Supplementary materials:

**Table S1. The length distribution of IS elements and their IRs and Tpase ORFs annotated for IS families in ISfinder.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| family | minIR | maxIR | minIS | maxIS | minPEP |
| IS1 | 8 | 67 | 732 | 4601 | 252 |
| IS110 | 2 | 31 | 969 | 4105 | 156 |
| IS1182 | 8 | 44 | 1330 | 1980 | 570 |
| IS1380 | 7 | 39 | 1474 | 4160 | 1158 |
| IS1595 | 10 | 43 | 701 | 7915 | 426 |
| IS1634 | 11 | 12 | 1511 | 2089 | 1314 |
| IS200/IS605 | 0\* | 0\* | 407 | 2223 | 147 |
| IS21 | 8 | 76 | 1924 | 3533 | 231 |
| IS256 | 8 | 48 | 1124 | 1629 | 990 |
| IS3 | 7 | 54 | 435 | 1814 | 120 |
| IS30 | 11 | 50 | 1027 | 8273 | 189 |
| IS4 | 8 | 67 | 521 | 5396 | 219 |
| IS481 | 5 | 52 | 553 | 3451 | 447 |
| IS5 | 7 | 45 | 789 | 5396 | 75 |
| IS6 | 12 | 36 | 696 | 1648 | 246 |
| IS607 | 12 | 46 | 1415 | 2607 | 453 |
| IS630 | 3 | 92 | 895 | 2009 | 318 |
| IS66 | 11 | 144 | 1364 | 3481 | 165 |
| IS701 | 12 | 38 | 1016 | 2207 | 921 |
| IS91 | 11 | 21 | 712 | 2604 | 648 |
| IS982 | 11 | 35 | 845 | 1282 | 429 |
| ISAS1 | 12 | 34 | 1139 | 3041 | 189 |
| ISAZO13 | 18 | 48 | 1284 | 2171 | 513 |
| ISH3 | 11 | 31 | 1200 | 1509 | 549 |
| ISKRA4 | 15 | 40 | 1164 | 3746 | 114 |
| ISL3 | 6 | 50 | 536 | 9109 | 408 |
| ISNCY | 4 | 52 | 786 | 3989 | 123 |
| new | 10 | 50 | 400 | 10000 | 50 |

nISs: number of IS; minIR: length of shortest IR; maxIR: length of longest IR, minIS: length of shortest IS; maxIS: length of longest IS; minPEP: length of shortest transposase ORF. \*: All ISs in family IS200/IS605 have no annotated IR except IS1341 with IR of 11 bp.

**Table S2. Performance variation over different overlap tolerances on ISbrowser data set**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ISEScan | | | | | | OASIS | | | | | |
| Overlap | Sn (%) | Ntp | N | Fdr (%) | Nh-Ntp | Nh | Sn (%) | Ntp | N | Fdr (%) | Nh-Ntp | Nh |
| 0.50 | 92 | 790 | 858 | 20 | 198 | 988 | 66 | 568 | 858 | 25 | 185 | 753 |
| 0.60 | 91 | 783 | 858 | 21 | 205 | 988 | 63 | 539 | 858 | 28 | 214 | 753 |
| 0.70 | 89 | 767 | 858 | 22 | 221 | 988 | 62 | 529 | 858 | 30 | 224 | 753 |
| 0.80 | 88 | 753 | 858 | 24 | 235 | 988 | 60 | 518 | 858 | 31 | 235 | 753 |
| 0.90 | 86 | 736 | 858 | 26 | 252 | 988 | 58 | 496 | 858 | 34 | 257 | 753 |
| 0.98 | 77 | 662 | 858 | 33 | 326 | 988 | 48 | 413 | 858 | 45 | 340 | 753 |
| 0.99 | 65 | 557 | 858 | 44 | 431 | 988 | 38 | 329 | 858 | 56 | 424 | 753 |

Columns: Overlap is the overlap rate between IS annotated by ISEScan (or OASIS) and the corresponding IS annotated in ISbrowser; Sn and Fdr are sensitivity and false discovery rate respectively; N and Nh are the total numbers of ISs annotated by ISbrowser and ISEScan (or OASIS) respectively; Ntp is the number of the matched ISs in ISbrowser; Nh-Ntp is the number of the ISs falsely annotated by ISEScan (or OASIS).

**Table S3. Performance variation over different overlap tolerances on E. coli MG1655 genome**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ISEScan | | | | | | OASIS | | | | | |
| Overlap | Sn (%) | Ntp | N | Fdr (%) | Nh-Ntp | Nh | Sn (%) | Ntp | N | Fdr (%) | Nh-Ntp | Nh |
| 0.50 | 100 | 40 | 40 | 20 | 10 | 50 | 95 | 38 | 40 | 12 | 5 | 43 |
| 0.60 | 100 | 40 | 40 | 20 | 10 | 50 | 95 | 38 | 40 | 12 | 5 | 43 |
| 0.70 | 100 | 40 | 40 | 20 | 10 | 50 | 95 | 38 | 40 | 12 | 5 | 43 |
| 0.80 | 100 | 40 | 40 | 20 | 10 | 50 | 93 | 37 | 40 | 14 | 6 | 43 |
| 0.90 | 100 | 40 | 40 | 20 | 10 | 50 | 93 | 37 | 40 | 14 | 6 | 43 |
| 0.98 | 98 | 39 | 40 | 22 | 11 | 50 | 88 | 35 | 40 | 19 | 8 | 43 |
| 0.99 | 75 | 30 | 40 | 40 | 20 | 50 | 88 | 35 | 40 | 19 | 8 | 43 |

Columns: Overlap is the overlap rate between IS annotated by ISEScan (or OASIS) and the corresponding IS annotated in the curated IS annotation data; Sn and Fdr are sensitivity and false discovery rate respectively; N and Nh are the total numbers of ISs annotated by the curated IS annotation data and ISEScan (or OASIS) respectively; Ntp is the number of the matched ISs in the curated IS annotation data; Nh-Ntp is the number of the ISs falsely annotated by ISEScan (or OASIS).

**Table S4. Performance on ISbrowser benchmark data set**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ISEScan | | | | | | OASIS | | | | | |
| Sequence | Sn (%) | Ntp | N | Fdr (%) | Nh-Ntp | Nh | Sn (%) | Ntp | N | Fdr (%) | Nh-Ntp | Nh |
| NC\_000915 | 83 | 5 | 6 | 17 | 1 | 6 | 83 | 5 | 6 | 50 | 5 | 10 |
| NC\_001264 | 100 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| NC\_001763 | 100 | 2 | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 |
| NC\_001869 | 68 | 17 | 25 | 19 | 4 | 21 | 24 | 6 | 25 | 40 | 4 | 10 |
| NC\_002570 | 100 | 85 | 85 | 4 | 4 | 89 | 95 | 81 | 85 | 17 | 17 | 98 |
| NC\_002578 | 50 | 1 | 2 | 75 | 3 | 4 | 0 | 0 | 2 | 0 | 0 | 0 |
| NC\_002607 | 70 | 16 | 23 | 6 | 1 | 17 | 0 | 0 | 23 | 0 | 0 | 0 |
| NC\_002608 | 94 | 29 | 31 | 26 | 10 | 39 | 35 | 11 | 31 | 39 | 7 | 18 |
| NC\_002689 | 83 | 5 | 6 | 38 | 3 | 8 | 33 | 2 | 6 | 60 | 3 | 5 |
| NC\_002754 | 90 | 132 | 146 | 29 | 54 | 186 | 58 | 85 | 146 | 44 | 67 | 152 |
| NC\_003106 | 94 | 32 | 34 | 27 | 12 | 44 | 9 | 3 | 34 | 73 | 8 | 11 |
| NC\_003413 | 97 | 28 | 29 | 7 | 2 | 30 | 93 | 27 | 29 | 4 | 1 | 28 |
| NC\_003454 | 100 | 17 | 17 | 15 | 3 | 20 | 65 | 11 | 17 | 48 | 10 | 21 |
| NC\_005024 | 100 | 7 | 7 | 0 | 0 | 7 | 100 | 7 | 7 | 0 | 0 | 7 |
| NC\_005054 | 100 | 6 | 6 | 0 | 0 | 6 | 100 | 6 | 6 | 0 | 0 | 6 |
| NC\_005127 | 100 | 2 | 2 | 0 | 0 | 2 | 0 | 0 | 2 | 0 | 0 | 0 |
| NC\_005838 | 100 | 9 | 9 | 18 | 2 | 11 | 11 | 1 | 9 | 0 | 0 | 1 |
| NC\_006392 | 100 | 3 | 3 | 0 | 0 | 3 | 0 | 0 | 3 | 0 | 0 | 0 |
| NC\_006393 | 75 | 9 | 12 | 55 | 11 | 20 | 33 | 4 | 12 | 33 | 2 | 6 |
| NC\_006395 | 100 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| NC\_006396 | 67 | 6 | 9 | 54 | 7 | 13 | 56 | 5 | 9 | 17 | 1 | 6 |
| NC\_006397 | 100 | 6 | 6 | 14 | 1 | 7 | 83 | 5 | 6 | 29 | 2 | 7 |
| NC\_006425 | 100 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| NC\_006461 | 92 | 11 | 12 | 0 | 0 | 11 | 33 | 4 | 12 | 0 | 0 | 4 |
| NC\_006462 | 89 | 8 | 9 | 20 | 2 | 10 | 67 | 6 | 9 | 0 | 0 | 6 |
| NC\_006493 | 100 | 1 | 1 | 50 | 1 | 2 | 0 | 0 | 1 | 0 | 0 | 0 |
| NC\_006624 | 100 | 1 | 1 | 83 | 5 | 6 | 0 | 0 | 1 | 0 | 0 | 0 |
| NC\_007349 | 100 | 1 | 1 | 67 | 2 | 3 | 0 | 0 | 1 | 0 | 0 | 0 |
| NC\_007351 | 100 | 3 | 3 | 0 | 0 | 3 | 100 | 3 | 3 | 0 | 0 | 3 |
| NC\_007352 | 100 | 3 | 3 | 0 | 0 | 3 | 100 | 3 | 3 | 0 | 0 | 3 |
| NC\_007426 | 50 | 1 | 2 | 94 | 17 | 18 | 100 | 2 | 2 | 78 | 7 | 9 |
| NC\_007427 | 100 | 3 | 3 | 67 | 6 | 9 | 33 | 1 | 3 | 67 | 2 | 3 |
| NC\_007641 | 100 | 2 | 2 | 0 | 0 | 2 | 100 | 2 | 2 | 0 | 0 | 2 |
| NC\_007643 | 100 | 7 | 7 | 30 | 3 | 10 | 71 | 5 | 7 | 0 | 0 | 5 |
| NC\_007681 | 100 | 3 | 3 | 40 | 2 | 5 | 0 | 0 | 3 | 0 | 0 | 0 |
| NC\_007796 | 98 | 49 | 50 | 13 | 7 | 56 | 86 | 43 | 50 | 39 | 28 | 71 |
| NC\_008010 | 97 | 28 | 29 | 15 | 5 | 33 | 79 | 23 | 29 | 15 | 4 | 27 |
| NC\_008025 | 100 | 30 | 30 | 6 | 2 | 32 | 80 | 24 | 30 | 0 | 0 | 24 |
| NC\_008541 | 100 | 6 | 6 | 33 | 3 | 9 | 83 | 5 | 6 | 0 | 0 | 5 |
| NC\_008711 | 100 | 8 | 8 | 11 | 1 | 9 | 100 | 8 | 8 | 0 | 0 | 8 |
| NC\_008712 | 83 | 5 | 6 | 29 | 2 | 7 | 50 | 3 | 6 | 40 | 2 | 5 |
| NC\_008713 | 0 | 0 | 1 | 100 | 3 | 3 | 0 | 0 | 1 | 100 | 1 | 1 |
| NC\_009012 | 84 | 81 | 96 | 7 | 6 | 87 | 93 | 89 | 96 | 7 | 7 | 96 |
| NC\_009138 | 100 | 13 | 13 | 0 | 0 | 13 | 69 | 9 | 13 | 0 | 0 | 9 |
| NC\_009675 | 88 | 7 | 8 | 36 | 4 | 11 | 13 | 1 | 8 | 67 | 2 | 3 |
| NC\_009939 | 100 | 13 | 13 | 19 | 3 | 16 | 92 | 12 | 13 | 14 | 2 | 14 |
| NC\_010655 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| NC\_010943 | 97 | 28 | 29 | 15 | 5 | 33 | 86 | 25 | 29 | 7 | 2 | 27 |
| NC\_011071 | 100 | 16 | 16 | 0 | 0 | 16 | 94 | 15 | 16 | 0 | 0 | 15 |
| NC\_011145 | 100 | 17 | 17 | 0 | 0 | 17 | 65 | 11 | 17 | 0 | 0 | 11 |
| NC\_011666 | 100 | 6 | 6 | 0 | 0 | 6 | 0 | 0 | 6 | 0 | 0 | 0 |
| NC\_011886 | 100 | 1 | 1 | 50 | 1 | 2 | 0 | 0 | 1 | 0 | 0 | 0 |
| NC\_011891 | 100 | 18 | 18 | 0 | 0 | 18 | 83 | 15 | 18 | 6 | 1 | 16 |
| total | 92 | 790 | 858 | 20 | 198 | 988 | 66 | 568 | 858 | 25 | 185 | 753 |

Columns: Sequence is the accession ID of NCBI reference genome sequence; Sn and Fdr are sensitivity and false discovery rate respectively; N and Nh are the total numbers of ISs annotated by ISbrowser and ISEScan (or OASIS) respectively; Ntp is the number of the matched ISs in ISbrowser; Nh-Ntp is the number of the ISs falsely annotated by ISEScan (or OASIS).

**Table S5. The manually curated IS elements in E.coli K-12 MG1655**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| name | family | strand | start | end | length | type |
| IS186A | IS4 | + | 15387 | 16731 | 1345 | f |
| IS1A | IS1 | - | 19796 | 20563 | 768 | f |
| IS1I | IS1 | - | 257908 | 258675 | 768 | f |
| IS911A | IS3 | + | 270206 | 270540 | 335 | p |
| IS30A | IS30 | + | 270541 | 271761 | 1221 | f |
| IS911A | IS3 | + | 271762 | 272190 | 429 | p |
| IS5A | IS5 | - | 273955 | 275149 | 1195 | f |
| IS1B | IS1 | - | 279163 | 279930 | 768 | f |
| IS30B | IS30 | + | 279931 | 280111 | 181 | p |
| ISX | IS3 | - | 280114 | 280425 | 312 | p |
| IS1C | IS1 | - | 290634 | 291401 | 768 | f |
| IS3A | IS3 | + | 315229 | 316483 | 1255 | f |
| IS2A | IS3 | + | 381260 | 382590 | 1331 | f |
| IS3B | IS3 | - | 391709 | 392966 | 1258 | f |
| IS3C | IS3 | + | 566777 | 568034 | 1258 | f |
| IS5B | IS5 | - | 574591 | 575785 | 1195 | f |
| IS186B | IS4 | + | 608007 | 609351 | 1345 | f |
| IS5D | IS5 | - | 687851 | 689045 | 1195 | f |
| IS1D | IS1 | + | 1049778 | 1050545 | 768 | f |
| IS3D | IS3 | - | 1094245 | 1095502 | 1258 | f |
| ISZ | IS4 | - | 1294426 | 1295405 | 980 | p |
| IS5U | IS5 | + | 1299499 | 1300693 | 1195 | f |
| IS5F | IS5 | + | 1396044 | 1397238 | 1195 | f |
| IS5Y | IS5 | - | 1427599 | 1428794 | 1196 | f |
| IS2D | IS3 | - | 1467910 | 1469240 | 1331 | f |
| IS30C | IS30 | + | 1469296 | 1470516 | 1221 | f |
| IS609 | IS200/IS605 | + | 1503161 | 1504908 | 1748 | f |
| IS2E | IS3 | + | 1650843 | 1651548 | 706 | p |
| IS1H | IS1 | - | 1978503 | 1979270 | 768 | f |
| IS5H | IS5 | - | 2066159 | 2067353 | 1195 | f |
| IS2F | IS3 | - | 2068941 | 2070271 | 1331 | f |
| IS5I | IS5 | - | 2101749 | 2102943 | 1195 | f |
| IS3E | IS3 | + | 2170171 | 2171428 | 1258 | f |
| IS5K | IS5 | - | 2288919 | 2290113 | 1195 | f |
| IS186C | IS4 | + | 2514273 | 2515617 | 1345 | f |
| IS2H | IS3 | - | 2996361 | 2997691 | 1331 | f |
| IS5LO | IS5 | + | 3130146 | 3131340 | 1195 | f |
| IS2I | IS3 | + | 3186096 | 3187426 | 1331 | f |
| IS5R | IS5 | - | 3365556 | 3366750 | 1195 | f |
| IS1E | IS1 | + | 3583428 | 3584195 | 768 | f |
| IS5T | IS5 | - | 3652036 | 3653230 | 1195 | f |
| IS150 | IS3 | + | 3720633 | 3722075 | 1143 | f |
| IS2K | IS3 | + | 4498181 | 4499511 | 1331 | f |
| IS4 | IS4 | - | 4502090 | 4503515 | 1426 | f |
| IS911B | IS3 | + | 4507125 | 4507458 | 334 | p |
| IS30D | IS30 | - | 4507459 | 4508679 | 1221 | f |
| IS600 | IS3 | - | 4508680 | 4509006 | 327 | p |
| IS911B | IS3 | + | 4509007 | 4509801 | 795 | p |
| IS1F | IS1 | + | 4518472 | 4519239 | 768 | f |

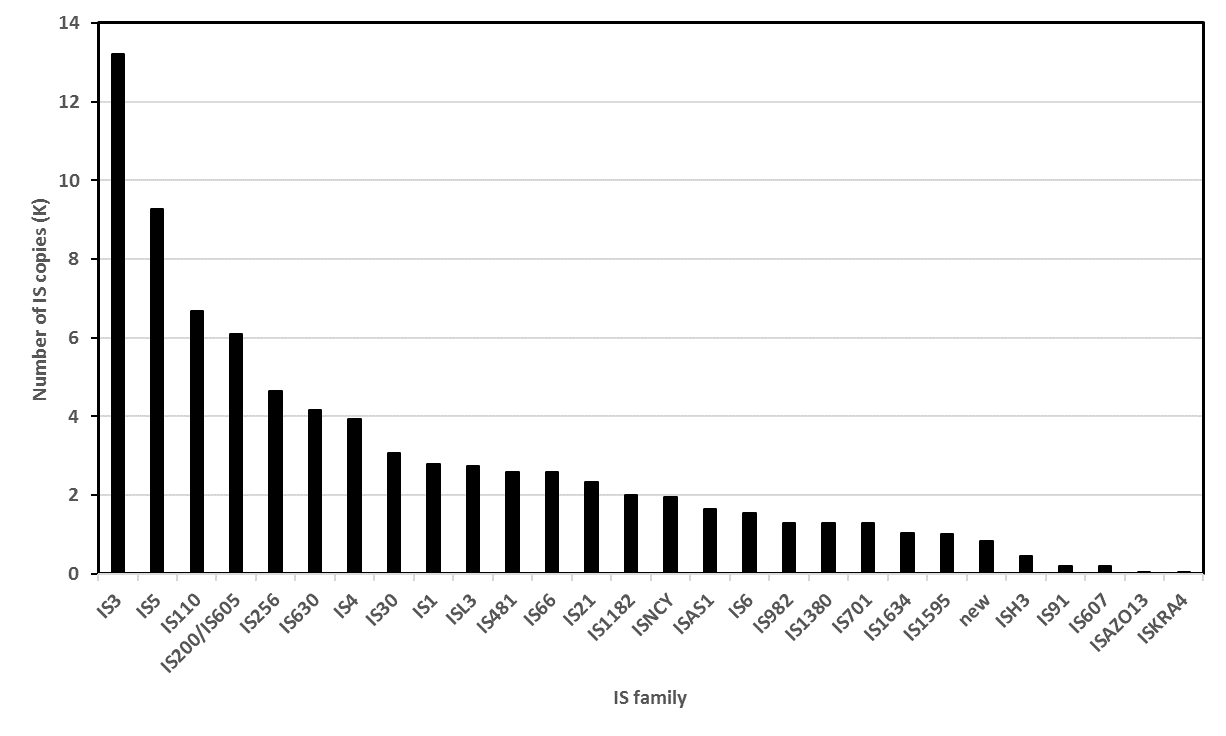
Note: name, name of IS element; family, family name of IS element; strand, DNA strand where transposase is located, + for positive strand while – for complementary strand; start and end, genome coordinates of IS element; length, length of IS element; type, type of IS, f for full-length IS while p for partial IS.

**Table S6. The IS elements reported by ISEScan but not matched with the MG1655 data set**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| strand | start | end | family | length | Roles |
| **+** | **146686** | **147773** | **ISNCY** | **1088** | **gene yadD. Transposase\_31 family protein, function unknown.** |
| - | 528982 | 530045 | IS481 | 1064 | gene ylbG'. Pseudogene, Rhs family protein, function unknown. |
| + | 736682 | 737971 | ISAS1 | 1290 | gene ybfL'. Pseudogene reconstruction, H repeat-associated protein; rhsC-linked. Full length YhhI is putative transposases. |
| + | 1530233 | 1531915 | ISAS1 | 1683 | gene yncI'. Pseudogene reconstruction, H repeat-associated putative transposase, RhsE-linked. |
| **+** | **1531672** | **1532962** | **ISAS1** | **1291** | **gene ydcC. H repeat-associated putative transposase, RhsE-linked, function unknown.** |
| **+** | **3542858** | **3544210** | **ISNCY** | **1353** | **gene yhgA, Transposase\_31 family protein, function unknown** |
| + | 3624234 | 3625524 | ISAS1 | 1291 | gene yhhI. H repeat-associated protein, RhsE-linked, function unknown. Full length paralogs YhhI and YdcC are putative transposases. |
| + | 4506742 | 4507804 | IS3 | 1063 | IS911B, the partial IS element of IS3 family. |
| + | 4509030 | 4510045 | IS3 | 1016 | IS911B, the partial IS element of IS3 family. |
| + | 4568960 | 4569847 | ISNCY | 888 | gene yjiP'. Pseudogene reconstruction, transposase\_31 family. |

Note: strand, DNA strand where transposase is located, + for positive strand while - for complementary strand; start and end, genome coordinates of IS element; family, family name of IS element; length, length of IS element; roles, overlapped genes from EcoGene3.0 and the suggested roles.

**Summarization**: Out of 10 false positives, 3 predictions (30%, highlighted) includes the function unknown transposase protein and might be IS element, 6 predictions (60%) are incomplete IS element, 1 prediction (10%) includes the function unknown pseudogene ylbG'.



**Fig. S1. The distribution of IS elements identified in NCBI prokaryotic genomes in different IS families by using ISEScan.** IS3 is the most abundant family, with 13,206 elements identified in these genomes, followed by IS5 with 9,271 identified elements.

**Curated ACLAME dataset**

To build profile HMMs for ISEScan pipeline, we use ACLAME as the starting point to build a transposase database, namely, curated ACLAME dataset. We first downloaded the most update protein sequences (ACLAME version 0.4) from all categories on ACLAME web site (<http://aclame.ulb.ac.be/perl/Aclame/Tools/exporter.cgi#proteins>), and saved the protein sequence database as a plain text file in fasta format. We then use the IS element and transposase relevant criteria to retrieve all transposases from the downloaded protein sequence database. The approach used to retrieve all transposases is to compare IS element and transposase relevant key words to the headline (beginning with ‘>’) of each protein sequence in protein sequence database (a file in fasta format) to get all transposase sequence candidates.

The first step is to get the headline beginning with ‘>’ for each sequence in the fasta file, namely, header, and the lowercase version of the header, namely, headerL. The second step is to scan each header and headerL we got in the first step to see if the sequence meets the following criteria: 1) ‘ IS’ is header, or any of these key words (‘insert’, 'ransposas', 'transpoase', 'ranspositio') is in headerL; 2) any of these key words ('helper', 'accessory', 'binding', 'resolvase', 'recombinase', 'repressor', 'inactiv') is not in headerL. We can get all transposase candidates from the second step, namely, tpase\_candidate data set. In the third step, we filter out from the tpase\_candidate data set the fake transposase sequences in a negative list of transposases (see the sheet of ‘negative\_list’ in the supplementary excel file), and get a final results, namely, a curated transposase database created from ACLAME protein sequence database.

The most update ACLAME database is the version 0.4 released in 2010. We therefore collected additional 66 transposases (see sheet of ‘Tpases\_from\_NCBI’ in the supplementary excel file) from NCBI protein database add them to the curated transposase database created in the third step above to create the curated ACLAME dataset. This curated ACLAME dataset is used to build profile HMMs.

**Supplementary discussion**

ISEScan scanned 2784 genomes and found 453 (in 34 genomes) and 3 (in 3 genomes) ISH3 copies in Archaea and Bacteria, respectively. In contrast, all other IS families are always found more often in Bacteria than Archaea though it is partly due to the bacterial species-bias existing in the sequenced prokaryotic genomes. Are these ISH3 copies found in Bacteria genomes due to the false report from ISEScan? To answer this question, we put all three ISH3 predictions in a table (ISH3 sheet in supplementary excel file) for the further validation and examined our results about the ISH3 copies in the bacterial genomes and described the details in the following texts.

ISEScan identified three ISH3 copies in three bacterial genomes (see ISH3 sheet in supplementary excel file), Chamaesiphon minutus PCC 6605 (uid183005), Anabaena variabilis ATCC 29413 (uid58043), Singulisphaera acidiphila DSM 18658 (uid81777), respectively. In Chamaesiphon minutus PCC 6605, the ISH3 copy is located on chromosome DNA sequence NC\_019697, ISEScan identified a full IS element with E-value of 8.7e-141 and the length of 1238 bps (1850179 - 1851416) and the IRs (16/16) flanking the predicted transposase. Furthermore, two Y(2)R(3)E(3)R motifs (38-49, 291-302) are found in the protein sequence of the identified Tpase. Based on the above signals about the transposase and IRs, the ISH3 copy found in NC\_019697 is likely to be a full IS element instead of the false report. To cross-validate whether it is ISH3 or the potential new family, we used blast to search for similar sequences in ISfinder. Blastn didn’t find any similar IS element in ISfinder database when querying the nucleic sequence of the ISH3 copy against ISfinder. The best blastn hits are two IS elements, ISTco4 (ISLre2 family, a newly published family in ISfinder and therefore not included in the profile HMM models used by ISEScan) and ISSto14 (ISH3 family), both of which have the same E-value returned by blastn, 0.12. However, the best hit returned by blastp when querying the predicted Tpase in the genome against ISfinder is an ISH3 element (ISMbu7) with E-value = 2e-52 and aa similarity of 51%. The blastn and blastp searchings suggest that it might be a novel IS element not available in ISfinder. The simultaneous existence of the transposase, the Y(2)R(3)E(3)R motif, the valid IRs and the novel nucleotide sequence implies that the ISH3 copy found on NC\_019697 (genomic coordinates: 1850179 - 1851416) is likely to be a new ISH3 element or a potential new IS family which is remotely related to ISH3 family.

We used the same process to examine the other two ISH3 copies in Anabaena variabilis ATCC 29413 (NC\_007410) and Singulisphaera acidiphila DSM 18658 (NC\_019892) listed in the table in the supplementary excel file. The blastn searches indicated that both ISH3 copies might be the novel IS elements not available in ISfinder database. The simultaneous existence of the transposase, the Y(2)R(3)E(3)R motif, the valid IRs and the novel nucleotide sequence shown in the table (see in ISH3 sheet in supplementary excel file) suggest that both ISH3 copies in bacterial genomes are not false reports. We simply list in a table (ISH3 sheet in supplementary excel file) the IS element signals for each ISH3 copy identified by ISEScan for the further reference. All three ISH3 examples found in bacteria genomes demonstrated and confirmed the capability of ISEScan to identify the novel or new IS element in prokaryotic genomes.