

Rebase 2020 Year in review

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

Affiliation: Genetic Information Research Institute, 20380 Town Center Lane Suite 240, Cupertino, CA 95014, USA

Correspondence: Kenji K. Kojima, kojima@girinst.org

Published: 2021/02/18

Update of Rebase 2020

In the year 2020, Rebase increased the number of entries by 5,624, including 4,342 sequences published on Rebase Reports. The total number of entries reached 58,126 at the end of the year 2020. These numbers include not only consensus and representative sequences of transposable elements (TEs) but also reference sequences for satellite repeats and microsatellites (593 entries), multicopy genes (195 entries), integrated viruses (196 entries), and uncharacterized repeats. The total number of TE sequences is 56,652.

Rebase transitioned to the subscription-based model on April 12, 2019. Since then, we have (1) increased the number of repeats published in each issue of Rebase Reports, and (2) covered more diverse organisms each month (Table 1). The choice of organisms is based on (1) the economic and scientific importance, (2) the quality of genome sequence, and (3) the phylogenetical distance from well-analyzed organisms. We also have continually updated, corrected and refined existing repeats that are with sequence flaws or ambiguously classified. We updated some 200 entries in the year 2020.

At the end of the year 2020, Rebase includes >100 entries from >100 species or >10 entries from >300 species. The numbers of entries for three TE-rich species (yellow fever mosquito, Asian rice, and zebrafish) exceed 2,000 (Table 2). Sometimes, multiple genomes in the same genus, such as *Drosophila* and *Arabidopsis*, have been sequenced and analyzed. Such analysis helps characterize very low-copy TE families or single-copy TE families in model organisms and reveal the genome evolution for longer time scale. Table 3 shows the total numbers of Rebase entries for each genus. In some cases, such as *Chondrus* and *Locusta*, only one species is analyzed, while in other cases, such as *Drosophila* and *Puccinia*, more than five species have been examined (Tables 4 and 5). Table 6 shows the top 10 species whose entries increased in

the year 2020, and Table 7 shows the updates of some model organisms. Even though the genomes of model organisms are very well analyzed, they still contain TE families to be discovered. The characteristics of several species we focused on in the year 2020 are described below.

Aedes aegypti (yellow fever mosquito)

The yellow fever mosquito is the pest that can spread devastating diseases including yellow fever, dengue fever, and Zika fever. The genome of this pest insect is very TE family-rich; especially LTR retrotransposons are very diverse. Reported 2,726 LTR retrotransposons from *A. aegypti* are now classified into 1,165 *Gypsy*, 366 *Copia*, and 1,195 *BEL* LTR retrotransposons. It is noteworthy that LTR retrotransposons deposited in Repbase are split into two parts, LTR and internal portion.

Arabidopsis thaliana (thale cress) and *Arabidopsis lyrata*

Arabidopsis thaliana is the most well-studied land plant species. It belongs to the family Brassicaceae, which includes several economically important crops, such as cabbage, Chinese cabbage, rapeseed, and radish. In Repbase, there are 222 entries from radish (*Raphanus sativus*), 381 from Chinese cabbage (*Brassica rapa*), 99 from cabbage (*Brassica oleracea*), and 84 from *Schrenkiella parvula/Eutrema parvum*. Repbase now contains 956 and 769 repeat entries from *A. thaliana* and its close relative *A. lyrata*, respectively.

Crassostrea gigas (Pacific oyster) and *Crassostrea virginica* (eastern oyster)

The Pacific oyster (*Crassostrea gigas*) is now the most widely farmed and the most commercially important oyster globally. It is native to Asia. It is also considered an invasive species where it has been introduced intentionally or accidentally. The eastern oyster (*Crassostrea virginica*) is native to the Atlantic Ocean, and is also cultivated in some parts of North America. In the year 2020, the entries from *C. gigas* increased by 281, and *C. virginica* entries increased by 167.

Oryzias latipes (medaka, Japanese ricefish)

Medaka or Japanese ricefish is one of the well-established model fish species. Despite the long history of research and the *Tol2* transposition system, the mobilome of medaka has not yet been well analyzed. We published 281 repeat families in the year 2020.

Panicum miliaceum (proso millet) and *Echinochloa crus-galli* (Japanese millet)

P. miliaceum is one of the earliest domesticated crops, cultivated already before 10,000 BCE in China. Japanese barnyard millet, *Echinochloa esculenta*, is considered to have been domesticated from *E. crus-galli* in East Asia, which itself is regarded as one of the worst weeds that reduce crop yields. These two millet species belong to the subfamily Panicoideae in the family Poaceae. Panicoideae includes the maize *Zea mays*, the sorghum *Sorghum bicolor*, and the sugarcane *Saccharum*. The analyses of the two millet species helps the studies of these economically important crop species too.

Puccinia hordei, *P. striiformis*, and other *Puccinia* species (rust)

All species in the genus *Puccinia* are obligate plant pathogens and are known as rusts. The genus *Puccinia*, which contains ~4,000 species, is considered the most economically destructive genus of biotrophic fungi; members of this genus are serious pathogens on all major cereal crop species except rice. Repbase contains entries from several *Puccinia* species and many of them were added in the year 2020 (Table 5).

Mus musculus (domestic mouse)

The domestic mouse *Mus musculus* is a widely used model mammalian species. Despite the fact that the draft mouse genome was published 18 years ago, there remain uncharacterized repeats in the mouse genome, and 35 new repeat sequences were added in the year 2019. In the year 2020, additional 115 new repeat sequences were reclaimed. Besides 463 entries just for *Mus musculus*, there are many entries of repeats distributed more widely in rodents, such as 48 for the genus *Mus*, 101 for Muridae, 44 for Rodentia, 229 for Eutheria, and 225 for Mammalia. In total, 1,347 repeat families deposited in Repbase might be found in the mouse genome, whereas some ancient families may have been eliminated from the genome completely. Repbase also contains 38 entries from the Gairdner's shrewmouse (*Mus pahari*) and two from *Mus caroli* (Ryukyu mouse).

Drosophila melanogaster (fruit fly) and other fruit flies

The genus *Drosophila* contains over 1,450 species, among which ~1,100 species belong to the subgenus *Drosophila*. *D. melanogaster* belongs to the subgenus *Sophophora*. In the year 2020, 41 sequences from *Drosophila melanogaster* were added to Repbase, as well as 102 TE families from *D. azteca*, 92 from *D. mojavensis*, 37 from *D. hydei*, and 2 from *D. willistoni*. In

total, Repbase contains 2,701 repeat entries from the genus *Drosophila*. Table 4 shows the *Drosophila* species whose repeat families in Repbase are over 10.

TE classification

We have 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 2019). The number of Repbase entries in each category is shown in Table 8, including the increase in the number of entries in the last three years. The number of entries in the *Academ* superfamily was more than tripled. It is noteworthy that our continuous efforts for classification result in a decrease in the number of unclassified DNA transposons.

References

Bao W, Kojima KK, Kohany O.

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

Mob DNA, 2015;6:11.

Kojima KK

Structural and sequence diversity of eukaryotic transposable elements

Genes and Genetic Systems, 2020; 94(6):233-252.

Table 1. Organisms published in each issue of *Rebase Reports*.

RR issue	Published entries	Vertebrate	Invertebrate	Plant	Fungi
20(1)	272	<i>Numida</i> (guineafowl), <i>Chelonia</i> (sea turtle), <i>Ceratotherium</i> (rhinoceros)	<i>Aedes</i> (mosquito), <i>Cimex</i> (bed bug)	<i>Vaccinium</i> (cranberry)	
20(2)	310	<i>Xenopus</i> (frog), <i>Mus</i> (mouse)	<i>Eulimnadia</i> (clam shrimp), <i>Crassostrea</i> (oyster)	<i>Ficus</i> (fig), <i>Citrus</i> (orange), <i>Morus</i> (mulberry)	<i>Margaritispora</i> , <i>Monascus</i> , <i>Rhodotorula</i>
20(3)	315		<i>Drosophila</i> (fruit fly)	<i>Panicum</i> (millet)	<i>Calonectria</i> , <i>Penicillium</i> , <i>Parastagonospora</i> , <i>Raffaelea</i>
20(4)	333	<i>Strigops</i> (kakapo), <i>Ovis</i> (sheep), <i>Cavia</i> (guinea pig)	<i>Crassostrea</i> (oyster)	<i>Echinochloa</i> (millet)	<i>Puccinia</i> , <i>Verticillium</i> , <i>Xylaria</i> , <i>Byssochlamys</i> , <i>Alternaria</i> , <i>Colletotrichum</i>
20(5)	296	<i>Homo</i> (human)	<i>Crassostrea</i> (oyster)	<i>Oryza</i> (rice), <i>Arachis</i> (peanut), <i>Echinochloa</i> (millet), <i>Saccharum</i> (sugarcane)	<i>Rhizophagus</i> , <i>Aspergillus</i>
20(6)	313	<i>Mus</i> (mouse), <i>Leptobrachium</i> (toad), <i>Zalophus</i> (sea lion)	<i>Crassostrea</i> (oyster)	<i>Arachis</i> (peanut), <i>Vaccinium</i> (cranberry, blueberry)	<i>Teratosphaeria</i> , <i>Sarocladium</i> , <i>Cutaneotrichosporon</i> , <i>Amphirosellinia</i> , <i>Fragosphaeria</i> , <i>Floccularia</i> , <i>Paecilomyces</i> , <i>Hortaea</i>
20(7)	311	<i>Amblyraja</i> (skate)	<i>Aedes</i> (mosquito)		<i>Puccinia</i> , <i>Elsinoe</i> , <i>Grosmannia</i>
20(8)	304	<i>Amblyraja</i> (skate)	<i>Drosophila</i> (fruit fly)	<i>Arabidopsis</i>	<i>Cladobotryum</i> , <i>Colletotrichum</i> , <i>Raffaelea</i>
20(9)	360	<i>Oryzias</i> (medaka)		<i>Arabidopsis</i>	<i>Microbotryum</i>
20(10)	409	<i>Amblyraja</i> (skate)	<i>Aedes</i> (mosquito)	<i>Raphanus</i> (radish)	
20(11)	539	<i>Oryzias</i> (medaka), <i>Anguilla</i> (eel), <i>Varanus</i> (Komodo dragon)	<i>Notospermus</i> (ribbon worm)	<i>Zea</i> (maize), <i>Raphanus</i> (radish), <i>Oryza</i> (rice)	<i>Puccinia</i>
20(12)	549	<i>Chanos</i> (milkfish), <i>Sphenodon</i> (tuatara)	<i>Crassostrea</i> (oyster), <i>Spodoptera</i> (moth)	<i>Arabidopsis</i> , <i>Raphanus</i> (radish), <i>Daucus</i> (carrot), <i>Panicum</i> (millet), <i>Echinochloa</i> (millet)	<i>Puccinia</i>

Species whose repeats are published first in Rebase Reports are colored in red.

Table 2. Top 20 species based on the number of Rebase entries.

Species	Total	2020 Increase
<i>Aedes aegypti</i> (yellow fever mosquito)	3319	727
<i>Oryza sativa</i> (Asian rice)	3075	62
<i>Danio rerio</i> (zebrafish)	2323	0
<i>Chondrus crispus</i> (Irish moss)	1183	0
<i>Locusta migratoria</i> (migratory locust)	1127	0
<i>Crassostrea gigas</i> (Pacific oyster)	1072	281
<i>Zea mays</i> (maize)	1012	126
<i>Sorghum bicolor</i> (sorghum)	971	0
<i>Arabidopsis thaliana</i> (thale cress)	956	423
<i>Chrysemys picta bellii</i> (painted turtle)	843	0
<i>Arabidopsis lyrata</i> (lyrate rockcress)	769	216
<i>Nematostella vectensis</i> (starlet sea anemone)	758	1
<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)	582	-1
<i>Glycine max</i> (soybean)	558	0
<i>Xenopus tropicalis</i> (western clawed frog)	555	21
<i>Salmo salar</i> (Atlantic salmon)	535	0
<i>Malus domestica</i> (apple)	526	0

Table 3. Top 20 genera based on the number of Rebase entries.

Genus (Species with >10 entries)	Total	2020 Increase
<i>Aedes (A. aegypti)</i>	3320	727
<i>Oryza (O. sativa)</i>	3084	62
<i>Drosophila</i> ¹	2801	274
<i>Danio (D. rerio)</i>	2323	0
<i>Arabidopsis (A. thaliana, A. lyrata)</i>	1725	639
<i>Crassostrea (C. gigas, C. virginica)</i>	1244	448
<i>Chondrus (C. crispus)</i>	1183	0
<i>Locusta (L. migratoria)</i>	1127	0
<i>Zea (Z. mays)</i>	1016	126
<i>Sorghum (S. bicolor)</i>	971	0
<i>Puccinia</i> ²	886	399
<i>Xenopus (X. tropicalis, X. laevis)</i>	846	21
<i>Chrysemys (C. picta)</i>	843	0
<i>Phytophthora (P. infestans, P. sojae, P. ramorum, P. parasitica)</i>	783	0
<i>Nematostella (N. vectensis)</i>	759	1
<i>Hydra (H. vulgaris)</i>	745	0
<i>Culex (C. quinquefasciatus)</i>	669	0
<i>Lepeophtheirus (L. salmonis)</i>	654	0
<i>Homo (H. sapiens)</i>	582	0
<i>Solanum (S. tuberosum, S. lycopersicum, S. demissum)</i>	575	0

1 Species are shown in Table 4.

2 Species are shown in Table 5.

Table 4. Repbase entries from the genus *Drosophila*.

Subgenus - Group (subgroup)	Species	Abbr.	Repbase entries	2020 Increase	
<i>Sophophora - melanogaster</i> (<i>melanogaster</i> subgroup)	<i>D. melanogaster</i>	DM	289	41	
	<i>D. simulans</i>	DSim	43	0	
	<i>D. sechellia</i>	DSe	24	0	
	<i>D. yakuba</i>	DY	136	0	
	(<i>suzukii</i> subgroup)	<i>D. biarmipes</i>	DBi	80	0
	(<i>ananassae</i> subgroup)	<i>D. ananassae</i>	DAn	240	0
		<i>D. bipectinata</i>	DBp	158	0
	(<i>elegans</i> subgroup)	<i>D. elegans</i>	DEl	169	0
	(<i>takahashii</i> subgroup)	<i>D. takahashii</i>	DTa	144	0
	(<i>rhopaloo</i> subgroup)	<i>D. rhopaloo</i>	DRh	116	0
	(<i>montium</i> subgroup)	<i>D. kikkawai</i>	DKi	97	0
		<i>D. serrata</i>	DSer	71	0
	<i>Sophophora - obscura</i>	<i>D. azteca</i>	DAzt	102	102
<i>D. pseudoobscura</i>		DPse	80	0	
<i>D. persimilis</i>		DPer	106	0	
<i>Sophophora - ficusphila</i>	<i>D. ficusphila</i>	DF	129	0	
<i>Sophophora - eugracilis</i>	<i>D. eugracilis</i>	DEu	78	0	
<i>Sophophora - willistoni</i>	<i>D. willistoni</i>	DWil	241	2	
<i>Drosophila - replete</i>	<i>D. mojavensis</i>	DMoj	154	92	
	<i>D. hydei</i>	DHyd	39	37	
<i>Drosophila - virilis</i>	<i>D. virilis</i>	DVi	88	0	
<i>Hawaiian Drosophila</i>	<i>D. grimshawi</i>	DGri	41	0	

¹ Only species with >10 entries are shown.

² Subgroups are shown only in the *melanogaster* group.

Table 5. Repbase entries from the genus *Puccinia*.

Species	Abbr.	Repbase entries	2020 Increase
<i>P. striiformis</i> (stripe rust, or yellow rust)	PSt	291	130
<i>P. graminis</i> (stem rust, or black rust)	PG/PGra	233	0
<i>P. hordei</i> (barley brown rust)	PHord	157	157
<i>P. triticina</i> (wheat leaf rust, or brown rust)	PTrit	120	27
<i>P. coronata</i> (barley crown rust)	PCor	59	59
<i>P. horiana</i> (chrysanthemum white rust)	PHor	25	25

¹ Only species with >10 entries are shown.

Table 6. Top 20 species whose entries increased in 2020.

Species	2020 Increase	Total
<i>Aedes aegypti</i> (yellow fever mosquito)	727	3319
<i>Arabidopsis thaliana</i> (thale cress)	423	956
<i>Crassostrea gigas</i> (Pacific oyster)	281	1072
<i>Oryzias latipes</i> (medaka)	281	374
<i>Agaricus bisporus</i> (portabella mushroom)	232	246
<i>Raphanus sativus</i> (radish)	222	222
<i>Arabidopsis lyrata</i> (lyrate rockcress)	216	769
<i>Amblyraja radiata</i> (thorny skate)	182	182
<i>Crassostrea virginica</i> (eastern oyster)	167	172
<i>Echinochloa crus-galli</i> (Japanese millet)	165	165
<i>Panicum miliaceum</i> (proso millet)	159	159
<i>Puccinia hordei</i> (rust)	157	105
<i>Spodoptera exigua</i> (beet armyworm)	138	138
<i>Zea mays</i> (maize)	126	1012
<i>Puccinia striiformis</i> (rust)	124	291
<i>Arachis hypogaea</i> (peanut)	119	119
<i>Mus musculus</i> (house mouse)	114	463
<i>Drosophila azteca</i> (fruit fly)	102	102
<i>Notospermus geniculatus</i>	88	89
<i>Daucus carota</i> (carrot)	84	160

Table 7. Update of model organisms in 2020.

Species	2020 Increase	Total
<i>Arabidopsis thaliana</i> (thale cress)	423	956
<i>Oryzias latipes</i> (medaka)	281	374
<i>Zea mays</i> (maize)	126	1012
<i>Mus musculus</i> (house mouse)	114	463
<i>Oryza sativa</i> (Asian rice)	62	3075
<i>Drosophila melanogaster</i> (fruit fly)	41	289
<i>Xenopus tropicalis</i> (western clawed frog)	21	555

Table 8. TE classification and the numbers of entries in Repbase.

A. LTR retrotransposon

Superfamily	Total	2020 increase	2019 increase	2018 increase
<i>Gypsy</i>	13460	1843	432	590
<i>Copia</i>	7957	569	314	380
<i>BEL</i>	2615	432	311	17
<i>DIRS</i>	645	186	41	0
Endogenous Retrovirus (ERV)				
<i>ERV1</i>	2262	115	160	20
<i>ERV2</i>	1563	102	182	13
<i>ERV3</i>	817	18	129	13
<i>ERV4</i>	193	0	6	0
<i>Lentivirus</i>	4	0	0	0
Unclassified ERV	328	0	3	1
Unclassified LTR retrotransposon	728	44	-37	2
Total	30573	3310	1541	1036

B. Non-LTR retrotransposon

Superfamily (clade)	Total	2020 increase	2019 increase	2018 increase
SINE				
<i>SINE1/7SL</i>	100	1	4	0
<i>SINE2/tRNA</i>	678	60	74	5
<i>SINE3/5S</i>	32	0	1	1
<i>SINEU/snRNA</i>	17	0	0	0
Unclassified SINE	112	1	0	-1
<i>CRE</i>	50	0	7	0
<i>R4</i>	53	1	6	0
<i>Hero</i>	28	4	1	0
<i>NeSL</i>	118	0	0	12
<i>R2</i>	160	0	1	0
<i>RandI/Dualen</i>	13	0	0	0
<i>Proto1</i>	6	0	0	0
<i>L1</i>	1971	184	56	41
<i>Tx1</i>	304	21	5	5
<i>RTETP</i>	1	0	0	0
<i>Proto2</i>	52	5	0	0
<i>RTEX</i>	148	4	4	2
<i>RTE</i>	553	12	14	40
<i>Outcast</i>	23	0	0	0
<i>Ingi</i>	22	0	5	0
<i>Vingi</i>	153	10	0	2
<i>I</i>	204	1	0	8
<i>Nimb</i>	117	2	3	4
<i>Tad1</i>	310	19	108	42
<i>Loa</i>	87	6	5	2
<i>R1</i>	266	9	10	10
<i>Jockey</i>	279	18	5	13
<i>Rex1</i>	126	26	5	0
<i>CR1</i>	873	17	5	48
<i>Kiri</i>	92	1	0	0
<i>L2</i>	297	8	3	1
<i>L2A</i>	5	0	0	0
<i>L2B</i>	29	0	0	2
<i>Crack</i>	142	2	0	0
<i>Daphne</i>	238	2	6	3
<i>Ambal</i>	8	0	0	0
<i>Penelope</i>	532	11	27	17
Unclassified non-LTR retrotransposon	185	5	0	-4
Total	8384	430	360	253

C. DNA transposon

Superfamily	Total	2020 increase	2019 increase	2018 increase
<i>EnSpm/CACTA</i>	913	156	10	32
<i>Transib</i>	143	8	13	-1
<i>hAT</i>	4072	432	529	156
<i>MuDR</i>	1730	182	166	37
<i>P</i>	214	15	9	1
<i>Kolobok</i>	334	6	34	8
<i>Dada</i>	51	15	0	0
<i>Mariner/Tc1</i>	3020	171	229	81
<i>Zator</i>	83	22	7	0
<i>piggyBac</i>	421	21	23	0
<i>Merlin</i>	90	1	14	0
<i>Harbinger</i>	1559	279	115	68
<i>ISL2EU</i>	142	29	3	22
<i>Ginger1</i>	39	0	0	0
<i>Ginger2/TDD</i>	39	6	12	1
<i>IS3EU</i>	45	10	8	4
<i>Sola</i>				
<i>Sola1</i>	107	5	2	0
<i>Sola2</i>	98	6	0	2
<i>Sola3</i>	42	13	1	0
<i>Academ</i>	329	239	0	0
<i>Novosib</i>	9	0	0	0
<i>Zisupton</i>	41	19	3	1
<i>Helitron</i>	1222	201	41	25
<i>Polinton</i>	142	10	13	11
<i>Crypton</i>				
<i>CryptonF</i>	24	1	0	0
<i>CryptonA</i>	18	0	1	0
<i>CryptonI</i>	10	0	1	0
<i>CryptonS</i>	59	0	0	0
<i>CryptonV</i>	53	0	3	4
Unclassified <i>Crypton</i>	89	2	0	7
Unclassified DNA transposon	2309	-53	-53	58
Total	17447	1796	1183	508