D298–D299 *Nucleic Acids Research, 2015, Vol. 43, Database issue* *Published online 05 November 2014*
doi: 10.1093/nar/gku1046

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, Ipswich, MA 01938, USA

Received October 02, 2014; Accepted October 11, 2014

ABSTRACT

REBASE is a comprehensive and fully curated database of information about the components of restriction-modification (RM) systems. It contains fully referenced information about recognition and cleavage sites for both restriction enzymes and methyltransferases as well as commercial availability, methylation sensitivity, crystal and sequence data. All genomes that are completely sequenced are analyzed for RM system components, and with the advent of PacBio sequencing, the recognition sequences of DNA methyltransferases (MTases) are appearing rapidly. Thus, Type I and Type III systems can now be characterized in terms of recognition specificity merely by DNA sequencing. The contents of REBASE may be browsed from the web http://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 2010 NAR Database Issue (1) described 3945 biochemically or genetically characterized restriction-modification (RM) systems and included an analysis of just 400 bacterial and archaeal genomes that were present in GenBank (2,3). Since then the number of completely sequenced genomes has risen to more than 5000 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. If and when they become biochemically characterized, then the appropriate Roman numeral is used to replace the Open Reading Frame (ORF)-designation initially assigned.

The fastest growing segment of biochemically characterized RM system components are the DNA methyltransferases, currently being found using real-time Single Molecule Real Time (SMRT) sequencing (5–7). In a large number of cases the recognition motifs for DNA methyltransferases arising from such sequencing can be matched with the genes that encode them. For instance, the recognition sequences of Type I systems have a very characteristic bipartite structure and when the genome contains only one such Type I set of genes, then it is clear that the methyltransferase gene and its associated specificity gene are both active and the recognition motif can be assigned. This in turn allows propagation of these specificities both to putative Type I systems and also in many cases allows the identification of 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 and again, is made possible because the recognition specificity for the system is encoded in the methyltransferase gene. The enormous growth in these systems is shown in Figure 1, which documents the rate of discovery of Type I and Type III RM system specificities since the very first one was discovered and characterized in 1968 (8).

Given the unique ability of SMRT sequencing to detect methylated bases it is almost always extremely useful to run the 'Modification and Motif Analysis' protocol on such data, to generate the motif_summary.csv file that summarizes the methylation patterns that are present. We encourage everyone sequencing bacterial 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 an interface specifically for this purpose (http://rebase.neb.com). Such data may be submitted prior to publication and a matching analysis will be performed upon receipt. 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 RM systems are present and might cause problems during transformation.

Another new feature of REBASE is the identification of a Gold Standard Set of RM system components where each component having a known sequence has been experimentally characterized as having a defined restriction or modification activity. This Gold Standard Set then allows accurate and traceable propagation of annotation into putative genes

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

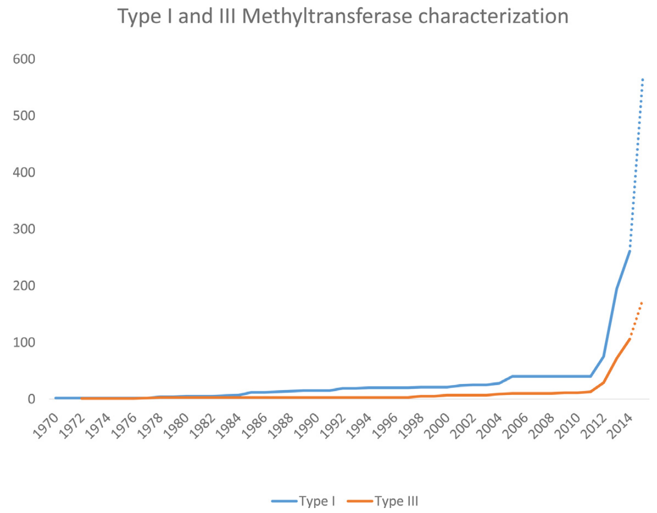

Figure 1. Dotted lines signify methylated motifs awaiting gene assignment. The sharp increase starting in 2012 is the result of the introduction of SMRT sequencing.

present in newly sequenced genomes. In future, all RM system specificity predictions will contain traceable annotation back to this reference Gold Standard Set.

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 to 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. Over the last 2 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 Inc.

Conflict of interest statement. None declared.

REFERENCES

1. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2010) REBASE–a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **38**, D234–D236.
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. 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. 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.
6. 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.
7. Korlach,J. and Turner,S.W. (2012) Going beyond five bases in DNA sequencing. *Curr. Opin. Struct. Biol.*, **22**, 251–261.
8. Linn,S. and Arber,W. (1968) Host specificity of DNA produced by *Escherichia coli*, X. In vitro restriction of phage fd replicative form. *Proc. Natl. Acad. Sci. U.S.A.*, **59**, 1300–1306.