5 years of community annotation using Wikipedia

Jennifer Daub[1], Ruth Eberhardt[1], John Tate[1], Paul Gardner[2], Sarah Burge[1], Alex Bateman[1]

Affiliations: [1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. [2]University of Canterbury, New Zealand.

Rfam is a database of non-coding RNA families, each represented by a multiple sequence alignment and a covariance model (CM)[1]. Our multiple sequence alignments and secondary structure annotations are generated manually by specialist biocurators. However, since June 2007 (Release 8.1) Rfam has been successfully coordinating community annotation via the online encyclopedia Wikipedia to provide textual annotation for these families[2].

The last Rfam release 10.1 (June 2011) contains 1,973 families which are linked to around 900 Wikipedia articles. New edits to these articles are downloaded on a daily basis and displayed on our website along side our specialist information.

Each Rfam family has a link to one Wikipedia article

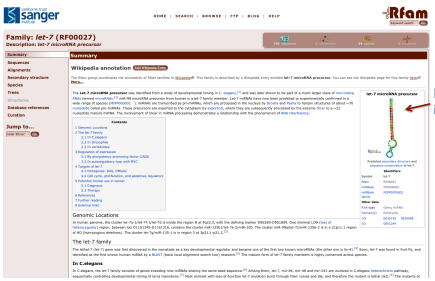
Wikipedia article is chosen by curator at time of family build

If no suitable article is available, a curator may:

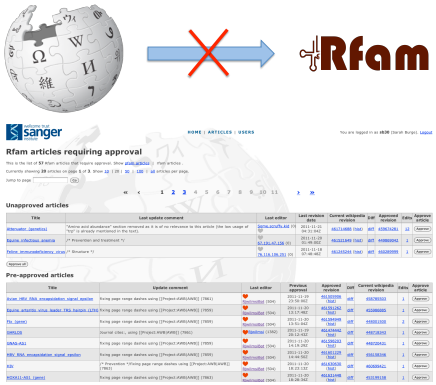
- Create a new article
- Link to a generic RNA article, e.g. miRNA, ncRNA, snoRNA
- Each article has a link back to Rfam

This article mapping is maintained by Rfam

Wikipedia article is mirrored on the Rfam website alongside our specialist information



Wikipedia content on rfam.sanger.ac.uk is not live



The Rfam wiki approval app allows us to approve content before displaying on our website.

The Rfambot

Wikipedia allows for automation using bots

Rfambot automates Rfam-related routine tasks e.g. Rfam infobox, release & image updates

Many non-Rfam specific bots edit & improve our mapped articles

Rfam specific information is not open to Wikipedia editing

Text annotation

Rfam infobox

EVERYTHING ELSE

e.g. Sequence alignments, Covariance models, Secondary structures, Types, Names, Article mapping

Wikipedia page edits and page views

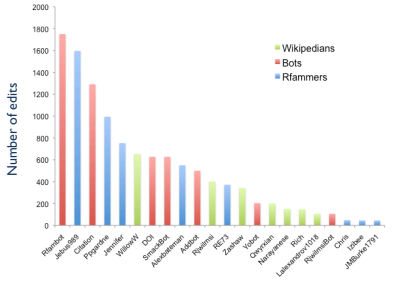
Since **June 2007**

- **16, 607** edits have been made by **2,213** editors
- **~351** edits were reverted – mainly by external editors
- **~600** articles have been edited **> 10** times
- **15** articles have received **> 120** edits

February 2012

- Each articles was viewed **>10**
- tRNA, microRNA, Ribosomal RNA and Ribozyme articles were each viewed **>1000** times
- **15%** of Rfam web access came via Wikipedia

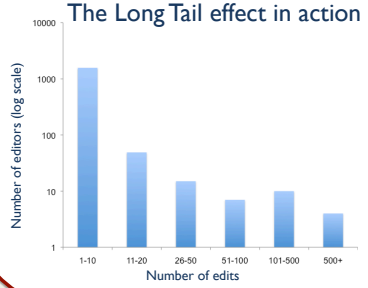
Who makes these edits?



Editors ranked by number of edits made since June 2007. Only the top editors are shown

Edits to our articles of interest are made by many people and bots not associated with Rfam

The Long Tail effect in action



Distribution of the number of edits each editor makes

Most editors make only a few edits but we have lots of editors

Why Wikipedia works for Rfam

- Stable editing infrastructure
- **Huge** user community-
  - scientific and non-scientific editors
  - excellent quality and vandalism policing
- The "Google factor"
- "Chores" can be automated by bots
- Rfam specialist data is untouched
- Editor contributions exceed Rfam manpower!

References:

1 Rfam: Wikipedia, clans and release Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, Bateman A. Nucleic Acids (2011) doi: 10.1093/nar/gkq1129

2 The RNA WikiProject: community annotation of RNA families. Daub J, Gardner PP, Tate J, Ramsköld D, Manske M, Scott WG, Weinberg Z, Griffiths-Jones S, Bateman A.RNA (2008) 12:2462-2464

Get a PDF of this poster