Thus, rarer species may be more buffered from extinction than expected from neutral sampling effects. However, time-lagged extinctions due to extinction debt may lead to additional species loss (*31*).

Although an examination of how hundreds of common and rare species were disproportionately influenced by invaders is beyond the scope of this study, we can glean insights by examining the traits of common and rare species at the study sites. For example, in Hawai'i the native sedge *Carex wahuensis* was rare in the absence of the invader but became proportionately more common in the presence of the invader, likely because it could tolerate lower light and/or take advantage of higher nitrogen imposed by the invasive *M. faya* (*32*). Likewise, in Missouri several native species known to be shade tolerant (such as *Desmodium glutinosum* and *Trillium recurvatum*) (*33*) were proportionately less influenced by the invasive *L. maacki* than were shade-intolerant species.

Overall, by explicitly focusing on scale-dependent processes, the results from our study reconcile the differences observed among local- and broad-scale effects of invasive plant species on biodiversity. Decreased intercepts (c) and increased slopes (z) of the SAR were primarily caused by neutral sampling effects. In addition, disproportionately smaller effects on rare species' abundances moderated species loss at the broadest spatial scale. Understanding the mechanisms by which invasive species shift species abundance distributions could improve our ability to forecast future invasion-induced extinctions. Although particularly harmful to native biodiversity at small spatial scales, invasive species' effects may be reversed through targeted control to increase native species abundances, at least until future extinction debt is paid.

References and Notes

1. M. Gaertner, A. D. Breeyen, C. Hui, D. M. Richardson, *Prog. Phys. Geogr.* **33**, 319 (2009).
2. K. I. Powell, J. M. Chase, T. M. Knight, *Am. J. Bot.* **98**, 539 (2011).
3. M. Vilà *et al.*, *Ecol. Lett.* **14**, 702 (2011).
4. D. F. Sax, S. D. Gaines, J. H. Brown, *Am. Nat.* **160**, 766 (2002).
5. J. Gurevitch, D. K. Padilla, *Trends Ecol. Evol.* **19**, 470 (2004).
6. L. C. Maskell, G. Firbank, K. Thompson, J. M. Bullock, S. M. Smart, *J. Ecol.* **94**, 1052 (2006).
7. T. J. Stohlgren, D. T. Barnett, C. S. Jarnevich, C. Flather, J. Kartesz, *Ecol. Lett.* **11**, 313, discussion 322 (2008).
8. M. A. Davis, *Bioscience* **53**, 481 (2003).
9. D. F. Sax, S. D. Gaines, *Proc. Natl. Acad. Sci. U.S.A.* **105** (suppl. 1), 11490 (2008).
10. M. A. Davis *et al.*, *Nature* **474**, 153 (2011).
11. C. S. Kolar, D. M. Lodge, *Trends Ecol. Evol.* **16**, 199 (2001).
12. L. Valéry, H. Fritz, J.-C. Lefeuvre, D. Simberloff, *Biol. Inv.* **10**, 1345 (2008).
13. J. Gurevitch, G. A. Fox, G. M. Wardle, D. Inderjit, D. Taub, *Ecol. Lett.* **14**, 407 (2011).
14. J. T. Hutchinson, E. A. Gandy, K. A. Langeland, *Inv. Plant Sci. Manage.* **4**, 349 (2011).
15. M. H. Collier, J. L. Vankat, M. R. Hughes, *Am. Midl. Nat.* **147**, 60 (2002).
16. M. Dorning, D. Cipollini, *Plant Ecol.* **184**, 287 (2006).
17. P. M. Vitousek, L. R. Walker, *Ecol. Monogr.* **59**, 247 (1989).
18. Materials and methods are available as supplementary materials on *Science* online.
19. F. W. Preston, *Ecology* **43**, 185 (1962).
20. R. M. May, in *Ecology and Evolution of Communities*, M. L. Cody, J. M. Diamond, Eds. (Harvard Univ. Press, Cambridge, MA, 1975), pp. 81–120.
21. F. He, P. Legendre, *Ecology* **83**, 1185 (2002).
22. S. J. Meiners, S. T. A. Pickett, M. L. Cadenasso, *Ecography* **25**, 215 (2002).
23. T. P. Rooney, S. M. Wiegmann, D. A. Rogers, D. M. Waller, *Conserv. Biol.* **18**, 787 (2004).
24. T. D. Olszewski, *Oikos* **104**, 377 (2004).
25. J. L. Green, A. Ostling, *Ecology* **84**, 3090 (2003).
26. M. D. Collins, D. Simberloff, *Environ. Ecol. Stat.* **16**, 89 (2009).
27. J. M. Chase, N. J. B. Kraft, K. G. Smith, M. Vellend, B. D. Inouye, *Ecosphere* **2**, art24 (2011).
28. A. S. MacDougall, B. Gilbert, J. M. Levine, *J. Ecol.* **97**, 609 (2009).
29. C. C. Daehler, *Annu. Rev. Ecol. Evol. Syst.* **34**, 183 (2003).
30. O. Chabrerie, J. Loinard, S. Perrin, R. Saguez, G. Decocq, *Biol. Invasions* **12**, 1891 (2010).
31. D. Tilman, R. M. May, C. L. Lehman, M. A. Nowak, *Nature* **371**, 65 (1994).
32. P. B. Adler, C. M. D'Antonio, J. T. Tunison, *Pac. Sci.* **52**, 69 (1998).
33. P. Bierzychudek, *New Phytol.* **90**, 757 (1982).

Acknowledgments: We are grateful to H. Bailey, J. Hidalgo, J. Powell, and M. Schutzenhofer for field assistance. We thank B. Allan, E. Gandy, T. Hingtgen, R. Loh, T. Mohrman, and J. Shaw for support with permits and logistics at each field site. The members of the Chase and Knight labs and three anonymous reviewers provided invaluable feedback and greatly improved the analysis and presentation of the manuscript. J.M.C. is an independent researcher. Funding was provided by NSF DGE 1143954 (to K.I.P.), DEB 1110629, and the Tyson Research Center at Washington University in St. Louis. Original data for species richness and area available on Dryad (doi: 10.5061/dryad.qq08m).

Supplementary Materials

www.sciencemag.org/cgi/content/full/339/6117/316/DC1
Materials and Methods
Figs. S1 to S5
Table S1
Reference (*34*)

2 July 2012; accepted 20 November 2012
10.1126/science.1226817

Structure of Histone mRNA Stem-Loop, Human Stem-Loop Binding Protein, and 3′hExo Ternary Complex

Dazhi Tan,[1] William F. Marzluff,[2,3] Zbigniew Dominski,[2,3] Liang Tong[1]*

Metazoan replication-dependent histone messenger RNAs (mRNAs) have a conserved stem-loop (SL) at their 3′-end. The stem-loop binding protein (SLBP) specifically recognizes the SL to regulate histone mRNA metabolism, and the 3′-5′ exonuclease 3′hExo trims its 3′-end after processing. We report the crystal structure of a ternary complex of human SLBP RNA binding domain, human 3′hExo, and a 26-nucleotide SL RNA. Only one base of the SL is recognized specifically by SLBP, and the two proteins primarily recognize the shape of the RNA. SLBP and 3′hExo have no direct contact with each other, and induced structural changes in the loop of the SL mediate their cooperative binding. The 3′ flanking sequence is positioned in the 3′hExo active site, but the ternary complex limits the extent of trimming.

Metazoan replication-dependent histone mRNAs have a conserved stem-loop (SL) structure at their 3′-end (*1*, *2*), distinct from the polyadenylate tail found on all other known eukaryotic mRNAs (*3*, *4*). The stem-loop binding protein (SLBP) (*5*), also known as hairpin binding protein (*6*), is a central regulator of histone mRNA metabolism. SLBP and the U7 small nuclear ribonucleoprotein (snRNP) (*7*) are required for the 3′-end processing of histone pre-mRNAs (fig. S1). SLBP is also required for the export, stability, and translation of mature mRNAs. The 3′-5′ exonuclease 3′hExo (also known as Eri-1) forms a tight ternary complex with SL and SLBP. 3′hExo can trim three nucleotides in vitro from the processed histone mRNA 3′-end, and SLBP protects against further trimming (*8*–*11*). 3′hExo is required for replication-dependent histone mRNA degradation (*12*). It is also involved in microRNA homeostasis (*13*) and 5.8*S* rRNA 3′-end maturation (*14*, *15*). The stem-loop RNA consists of a six–base pair stem and a four-base loop, as well as flanking sequences at both ends (fig. S1). SLBP (31 kD) has high affinity for the SL (dissociation constant K_d = 1 to 10 nM) (*16*–*20*). It contains a ~70-residue RNA binding domain (RBD) (Fig. 1A and fig. S2). 3′hExo (40 kD) consists of an N-terminal SAP domain (~60 residues) followed by a nuclease domain (~220 residues) that belongs to the DEDDh superfamily (Fig. 1A and fig. S3) (*9*–*11*, *21*).

We report here the crystal structure at 2.6 Å resolution of the ternary complex of human SLBP RBD, human 3′hExo (SAP and nuclease domains), and a 26-nucleotide SL with consensus sequence (Fig. 1, A and B, and table S1) (*22*). Clear electron density was observed for all 26 nucleotides of the SL (Fig. 1C). The stem (nucleotides 6 to 11 and 16 to 21) has a slightly flattened

[1]Department of Biological Sciences, Columbia University, New York, NY 10027, USA. [2]Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599, USA. [3]Program in Molecular Biology and Biotechnology, University of North Carolina, Chapel Hill, NC 27599, USA.

*To whom correspondence should be addressed. E-mail: ltong@columbia.edu

classical A-form structure (fig. S4 and table S2). Of the four nucleotides in the loop, the first (U12), second (U13), and fourth (C15) bases are flipped out (Fig. 1D). In the 3′ flanking sequence, nucleotides 22 to 25 continue the helical structure of the stem, but the base of the last nucleotide (A26) is flipped by ~180° relative to C25 (Fig. 1C). The riboses of all four nucleotides in the loop (nucleotides 12 to 15) and C25 are in the 2′ endo configuration, and the RNA backbone adopts sharp turns at these nucleotides.

The structure of SLBP RBD contains three helices (αA, αB, and αC). Helices αA and αC interact with the 5′ flanking sequence, the 5′ arm of the stem, and the loop of the RNA (Fig. 1B), consistent with earlier data (*9*, *16*, *18*). In particular, helix αC is positioned closest to the SL and may function as a ruler that can measure the length of the stem, with residues near its N terminus (conserved Lys^{177}-Tyr^{178}-Ser^{179}-Arg^{180}-Arg^{181} motif, fig. S2) contacting the 5′-end of the stem and the 5′ flanking sequence and its C-terminal region contacting the loop.

The only direct recognition between SL and SLBP is through the guanine base of the second nucleotide of the stem (G7), via two hydrogen bonds with the side-chain guanidinium group of Arg^{181} (Fig. 2A). The side chain of Tyr^{144} (αA) is π-stacked with the first and the third base, and the side chain of His^{195} (αC) with the fourth base of the loop (Fig. 2B). Other interactions are primarily between the RNA backbone and the SLBP RBD (Fig. 2C and fig. S5).

Nucleotides 3 to 5 in the 5′ flanking sequence, also implicated in binding to SLBP (*9*, *16*, *18*), have interactions with the RBD (fig. S6). Besides residues Tyr^{178} and Ser^{179}, the connection between αA and αC is not in direct contact with the RNA (Fig. 1B). This segment contains the conserved Thr^{171}-Pro^{172}-Asn^{173}-Lys^{174} sequence, and Thr^{171} phosphorylation produces a factor of 7 enhancement in the affinity for SL (*19*). This residue is located near the side chains of Lys^{146} (αA), Tyr^{151} (αA), and Trp^{190} (αC), and its phosphorylation may affect the positioning of the αA and αC helices (fig. S6). Tyr^{151} is part of the conserved Tyr-Asp-Arg-Tyr motif (fig. S2), and the affinity of the Tyr^{151} → Phe mutant for SL is lower by a factor of ~10 relative to wild-type SLBP (*23*).

3′hExo contacts the loop, the 3′ arm of the stem, and the 3′ flanking sequence of the SL (Fig. 1B), as suggested by earlier studies (*9*–*11*). The SAP domain contains three helices (α1 to α3) and interacts primarily with the loop of the SL through α1 (Figs. 1B and 2B). The U13 base interacts with the side chains of Tyr^{66} (α1) and Lys^{111} (α3), and the C15 base has a hydrogen bond to the side chain of Arg^{78} (α1). Additional interactions are with the backbone of the RNA (Fig. 2C and fig. S7).

Nucleotides 24 to 26 at the 3′-end of the SL are located in the active site of the nuclease domain of 3′hExo (Fig. 1B). The C25 base is π-stacked with that of C24 on one face and the side chain of Trp^{233} on the other (Fig. 3A), thereby breaking the helical pattern of the RNA. The side chain of Arg^{261} is located close to the base and ribose of both C24 and C25 (Fig. 3A). In comparison, the first two nucleotides of the 3′ flanking sequence (A22 and C23) do not make direct contacts with 3′hExo (Fig. 3B). The binding mode of the last nucleotide (A26) is similar to that of AMP in the nuclease domain reported earlier (Fig. 3A) (*21*). The phosphate group of A26 is located near the cluster of acidic side chains that coordinate two metal ions for catalysis.

The crystal also contained a 3′hExo-SL binary complex (Fig. 3C and fig. S8). SLBP RBD has low solubility, and some of it precipitated during the preparation of the complex. The nuclease domains of 3′hExo in the two complexes have essentially the same conformation (root-mean-square distance 0.4 Å). The SAP domain shows a small movement (~5° rotation), together with a movement of the RNA (Fig. 3C). However, the first eight nucleotides of the SL, including three at the base of the stem, are disordered in this binary complex (Fig. 3C). The structure of this binary complex is similar to that of 3′hExo in complex with a stem-loop RNA without any flanking sequences reported earlier (PDB entry 1ZBH), although the SAP domain in that crystal comes from another 3′hExo molecule of a domain-swapped dimer (fig. S9).

Transversion of the second base pair of the stem led to a factor of >200 reduction in affinity for SLBP, whereas transversion of the first, third, fourth, or fifth base pair led to a factor of <5 reduction (*18*), consistent with the structural observations (Fig. 2A and fig. S5). Mutation of Arg^{181} in SLBP also inhibited SL binding in yeast three-hybrid assays (*24*, *25*). In comparison, transversion of the second base pair had little effect on 3′hExo binding (*9*), also consistent with the structure (table S3). To further validate the structural observations, we introduced mutations in the SL-SLBP RBD and SL-3′hExo interfaces and determined their effects on the formation of the binary and ternary complexes. Overall, the mutagenesis results are in good agreement with the structure (fig. S10 and table S4).

The modes of SL recognition by the RBD of SLBP and the SAP domain of 3′hExo appear to be distinct from other RNA binding proteins. The SLBP RBD does not have a close structural homolog in the PDB. Although the SAP domain has structural similarity to a domain in the recombination endonuclease VII (*26*), it does not share a common mode of nucleic acid interaction with that domain.

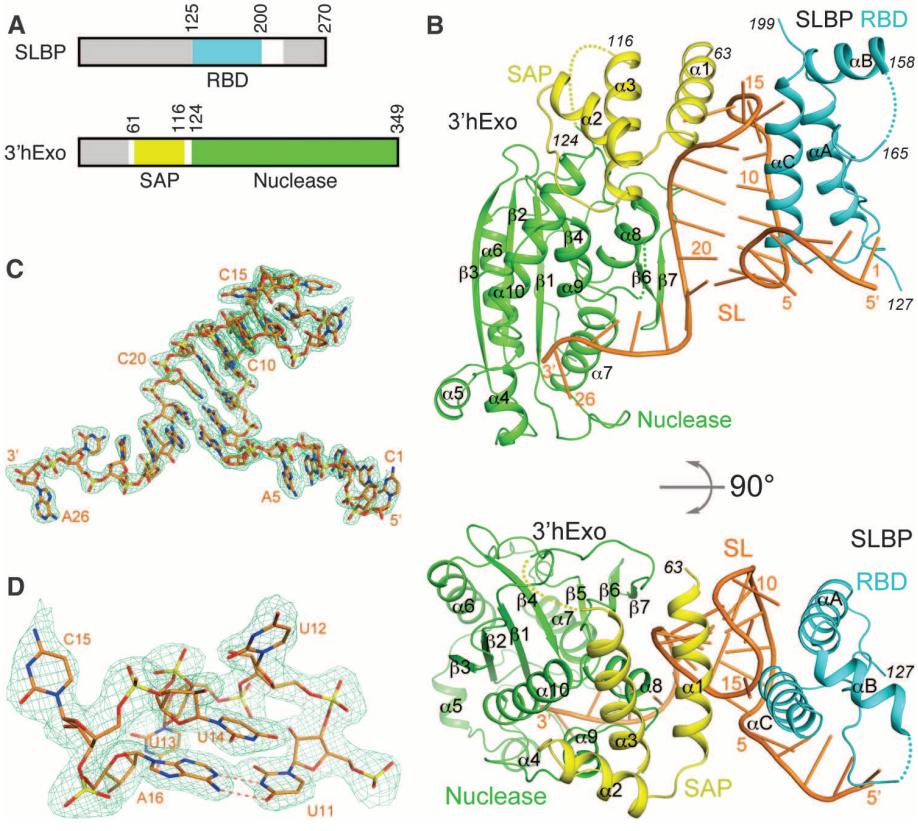

Fig. 1. Structure of human SLBP RBD, human 3′hExo, and SL RNA ternary complex. (**A**) Domain organizations of human SLBP and human 3′hExo. Residues not included in the expression constructs are shown in gray. (**B**) Schematic drawings of the structure of the ternary complex of human SLBP RBD (cyan), human 3′hExo (SAP domain in yellow and nuclease domain in green), and 26-nucleotide SL RNA (orange). (**C**) Simulated annealing omit $F_{obs} - F_{calc}$ electron density (light green) for the SL RNA at 2.6 Å resolution, contoured at 3σ. Phosphorus atoms are in yellow, oxygens in red, and nitrogens in blue. (**D**) Close-up of the loop region of the SL RNA. All the structure figures were produced with PyMOL (www.pymol.org).

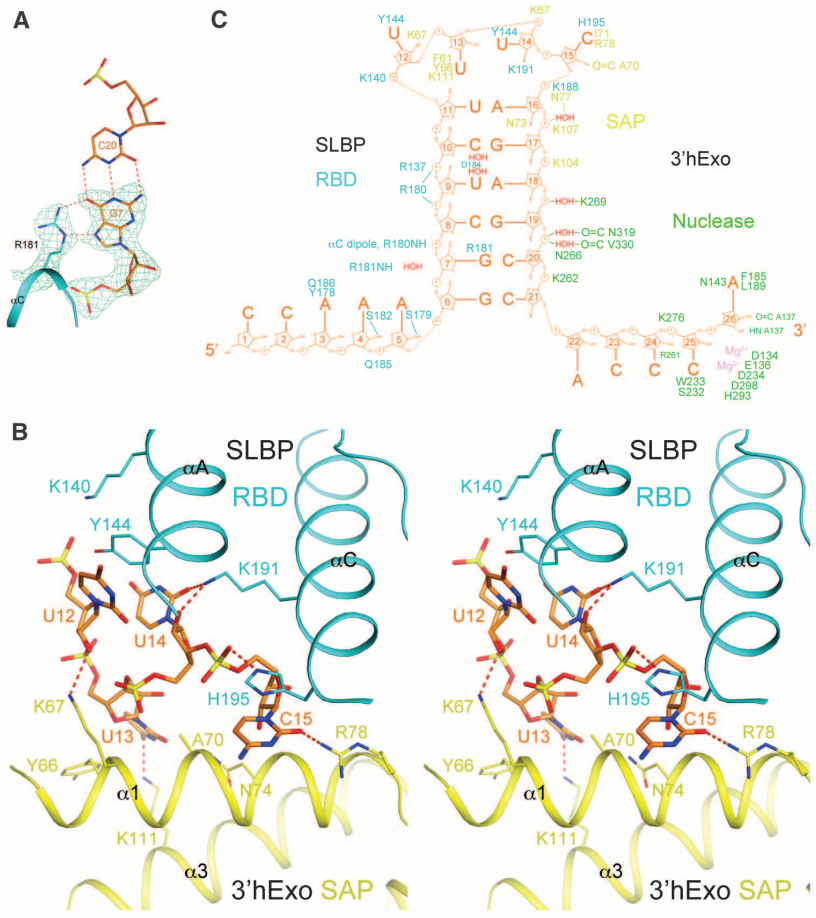

Fig. 2. Interactions between the SL RNA and SLBP RBD and 3′hExo. (**A**) Specific recognition of G7 in the second base pair of the stem (orange) by hydrogen bonding (dashed lines in red) with the side chain of Arg^{181} (cyan) of SLBP. Simulated annealing omit $F_{obs} - F_{calc}$ electron density for G7 and Arg^{181} is also shown, contoured at 5σ. (**B**) Stereopair showing interactions of the loop of the SL RNA (orange) with the SLBP RBD (cyan) and the 3′hExo SAP domain (yellow). (**C**) Schematic drawing summarizing the interactions between SL and SLBP RBD (cyan) and 3′hExo.

Our studies suggest that SLBP RBD and 3′hExo recognize the overall shape of the SL (especially its loop) rather than the sequence; this idea is also supported by observations from the single-transversion studies (*9*, *18*). At the same time, the sequence of the SL plays a role in determining its shape. The first (U12) and third (U14) nucleotides of the loop are highly conserved (fig. S11) and contribute to the specificity of recognition (*9*, *18*). *Caenorhabditis elegans* SLBP is more selective for a C at the first position of the loop, whereas human SLBP binds RNAs with C or U at the first position with comparable affinity (*17*). Tyr^{144} of human SLBP is replaced by an Arg residue in *C. elegans* SLBP, and this may result in a distinct mechanism of recognizing the C in the first nucleotide of the loop (Fig. 2B).

There are no direct contacts between SLBP RBD and 3′hExo in the ternary complex (Fig. 1B and fig. S12). The RBD and SAP domain are arranged on opposite sides of the loop, and they approach each other most closely there. Cooperative binding between the two proteins (*9*–*11*) is likely attributable to induced structural changes in the loop, such that binding of one protein induces a conformation of the loop that promotes the binding of the other protein. In structures of the SL alone in solution (*27*, *28*), the conformation of the loop region is different from that in the complex observed here (fig. S13).

3′hExo has primarily bipartite interactions with the SL. The SAP domain recognizes the loop while the nuclease domain binds the 3′ flanking sequence (Fig. 1B). Disruption of interactions at either of these two sites leads to reduced binding (*9*, *10*). Nucleotide A26 would be the leaving group for the 3′-5′ exonuclease activity (Fig. 3A), which does not show sequence preference (*9*),

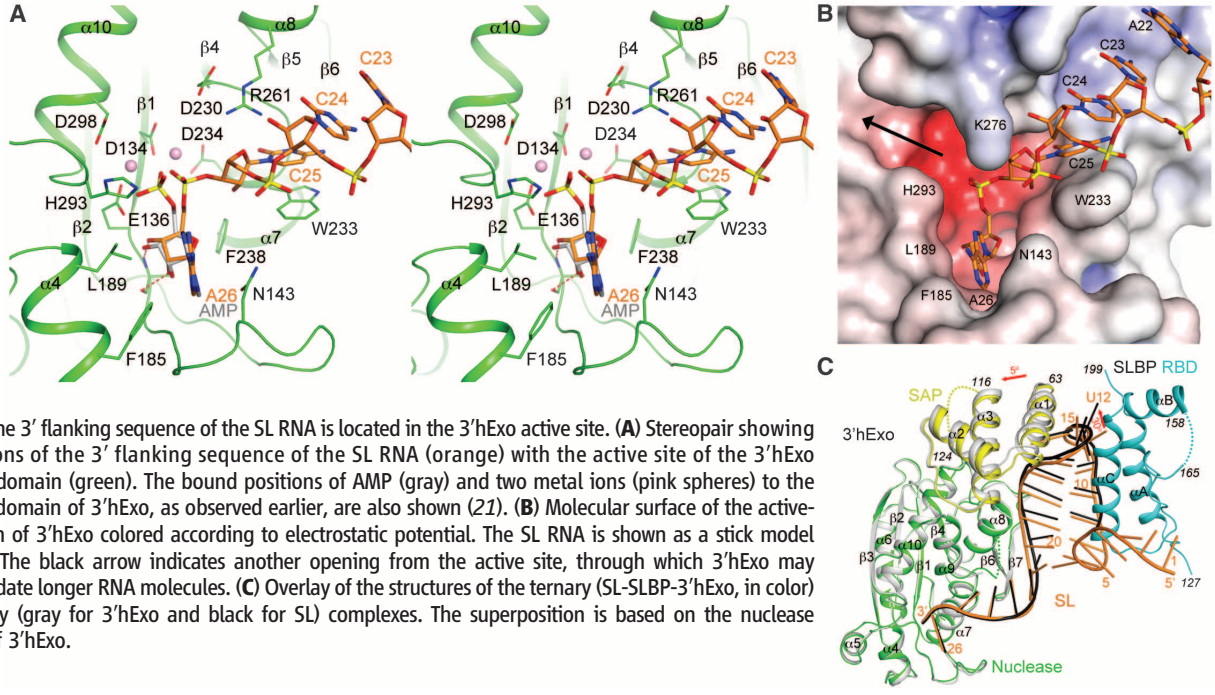

Fig. 3. The 3′ flanking sequence of the SL RNA is located in the 3′hExo active site. (**A**) Stereopair showing interactions of the 3′ flanking sequence of the SL RNA (orange) with the active site of the 3′hExo nuclease domain (green). The bound positions of AMP (gray) and two metal ions (pink spheres) to the nuclease domain of 3′hExo, as observed earlier, are also shown (*21*). (**B**) Molecular surface of the active-site region of 3′hExo colored according to electrostatic potential. The SL RNA is shown as a stick model (orange). The black arrow indicates another opening from the active site, through which 3′hExo may accommodate longer RNA molecules. (**C**) Overlay of the structures of the ternary (SL-SLBP-3′hExo, in color) and binary (gray for 3′hExo and black for SL) complexes. The superposition is based on the nuclease domain of 3′hExo.

as neither C25 nor A26 is recognized specifically. Although 3′hExo can remove the last three nucleotides of the SL (*9*), further degradation is not possible because the 3′-end of the shortened SL can no longer reach the active site of 3′hExo in the ternary complex (Fig. 3B), thereby explaining how SLBP protects histone mRNAs from excessive trimming by 3′hExo.

Besides recognizing the SL RNA, another function of SLBP is the recruitment of U7 snRNP and stabilization of its interaction with the histone pre-mRNA for 3′-end processing (fig. S1) (*23*, *29*). The 20 residues immediately C-terminal to the RBD of SLBP are required for this processing (*29*). These residues are present in the recombinant SLBP used in the current structural studies, but they are disordered. A second region required for processing is located in helix αB of the RBD, especially the Tyr-Asp-Arg-Tyr motif (Fig. 1B and fig. S6), where mutation of the Asp and Arg residues to Gln and Cys, respectively, did not affect binding but abolished processing (*23*). Our structure shows that these two regions are likely located close to each other (fig. S6) and therefore also identifies a surface feature of SLBP that is involved in histone pre-mRNA 3′-end processing (fig. S14).

References and Notes

1. Z. Dominski, W. F. Marzluff, *Gene* **396**, 373 (2007).
2. W. F. Marzluff, E. J. Wagner, R. J. Duronio, *Nat. Rev. Genet.* **9**, 843 (2008).
3. J. Zhao, L. Hyman, C. L. Moore, *Microbiol. Mol. Biol. Rev.* **63**, 405 (1999).
4. C. R. Mandel, Y. Bai, L. Tong, *Cell. Mol. Life Sci.* **65**, 1099 (2008).
5. Z. F. Wang, M. L. Whitfield, T. C. Ingledue 3rd, Z. Dominski, W. F. Marzluff, *Genes Dev.* **10**, 3028 (1996).
6. F. Martin, A. Schaller, S. Eglite, D. Schümperli, B. Müller, *EMBO J.* **16**, 769 (1997).
7. K. L. Mowry, J. A. Steitz, *Science* **238**, 1682 (1987).
8. T. E. Mullen, W. F. Marzluff, *Genes Dev.* **22**, 50 (2008).
9. Z. Dominski, X.-C. Yang, H. Kaygun, M. Dadlez, W. F. Marzluff, *Mol. Cell* **12**, 295 (2003).
10. X.-C. Yang, M. Purdy, W. F. Marzluff, Z. Dominski, *J. Biol. Chem.* **281**, 30447 (2006).
11. X.-C. Yang, M. P. Torres, W. F. Marzluff, Z. Dominski, *Mol. Cell. Biol.* **29**, 4045 (2009).
12. K. P. Hoefig *et al.*, *Nat. Struct. Mol. Biol.* 10.1038/nsmb.2450 (2012).
13. M. F. Thomas *et al.*, *Blood* **120**, 130 (2012).
14. K. M. Ansel *et al.*, *Nat. Struct. Mol. Biol.* **15**, 523 (2008).
15. H. W. Gabel, G. Ruvkun, *Nat. Struct. Mol. Biol.* **15**, 531 (2008).
16. A. S. Williams, W. F. Marzluff, *Nucleic Acids Res.* **23**, 654 (1995).
17. F. Michel, D. Schümperli, B. Müller, *RNA* **6**, 1539 (2000).
18. D. J. Battle, J. A. Doudna, *RNA* **7**, 123 (2001).
19. C. H. Borchers *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 3094 (2006).
20. M. Zhang, T. T. Lam, M. Tonelli, W. F. Marzluff, R. Thapar, *Biochemistry* **51**, 3215 (2012).
21. Y. Cheng, D. J. Patel, *J. Mol. Biol.* **343**, 305 (2004).
22. See supplementary materials on *Science* Online.
23. Z. Dominski, J. A. Erkmann, J. A. Greenland, W. F. Marzluff, *Mol. Cell. Biol.* **21**, 2008 (2001).
24. F. Martin, F. Michel, D. Zenklusen, B. Müller, D. Schümperli, *Nucleic Acids Res.* **28**, 1594 (2000).
25. S. Jaeger, G. Eriani, F. Martin, *FEBS Lett.* **556**, 265 (2004).
26. C. Biertümpfel, W. Yang, D. Suck, *Nature* **449**, 616 (2007).
27. E. S. DeJong, W. F. Marzluff, E. P. Nikonowicz, *RNA* **8**, 83 (2002).
28. K. Zanier *et al.*, *RNA* **8**, 29 (2002).
29. Z. Dominski, L. X. Zheng, R. Sanchez, W. F. Marzluff, *Mol. Cell. Biol.* **19**, 3561 (1999).

Acknowledgments: We thank N. Whalen, S. Myers, R. Jackimowicz, and H. Robinson for access to the X29A beamline at the National Synchrotron Light Source. Supported by NIH grants GM077175 (L.T.) and GM029832 (W.F.M. and Z.D.). The structure has been deposited at the Protein Data Bank (accession code 4HXH).

Supplementary Materials

www.sciencemag.org/cgi/content/full/339/6117/318/DC1
Materials and Methods
Figs. S1 to S14
Tables S1 to S4
References (*30*–*40*)

10 August 2012; accepted 14 November 2012
10.1126/science.1228705

Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Surnames are paternally inherited in most human societies, resulting in their co-segregation with Y-chromosome haplotypes (*1*–*5*). Based on this observation, multiple genetic genealogy companies offer services to reunite distant patrilineal relatives by genotyping a few dozen highly polymorphic short tandem repeats across the Y chromosome (Y-STRs). The association between surnames and haplotypes can be confounded by nonpaternity events, mutations, and adoption of the same surname by multiple founders (*5*). The genetic genealogy community addresses these barriers with massive databases that list the test results of Y-STR haplotypes along with their corresponding surnames. Currently, there are at least eight databases and numerous surname project Web sites that collectively contain hundreds of thousands of surname-haplotype records (table S1).

The ability of genetic genealogy databases to breach anonymity has been demonstrated in the past. In a number of public cases, male adoptees and descendants of anonymous sperm donors used recreational genetic genealogy services to genotype their Y-chromosome haplotypes and to search the companies' databases (*6*–*9*). The genetic matches identified distant patrilineal relatives and pointed to the potential surnames of their biological fathers. By combining other pieces of demographic information, such as date and place of birth, they fully exposed the identity of their biological fathers. Lunshof *et al.* (*10*) were the first to speculate that this technique could expose the full identity of participants in sequencing projects. Gitschier (*11*) empirically approached this hypothesis by testing 30 Y-STR haplotypes of CEU participants in these databases and reported that potential surnames can be detected. [CEU participants are multigenerational families of northern and western European ancestry in Utah who had originally had their samples collected by CEPH (Centre d'Etude du Polymorphisme Humain) and were later reconsented to participate in the HapMap project.] However, these surnames could match thousands of individuals, and the study did not pursue full re-identification at a single-person resolution.

Our goal was to quantitatively approach the question of how readily surname inference might be possible in a more general population, apply this approach to personal genome data sets, and demonstrate end-to-end identification of individuals with only public information. We show that full identities of personal genomes can be exposed via surname inference from recreational genetic genealogy databases followed by Internet searches. In all cases in which individuals were studied who had donated DNA samples, the informed consent statements they had signed stated privacy breach as a potential risk and the data usage terms did not prevent re-identification. Representatives of relevant organizations that funded the original studies were notified and confirmed the compliance of this study with their guidelines (*12*).

As a primary resource for surname inference, we focused on Ysearch (www.ysearch.org) and

[1]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA. [2]Harvard–Massachusetts Institute of Technology (MIT) Division of Health Sciences and Technology, MIT, Cambridge, MA 02139, USA. [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [4]Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, MA 02114, USA. [5]Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, TX 77030, USA. [6]Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel. [7]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. [8]Department of Molecular Microbiology and Biotechnology, Tel-Aviv University, Tel Aviv 69978, Israel. [9]The International Computer Science Institute, Berkeley, CA 94704, USA.

*To whom correspondence should be addressed. E-mail: yaniv@wi.mit.edu