providing salient categorical information in greater detail than the officially reported category. Finally, we show how the database can be used for unbiased discovery of research trends, and we document the remarkable increase in funding for research on microRNA biology from 2007 to 2009. Changes in topics associated with this burgeoning area demonstrate a transition in the nature of the research, from basic cellular and molecular biology to investigations of complex physiological processes and disease diagnoses.

In each case, the machine-learned topics are robustly correlated with funding by specific NIH Institutes, highlighting the importance of the underlying categories to the NIH. The patterns elucidated in this framework are consistent with Institute policies, but obtaining similar information in the absence of the current database would require extensive exploration of Institute websites, followed by time-consuming research on appropriate keywords for queries of specific categories. Our database offers an alternative approach that enables rapid and reproducible retrieval of meaningful categorical information.

To ensure transparent and accurate representations of the algorithm-derived topics, we provide extensive contextual information derived from the documents associated with each topic, in a format conducive to spot checks and to detailed examination for cases requiring precise categorical distinctions. Additionally, we implemented a new technique for automatically assessing topic quality using statistics of topic word co-occurrence (**Supplementary Methods**), which we used for curating the database to identify poor quality topics.

Our use of this graphing algorithm is somewhat different from previous gene expression analyses and scientometric studies based on journal citation linkages (see **Supplementary Methods** for references). We assessed the information-retrieval capabilities of the graphs and found that they performed well relative to the document similarity measures that served as inputs. Notably, rather than forming isolated clusters, in this case the algorithm produced a lattice-like structure, in which clusters are linked by strings of aligned documents whose topical content is jointly relevant to the clusters at either end of each string (**Supplementary Fig. 1**). In addition to providing extra 'subcluster' resolution of content that falls between clusters, this lattice-like framework formed a logical organizational structure, merging the local, intermediate and global levels of the graph.

The categories and clusters represented in this database are comprehensive and thus provide reference points from which various information requirements can be addressed by users with divergent interests and needs. Perhaps more importantly, they provide a basis for discovery of interrelationships among concepts and documents that otherwise would be obscure.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We acknowledge assistance and support from G. LaRowe and N. Skiba at ChalkLabs, and input and feedback from NIH staff during the project. We thank S. Silberberg, C. Cronin, K. Boyack and K. Borner for helpful advice and comments on the manuscript. This project has been supported through small contracts from the NIH to University of Southern California (271200900426P and 271200900244P), University of Massachusetts (271201000758P, 271200900640P, 271201000704P and 271200900639P), ChalkLabs LLC (271200900695P and 271201000701P) and TopicSeek LLC (271201000620P and 271200900637P).

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Edmund M Talley[1], David Newman[2], David Mimno[3,6], Bruce W Herr II[4], Hanna M Wallach[3], Gully A P C Burns[5], A G Miriam Leenders[1] & Andrew McCallum[3]

[1]National Institute of Neurological Disorders and Stroke, Bethesda, Maryland, USA. [2]University of California, Irvine, Irvine, California, USA. [3]University of Massachusetts, Amherst, Amherst, Massachusetts, USA. [4]ChalkLabs, Bloomington, Indiana, USA. [5]Information Sciences Institute, University of Southern California, Marina del Rey, California, USA. [6]Present address: Princeton University, Princeton, New Jersey, USA.
e-mail: talleye@ninds.nih.gov

Predicting protein associations with long noncoding RNAs

To the Editor: Only a small fraction of the human transcriptome (~1%) encodes proteins[1], but a large portion of transcripts is long noncoding RNAs (lncRNAs) and is an unexplored component of mammalian genomes[2]. Here we introduce a method to perform large-scale predictions of protein-RNA associations. Our algorithm, 'fast predictions of RNA and protein interactions and domains at the Center for Genomic Regulation, Barcelona, Catalonia' (catRAPID), evaluates the interaction propensities of polypeptide and nucleotide chains using their physicochemical properties. The algorithm is freely available at http://big.crg.cat/gene_function_and_evolution/services/catrapid.

We trained catRAPID on 592 protein-RNA pairs available in the Protein Data Bank to discriminate interacting and non-interacting molecules using only information contained in their sequences (**Supplementary Table 1**). Secondary structure propensities accounted for 72% of catRAPID ability to predict protein-RNA associations, followed by hydrogen bonding (58%) and van der Waals (26%) contributions. Occurrence of hairpin loops in nucleotide sequences and presence of helical elements in polypeptide sequences positively correlated with interaction propensities. Protein and RNA binding sites had higher interaction propensities than other regions in complexes (**Fig. 1a**, **Supplementary Methods** and **Supplementary Tables 2** and **3**).

We validated our algorithm on a large collection of protein associations with lncRNAs[3], the NPInter dataset (**Supplementary Methods** and **Supplementary Table 4**). Using catRAPID we correctly predicted 89% of experimentally supported interactions linked to physical evidence of binding (**Fig. 1b**). We observed less significant performance (P = ~0.1) for interactions inferred from indirect evidence (**Supplementary Methods**). To test catRAPID's ability to identify non-interacting molecules, we generated random lists of RNA associations with proteins involved in DNA and protein-binding (DNA BP and protein BP datasets, respectively; **Fig. 1c** and **Supplementary Table 5**). We predicted interactions only for <40% of cases, which suggests that these associations are unlikely to take place (RNA BP dataset; **Fig. 1c** and **Supplementary Table 5**). With regard to random associations with RNA-binding proteins, we observed slightly higher interaction propensities (~52%), which indicates occurrence of spurious binding.

To validate the ability of catRAPID to identify binding regions, we analyzed the human ribonuclease mitochondrial RNA processing (MRP) complex[4] (**Supplementary Methods**). The MRP assembly comprises ten protein subunits: hPop1, hPop5, Rpp14, Rpp20, Rpp21, Rpp25, Rpp29, Rpp30, Rpp38 and Rpp40. We predicted

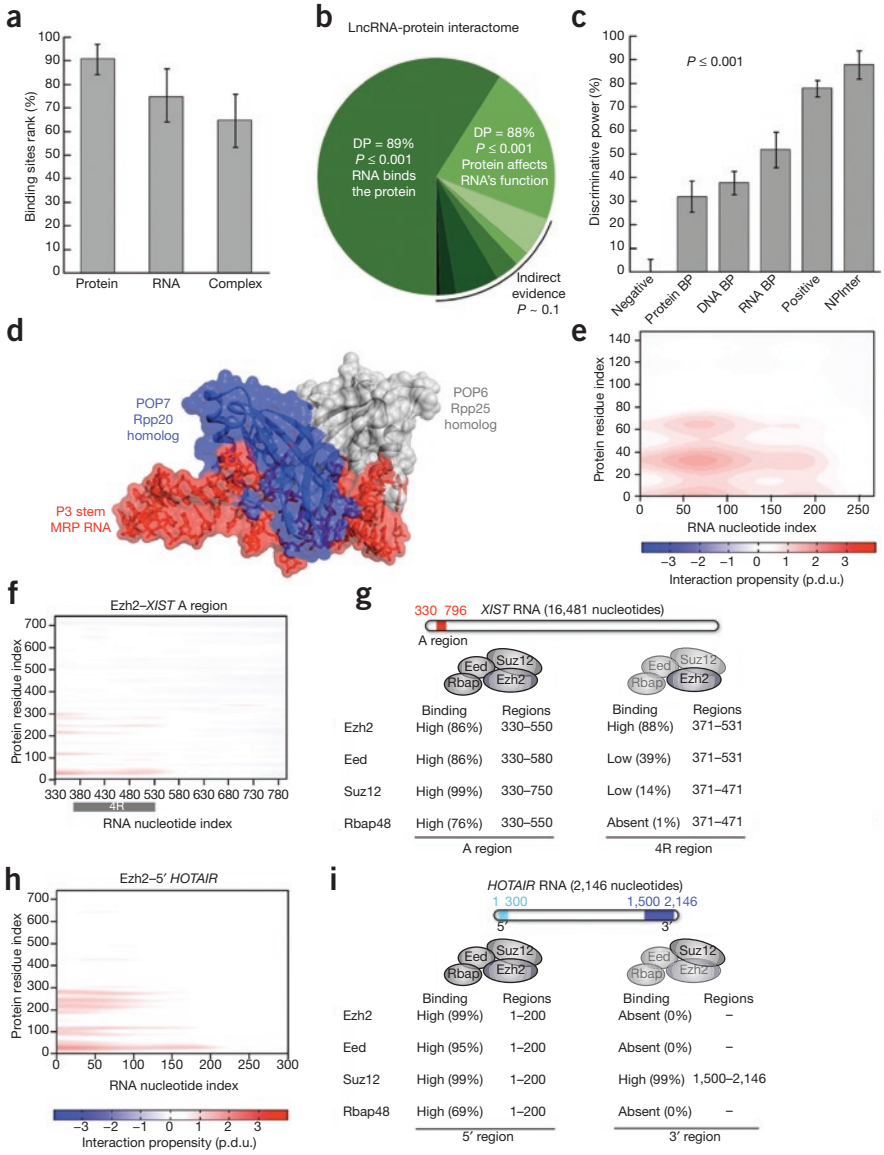

Figure 1 | Performance, validation and applications of catRAPID. (**a**) Interaction propensity rank of binding regions in the positive set (410 interacting pairs). Error bars, s.d. (**b**) Performance on NPInter dataset (405 interacting pairs) measured by discriminative power (DP). (**c**) Performance (discriminative power) on training (positive and negative sets are lists of interacting and non-interacting protein-RNA pairs) and indicated test sets. Error bars, s.d. (**d**) P3 stem RNA in complex with Rpp20 and Rpp25 homologs (yeast MRP structure). (**e**) Predictions of interactions between human *MRP* RNA and Rpp20. Interaction propensity is measured in procedure defined units (p.d.u.). (**f**) Interaction propensity for binding regions of Ezh2 with *XIST* A region. (**g**) Predictions of associations between PCR2 complex and *XIST* A region. Analysis of 4R region is displayed. Regions are listed by nucleotide positions. (**h**) Interaction propensity for binding regions of Ezh2 with 5′ *HOTAIR*. (**i**) Predictions of interactions between PCR2 complex with 5′ and 3′ domains of *HOTAIR*.

that Rpp20 binds the P3 stem as reported in the crystal structure of the MRP complex yeast homolog (Protein Data Bank code: 3IAB, **Fig. 1d**,**e** and **Supplementary Methods**). Rpp14, Rpp30, Rpp40 and hPop5 were predicted to be low- or non-interacting, in agreement with experimental evidence[4]. By contrast, Rpp21, Rpp25, Rpp29 and Rpp38 had high propensity to interact with RNA, as is the case *in vitro*[4] (**Supplementary Fig. 1** and **Supplementary Table 6**). We also analyzed the human RNase P that shares proteins with the MRP system[4]. All the subunits were predicted to bind RNA, and Rpp20, Rpp21, Rpp25, Rpp29 and Rpp38 had the highest interaction propensities (**Supplementary Fig. 2**).

Additionally, we calculated the interaction propensities of human *XIST* and *HOTAIR* lncRNAs with the chromatin-modifying polycomb repressive complex 2 (PRC2). In agreement with experimental evidence[5,6] we predicted Ezh2, Eed, Suz12 and Rbap48 to bind the A region of *XIST* and the 5′ domain of *HOTAIR* with confidence >90% (**Fig. 1f–i** and **Supplementary Figs. 3** and **4**). The majority of PRC2 components showed poor propensity to bind the 3′ domain of *HOTAIR*, confirming previous observations[6] (**Fig. 1i** and **Supplementary Fig. 4**). Strong interaction propensity was predicted only for Suz12 at the 3′ domain of *HOTAIR*, but no experimental data are available for comparison. We also performed predictions on portions 4R and 2R in the A region of *XIST*[5]. Our calculations indicated high interaction propensities for Ezh2 and suggested that the nucleation center is probably the 4R region. Suz12 had low interaction propensities with 2R and 4R, in agreement with experimental data[5], indicating that its binding might require the entire A region (**Supplementary Fig. 3**).

In conclusion, prediction of lncRNAs function is generally hampered by poor sequence homology and lack of interaction data. We expect that catRAPID will guide experimental approaches and facilitate a deeper understanding of the role of lncRNAs in post-transcriptional regulatory networks.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank R. Guigò, members of the Bioinformatics and Genomics Program at Centre for Genomic Regulation and the Bioinformatics Core Facility, and A. Hermoso and P. Di Tommaso.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Matteo Bellucci, Federico Agostini, Marianela Masin & Gian Gaetano Tartaglia

Center for Genomic Regulation, Universitat Pompeu Fabra, Barcelona, Spain.
e-mail: gian.tartaglia@crg.es

1. Ørom, U.A. *et al. Cell* **143**, 46–58 (2010).
2. Ponting, C.P., Oliver, P.L. & Reik, W. *Cell* **136**, 629–641 (2009).
3. Wu, T. *et al. Nucleic Acids Res.* **34**, D150–D152 (2006).
4. Welting, T.J.M., van Venrooij, W.J. & Pruijn, G.J.M. *Nucleic Acids Res.* **32**, 2138–2146 (2004).
5. Maenner, S. *et al. PLoS Biol.* **8**, e1000276 (2010).
6. Tsai, M. *et al. Science* **329**, 689–693 (2010).