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: 5-100

CPU	RAM	Time	Timestamp_of_completion
1	879.99 MB	00:03:38	2024-02-21 14:16:28

In this simulation, TEgenomeSimulator did the followings:

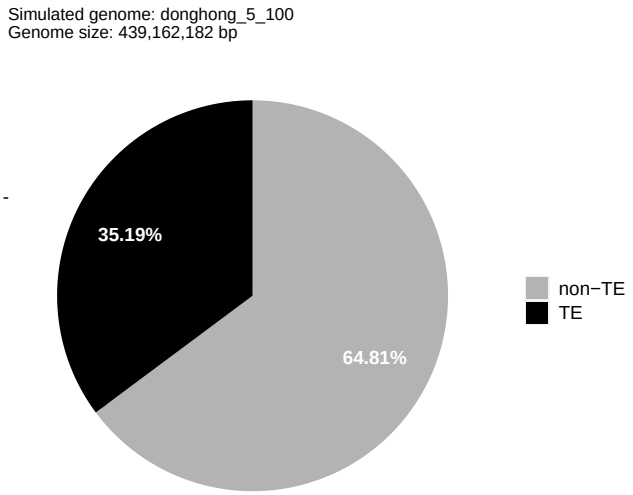
1. It took the TE library file, **combined_curated_TE_lib_ATOSZM_selected.fasta**, which comprises curated TE family sequences from ***A. thalian***, ***Z. maize*** and ***O. sativa***, and simulates multiple TE copies with sequence variations depending on the parameters specified in the table: **TElib_sim_list_5_100.table**. 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_100_genome_sequence_out_nest.fasta.fai
- All TE fasta index file: donghong_5_100_repeat_sequence_out_nest.fasta.fai
- All TE gff file: donghong_5_100_repeat_annotation_out_nest.gff

3 The proportion of the TE/nonTE sequence

- Total simulated genome size: 439,162,182 bp
- Total simulated TE bases: 154,521,730 bp



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_100_genome/report/

File name: 001_sim_genome_size_donghong_5_100.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.

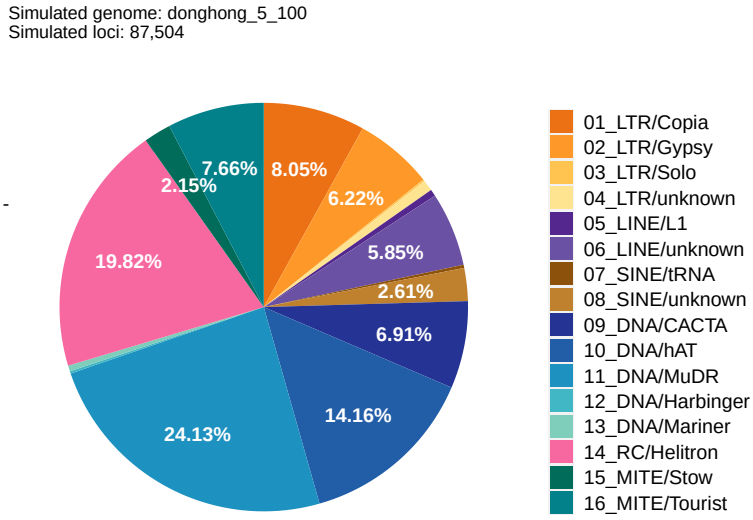
TE_superfamily	loci	bp	family_count	loci_percentage	bp_percentage	family_percentage
01_LTR/Copia	7044	20653831	139	8.05	13.37	8.44
02_LTR/Gypsy	5440	15956044	109	6.22	10.33	6.62
03_LTR/Solo	119	123081	3	0.14	0.08	0.18
04_LTR/unknown	767	1022315	11	0.88	0.66	0.67
05_LINE/L1	474	1783853	12	0.54	1.15	0.73
06_LINE/unknown	5121	9461178	94	5.85	6.12	5.71
07_SINE/tRNA	221	38678	5	0.25	0.03	0.30
08_SINE/unknown	2282	1159322	43	2.61	0.75	2.61
09_DNA/CACTA	6044	17198216	110	6.91	11.13	6.68
10_DNA/hAT	12389	10544455	249	14.16	6.82	15.12
11_DNA/MuDR	21113	39076681	393	24.13	25.29	23.86
12_DNA/Harbinger	164	515607	3	0.19	0.33	0.18
13_DNA/Mariner	398	216767	7	0.45	0.14	0.43
14_RC/Helitron	17347	34834469	314	19.82	22.54	19.06
15_MITE/Stow	1879	392862	40	2.15	0.25	2.43
16_MITE/Tourist	6702	1552360	115	7.66	1.00	6.98
Total	87504	154529719	1647	100.00	100.00	100.00

Table saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_100_genome/report/

File name: sim_genome_summary_TEsuperfamily_donghong_5_100.csv

4.1 Total simulated TE loci categorised by TE superfamily

- Total simulated TE loci: 87,504 loci



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_100_genome/report/

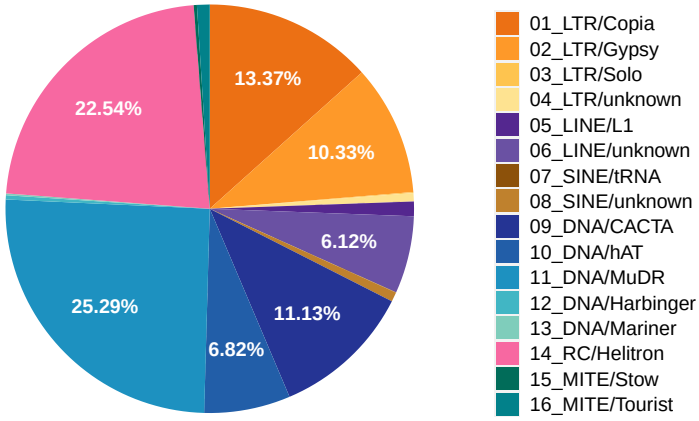
File name: 002_sim_te_loci_donghong_5_100.png

4.2 Total simulated TE bases categorised by TE superfamily

- Total simulated TE bases: 154,529,719 bp

Simulated genome: donghong_5_100
Simulated TE bases: 154,529,719 bp

-



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_100_genome/report/

File name: 003_sim_te_bp_donghong_5_100.png

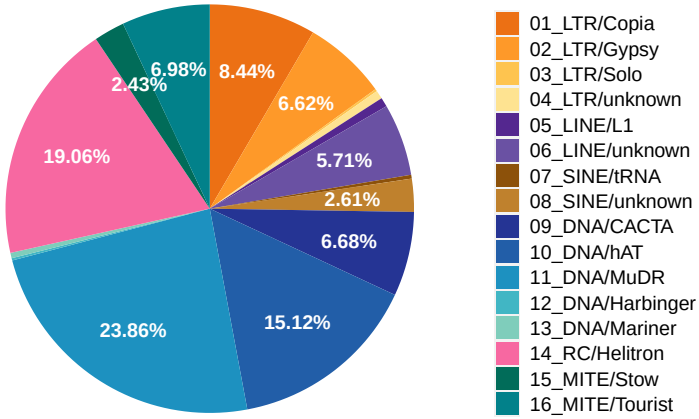
4.3 Simulated TE family per superfamily

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

- Total simulated TE families: 1647 families

Simulated genome: donghong_5_100
Simulated families: 1,647

-



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_100_genome/report/

File name: 004_sim_te_family_donghong_5_100.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	491216	493425	+	06_LINE/unknown	Os0948#LINE/unknown_TE0000099
102	chr1	493878	496076	-	10_DNA/hAT	Os2295#DNAauto/hAT_TE0000100
103	chr1	497440	500513	+	06_LINE/unknown	Os3659#LINE/unknown_TE0000101
104	chr1	507121	515653	+	14_RC/Helitron	Os1887#DNAauto/Helitron_TE0000102
105	chr1	516343	516712	-	09_DNA/CACTA	Os1329#DNAnona/CACTA_TE0000103
106	chr1	518020	518439	+	11_DNA/MuDR	Os0986#DNAnona/MULE_TE0000104
107	chr1	520368	520791	-	11_DNA/MuDR	Os0182#DNAnona/MULE_TE0000105
108	chr1	527733	528311	-	09_DNA/CACTA	Os3007#DNAnona/CACTA_TE0000106
109	chr1	528357	528573	-	10_DNA/hAT	Os0780#DNAnona/hAT_TE0000107
110	chr1	531644	532527	+	14_RC/Helitron	Os1070#DNAnona/Helitron_TE0000108

	TE_id	TE_family	Subclass
101	TE0000099	Os0948	LINE_retrotransposon
102	TE0000100	Os2295	TIR_transposon
103	TE0000101	Os3659	LINE_retrotransposon
104	TE0000102	Os1887	Helitron
105	TE0000103	Os1329	TIR_transposon
106	TE0000104	Os0986	TIR_transposon
107	TE0000105	Os0182	TIR_transposon
108	TE0000106	Os3007	TIR_transposon
109	TE0000107	Os0780	TIR_transposon
110	TE0000108	Os1070	Helitron

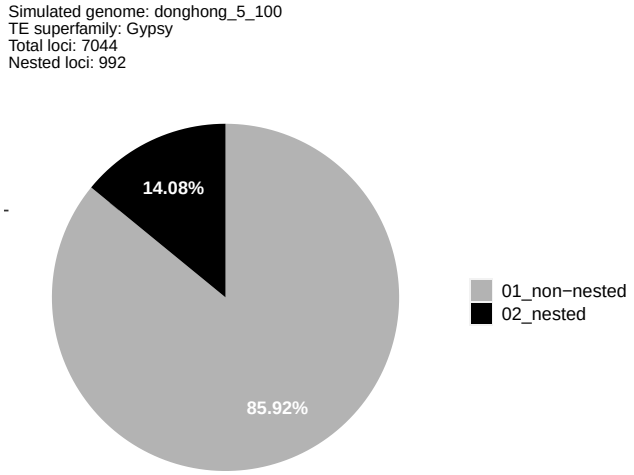
	Identity	Integrity	Nested_in	Cut_at	Cut_by	chrid
101	0.95	0.71	NA	NA	NA	1
102	0.58	0.66	NA	NA	NA	1
103	0.82	0.84	NA	NA	NA	1
104	0.9	0.71	NA	NA	NA	1
105	0.67	0.54	NA	NA	NA	1
106	0.86	0.97	NA	NA	NA	1
107	0.75	0.8	NA	NA	NA	1
108	0.81	0.55	NA	NA	NA	1
109	0.6	0.9	NA	NA	NA	1
110	0.89	0.75	NA	NA	NA	1

Full table saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_100_genome/report/

File name: sim_genome_TE_insertion_info_donghong_5_100.csv

6 Simulation of nested TE insertions

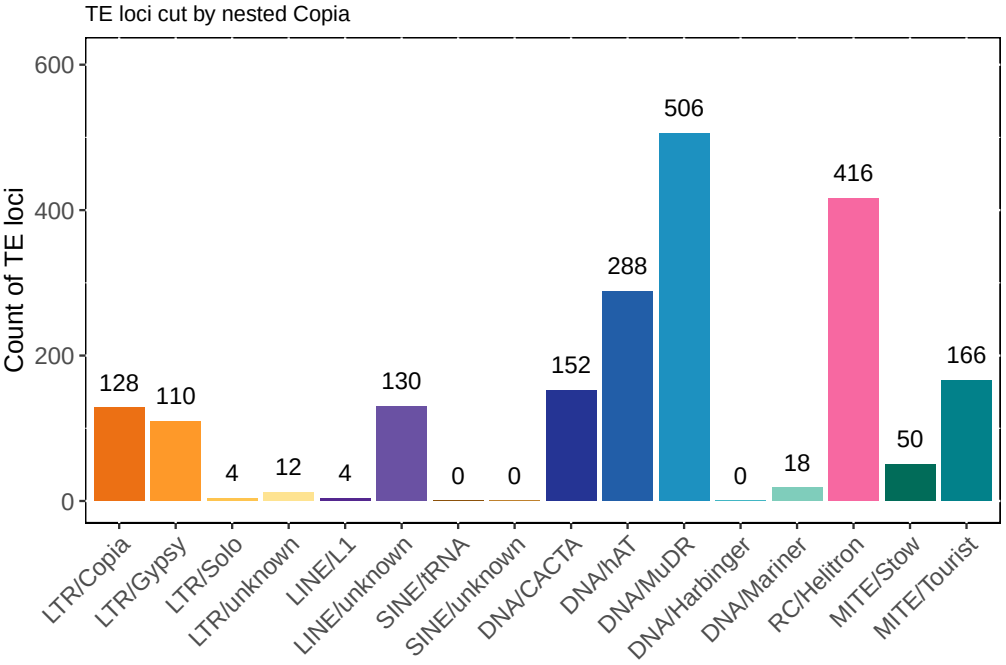
6.1 Nested and non-nested Copia



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_100_genome/report/

File name: 005_sim_te_nested_copia_donghong_5_100.png

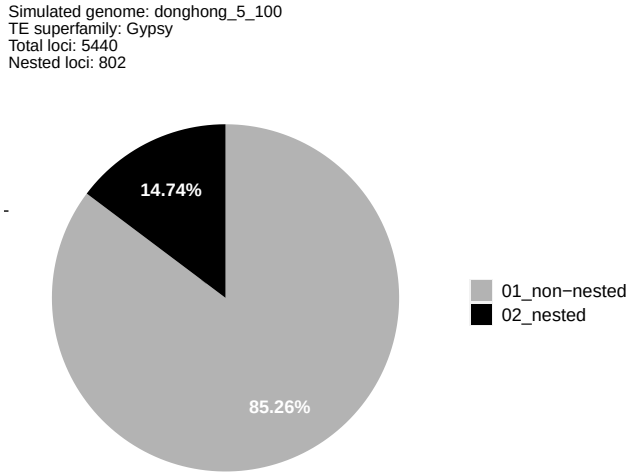
6.2 TE loci cut by nested Copia



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_100_genome/report/

File name: 006_sim_te_cutby_nested_copia_donghong_5_100.png

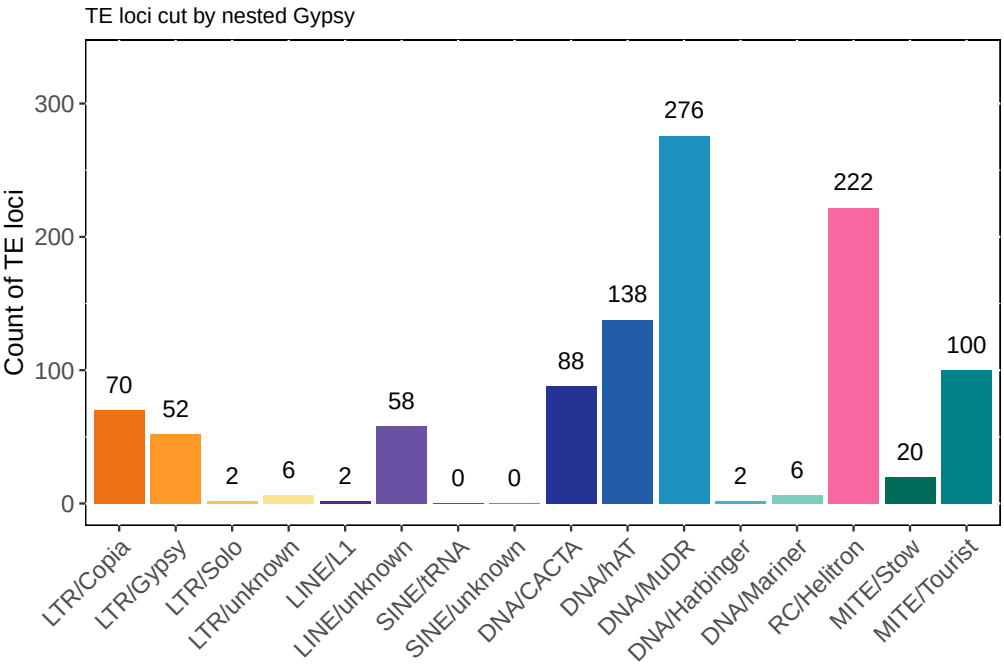
6.3 Nested and non-nested Gypsy



High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_100_genome/report/

File name: 007_sim_te_nested_gypsy_donghong_5_100.png

6.4 TE loci cut by nested Gypsy



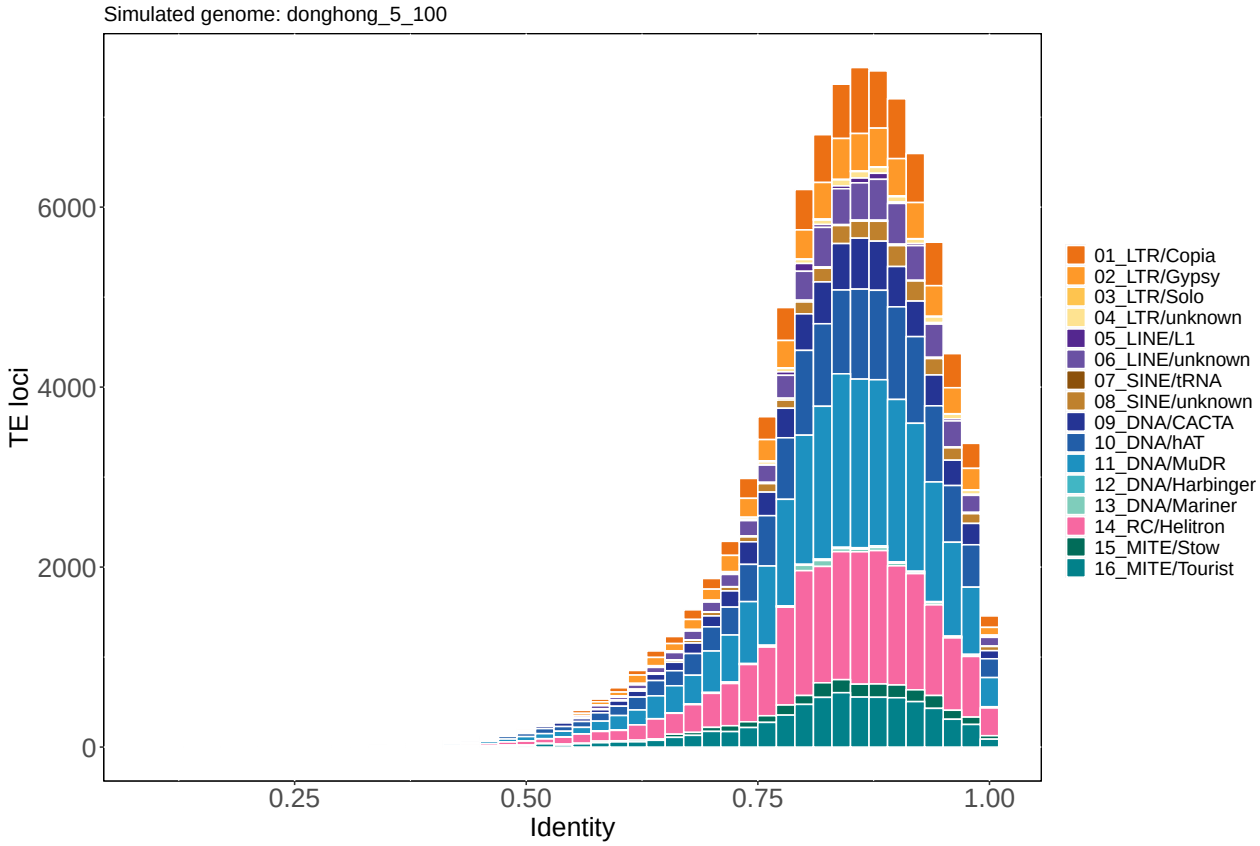
High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_100_genome/report/

File name: 008_sim_te_cutby_nested_gypsy_donghong_5_100.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_100.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_100.table**.



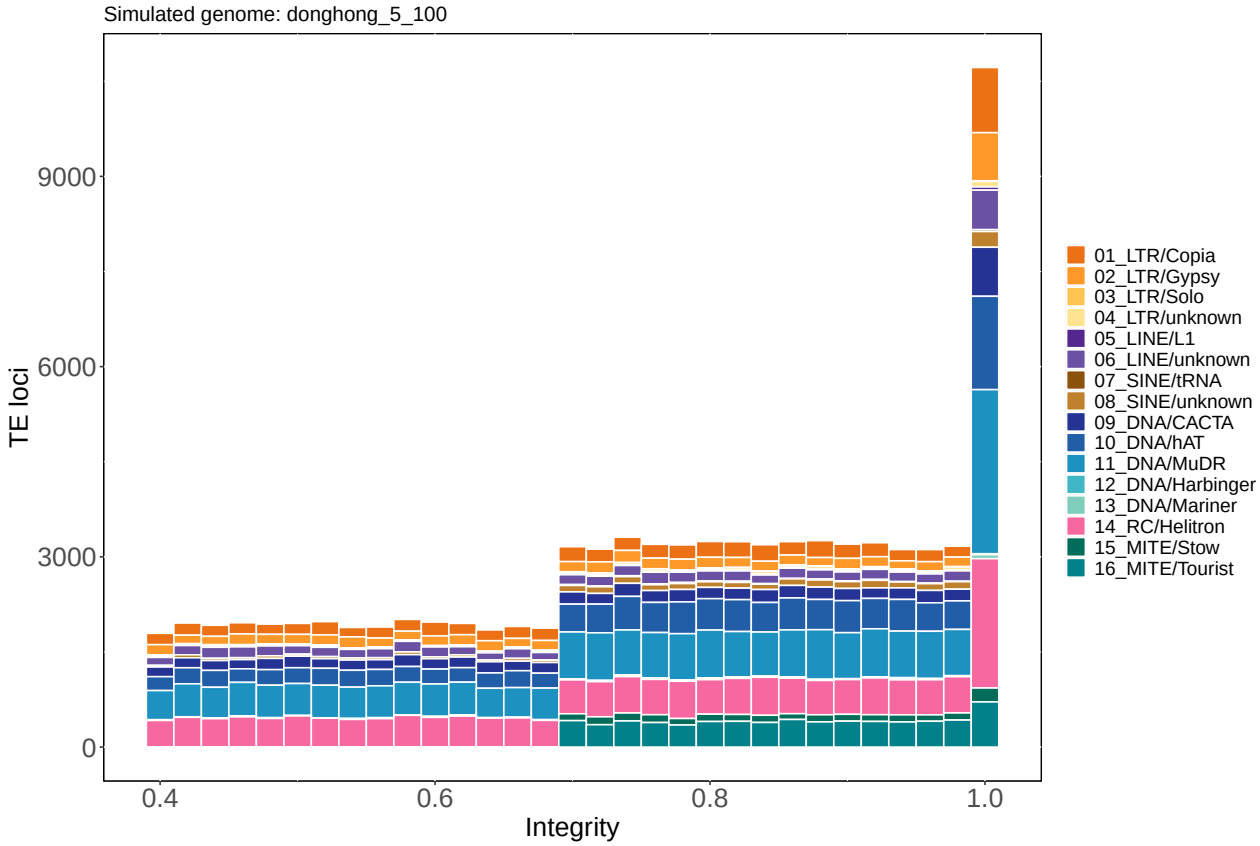
High quality image saved in
/workspace/cflthc/scratch/2022_Actinidia_TE/10.01_TEgenomeSimulator_output/sim_donghong_5_100_genome/report/

File name: 009_sim_te_loci_identity_donghong_5_100.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_100.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_100_genome/report/

File name: 010_sim_te_loci_integrity_donghong_5_100.png