Supplemental marterials for the article

Guideline for the Genome Transposon Annotation derived from evaluation of popular TE tools

Haidong Yan 1, Federica Torchiana 2 and Aureliano Bombarely 2,\*

1 School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA 24061 USA

2 Department of Bioscience, Universita degli Studi di Milano Milan, Italy, 20133

**\*** Correspondence: Aureliano.bombarely@unimi.it;

Received: date; Accepted: date; Published: date

**Table S1** TE identification tools used in plant genome sequencing studies in 2019.

|  |  |  |  |
| --- | --- | --- | --- |
| **Species** | **De novo-based strategy** | **Homology-based strategy** | **Time** |
| *Malania oleifera* | RepeatModeler | RepBase | Jan, 2019 |
| *Cucurbita argyrosperma* | REPET pipeline | REPET pipeline | Jan, 2019 |
| *Cinnamomum kanehirae* | RepeatModeler | None | Jan, 2019 |
| *Malania oleifera* | RepeatModeler | RepBase | Jan, 2019 |
| *Panicum miliaceum* | RepeatModeler; LTR\_FINDER; PILER | RepBase | Jan, 2019 |
| *Antirrhinum majus* | RepeatModeler; LTR\_FINDER | PGSB repeat element database | Jan, 2019 |
| *garden strawberry (Fragaria × ananassa)* | RepeatModeler; LTR\_FINDER; LTRharvest; LTR\_retriever; MITE-Hunter | None | Feb, 2019 |
| *Arachis hypogaea* | None | RepBase; Dfam | Mar, 2019 |
| *Crucihimalaya himalaica* | RepeatModeler; PILER; LTR\_FINDER | RepBase | Mar, 2019 |
| *Caulerpa lentillifera* | RepeatModeler | None | Mar, 2019 |
| *Actinidia eriantha* | RepeatModeler; LTRharvest; MITE-Hunter | None | April, 2019 |
| *Scutellaria baicalensis* | RepeatModeler; LTR\_FINDER; PILER | RepBase | April, 2019 |
| *Carya illinoinensis* | RepeatModeler | RepBase | May, 2019 |
| *Carya cathayensis* | RepeatModeler | RepBase | May, 2019 |
| *Beta patula* | None | Consensus TEs from *Beta vulgaris* and *Spinacia oleracea* | May, 2019 |
| *Mercurialis annua* | RepeatModeler | Consensus TEs from *Mercurialis annua* (from RepeatModeler), Euphorbiacae, and *Vitis vinifera* | May, 2019 |
| *Malus baccata* | RepeatModeler | RepBase | May, 2019 |
| *richopus zeylanicus* | RepeatModeler | RepBase | June, 2019 |
| *Sedum album* | RepeatModeler; LTR\_FINDER; LTRharvest; LTR\_retriever; MITE-Hunter | None | June, 2019 |
| *Abies alba* | RepeatModeler | Consensus TEs from Pinaceae-specific library of RepBase | June, 2019 |
| *Xanthoceras sorbifolium* | RepeatModeler | RepBase plant repeat database | June, 2019 |
| *Spatholobus suberectus* | RepeatModeler; LTR\_FINDER; PILER | RepBase | July, 2019 |
| *Dactylis glomerata* | RepeatModeler; LTR\_FINDER | RepBase | July, 2019 |
| *Acer yangbiense* | RepeatModeler | None | July, 2019 |
| *Eragrostis curvula* | RepeatModeler | Consensus TEs from *Zea mays*, *Sorghum bicolor* and *Oryza sativa* library of RepBase | July, 2019 |
| *Pyrus betuleafolia* | LTR\_FINDER; RepeatScout; PILER | RepBase | Aug, 2019 |
| *Phytolacca americana* | RepeatModeler | None | Aug, 2019 |
| *Persea americana* | REPET pipeline | REPET pipeline | Aug, 2019 |
| *Gelsemium elegans* | RepeatModeler | RepBase and Mips-REdat libraries | Aug, 2019 |
| *Gossypium raimondii* | None | Consensus TEs from cotton-specific repeats library of RepBase | Aug, 2019 |
| *Gossypium turneri* | None | Consensus TEs from cotton-specific repeats library of RepBase | Aug, 2019 |
| *Pisum sativum* | REPET pipeline | REPET pipeline | Sep, 2019 |
| *Chloropicon primus* | RepeatModeler | None | Sep, 2019 |
| *Castanea mollissima* | RepeatModeler | None | Sep, 2019 |
| *Suaeda aralocaspica* | RepeatModeler; PILER | Consensus TEs from Viridiplantae repeats library of RepBase | Sep, 2019 |
| *Kochia scoparia* | None | Consensus TEs from Viridiplantae repeats library of RepBase | Sep, 2019 |
| *Asclepias syriaca* | RepeatModeler | None | Sep, 2019 |
| *Ananas comosus* | None | RepBase | Sep, 2019 |
| *Solanum aethiopicum* | RepeatModeler; LTRharvest | RepBase | Oct, 2019 |
| *Piper nigrum* | RepeatModeler; LTRharvest | TransposonPSI | Oct, 2019 |
| *Polygonum cuspidatum* | LTR\_FINDER; MITE-Hunter; RepeatScout; PILER | RepBase | Oct, 2019 |
| *Rhododendron williamsianum* | RepeatModeler | RepBase | Nov, 2019 |
| *Chenopodium pallidicaule* | RepeatModeler | RepBase | Nov, 2019 |
| *Pachycladon cheesemanii* | RepeatModeler; LTR\_FINDER; MITE-Hunter; TransposonPSI | Consensus TEs from Viridiplantae repeats library of RepBase | Nov, 2019 |
| *Benincasa hispida* | RepeatModeler | TIGR plant repeats database | Nov, 2019 |
| *Trochodendron aralioides* | RepeatModeler; LTR\_FINDER | Repbase | Nov, 2019 |
| *Salix brachista* | RepeatModeler | None | Nov, 2019 |
| *Nyssa sinensis* | RepeatModeler; LTR\_FINDER; LTRharvest; LTR\_retriever | Repbase | Nov, 2019 |
| *Gossypioides kirkii* | RepeatModeler | None | Nov, 2019 |
| *Ceratopteris richardii* | RepeatModeler; LTRharvest | TESeeker; RepBase | Dec, 2019 |
| *Coriandrum sativum* | RepeatModeler; LTR\_FINDER; PILER | RepBase | Dec, 2019 |
| *Chlorokybus atmophyticus* | RepeatModeler; LTRharvest; MITE-Hunter | RepBase | Dec, 2019 |
| *Mesostigma viride* | RepeatModeler; LTRharvest; MITE-Hunter | RepBase | Dec, 2019 |
| *Utricularia reniformis* | REPET pipeline | REPET pipeline | Dec, 2019 |
| *Diospyros oleifera* | LTR\_FINDER; MITE-Hunter; PILER; RepeatScout | RepBase | Dec, 2019 |
| *Nymphaea colorata* | RepeatModeler; LTR\_FINDER | None | Dec, 2019 |
| *Artocarpus heterophyllus* | RepeatModeler; LTRharvest; MITE-Hunter | RepBase | Dec, 2019 |
| *Artocarpus altilis* | RepeatModeler; LTRharvest; MITE-Hunter | RepBase | Dec, 2019 |

**Table S2** Summary of tools and sources used for TE identification in 2019.

|  |  |  |  |
| --- | --- | --- | --- |
| **De novo-based strategy** | **Number of Studies** | **Homology-based strategy** | **Number of Studies** |
| RepeatModeler | 45 | All libraries from RepBase | 28 |
| LTR\_FINDER | 16 | Species- or genus-specific libraries from RepBase | 8 |
| LTRharvest | 11 | Other sources | 9 |
| MITE-Hunter | 10 | None | 13 |
| PILER | 9 |  |  |
| None | 6 |  |  |
| REPET pipeline | 4 |  |  |
| LTR\_retriever | 3 |  |  |
| RepeatScout | 3 |  |  |
| TransposonPSI | 1 |  |  |

**Table S3** TE copy number for each mutation type.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type** | **Level** | **Copy number** | **Fusion copy number** | **Total number** |
| Mutation | 1% | 250 | 20 | 270 |
| 5% | 250 | 20 | 270 |
| 10% | 250 | 20 | 270 |
| Insertion | 1% | 250 | 20 | 270 |
| 5% | 250 | 20 | 270 |
| 10% | 250 | 20 | 270 |
| Deletion | 1% | 250 | 20 | 270 |
| 5% | 250 | 20 | 270 |
| 10% | 250 | 20 | 270 |
| Mutation + Deletion | 1% | 250 | 20 | 270 |
| Mutation + Insertion | 1% | 250 | 20 | 270 |
| Insertion + Deletion | 1% | 250 | 20 | 270 |
| Mutation + Insertion + Deletion | 1% | 250 | 20 | 270 |
| Non-variation |  | 250 | 20 | 270 |
| Total |  |  |  | 3780 |

**Table S4** Library collection from 22 genera from RepBase and PGSB datasets.

|  |  |  |  |
| --- | --- | --- | --- |
| **Genera** | **Family** | **Class** | **Library number** |
| *Arabidopsis* | Brassicaceae | Dicots | 2499 |
| *Brassica* | Brassicaceae | Dicots | 522 |
| *Lotus* | Nelumbonaceae | Dicots | 816 |
| *Gossypioides* | Malvaceae | Dicots | 584 |
| *Medicago* | Fabaceae | Dicots | 1184 |
| *Glycine* | Fabaceae | Dicots | 5998 |
| *Lycopersicon* | Solanaceae | Dicots | 1786 |
| *Prunus* | Rosaceae | Dicots | 163 |
| *Gossypium* | Malvaceae | Dicots | 11653 |
| *Populus* | Salicaceae | Dicots | 324 |
| *Malus* | Rosaceae | Dicots | 514 |
| *Solanum* | Solanaceae | Dicots | 1849 |
| *Saccharum* | Poaceae | Monocots | 122 |
| *Oryza* | Poaceae | Monocots | 4091 |
| *Brachypodium* | Poaceae | Monocots | 1105 |
| *Zea* | Poaceae | Monocots | 14382 |
| *Sorghum* | Poaceae | Monocots | 7829 |
| *Hordeum* | Poaceae | Monocots | 4725 |
| *Triticum* | Poaceae | Monocots | 4630 |
| *Physcomitrella* | Funariaceae | None | 1109 |
| *Chlamydomonas* | Chlamydomonadaceae | None | 125 |
| *Picea* | Pinaceae | None | 247 |

**Figure S1** Comparison of tool performances with ref-based method for three TE copy types in LTR TEs. REF: Ref-based method. RPMD: RepeatModeler. FIND: LTR\_FINDER. HART: LTRharvest. ‘\_’ indicates tool combination. 1X: 1 copy time. 10X: 10 copy times. 25X: 25 copy times. The color of bars indicates different tools and their combinations.

**Figure S2** Comparison of tool performances for three TE copy types in MITE TEs. REF: Ref-based method. PRMD: RepeatModeler. HUNT: MITE\_Hunter. ‘\_’ indicates tool combination. 1X: 1 copy time. 10X: 10 copy times. 25X: 25 copy times. The color of bars indicates different tools and their combinations.

**Figure S3** Comparison of performances of tools with ref-based method for LTRs detection. a-c indicates TE copy number is one, ten, and twenty-five, respectively. REF: Ref-based method. PRMD: RepeatModeler. FIND: LTR\_FINDER. HART: LTRharvest. ‘\_’ indicates tool combination. Red bar indicates Accuracy. Green bar indicates Precision. Light blue bar indicates Sensitivity. Purple bar indicates Specificity.

|  |  |
| --- | --- |
|  | © 2020 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). |