Both Binding Strength and Evolutionary Accessibility Affect the Population Frequency of Transcription Factor Binding Sequences in *Arabidopsis thaliana*

Gabriel Schweizer [1,2] and Andreas Wagner [1,2,3,4,*]

[1]Department of Evolutionary Biology and Environmental Studies, University of Zürich, Switzerland

[2]Swiss Institute of Bioinformatics, Quartier Sorge-Batiment Genopode, Lausanne, Switzerland

[3]Santa Fe Institute, Santa Fe, New Mexico, USA

[4]Stellenbosch Institute for Advanced Study (STIAS), Wallenberg Research Centre at Stellenbosch University, South Africa

*Corresponding author: E-mail: andreas.wagner@ieu.uzh.ch

Accepted: 6 December 2021

Abstract

Mutations in DNA sequences that bind transcription factors and thus modulate gene expression are a source of adaptive variation in gene expression. To understand how transcription factor binding sequences evolve in natural populations of the thale cress *Arabidopsis thaliana*, we integrated genomic polymorphism data for loci bound by transcription factors with in vitro data on binding affinity for these transcription factors. Specifically, we studied 19 different transcription factors, and the allele frequencies of 8,333 genomic loci bound in vivo by these transcription factors in 1,135 *A. thaliana* accessions. We find that transcription factor binding sequences show very low genetic diversity, suggesting that they are subject to purifying selection. High frequency alleles of such binding sequences tend to bind transcription factors strongly. Conversely, alleles that are absent from the population tend to bind them weakly. In addition, alleles with high frequencies also tend to be the endpoints of many accessible evolutionary paths leading to these alleles. We show that both high affinity and high evolutionary accessibility contribute to high allele frequency for at least some transcription factors. Although binding sequences with stronger affinity are more frequent, we did not find them to be associated with higher gene expression levels. Epistatic interactions among individual mutations that alter binding affinity are pervasive and can help explain variation in accessibility among binding sequences. In summary, combining in vitro binding affinity data with in vivo binding sequence data can help understand the forces that affect the evolution of transcription factor binding sequences in natural populations.

Key words: transcription factor, binding affinity, evolutionary path accessibility, epistasis, *Arabidopsis thaliana*.

Significance

Frequent genotypes in a population may have two evolutionary origins. They may be high fitness genotypes, and thus directly favored by natural selection. Alternatively, they may be easily accessible by mutational paths through genotype space. Here we show that both origins play a role in the evolution of DNA sequences bound by transcription factors in the thale cress *Arabidopsis thaliana*. Our work shows that high fitness is not always the only explanation for the prevalence of frequent genotypes.

Introduction

Regulated gene expression is crucial to shape and maintain organismal phenotypes (Wray 2007; Wittkopp and Kalay 2011; Rice and Rebeiz 2019). Such regulation is controlled by multiple processes, including the binding of a transcription factor to specific loci in a genome (Coulon et al. 2013). Mutations of the DNA sequences at such a locus can cause variation in gene regulation and contribute to phenotypic

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

diversity within and between species (Romero et al. 2012; Signor and Nuzhdin 2018). An especially important kind of mutation alters the DNA sequence at an individual genomic locus bound by a transcription factor, and changes the sequence's affinity to the transcription factor (Wray 2007; Kwasnieski et al. 2012; Ichihashi et al. 2014). Such regulatory variation can be an important raw material for adaptive evolution (Wray 2007; Romero et al. 2012). Patterns of selection in genomic loci bound by transcription factors have been studied within and between several species, including humans (Torgerson et al. 2009; Mu et al. 2011; Vernot et al. 2012; Naidoo et al. 2018), *Drosophila* (He et al. 2011), and *Saccharomyces cerevisiae* (Connelly et al. 2013).

How regulatory differences contribute to adaptive evolution has been studied in several plant species, including the thale cress *Arabidopsis thaliana* (West et al. 2007; Zhang et al. 2011; Ichihashi et al. 2014; Lasky et al. 2014). Existing studies in *A. thaliana* illuminate the genomic architecture of transcriptional control (Verma 2019; Nakamichi 2020). They also demonstrate how rewiring a gene regulatory network can affect phenotypes (Lv et al. 2014; Jiang et al. 2016). However, they do not help us understand how mutations in regulatory sequences affect the binding of a transcription factor to DNA, and how the strength of such binding affects sequence evolution.

A protein-binding microarray (PBM) is a powerful technology to understand how DNA mutations alter the affinity of regulatory DNA for a transcription factor (Berger et al. 2006; Berger and Bulyk 2009). Such an array typically measures how strongly each oligonucleotide with a length of ten base pairs binds to a transcription factor in vitro (Berger et al. 2006). PBM experiments were previously used to study affinity landscapes of transcription factor binding sequences (Payne and Wagner 2014; Aguilar-Rodríguez et al. 2017; Cano and Payne 2020). Here we use them to relate the binding affinities of specific transcription factor binding sequences to the allele frequencies of these sequences in natural *Arabidopsis* populations.

At least two characteristics of a transcription factor binding sequence may explain its frequency in a population. The first is the affinity of the sequence to its cognate transcription factor, which can be estimated with PBMs. Several studies suggest that high affinity may be favored by natural selection. For example, a study that integrated human whole genome sequences with genome-wide chromatin immunoprecipitation and sequencing data found that DNA sequences with strong affinity to a transcription factors experience strong selection to maintain this DNA sequence (Arbiza et al. 2013). Unrelated work shows that the binding affinity of regulatory DNA to transcription factors is subject to adaptive evolution in 29 human tissues. Especially strong positive selection on this affinity exists in the brain, suggesting that adaptive changes in gene regulation contributed to human brain evolution (Liu and Robinson-Rechavi 2020). In addition, the affinities of at least some binding sequences to transcription factors are subject to positive selection in other mammalian species (Molineris et al. 2011). Such preference for strong binding may not be universal, however, because some eukaryotic transcription factors require low affinity binding sites to function correctly (Ramos and Barolo 2013; Crocker et al. 2015; Delker et al. 2019).

A second characteristic that may help explain a binding sequence's frequency is that it may be easily "accessible" by Darwinian evolution. Multiple evolutionary paths of successive single mutations typically lead from an ancestral genotype to any one beneficial genotype in an extant population. In only a fraction of these paths may each individual mutational step be favored by natural selection. This fraction can be used as a proxy for the evolutionary accessibility of a genotype, and it may differ among beneficial genotypes (Weinreich et al. 2006; Poelwijk et al. 2007). In other words, even high fitness genotypes may not be easily "findable" (McCandlish 2013; Schaper and Louis 2014). For the paths leading to a transcription factor binding sequence with a given affinity to its cognate transcription factor, PBMs can help quantify accessibility of transcription factor binding sequences, because they link each short DNA sequence with an affinity value.

Here, we integrate multiple sources of functional genomic data with PBM data to assess the relative importance of binding affinity and evolutionary accessibility to help explain the frequency of transcription factor binding sequences in *A. thaliana*. Specifically, we ask if the frequency of different binding sequences can be explained by their affinity, their accessibility, or both. To this end, we first identify 8,333 genomic *A. thaliana* loci that are specifically bound in vivo by at least one of 19 transcription factors. Using genomic polymorphism data from a collection of 1,135 *A. thaliana* accessions, we find that variation at bound loci is under purifying selection, and that high frequency alleles (binding sequences) at bound loci tend to have high DNA binding affinity. Epistatic (nonadditive) interactions among mutations that affect binding affinity are frequent, which can help explain variation in evolutionary accessibility among binding sequences. Finally, we show that both binding affinity and accessibility can help explain why some alleles have higher frequencies than others do.

Results

Transcription Factor Binding Sequences Are Less Diverse Than Random Genomic Sequences

In a first analysis, we wanted to characterize the genomic diversity of genomic loci bound by transcription factors. Exceptionally low diversity may indicate purifying selection, whereas high diversity may indicate diversifying selection. To study genomic diversity, we first identified 8,333 loci bound

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

in vivo by one of 19 transcription factors in the reference genome of the *A. thaliana* accession Col-0 (supplementary tables 1 and 2, Supplementary Material online; fig. 1*A* and *B*). To identify bound loci, we combined results from in vivo DNase I footprint experiments with data from in vitro DAPseq studies (fig. 1*B*). Because binding affinity measurements are only available for transcription factor binding sequences that are at most eight base pairs long, we restricted our analysis to bound loci spanning eight genomic positions. We note that most eukaryotic transcription factor binding sequences span no more than ten base pairs (Stewart et al. 2012). We considered a nucleotide sequence as specifically bound if its affinity value (*E*-score) exceeded 0.35 in two replicate experiments, because this cutoff had been shown to be associated with a false binding discovery rate smaller than 0.001 (Badis et al. 2009; Zhu et al. 2009).

The number of bound loci varied dramatically among transcription factors. It ranged from 29 bound loci for the AP2-EREBP transcription factor CRF4 (AT4G27950), to 1,864 bound loci for the CPP transcription factor SOL1 (AT3G22760) (supplementary table 2, Supplementary Material online). We then obtained, for each bound locus of each transcription factor, orthologous binding sequences from 1,134 additional *A. thaliana* accessions. Out of the 8,333 binding loci of all transcription factors, we excluded 1,096 bound loci from further analysis, because all 1,134 orthologous binding sequences at these loci contained ambiguous nucleotides or indels. This prevents us from linking binding sequences with their in vitro binding affinity, that is, with the *E*-score from PBM experiments. The remaining 7,237 bound loci contained between one and 1,097 binding sequences without indels or ambiguous nucleotides.

Each bound locus could in principle contain 1,135 different binding sequences, because we studied 1,135 accessions. However, we found many fewer such binding sequences, which hinted at a low genetic diversity of transcription factor binding sequences. In total, 53.7% (3,888) of bound loci were completely monomorphic, that is, they harbored only a single binding sequence (fig. 2*A*; supplementary table 2, Supplementary Material online). The proportion of monomorphic bound loci varied among transcription factors. It ranged between 37.9% for the transcription factor CRF4, and 61.7% for the WRKY transcription factor WRKY75 (AT5G13080; supplementary table 2, Supplementary Material online). In addition, no bound locus harbored more than 11 out of the maximally possible 1,135 binding sequences (fig. 2*A*; supplementary table 2, Supplementary Material online). This maximum holds for a locus binding the TCR/CxC transcription factor TCX3 (AT3G22760). More generally, bound loci harbored on average only between 1.501 different binding sequences for the transcription factor WRKY75, and 1.919 different binding sequences for the AP2-EREBP transcription factor RAP2.6 (AT1G43160; supplementary table 2, Supplementary Material online). We emphasize that the low

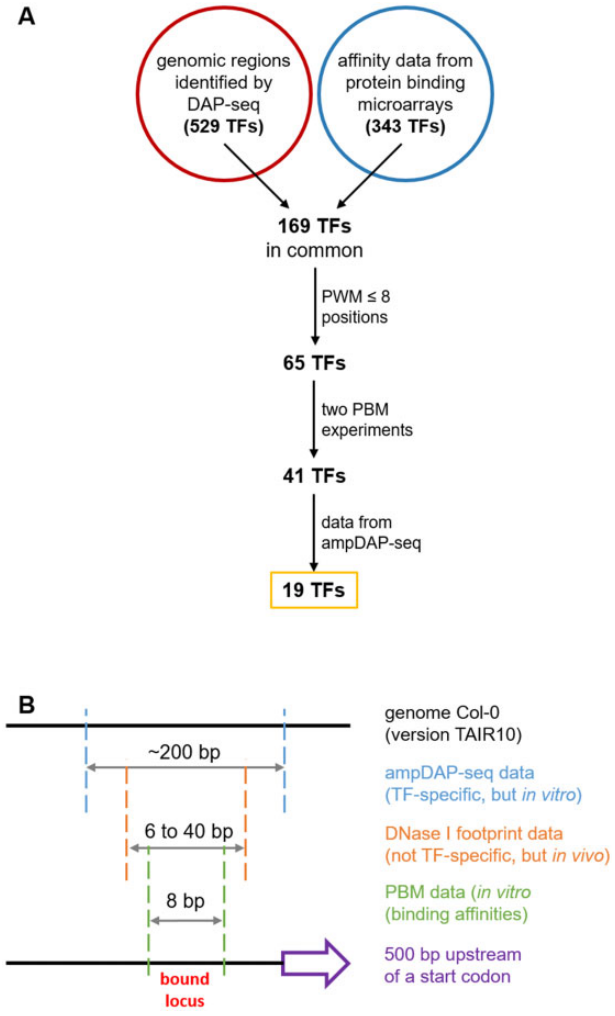

Fig. 1.—Identification of transcription factors and bound loci to study the evolution of transcription factor binding sequences in *A. thaliana*. (*A*) We used publicly available data of binding regions that were identified by DNA affinity purification and sequencing (red circle) and binding affinities that were measured using PBMs (blue circle). Both data sets had 169 transcription factors (TFs) in common. After filtering for transcription factors that had only up to eight informative positions, binding affinity data from two replicate experiments, and genomic data from DNA amplification and sequencing (ampDAP-seq) experiments, we retained 19 transcription factors for our analysis (yellow box). (*B*) To identify in vivo bound genomic loci in the *A. thaliana* accession Col-0, we first identified all genomic regions of length 200 bp identified as bound by a transcription factor in ampDAP-seq experiments (blue dashed lines), because results reported from such experiments equal approximately 200 bp. Second, within regions covered by ampDAP-seq data, we identified regions of 6–40 base pairs that were covered by in vivo DNase I hypersensitivity experiments (orange dashed lines). Third, within regions that showed such a DNAse I footprint, we identified loci of length eight (green dashed lines) that can be bound by a specific transcription factor (*E*-score exceeding 0.35 in a PBM experiment). Fourth, we retained only loci that are located within 500 bp upstream of a gene's start codon (purple arrow) or within the entire intergenic region if this region was shorter than 500 bp. We refer to such loci as bound loci (red label).

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

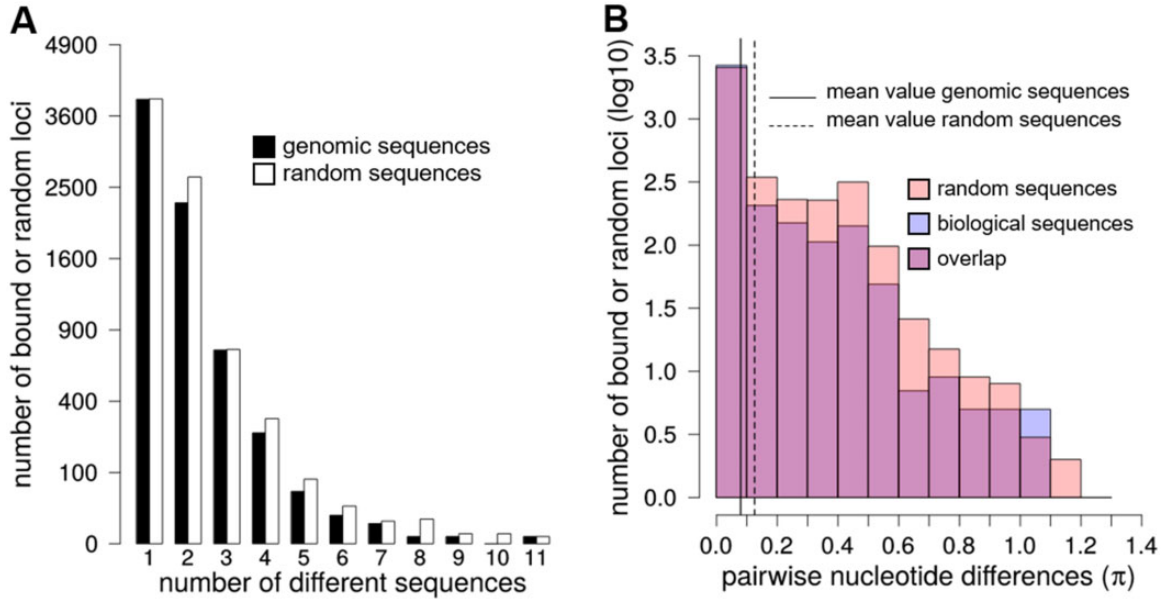

Fig. 2.—Nucleotide diversity in biologically bound loci and random genomic sequences that are under weak or no selection (see Materials and Methods). (*A*) We detected up to 11 different biological sequences at bound loci (black bars) as well as at random genomic sequences per locus (white bars, see Materials and Methods). The vertical axis shows the number of loci on a square root scale, because this scale is better suited than a logarithmic scale for binned data with a highly skewed distribution, where some bins contain zero data points (like that of bound loci with ten binding sequences in our data). In the random sequence data, a binding sequence corresponds to a concatenation of eight nucleotides from genomic positions chosen at random from all third positions of 4-fold degenerated codons. (*B*) Pairwise nucleotide differences (π) between all sequences at a bound locus. The histogram shows the distribution of π in biological and random sequences in different colors, as indicated by the legend (salmon, random sequences; lavender, biological sequences). The solid black line indicates the mean value of π for biological sequences, and the dashed black line represents the mean value of π for random sequences.

number of different binding sequences at a bound locus does not result from our exclusion of sequences with indels or ambiguous nucleotides (supplementary text 1, Supplementary Material online).

We hypothesized that this low diversity might result from purifying selection that maintains the function of transcription factor binding sequences (Heyndrickx et al. 2014). To test this hypothesis, we first asked how many different binding sites one would expect per locus, based on the genome-wide average nucleotide diversity of 0.5% (The 1001 Genomes Consortium 2016). The expected number of unique alleles at each bound locus is expected to follow a binomial *B(n, p)* distribution whose parameters we estimated from our data. Specifically, the probability for finding eight identical nucleotides between two orthologous binding sequences equals $(1-0.005)^8$ because the average genomic nucleotide diversity equals 0.5% and the binding sequences we study comprise no more than eight nucleotides (see Materials and Methods). Thus, the probability *P* of finding two different binding sequences equals $1-(1-0.005)^8 = 0.0393$. The number of trials (*n*) in this binomial distribution equals the number of accessions (1,135). Thus, we modeled the number of different binding sequences at each bound locus with the binomial distribution *B*(1135, 0.0393). For each bound locus, we then identified the number *w* of different alleles at this locus, and calculated the probability of finding *w* or fewer alleles according to this binomial distribution. We repeated this procedure for all 7,237 bound loci, and found (after Bonferroni-correction for multiple testing) a significantly lower number of alleles than expected by chance at each locus (*P* values $\leq 1.15 \times 10^{-9}$).

In a further, complementary test of this hypothesis, we compared the allelic diversity of bound loci with those of random genomic sequences of equal length (eight nucleotides) that are likely to be under weak or no selection. We obtained these random sequences by concatenating third positions of 4-fold degenerated codons in protein-coding regions throughout the genome of accession Col-0, and then obtaining genotype of orthologous nucleotides for all other 1,134 *A. thaliana* accessions (see Materials and Methods). We assembled for each bound locus of each transcription factor a set of such random genomic sequences. As in the data of bound loci, we found up to 11 different alleles in the random sequences for a bound locus (fig. 2*A*; supplementary table 3, Supplementary Material online). However, the average number of different alleles in the random sequences (1.739) was significantly higher than at the bound loci (1.672 alleles; *P* value $= 8.989 \times 10^{-5}$, Wilcoxon rank-sum test).

To identify further differences between bound loci and random genomic sequences, we also computed the average

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

number π of pairwise nucleotide differences between all alleles at a locus (Nei and Li 1979). We restricted this analysis to the 3,347 bound loci and 3,832 random loci with at least two different sequences per locus (supplementary tables 2 and 3, Supplementary Material online, respectively), and calculated one value of π per locus. Again, random genomic sequences were significantly more diverse than bound loci (fig. 2*B*; mean $\pi_{random} = 0.1259$ vs $\pi_{biological} = 0.0795$; *P* value $< 2.2 \times 10^{-16}$, Wilcoxon Rank-Sum Test). In sum, our analysis demonstrates a significantly low diversity of loci bound by transcription factors. This observation supports the hypothesis that purifying selection acts on bound loci.

Minor Alleles at Bound Loci Are Geographically Restricted

We next aimed to identify basic distinguishing features between high and low frequency alleles (binding sequences) at bound loci. To this end, we calculated for each polymorphic bound locus the frequency of all alleles (supplementary table 2, Supplementary Material online). To allow us to compare results among loci, we based these calculations on 1,135 possible binding sequences at each bound locus, even if some sequences contained indels or ambiguous nucleotides, and were thus excluded from our analysis. We found that the distribution of binding sequence frequencies is highly skewed toward one major allele, that is, the binding sequence with the highest frequency. Specifically, for 91.9% of bound loci (3,076 of 3,347), the major allele had a frequency of at least 50%, and for 64.3% of bound loci (2,153 of 3,347) this frequency exceeded 80% (fig. 3*A*; supplementary table 4, Supplementary Material online).

Previous work identified ten admixture groups in the collection of 1,135 *A. thaliana* accessions, which are explained by the geographic origin of the accessions (The 1001 Genomes Consortium 2016). We identified the admixture group to that each allele at each bound locus belongs. Although the major allele occurred in all ten admixture groups for 98.4% of bound loci (3,292 of 3,347) (fig. 3*B*; supplementary table 4, Supplementary Material online), the minor allele(s), that is, the allele(s) with the lowest frequency occurred in only one admixture group for 56.8% of bound loci (1,900 of 3,347 loci; fig. 3*C*). For 1,432 of the 1,900 loci the minor allele occurred only in one accession. Conversely, we found that different admixture groups are associated with different numbers of loci and minor alleles. For example, we found that minor alleles at a maximum of 495 among 1,900 loci belong to the admixture group "Italy–Balkan–Caucasus," whereas the minor alleles of a minimum of 38 loci are part of the admixture group "North Sweden" (inset of fig. 3*C*; supplementary table 4, Supplementary Material online).

To compare the observed distribution of sequences in each admixture group with a random expectation, we used multinomial sampling with a sample size of 1,900 (the number of bound loci), ten bins (the number of admixture groups), and a probability of belonging to one bin that equals the fraction of accessions in each admixture group (Admixed, 137/1135; Asia, 79/1135; Central Europe, 184/1135; Germany, 171/1135; Italy–Balkan–Caucasus, 92/1135; North Sweden, 64/1135; Relict, 25/1135; South Sweden, 156/1135; Spain, 110/1135; Western Europe, 117/1135). We repeated this random sampling 10,000 times, and found that the observed distribution of admixture groups differs significantly from a random expectation (*P* value $< 2.2 \times 10^{-16}$; χ^2 test). In summary, minor alleles have specific geographic origins. We speculate that they may be linked to environmental conditions at these origins.

High Frequency Binding Sequences Tend to Have High Affinity to Their Cognate Transcription Factor

Mutations in bound loci can help fine-tune binding affinities and gene expression levels (Sharon et al. 2012; Inukai et al. 2017; Rastogi et al. 2018). Models of regulatory evolution commonly assume a direct link between binding affinity and gene expression, that is, high affinity implies high expression (Gao and Stock 2015; Grassi et al. 2015). Indeed, a recent study with the yeast transcription factors GCN4 and FHL1 showed that high in vitro binding affinities entail high in vivo gene expression (Sharon et al. 2012; Aguilar-Rodríguez et al. 2017). Other work shows that mutations reducing binding affinities are under negative selection, and that mutations increasing binding affinities are positively selected (Mustonen and Lässig 2009; Arbiza et al. 2013). Conversely, some studies show that strong binding affinity is not always linked to high gene expression and that strong binding affinity can be deleterious (Ramos and Barolo 2013; Crocker et al. 2015; Delker et al. 2019).

Motivated by such conflicting reports, we aimed to investigate if binding affinity may be subject to selection in natural accessions of *A. thaliana*. To this end, we asked whether transcription factor binding sequences (alleles) with high affinity, that is, large *E*-scores in a PBM experiment, tend to have high frequencies in our study population. This is indeed the case. Remarkably, it holds for each transcription factor (after Bonferroni-correction for multiple testing; Kendall's tau between 0.215 and 0.527, *P* values between 9.941×10^{-57} and 1.794×10^{-6}, sample size between 42 and 1,948; supplementary table 5, Supplementary Material online). To further investigate a link between binding affinities and frequencies of binding sequences, we ranked all eight-mers present on a PBM according to their binding affinity, such that the most favored sequence has rank one and the most disfavored sequences has rank 32,896. Across all transcription factors, sequences at bound loci had at least rank 160, that is, all such sequences occur in the top 0.486% of favored sequences. Both results suggest that stronger binding affinities are favored by natural selection in *A. thaliana*. In a

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

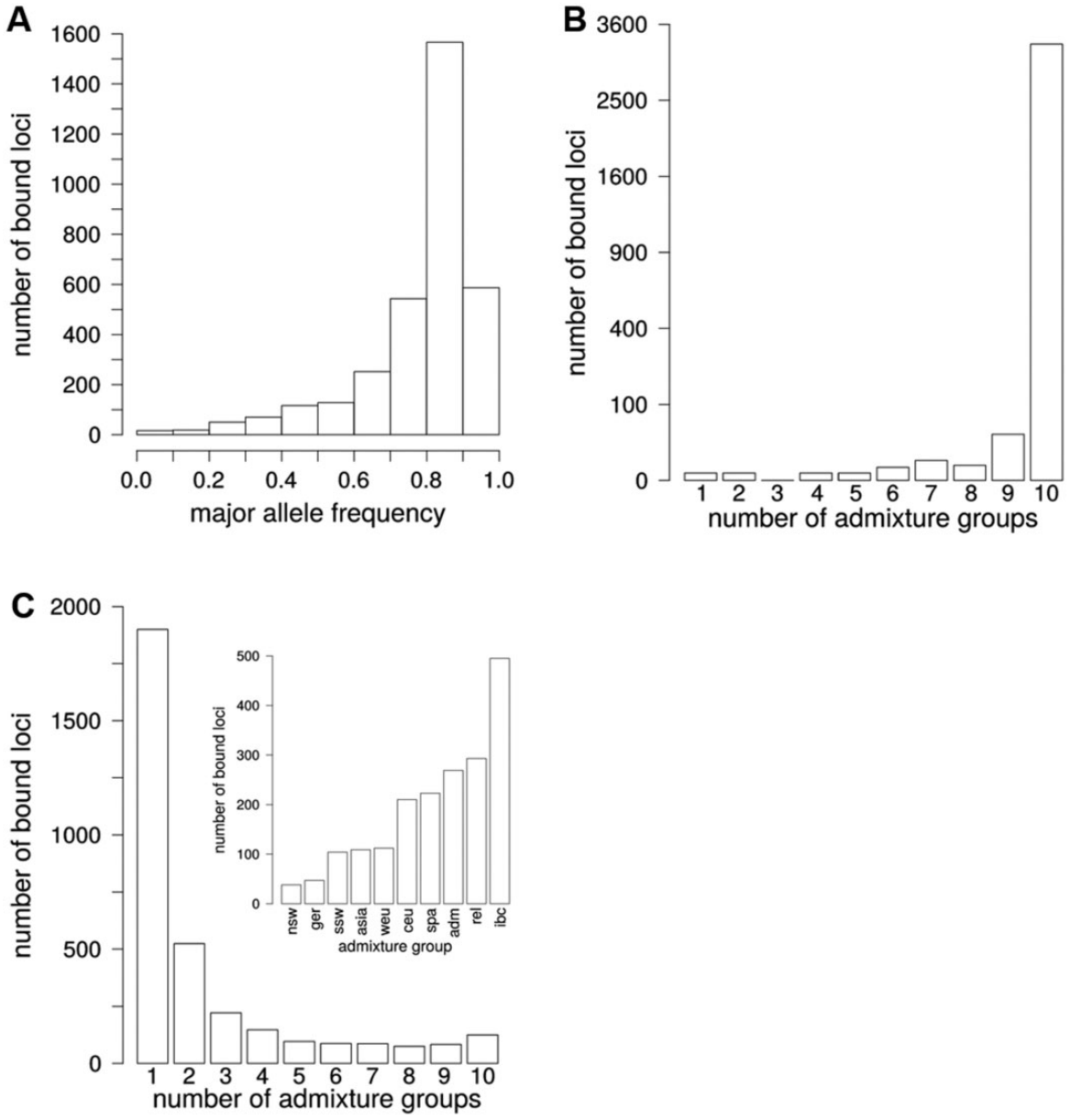

Fig. 3.—Most minor alleles belong to one admixture group. (*A*) Histogram of major allele frequencies. The horizontal axis shows the frequency of the major allele, that is, the allele with the largest frequency among all alleles, for $n = 3{,}347$ polymorphic bound loci. The vertical axis shows the number of bound loci with the respective allele frequency. (*B*) The number of admixture groups in which a major allele is found (horizontal axis) is plotted against the number of polymorphic loci bound by a transcription factor (vertical axis). For most bound loci, the major allele was found in all ten previously defined admixture groups. The data are shown on a square root scale, which is better suited than a logarithmic scale for categorical data with a highly skewed distribution, where some categories contain zero data points (no major alleles are found in three admixture groups). (*C*) The number of admixture groups in which a minor allele is found (horizontal axis) is plotted against the number of polymorphic bound loci (vertical axis). The minor alleles of most bound loci belong to one admixture group. The inset shows to which admixture groups minor alleles that occur in only one admixture group belong (horizontal axis). Admixture groups were previously defined (The 1001 Genomes Consortium 2016) and are abbreviated as follows: nsw, North Sweden; ger, Germany; ssw, South Sweden; weu, Western Europe; asia, Asia; ceu, Central Europe; spa, Spain; adm, Admixed; rel, Relict; ibc, Italy–Balkan–Caucasus. Most minor alleles that belong to one admixture group are found in the Italy–Balkan–Caucasus group.

separate analysis, however, we also showed that frequent binding sequences with strong binding affinities are not necessarily associated with high gene expression levels (supplementary text 2, Supplementary Material online), which is consistent with previous reports (Ramos and Barolo 2013; Crocker et al. 2015; Delker et al. 2019). Thus, whereas selection tends to favor high affinity for the bound loci we examined, it may not necessarily favor high expression.

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

Frequent Binding Sequences Tend to Have Many Neighbors with Lower Binding Affinity

In models of adaptive landscapes, evolution is described as a hill-climbing process, where genomic sequences that occupy peaks are favored by natural selection (Kauffman and Levin 1987). We wanted to find out whether many naturally occurring binding sequences constitute a local peak in the affinity landscape of each of our 19 transcription factors. To this end, we determined, for each unique binding sequence S at each bound locus, all $8 \times 3 = 24$ one-mutant neighbors. We then recorded the fraction of neighbors with lower binding affinity (E-score) than S itself, and refer to this quantity as the "peakness" of S (see Materials and Methods). If S has a peakness of one, it is a local peak in the affinity landscape. We found that between 18.5% and 41.8% of binding sequences are local peaks, depending on the transcription factor (supplementary table 6, Supplementary Material online). Thus, only a minority of binding sequences are local peaks.

Although few frequent binding sequences are local peaks, they may still exist near such peaks. We used peakness to help us quantify this proximity, reasoning that a binding sequence close to a peak may have many neighboring binding sites with lower affinity. Indeed, for all 19 analyzed transcription factors, frequent binding sequences have high peakness (after Bonferroni-correction for multiple testing; Kendall's tau between 0.209 and 0.540, P values 1.858×10^{-54} and 3.414×10^{-6}, sample size between 42 and 1,948; supplementary table 5, Supplementary Material online). This observation is consistent with the view that adaptive evolution drives binding sequences toward local affinity peaks in natural populations of *A. thaliana*.

Frequent Binding Sequences Are Accessible through Many Mutational Paths

Each binding sequence at a bound locus in the 1,135 *A. thaliana* accessions could in principle have been created through a path of successive single DNA mutations starting from any sequence in genotype space—the collection of all possible binding sequences (Payne and Wagner 2014; Aguilar-Rodríguez et al. 2017). We call such a mutational path evolutionarily accessible if the binding affinities (E-scores) of the sequences along this path increase monotonically. In other words, we interpret each mutation that increases affinity as an adaptive mutation, because of our observation that binding sequences with high frequencies have high binding affinities.

It is possible that at least some high frequency transcription factor binding sequences in our study population exist not just because of their high affinity, but also because they are easily accessible, that is, through multiple mutational paths. To test this hypothesis, we first determined the fraction of accessible paths for all sequences that differ by one, two, three, or four mutations from each binding sequence at each bound locus. We further distinguished between mutational paths with monotonically and strictly monotonically increasing binding affinities (see Materials and Methods). We here describe our observations for monotonically increasing binding affinities, and note that analogous observations hold for strictly monotonically increasing affinities (supplementary fig. 1, Supplementary Material online). Not unexpectedly, the fraction of accessible mutational paths decreases with increasing path length (fig. 4; supplementary table 7, Supplementary Material online; Kendall's tau $= -0.6325$, P value $< 2.2 \times 10^{-16}$, sample size $= 32{,}820$). Five binding sequences of four transcription factors at four bound loci had no accessible path leading to them, and no binding sequence was accessible by all mutational paths (supplementary table 7, Supplementary Material online). On average across all path lengths, genomic binding sequences for the AP2-EREBP transcription factor CBF2 (AT4G25470) showed the lowest fraction (0.1925) of accessible paths. Conversely, binding sequences for the CPP transcription factor AT2G20110 showed the highest fraction (0.4297) of accessible paths (supplementary table 7, Supplementary Material online). For all 19 analyzed transcription factors, and across all mutational path lengths, we observed after Bonferroni-correction for multiple testing that more frequent binding sequences are accessible through a greater number of mutational paths (supplementary table 5, Supplementary

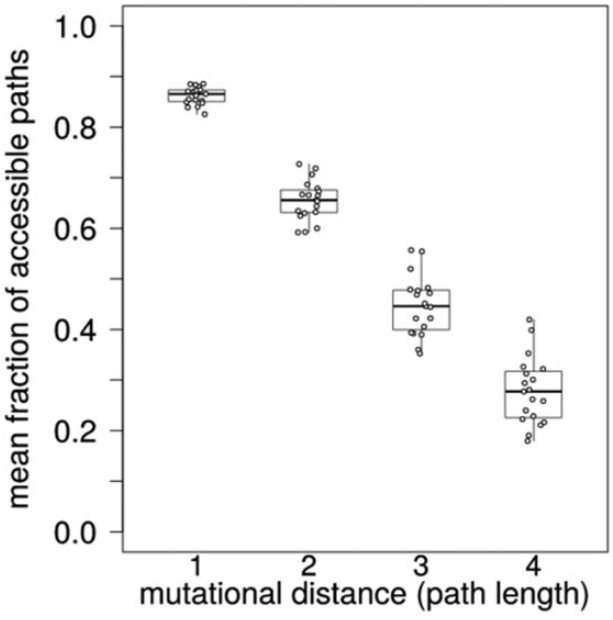

Fig. 4.—Fraction of accessible mutational paths. We calculated the fraction of paths with monotonically increasing binding affinities (vertical axis) for path lengths of one, two, three, and four mutational steps (horizontal axis). Results are represented as a box plot. The central bold horizontal line indicates the median value, and the lower and upper box limits represent the first and third quartile, respectively. Whiskers indicate values within the 1.5-fold interquartile range, and open circles show data points for individual transcription factors ($n = 19$ in each box plot).

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

Material online; Kendall's Tau between 0.2017 and 0.5174, *P* values between 1.436×10^{-58} and 9.253×10^{-6}, sample sizes between 42 and 1,948).

We next wanted to disentangle the relative contributions of binding affinity and accessibility to high frequency binding sequences. Both factors may play a role in the evolution of anyone binding sequence, but one of them may be more important than the other. Disentangling their relative contributions is not trivial, because they are correlated with each other (supplementary table 5, Supplementary Material online; Kendall's tau between 0.5751 and 0.8603 depending on the transcription factor, *P* values between $<2.2 \times 10^{-16}$ and 1.140×10^{-15}, sample size between 42 and 1,948). We thus used a partial correlation analysis to determine the statistical association between frequency and affinity while controlling for accessibility, or vice versa. We performed this analysis separately for each transcription factor (supplementary table 8, Supplementary Material online).

For eight transcription factors, the frequency of binding sequences remained associated with affinity after controlling for accessibility, and vice versa. In these transcription factors, both affinity and accessibility contribute to the observed frequency of binding sequences. The transcription factor Dof3.2 (AT3G45610) had the highest association between affinity and frequency after controlling for accessibility (supplementary table 8, Supplementary Material online; Kendall's partial tau = 0.3026, *P* value = 1.632×10^{-22}, sample size = 467; Bonferroni-correction for multiple testing). Conversely, the WRKY transcription factor WRKY25 (AT2G30250) had the highest association between accessibility and frequency after controlling for affinity (supplementary table 8, Supplementary Material online; Kendall's partial tau = 0.1723, *P* value = 7.375×10^{-15}, sample size = 910; Bonferroni-correction for multiple testing). For six out of 19 transcription factors, binding sequence frequency did not remain associated with accessibility when controlling for affinity (supplementary table 8, Supplementary Material online). In these transcription factors, affinity but not accessibility can help explain binding sequence frequency. Conversely, for the AP2 transcription factor ERF4 (AT3G15210) binding sequence frequency did not remain associated with affinity after controlling for accessibility (supplementary table 8, Supplementary Material online; Kendall's partial tau = 0.0714; *P* value = 0.090, sample size = 255). In this transcription factor accessibility but not affinity can help explain the frequencies of binding sequences. For four transcription factors, sequences with high frequencies do not remain associated with affinity or evolvability after controlling the other quantity (supplementary table 8, Supplementary Material online). In sum, depending on the transcription factor, both affinity and accessibility contribute to the evolution of transcription factor binding sequences. However, based on the above numbers' affinity is more important for the majority of transcription factors.

Pervasive Epistasis in Transcription Factor Binding Sequences

Similar to what has been found in previous work (Payne and Wagner 2014; Aguilar-Rodríguez et al. 2017), inaccessible mutational paths to binding sequences at bound loci of our study populations are frequent in our data (fig. 4). One factor that can reduce the accessibility of mutational paths is epistasis (Weinreich et al. 2005; Poelwijk et al. 2007; Kvitek and Sherlock 2011). Thus, we sought to quantify the prevalence, type, and strength of epistasis in the binding of transcription factors to the DNA sequences we studied. Specifically, we analyzed how mutations interact in their effect on binding affinity for a focal sequence, two of its one-mutant neighbors, and the corresponding two-mutant neighbor. For ease of reference, we designated the wild-type sequence as ab, the two single-mutants as Ab and aB, and the double-mutant as AB (fig. 5*A*). We chose sequence AB such that it shows the strongest binding affinity among all four sequences. This choice reflects the assumption that binding sequences evolve toward higher affinity, which is supported by our observation that more frequent binding sequences show stronger affinity.

Like other authors, we distinguish two main types of epistasis (Poelwijk et al. 2007). The first is magnitude epistasis, where the binding affinity of sequence AB is higher (positive magnitude epistasis) or lower (negative magnitude epistasis) than the sum of the binding affinities of sequences aB and Ab. The second is sign epistasis, where the binding affinity of one sequence (aB or Ab—simple sign epistasis) or the affinity of both sequences (reciprocal sign epistasis) is lower than the affinity of sequence ab (fig. 5*A*). To identify epistasis, we built genotype networks, that is, graphs in which nodes represent binding sequences of a bound locus, and in which edges connect two binding sequences if they differ by a single mutation. We scanned each genotype network for the presence of "squares," that is, cycles of length four that connect a wild-type sequence to a two-mutant neighbor of the sequence (Aguilar-Rodríguez et al. 2017). Because only few bound loci (323 out of 7,237; supplementary table 2, Supplementary Material online) have at least four different binding sequences, the number of complete squares in our data are small. For example, the bound loci of eight transcription factors did not contain any complete squares (supplementary table 9, Supplementary Material online). Moreover, we identified only up to five complete squares for the remaining 11 transcription factors (supplementary table 9, Supplementary Material online). However, there are two more motifs in genotype networks from which squares can be unambiguously inferred, such that epistatic interactions can be studied for them. First, if three binding sequences are connected but do not form a cycle, the missing fourth binding sequence can be inferred to complete the square. Second, if two observed binding sequences alleles differ by two mutations, the missing two single-mutant neighbors can also be unambiguously inferred. To investigate the incidence of epistasis, we pooled all mutation pairs

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

(squares), regardless of how many binding sequences have to be inferred (zero, one, or two). In this way, we identified 1,891 mutation pairs (supplementary table 9, Supplementary Material online). We excluded 40 of these pairs from further analysis, because none of the four sequences met our minimal threshold for specific binding to their cognate transcription factor (*E*-score $>$ 0.35, see Materials and Methods, supplementary table 10, Supplementary Material online). The resulting data contained 19 mutant squares for which all four sequences existed in the genomic data, as well as 1,708 and 124 squares where one and two sequences had to be inferred, respectively (supplementary table 10, Supplementary Material online). The total number of squares ranged between six squares for the transcription factor CRF4 and 463 squares for the TCR-CxC transcription factor TCX3 (AT3G22760) (supplementary table 10, Supplementary Material online).

We found that 390 of 1,851 mutation pairs (21.1%) showed only additive interactions. Positive magnitude epistasis was the most common type of epistasis in 18 out of the 19 analyzed transcription factors. An average fraction of 40.7% of mutation pairs showed this type of epistasis (fig. 5*B*). The transcription factor DREB1C (AT4G25470) from the AP2 family showed an especially high fraction of positive magnitude epistasis (66.7% of all mutation pairs). The second most common type of epistasis was negative magnitude epistasis, with an average fraction of 21.1% of mutation pairs (fig. 5*B*). An exception is the transcription factor CRF4, where no mutation pair showed negative magnitude epistasis (supplementary table 10, Supplementary Material online). Next was simple sign epistasis (15.9% of mutation pairs, and reciprocal sign epistasis [2.7%, fig. 5*B*]). Reciprocal sign epistasis was most rare for 17 of the 19 analyzed transcription factors, but this was not the case for the transcription factors CRF4 and the AP2 transcription factor DREB1C (AT4G25470), where reciprocal sign epistasis occurred in at least 9.5% of squares (supplementary table 10, Supplementary Material online).

In sum, our analysis of epistasis based on mutation pairs revealed that between 7.9% and 29.7% of mutation pairs show sign epistasis. Because sign epistasis creates local valleys in an adaptive landscape, it reduces path accessibility, and can thus help explain the existence of inaccessible paths to high affinity binding sites in our data (Poelwijk et al. 2011). This raises the possibility that natural selection has affected the incidence of sign epistasis in *A. thaliana* populations, by favoring sequences that show little sign epistasis. However, we found no evidence that the incidence of epistasis itself is subject to adaptive evolution (supplementary text 3, Supplementary Material online).

Inferred Binding Sequences Tend to Have Lower Binding Affinities

Our analysis of epistatic interactions among four binding sequence variants (ab, Ab, aB, and AB, fig. 5*A*) identified many quadruplets (squares) of sequences where at least one

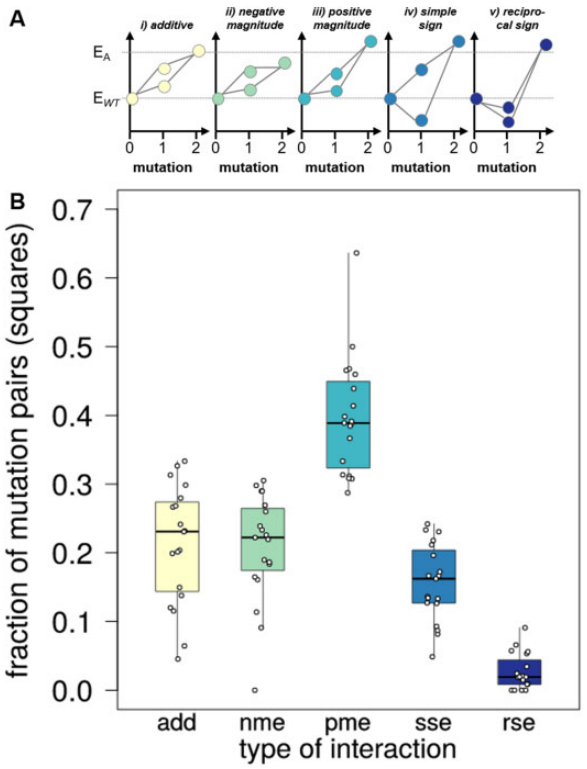

Fig. 5.—Prevalence of epistasis in transcription factor binding sequences. (*A*) Schematic for the five kinds of mutation pairs we distinguished according to the type of epistasis between them. The vertical axis represents binding affinities (*E*-scores). E_{WT} denotes the *E*-score of the wild-type, and E_A represents the *E*-score of the double mutant when the two mutations interact additively (gray dashed lines). The horizontal axes show the number of mutational steps between sequences, and the circles in each plot represent hypothetical DNA binding affinities for the wild-type sequence (zero mutational steps), two one-mutant neighbors, and one two-mutant neighbor. (*i*) Additivity (absence of epistasis), where affinities (*E*-scores) of both single mutant neighbors are higher than the *E*-score of the wild-type, and their sum equals the *E*-score of the double mutant. (*ii*) Negative magnitude epistasis. Again, both single mutant neighbors have a higher *E*-score than the wild-type, but the double mutant has an *E*-score that is lower than the sum of the *E*-scores of the single-mutant neighbors. (*iii*) Positive magnitude epistasis, which is analogous to negative magnitude epistasis, but the *E*-score of the double mutant is higher than the sum of the two single mutant neighbors. (*iv*) Simple sign epistasis, where one single-mutant neighbor has a higher and the other single mutant neighbor has a lower *E*-score than the wild-type. In (*v*), both single mutant neighbors have a lower *E*-score than the wild-type which illustrates reciprocal sign epistasis. (*B*) Empirical data on the fraction of mutation pairs (vertical axis) that fall into each of the five categories (horizontal axis) from (*A*). Each open circle corresponds to data from one of our 19 transcription factors. Acronyms: add, additive (no epistasis); nme, negative magnitude epistasis; pme, positive magnitude epistasis; sse, simple sign epistasis; rse, reciprocal sign epistasis. Colors in each plot correspond to the corresponding color in (*A*), representing different types of epistasis. The central bold horizontal line in the box plot indicates the median of all 19 individual data points, and the lower and upper box limits represent the first and third quartiles, respectively. Whiskers indicate values within the 1.5-fold interquartile range.

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

sequence had to be inferred, because it does not occur in the *Arabidopsis* population we studied. Because high affinity binding sequences appear to be favored by natural selection, we reasoned that the inferred sequences may not occur in vivo because they are weak binders and have been eliminated by natural selection. To test this hypothesis, we first focused on squares with one inferred sequence. Specifically, we compared the binding affinity of the three present genomic sequences with that of the single inferred sequence, and did so for all sequence squares of each transcription factor. For 18 out of our 19 transcription factors, the missing sequences indeed have significantly (after Bonferroni-correction for multiple testing) weaker binding affinities (*P* values between 4.045×10^{-27} and 0.0008; sample size between 21 and 415 mutation pairs). Only in transcription factor CRF4 did the missing sequences not have significantly weaker binding affinity, possibly because this factor had a small sample size of only six squares. However, selective purging of weak binders cannot be the only reason for the absence of some sequences, because in 3.6% of squares, the inferred sequence had greater binding affinity than the existing sequences.

We repeated this analysis for squares with two missing sequences. Perhaps partly due to the modest sample size (between 1 and 41 squares per transcription factor), we did not find any significant differences (after correcting for multiple testing) in affinity between existing and inferred sequences for any one transcription factor in this subset of our data. In sum, where sample sizes are sufficiently large, most evidence is consistent with the possibility that selection eliminates low affinity binding sequences. However, we also note that transcription factor binding sequences may be absent from a population for a variety of other reasons, such as selection against transcriptional cross-talk (Friedlander et al. 2016), mutational biases (Long et al. 2018; Svensson and Berger 2019; Cano and Payne 2020), or simply historical accidents.

Discussion

Variation at genomic loci that are bound by transcription factors contributes to phenotypic differences between and within species, and is subject to natural selection (Wray 2007; Romero et al. 2012). The population genetics of transcription factor binding sites has often been studied by theoretical approaches (Khatri and Goldstein 2015; Tuğrul et al. 2015). In contrast, we aimed to understand how such sites evolve in natural populations of *A. thaliana*. To this end, we first identified 8,333 genomic loci that are bound by 1 of 19 analyzed transcription factors in vivo. Consistent with previous reports in *A. thaliana* and other species, we found low genetic diversity at bound loci, suggesting that they are subject to purifying selection (Heyndrickx et al. 2014; Wang et al. 2018; Radke et al. 2021). We then studied two factors that may contribute to the high population frequency of transcription factor binding sequences, namely a sequence's binding affinity and its evolutionary accessibility.

Mutations in a DNA sequence bound by a transcription factor can alter the sequence's affinity to this transcription factor. In species as different as yeast, fruit flies and humans, natural selection usually favors high affinity (Mustonen and Lässig 2009; Sharon et al. 2012; Arbiza et al. 2013). However, exceptions from this principle exist. For example, several studies in *Drosophila melanogaster* highlighted that weak binding is favored by natural selection for some transcription factors (Ramos and Barolo 2013; Crocker et al. 2015; Delker et al. 2019). In *A. thaliana*, little pertinent information is available, especially about the importance of low-affinity transcription factor binding (Lai et al. 2019). To estimate the importance of binding affinity for the evolution of transcription factor binding sequences in *A. thaliana*, we estimated the statistical association between their affinity and their frequency in the *A. thaliana* population we studied. For all 19 transcription factors we analyzed, the binding sequences with the highest frequencies also tended to bind their cognate transcription factor with the highest affinities. Strong binding is thus favored by natural selection.

The evolutionary accessibility of genotypes has been the subject of both theoretical (Berestycki et al. 2016; Zagorski et al. 2016) and empirical work (Chevereau et al. 2015; Lukačišinová et al. 2020). Previous work examined evolutionary accessibility in proteins (Weinreich et al. 2006; Wu et al. 2016; Hartman and Tullman-Ercek 2019), transcription factor binding sequences (Payne and Wagner 2014; Aguilar-Rodríguez et al. 2017), and metabolic phenotypes (Josephides and Swain 2017). These studies highlighted the interplay between the ruggedness of an adaptive landscape, the number of global and local peaks in such a landscape, and the fraction of accessible mutational paths.

Our work goes beyond previous studies because it examines accessibility in natural populations. We show that 77.3% of pairs of single mutations in transcription factor binding sites interact epistatically, and 17.7% show sign epistasis, which can render some evolutionary paths to a genotype inaccessible (Poelwijk et al. 2007, 2011). Not surprisingly then, different binding sequences vary in their accessibility. Most importantly, highly accessible binding sequences also tend to have high population frequency. The reason is not just that accessibility is correlated with affinity. For six of our 19 transcription factors, the accessibility of a binding sequence remains correlated with frequency after controlling for affinity. For one transcription factor, accessibility even plays the dominant role in explaining frequencies of binding sequences. Taken together, our observations provide empirical evidence that both fitness and accessibility can matter in the evolution of regulatory genotypes.

An open question is how affinity and accessibility interact in the evolution of binding sequences. For example, high accessibility may help create a sequence that is weakly bound by

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

a transcription factor, and selection may subsequently increase the affinity of this sequence. Another open question regards the role of genetic drift in the evolution of binding sequences. The effective population size of *A. thaliana* is modest, because of a strong bottleneck imposed by the last ice age and a high incidence of inbreeding (Gossmann et al. 2010; Cao et al. 2011). In small populations, genetic drift is strong, and may help populations traverse adaptive valleys caused by epistasis. It may thus render otherwise inaccessible mutational paths to high affinity binding sequences accessible (Jain et al. 2011; Lobkovsky et al. 2011). In addition, strong drift may reduce the efficiency of selection in promoting high affinity binding. In a species with higher effective population size, both path accessibility and selection for high affinity may play a greater role in binding site evolution, but they may be affected to a different extend by population size. In addition to genetic drift, developmental systems drift, which permits the divergence of gene regulatory logic even when gene expression phenotypes are preserved, may influence the evolution of transcription factor binding sequences (Townsley and Sinha 2012). Its effects may also be stronger in small populations.

Our analysis of accessible mutational paths has further limitations. Most notably, multiple additional factors may affect path accessibility. For example, some paths may be favored because of biases in mutation rates (Long et al. 2018; Svensson and Berger 2019; Cano and Payne 2020). Also, some of the sequences we study may be bound by multiple transcription factors, perhaps as a consequence of unavoidable transcriptional cross-talk (Friedlander et al. 2016), or to allow combinatorial control of gene expression. Fewer accessible paths may exist to such sequences because their accessibility requires that the binding affinity of more than one transcription factor must increase monotonically along an evolutionary path. In consequence, path accessibility may actually be smaller than suggested by our analysis. If this is the case, then affinity may play an even more important role in the evolution of binding sequences.

A more general limitation of our work comes from the low genomic diversity of *A. thaliana* accessions (0.5% nucleotide diversity; The 1001 Genomes Consortium 2016). This low diversity is reduced further by purifying selection on transcription factor binding sequences. On the one hand, this makes it unlikely that our results and conclusions are affected by ascertainment biases. We used the genome of accession Col-0 to identify bound loci. Because genetic diversity is low, it is unlikely that we would identify many other bound loci when using other genomes. On the other hand, the result of low genetic diversity is that only a small number of alleles exist at each locus bound by a transcription factor, which limits our ability to study the prevalence of epistasis and other quantities that involve interactions among different mutations. For example, epistatic interactions between more than two genomic positions can convert inaccessible to accessible mutational paths, but we were unable to analyze such polymorphisms for a lack of pertinent data (Weinreich et al. 2013; Wu et al. 2016). Among all 8,333 genomic loci bound by at least one of our 19 transcription factors, we found only 19 complete "squares" of a binding site, two one-mutant neighbors, and the two-mutant neighbor they can form. To obtain more data, one would need to analyze more diverse genomes, but in such genomes, it can quickly become challenging to identify orthologous regulatory sequences (Liang et al. 2008; Baker et al. 2011).

Another limitation is that PBM experiments are performed with those variants of a transcription factor that are encoded by accession Col-0. Even though the extent of DNA sequence variation among all accessions is small (0.5%), different accessions may encode a slightly different variant of each transcription factor, which may differ in their affinity for DNA binding. To exclude this possibility, one would have to quantify DNA binding affinity for each transcription factor variant separately.

Fourth, our work is only based on bound loci identified in root tissue, and we cannot exclude the possibility that loci bound in different tissues or developmental stages are subject to different evolutionary forces.

A final limitation stems from the paucity of relevant gene expression data, which preclude a reliable mapping of binding affinity to gene expression (supplementary text 2, Supplementary Material online). For example, the DNAse I hypersensitivity data we used to identify loci bound by transcription factors in vivo is available for root tissue, but the most suitable gene expression data that cover a large proportion of our analyzed 1,135 accessions is available for leaf tissue (supplementary text 2, Supplementary Material online). Similar data limitations preclude an analysis of other mechanisms of gene regulation. Among them is DNA methylation (Zhang et al. 2018), which is important for the regulation of *A. thaliana* genes (O'Malley et al. 2016). Also among them is regulation mediated by noncoding RNAs (Heo et al. 2013), and cross-family transcription factor interactions (Bemer et al. 2017). They constitute additional layers of regulation that may diminish the importance of transcription factor binding.

In sum, most limitations of our work stem from limited data availability. To obtain deeper insights into the structure of the adaptive landscapes on which transcription factor binding sequences evolve, more functional in vivo data will be particularly important. Given the limited diversity of *A. thaliana* accessions, genomic data from closely related species would also be useful. However, even the limited data we have suggest that high fitness may not always be the reason why some genotypes are frequent in a population. If evolutionary accessibility matters even in the simple genotypes we study, the contingencies it can create may play an even greater role in more complex genotypes, such as those of regulatory circuits and whole organisms (Blount et al. 2018; Edwards 2019).

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

Materials and Methods

We used only previously published and publicly available data sets. These included 1) data on in vitro binding affinities of a transcription factor to double-stranded DNA, as determined through PBMs (Franco-Zorrilla et al. 2014; Weirauch et al. 2014; Hume et al. 2015); 2) data on genomic binding regions of transcription factors identified through in vitro DNA affinity purification and sequencing (O'Malley et al. 2016); 3) genomic footprint data from a DNase I hypersensitivity in vivo experiment with root tissue from *A. thaliana* (Sullivan et al. 2014); and 4) genomic polymorphism data from a worldwide collection of 1,135 *A. thaliana* accessions (The 1001 Genomes Consortium 2016). This collection also contains the genome sequence of accession Col-0, which serves as a reference genome of the collection (The 1001 Genomes Consortium 2016). We obtained annotation information for accession Col-0 from The Arabidopsis Information Resource (Lamesch et al. 2012; Berardini et al. 2015; https://www.arabidopsis.org/, last accessed August 30, 2021), version TAIR10. We next describe these data sets in more detail.

In Vitro DNA Binding Affinity Data of Transcription Factors

In vitro binding affinities of a transcription factor to sequences of double-stranded DNA can be measured with PBMs. These microarrays contain DNA fragments that are ten nucleotides long and that cover all 32,896 possible nucleotide sequences of length eight multiple times, where "reverse-complement" sequences are counted only once (Berger et al. 2006; Berger and Bulyk 2009). Affinity data are typically reported for nucleotide eight-mers, because the multiple representation of each eight-mer within all ten-mers allows for more robust binding measurements (Berger et al. 2006; Weirauch et al. 2014). We thus note that all binding sequences and bound genomic loci that we analyze in this work are no longer than eight nucleotides. In a PBM experiment, transcription factor-DNA binding affinities are measured by applying an epitope-tagged transcription factor to the microarray, followed by incubation of the bound transcription factor with fluorophore-coupled antibodies, and measurement of fluorescence. The fluorescence signal measured for each spot (DNA fragment) on the array is then converted to a binding affinity value (*E*-score) with a rank-based statistic. In this way, the most favored eight-mer is assigned an *E*-score of $+0.5$ and the most disfavored eight-mer is assigned an *E*-score of -0.5 (Berger et al. 2006; Berger and Bulyk 2009). *E*-scores allow comparisons across transcription factors and correlate with binding affinities. Thus, they can be considered as relative binding affinities (Berger et al. 2006; Payne and Wagner 2014).

We used three sources of PBM data (Franco-Zorrilla et al. 2014; Weirauch et al. 2014; Hume et al. 2015) and analyzed only binding data for transcription factors that met the following three requirements: 1) results from two replicate PBM experiments are available for the transcription factor; 2) the transcription factor's binding sequence is no longer than eight base pairs, as determined by position weight matrix data obtained through PBM experiments (Weirauch et al. 2014); and 3) results from in vitro DNA affinity purification and sequencing experiments (O'Malley et al. 2016) are available. In contrast to DNase I hypersensitivity experiments, the latter data provide a link between genomic regions and individual transcription factors that bind to this region. Nineteen transcription factors from seven families fulfilled all three requirements (fig. 1*A*; supplementary table 1, Supplementary Material online). We obtained all binding affinities (*E*-scores) for these transcription factors from the database CIS-BP version 2.00 (Weirauch et al. 2014).

We considered a DNA binding sequence as specifically bound by a transcription factor if its *E*-score exceeded a threshold of 0.35 in both replicate PBM experiments, because such high *E*-scores indicate high affinity binding, and they are associated with a false discovery rate of binding that is smaller than 0.001 (Badis et al. 2009; Zhu et al. 2009; Payne and Wagner 2014; Aguilar-Rodríguez et al. 2017). Depending on the transcription factor, we identified between 14 and 647 nucleotide eight-mers as specifically bound according to this criterion (supplementary table 1, Supplementary Material online). Because PBM experiments are inherently noisy (Berger and Bulyk 2009), we calculated separately for each transcription factor a noise value δ that quantifies experimental noise originating from two replicate binding affinity measurements (supplementary table 1, Supplementary Material online). For our calculations, we first determined binding affinity values of all bound sequences (*E*-score >0.35 in both replicate experiments). Then, we fitted a linear regression between affinity values of both experiments. Specifically, we used data from each replicate experiment once as the explanatory variable, and once as the dependent variable. We computed measurement noise δ as the average of the two residual standard errors of each linear regression (Aguilar-Rodríguez et al. 2017).

Genomic Transcription Factor Binding Regions

Short genomic regions bound by transcription factors were previously identified in vitro by DNA affinity purification and sequencing in the *A. thaliana* accession Col-0. In such experiments, a haloalkane dehalogenase-tagged and in vitro expressed transcription factor is immobilized on beads and incubated with fragmented genomic DNA. DNA fragments bound to the transcription factor are then sequenced, and a genomic binding region with a typical length of 200 base pairs is identified (O'Malley et al. 2016). We obtained data on the bound genomic regions from the PlantCistromeDB (http://neomorph.salk.edu/PlantCistromeDB; release 1 from May 2016, last accessed August 30, 2021), including transcription factors in our analysis whose data meet the following two

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

requirements. First, genomic binding regions had been identified with PCR-amplified DNA fragments. This step eliminates epigenetic DNA modifications like methylation. We thereby ensure comparability with data obtained from PBM experiments, where DNA fragments on the array also contain no epigenetic information. Second, we require that a transcription factor was analyzed in both PBM and affinity purification/sequencing experiments (fig. 1*A*).

Identification of In Vivo Bound Genomic Loci

To detect high quality bound loci within the genomic binding regions identified by DNA affinity purification/sequencing (O'Malley et al. 2016), we first filtered all binding regions, and retained only those binding regions that overlapped with genomic footprints identified by in vivo DNase I hypersensitivity assays (Sullivan et al. 2014; fig. 1*B*). Footprint data do not allow one to link a bound genomic region to a specific transcription factor, but DNA affinity purification/sequencing data provide this link, because such experiments are performed with a single known transcription factor. Next, we scanned each genomic region for the presence of eight-mers with high binding affinities (*E*-score $>$ 0.35 in both replicate experiments). To this end, we used a sliding window approach with a window length of eight nucleotides and a step size of one nucleotide (fig. 1*B*). If we found such an eight-mer, we designated this genomic region a bound locus. We considered binding sequences on both DNA strands, and if we detected more than one bound locus in one genomic region, we randomly chose one bound locus for further analyses.

To assign a bound locus to a target gene, we retained only bound loci located within 500 base pairs upstream of a gene's start codon, or up to the stop or start codon of the nearest upstream gene, if the intergenic region was shorter than 500 base pairs (fig. 1*B*). In this way, we identified 8,333 bound loci (between 29 and 1,864 bound loci per transcription factor; supplementary table 2, Supplementary Material online). In summary, we identified all loci in the reference genome of accession Col-0 that are bound in vivo by at least one of the analyzed 19 transcription factors. Whether each of these bound loci is also bound by any of its orthologous sequences in the other 1,134 *A. thaliana* accessions depends on a variety of factors, such as the affinity (*E*-score) of the orthologous sequence to the focal transcription factor, and chromatin accessibility at the orthologous locus.

We note that for 18 out of 19 analyzed transcription factors, only a subset of all sequences that can in principle be bound, as determined by in vitro PBM experiments, actually occur at bound loci in the reference accession Col-0. Specifically, the fraction of sequences that can be bound and that occur at bound genomic loci ranges between 11.9% for the transcription factor CRF4, and 89.5% for the trihelix transcription factor AT5G47660 in accession Col-0 (supplementary table 1, Supplementary Material online). This observation may have at least two explanations. First, our analysis was restricted to loci bound by transcription factors in root tissue. Additional bound loci with different binding sequences may be identified in other tissues or developmental stages. Second, binding sequences at any one bound locus may have evolved to avoid detrimental regulatory crosstalk that occurs when multiple transcription factors interact with the same binding sequence (Friedlander et al. 2016). Such crosstalk avoidance may reduce the set of binding sequences used by an organism.

Population Genomic Data

In addition to the genome sequence of the accession Col-0, we used 1,134 *A. thaliana* genome sequences from a worldwide collection of accessions (version 3.1; The 1001 Genomes Consortium 2016). These sequences had been previously obtained by combining a reference genome sequence from accession Col-0 (version TAIR 10) with information on nucleotide variants specific to each other accession, including indels. In this sequence data, a nucleotide whose identity had not been unambiguously identified in any one accession is designated by the letter N (The 1001 Genomes Consortium 2016). The 1,134 accession's genome sequences, as well as the genomic binding regions of transcription factors determined with DNA affinity and purification had previously been mapped onto the same reference genome version TAIR10 (O'Malley et al. 2016; The 1001 Genomes Consortium 2016). This allowed us to identify positions orthologous to specific bound loci in each of the 1,134 genome sequences. For this purpose, we used files that indicate, for each position in the reference genome (TAIR10), the corresponding position in the accession of interest (https://1001genomes.org/data/GMI-MPI/releases/v3.1/pseudogenomes/dat/, last accessed August 30, 2021). Through this procedure, we identified for each bound locus of each transcription factor in the accession Col-0 orthologous binding sequences in the other 1,134 sequenced *A. thaliana* accessions. Next, we removed orthologous sequences of a bound locus from our data set if they spanned an indel or contained ambiguous nucleotides (Ns), because such sequences cannot be linked to binding affinity data obtained from PBMs. This procedure led us to exclude 1,096 bound loci (between three and 308 loci, depending on the transcription factor; supplementary table 2, Supplementary Material online), because *all* of the 1,134 orthologous binding sequences of a bound locus contained Ns or indels. More generally, depending on the transcription factor and bound locus, we could use between 75.6% and 84.4% of all orthologous binding sequences across all bound loci of a transcription factor for further analysis (supplementary table 2, Supplementary Material online). Overall, this filtering of bound loci with sequences containing ambiguous nucleotides and indels resulted in a reduced data set of 7,237 bound loci.

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

Reconstruction of Genotype Networks

We constructed one genotype network for each individual transcription factor and for each of its bound loci, based on the 1,135 *Arabidopsis* accessions we consider. Each node (or vertex) of such a genotype network corresponds to one of the orthologous nucleotide sequences of one bound locus. Two nodes are neighbors (connected by an edge) if the associated genotypes (binding sequences) differ in exactly one nucleotide. To construct such a network, we first retained, for each bound locus, only sequences without indels or ambiguous nucleotides. The maximally possible number of retained sequences is 1,135, one from the Col-0 reference, and 1,134 from the worldwide collection of other accessions. However, because many sequences contained indels or ambiguous nucleotides, we were able to retain only between one and 1,097 sequences, depending on the bound locus.

Second, we created a list of *unique* orthologous sequences for each bound locus, treating sequences that are reverse-complements of each other as identical. This list comprised between one and eleven binding sequences. To construct a meaningful genotype network, one needs at least two different binding sequences per bound locus, that is, the locus needs to be polymorphic in the 1,135 accessions. This was the case for 3,349 bound loci (supplementary table 2, Supplementary Material online), and we aimed to construct genotype networks only for these loci.

Third, we scanned the list of unique binding sequences to identify all pairs of sequences that differ in exactly one position. We call two such genotypes one-mutant neighbors, and retained only sequences that had at least one such neighbor for further analyses. In graph-theoretical language (Brandes and Erlebach 2005), this means that we only studied connected components of a genotype network with at least two nodes. Exceptions to this general approach are sequences that differ at two positions, which we used for the detection of epistasis (see Detection and Classification of Epistasis). We built all genotype networks using the R (R Core Team 2020) package igraph (Csárdi and Nepusz 2006).

The above procedure yielded between 18 and 768 such networks, depending on the transcription factor. These networks comprised between two and eleven binding sequences, consistent with the observation that a maximum of 11 unique binding sequences exist at any one bound locus we studied (supplementary table 2, Supplementary Material online). For 3,289 bound loci, networks had a single connected component (Brandes and Erlebach 2005), and for one bound locus, the network consisted of two components. For the remaining 59 of the 3,349 analyzed binding loci, we found at least two different putative binding sequences, but none of these sequences differed at one nucleotide (supplementary table 11, Supplementary Material online).

Detection and Classification of Epistasis

Our approach to detect epistasis, that is, nonadditive effects of DNA mutations on binding affinities, relies on the presence of mutation pairs. Such mutation pairs can be represented as "squares" in a genotype network, that is, four nodes that form a cycle of edges in a genotype network (Poelwijk et al. 2007, 2011; Aguilar-Rodríguez et al. 2017). We first scanned each genotype network at all bound loci for the presence of such squares. We then filtered the list of squares and kept only those that contained at least one binding sequence with an *E*-score >0.35 in both replicate experiments to ensure that we do not analyze unbound sequences (none of the four sequences may originate from the reference accession Col-0). Using a previously established method (Poelwijk et al. 2011; Aguilar-Rodríguez et al. 2017), we then investigated the prevalence, type, and strength of epistasis. To this end, we studied how a wild-type sequence would evolve via a succession of two single mutations to a double mutant. We designated the wild-type sequence as genotype "ab," the two single-mutants as "Ab" and "aB," and the double-mutant as "AB." We assumed that sequences evolve toward stronger binding such that sequence AB shows the strongest binding affinity (*E*-score) among all four sequences. We then calculated the strength of epistasis ε (i.e., the quantitative deviation from additive interactions between the mutants) as

$$\varepsilon = E_{AB} + E_{ab} - E_{ab} - E_{aB} \quad (1)$$

where E denotes the binding affinity of a sequence. For example, E_{AB} is the binding affinity (mean *E*-score from two replicate experiments) of binding sequence AB. We considered two mutants as interacting epistatically only if the absolute strength of epistasis $|\varepsilon|$ was greater than the noise threshold δ. If this condition was met, we classified the epistatic interaction as magnitude epistasis (positive or negative), simple sign epistasis, or reciprocal sign epistasis (Poelwijk et al. 2007), using the following criteria. Magnitude epistasis requires that

$$\Delta E_{ab\to Ab} + \Delta E_{aB\to AB} = |\Delta E_{ab\to Ab}| + |\Delta E_{aB\to AB}|. \quad (2)$$

We classified such mutation pairs as displaying positive magnitude epistasis ($\varepsilon > 0$) or negative magnitude epistasis ($\varepsilon < 0$). Simple sign epistasis requires that

$$\Delta E_{ab\to Ab} + \Delta E_{aB\to AB} < |\Delta E_{ab\to\ Ab}| + E_{aB\to\ AB}. \quad (3)$$

Reciprocal sign epistasis requires that equation (3) and that

$$\Delta E_{ab\to AB} + \Delta E_{Ab\to AB} < |\Delta E_{ab\to\ aB}| + E_{Ab\to\ AB}. \quad (4)$$

In (2–4), ΔE denotes the effect of single mutations (e.g., ab → Ab) on binding affinity. Wherever ΔE was smaller than the noise threshold δ, we assigned ΔE a value of zero. If all four mutational effects were smaller than δ (even if

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

$|\varepsilon| > \delta$), we classified the mutational interaction as nonepistatic. These restrictions ensure a conservative quantification of epistasis (Aguilar-Rodríguez et al. 2017).

We followed this procedure to detect epistasis also for genotype networks where one or two sequences of a cyclic path with length four were missing, that is, they were absent at a bound locus from any of the 1,135 accessions. Such incomplete squares include triplets, that is, three connected genotypes that do not form a cycle. They also included pairs of genotypes that differ by two mutations but are not connected by an edge to a shared one-mutant neighbor. We note that every incomplete square can be unambiguously completed in silico by inferring the missing single sequence (triplets) or the missing two sequences. In total, our data set contained 1,891 squares, among which 19 contained no inferred sequence, 1,739 contained one inferred sequence, and 133 contained two inferred sequences. We excluded 40 squares from further analysis because none of the four sequences exceeded an *E*-score of 0.35.

"Peakness" of Binding Sequences

We define the "peakness" P of a DNA sequence S that can be bound by a transcription factor as the fraction of its one-mutant neighbors with an affinity for the transcription factor that is not higher than that of S itself. If P equals one (all neighbors have lower affinity values) then S is part of a local peak in the affinity landscape of the transcription factor.

To quantify P for any one focal binding sequence S of one of our 19 transcription factors, we determined the nucleotide sequences of all $8 \times 3 = 24$ one-mutant neighbors of S. We then counted the number of these neighbors with a binding affinity (E_N) that is lower than or equal to the binding affinity of the focal sequence E_S, that is, if $E_N \leq E_S$, or if $\text{abs}(E_N - E_S) \leq \delta$, where δ is the noise threshold. We then divided the resulting number by 24. A high value of this quantity indicates that many one-mutant neighbors have a lower or equal binding affinity (*E*-score) compared with the focal sequence S.

Calculating the Fraction of Accessible Mutational Paths

The complete genotype space with 32,896 sequences is too large for exhaustive computational explorations of all possible mutational paths between any two sequences. Moreover, the fraction of accessible paths to a focal binding sequence decreases with increasing path length (Aguilar-Rodríguez et al. 2017). For computational feasibility, we thus restricted our analysis of path accessibility to paths with a length not exceeding four mutational steps, that is, to sequences that differ from a focal sequence by no more than four mutations. Any one focal binding sequence has $\binom{8}{L} \times 3^L$ L-mutant neighbors, and each mutant neighbor can be reached via $L!$ mutational paths, where L denotes the number of nucleotide differences between two sequences. Specifically, a focal binding sequence has $8 \times 3 = 24$ one-mutant neighbors, and one path (the single mutation) leads from each of them to the focal sequence. In addition, it has $\binom{8}{2} \times 3^2 = 252$ two-mutant neighbors, and $2! = 2$ associated paths, that is, paths leading from each mutant neighbor to the focal sequence. It has also $\binom{8}{3} \times 3^3 = 1{,}512$ three-mutant neighbors with $3! = 6$ associated paths, and $\binom{8}{4} \times 3^4 = 5{,}670$ four-mutant neighbors with $4! = 24$ paths leading to each such neighbor.

To construct all mutational paths for any one focal binding sequence, we first created in silico the set S_{fw} of all one-, two-, three-, and four-mutant neighbors, as well as the set S_{rv} of all one-, two-, three-, and four-mutant neighbors of the reverse-complement of the focal sequence. We then created the union S_U of S_{fw} and S_{rv}, treating sequences that are reverse complements of each other as identical, that is, they are only present once in S_U. Next, we iterated over all sequences s in S_U and calculated the mutational distance of s to the focal sequence f, as well as the mutational distance between the reverse complement of sequence s and the focal sequence f. If the two distances differed, we analyzed the sequence pair with the smaller distance. If this smaller distance was between one and four, we enumerated all shortest mutational paths from sequence s to sequence f. We then scanned each path for sequences that are reverse complements of each other. Because we treat such sequences as identical, we discarded such paths from subsequent analyses, because they would lead us to misestimate the path length. We then called a mutational path accessible if each mutational step leading from some sequence S_n to a neighboring sequence S_{n+1} along this path was accessible, that is, if the associated *E*-scores (E_{Sn} and E_{Sn+1}) increased monotonically or strictly monotonically (Weinreich et al. 2006; Poelwijk et al. 2007; Aguilar-Rodríguez et al. 2017). More specifically, for a monotonically increasing path, we required that $E(S_{n+1}) \geq E(S_n)$ or that $\text{abs}(E(S_{n+1}) - E(S_n)) \leq \delta$ for all sequences along the path. For strictly monotonically increasing paths, we required that $E(S_{n+1}) > E(S_n) + \delta$. In both cases, δ denotes the experimentally determined noise value. We quantified the fraction of accessible paths that lead to a specific focal sequence, and refer to this fraction as the (evolutionary) accessibility of the sequence.

Obtaining Nucleotide Sequences of Length Eight from Random Genomic Positions

Following previous work (Heyndrickx et al. 2014), we based our analysis of random genomic positions on the third position of 4-fold degenerated codons as a proxy for genomic DNA that is under weak or no selection (Li et al. 1985). The genetic code has 32 codons with 4-fold degenerated third positions. These codons encode the amino acids alanine, arginine, glycine, leucine, proline, serine, threonine, and valine. We scanned all protein-coding genes annotated on the

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

"forward" strand of the reference genome (accession Col-0) for the presence of one of these 32 codons, and noted the genomic location of the third codon position. This yielded a total set Q of 1,805,469 third-codon positions across all chromosomes. We then randomly selected once, for each chromosome and without replacement, 8κ genomic positions from Q, where κ denotes the number of bound loci of all transcription factors on each chromosome (chromosome 1: 2,178 bound loci; chromosome 2: 1,180 bound loci; chromosome 3: 1,706 bound loci; chromosome 4: 1,367 bound loci; and chromosome 5, 1,902 bound loci; supplementary table 2, Supplementary Material online). We then concatenated the randomly selected positions to create sequences of length eight, thereby creating random sequences of the same length as our transcription factor binding sequences. We refer to each such sequence as a random genomic sequence.

We then used the other 1,134 *A. thaliana* accessions to obtain orthologous sequences for each random genomic sequence as described for genomic loci bound by transcription factors (see Population Genomic Data). The relevant genomic positions and sequences are summarized in supplementary table 3, Supplementary Material online.

Frequency of Epistasis in Randomly Selected Mutation Pairs

For each of our 19 transcription factors, we wanted to compare the incidence of epistasis in the binding sequences that occur at bound loci in the *A. thaliana* genome (henceforth: in vivo data) to that determined in vitro from all 32,896 possible eight-mers on a PBM. To quantify epistasis in in vitro sequences, we used affinity data from PBM experiments for all mutation pairs ("squares") of eight-mers that differed by at most two nucleotides as we describe next.

To identify all possible squares in the genotype space of all eight-mers, we first collected all nucleotide eight-mers that differ by two nucleotides. Next, we identified in silico the sequences of both single mutant neighbors that allow the formation of a square. Note that this needs to be done only once, because all pertinent PBM experiments were performed with the same collection of nucleotide eight-mers present on the chip (Weirauch et al. 2014). Finally, we filtered the data for unique squares, and for squares that consist of four truly different sequences, excluding sequences that are reverse complements of each other. This procedure yielded 2,148,416 unique squares.

In a next step, we filtered all squares by GC content and down-sampled the number of squares for which we had in vitro data to match the sample size of the in vivo data. Specifically, we first sorted the squares from the in vivo data into bins according to the GC content of the four sequences in it. To this end, we created 33 bins representing a GC-content from zero to 32 (i.e., four sequences with eight base pairs each), and assigned each square to one bin. We binned the squares for the in vitro data analogously. Then we randomly sampled sequences from each bin for the in vitro data, where the total sample size equaled the total number of squares from the in vivo data. The probability with which we selected a square equaled the fraction of squares in that bin for the in vivo data.

We performed this analysis in 10,000 replicates, and separately for each transcription factor. We categorized the sampled squares by the type of epistasis we described in the main text (fig. 5*A*). The results of this analysis are provided in supplementary file 1, Supplementary Material online. We used the replicate samples for a randomization test of the null hypothesis that the in vivo incidence of epistasis is not significantly different from that expected in vitro. To test if an observed incidence of epistasis was significantly different from the incidence expected by chance, we calculated the number of random samples, where the incidence of epistasis was smaller or greater than in the biological data. Dividing these numbers by 10,000 (the total number of random samples) yielded a P value for the randomization test. We Bonferroni-corrected for multiple testing by dividing the alpha level of 5% by $19 \times 5 \times 2 = 190$, because we performed the analysis for 19 transcription factors, discriminated between five types of mutation pairs (fig. 5*A*), and considered higher or smaller incidences of epistasis in random samples compared with biological data.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by a research fellowship of the German Research Foundation (DFG, SCHW 1968/1-1) to G.S. and by grants from the European Research Council under Grant Agreement No. 739874, the Swiss National Science Foundation (grant 31003A_172887), and the University Priority Research Program in Evolutionary Biology to A.W. We thank all members of the Andreas Wagner laboratory and Joshua L. Payne (ETH Zürich) for stimulating discussions.

Author Contributions

G.S. and A.W. have done the conceptualization, funding acquisition, methodology, visualization, and also wrote the manuscript (original draft preparation, review, and editing); G.S. did the formal analysis, investigation, and project administration; A.W. did the supervision.

Data Availability

All data that we used in this study are available from public repositories. Binding affinity data were obtained from CIS-BP

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

(http://cisbp.ccbr.utoronto.ca/, last accessed August 30, 2021), data for in vitro bound loci were obtained from http://neomorph.salk.edu/PlantCistromeDB, last accessed August 30,2021 information for in vivo bound loci were downloaded from http://plantregulome.org/ (last accessed August 30, 2021), and information about the 1,135 *A. thaliana* accessions (including genomic polymorphism data) were collected from https://1001genomes.org/ (last accessed August 30, 2021). Data that were generated during the course of this study are available as Supplementary Material online. Scripts that are needed to reproduce analyses are provided in supplementary file 1, Supplementary Material online, which is available at Zenodo (https://zenodo.org/record/5718712#.YZu96SYo_kw; doi:10.5281/zenodo.5718712, last accessed December 15, 2021).

Literature Cited

Aguilar-Rodríguez J, Payne JL, Wagner A. 2017. A thousand empirical adaptive landscapes and their navigability. Nat Ecol Evol. 1(2):45.

Badis G et al. 2009. Diversity and complexity in DNA recognition by transcription factors. Science 324:1720–1723.

Baker CR, Tuch BB, Johnson AD. 2011. Extensive DNA-binding specificity divergence of a conserved transcription regulator. Proc Natl Acad Sci USA. 108(18):7493–7498.

Bemer M, van Dijk ADJ, Immink RGH, Angenent GC. 2017. Cross-family transcription factor interactions: an additional layer of gene regulation. Trends Plant Sci. 22(1):66–80.

Berardini TZ, et al. 2015. The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. Genesis 53(8):474–485.

Berestycki J, Brunet E, Shi Z. 2016. The number of accessible paths in the hypercube. Bernoulli 22(2):653–680.

Berger MF, et al. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol. 24(11):1429–1435.

Berger MF, Bulyk ML. 2009. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc. 4(3):393–411.

Blount ZD, Lenski RE, Losos JB. 2018. Contingency and determinism in evolution: replaying life's tape. Science 362(6415):aam5979.

Brandes U, Erlebach T, editors. 2005. Netwwork analysis—methological foundations. Berlin (Germany): Springer.

Cano AV, Payne JL. 2020. Mutation bias interacts with composition bias to influence adaptive evolution. PLOS Comput Biol. 16(9):e1008296.

Cao J, et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet. 43(10):956–965.

Chevereau G, et al. 2015. Quantifying the determinants of evolutionary dynamics leading to drug resistance. PLoS Biol. 13(11):e1002299.

Connelly C, Skelly DA, Dunham MJ, Akey JM. 2013. Population genomics and transcriptional consequences of regulatory motif variation in globally diverse *Saccharomyces cerevisiae* strains. Mol Biol Evol. 30(7):1605–1613.

Coulon A, Chow CC, Singer RH, Larson DR. 2013. Eukaryotic transcriptional dynamics: from single molecules to cell populations. Nat Rev Genet. 14(8):572–584.

Crocker J, et al. 2015. Low affinity binding site clusters confer hox specificity and regulatory robustness. Cell 160(1–2):191–203.

Csárdi G, Nepusz T. 2006. The igraph software package for complex network research. InterJ Comp Syst. 1695.

Cunto F, Provero P. 2011. Evolution of promoter affinity for transcription factors in the human lineage. Mol Biol Evol 28:2173–2183.

Delker RK, Ranade V, Loker R, Voutev R, Mann RS. 2019. Low affinity binding sites in an activating CRM mediate negative autoregulation of the *Drosophila* Hox gene Ultrabithorax. PLoS Genet. 15(10):e1008444.

Edwards EJ. 2019. Evolutionary trajectories, accessibility and other metaphors: the case of C4 and CAM photosynthesis. New Phytol. 233:1742–1755.

Franco-Zorrilla JM, et al. 2014. DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc Natl Acad Sci USA. 111(6):2367–2372.

Friedlander T, Prizak R, Guet CC, Barton NH, Tkačik G. 2016. Intrinsic limits to gene regulation by global crosstalk. Nat Comm. 7:12307.

Gao R, Stock AM. 2015. Temporal hierarchy of gene expression mediated by transcription factor binding affinity and activation dynamics. mBio 6(3):e00686-15.

Gossmann TI, et al. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol Biol Evol. 27(8):1822–1832.

Grassi E, Zapparoli E, Molineris I, Provero P. 2015. Total binding affinity profiles of regulatory regions predict transcription factor binding and gene expression in human cells. PLoS One 10(11):e0143627.

Hartman EC, Tullman-Ercek D. 2019. Learning from protein fitness landscapes: a review of mutability, epistasis, and evolution. Curr Opin Syst Biol. 14:25–31.

He BZ, Holloway AK, Maerkl SJ, Kreitmann M. 2011. Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules. PLoS Genet. 7:2011.

Heo JB, Lee YS, Sung S. 2013. Epigenetic regulation by long noncoding RNAs in plants. Chromosome Res. 21(6–7):685–693.

Heyndrickx KS, Van de Velde J, Wang C, Weigel D, Vandepoele K. 2014. A functional and evolutionary perspective on transcription factor binding in *Arabidopsis thaliana*. Plant Cell 26(10):3894–3910.

Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML. 2015. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res. 43(Database Issue):D117–D122.

Ichihashi Y, et al. 2014. Evolutionary developmental transcriptomics reveals a gene network module regulating interspecific diversity in plant leaf shape. Proc Natl Acad Sci USA. 112:12758–12763.

Inukai SI, Kock KH, Bulyk ML. 2017. Transcription factor–DNA binding: beyond binding site motifs. Curr Opin Genet Dev. 43:110–119.

Jain K, Krug J, Park SC. 2011. Evolutionary advantage of small populations on complex fitness landscapes. Evolution 65(7):1945–1955.

Jiang Z, Dong X, Li Z-G, He F, Zhang Z. 2016. Differential coexpression analysis reveals extensive rewiring of *Arabidopsis* gene coexpression in response to *Pseudomonas syringae* infection. Sci Rep. 6:35064.

Josephides C, Swain P. 2017. Predicting metabolic adaptation from networks of mutational paths. Nat Comm. 8:685.

Kauffman S, Levin S. 1987. Towards a general theory of adaptive walks on rugged landscapes. J Theor Biol. 128(1):11–45.

Khatri BS, Goldstein RA. 2015. A coarse-grained biophysical model of sequence evolution and the population size dependence of the speciation rate. J Theor Biol. 378:56–64.

Kwasnieski JC, et al. 2012. Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc Natl Acad Sci USA. 109(47):19498–19503.

Kvitek DJ, Sherlock G. 2011. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. PLoS Genet. 7(4):e1002056.

Lai X, et al. 2019. Building transcription factor binding site models to understand gene regulation in plants. Mol Plant. 12(6):743–763.

Lamesch P, et al. 2012. The *Arabidopsis* information resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 40(Database Issue):D1202–D1210.

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023

Lasky JR, et al. 2014. Natural variation in abiotic stress responsive gene expression and local adaptation to climate in *Arabidopsis thaliana*. Mol Biol Evol. 31(9):2283–2296.
Li W-H, Wu C-I, Luo C-C. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol. 2(2):150–174.
Liang H, Lin Y-S, Li W-H. 2008. Fast evolution of core promoters in primate genomes. Mol Biol Evol. 25(6):1239–1244.
Liu J, Robinson-Rechavi M. 2020. Robust inference of positive selection on regulatory sequences in the human brain. Sci Adv. 6:eabc9863.
Lobkovsky AE, Wolf YI, Koonin EV. 2011. Predictability of evolutionary trajectories in fitness landscapes. PLoS Comput Biol. 7(12):e1002302.
Long H, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. Nat Ecol Evol. 2(2):237–240.
Lukačišinová M, Fernando B, Bollenbach T. 2020. Highly parallel lab evolution reveals that epistasis can curb the evolution of antibiotic resistance. Nat Comm. 11:3105.
Lv Q, Cheng R, Shi T. 2014. Regulatory network rewiring for secondary metabolism in *Arabidopsis thaliana* under various conditions. BMC Plant Biol. 14:180.
McCandlish DM. 2013. On the findability of genotypes. Evolution 67(9):2592–2603.
Molineris I, Grassi E, Ala U, Di Cunto F, Provero P. 2011. Evolution of promoter affinity for transcription factors in the human lineage. Mol Biol Evol. 28(8):2173–2183.
Mu XJ, et al. 2011. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 genomes project. Nucleic Acids Res. 39(16):7058–7076.
Mustonen V, Lässig M. 2009. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. Trends Genet. 25(3):111–119.
Naidoo T, Sjödin P, Schlebusch C, Jakobsson M. 2018. Patterns of variation in cis-regulatory regions: examining evidence of purifying selection. BMC Genomics 19(1):95.
Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA. 76(10):5269–5273.
Nakamichi N. 2020. The transcriptional network in the *Arabidopsis* circadian clock system. Genes 11(11):1284.
O'Malley RC, et al. 2016. Cistrome and epicistrome features shape the regulatory DNA landscape. Cell 165(5):1280–1292.
Payne JL, Wagner A. 2014. The robustness and evolvability of transcription factor binding sites. Science 343(6173):875–877.
Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ. 2007. Empirical fitness landscapes reveal accessible evolutionary paths. Nature 445(7126):383–386.
Poelwijk FJ, Tănase-Nicola S, Kiviet DJ, Tans SJ. 2011. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. J Theor Biol. 272(1):141–144.
R Core Team. 2020. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Radke DW, et al.; Alzheimer's Disease Neuroimaging Initiative. 2021. Purifying selection on noncoding deletions of human regulatory loci detected using their cellular pleiotropy. Genome Res. 31(6):935–946.
Ramos AI, Barolo S. 2013. Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. Philos Trans R Soc Lond B Biol Sci. 368(1632):20130018.
Rastogi C, et al. 2018. Accurate and sensitive quantification of protein-DNA binding affinity. Proc Natl Acad Sci USA. 115(16):E3692–E3701.
Rice G, Rebeiz M. 2019. Evolution: how many phenotypes do regulatory mutations affect? Curr Biol. 29(1):R21–R23.
Romero IG, Ruvinsky I, Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. Nat Rev Genet. 13(7):505–516.
Schaper S, Louis AA. 2014. The arrival of the frequent: how bias in genotype-phenotype maps can steer populations to local optima. PLoS One. 9(2):e86635.
Sharon E, et al. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat Biotechnol. 30(6):521–530.
Signor SA, Nuzhdin SV. 2018. The evolution of gene expression in cis and trans. Trends Genet. 34(7):532–544.
Stewart AJ, Hannenhalli S, Plotkin JB. 2012. Why transcription factor binding sites are ten nucleotides long. Genetics 192(3):973–985.
Sullivan AM, et al. 2014. Mapping and dynamics of regulatory DNA and transcription factor networks in *A. thaliana*. Cell Rep. 8(6):2015–2030.
Svensson EI, Berger D. 2019. The role of mutation bias in adaptive evolution. Trends Ecol Evol. 34(5):422–434.
The 1001 Genomes Consortium. 2016. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell. 166:481–491.
Torgerson DG, et al. 2009. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genet. 5(8):e1000592.
Townsley BT, Sinha NR. 2012. A new development: evolving concepts in leaf ontogeny. Annu Rev Plant Biol. 63:535–562.
Tuğrul M, Paixão T, Barton NH, Tkačik G. 2015. Dynamics of transcription factor binding site evolution. PLoS Genet. 11(11):e1005639.
Verma N. 2019. Transcriptional regulation of anther development in *Arabidopsis*. Gene. 689:202–209.
Vernot B, et al. 2012. Personal and population genomics of human regulatory variation. Genome Res. 22(9):1689–1697.
Wang X, et al. 2018. Analysis of genetic variation indicates DNA shape involvement in purifying selection. Mol Biol Evol. 35(8):1958–1967.
Weinreich DM, Watson RA, Chao L. 2005. Perspective: sign epistasis and genetic constraint on evolutionary trajectories. Evolution 59(6):1165–1174.
Weinreich DM, Delaney NF, Depristo M, Hartl DL. 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312(5770):111–2007.
Weinreich DM, Lan Y, Wylie CS, Heckendorn RB. 2013. Should evolutionary geneticists worry about higher-order epistasis? Curr Opin Genet Dev. 23(6):700–707.
Weirauch MT, et al. 2014. Determination and inference of Eukaryotic transcription factor sequence specificity. Cell 158(6):1431–1443.
West MAL, et al. 2007. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. Genetics 175(3):1441–1540.
Wittkopp PJ, Kalay G. 2011. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet. 13(1):59–69.
Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 8(3):206–216.
Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. 2016. Adaptation in protein fitness landscapes is facilitated by indirect paths. eLife 5:e16965.
Zagorski M, Burda Z, Waclaw B. 2016. Beyond the Hypercube: evolutionary accessibility of fitness landscapes with realistic mutational networks. PLoS Comput Biol. 12(12):e1005218.
Zhang X, Cal AJ, Borevitz JO. 2011. Genetic architecture of regulatory variation in *Arabidopsis thaliana*. Genome Res. 21(5):725–733.
Zhang H, Lang Z, Zhu J-K. 2018. Dynamics and function of DNA methylation in plants. Nat Rev Mol Cell Biol. 19(8):489–506.
Zhu C, et al. 2009. High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res. 19(4):556–566.

Associate editor: Soojin Yi

Downloaded from https://academic.oup.com/gbe/article/13/12/evab273/6459646 by University Library Zurich / Zentralbibliothek Zurich user on 03 January 2023