EXTENDED PDF FORMAT
SPONSORED BY
Travel Grants Available
AACR Annual Meeting 2014
Application Deadline:
January 31st
>Apply Now
R&D
http://www.rndsystems.com

Science
AAAS

Exonic Transcription Factor Binding Directs Codon Choice and Affects Protein Evolution

Andrew B. Stergachis *et al.*

Science **342**, 1367 (2013);
DOI: 10.1126/science.1243490

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by clicking here.

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines here.

The following resources related to this article are available online at www.sciencemag.org (this information is current as of February 3, 2014):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
http://www.sciencemag.org/content/342/6164/1367.full.html

Supporting Online Material can be found at:
http://www.sciencemag.org/content/suppl/2013/12/11/342.6164.1367.DC1.html

A list of selected additional articles on the Science Web sites **related to this article** can be found at:
http://www.sciencemag.org/content/342/6164/1367.full.html#related

This article **cites 61 articles**, 32 of which can be accessed free:
http://www.sciencemag.org/content/342/6164/1367.full.html#ref-list-1

This article has been **cited by** 1 articles hosted by HighWire Press; see:
http://www.sciencemag.org/content/342/6164/1367.full.html#related-urls

This article appears in the following **subject collections:**
Evolution
http://www.sciencemag.org/cgi/collection/evolution

Downloaded from www.sciencemag.org on February 3, 2014

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2013 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

References and Notes

1. R. E. Lenski, M. R. Rose, S. C. Simpson, S. C. Tadler, *Am. Nat.* **138**, 1315–1341 (1991).
2. C. L. Burch, L. Chao, *Genetics* **151**, 921–927 (1999).
3. D. M. Weinreich, N. F. Delaney, M. A. Depristo, D. L. Hartl, *Science* **312**, 111–114 (2006).
4. S. Kryazhimskiy, G. Tkačik, J. B. Plotkin, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 18638–18643 (2009).
5. H.-H. Chou, H.-C. Chiu, N. F. Delaney, D. Segrè, C. J. Marx, *Science* **332**, 1190–1192 (2011).
6. A. I. Khan, D. M. Dinh, D. Schneider, R. E. Lenski, T. F. Cooper, *Science* **332**, 1193–1196 (2011).
7. I. G. Szendro, M. F. Schenk, J. Franke, J. Krug, J. A. G. M. de Visser, *J. Stat. Mech.* **2013**, P01005 (2013).
8. T. J. Kawecki *et al.*, *Trends Ecol. Evol.* **27**, 547–560 (2012).
9. R. E. Lenski, M. Travisano, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 6808–6814 (1994).
10. J. E. Barrick *et al.*, *Nature* **461**, 1243–1247 (2009).
11. Materials and methods and supplementary text are available as supporting material on *Science* Online.
12. P. Sibani, M. Brandt, P. Alstrøm, *Intl. J. Mod. Phys.* **12**, 361–391 (1998).
13. P. J. Gerrish, R. E. Lenski, *Genetica* **102-103**, 127–144 (1998).
14. M. Hegreness, N. Shoresh, D. Hartl, R. Kishony, *Science* **311**, 1615–1617 (2006).
15. S.-C. Park, J. Krug, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 18135–18140 (2007).
16. G. I. Lang *et al.*, *Nature* **500**, 571–574 (2013).
17. S. Wielgoss *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 222–227 (2013).
18. R. J. Woods *et al.*, *Science* **331**, 1433–1436 (2011).
19. M. M. Desai, D. S. Fisher, A. W. Murray, *Curr. Biol.* **17**, 385–394 (2007).
20. F. Vasi, M. Travisano, R. E. Lenski, *Am. Nat.* **144**, 432–456 (1994).
21. R. G. Eagon, *J. Bacteriol.* **83**, 736–737 (1962).
22. S. Goyal *et al.*, *Genetics* **191**, 1309–1319 (2012).
23. S. Wielgoss *et al.*, *G3 (Bethesda)* **1**, 183–186 (2011).
24. S. F. Elena, R. E. Lenski, *Evolution* **51**, 1058–1067 (1997).
25. C. E. Paquin, J. Adams, *Nature* **306**, 368–370 (1983).

Acknowledgments: This work was supported by grants from the National Science Foundation (DEB-1019989) including the BEACON Center for the Study of Evolution in Action (DBI-0939454), and by funds from the Hannah Chair Endowment at Michigan State University. We thank three reviewers for comments; I. Dworkin, J. Krug, A. McAdam, C. Wilke, and L. Zaman for discussions; and N. Hajela for technical assistance. R.E.L. will make strains available to qualified recipients, subject to completion of a material transfer agreement that can be found at www.technologies.msu.edu/inventors/mta-cda/mta/mta-forms. Datasets and analysis scripts are available at the Dryad Digital Repository (doi:10.5061/dryad.0hc2m).

Supplementary Materials

www.sciencemag.org/content/342/6164/1364/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S7
Tables S1 to S4
References (*26–40*)

17 July 2013; accepted 4 November 2013
Published online 14 November 2013;
10.1126/science.1243357

Exonic Transcription Factor Binding Directs Codon Choice and Affects Protein Evolution

Andrew B. Stergachis,[1] Eric Haugen,[1] Anthony Shafer,[1] Wenqing Fu,[1] Benjamin Vernot,[1] Alex Reynolds,[1] Anthony Raubitschek,[2,3] Steven Ziegler,[3] Emily M. LeProust,[4]* Joshua M. Akey,[1] John A. Stamatoyannopoulos[1,5]†

Genomes contain both a genetic code specifying amino acids and a regulatory code specifying transcription factor (TF) recognition sequences. We used genomic deoxyribonuclease I footprinting to map nucleotide resolution TF occupancy across the human exome in 81 diverse cell types. We found that ~15% of human codons are dual-use codons ("duons") that simultaneously specify both amino acids and TF recognition sites. Duons are highly conserved and have shaped protein evolution, and TF-imposed constraint appears to be a major driver of codon usage bias. Conversely, the regulatory code has been selectively depleted of TFs that recognize stop codons. More than 17% of single-nucleotide variants within duons directly alter TF binding. Pervasive dual encoding of amino acid and regulatory information appears to be a fundamental feature of genome evolution.

The genetic code, common to all organisms, contains extensive redundancy, in which most amino acids can be specified by two to six synonymous codons. The observed ratios of synonymous codons are highly nonrandom, and codon usage biases are fixtures of both prokaryotic and eukaryotic genomes (*1*). In organisms with short life spans and large effective population sizes, codon biases have been linked to translation efficiency and mRNA stability (*2–7*). However, these mechanisms explain only a small fraction of observed codon preferences in mammalian genomes (*7–11*), which appear to be under selection (*12*).

Genomes also contain a parallel regulatory code specifying recognition sequences for transcription factors (TFs) (*13*), and the genetic and regulatory codes have been assumed to operate independently of one another and to be segregated physically into the coding and noncoding genomic compartments. However, the potential for some coding exons to accommodate transcriptional enhancers or splicing signals has long been recognized (*14–18*).

To define intersections between the regulatory and genetic codes, we generated nucleotide-resolution maps of TF occupancy in 81 diverse human cell types using genomic deoxyribonuclease I (DNaseI) footprinting (*19*). Collectively, we defined 11,598,043 distinct 6– to 40–base pair (bp) footprints genome-wide (~1,018,514 per cell type), 216,304 of which localized completely within protein-coding exons (~24,842 per cell type) (Fig. 1, A and B; fig. S1A; and table S1). Approximately 14% of all human coding bases contact a TF in at least one cell type (average 1.1% per cell type) (Fig. 1C and fig. S1B), and 86.9% of genes contained coding TF footprints (average 33% per cell type) (fig. S1, C and D).

The exonic TF footprints we observed likely underestimate the true fraction of protein-coding bases that contact TFs because (i) TF footprint detection increases substantially with sequencing depth (*13*), and (ii) the 81 cell types sampled, although extensive, is far from complete; we saw little evidence of saturation of coding TF footprint discovery (fig. S2).

To ascertain coding footprints more completely, we developed an approach for targeted exonic footprinting via solution-phase capture of DNaseI-seq libraries using RNA probes complementary to human exons (*19*). Targeted capture footprinting of exons from abdominal skin and mammary stromal fibroblasts yielded ~10-fold increases in DNaseI cleavage—equivalent to sequencing >4 billion reads per sample by using conventional genomic footprinting (fig. S3A)—quantitatively exposing many additional TF footprints (fig. S3, B to D). Overall, we identified an average of ~175,000 coding footprints per cell type (fig. S1E), which is 7- to 12-fold more than with conventional footprinting.

Although coding sequences are densely occupied by TFs in vivo, the density of TF footprints at different genic positions varied widely, with many genes exhibiting sharply increased density in the translated portion of their first coding exon (Fig. 1D and fig. S4A). In contrast, internal coding exons were as likely as flanking intronic sequences to harbor TF footprints (Fig. 1D). The total number of coding DNaseI footprints within a gene was related both to the length of the gene and to its expression level (fig. S4, B to D).

Given their abundance, we sought to determine whether exonic TF binding elements were under evolutionary selection. Fourfold degenerate coding bases are frequently used as a model of neutral (or nearly neutral) evolution (*20*) but

[1]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA. [2]Department of Immunology, University of Washington, Seattle, WA 98109, USA. [3]Benaroya Research Institute, Seattle, WA 98101, USA. [4]Agilent Technologies, Santa Clara, CA 95051, USA. [5]Department of Medicine, University of Washington, Seattle, WA 98195, USA.

*Present address: Twist Bioscience, San Francisco, CA 94158, USA.
†Corresponding author. E-mail: jstam@uw.edu

may exhibit constraint when a functional signal impinges on coding sequence (*11*). Across the coding compartment, fourfold degenerate bases (4FDBs) within TF footprints show significantly greater evolutionary constraint versus nonfootprinted 4FDBs (Fig. 1E and fig. S5, A and B), indicating that TF-DNA recognition constrains the third codon position.

To test for evolutionary constraint at coding footprints in modern human populations, we quantified the age of mutations arising within or outside of coding footprints using exome sequencing data from 4298 individuals of European ancestry (fig. S5C) and 2217 individuals of African American ancestry (fig. S5D) (*21*). This analysis revealed that mutations within coding footprints were on average 10.2% younger than those outside of footprints (Fig. 1F and fig. S5E), signaling influence of coding TF elements on human fitness.

Both synonymous and nonsynonymous mutations within coding footprints were significantly younger than those outside of footprints (Fig. 1F and fig. S5E), indicating that coding TF binding constrains both codon and amino acid evolution. The genome-wide recognition sequence landscape of each TF has evolved to fit the molecular topography of its protein-DNA binding interface (Fig. 1G) (*13*). To study how specific TFs influence codon and amino acid choice at

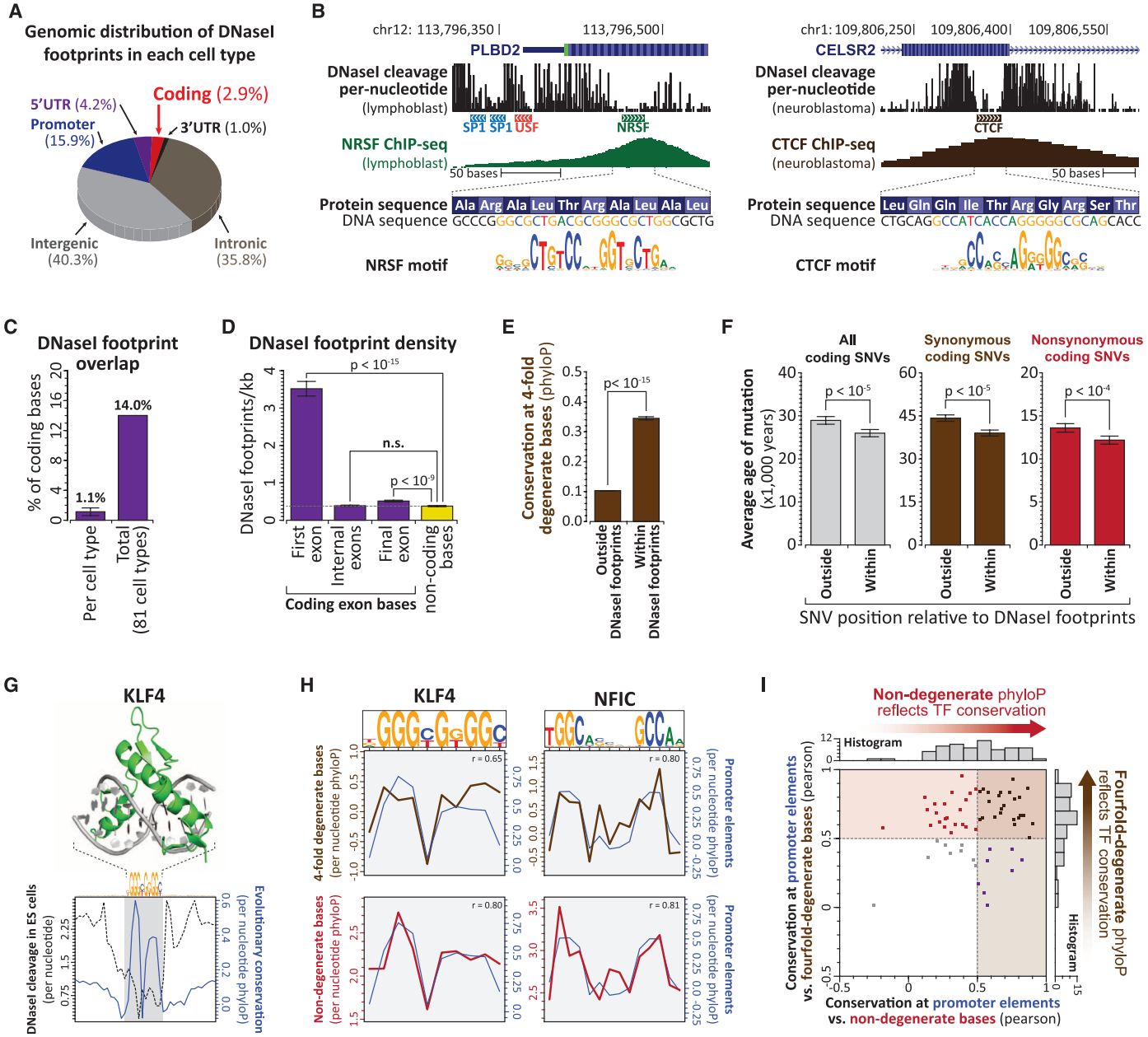

Fig. 1. TFs densely populate and evolutionarily constrain protein-coding exons. (**A**) Distribution of DNaseI footprints. (**B**) Per-nucleotide DNaseI cleavage and chromatin immunoprecipitation sequencing (ChIP-seq) signal for coding CTCF (left) and NRSF (right) binding elements. (**C**) Proportion of coding bases within DNaseI footprints in each of 81 cell types (left), or any cell type (right). (**D**) Average footprint density within first, internal, or final coding exons [mean ± SEM; *P* value, paired *t* test; nonsignificant (n.s.) indicates $P > 0.1$]. (**E**) PhyloP conservation at 4FDBs within and outside footprints. (**F**) Estimated mutational age at all (gray), synonymous (brown), and nonsynonymous (red) coding SNVs (European) within and outside footprints [*P* values per (*21*)] (**G**) Structure of DNA-bound KLF4 versus average per-nucleotide DNaseI cleavage and evolutionary constraint at KLF4 footprints. (**H**) Average per-nucleotide conservation at 4FDBs (brown) and NDBs (red) overlapping KLF4 (left) and NFIC (right) footprints [*r*, Pearson correlation; conservation at promoter bases versus 4FDBs (top) or NDBs (bottom)]. (**I**) Evolutionary constraint imparted by 63 TFs at promoter elements, 4FDBs and NDBs (Pearson correlations).

their recognition sites, we compared the per-nucleotide evolutionary conservation profiles of TF recognition sequences at noncoding 4FDBs and nondegenerate coding bases (NDBs). For example, the conservation profiles at 4FDBs and NDBs at KLF4 and NFIC recognition sites closely mirror those of recognition sites in noncoding regions (promoter) (Fig. 1H). As such, these TFs constrain both codon choice (via constraint on 4FDBs) and amino acid choice (via NDBs) encoded at their recognition sites. Analysis of conservation profiles for 63 TFs with prevalent occupancy within coding regions (*19*) showed that 73% constrain 4FDBs and 51% constrain NDBs (Fig. 1I and figs. S6 and S7). Thus, individual TFs may influence both codon and amino acid choice.

To examine how TF binding relates to codon usage patterns, we examined binding at preferred (biased) versus nonpreferred codons. For example, across all human proteins asparagine is encoded by the AAC codon 52% of the time (versus AAT, 48%), indicating a generalized 4% bias in favor of this codon. However, genome-wide, 60.4% of asparagine codons within footprints are AAC, versus only 50.8% outside of footprints (a 9.6% occupancy bias toward the preferred codon) (Fig. 2A). Apart from arginine (see below), for all amino acids encoded by two or more codons the codon that is preferentially used genome-wide is also preferentially occupied by TFs (Fig. 2B and table S2).

To determine whether preferential occupancy of biased codons is inherent to TF recognition sequences, we compared trinucleotide frequencies within coding versus noncoding footprints. Trinucleotide combinations favored by TFs within coding sequence were equivalent to those favored in noncoding sequence (Fig. 3C), indicating

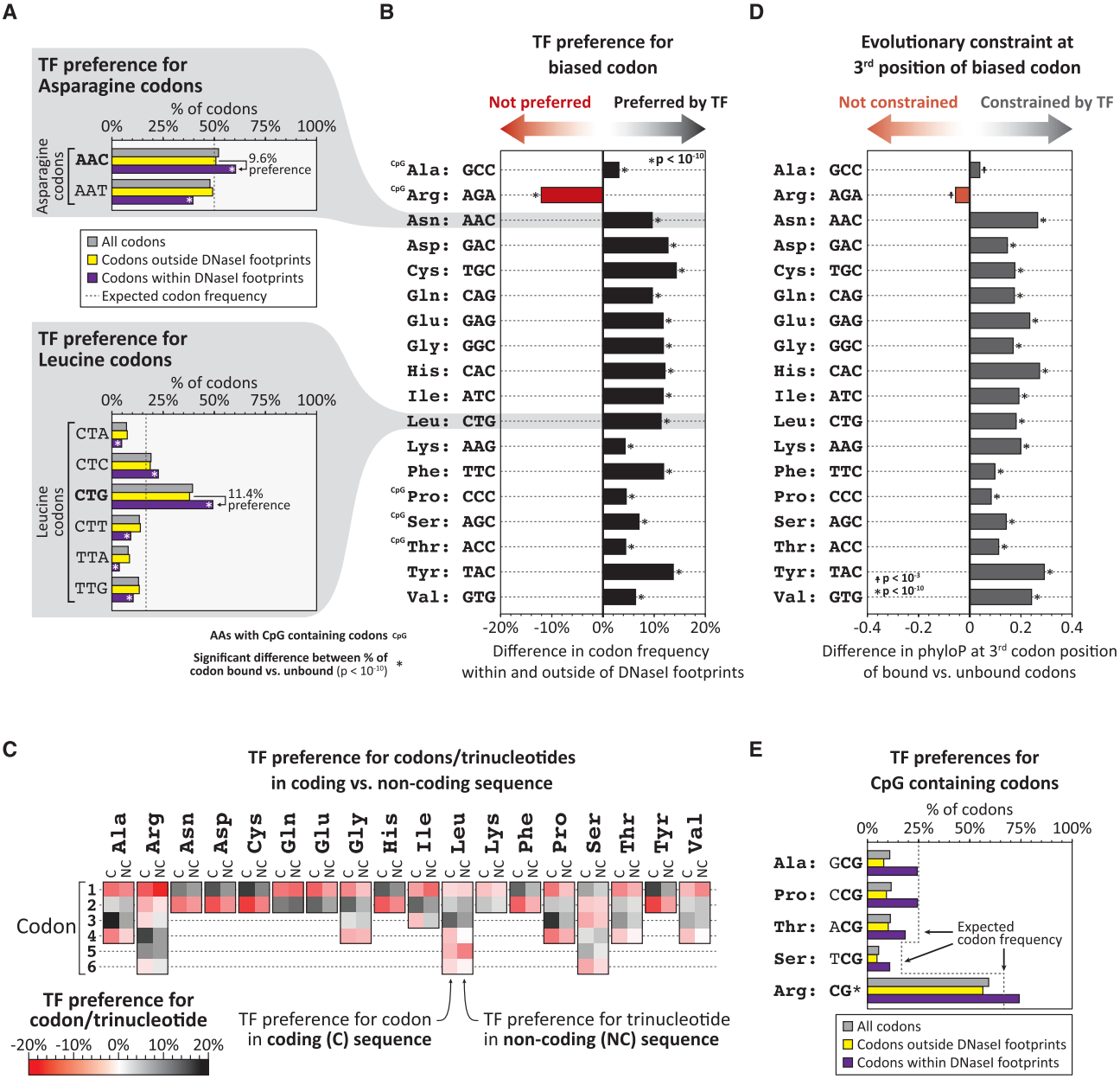

Fig. 2. TFs modulate global codon biases. (**A**) Proportions of all codons (gray), or codons outside of (yellow) or within (purple) footprints, that encode asparagine (top) or leucine (bottom). Codons with bias (AAC for asparagine and CTG for leucine) preferentially localize within footprints. (**B**) Preferential footprinting of biased codons, calculated as in (A) (*P* values, Pearson's χ^2 test). (**C**) Preferential footprinting of each codon trinucleotide in coding versus noncoding regions (C, coding; NC, noncoding). (**D**) Difference in average evolutionary constraint at third positions of biased codons outside versus within footprints (*P* values, Mann-Whitney test). (**E**) Proportions of amino acids encoded by CpG-containing codons among all codons (gray), codons outside footprints (yellow), or codons within footprints (purple).

that global TF binding preferences are directly reflected in the frequency of different codons. Of note, baseline trinucleotide frequencies within coding and noncoding sequence are largely independent of one another (table S2). The fact that the third position of preferred codons overlapping footprints is under excess evolutionary constraint (Fig. 2D and table S2) supports a general role for TFs in potentiating codon usage biases through the selective preservation of preferred codons.

Although nearly all codon biases parallel TF recognition preferences genome-wide, arginine—one of the five amino acids encoded by codons containing CpGs (four out of six codons)—was a notable exception. CpGs frequently occur in regulatory DNA (table S2) yet have an elevated mutational rate (*22*). Consequently, although TFs may favor CpG-containing codons (Fig. 2E) and impart excess constraint thereto (table S2), the higher mutational rate at such codons is likely incompatible with preferential use.

Codons outside footprints still exhibit usage biases (Fig. 2A and table S2); however, it is likely that these biases also reflect the actions of TFs. First, our conclusions above are drawn from a conservative and incomplete annotation of duons. Second, because TF trinucleotide preferences and codon biases have not changed substantially since the divergence of humans and mice (fig. S8), preferences at any given codon may result from a TF binding element extant in some ancestral species to human. Third, codon usage bias can be exaggerated because of mutual reinforcement with other cellular factors such as tRNA abundances (*23*, *24*). Indeed, such mechanisms could be linked to codon biases created by exonic TF occupancy through a feedback mechanism that potentiates intrinsic TF-imposed biases, resulting in both abundant and rare codons and associated tRNAs, differences that could in turn affect protein synthesis and stability (*25*–*27*).

To analyze positional occupancy patterns of specific TFs within coding sequence, we systematically matched TF recognition sequences with footprints, providing an accurate measure of a TF's in vivo occupancy (*13*, *28*). This analysis revealed that a subset of TFs selectively avoid coding sequences (Fig. 3A). Intriguingly, TFs involved in positioning the transcriptional preinitiation complex, such as NFYA and SP1 (*29*), preferentially avoid the translated region of the first coding exon (Fig. 3A) and typically occupy elements immediately upstream of the methionine start codon (Fig. 3B and fig. S9A). Conversely, TFs involved in modulating promoter activity, such as YY1 and NRSF, preferentially occupy the translated region of the first coding exon (Fig. 3, A and C) (*30*, *31*). These findings indicate that the translated portion of the first coding exon may serve functionally as an extension of the canonical promoter.

More broadly, the repressor NRSF preferentially occupies and evolutionarily constrains sequences coding for leucine-rich protein domains, such as signal peptide and transmembrane domains (Fig. 3D and fig. S9, B and C). Also, TFs such as CTCF and SREBP1 preferentially occupy and constrain splice sites (fig. S10, A to D), which are otherwise generally depleted of DNaseI footprints (fig. S10E). The above results suggest that specific protein structural and splicing features may undergo exaptation for specific regulatory purposes.

We also found that the occupancy of specific TFs within coding sequence parallels the extent of CpG methylation at their binding site (fig. S11). This raises the possibility that gene body methylation, which is paradoxically extensive at actively transcribed genes (*32*, *33*), may provide a tunable mechanism for thwarting opportunistic TF occupancy within coding sequence during transcription.

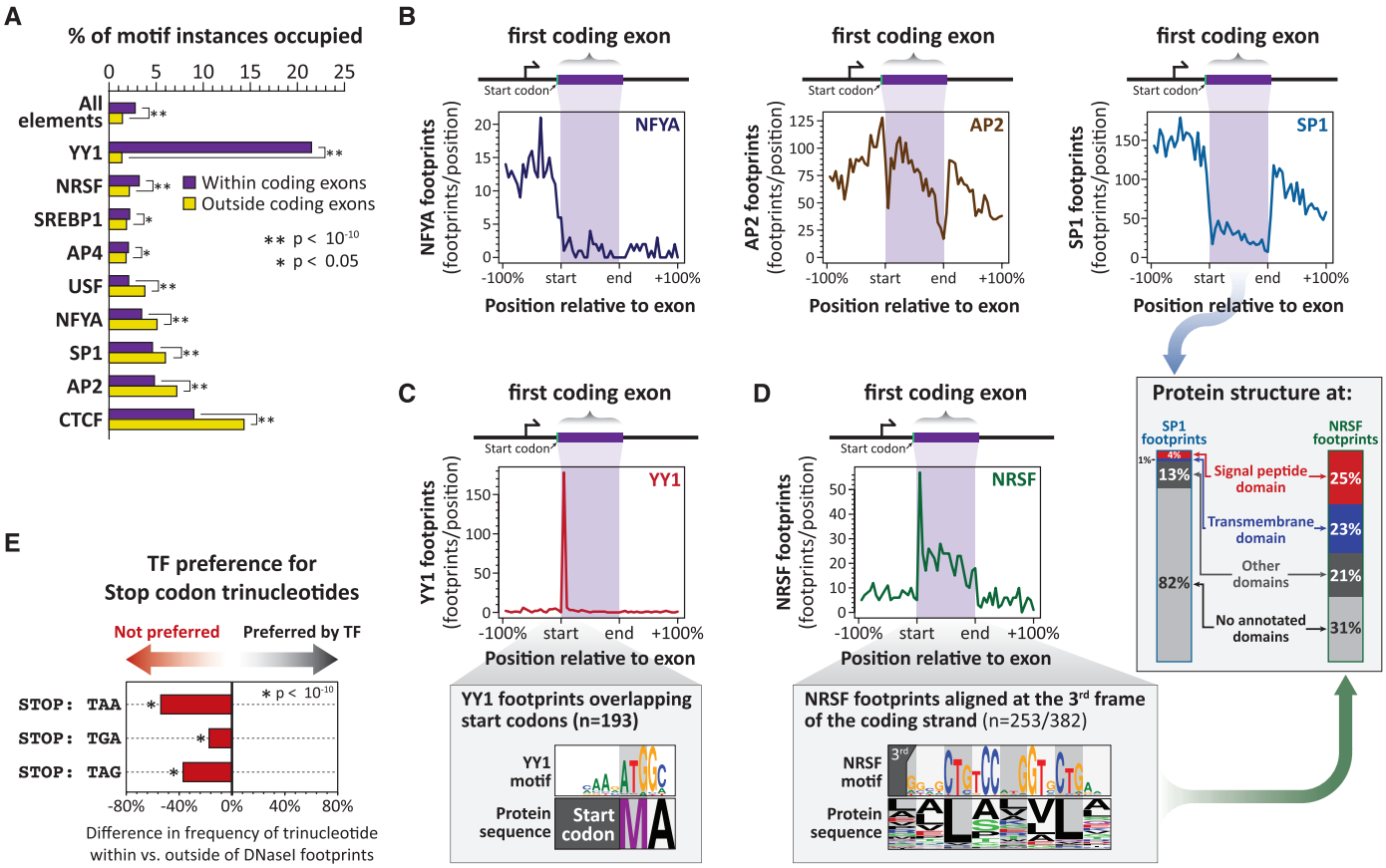

Fig. 3. TFs exploit and avoid specific coding features. (**A**) Percentage of TF motifs occupied in coding versus noncoding regions (*P* values, paired *t* test). (**B**) Density of NFYA (left), AP2 (middle), and SP1 (right) footprints relative to translated region of first coding exons. (**C**) (Top) Density of YY1 footprints across first coding exons. (Bottom) YY1 recognition sequence and corresponding amino acid sequence within YY1 footprints overlapping start codons. (**D**) (Top left and bottom) For NRSF as per (C). (Right, arrow) Protein domain annotation of first exon third-frame NRSF footprints versus SP1 footprints. (**E**) TF preference (avoidance) of stop codon trinucleotides within versus outside footprints in noncoding regions (*P* values, Pearson's χ^2 test).

If TFs, through selective recognition sequences, could impose changes in protein sequence, deleterious consequences could arise if such changes resulted in a nonsense substitution. We observed that TFs generally avoid stop codons (fig. S10E). This finding extends to noncoding regions, in which stop codon trinucleotides (TAA, TAG, and TGA) are selectively depleted within footprints. This indicates that the global TF repertoire has been selectively purged of DNA binding domains capable of recognizing—and thus preferentially stabilizing—nonsense codons (Fig. 3E and fig. S10F).

The high sequencing coverage provided by genomic footprinting revealed 592,867 heterozygous single-nucleotide variants (SNVs) across the 81 cell type samples, and 3% of coding footprints harbored heterozygous SNVs (Fig. 4A). Functional SNVs that disrupt TF occupancy quantitatively skew the allelic origins of DNaseI cleavage fragments (*13*), and 17.4% of all heterozygous coding SNVs within footprints showed this signature (Fig. 4B and fig. S12), including both synonymous and nonsynonymous variant classes (Fig. 4C). The potential of a coding SNV to disrupt overlying TF occupancy was independent of the class of variant (Fig. 4D) or whether a nonsynonymous variant was predicted to be deleterious to protein function (Fig. 4, E and F).

Of common disease- and trait-associated SNVs identified by genome-wide associated studies (GWASs) in coding sequence (*19*), 13.5% fall within duons (fig. S13A). GWAS single-nucleotide polymorphisms in duons encompass both synonymous (12%) and nonsynonymous (88%) substitutions (fig. S13A) and may directly affect pathogenetic mechanisms (fig. S13, B to F, and table S3). As such, disease-associated variants within duons may compromise both regulatory and/or protein-structural functions. These findings have substantial practical implications for the interpretation of genetic variation in coding regions.

Our results indicate that simultaneous encoding of amino acid and regulatory information within exons is a major functional feature of complex genomes. The information architecture of the received genetic code is optimized for superimposition of additional information (*34*, *35*),

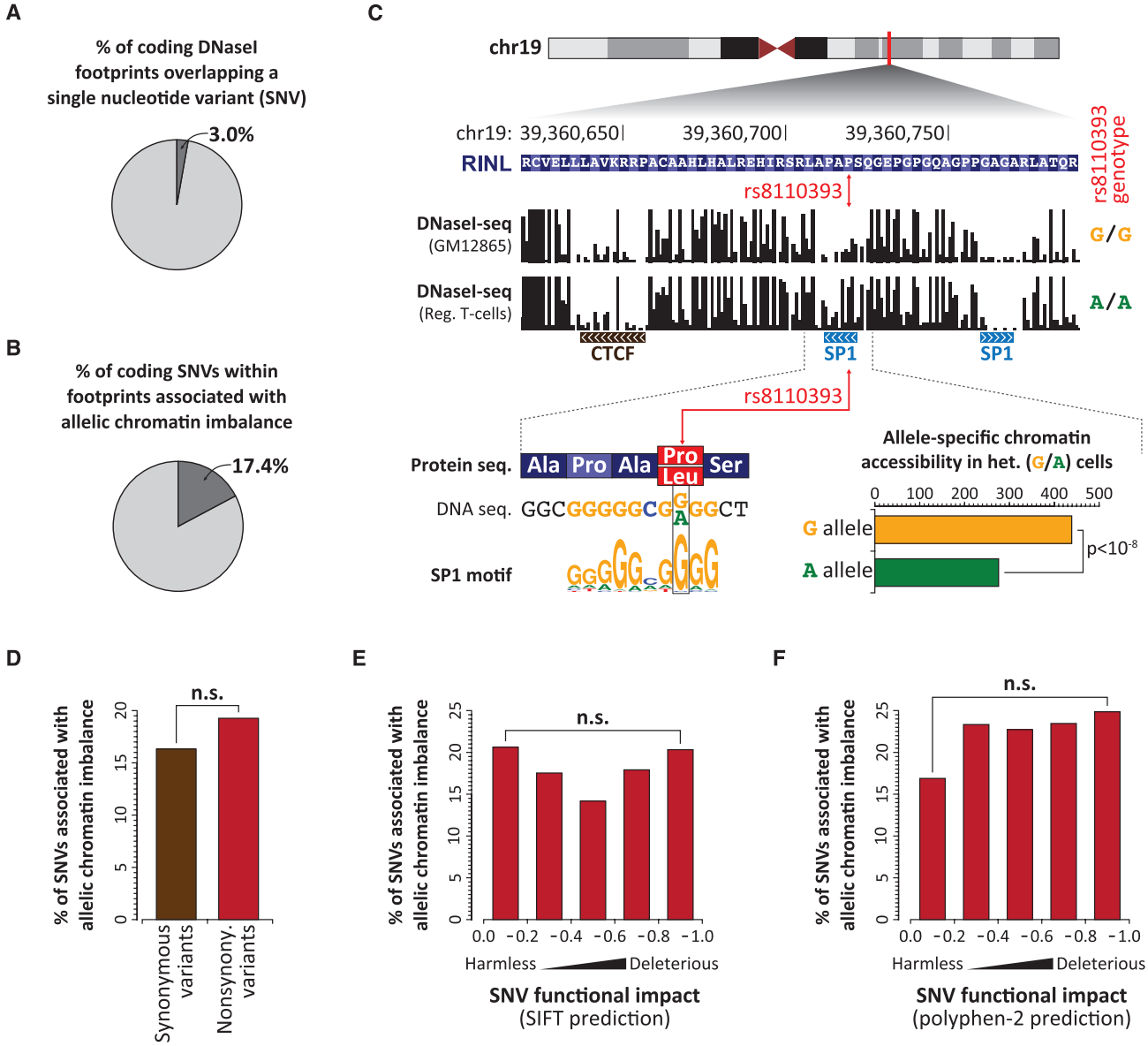

Fig. 4. Genetic variation in duons frequently alters TF occupancy. (**A**) Proportion of coding footprints overlapping a SNV in any of 81 cell types. (**B**) Proportion of SNVs in duons that allelically alter TF occupancy. (**C**) (Top) Per-nucleotide DNaseI cleavage at common nonsynonymous G→A SNV (rs8110393) in G/G and A/A homozygous cells. (Bottom) Allelic SP1 occupancy in heterozygous (G/A) cells. (**D**) Proportion of synonymous and nonsynonymous variants in duons that allelically alter TF occupancy. (**E** and **F**) Proportion of nonsynonymous variants from (D) grouped by predicted impact of coding variant on protein function using (E) SIFT or (F) Polyphen-2. None of the bins are significantly different (Fisher's exact test; n.s. indicates $P > 0.1$).

and this intrinsic flexibility has been extensively exploited by natural selection. Although TF binding within exons may serve multiple functional roles, our analyses above is agnostic to these roles, which may be complex (*36*).

References and Notes

1. R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pavé, *Nucleic Acids Res.* **8**, r49–r62 (1980).
2. T. Ikemura, *J. Mol. Biol.* **151**, 389–409 (1981).
3. R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, R. Mercier, *Nucleic Acids Res.* **9**, r43–r74 (1981).
4. M. Gouy, C. Gautier, *Nucleic Acids Res.* **10**, 7055–7074 (1982).
5. A. Eyre-Walker, M. Bulmer, *Nucleic Acids Res.* **21**, 4599–4603 (1993).
6. D. B. Carlini, W. Stephan, *Genetics* **163**, 239–243 (2003).
7. M. dos Reis, R. Savva, L. Wernisch, *Nucleic Acids Res.* **32**, 5036–5044 (2004).
8. J. L. Parmley, J. V. Chamary, L. D. Hurst, *Mol. Biol. Evol.* **23**, 301–309 (2006).
9. T. Warnecke, C. C. Weber, L. D. Hurst, *Biochem. Soc. Trans.* **37**, 756–761 (2009).
10. W. Gu, T. Zhou, C. O. Wilke, *PLOS Comput. Biol.* **6**, e1000664 (2010).
11. M. F. Lin *et al.*, *Genome Res.* **21**, 1916–1928 (2011).
12. Z. Yang, R. Nielsen, *Mol. Biol. Evol.* **25**, 568–579 (2008).
13. S. Neph *et al.*, *Nature* **489**, 83–90 (2012).
14. S. M. Hyder, Z. Nawaz, C. Chiappetta, K. Yokoyama, G. M. Stancel, *J. Biol. Chem.* **270**, 8506–8513 (1995).
15. G. Lang, W. M. Gombert, H. J. Gould, *Immunology* **114**, 25–36 (2005).
16. D. I. Ritter, Z. Dong, S. Guo, J. H. Chuang, *PLOS ONE* **7**, e35202 (2012).
17. A. H. Khan, A. Lin, D. J. Smith, *PLOS ONE* **7**, e46098 (2012).
18. R. Y. Birnbaum *et al.*, *Genome Res.* **22**, 1059–1068 (2012).
19. Materials and methods are available as supplementary materials on *Science* Online.
20. W.-H. Li, *Molecular Evolution* (Sinauer Associates, Sunderland, MA, 1997).
21. W. Fu *et al.*, *Nature* **493**, 216–220 (2013).
22. C. Coulondre, J. H. Miller, P. J. Farabaugh, W. Gilbert, *Nature* **274**, 775–780 (1978).
23. M. Bulmer, *Nature* **325**, 728–730 (1987).
24. M. Bulmer, *Genetics* **129**, 897–907 (1991).
25. J. Duan *et al.*, *Hum. Mol. Genet.* **12**, 205–216 (2003).
26. J. zur Megede *et al.*, *J. Virol.* **74**, 2628–2635 (2000).
27. J. R. Coleman *et al.*, *Science* **320**, 1784–1787 (2008).
28. R. M. Samstein *et al.*, *Cell* **151**, 153–166 (2012).
29. S. McKnight, R. Tjian, *Cell* **46**, 795–805 (1986).
30. C. Zhang *et al.*, *Nucleic Acids Res.* **34**, 2238–2246 (2006).
31. H. Xi *et al.*, *Genome Res.* **17**, 798–806 (2007).
32. A. Hellman, A. Chess, *Science* **315**, 1141–1143 (2007).
33. D. Zilberman, M. Gehring, R. K. Tran, T. Ballinger, S. Henikoff, *Nat. Genet.* **39**, 61–69 (2007).
34. S. Itzkovitz, U. Alon, *Genome Res.* **17**, 405–412 (2007).
35. S. Itzkovitz, E. Hodis, E. Segal, *Genome Res.* **20**, 1582–1589 (2010).
36. T. R. Mercer *et al.*, *Nat. Genet.*, published online 23 June 2013 (10.1038/ng.2677).

Acknowledgments: We thank many colleagues for their insightful comments and critical readings of the manuscript. We also thank many colleagues who provided individual cell samples for DNaseI analysis. We also thank E. Rynes for his technical assistance. This work was supported by NIH grants U54HG004592, U54HG007010, and U01ES01156 to J.A.S. A.B.S. was supported by grant FDK095678A from the National Institute of Diabetes and Digestive and Kidney Diseases. J.M.A. is a paid consultant for Glenview Capital. All data from this study are available through the ENCODE data repository at UCSC (www.encodeproject.org) and the Roadmap Epigenomics data repository at NCBI (www.ncbi.nlm.nih.gov/epigenomics).

Supplementary Materials

www.sciencemag.org/content/342/6164/1367/suppl/DC1
Materials and Methods
Figs. S1 to S13
Tables S1 to S3
References (*37–63*)

19 July 2013; accepted 23 October 2013
10.1126/science.1243490

Cryptic Variation in Morphological Evolution: HSP90 as a Capacitor for Loss of Eyes in Cavefish

Nicolas Rohner,[1] Dan F. Jarosz,[2]* Johanna E. Kowalko,[1] Masato Yoshizawa,[3] William R. Jeffery,[3,4] Richard L. Borowsky,[5] Susan Lindquist,[2,6,7] Clifford J. Tabin[1]†

In the process of morphological evolution, the extent to which cryptic, preexisting variation provides a substrate for natural selection has been controversial. We provide evidence that heat shock protein 90 (HSP90) phenotypically masks standing eye-size variation in surface populations of the cavefish *Astyanax mexicanus*. This variation is exposed by HSP90 inhibition and can be selected for, ultimately yielding a reduced-eye phenotype even in the presence of full HSP90 activity. Raising surface fish under conditions found in caves taxes the HSP90 system, unmasking the same phenotypic variation as does direct inhibition of HSP90. These results suggest that cryptic variation played a role in the evolution of eye loss in cavefish and provide the first evidence for HSP90 as a capacitor for morphological evolution in a natural setting.

A longstanding question in evolutionary biology is the extent to which selection acts on preexisting "standing variation" in a population, as opposed to de novo mutations. Recent studies have indicated that both mechanisms have contributed to morphological evolution (*1*, *2*). Thus, although de novo mutations may exist and contribute to phenotypic evolution, repeated use of standing variation has played an important role in the evolution in these fish. However, these observations also raise a critical question: How is genetic variation maintained in a population if it is not adaptive before new selective conditions?

Waddington proposed that developmental processes are quite robust and produce the same phenotype regardless of minor genotypic variation, a phenomenon he termed "canalization" (*3*). In such conditions, cryptic variation can accumulate and can be maintained without consequence. He further proposed that under certain environmental conditions, this property could be lost ("decanalization"), resulting in expression of the cryptic variation on which selection could act (*4*).

More recently, Lindquist demonstrated that HSP90 (heat shock protein 90) provides a molecular mechanism for buffering genetic variation and releasing it in response to environmental stress (*5–10*). The HSP90 chaperone assists in the folding of proteins that are metastable signal transducers, such as kinases, transcription factors, and ubiquitin ligases. HSP90 is normally present at much higher concentrations than needed to maintain these proteins, allowing it to act as a buffer, protecting organisms from phenotypic consequences that would otherwise be caused by genetic variants of these proteins. Because protein folding is so sensitive to environmental stress, changes in the environment can exhaust the chaperone buffer, unmasking vulnerable polymorphisms. And because multiple variants can be unmasked at the same time, this system provides a mechanism to create complex traits in a single step (*11*).

Besides changes in the activities of kinases, phosphatases, transcription factors, and ubiquitin ligases, other distinct mechanisms have been reported by which changes in HSP90 function can lead to changes in phenotype (*5*, *10*, *12–16*)

Evidence strongly suggests that this mechanism has operated in microbial populations (*7*, *8*), but its relevance to the evolution of natural populations of higher organisms remains highly controversial. Thus far, examples of HSP90-mediated canalization in multicellular eukaryotes have been limited to lab strains of various model organisms. Moreover, with the exception of some phenotypes in *Arabidopsis*, the phenotypes of HSP90-released canalization in higher organisms are not

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [2]The Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA. [3]Department of Biology, University of Maryland, College Park, MD 20742, USA. [4]Marine Biological Laboratory, Woods Hole, MA 02543, USA. [5]Department of Biology, New York University, New York, NY 10003, USA. [6]Howard Hughes Medical Institute, Cambridge, MA 02142, USA. [7]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA.

*Present address: Departments of Chemical and Systems Biology and Developmental Biology, School of Medicine, Stanford University, Stanford, CA 94305, USA.
†Corresponding author. E-mail: tabin@genetics.med.harvard.edu