Quantitative predictions of protein interactions with long noncoding RNAs

To the Editor: Long noncoding RNAs (lncRNAs, which comprise 68% of the human transcriptome with average length of 1,000 nt) interact with various RNA-binding proteins (RBPs) to mediate cellular functions[1]. Identification of lncRNA interactions through experimental methods is challenging and can be complemented by the use of computational models to predict protein partners. Yet it is currently impractical to calculate RBP–lncRNA interactions for large transcripts[2].We introduce Global Score as a means to predict protein interactions with large RNAs (http://service.tartaglialab.com/new_submission/globalscore).

Exploiting sequence information, this algorithm integrates local properties of protein and RNA structures into an overall binding propensity (see **Supplementary Methods** and **Fig. 1a**). Calibrated on high-throughput data (training, 14 PAR-CLIP or HITS-CLIP experiments; testing, 8 protein arrays), Global Score is the first approach to quantitatively predict RBP partners of RNA >1,000 nt (>80% discrimination between interacting and noninteracting pairs; **Fig. 1b, Supplementary Fig. 1a**, and **Supplementary Tables 1–4**). Large-scale analysis of eCLIP assays (>17,000 RBPs–RNAs)[3] reveals that Global Score performances significantly increase with the binding strength (*P* values < 10^{-10}; Wilcoxon signed-rank test; **Fig. 1c**, **Supplementary Methods, Supplementary Fig. 2**, and **Supplementary Table 5**).

Using >600 RBPs available from five proteomic and genomic screens, we investigated *Xist* interactions (**Supplementary Methods**,

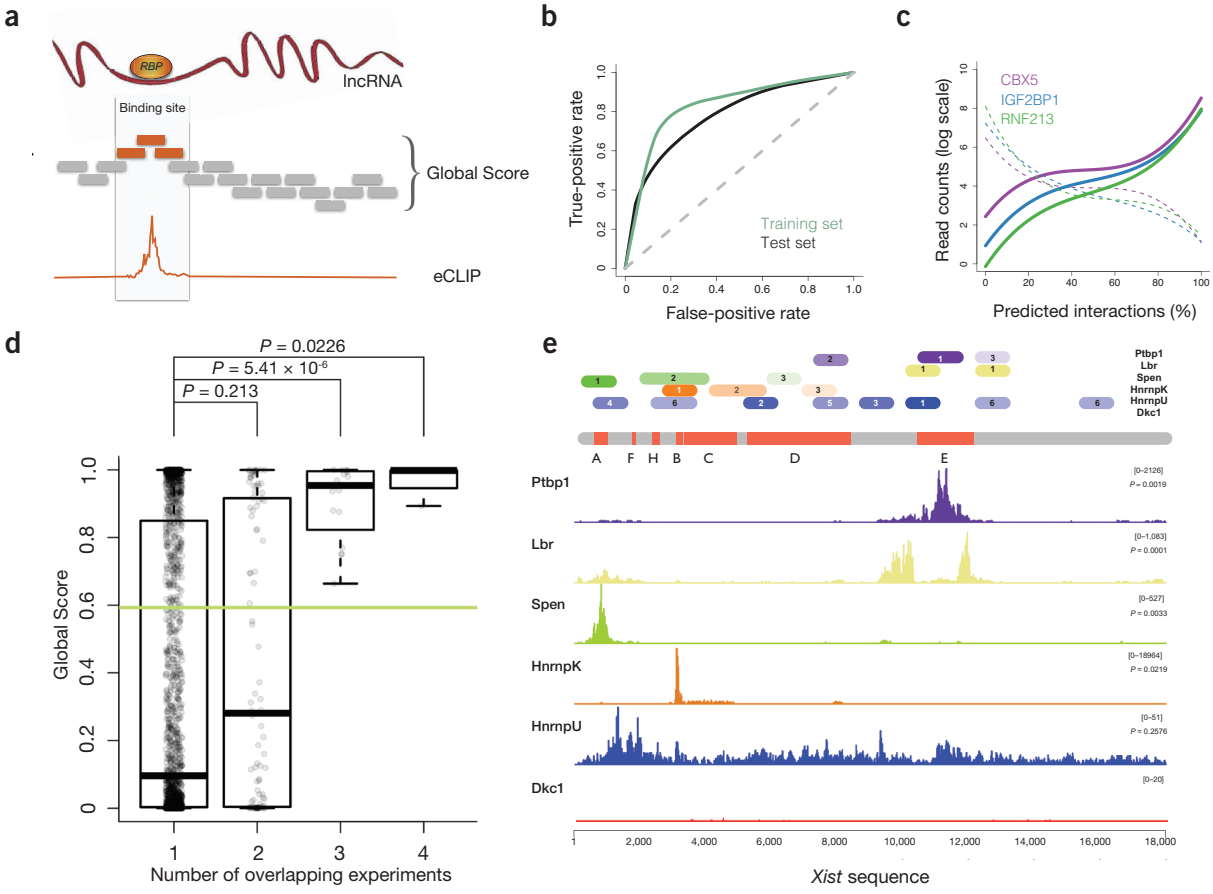

Figure 1 | Training and testing of the Global Score for prediction of protein interactions with large RNAs. (**a**) Global Score predicts RBP interactions with large RNAs. (**b**) Training (green curve, area under the ROC curve (AUC) = 0.84) and testing (black curve, AUC = 0.80) performances. The dashed line indicates performances of a random predictor (AUC = 0.50). (**c**) Transcript analysis: interaction-prone RBPs (continuous lines) increase from low to high read counts, while low-propensity RBPs (dashed lines) decrease (*P* values < 10^{-30}; three RNA shown; eCLIP assays). (**d**) Global Score predictions correlate with the number of experiments reporting *Xist* interaction with a specific RBP (green line, significance at *P* value < 0.01). Box plot limits are upper and lower quartiles; center lines represents the median and whiskers indicate minimum and maximum values. (**e**) Agreement between predicted binding sites (top; numbers and shading represent the ranking of the predictions) and eCLIP experiments (bottom; read counts and significance of the match). *Xist* tandem repeats are indicated with orange boxes and letters (A, F, H, B, C, D, and E). Nucleotide coordinates are reported at the bottom.

Supplementary Fig. 1b, and Supplementary Tables 1–3, 6 and **7**)[4–8]. Global Score predictions correlate with the number of independent experiments reporting *Xist* association with a specific protein (**Fig. 1d**) and identify 38 high-confidence RBPs (empirical *P* value < 0.01) reported in two or more independent experiments (**Supplementary Fig. 3**). We used eCLIP to test *Xist* binding to the RBPs HnrnpK (Global Score of 0.99), Ptbp1 (0.99), Lbr (0.79), HnrnpU (Saf-A) (0.66), Spen (Sharp) (0.59; **Supplementary Methods** and **Supplementary Figs. 4** and **5**), and negative control Dkc1 (0.01). Global Score prediction correlates with the eCLIP binding profile (Pearson correlation = 0.93; **Supplementary Fig. 6**): Spen and HnrnpK, Ptbp1, and Lbr interact respectively with A, B, and E repeats and adjacent regions, while HnrnpU binds across the whole transcript, and Dkc1 does not interact with *Xist* (**Fig. 1e** and **Supplementary Table 8**).

The ability to predict global (overall binding ability; **Fig. 1b**) and local (protein and RNA contacts; **Fig. 1e**) interactions makes Global Score the first computational algorithm (**Supplementary Fig. 1**) to prioritize lncRNAs partners for experimental validation (**Fig. 1c, Supplementary Fig. 7**, and **Supplementary Table 9**).

Data availability.

The webserver, documentation, and compiled stand-alone version of Global Score are available at http://service.tartaglialab.com/new_submission/globalscore. Data from HnrnpK, Ptbp1, Lbr, HnrnpU, Sharp, and Dkc1 eCLIP experiments are deposited at the GEO with codes (respectively) GSM2325771, GSM2299086, GSM2299085, GSM2325772, GSM2299084, and GSM2325770.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Whitworth and K. Havas for critical reading of the manuscript, the European Research Council (RIBOMYLOME_309545), MINECO (BFU2014-55054-P), Centro de Excelencia Severo Ochoa 2013–2017, and the European Molecular Biology Laboratory Grant-50800.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Davide Cirillo[1,2], Mario Blanco[3], Alexandros Armaos[1,2], Andreas Buness[4], Philip Avner[4], Mitchell Guttman[3], Andrea Cerase[4] & Gian Gaetano Tartaglia[1,2,5]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain. [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [3]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California. [4]EMBL-Monterotondo, Rome, Italy. [5]Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.
e-mail: andrea.cerase@embl.it or gian.tartaglia@crg.eu

1. Iyer, M.K. *et al. Nat. Genet.* **47**, 199–208 (2015).
2. Bellucci, M., Agostini, F., Masin, M. & Tartaglia, G.G. *Nat. Methods* **8**, 444–445 (2011).
3. Van Nostrand, E.L. *et al. Nat. Methods* **13**, 508–514 (2016).
4. Chu, C. *et al. Cell* **161**, 404–416 (2015).
5. McHugh, C.A. *et al. Nature* **521**, 232–236 (2015).
6. Minajigi, A. *et al. Science* **349**, aab2276 (2015).
7. Moindrot, B. *et al. Cell Rep.* **12**, 562–572 (2015).
8. Monfort, A. *et al. Cell Rep.* **12**, 554–561 (2015).