

## Rebase 2018 Year in review

**Authors:** Kenji K. Kojima, Weidong Bao, and Oleksiy Kohany

**Affiliation:** Genetic Information Research Institute, 465 Fairchild Drive, Suite 201, Mountain View, CA 94043, USA

**Correspondence:** Kenji K. Kojima, [kojima@girinst.org](mailto:kojima@girinst.org)

**Published:** 2019/02/19

### *Update of Rebase 2018*

In the year 2018, Rebase increased the number of entries by 1,801 including 1,400 sequences published on Rebase Reports. The total number of entries reached 49,400 at the end of the year 2018. They include not only consensus and representative sequences of transposable elements (TEs), but also reference sequences for multicopy genes and integrated viruses, and uncharacterized repeats. The total number of TE sequences is 48,041.

### *Species coverage*

At the end of the year 2018, Rebase includes >100 entries from >100 species or >10 entries from >250 species. We have nearly evenly worked on vertebrates, invertebrates, and plants. Rebase contains ~14,000 entries from vertebrates, ~11,000 entries from arthropods, ~14,000 entries from land plants. The numbers of entries for some TE-rich species, such as rice, zebrafish, and mosquito, exceed 2,000 (Table 1). Fungi occupy a relatively small portion with ~2,500 entries. The underrepresentation of protists and fungi in the numbers of entries is partly because of their smaller genome sizes.

Table 2 shows the top 10 species whose entries increased in the year 2018. The characteristics of several species we focused on last year are described below.

### *Oryza sativa* (Asian rice)

In the year 2017, 358 repeat sequences from Asian rice were published. Additional 594 repeat sequences were published last year. At the end of 2018, over 3000 rice repeat sequences are registered in Rebase. Our preliminary analysis suggests the repeat (TE) ratio of reference rice genome “Nipponbare” is around 55%: 20% more than the

estimate done by the International Rice Genome Sequencing Project (2005). This underestimate is still present in the analysis of the 3000 Rice Genomes Project (Wang W. et al. 2018). Because TEs are the major contributors of intraspecies variations, the exhaustive characterization of TEs from this economically important crop & model plant organism definitely contributes to the human welfare.

#### *Bemisia tabaci* (silverleaf whitefly)

This agricultural pest insect belongs to the order Hemiptera. Among Hemiptera, this species has the third most repeat entries (282) in Repbase, following the pea aphid *Acyrtosiphon pisum* (345) and *Rhodnius prolixus*, a triatomine vector of the Chagas disease parasite (296), and the 7th among insects. Silverleaf whitefly has a devastating impact on agriculture because of its nature as a vector of various plant diseases, including African cassava mosaic virus and tomato yellow leaf curl virus, and of its broad host range – its common hosts are crops including tomatoes, squash, broccoli, cotton, sweet potato, and ornamental plants such as poinsettia and roses –.

#### *Gossypium arboreum* (tree cotton)

At the end of 2018, Repbase contains 197 entries from *G. arboreum*. Its close relative, *G. raimondii* has 246 entries and its distant relative, cacao (*Theobroma cacao*), has 143 entries. Cottons and cacao belong to the family Malvales. The genomes of allotetraploid cotton *G. hirsutum* and *G. barbadense* were also sequenced (Wang M. et al. 2018). Both allotetraploid cotton species have the A subgenome related to *G. arboreum* and the D subgenome related to *G. raimondii*. The repeat datasets of *G. arboreum* and *G. raimondii* together contribute to the studies of genome dynamics and evolution of allotetraploid cottons.

#### *Musa acuminata* (wild banana)

Almost all modern edible seedless bananas come from two wild species – *M. acuminata* and *M. balbisiana*. Cavendish bananas are a subgroup of the triploid cultivars of *M. acuminata*. Cultivated bananas contain tiny seeds and thus propagate vegetatively; wild bananas contain large seeds. The genome of *M. acuminata* has few recently active families of TEs.

### ***TE classification***

We have basically not updated our classification scheme since Bao et al. (2015). For more information and comparison with other classification schemes, please read our most recent review article (Kojima 2018). The numbers of Repbase entries in each category is shown in Table 3, including the increase in the number of entries in the last three years.

### ***References***

Bao W, Kojima KK, Kohany O.

Repbase Update, a database of repetitive elements in eukaryotic genomes.

Mob DNA, 2015;6:11.

Kojima KK

Human Transposable Elements in Repbase: Genomic Footprints from Fish to Humans

Mobile DNA, 2018; 9:2.

Kojima KK

Structural and sequence diversity of eukaryotic transposable elements

Genes and Genetic Systems, 2018 advanced online publication on Nov. 9

**Table 1. Top 20 species of the number of Rebase entries.**

<b>Species</b>	<b>Total</b>	<b>2018 Increase</b>
<i>Oryza sativa</i> (Asian rice)	3011	678
<i>Danio rerio</i> (zebrafish)	2322	8
<i>Aedes aegypti</i> (yellow fever mosquito)	2141	0
<i>Chondrus crispus</i> (Irish moss)	1181	0
<i>Locusta migratoria</i> (migratory locust)	1127	-1
<i>Sorghum bicolor</i> (sorghum)	971	0
<i>Chrysemys picta bellii</i> (painted turtle)	843	0
<i>Zea mays</i> (maize)	827	0
<i>Crassostrea gigas</i> (Pacific oyster)	798	0
<i>Nematostella vectensis</i> (starlet sea anemone)	757	0
<i>Hydra vulgaris</i> (hydra)	745	0
<i>Culex quinquefasciatus</i> (southern house mosquito)	660	0
<i>Lepeophtheirus salmonis</i> (salmon louse)	654	0
<i>Homo sapiens</i> (human)	583	2
<i>Glycine max</i> (soybean)	560	3
<i>Xenopus tropicalis</i> (western clawed frog)	534	-1
<i>Salmo salar</i> (Atlantic salmon)	535	0
<i>Malus domestica</i> (apple)	526	0
<i>Arabidopsis thaliana</i> (thale cress)	527	2
<i>Alligator mississippiensis</i> (American alligator)	508	0

**Table 2. Top 10 species whose entries increased in 2018.**

<b>Species</b>	<b>2018 Increase</b>	<b>Total</b>
<i>Oryza sativa</i> (Asian rice)	594	3011
<i>Bemisia tabaci</i> (silverleaf whitefly)	282	282
<i>Gossypium arboreum</i> (tree cotton)	197	197
<i>Musa acuminata</i> (wild banana)	138	141
<i>Schmidtea mediterranea</i> (freshwater planarian)	122	361
<i>Saccharum</i> hybrid cultivar SP80-3280 (sugarcane)	55	55
<i>Pelodiscus sinensis</i> (Chinese soft-shell turtle)	50	52
<i>Solanum tuberosum</i> (potato)	42	427
<i>Erysiphe pisi</i> (powdery mildew)	42	43
<i>Solanum lycopersicum</i> (tomato)	38	64

**Table 3. TE classification and the numbers of entries in Repbase.**

***A. LTR retrotransposon***

<b>Superfamily</b>	<b>Total</b>	<b>2018 increase</b>	<b>2017 increase</b>	<b>2016 increase</b>
<i>Gypsy</i>	11185	590	286	304
<i>Copia</i>	7074	380	170	239
<i>BEL</i>	1872	17	2	26
<i>DIRS</i>	418	0	38	4
Endogenous Retrovirus (ERV)				
<i>ERV1</i>	1987	20	17	71
<i>ERV2</i>	1279	13	3	35
<i>ERV3</i>	670	13	0	8
<i>ERV4</i>	187	0	0	0
<i>Lentivirus</i>	4	0	0	0
Unclassified ERV	325	1	16	0
Unclassified LTR retrotransposon	721	2	132	54
<b>Total</b>	<b>25723</b>	<b>1036</b>	<b>664</b>	<b>741</b>

**B. Non-LTR retrotransposon**

<b>Superfamily (clade)</b>	<b>Total</b>	<b>2018 increase</b>	<b>2017 increase</b>	<b>2016 increase</b>
SINE				
<i>SINE1/7SL</i>	95	0	0	0
<i>SINE2/tRNA</i>	544	5	20	45
<i>SINE3/5S</i>	31	1	0	2
<i>SINEU/snRNA</i>	17	0	0	2
Unclassified SINE	111	-1	0	3
<i>CRE</i>	43	0	0	7
<i>R4</i>	46	0	1	1
<i>Hero</i>	23	0	0	1
<i>NeSL</i>	118	12	9	0
<i>R2</i>	159	0	0	81
<i>RandI/Dualen</i>	13	0	0	5
<i>Proto1</i>	6	0	0	0
<i>L1</i>	1731	41	42	109
<i>Tx1</i>	278	5	1	9
<i>RTETP</i>	1	0	0	0
<i>Proto2</i>	47	0	2	1
<i>RTEX</i>	140	2	3	2
<i>RTE</i>	527	40	12	34
<i>Outcast</i>	23	0	0	0
<i>Ingi</i>	17	0	3	0
<i>Vingi</i>	143	2	8	0
<i>I</i>	203	8	1	0
<i>Nimb</i>	112	4	3	2
<i>Tad1</i>	183	42	1	0
<i>Loa</i>	76	2	0	5
<i>R1</i>	247	10	0	0
<i>Jockey</i>	256	13	1	0
<i>Rex1</i>	95	0	0	8
<i>CR1</i>	851	48	32	39
<i>Kiri</i>	91	0	0	0
<i>L2</i>	286	1	9	56
<i>L2A</i>	5	0	0	0
<i>L2B</i>	29	2	0	1
<i>Crack</i>	140	0	0	3
<i>Daphne</i>	230	3	3	43
<i>Ambal</i>	8	0	0	0
<i>Penelope</i>	494	17	108	18
Unclassified non-LTR retrotransposon	175	-4	57	13
<b>Total</b>	<b>7341</b>	<b>253</b>	<b>316</b>	<b>490</b>

**C. DNA transposon**

<b>Superfamily</b>	<b>Total</b>	<b>2018 increase</b>	<b>2017 increase</b>	<b>2016 increase</b>
<i>EnSpm/CACTA</i>	747	32	29	90
<i>Transib</i>	122	-1	0	0
<i>hAT</i>	3111	156	280	260
<i>MuDR</i>	1382	37	284	362
<i>P</i>	190	1	0	2
<i>Kolobok</i>	294	8	35	9
<i>Dada</i>	36	0	0	0
<i>Mariner/Tc1</i>	2620	81	91	79
<i>Zator</i>	54	0	1	0
<i>piggyBac</i>	377	0	6	18
<i>Merlin</i>	75	0	1	0
<i>Harbinger</i>	1165	68	103	137
<i>ISL2EU</i>	110	22	0	0
<i>Ginger1</i>	39	0	19	0
<i>Ginger2/TDD</i>	21	1	0	1
<i>IS3EU</i>	27	4	0	1
<i>Sola</i>				
<i>Sola1</i>	100	0	0	0
<i>Sola2</i>	92	2	0	1
<i>Sola3</i>	28	0	0	0
Unclassified <i>Sola</i>	1	0	0	0
<i>Academ</i>	90	0	1	1
<i>Novosib</i>	9	0	0	0
<i>Zisupton</i>	19	1	0	2
<i>Helitron</i>	980	25	55	288
<i>Polinton</i>	119	11	4	3
<i>Crypton</i>				
<i>CryptonF</i>	23	0	0	0
<i>CryptonA</i>	17	0	0	0
<i>CryptonI</i>	9	0	0	0
<i>CryptonS</i>	59	0	0	0
<i>CryptonV</i>	50	4	0	2
Unclassified <i>Crypton</i>	87	7	0	0
Unclassified DNA				
transposon	2415	58	135	167
<b>Total</b>	<b>13960</b>	<b>508</b>	<b>1044</b>	<b>1423</b>