_SINE/tRNA	2633	486823	5	0.31	0.03	0.30
08_SINE/unknown	19902	9098057	43	2.36	0.59	2.61
09_DNA/CACTA	54445	145568262	110	6.46	9.46	6.68
10_DNA/hAT	123457	109389602	249	14.64	7.11	15.12
11_DNA/MuDR	204892	409134117	393	24.30	26.59	23.86
12_DNA/Harbinger	1357	5018316	3	0.16	0.33	0.18
13_DNA/Mariner	3474	1912591	7	0.41	0.12	0.43
14_RC/Helitron	167031	331140759	314	19.81	21.52	19.06
15_MITE/Stow	22152	4601231	40	2.63	0.30	2.43
16_MITE/Tourist	58471	13928354	115	6.93	0.91	6.98
Total	843322	1538476747	1647	100.00	100.00	100.00

Table saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_1000_genome/report/

File name: sim_genome_summary_TEsuperfamily_donghong_5_1000.csv

4.1 Total simulated TE loci categorised by TE superfamily

- Total simulated TE loci: 843,322 loci

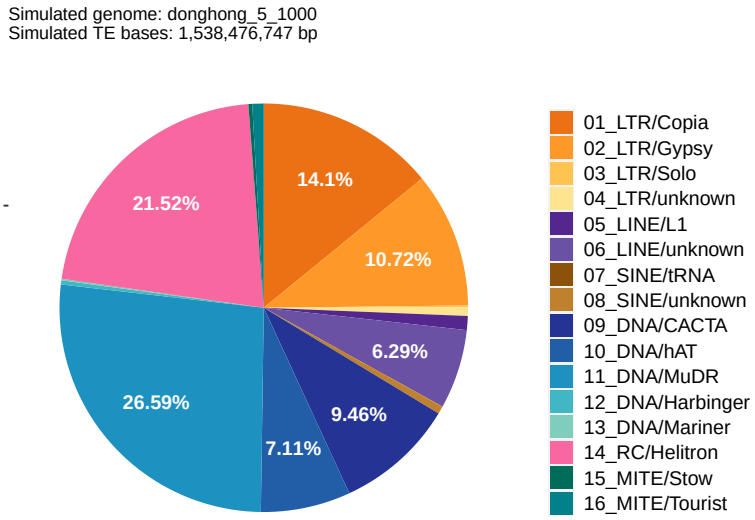


High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_1000_genome/report/

File name: 002_sim_te_loci_donghong_5_1000.png

4.2 Total simulated TE bases categorised by TE superfamily

- Total simulated TE bases: 1,538,476,747 bp



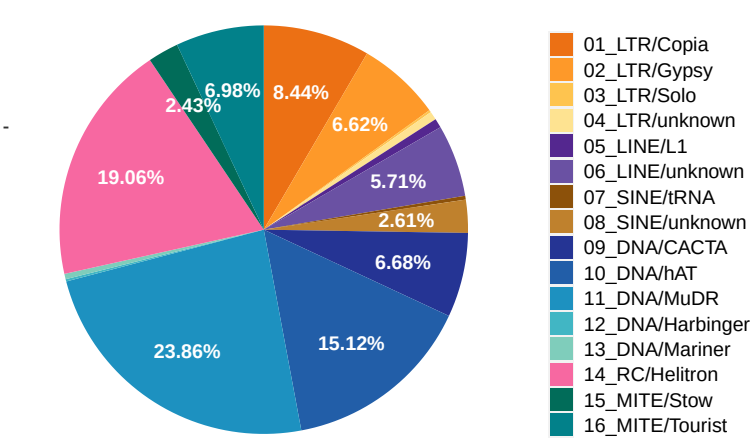
High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_1000_genome/report/

File name: 003_sim_te_bp_donghong_5_1000.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_1000_genome/report/

File name: 004_sim_te_family_donghong_5_1000.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	178826	182929	+	01_LTR/Copia	ATCOPIA22#LTR/Copia_TE0000099
102	chr1	183728	183981	-	16_MITE/Tourist	Os1715#MITE/Tourist_TE0000100
103	chr1	184358	184984	-	09_DNA/CACTA	Os2647#DNAnona/CACTA_TE0000101
104	chr1	185152	186264	-	14_RC/Helitron	Os2969#DNAnona/Helitron_TE0000102
105	chr1	186399	186653	-	14_RC/Helitron	Os2221#DNAnona/Helitron_TE0000103
106	chr1	186991	189958	-	01_LTR/Copia	ATCOPIA3#LTR/Copia_TE0000104
107	chr1	190281	190590	+	16_MITE/Tourist	Os0328#MITE/Tourist_TE0000105
108	chr1	190932	191117	+	10_DNA/hAT	Os0517#DNAnona/hAT_TE0000106
109	chr1	191671	192106	-	11_DNA/MuDR	Os0105#DNAnona/MULE_TE0000107
110	chr1	192669	192928	-	16_MITE/Tourist	Os1114#MITE/Tourist_TE0000108

	TE_id	TE_family	Subclass
101	TE0000099	ATCOPIA22	LTR_retrotransposon
102	TE0000100	Os1715	MITE
103	TE0000101	Os2647	TIR_transposon
104	TE0000102	Os2969	Helitron
105	TE0000103	Os2221	Helitron
106	TE0000104	ATCOPIA3	LTR_retrotransposon
107	TE0000105	Os0328	MITE
108	TE0000106	Os0517	TIR_transposon
109	TE0000107	Os0105	TIR_transposon
110	TE0000108	Os1114	MITE

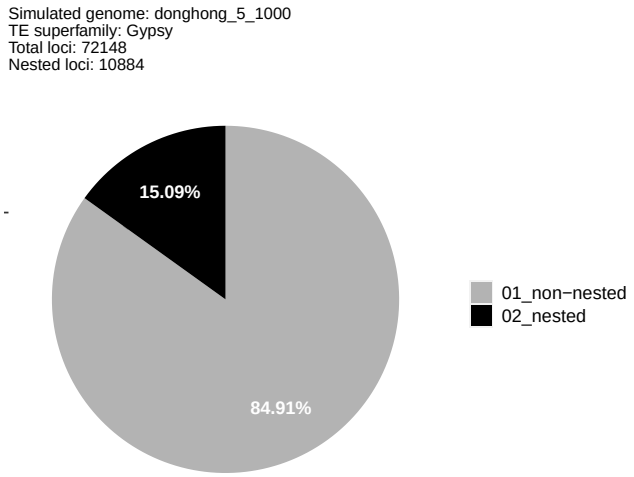
	Identity	Integrity	Nested_in	Cut_at	Cut_by	chrid
101	0.83	0.84	NA	NA	NA	1
102	0.91	0.89	NA	NA	NA	1
103	0.78	0.75	NA	NA	NA	1
104	0.94	0.41	NA	NA	NA	1
105	0.69	0.87	NA	NA	NA	1
106	0.91	0.64	NA	NA	NA	1
107	0.84	0.97	NA	NA	NA	1
108	0.9	1.0	NA	NA	NA	1
109	0.53	0.73	NA	NA	NA	1
110	0.93	0.82	NA	NA	NA	1

Full table saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_1000_genome/report/

File name: sim_genome_TE_insertion_info_donghong_5_1000.csv

6 Simulation of nested TE insertions

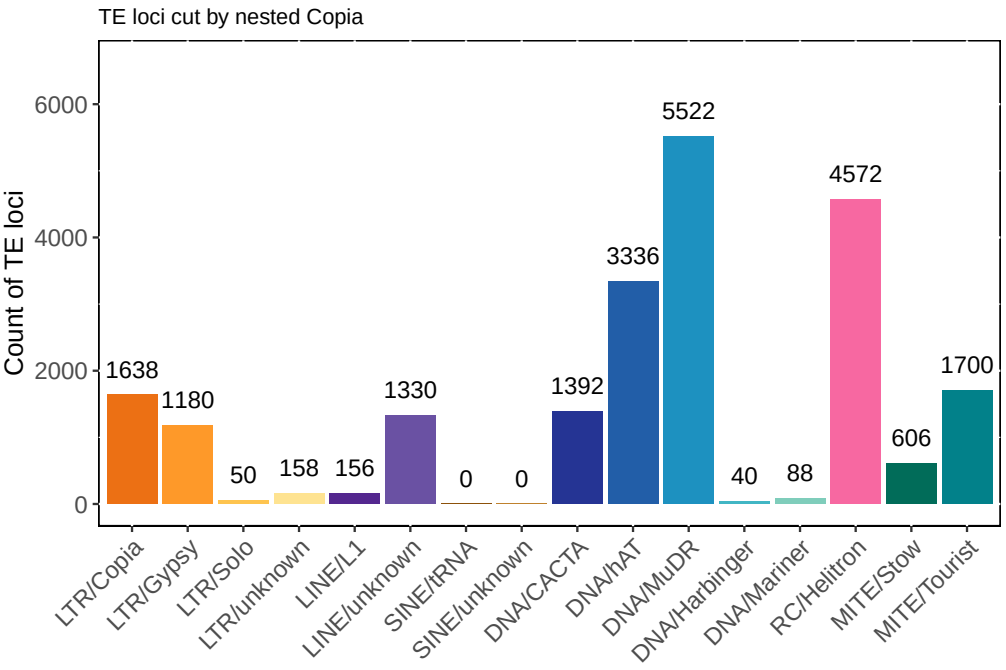
6.1 Nested and non-nested Copia



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_1000_genome/report/

File name: 005_sim_te_nested_copia_donghong_5_1000.png

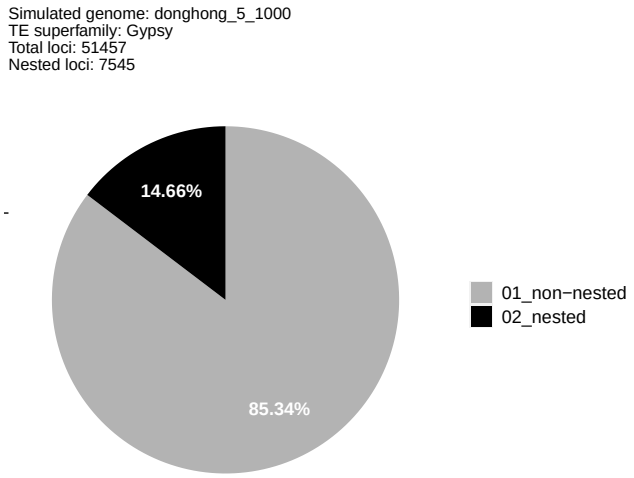
6.2 TE loci cut by nested Copia



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_1000_genome/report/

File name: 006_sim_te_cutby_nested_copia_donghong_5_1000.png

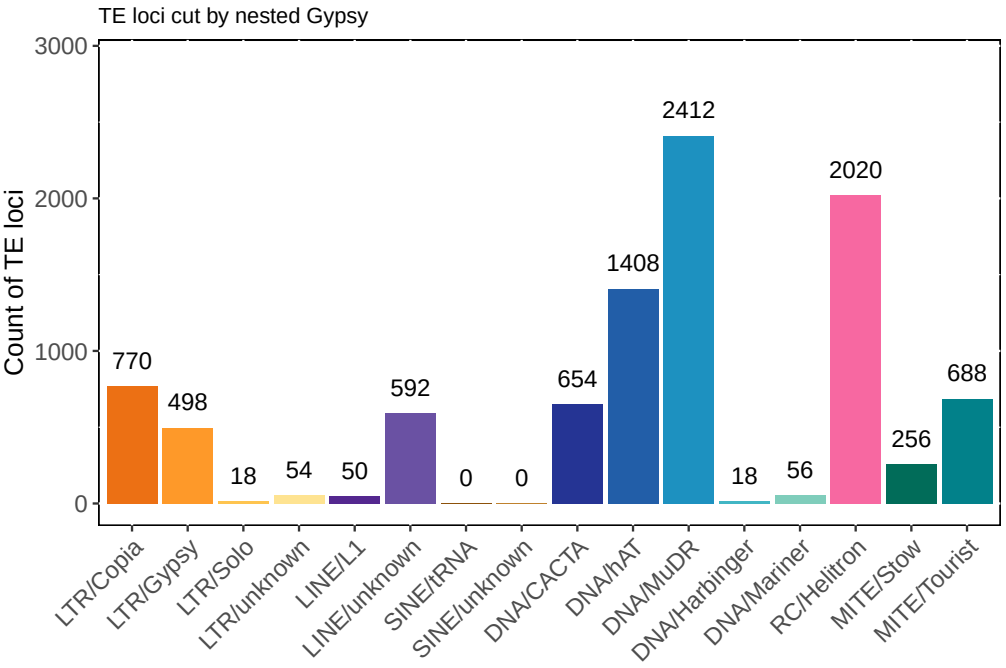
6.3 Nested and non-nested Gypsy



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_1000_genome/report/

File name: 007_sim_te_nested_gypsy_donghong_5_1000.png

6.4 TE loci cut by nested Gypsy



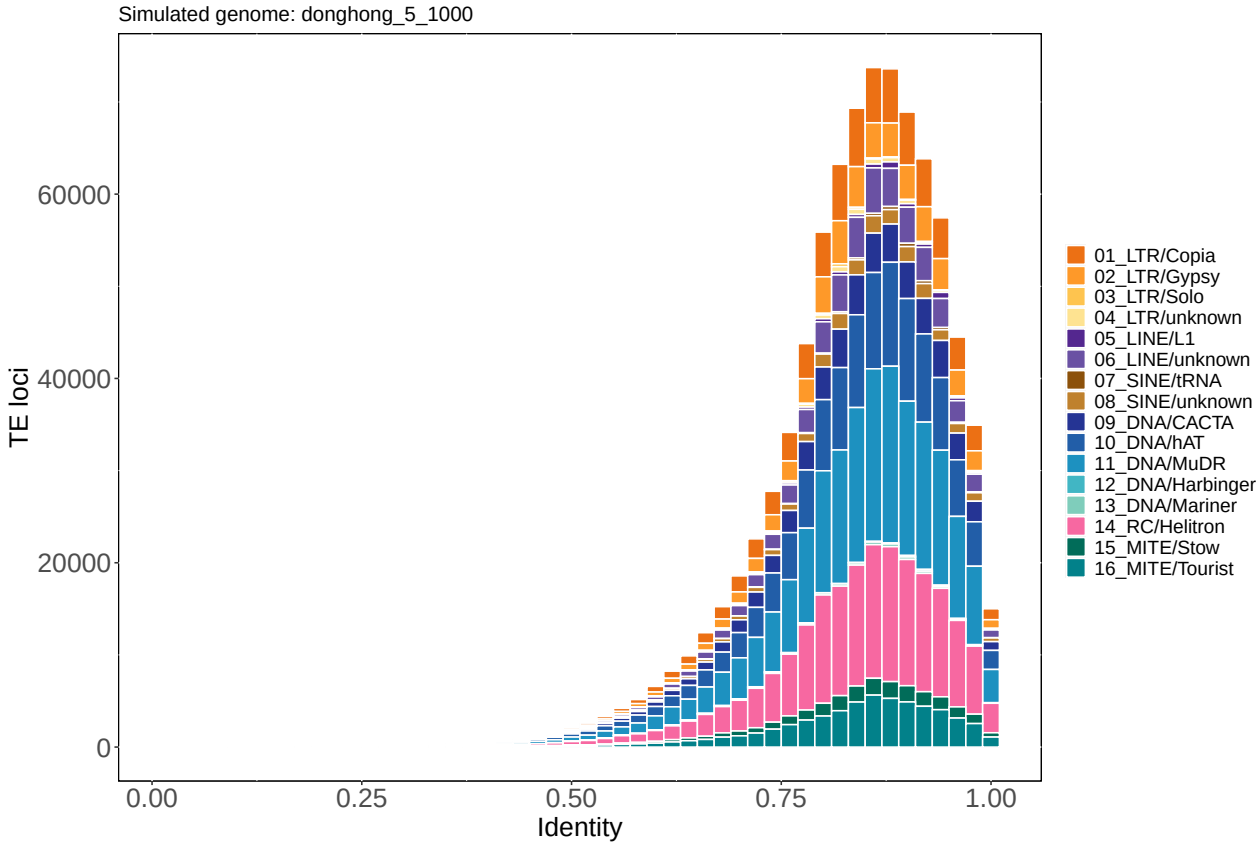
High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_1000_genome/report/

File name: 008_sim_te_cutby_nested_gypsy_donghong_5_1000.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_1000.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_1000.table**.



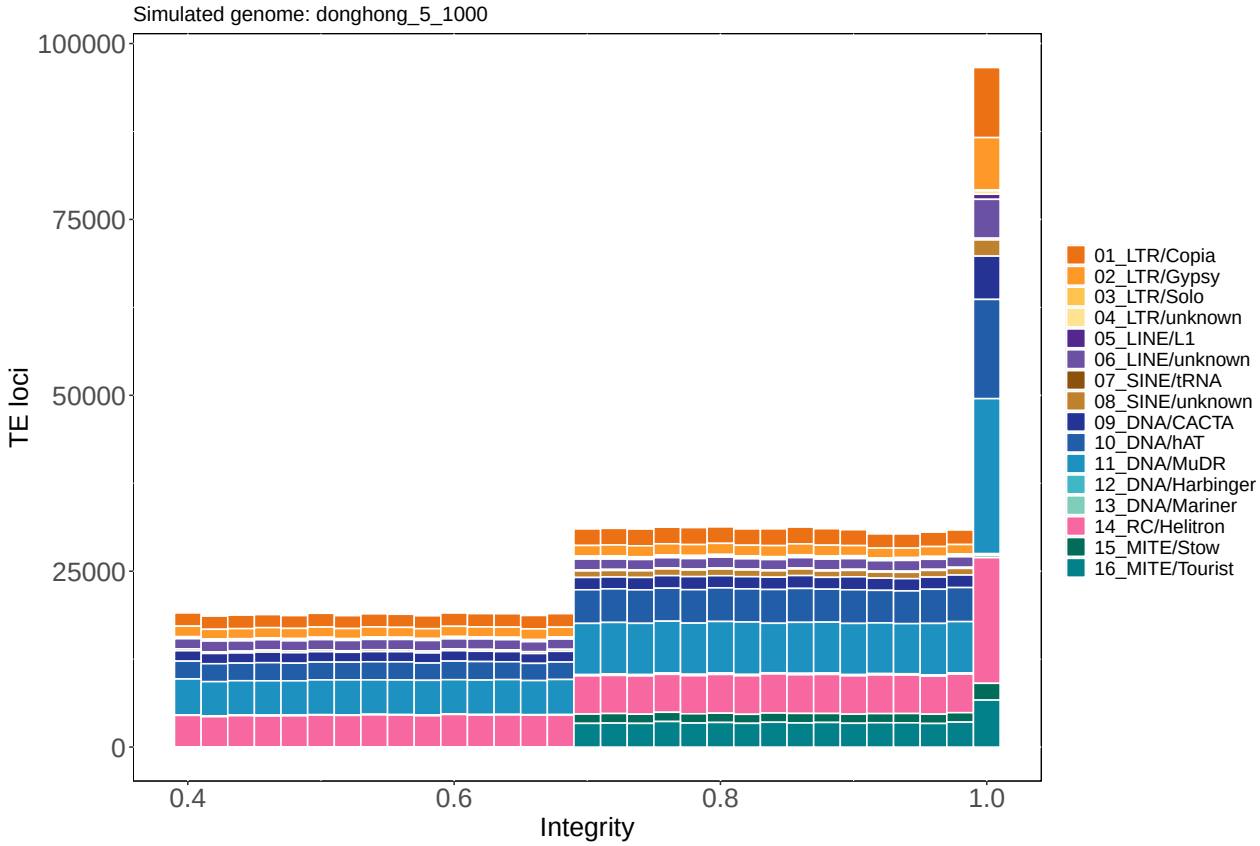
High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_1000_genome/report/

File name: 009_sim_te_loci_identity_donghong_5_1000.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__1000.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_1000_genome/report/

File name: 010_sim_te_loci_integrity_donghong_5_1000.png