RNA-mediated gene regulation is less evolvable than transcriptional regulation

Joshua L. Payne, Fahad Khalid, and Andreas Wagner

Gene expression is regulated at multiple levels ranging from the accessibility of chromatin to the posttranslational modifications of proteins. Much of this regulation is carried out by sequence-specific, nucleotide-binding proteins that target DNA or RNA molecules. Transcription factors (TFs) are one such class of proteins. They bind short DNA sequences to regulate gene expression at the level of transcription by activating or blocking the recruitment of RNA polymerase to the transcription start site (1). RNA binding proteins (RBPs) are another such class of proteins. They bind short RNA sequences to regulate gene expression posttranscriptionally by regulating the splicing of precursor mRNA as well as the stability, transport, translation, and decay of mature mRNA (2).

Mutations that affect the regulation of gene expression are often deleterious. This is evidenced by the numerous diseases associated with mutations in the nucleic acid binding sites of regulatory proteins (3–6). For example, spinal muscular atrophy, a pediatric neurodegenerative disorder, is caused by a point mutation in an exonic splicing element, which abrogates binding of an RNA binding protein than they are to create interactions of a gene regulatory region with a new transcription factor. Our observations are consistent with the high level of conservation observed for interactions between RNA binding proteins and their target molecules as well as the evolutionary plasticity of regulatory regions bound by transcription factors. They may help explain why transcriptional regulation is implicated in many more evolutionary adaptations and innovations than RNA-mediated gene regulation.

Much of gene regulation is carried out by proteins that bind DNA or RNA molecules at specific sequences. One class of such proteins is transcription factors, which bind short DNA sequences to regulate transcription. Another class is RNA binding proteins, which bind short RNA sequences to regulate RNA maturation, transport, and stability. Here, we study the robustness and evolvability of these regulatory mechanisms. To this end, we use experimental binding data from 172 human and fruit fly transcription factors and RNA binding proteins as well as human polymorphism data to study the evolution of binding sites in vivo. We find little difference between the robustness of regulatory protein–RNA interactions and transcription factor–DNA interactions to DNA mutations. In contrast, we find that RNA-mediated regulation is less evolvable than transcriptional regulation, because mutations are less likely to create interactions of an RNA molecule with a new RNA binding protein than they are to create interactions of a gene regulatory region with a new transcription factor. Our observations are consistent with the high level of conservation observed for interactions between RNA binding proteins and their target molecules as well as the evolutionary plasticity of regulatory regions bound by transcription factors. They may help explain why transcriptional regulation is implicated in many more evolutionary adaptations and innovations than RNA-mediated gene regulation.

Significance

Cells regulate the activity of genes in a variety of ways. For example, they regulate transcription through DNA binding proteins called transcription factors, and they regulate mRNA stability and processing through RNA binding proteins. Based on current knowledge, transcriptional regulation is more widespread and is involved in many more evolutionary adaptations than posttranscriptional regulation. The reason could be that transcriptional regulation is studied more intensely. We suggest instead that transcriptional regulation harbors an intrinsic evolutionary advantage: when mutations change transcriptional regulation, they are more likely to bring forth novel patterns of such regulation. That is, transcriptional regulation is more evolvable. Our analysis suggests a reason why a specific kind of gene regulation is especially abundant in the living world.

Edited by Rafael Sanjuán, Universitat de Valencia, Valencia, Spain, and accepted by Editorial Board Member Daniel L. Hartl February 28, 2018 (received for review November 2, 2017)

This open access article is distributed under a Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Author contributions: J.L.P. and A.W. designed research; J.L.P. and F.K. performed research; J.L.P., F.K., and A.W. analyzed data; and J.L.P. and A.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. R.S. is a guest editor invited by the Editorial Board.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

1To whom correspondence should be addressed. Email: Josha Payne@env.ethz.ch.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1719138115/-/DCSupplemental.

Published online March 26, 2018.
mutation to any one genotype on the network has the potential to produce another genotype that is also on the network and therefore, also has the same phenotype. Second, genotype networks tend to spread far throughout the space of all possible genotypes, which provides mutational access to the genotype networks of different phenotypes. Genotype networks, therefore, confer both robustness and evolvability to phenotypes (25).

The study of genotype-phenotype maps is currently being transformed by technological advances in high-throughput sequencing and chip-based technologies (26–43) as well as by synthetic biology (44). These approaches facilitate the assignment of phenotypes to a large number of genotypes and thus, provide the opportunity to empirically characterize a genotype–phenotype map. Examples of such phenotypes include the ability of RNA molecules to bind an aptamer (29), the fluorescent activity of proteins (39), and the formation of a spatial stripe by synthetic gene regulatory circuits (44). For some genotype–phenotype maps, a phenotype can be assigned to all possible genotypes. Specifically, protein binding microarrays provide measurements of the affinity with which a TF binds to all possible 8-nt DNA sequences (45), and RNAcompete provides measurements of the affinity with which an RBP binds to all but two of the possible 7-nt RNA sequences (46). Data like these can be thought of as exhaustively enumerated, empirically derived genotype–phenotype maps, where the genotype is a DNA or RNA sequence and the phenotype is the molecular capacity of the sequence to bind a TF or RBP, respectively (31).

Here, we use experimental data from protein binding microarrays and RNAcompete to study and compare the robustness and evolvability of the nucleic acid binding sites of TFs and RBPs via a comparative analysis of their genotype–phenotype maps. In these maps, a genotype is a nucleic acid sequence, and a phenotype is the molecular capacity of the sequence to bind a TF or RBP. These genotype–phenotype maps are particularly amenable to comparative analysis because of the many similarities between these two forms of protein–nucleotide interactions. For example, the biophysics of binding are similar for the two nucleic acids, such that binding affinity between nucleotides in the DNA or RNA sequence and amino acids in the protein’s binding domain is largely determined via intermolecular interactions, such as ionic and hydrogen bonding (47, 48). Additionally, the number of DNA sequences profiled by protein binding microarrays is of the same order of magnitude as the number of RNA sequences profiled by RNAcompete, and this number is small enough to facilitate the exhaustive analysis of binding preferences for many TFs and RBPs (49, 50). Finally, both of these datasets exhibit the three hallmark characteristics of genotype–phenotype maps: (i) multiple distinct sequences bind the same protein (49, 50); (ii) some proteins are bound by many distinct sequences, whereas others are bound by very few (49, 50); and (iii) the sequences that bind a protein tend to form a single-genotype network (31, 51). Despite these similarities, we show here that these two genotype–phenotype maps exhibit pronounced architectural differences, which suggest that the nucleic acid binding sites of RNA-mediated gene regulation are less evolvable than those of transcriptional regulation.

**Results**

**Protein Binding Microarray and RNAcompete Data.** We use experimental data from protein binding microarrays (50) and RNAcompete (49) to construct and compare two empirical genotype–phenotype maps. These data are comparable, because both technologies use the same analysis pipeline to transform the fluorescent intensities of microarray spots into an enrichment score (E-score) for all possible sequences of a short length. The E-score is a variant of the Wilcoxon–Mann–Whitney statistic that ranges from −0.5 to 0.5 and describes the relative binding preferences of a protein to its nucleic acid ligands, with larger numbers indicating increased preference. Protein binding microarrays profile the binding specificity of a TF by assigning an E-score to all 32,896 possible 8-nt DNA sequences, whereas RNAcompete profiles the binding specificity of an RBP by assigning an E-score to 16,382 7-nt RNA sequences (Materials and Methods). By setting a threshold on the E-score distribution, one can delineate sequences that specifically bind a TF or RBP from those that do not. Thus, we can assign the binary phenotype of bound or unbound to each sequence respective to each TF and RBP. To do so, we use an E-score threshold of 0.35 following earlier work (31, 41, 52, 53) (Materials and Methods). In SI Appendix, we consider more stringent binding affinity thresholds to show that our findings are qualitatively insensitive to this parameter.

Because there are currently far more protein binding microarray data available than RNAcompete data, we choose our study species according to the availability of the RNAcompete data. Specifically, we study proteins from *Drosophila melanogaster* and *Homo sapiens*, because these species have more RNAcompete data available than any other species (Materials and Methods, Table 1, and Dataset S1). These proteins include 38 TFs and 71 RBPs from *H. sapiens* as well as 30 TFs and 33 RBPs from *D. melanogaster*.

We construct a genotype network for each TF and for each RBP. In such a network, vertices represent sequences that bind the TF or RBP (E-score > 0.35), and edges connect vertices if their corresponding sequences differ by a single small mutation (Materials and Methods) (31). Since some sequences bind multiple TFs or RBPs, genotype networks may overlap. Additionally, the sequences that bind a TF or RBP sometimes form multiple, disconnected genotype networks (SI Appendix, Table S1). In this case, there is always one genotype network that is much larger than the others (Dataset S1). We restrict our analyses to this largest genotype network for each TF and RBP as in our previous work (31, 41, 51).

Fig. 1 shows the sizes of the genotype networks of TF and RBP binding sites, expressed as fractions of genotype space, for both *D. melanogaster* and *H. sapiens*. We consider this fractional size rather than absolute size to facilitate the comparison of these two genotype–phenotype maps, which differ in the sizes of their respective genotype spaces. The range of fractional genotype network sizes is similar for the two classes of proteins, although there

---

**Table 1. Data analyzed in this study**

<table>
<thead>
<tr>
<th>Type of protein</th>
<th>Species</th>
<th>No. of proteins in the database</th>
<th>No. of proteins in our dataset</th>
<th>No. of binding domain classes in our dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF</td>
<td><em>H. sapiens</em></td>
<td>41</td>
<td>38 (29)</td>
<td>11 (9)</td>
</tr>
<tr>
<td>TF</td>
<td><em>D. melanogaster</em></td>
<td>30</td>
<td>30 (23)</td>
<td>10 (7)</td>
</tr>
<tr>
<td>RBP</td>
<td><em>H. sapiens</em></td>
<td>77</td>
<td>71 (38)</td>
<td>9 (5)</td>
</tr>
<tr>
<td>RBP</td>
<td><em>D. melanogaster</em></td>
<td>48</td>
<td>33 (20)</td>
<td>4 (4)</td>
</tr>
</tbody>
</table>

A protein was included in our dataset if it bound at least one sequence (E-score > 0.35). Numbers in parentheses correspond to a reduced dataset, in which we only consider proteins that have a genotype network that occupies >0.5% of the genotype space.
are fewer small genotype networks of TF binding sites than of RBP binding sites. For example, 16 genotype networks of TF binding sites (24%) occupy less than 0.5% of genotype space, whereas 46 genotype networks of RBP binding (44%) occupy less than 0.5% of genotype space (dashed vertical lines in Fig. 1).

**Robustness.** The mutational robustness of an individual binding site refers to the likelihood that a mutation in this site leaves binding intact. We quantify this robustness as the fraction of all possible mutations that yield a sequence that is part of the same genotype network. We then generalize this quantity by computing the average robustness of all sites that bind a given TF or RBP (31, 54), which provides a measure of binding site robustness specific to any one nucleic acid binding protein.

The distributions of robustness are similar for binding sites of TFs and RBPs (Fig. 2 A–D). On average, the robustness of *D. melanogaster* TFs is 0.33, whereas the average robustness of RBPs is 0.27. A nearly identical pattern is found in *H. sapiens*, where the average robustness for TFs is 0.33 compared with 0.26 for RBPs. However, the left-hand tail of the distribution is longer for RBPs. This means that some RBPs have considerably lower robustness than any of the TFs (31, 54), which provides a measure of binding site robustness specific to any one nucleic acid binding protein.

The distributions of robustness are similar for binding sites of TFs and RBPs (Fig. 2 A–D). On average, the robustness of *D. melanogaster* TFs is 0.33, whereas the average robustness of RBPs is 0.27. A nearly identical pattern is found in *H. sapiens*, where the average robustness for TFs is 0.33 compared with 0.26 for RBPs. However, the left-hand tail of the distribution is longer for RBPs. This means that some RBPs have considerably lower robustness than any of the TFs (31, 54), which provides a measure of binding site robustness specific to any one nucleic acid binding protein.

To determine why robustness can be higher for TFs than for RBPs, we studied the relationship between robustness and genotype network size. Fig. 2E shows that robustness increases approximately logarithmically with the size of the genotype network for both classes of proteins and in both species. This relationship may be a generic property of genotype–phenotype maps as computational models suggest (23, 24, 55, 56). The white symbols in the lower left-hand corner of the plot in Fig. 2E show that the low-robustness RBPs have small genotype networks. This indicates that robustness is often higher for TFs, because they do not have genotype networks as small as RBPs.

**Evolvability.** In the context of nucleic acid binding sites, evolvability is the propensity of a mutation to bring forth a new binding phenotype (31). We quantify the evolvability of a TF’s or RBP’s binding sites in two steps. First, we determine the set of sequences that differ by a single mutation from any of the protein’s binding sites but that are not part of the protein’s genotype network. Second, we determine the fraction of TFs or RBPs in our dataset that these one-mutant neighbors bind (31, 54). This fraction is our measure of evolvability. It takes on high values when the one-mutant neighbors of a genotype network bind many TFs or RBPs and low values when the sequences neighboring a genotype network bind few TFs or RBPs.

We find that TF binding sites are considerably more evolvable than RBP binding sites in both *D. melanogaster* and *H. sapiens* (Fig. 3 A–D) (*P* < 10⁻⁵ for both comparisons, one-sided Wilcoxon rank sum test). For example, in *D. melanogaster*, 29 of 30 TFs are more evolvable than even the most evolvable RBP. More generally, mutations to TF binding sites have the potential to create binding sites for 96% of the TFs in our dataset (65 TFs) on average, whereas mutations to RBP binding sites have the potential to create binding sites for 60% of the RBPs in our dataset (62 RBPs) on average. This means that the average evolvability of TFs is more than 1.5 times that of RBPs in both species. Mutations to TF binding sites are, therefore, more likely to bring forth new binding phenotypes than are mutations to RBP binding sites.

We next studied the relationship between evolvability and genotype network size to better understand how the genotype networks of different binding phenotypes are distributed in the genotype spaces of the two genotype–phenotype maps. Fig. 3E shows that evolvability increases with genotype network size for both TFs and RBPs but that the rate of increase is slower for RBPs. This raises the possibility that nucleic acid binding phenotypes are arranged differently in the two genotype spaces, because genotype networks of similar size confer different levels of evolvability as do phenotypes of similar robustness (SI Appendix, Fig. S1). Perhaps genotype networks of TF binding sites are nearer to one another in genotype space than are those of RBP binding sites? To find out, we calculated the average mutual distance between all pairs of genotypes from different genotype networks for all pairs of TFs and for all pairs of RBPs. This measure provides a sense of the proxim-
sequences may also bind other proteins, this analysis amounts to a further characterization of genotype network overlap. Fig. 5 shows that, for any number \( n \) of mutations, a greater proportion of binding phenotypes can be reached in the genotype space of transcriptional regulation than in that of RNA-mediated gene regulation. For example, with just a single mutation \( (n = 1) \), 39% of \( D. \) melanogaster TFs and 23% of \( H. \) sapiens TFs can be reached compared with 25% of \( D. \) melanogaster RBPs and 9% of \( H. \) sapiens RBPs. This difference is even more pronounced at the largest number of mutations \( (n = 8 \text{ for TFs and } n = 7 \text{ for RBPs}) \), where 75% of \( D. \) melanogaster TFs and 67% of \( H. \) sapiens TFs can be reached compared with just 44% of \( D. \) melanogaster RBPs and 23% of \( H. \) sapiens RBPs. In the second variant of this analysis, we considered all genotypes within a given number of \( n \) mutations of a focal genotype, regardless of whether these genotypes belong to the same genotype network as the focal genotype (i.e., regardless of whether the nucleic acid sequences bind the same protein). This measure, therefore, simultaneously characterizes the overlap and juxtaposition of genotype networks. SI Appendix, Fig. S4 shows that, at any number of \( n \) mutations, one can reach at least as many and usually more novel binding phenotypes for TFs than for RBPs. As the mutational radius approaches the diameter of the genotype space, the fraction of reachable phenotypes will necessarily approach one. However, the number of mutations at which this occurs is always smaller in the genotype space of TF binding sites than in the genotype space of RBP binding sites. For example, in \( D. \) melanogaster, all phenotypes only become reachable for RBP binding sites at the largest possible number of \( n = 7 \) mutations, whereas \( n = 4 \) mutations—just one-half of the maximum distance—suffice to reach all TF binding phenotypes. In sum, both variants of this analysis show that the nucleic acid binding sites of transcriptional regulation are more evolvable, because a given number of mutations can reach a larger number of new binding phenotypes.

At least some of the reduced robustness and evolvability of RBP binding sites relative to TF binding sites stems from the presence of small genotype networks in our RBP dataset. This raises the question of whether the abundance of small genotype networks is truly a characteristic feature of this genotype–phenotype map or is actually the result of a sampling bias in our dataset. Such bias could occur if our dataset includes a nonrepresentative proportion of RBPs that have small genotype networks. One way to determine if this bias exists is to compare the distributions of the number of bound sequences per RBP in our dataset with those of the remaining 68 RBPs for which binding data are available (Materials and Methods). The two distributions are statistically indistinguishable \( (P = 0.24, \text{Wilcoxon rank sum test}) \) (SI Appendix, Fig. S5). To the extent that the RNAcompete data are representative of RBP binding preferences in general, this suggests that small genotype networks are indeed a characteristic feature of the genotype–phenotype map of RBP binding sites. In SI Appendix, we also show that our main conclusions are insensitive to the removal of proteins with small genotype networks (SI Appendix, Figs. S6–S8) and to the single-nucleotide difference in the lengths of the TF and RBP binding sites that we study (SI Appendix, Figs. S9–S12). Additionally, we show that our main conclusions are insensitive to the relaxation of some of our modeling assumptions, including the affinity threshold used to delineate bound from unbound sequences (SI Appendix, Figs. S13–S18) and the exclusion of small genotype network components from our evolvability measure (SI Appendix, Fig. S19).
TF binding sites are more evolvable than RBP binding sites. Histograms of the distribution of evolvability for TF binding sites in (A) D. melanogaster and (B) H. sapiens and for RBP binding sites in (C) D. melanogaster and (D) H. sapiens. (E) Evolvability is shown in relation to the fractional size of the genotype network, showing that evolvability increases more rapidly with genotype network size and reaches a higher maximum value for TF binding sites than for RBP binding sites.

Binding Site Variants in the Human Population. The measures of robustness and evolvability that we studied here take into consideration all of the DNA and RNA sequences that bind a TF or RBP, respectively. Moreover, they assume that all types of point mutations to these sequences are equally likely. However, only a subset of all DNA and RNA sequences is used for gene regulation in vivo, and mutations to these sequences may be subject to biases. These include context-dependent mutation rates (57) as well as simple transition:transversion biases (58). In other words, our observations above need not hold for the binding sites encountered in vivo or for their mutational variants.

To find out if they do, we studied putative TF and RBP binding sites in humans and the single-nucleotide mutants of these sequences that exist as standing variation (Materials and Methods). Specifically, we collected DNaseI footprint data from 41 diverse cell and tissue types (59). These data demarcate protein-bound regions of the genome at single-nucleotide resolution genome-wide and can, therefore, be used to predict TF binding sites. We focused on footprints that are likely to be involved in gene regulation by filtering the footprints to only include those that overlap the promoter regions of protein-coding genes. We also collected RNase footprints from HeLa cells, which demarcate protein-bound regions of the transcriptome at single-nucleotide resolution transcriptome-wide and can, therefore, be used to predict RBP binding sites (60). We focused on footprints that are likely to be involved in gene regulation by filtering the footprints to only include those that were in the 5′ or 3′ UTRs of protein-coding transcripts, because these are common regulatory targets of RBPs (49). For each of the DNaseI and RNase footprints in these regulatory regions, we determined the likely bound TFs or RBPs based on the available binding data (49, 50) (Materials and Methods). This resulted in a list of putative binding sites for each TF and RBP. To determine which mutational variants of these binding sites exist as standing variation in the human population, we queried the 1000 Genomes Project Consortium data for SNPs (61). This allowed us to determine which of a binding site’s mutational variants likely abrogate binding to the focal protein and which do not. For those that are likely to abrogate binding, we determined which other TFs or RBPs the mutational variant might bind. In this way, we determined the robustness and evolvability of the set of nucleic acid sequences that are encountered as putative binding sites in vivo and how these properties relate to the size of a protein’s genotype network.

Fig. 6E shows what as the size of a genotype network increases, so does the number of a binding site’s mutational variants that exist as standing variation in the human population and that do not abrogate binding to the focal protein (i.e., the mutational variants of a binding site are on the same genotype network as the binding site). This pattern is similar to that of Fig. 2E, indicating that the relationship between robustness and genotype network size is consistent among the subset of binding sites encountered in vivo and the superset of sequences characterized in vitro. Moreover, the distributions of robustness are statistically indistinguishable among TFs and RBPs ($P = 0.83$, one-sided Wilcoxon rank sum test), consistent with our observation that their binding sites are similarly robust to mutation (Fig. 2). Much of the standing variation in human nucleic acid binding sites is, therefore, unlikely to abrogate binding, especially for TFs and RBPs with large genotype networks.

Fig. 6B shows that, as the size of a genotype network increases, so does the number of TFs or RBPs that are bound by the one-mutant neighbors of binding sites that abrogate binding to the focal protein (i.e., the mutational variants of a binding site are not on the same genotype network as the binding site but
Genotype networks of TF binding sites are typically separated by fewer mutations and exhibit more overlap than genotype networks of RBP binding sites. (A) Bar heights show the number of mutations that are needed to convert two binding sites from different genotype networks into one another averaged across all pairs of binding sites from all pairs of genotype networks and normalized by the length of the binding site. (B) Bar heights show the number of binding sites that are shared between two genotype networks averaged over all pairs of genotype networks and normalized by the number of possible binding sites. Note the logarithmic scale of the y axis. In both panels, the error bars represent a single SD.

Fig. 4. Genotype networks of TF binding sites are typically separated by fewer mutations and exhibit more overlap than genotype networks of RBP binding sites. (A) Bar heights show the number of mutations that are needed to convert two binding sites from different genotype networks into one another averaged across all pairs of binding sites from all pairs of genotype networks and normalized by the length of the binding site. (B) Bar heights show the number of binding sites that are shared between two genotype networks averaged over all pairs of genotype networks and normalized by the number of possible binding sites. Note the logarithmic scale of the y axis. In both panels, the error bars represent a single SD.

Discussion

The nucleic acid sequences that bind TFs dictate when, where, and to what extent genes are transcribed. The resulting transcripts comprise nucleic acid sequences that bind RBPs to regulate much of the transcripts’ lifecycle, including splicing, transport, and decay. Interactions between nucleic acid binding sites and their cognate regulatory proteins are, therefore, fundamental to the regulation of gene expression. Here, we used experimental data to study the robustness and evolvability of these interactions. We did so via a comparative analysis of their genotype-phenotype maps. In these maps, a genotype is a nucleic acid sequence, and a phenotype is the molecular capacity of the sequence to bind a protein (31). For each TF and RBP, we constructed a genotype network from the sequences that bind the protein, and we studied how these genotype networks confer robustness and evolvability to the binding sites they harbor.

These empirical genotype-phenotype maps exhibit two architectural differences. First, the genotype networks of RBP binding sites are often much smaller than those of TF binding sites, which means that the bindings sites of some RBPs are less robust than those of TFs. Second, the genotype networks of TF binding sites interface and overlap with one another more than the genotype networks of RBP binding sites, rendering the space of TF binding sites more conducive to the evolution of new binding phenotypes. This observation is consistent with our analysis of standing genetic variation in the human population as reported by the 1000 Genomes Project Consortium (61). Specifically, standing variation in TF binding sites is more likely to confer new binding phenotypes than standing variation in RBP binding sites.

There are additional facets to the robustness and evolvability of transcriptional and RNA-mediated gene regulation that we do not study here. For example, the robustness of transcriptional regulation can be enhanced by homotypic clusters of TF binding sites, shadow enhancers, and redundant TFs (62), whereas the evolvability of RNA-mediated gene regulation can be enhanced by alternative splicing, alternative polyadenylation, and RNA editing (63). Even in the context of nucleic acid binding sites, which are the focus of this study, there is an additional facet to evolvability that we do not consider but that may further contribute to the greater evolvability of TF binding sites. It is the propensity for binding sites to evolve de novo in regulatory regions (64, 65). The reason is that, in higher eukaryotes, such as the species studied here, the transcriptional regulation of gene expression is typically implemented by a promoter and several enhancers (66, 67), and each of these can span well over 1 kb of DNA (68, 69). This is a larger mutational target than the 5′ and 3′ UTRs of transcripts, which in humans, tend to span ~200 bp and ~1 kbp pair, respectively (70). While RBP binding sites are known to evolve de novo in these regions (71), the larger size of promoters and enhancers makes it likely that TF binding sites are also more evolvable in the context of de novo evolution.

Our evolvability measure is specifically concerned with mutations to binding sites that change the binding partner from one protein to another. Such mutations can indeed lead to adaptive changes in gene expression as examples from transcriptional regulation show (11, 72, 73). For instance, the initiation and progression of cancer are parts of an evolutionary process, in which mutations facilitate the uncontrolled division and spread of abnormal cells throughout the body. Such mutations commonly arise in regulatory regions (72, 74, 75) as evidenced by recurrent mutations in the binding sites of CEBP TFs, which create high-affinity binding sites for other TFs (72). These mutations are found across a diversity of cancer types, which suggests that they drive a change in gene expression that is selectively advantageous for cancer cells. We do not know of an analogous example for RBP binding sites. The reason may be that they are not as well-studied, that they show lower evolvability, or both.

Our study has three limitations that are worth highlighting. The first is the size of our dataset. The human genome encodes ~1,400 TFs and ~700 mRNA-binding RBPs (6, 76), whereas our dataset comprises 38 human TFs and 71 human RBPs. The
Inclusion of more proteins will necessarily affect the architecture of the genotype–phenotype maps that we study. However, the consistency of our conclusions across the *D. melanogaster* and *H. sapiens* datasets, which comprise different binding domains and different numbers of proteins per binding domain, provides reassurance that our findings are general. The second limitation is that we use indirect evidence of in vivo protein–nucleotide interactions, because footprinting assays do not reveal which proteins are bound by a specific nucleic acid sequence. We ameliorate this limitation by only studying footprints that contain sequences that bind the proteins in our dataset. Ideally, however, we would study data from assays that provide direct evidence of protein–nucleotide interactions, but these data are available neither for most of the proteins in our dataset nor for the 1000 Genomes Project Consortium data. The third limitation is the length of the binding sites that we study, which are representative of the binding preferences of many but not all TFs and RBPs (49, 50). For example, Cys2-His2 zinc finger proteins typically bind longer sequences (79), to which we cannot extrapolate our findings. As our ability to predict (80) and measure (81) the binding preferences of such proteins continues to advance, it will become possible to extend our analyses to longer nucleic acid ligands.

In sum, our comparative analysis of two empirical genotype–phenotype maps suggests that the nucleic acid binding sites of RNA-mediated gene regulation are less evolvable than those of transcriptional regulation. This observation is consistent with the high levels of mRNA target conservation for RBPs (10, 82, 83) and the evolutionary plasticity of the regulatory regions involved in transcriptional regulation (84–86) as well as their role in the evolution of myriad adaptations and innovations (11, 87, 88).

**Materials and Methods**

**Protein Binding Microarray and RNAcompete Data.** The largest databases for protein binding microarray and RNAcompete data are CIS-BP (50) and CSBP-RNA (49), respectively. There is currently far more protein binding microarray data available than RNAcompete data. Specifically, the most recent build of the CIS-BP database (version 1.02) contains protein binding microarray data for 1,665 TFs from 132 species, whereas the most recent build of the CSBP-RNA database (version 0.6) contains RNAcompete data for 194 RBPs from 24 eukaryotic species. Our choice of study species was, therefore, guided by the availability of the RNAcompete data, because they are more limited. We chose to study proteins from *D. melanogaster* and *H. sapiens*, because these species have more RNAcompete data available than any of the other species (Table 1 and Dataset S1). We limited our study to two species, because the species with the next largest number of RBPs in the CISBP-RNA database was *Caenorhabditis elegans*, which had only 15 RBPs profiled. More specifically, we downloaded protein binding microarray data for 30 *D. melanogaster* and 38 *H. sapiens* TFs from the CIS-BP database (version 1.02) (50), and we downloaded RNAcompete data for 33 *D. melanogaster* and 71 *H. sapiens* RBPs from the web supplement of ref. 49. Both the protein binding microarray and RNAcompete data include a nonparametric rank-based enrichment score (E-score) that can be used to delineate sequences that specifically bind a TF or RBP from those that do not. The protein binding microarray data include an E-score for all possible 8-nt dsDNA sequences. The total number of such sequences is $4^8 = 65,536$, because each sequence is merged with its reverse complement and because there are $4^8$ sequences that are identical to their reverse complement and therefore, cannot be merged (89). The RNAcompete data include an E-score for nearly all possible 7-nt dsRNA sequences. The total number of such sequences is $4^7 = 12,862$, where the $-2$ accounts for the two sequences (GCTGCTTC and GAAGAAGG) that contain an Sapi restriction site, which is used during the RNA pool synthesis to remove linker sequences from PCR products (46). The data from these two experimental protocols are comparable, because the numbers of nucleic acid sequences profiled are of the same order of magnitude and because the E-score is calculated in the same way and has the same meaning in the two datasets. Following earlier work (31, 41, 52), we consider a sequence to bind a TF or RBP if its E-score exceeds 0.35. We chose this threshold not only because it has precedent (31, 41, 52, 53), but also because an analysis of the relationship between E-score and false discovery rate for 104 mouse TFs revealed that sequences with an E-score greater than 0.35 had a false discovery rate less than 0.001 (31). In *SI Appendix*, we perform a sensitivity analysis of our results to this binding affinity threshold.

For each TF and RBP, E-scores were provided from two distinct array designs. We considered the E-score of a DNA or RNA sequence to be the average of the two E-scores. To include a TF or RBP in our dataset, we required that it bind at least one sequence (i.e., that at least one sequence had an average E-score > 0.35). For 11 TF and for 12 RBPs, data were available from more than one experiment. For these proteins, we chose the experiment with the largest number of bound sequences (Dataset S1) to avoid any bias toward small genotype networks in the genotype–phenotype map of RBP binding sites. In a supplementary analysis reported in *SI Appendix*, Fig. S5, we studied RNAcompete data for an additional 68 RBPs. We processed these in the same way as the data from the 33 *D. melanogaster* and 71 *H. sapiens* RBPs. In supplementary analyses reported in *SI Appendix*, Figs. S9–S12, we studied the space of all possible $4^7/2 = 8,192$ RBP-nt dsDNA sequences (note that, in this case, we do not have to account for sequences that are
identical to their reverse complement, because such sequences only exist when the sequence length is an even number). We used the same analysis pipeline to calculate $E$-scores for these sequences as used for the 8-nt DNA sequences studied in the text.

Both the CIS-BP and the CISBP-RNA databases have been updated since their original release. This is why our calculation of the total number of TFs and RBPs in these databases does not match the numbers reported in the original manuscripts (49, 50). We calculated the number of TFs in our dataset that are bound by the binding site variants that do not remain on the focal TF's or RBP's genotype network, shown in relation to the fractional size of the genotype network (horizontal axis). Binding site variants are from phase 3 of the 1000 Genomes Project Consortium (61). Black symbols correspond to the robustness and evolvability of binding sites in DNaseI footprints that overlap promoters averaged across 41 cell and tissue types. White symbols correspond to the robustness and evolvability of binding sites in RNase footprints that overlap the 5′ and 3′ UTRs of transcripts in HeLa cells.

CISBP-RNA database by entering “RNAcompete” as By Evidence Type in the homepage of the CISBP-RNA website and downloading the resulting .csv file. We removed all RBPs labeled as “RNAcompete_CONSTRUCTS” in the Species column and counted the number of RBPs and the number of unique species. Analogously, we calculated the number of RBPs and species in the

**Human Polymorphism Data.** We used data from phase 3 of the 1000 Genomes Project, which includes 84.7 million SNPs from 2,504 individuals representing 26 human populations (61). We filtered the dataset to only include SNPs that passed quality control. These data and the data described below pertain to build hg19 of the human genome.

We sought to determine the amount of variation in TF and RBP binding sites. To do so, we first identified regions of the human genome that may be involved in gene regulation by TFs or RBPs. For TFs, we identified protein-bound regions of gene promoters using digital footprints from DNaseI hypersensitivity assays (59). These data provide genome-wide evidence of protein–DNA interactions across 41 cell and tissue types at single-nucleotide resolution. We used BEDTools to filter the footprints (90), such that we only retained footprints that overlapped promoter regions by at least 1 bp. We defined the promoter region of a gene to be the DNA sequence 1 kb upstream of the gene’s transcription start site. We restricted our attention to genes encoding mRNA as indicated by the prefix NM in the RefSeq database (91). For these same genes, we identified 5′ and 3′ UTRs as potential regulatory targets for RBPs. We identified protein-bound regions in mRNA transcripts using footprints from an RNase hypersensitivity assay (60), which provides transcriptome-wide evidence of protein–RNA interactions in HeLa cells. We used bedtools to filter the footprints, such that we only retained footprints that overlapped 5′ and 3′ UTRs by at least 1 bp.

For each TF, we scanned each promoter footprint for potential TF binding sites, and for each RBP, we scanned each 5′ UTR footprint and each 3′ UTR footprint for potential RBP binding sites. The scanning worked as follows. If a footprint contained at least one sequence with an $E$-score $> 0.35$ for the TF or RBP, then we considered this sequence to be a potential binding site. If the footprint contained more than one binding site, then we randomly chose one of them for later analysis. We chose to assign only one binding site per footprint per TF or RBP to avoid any bias that may be introduced by the different lengths of DNaseI footprints and RNase footprints.

The above procedure resulted in a list of genomic coordinates of potential binding sites for each TF and RBP. The number of such TF binding sites ranged from 266 for the TF FOXP6 in HFF cells to a maximum of 50,126 for the TF DNMT1 in NB4 cells. The number of such RBP binding sites ranged from 0 for the RBPs ENOX1, FXR2, and ZNF638 to a maximum of 32,272 for the RBP HNRNPH2. For each of these potential binding sites, we used the Samtools function tabix to query the 1000 Genomes Consortium data for SNPs (92). We used the results of these queries to calculate the proportion of binding site variants that is on the focal TF’s or RBPs genotype network and the proportion that is not. For those variants that are not on the genotype network, we calculated the fraction of TFs or RBPs in our dataset that these variants have the potential to bind.

**Genotype Networks.** We constructed a genotype network for each TF and RBP following the procedure of Payne and Wagner (31). Specifically, for each protein, we used the Smith–Waterman algorithm to perform a pairwise alignment on all pairs of bound sequences ($E$-score $> 0.35$), in which we prohibited gaps. We then calculated the mutational distance $m(s_i, s_j)$ between two sequences $s_i$ and $s_j$ as the number of mismatches in the alignment of these sequences. We represented bound sequences as vertices in the network, and we connected two vertices by an edge if their corresponding sequences differed by a single small mutation. We considered point mutations and small indels that shift an entire contiguous binding site by a single nucleotide. We chose one of them for later analysis. We chose to assign only one binding site per footprint per TF or RBP to avoid any bias that may be introduced by the different lengths of DNaseI footprints and RNase footprints.

The above procedure resulted in a list of genomic coordinates of potential binding sites for each TF and RBP. The number of such TF binding sites ranged from 266 for the TF FOXP6 in HFF cells to a maximum of 50,126 for the TF DNMT1 in NB4 cells. The number of such RBP binding sites ranged from 0 for the RBPs ENOX1, FXR2, and ZNF638 to a maximum of 32,272 for the RBP HNRNPH2. For each of these potential binding sites, we used the Samtools function tabix to query the 1000 Genomes Consortium data for SNPs (92). We used the results of these queries to calculate the proportion of binding site variants that is on the focal TF’s or RBPs genotype network and the proportion that is not. For those variants that are not on the genotype network, we calculated the fraction of TFs or RBPs in our dataset that these variants have the potential to bind.

**Statistical Analysis.** We use the nonparametric, one-sided Wilcoxon rank sum test of the null hypothesis that the distributions of samples $x$ and $y$ come from distributions with equal medians against the alternative hypothesis that $y$ is sampled from a distribution with a median that is greater than that of the distribution from which $x$ was sampled. We do not distinguish between $P$ values that are less than $10^{-3}$. 

**Fig. 6.** Analysis of human DNA polymorphisms shows that the binding sites of TFs and RBPs exhibit similar robustness to mutation, yet those of TFs are more likely to bring forth new binding phenotypes. (A) Robustness (vertical axis), defined as the fraction of a TF’s or RBP’s binding site variants that remain on the focal protein’s genotype network, shown in relation to the fractional size of the genotype network (horizontal axis). Binding site variants are from phase 3 of the 1000 Genomes Project Consortium (61). Black symbols correspond to the robustness and evolvability of binding sites in DNaseI footprints that overlap promoters averaged across 41 cell and tissue types. White symbols correspond to the robustness and evolvability of binding sites in RNase footprints that overlap the 5′ and 3′ UTRs of transcripts in HeLa cells.

**B** Evolvability (vertical axis), defined as the fraction of TFs or RBPs in our dataset that are bound by the binding site variants that do not remain on the focal protein’s genotype network, shown in relation to the fractional size of the genotype network (horizontal axis). Binding site variants are from phase 3 of the 1000 Genomes Project Consortium (61). Black symbols correspond to the robustness and evolvability of binding sites in DNaseI footprints that overlap promoters averaged across 41 cell and tissue types. White symbols correspond to the robustness and evolvability of binding sites in RNase footprints that overlap the 5′ and 3′ UTRs of transcripts in HeLa cells.
ACKNOWLEDGMENTS. We thank Mihai Albu, Alejandro V. Cano, Paco Majic, and Matt Weiach for discussions as well as three anonymous reviewers for their critical reading and feedback. J.L.P. acknowledges support from Swiss National Science Foundation Grants PZ00P3_154773 and PP00P3_170604.

A.W. acknowledges support from European Research Council Advanced Grant 739874, Swiss National Science Foundation Grant 31003A_172887, and the University Priority Research Program in Evolutionary Biology at the University of Zurich.