Published online 1 November 2022

https://doi.org/10.1093/nar/gkac975

REBASE: a database for DNA restriction and modification: enzymes, genes and genomes

Richard J. Roberts *, Tamas Vincze, Janos Posfai and Dana Macelis

New England Biolabs, 240 County Road, Ipswich, MA 01938 USA

Received September 09, 2022; Revised October 09, 2022; Editorial Decision October 10, 2022; Accepted October 17, 2022

ABSTRACT

REBASE is a comprehensive and extensively curated database of information about the components of restriction-modification (RM) systems. It is fully referenced and provides information about the recognition and cleavage sites for both restriction enzymes and DNA methyltransferases together with their commercial availability, methylation sensitivity, crystal and sequence data. All completely sequenced genomes and select shotgun sequences are analyzed for RM system components. When PacBio sequence data is available, the recognition sequences of many DNA methyltransferases (MTases) can be determined. This has led to an explosive growth in the number of well-characterized MTases in REBASE. The contents of REBASE may be browsed from the web rebase.neb.com and selected compilations can be downloaded by FTP (ftp.neb.com). Monthly updates are also available via email.

OVERVIEW

The previous description of REBASE in the 2015 NAR Database Issue (1) described 6,804 biochemically or genetically characterized restriction-modification systems and included an analysis of just over 5,000 bacterial and archaeal genomes that were present in GenBank (2,3). Since then, the number of completely sequenced genomes has risen to >52,000 and comprehensive descriptions of the RM content of these fully sequenced genomes are available through REBASE in several different formats. In addition, selected genome shotgun sequence sets are also analyzed, and the RM system components extracted. The predicted proteins are named according to standard conventions (4) and labeled with the suffix 'P' to indicate they are putative. When they become biochemically characterized, then the appropriate Roman numeral is used to replace the ORF-designation initially assigned.

The new data going into REBASE comes from the published literature, but increasingly from the analysis of new sequences being deposited into GenBank. After computational prediction, likely candidates are examined manually before inclusion. It should be noted that because Type II restriction enzyme genes appear to evolve rapidly, in many cases genes adjacent to or close by Type II methyltransferase genes may be restriction enzyme genes that cannot be definitively assigned as such. Thus, the number of solitary Type II M genes is likely to be artificially high. Nevertheless, it is well known that there are many examples of solitary methylases that are genuine (5) and more are likely to exist, which may lead to some exciting new biological discoveries.

The increasing use of SMRT sequencing has led to large numbers of DNA methyltransferases being characterized both in terms of the sequences they recognize and the specific base methylated (6–8). This is exemplified by the Type I RM systems, the target sequences of which have a very characteristic bipartite structure. When the genome contains just a single Type I system, then it is clear that the methyltransferase gene and its associated specificity gene are both active and the recognition sequence can be assigned. This in turn allows the assignment of specificities to putative Type I systems in other genomes and frequently allows specificity assignments to be made for active Type I systems in genomes where more than one Type I system is present. These matches are made via REBASE and documented. The same is true for Type III RM systems where again, the recognition specificity for the system, including the endonuclease, is encoded in the methyltransferase gene. The enormous growth in these systems is illustrated in Table 1, which compares the numbers between 1 September 2015 and 1 September 2022.

Given the unique ability of SMRT sequencing to detect methylated bases we encourage users to run PacBio's 'Base Modification Analysis' protocol on such data with the 'Find Modified Base Motifs' switch set to ON. This will generate the motifs.csv file, labelled as 'Modified Base Motifs' in the SMRT Link GUI, that summarizes the methylation patterns that are present. We especially encourage everyone sequencing bacterial and archaeal genomes to generate these summaries and submit them to GenBank as part of their sequence submission. We also welcome their direct submission to REBASE and have a user interface specifically for this purpose. Such data may be submitted prior to publication and a matching analysis will be performed upon re-

*To whom correspondence should be addressed. Tel: +1 978 380 7405; Fax: +1 978 412 9910; Email: roberts@neb.com

Table 1. Statistics for RM systems in REBASE

	2015		2022	
RM system type	Active	Putative	Active	Putative
I MS	770	9,610	4,364	48,800P
II R	4,026[a]	14,048[b]	4,729[c]	67,663[d]
II M	1,793[a]	38,143[b]	5,611[c]	196,604[d]
III R	22	3,791	22	16,634
III M	193	4,185	924	17,996

[a] Note that 39 Type IIG enzymes are included in both the R and the M counts. Type IIG systems generally have both methyltransferase and endonuclease activities in a single polypeptide (4).
[b] Note that 191 putative Type IIG enzymes are included in both the R and the M counts.
[c] Note that 218 Type IIG enzymes are included in both the R and the M counts.
[d] Note that 4,882 putative Type IIG enzymes are included in both the R and the M counts.

ceipt. The results can be kept private until the submitter is ready to publish. For individuals mainly interested in the genome sequence, this further analysis can add much value to the sequence by indicating which restriction-modification systems are present and hence might cause problems during transformation.

Another important feature of REBASE is the identification of a Gold Standard Set of RM system components where the component has been experimentally characterized using a protein of known sequence. In addition to the traditional verification tools, PacBio methylation data has provided a very simple route for MTase characterization. This Gold Standard Set then allows accurate and traceable propagation of annotation into putative genes present in newly sequenced genomes.

Given the explosion of genome sequencing and the ever-increasing number of sequences being produced globally, it will be essential in the future to streamline the automatic processing of RM system candidates as is done for other genome components. For this reason, the Gold Standard RM system components will become increasingly important and REBASE will incorporate the similarity distance from individual members of the Gold Standard set as part of the annotation trail.

From the REBASE website users have a variety of resources available that facilitate the analysis of sequence information, including tools for analyzing sequences (REBASE TOOLS), that allow restriction enzyme recognition sites to be found in submitted sequences (NEBCUTTER) and an implementation of BLAST to allow searching against all sequences in REBASE. Specialty lists of sequence data (REBASE LISTS) such as all Type I specificity subunits or the Gold Standard Set are available for download. In addition, an advanced search feature enables REBASE to be queried by users for specific combinations of information about RM system components, including searching by date of entry. Additional features are added regularly and the site should be consulted for a brief description of any new features as they appear.

ACKNOWLEDGEMENTS

Special thanks are due to the many individuals who have so kindly contributed their unpublished results for inclusion in this compilation and to the REBASE users who continue to guide our efforts with their helpful comments. In recent years many individuals and sequencing centers have contributed their unpublished PacBio data and we are extremely grateful for their willingness to share their data with the wider biological community. We are especially grateful to Karen Otto for administrative help.

FUNDING

REBASE is funded completely by New England Biolabs. Funding for open access charge: New England Biolabs.

Conflict of interest statement. All authors are employees of New England Biolabs who sell restriction enzymes and DNA methylases.

REFERENCES

1. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2015) REBASE–a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **43**, D298–D299.
2. Benson,D.A., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2014) GenBank. *Nucleic Acids Res.*, **42**, D32–D37.
3. Kim D. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
4. Roberts,R.J., Belfort,M., Bestor,T., Bhagwat,A.S., Bickle,T.A., Bitinaite,J., Blumenthal,R.M., Degtyarev,S.K., Dryden,D.T.F., Dybvig,K. *et al.* (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.*, **31**, 1805–1812.
5. Anton,B.P. and Roberts,R.J. (2021) Beyond restriction modification: epigenomic roles of DNA methylation in prokaryotes. *Annu. Rev. Microbiol.*, **75**, 129–149.
6. Eid,J., Fehr,A., Gray,J., Luong,K., Lyle,J., Otto,G., Peluso,P., Rank,D., Baybayan,P., Bettman,B. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.
7. Flusberg,B.A., Webster,D.R., Lee,J.H., Travers,K.J., Olivares,E.C., Clark,T.A., Korlach,J. and Turner,S.W. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, **7**, 461–465.
8. Korlach,J. and Turner,S.W. (2012) Going beyond five bases in DNA sequencing. *Curr. Opin. Struct. Biol.*, **22**, 251–261.