Supplementary Materials for

Survey of variation in human transcription factors reveals prevalent DNA binding changes

Luis A. Barrera, Anastasia Vedenko*, Jesse V. Kurland*, Julia M. Rogers, Stephen S. Gisselbrecht, Elizabeth J. Rossin, Jaie Woodard, Luca Mariani, Kian Hong Kock, Sachi Inukai, Trevor Siggers, Leila Shokri, Raluca Gordân, Nidhi Sahni, Chris Cotsapas, Tong Hao, Song Yi, Manolis Kellis, Mark J. Daly, Marc Vidal, David E. Hill, Martha L. Bulyk
correspondence to: mlbulyk@receptor.med.harvard.edu

This PDF file includes:

Materials and Methods
SupplementaryText
Figs. S1 to S12
Table S5
Captions for Tables S1 to S4, S6, S7

Other Supplementary Materials for this manuscript includes the following:

Databases S1 to S4, S6, S7 as zipped archives: [paste data table titles in a list]

Table S1. Nonsynonymous (missense) SNPs in TF DNA-binding domains.
Table S2. Transcription factors analyzed for the presence of variants in this study.
Table S3. Nonsense SNPs identified in TF DNA-binding domains.
Table S4. PBM experimental conditions and DBD clone sequences.
Table S6: Specificity and affinity changes identified by PBMs.
Table S7. PBM 8-mer data for GST-only negative control PBM experiments.

Materials and Methods

Identification of DNA-binding domains of sequence-specific transcription factors

We identified genes for sequence-specific TFs based on a previously published, manually curated census of human TFs (*5*). Only TF genes from the two highest confidence categories, requiring direct functional evidence or the presence of domains never found in non-TF genes (encompassing 1,364 genes), were considered in this study. Protein sequences for the selected TF genes were retrieved from the Ensembl database (version 67) (*24*). To identify matches to DNA-binding domains (DBDs), we retrieved hidden Markov models (HMMs) from the Pfam database corresponding to DBD structural classes that have been identified in human genomes (*25*). The hmmscan tool, which is part of the HMMER 3.0 package (*26*), was used to scan human protein sequences for DBD instances. We used the default hmmscan parameters, except for a more stringent domain match threshold (E-value < 0.0001). In total, 1,254 genes from the original list had currently valid Ensembl Gene IDs and matches to one of the DBD classes that were retrieved. We used this reduced set of 1,254 TF genes for all subsequent analyses.

We used variant annotations obtained from dbNSFP v2.0b (*27*) to link nucleotide changes to amino acid substitutions and identify nsSNPs. For each SNP, its effect on all overlapping Ensembl transcript models was considered. DBDPs were identified as nsSNPs that affected the sequences of the DBDs identified by HMMER, as described above. In rare instances, DBD matches differed across transcripts due to alternative splicing. In such cases, the transcript with the best match score (lowest E-value) to the Pfam HMM was selected to represent the DBD for that gene.

Sources of SNPs and disease mutations

The nsSNPs selected for experimental testing were drawn from either the 1000 Genomes Project Phase 2 release (*2*) (1700 individuals) or the Exome Sequencing Project 6500 release (February, 2013; 6,503 individuals) (*7*). A later release of the 1000 Genomes Project data (Phase 3, September 2014 release; 2,535 individuals) was used for statistical analyses. Variants from the Exome Aggregation Consortium (ExAC) v0.2 release (61,486 individuals) (*1*) were also used for statistical analyses and determination of allele frequencies, but with the exception of HOXD13 N298S, were not considered for experimental testing. For analyses involving the number of variants found per individual, only the 1000 Genomes Project Phase 3 data were used because other datasets did not provide full genotype data. In all cases, only variants that passed the most stringent level of quality control filters ("PASS" value in the VCF file) were used for statistical analyses or selected for experimental testing. Mendelian variants were retrieved from the curated set of Online Mendelian Inheritance in Man (OMIM) mutations in the UniProtKB (*28*) (release 2013_05) database, with the exception of the HOXD13 Q325K mutation, which was only recently identified in an individual with complex hand and foot malformations (*12*). For all genes, the coordinates of amino acid substitutions were mapped to the canonical splice isoform selected by UniProtKB. The domain position affected by mutations was determined from the optimal alignment between the Pfam HMM and the protein sequence, as determined by the hmmscan tool in the HMMER 3.0 package (*26*).

Annotation of DNA-contacting residues in select Pfam domains

Four DBD structural classes were selected for detailed annotation of residues likely to participate in DNA-binding: C2H2 zinc-fingers (Pfam: zf-C2H2 / PF00096), homeodomains (Pfam: Homeobox / PF00046), forkhead (Pfam: Forkhead / PF00250), and basic helix-loop-helix domains (Pfam: HLH / PF00010). These classes were prioritized based on their occurrences in significant numbers of human TFs and the availability of prior knowledge about the amino acid residues that are involved in DNA-contacts. For all classes except homeodomains, backbone- and base-contacting domain positions were identified based on published studies (*29-37*). For each class, the positions of amino acids that had been described explicitly as base- or backbone-contacting in the literature were manually linked to the corresponding positions in the Pfam domain. The residues annotated as DNA-contacting in each case are shown in Fig. S2. If a residue at a given position in the domain was reported as making both base and backbone contacts, it was annotated as base-contacting. Residues at positions adjacent to base-contacting residues that were not identified as making backbone contacts were annotated as "adjacent to a base-contacting residue".

In the case of homeodomains, we analyzed structural data to comprehensively identify residues that may play a role in protein-DNA contacts. Ten homeodomain co-crystal structures in the Protein Data Bank (PDB IDs: 1IG7, 2H1K, 3LNQ, 3HDD, 9ANT, 1JGG, 1DU0, 2HDD, 2HOS, and 1APL) were chosen to sample a wide range of sequence diversity within the homeodomain family while excluding complexes that exhibited cooperative dimerization or included co-factors. When multiple identical proteins were contained within the same unit cell, a single instance was selected for analysis. Coordinates were extracted from PDB files using the "pdbread" function from the MATLAB Bioinformatics Toolbox, which was also used to calculate distances between amino acid residues and non-hydrogen atoms in DNA. We separately considered contacts between amino acid residues and DNA bases and amino acid residues and the phosphate backbone. The minimal distance between amino acid residues and DNA was used to define contact strength: contacts within 3.5 Å were assigned a score of 2, while contacts between 3.5 Å and 5 Å were assigned a score of 1. Contact maps for separate proteins were aligned using ClustalW v2.1 with default settings to perform a multiple sequence alignment of the corresponding protein sequences. DNA sequences were aligned by visual inspection. For each position in the domain and each position in the binding site, we calculated the mean contact score over all structures, creating an average contact map that summarizes the likelihood that a residue participates in DNA contacts. The average score obtained for each domain position was used for subsequent prioritization, as described below. All homeodomain positions with non-zero average scores for backbone or base contacts were annotated as putatively DNA-contacting in the Pfam domain annotation scheme (Fig. S2).

Prioritization of variants for experimental testing

We used several criteria to filter DBDPs found in population sequencing studies and identify variants that were likely to alter DNA-binding. These criteria can be summarized as (a) the prevalence of the variant in the population, (b) inferred proximity

of affected residues to DNA based on structural data, (c) deleteriousness of the mutation as predicted by published tools, and (d) known phenotypic associations of the affected gene, and are described in more detail below. To minimize the number of selected variants that may be due to sequencing errors, variants found in heterozygous form in only one individual were excluded. Otherwise, DBDPs that had certain combinations of features of interest were manually evaluated and curated. Ideally, we sought to find variants that were present in many individuals, affected genes with known phenotypes, and were predicted to have a significant potential to alter DNA-binding properties or disrupt protein stability. In practice, variants were considered for testing if they met the criteria for at least two categories. If multiple DBDPs were found in the same gene, variants that met just one criterion were sometimes tested alongside variants that met multiple criteria to allow comparisons of effect sizes between prioritized and non-prioritized variants. Similarly, we selected a few variants that were not predicted to alter DNA-binding but occurred in genes for which Mendelian disease mutations had been chosen for experimental testing.

Structural information was used to prioritize variants by determining if the affected residue was in an annotated DNA-contacting position. For the four DBD classes for which Pfam domains were annotated, the per-position annotations were used to evaluate whether residue changes were likely to affect protein-DNA contacts. This was done by finding the optimal match to the Pfam HMM for a given protein sequence and determining if the domain position in which the amino acid substitution occurred was annotated as DNA-contacting. A small subset of 8 DBDPs was prioritized by manual evaluation of the consequences of the amino acid substitution on homologous co-crystal structures.

Several tools designed to predict whether coding mutations are likely to be biochemically damaging were used to aid in the prioritization of variants: SIFT (*38*), PolyPhen-2 (*9*), LRT (*39*), MutationTaster (*40*), and MutationAssessor (*41*). Studies comparing the agreement between predictions made by different tools have reported significant discrepancies, but have also shown that combining predictions from different tools improves overall accuracy (*42*). Based on these observations, DBDPs that were predicted to be damaging by at least three of the five tools were assigned the highest priority. However, variants were also considered in cases where at least one predictor tool considered the variant as damaging and the effect of the substitution was deemed of high likelihood to impact DNA-binding through the methods described above.

In addition to residue-specific considerations, we integrated information about gene-level phenotypic associations into our prioritization scheme. DBDPs affecting genes with at least one associated OMIM code, as annotated in UniProt KB (*28*), were assigned a higher priority, as these variants are *a priori* more likely to have phenotypic consequences. We also considered whether genes harboring DBDPs were associated with variants found in genome-wide association studies (GWAS) in the NHGRI GWAS catalog (*11*). DBDPs in genes that were directly reported in association to traits (*i.e.*, in the "Reported Gene(s)" column in the GWAS catalog) were given higher priority. In addition, we considered whether DBDPs were in linkage disequilibrium (LD) with GWAS tag SNPs. We retrieved LD tables derived from the AFR (African), AMR (Admixed American), EUR (European) and ASN (Asian) populations in the 1000 Genomes Phase 1 data from the HaploReg tool (*43*). DBDPs in LD with GWAS SNPs

from the NHGRI catalog at a threshold of $R^2 > 0.5$ in any population were assigned a high priority for experimental testing.

Finally, we selected a set of DBDPs that were considered as unlikely to affect DNA-binding but were deemed to be interesting for other reasons. These included DBDPs that occurred in genes that were being assayed for the effect of other variants or that occurred at high minor allele frequencies in genes that had known Mendelian phenotypes.

We selected Mendelian disease mutations under two general categories: (a) mutations affecting genes for which DBDPs were prioritized for experimental testing, (b) mutations occurring in genes in which several Mendelian disease variants were known to affect the same DNA-binding domain. Whenever a gene harbored a DBDP that was prioritized for experimental testing and the same gene had known Mendelian disease mutations, at least one mutation was selected for experimental testing. Mendelian disease mutations were also chosen for testing in cases where different mutations within the DBD were associated with distinct OMIM codes (*i.e.*, phenotypes), particularly when certain mutations affected DNA-contacting residues. Conclusions from comparisons between the PBM binding profiles and distinct phenotypes depend on accurate diagnosis and resolution in distinguishing related diseases in patients.

In a few limited cases, additional variants were included that did not strictly match the criteria described above, but occurred in genes where other assayed mutations had already been selected in accordance with the above criteria. Three selected variants (PAX4 R133W, NR2E3 R77Q and ZNF655 E327G) altered the flanking sequence between Pfam domain matches, as opposed to the canonically defined DBDs themselves. These variants were excluded from analyses involving DNA-contacting positions. Two DBD variants were included based on observations reported in the HapMap Project (HOXC4 R158L and SNAI2 T324I) and one variant was included based on multiple observations in dbSNP (SNAI2 T324I).

Selection of TF subsequences for cloning

We identified TF amino acid sequences corresponding to the DBDs, as defined by Pfam HMM matches, plus 15 amino acid (a.a.) flanks expanding towards both the N-terminal and C-terminal ends. Previous studies have successfully used GST-tagged constructs comprising the DBD and 15 a.a. flanks in PBM experiments (*44, 45*). Here, we employed the same strategy. In cases where multiple DBDs were present in the same protein (e.g., PAX TFs, or proteins with multiple C2H2 zinc-finger domains), we created constructs that encompassed all DBDs plus 15 flanking amino acids of the DBDs located closest to the protein termini.

Generation of TF Entry clones and LR transfer into pDEST15 vector

We created N-terminal glutathione S-transferase (GST) fusion constructs of the 158 DBD alleles. Specifically, Entry clones carrying the selected TF subsequences were generated by PCR-based Gateway recombinational cloning. For PCR amplification, all the forward and reverse primers contained attB1 and attB2 sites, respectively, at their 5' ends. PCR reactions were performed using KOD Hot Start DNA polymerase according

to the manufacturer's protocol (Novagen), and using TF reference clones from human ORFeome version 7.1 (http://horfdb.dfci.harvard.edu/hv7/) as template. The resulting PCR products were then cloned into pDONR223 vector by Gateway BP reactions, yielding desired TF Entry clones. After bacterial transformation, miniprep plasmid DNA of all Entry clones was extracted, and then transferred individually by in vitro Gateway LR cloning into pDEST15 expression vector, deriving N-terminal GST-tagged TF fusions. All these expression clones were sequence-verified in two directions using universal primers pGEXfw and T7-Terminator, and no mutations were found. The primer sequences are as follows:

- pGEXfw: 5'-GGCAAGCCACGTTTGGTG-3'
- T7-Terminator: 5'-GCTAGTTATTGCTCAGCG-3'

A small subset of the TF DBDs (NR2E3, KLF1, GFI1B, VAX2, FOXC1, ARX, PAX3) were generated by gene synthesis (GenScript). The TF subsequences were codon optimized for *E. coli* expression, flanked by attB1 and attB2 sites at the 5' and 3' ends, respectively, and cloned into pUC57. These constructs were then transferred into pDONR221 using the Gateway BP reaction, to generate Entry clones. As above, these clones were then transferred into pDEST15 by Gateway LR cloning, and then sequence-verified.

Generation of mutant DBD clones

To generate mutant TF clones, we used an enhanced, two-stage, site-directed mutagenesis pipeline (*46*). Briefly, for a given TF mutation, the mutagenesis platform consisted of two "primary PCRs" to generate TF fragments, and one "fusion PCR" to obtain the mutated TF. For the primary PCRs, vector-specific universal primers were used in combination with the respective two TF-specific internal forward and reverse primers to generate overlapping fragments containing the desired nucleotide substitution. The universal primers allowed the Gateway recombination sites to be preserved on both ends of the TFs. The mutation-specific primers, MutF and MutR, harboring the desired nucleotide changes, were designed to be complementary to each other. Site-directed PCRs were performed on either TF domains already cloned into the Destination vector pDEST15 or on TF domain Entry clones in pDONR223. For TF domains in pDEST15, the two TF fragments flanking a given mutation were amplified using the primer pair Tag1-pGEXfw and MutR, and the primer pair Tag2-T7-Term and MutF, respectively. In the subsequent fusion PCR, the two primary fragments were fused together using the primer pair Tag1 and Tag2 to generate the mutated TFs, and the mutant TF PCR products were then introduced into pDONR223 by a BP reaction followed by bacterial transformation. For TF domains in pDONR223, the two TF fragments flanking a given mutation were amplified using the primer pair M13G-FOR and MutR, and the primer pair M13G-REV and MutF, respectively. In the subsequent fusion PCR, the two primary fragments were fused together using the primer pair M13G-FOR and M13G-REV to generate the mutated TFs, and the mutant TF PCR products were then introduced into pDEST15 by an LR reaction followed by bacterial transformation. At least two independent colonies per mutant TF were isolated. Following sequence confirmation by Sanger sequencing, the clones that had only the

desired mutations (no additional mutations) were selected and consolidated. Mutant TFs in pDONR223 were transferred to pDEST15 by Gateway LR reactions.

Primer sequences used are as follows:

- M13G-FOR: 5'-CCCAGTCACGACGTTGTAAAACG-3'
- M13G-REV: 5'-GTGTCTCAAAATCTCTGATGTTAC-3'
- Tag1-pGEXfw: 5'-GGCAGACGTGCCTCACTACTGGCAAGCCACGTTTGGTG-3'
- Tag2-T7-Term: 5'-CTGAGCTTGACGCATTGCTAGCTAGTTATTGCTCAGCG-3'
- Tag1: 5'-GGCAGACGTGCCTCACTACT-3'
- Tag2: 5'-CTGAGCTTGACGCATTGCTA-3'

Three of the HOXD13 mutant alleles (I297V, N298S, and Q325K) were generated through site-directed mutagenesis using a single PCR step. The reference allele of HOXD13 in pDEST15 was used as a template for PCR with a pair of complementary primers containing the desired substitution. The PCR product was digested with DpnI, and then directly transformed into bacteria. At least two colonies per mutant were assessed by sequencing to verify that only the desired mutation had been introduced.

Protein expression and quantification

In vitro transcription and translation (IVT) reactions were performed according to the manufacturer's protocol (NEB PURExpress IVT Kit). Western blots were used to estimate molar concentrations of all in vitro translated proteins by utilizing a dilution series of recombinant GST (Sigma) essentially as described previously (*13*). Equal volumes of IVT samples and known concentrations of GST were suspended in 4x XT Sample Buffer (BioRad), heated to 95 °C for 5 minutes, and loaded on a precast 4-12% Bis-Tris Criterion gel (Bio-Rad). Samples were subject to electrophoresis at 190 V for 35 minutes and then transferred to a nitrocellulose membrane (Sigma) at 100-115 mA for 2 hours. Membranes were visualized using the SuperSignal West Femto Maximum Sensitivity Substrate kit (Pierce) according to the manufacturer's protocols. Primary antibody was added to achieve a final concentration of 20 ng/ml (rabbit anti-GST antibody; Sigma cat #097K4767). Secondary antibody was added at a final concentration of 5 ng/ml (goat anti-rabbit secondary Ab; ThermoScientific #31460).

Films were scanned and concentrations of full-length proteins were determined using Quantity One software version 4.5.0 (BioRad), in accordance with the GST standard curve. All reference and alternative allele proteins were expressed in the same IVT batch.

Protein binding microarray (PBM) experiments

Oligonucleotide arrays were double-stranded and PBM experiments were performed following previously described experimental protocols (*13, 47*). The array design employed was an "all 10-mer" universal array in 8 x 60K, GSE format (Agilent Technologies; AMADID #030236). To minimize potential batch effects, reference and

mutant alleles for the same TF were assayed on separate chambers in the same PBM slide. All experiments comparing reference and alternative alleles used proteins expressed in the same batch and diluted to achieve equal TF concentrations across an allelic series. Experimental conditions used for all PBM experiments, including TF concentrations and buffers, are described in Table S4.

Negative control 'GST-only' PBM experiments were performed using GST expressed by the PURExpress IVT kit (New England Biolabs) according to the manufacturer's protocol using the empty pDEST15 expression vector (Invitrogen). 'GST-only' PBM experiments were performed in duplicate chambers at 100 nm and 600 nm GST in PBS buffers following previously described experimental protocols (*13, 47*).

Protein binding microarray (PBM) data processing

PBM scan images were obtained using a GenePix 4000A Microarray Scanner (Molecular Devices). The resulting image data were processed using GenePix Pro v7.2 to obtain signal intensity data for each spot. The data were then further processed by using Masliner software (v1.02) (*48*) to combine scans from different intensity settings, increasing the effective dynamic range of the signal intensity values. If a dataset had any negative background-subtracted intensity (BSI) values (which can occur if the region surrounding a spot is brighter than the spot itself), consistent pseudocounts were added to all BSI values such that they all became nonnegative. All BSI values were normalized using the software for spatial de-trending providing in the Universal PBM Analysis Suite (*47*), as previously described (*13, 47*).

PBM-based evaluation of DNA-binding changes

For each PBM experiment, we used the Seed-and-Wobble algorithm (*13*), which is part of the Universal PBM Analysis Suite (*47*), to calculate the PBM enrichment score (E-score) of each of 32,768 nonredundant, ungapped 8-mers for each protein (*13*). The E-score is a rank-based statistic that is closely related to the area under the receiver operating characteristic (ROC) curve and robust to technical variation across arrays (*13*). Larger E-score values reflect higher specificity for binding a particular 8-mer. Z-scores for each 8-mer and position weight matrices (PWMs) were also derived using the Universal PBM Analysis Suite and Seed-and-Wobble algorithm, respectively. Sequence logos for each allele were created by using the Seed-and-Wobble PWM as input for WebLogo v2.8.2 (*49*) with default parameters.

The presence of E-scores ≥ 0.45 has been reported as a viable quality control metric to identify successful PBM experiments (*45, 50*). Here, we deemed a PBM experiment to be of acceptable quality under a more stringent criterion of yielding $\geq$ five 8-mers with an E-score ≥ 0.45. Because some mutant TF alleles are expected to lose their ability to bind DNA specifically, we considered such experiments acceptable for publication as long as the reference allele protein expressed and tested in the same batch yielded $\geq$ five 8-mers with E-scores ≥ 0.45.

Identifying affinity differences

To determine if two alleles exhibited a difference in binding affinity, we compared the distribution of E-scores obtained for each allele. A high E-score value indicates a strong deviation from the null distribution for the ranks of probes containing instances of a particular 8-mer. As the affinity of a TF allele increases while the concentration is constant, more binding sites will be occupied at high frequencies. Therefore, with all other parameters remaining constant, a higher affinity allele should yield a PBM dataset with a larger number of high-scoring 8-mers.

To detect affinity changes, we used the Wilcoxon rank-sum test to determine whether a pair of experiments showed statistically significant differences in their top E-scores. We calculated the Wilcoxon rank-sum test P-value when comparing the highest 50 E-scores in each experiment. We corrected the P-values derived from comparing reference and alternative alleles using the Benjamini-Hochberg correction (*51*), which was calculated over all pairwise comparisons between reference and alternative alleles. Mutations were classified as changing affinity when $Q < 0.05$. The direction of the affinity change (*i.e.*, increase or decrease) was determined by comparing the median value among the top 50 E-scores for each allele and selecting the allele with the larger median value as the one with the predicted higher affinity. Our PBM-based criteria exhibited perfect specificity and moderate sensitivity (0.71) in detecting affinity changes relative to other experimental methods (Table S5, Fig. S5); therefore, our approach is conservative, identifying true changes with high confidence.

Identifying specificity differences

To detect specificity differences between alleles, we used a previously described method (*14*) for identifying statistically significant differences among 8-mer E-scores between two PBM datasets. Briefly, DNA 8-mers are placed into overlapping groups composed of all 8-mers that contain matches to a given DNA 6-mer. The E-scores corresponding to 8-mers in each of these groups are then compared across alleles using an intersection-union test (*14*), followed by the adjustment of P-values using the Benjamini-Hochberg correction (*51*). The result is a set of 6-mers that are bound preferentially by one TF allele over the other.

Here, we developed a stringent set of criteria for determining whether a mutant TF allele bound DNA with altered specificity relative to the reference allele. First, we excluded any experiments where the alternative allele significantly lost sequence-specific binding activity, as these cases might lead to confounded affinity and specificity changes. Therefore, only datasets from alternative alleles that met the same quality control criterion used for reference alleles (at least five 8-mers with E-scores ≥ 0.45) were tested for specificity differences. In addition, we excluded pairs of alleles where the number of 8-mers bound by the alternative allele at an E-score ≥ 0.45 was at least 2-fold less than the number bound by the reference allele. For the remaining pairs, we used the method described above to find preferred 6-mers with a q-value < 0.05. We found that pairwise comparisons between alleles where at least three 6-mers were bound preferentially by either allele and whose 8-mer E-scores had an R^2 value < 0.9 were highly reproducible across replicates (see below). Therefore, we considered alleles that matched all criteria described in this paragraph and for which pairwise comparisons with the reference allele yielded ≥ 3 preferred 6-mers and had 8-mer E-score $R^2 < 0.9$ to have altered specificity.

Reproducibility of affinity and specificity differences

E-scores have been previously shown to be highly reproducible across replicate PBM experiments (*47*). We verified that alleles identified as having altered affinity or specificity were consistently labeled as such in a set of 58 duplicate PBM experiments. The dataset for each replicate experiment was independently scored using the criteria described above. Affinity calls were found to be consistent across replicate experiments in 90% of replicate pairs, while specificity calls were consistent across 89% of replicates. In discordant cases, the replicate experiments with the largest total number of E-scores ≥ 0.45 were used to determine whether a particular allele had altered affinity or specificity.

Concordance with experimental data from prior studies

We searched the literature to identify cases where the same mutations selected for this study had been previously tested experimentally to determine their biochemical effects. Through manual curation, we collected a set of 19 experiments that directly or indirectly measured the binding affinities of mutant alleles, as summarized in Table S5. In most of these cases, only qualitative data were provided, such as gel images derived from non-quantitative electrophoretic mobility shift assays. Therefore, to enable systematic comparisons, we manually curated each reported experiment and assigned the mutant allele to one of three categories: (a) no effect on DNA binding (0), (b) partial loss of binding (-), and (c) complete loss of binding (--). The results of this comparison are summarized in Fig. S5.

Comparisons to HOXD13 ChIP-Seq and RNA-Seq data

For these in-depth analyses, we selected the highest quality (as defined above) PBM datasets obtained for the HOXD13 reference, Q325K and Q325R alleles. We used published ChIP-Seq and RNA-Seq data for the mouse Hoxd13 protein (which has an identical homeodomain sequence to the human HOXD13 protein), the mouse Hoxd13 Q317K variant (according to the OMIM amino acid residue numbering for the HOXD13 protein; in this paper, we use the ClinVar (*52*) amino acid residue numbering, derived from RefSeq transcript NM_000523.3 and refer to this position as Q325 throughout this paper) and the mouse Hoxd13 Q317R variant (human HOXD13 Q325R) overexpressed in chicken mesenchymal stem cells (GEO accession number: GSE44799) (*12*). ChIP-Seq peaks were based on two independent biological replicates and called using MACS2, as previously described (*12*). The RNA-Seq experiments provided transcriptome profiles (in RPKM) based on one RNA sample for each condition, mapped to the reference *Gallus gallus* (chicken) (galGal3) transcriptome using bowtie (*12*). Our analysis, described in the following paragraphs, compared the reference HOXD13 allele with each of these two HOXD13 mutant alleles separately.

Identification of HOXD13 allele-preferred and allele-common 8-mers

To distinguish allele-preferred 8-mers (blue and red points in Figure 3B and Figure S8A) versus allele-common 8-mers (black points in Figure 3B and Figure S8A), we identified allele-preferred 6-mers from PBM data using the same approach as described above for all TF allelic series. To select 8-mers preferentially-bound by either the reference or the mutant HOXD13 allele in the pairwise comparisons, we first filtered for

8-mers with E-score above 0.4 for at least one of the two alleles, and then identified the 8-mers that also (i) contained at least one allele-preferred 6-mer for the respective allele, and (ii) had a respective allele-preferred E-score that was at least 0.05 greater than the E-score for the other allele (blue and red points, Figure 3B and Figure S8A). To select 8-mers bound by both alleles ("allele-common 8-mers"), we filtered for 8-mers with E-score above 0.4 for at least one of the two alleles, but that did not contain any allele-preferred 6-mers and had an E-score difference between the two alleles of less than 0.1 (black points in Figure 3B and Figure S8A).

Enrichment of PBM 8-mers in ChIP-Seq peaks

To examine the enrichment of the identified allele-preferred or allele-common 8-mers in ChIP-Seq peaks, we calculated the area under receiver operating characteristic curve (AUROC, Figure 3C and Figure S8B) and the area under precision-recall curve (AUPR, Figure S9) statistics to evaluate enrichment in the top 1000 actual ChIP-Seq peak sequences as compared to a background set of 1000 randomized sequences (*53*). We defined "shared" ChIP-Seq peaks as those having a reciprocal overlap of peak boundaries >50% in pairwise comparisons of ChIP-Seq data for the mutant and reference alleles. The rest of the peaks were then classified as "allele-specific" for either the reference or mutant alleles (i.e., reference-only, Q325R-only, or Q325K-only ChIP-Seq peaks). To rank shared peaks with respect to their ChIP-Seq computed significance (*P*-value) of enrichment, we averaged the rankings from the two reciprocal pairwise comparisons. We trimmed the peaks to encompass a maximum of 500 bp by using the inferred position of the peak center provided by the original dataset. To compute the AUROC and the AUPR of each ChIP-Seq peak subset, we used the top 1000 peaks (ranked by ChIP-Seq computed significance (*P*-value) of enrichment) as the foreground set; to construct the background set, we generated an equally sized set of 1000 permuted sequences with identical dinucleotide frequencies and lengths. Each sequence in the foreground and background sets was then assigned a score corresponding to the E-score for its top-scoring 8-mer among the preferred 8-mers for each allele. The AUROC statistic was obtained by calculating sensitivity and specificity values as the E-score threshold for predicting that a sequence belonged in the foreground set (i.e., that the sequence was bound by that allele) was varied between 0.3 and 0.5. We calculated the precision and sensitivity values using a similar approach to obtain the AUPR statistic. The *P*-values associated with each AUROC and AUPR value were calculated by using a Wilcoxon signed-rank test comparing the scores for foreground and background sequences.

Enrichment of putative target genes regulated directly by the HOXD13 variants

To investigate the impact on gene expression caused by differences in DNA binding sites recognized by different HOXD13 alleles, we first quantified the enrichment of allele-specific and allele-shared ChIP-Seq peaks near genes that exhibited differential expression in cells in which the different HOXD13 alleles were overexpressed. We focused our analyses on the direct binding of each HOXD13 allele to its preferential binding sites by filtering ChIP-Seq peaks based on the presence of 8-mers preferentially bound by the same allele, as described above.

We first calculated the percentage of unique PBM-derived allele-preferred 8-mers found within the sequences of the top 1000 allele-specific ChIP-Seq peaks, ranked by ChIP-Seq computed significance (*P*-value) of enrichment (red and blue boxplots, Fig. S10A, S10B, S10D, S10E). Specifically, for each allele, we calculated the percentage of the unique, PBM-derived allele-preferred 8-mers found in the associated allele-specific ChIP-Seq peaks (e.g., reference-preferred 8-mers found within the reference-only ChIP-Seq peaks (red boxplot, Fig. S10A)) versus the set of unique, PBM-derived allele-preferred 8-mers of the other allele found in the same ChIP-Seq peaks (e.g., Q325K-preferred 8-mers found within the reference-only ChIP-Seq peaks (blue boxplot, Fig. S10A)). We used the mid-point between the 25th percentile of the former distribution and the 75th percentile of the latter distribution as our threshold (blue dashed horizontal line in Fig. S10) to infer allele-specific ChIP-Seq peaks enriched ($P < 2.2 \times 10^{-16}$, one-tailed Wilcoxon signed-rank test) for unique PBM-derived allele-preferred 8-mers of the same allele as compared to those of the other allele ("allele-preferred directly bound peaks"). Similarly, we identified shared ChIP-Seq peaks enriched for PBM-derived allele-common 8-mers as compared to PBM-derived allele-preferred 8-mers bound preferentially by either the reference or mutant allele in each pairwise comparison (i.e., reference versus Q325K allele, and reference versus Q325R allele) (black and violet boxplots in Fig. S10C, S10F).

We compiled the putative target genes associated with each ChIP-Seq peak by identifying all transcription start sites within the galGal3 genome and their corresponding genes ("proximal genes") within +/- 100 kb of each ChIP-Seq peak. To restrict our analysis to reliably detected genes, we filtered the RNA-seq data for transcripts with a minimum of 1 RPKM in at least one RNA-Seq experiment (i.e., HOXD13 reference or Q325K overexpression experiments). We added a pseudocount of 1 to all RPKM values in the filtered RNA-Seq data, and then calculated differential gene expression for each filtered RNA-Seq data set, using a differential expression threshold of 2-fold change in RPKM values as compared to the mock-infected control.

We then used the top 1000 peaks (ranked by ChIP-Seq peak computed significance (*P*-value) of enrichment) from each set of "reference-preferred directly bound peaks", "mutant-preferred directly bound peaks", and "allele-common directly bound peaks" to evaluate the enrichment of their respective putative target genes within the up- or down-regulated differentially expressed genes from overexpression of HOXD13 reference or mutant alleles (*12*) (Fig. 3D and Fig. S8C). By random circular permutation of transcript IDs (*54*), we created 100 random background gene sets to compute empirical *Z*-scores for the enrichment of differentially expressed genes associated with allele-preferred directly bound peaks; briefly, we did 100 circular permutations of the transcript ID column in the RNA-Seq RPKM table (with minimum RPKM = 1) to create 100 'shuffled' gene expression datasets, which we then used to identify mock sets of 'differentially expressed' genes (>2 fold-change) for each 'shuffled' gene expression dataset. We checked the normality of the distributions of the background gene sets using Shapiro-Wilk tests and Q-Q plots; all were normal or approximately normal. We calculated empirical P-values using a permutation test, identifying the rank of the actual test statistic (the actual number of proximal genes that were differentially expressed from

overexpression of HOXD13 reference or mutant alleles) within the distribution of test statistics computed for each of 100 background gene sets.

Identification of co-expressed paralogs and loss-of-function (LoF) tolerant genes

Paralogous gene pairs were identified using annotations from the Duplicated Genes Database (DGD) (February 25, 2015 release) (*8*). Any pair of human genes belonging to the same homology group, as defined by DGD, was considered to be paralogous. Co-expression was determined using the Hsa.v13 dataset obtained from COXPRESdb (*55*). COXPRESdb provides a matrix of Pearson correlation coefficients quantifying the similarity of expression pattern of gene pairs across a wide range of tissues. We identified gene pairs as being co-expressed when one of the genes was among the 25 genes with the highest correlation coefficients for the other gene. The results related to co-expressed paralogs were essentially unchanged when the threshold was varied to include the top 50 or top 100 most correlated genes as being co-expressed. Genes that were tolerant of LoF mutations were defined based on the results of Sulem *et al.* (*21*). Briefly, a gene was considered LoF-tolerant if at least one of the individuals studied was reported as being a homozygote or a complex heterozygote for frameshift or nonsense variants.

Statistical testing of DBDP enrichment in TF subsets

For each gene, we calculated the number of predicted base- or backbone-contacting residues that were altered by at least one genetic variant. To account for the fact that TFs can have different numbers of DNA-contacting residues, we normalized the number of residues affected by genetic variation by dividing by the number of DNA-contacting residues in each TF. We used a two-sample permutation test to determine whether certain subsets of TFs had a higher fraction of variable residues than others (e.g., genes with co-expressed paralogs vs. those without). For all permutation tests, we used the 'permTS' function in the *perm* R package with standard parameter values.

To determine whether the observed enrichments were statistically independent (e.g., not due to genes with co-expressed paralogs often being tolerant of LoF mutations), we fitted a standard linear regression model ('lm' function in R) with the fraction of variable DNA-contacting residues as the dependent variable, and binary values representing LoF-tolerance, the presence of a co-expressed paralog, and their interaction as independent variables (see statistical formula below). Both LoF-tolerance and paralog presence were highly significant predictive features independently ($P < 10^{-5}$, t-test), while the interaction term was not significant ($P = 0.296$, t-test).

```
Fraction.Variable.Residues ~  LoF.Tolerant +
Paralog.Present + LoF.Tolerant:Paralog.Present
```


1. E. A. Consortium, *bioRxiv*, (2015).
2. G. R. Abecasis *et al.*, *Nature* **467**, 1061-1073 (2010).
3. H.-J. Westra *et al.*, *Nat Genet* **45**, 1238-1243 (2013).
4. A. Veraksa, M. Del Campo, W. McGinnis, *Mol Genet Metab* **69**, 85-100 (2000).
5. J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, N. M. Luscombe, *Nat Rev Genet* **10**, 252-263 (2009).

6. Materials and methods are available as supplementary materials on *Science* Online.
7. W. Fu *et al.*, *Nature* **493**, 216-220 (2013).
8. U. Consortium., *Nucleic Acids Res* **43**, D204-212 (2015).
9. I. A. Adzhubei *et al.*, *Nat Meth* **7**, 248-249 (2010).
10. P. C. Ng, S. Henikoff, *Nucleic Acids Res* **31**, 3812-3814 (2003).
11. D. Welter *et al.*, *Nucleic Acids Res* **42**, D1001-1006 (2014).
12. D. M. Ibrahim *et al.*, *Genome Res.* **23**, 2091-2102 (2013).
13. M. F. Berger *et al.*, *Nat Biotech* **24**, 1429-1435 (2006).
14. B. Jiang, J. S. Liu, M. L. Bulyk, *Bioinformatics* **29**, 1390-1398 (2013).
15. J. I. Fuxman Bass *et al.*, *Cell* **161**, 661-673 (2015).
16. C. L. Freund *et al.*, *Cell* **91**, 543-553 (1997).
17. A. Swaroop *et al.*, *Hum. Mol. Genet.* **8**, 299-305 (1999).
18. P. K. Swain *et al.*, *Neuron* **19**, 1329-1336 (1997).
19. N. Brison, P. Debeer, P. Tylzanowski, *Dev Dyn* **243**, 37-48 (2014).
20. V. Salsi, M. A. Vigano, F. Cocchiarella, R. Mantovani, V. Zappavigna, *Dev Biol* **317**, 497-507 (2008).
21. P. Sulem *et al.*, *Nat Genet* **47**, 448-452 (2015).
22. M. Ouedraogo *et al.*, *PLoS One* **7**, e50653 (2012).
23. T. Lappalainen, S. B. Montgomery, A. C. Nica, E. T. Dermitzakis, *Am J Hum Genet* **89**, 459-463 (2011).
24. F. Cunningham *et al.*, *Nucleic Acids Res* **43**, D662-D669 (2015).
25. X. Guo, M. L. Bulyk, A. J. Hartemink, *Pac Symp Biocomput*, 104-115 (2012).
26. S. R. Eddy, *Bioinformatics (Oxford, England)* **14**, 755-763 (1998).
27. X. Liu, X. Jian, E. Boerwinkle, *Hum. Mutat.* **32**, 894-899 (2011).
28. C. The UniProt, *Nucleic Acids Res* **36**, D190-D195 (2008).
29. E. Boura, L. Rezabkova, J. Brynda, V. Obsilova, T. Obsil, *Acta Crystallogr. D Biol. Crystallogr.* **66**, 1351-1357 (2010).
30. M. M. Brent, R. Anand, R. Marmorstein, *Structure* **16**, 1407-1416 (2008).
31. K. L. Clark, E. D. Halay, E. Lai, S. K. Burley, *Nature* **364**, 412-420 (1993).
32. F. De Masi *et al.*, *Nucleic Acids Res* **39**, 4553-4563 (2011).
33. D. R. Littler *et al.*, *Nucleic Acids Res* **38**, 4527-4538 (2010).
34. J. C. Stroud *et al.*, *Structure* **14**, 159-166 (2006).
35. K.-L. Tsai *et al.*, *J. Biol. Chem.* **281**, 17400-17409 (2006).
36. K.-L. Tsai *et al.*, *Nucleic Acids Res* **35**, 6984-6994 (2007).
37. S. A. Wolfe, L. Nekludova, C. O. Pabo, *Annual Review of Biophysics and Biomolecular Structure* **29**, 183-212 (2000).
38. N.-L. Sim *et al.*, *Nucleic Acids Res* **40**, W452-457 (2012).
39. S. Chun, J. C. Fay, *Genome Res.* **19**, 1553-1561 (2009).
40. J. M. Schwarz, C. Rödelsperger, M. Schuelke, D. Seelow, *Nat Meth* **7**, 575-576 (2010).
41. B. Reva, Y. Antipin, C. Sander, *Nucleic Acids Res* **39**, e118 (2011).
42. A. Olatubosun, J. Väliaho, J. Härkönen, J. Thusberg, M. Vihinen, *Hum. Mutat.* **33**, 1166-1174 (2012).
43. L. D. Ward, M. Kellis, *Nucleic Acids Res* **40**, D930-D934 (2012).
44. G. Badis *et al.*, *Science* **324**, 1720-1723 (2009).

45. M. F. Berger *et al.*, *Cell* **133**, 1266-1276 (2008).
46. N. Sahni *et al.*, *Cell* **161**, 647-660 (2015).
47. M. F. Berger, M. L. Bulyk, *Nat Protoc* **4**, 393-411 (2009).
48. A. M. Dudley, J. Aach, M. A. Steffen, G. M. Church, *PNAS* **99**, 7554-7559 (2002).
49. G. E. Crooks, G. Hon, J.-M. Chandonia, S. E. Brenner, *Genome Res.* **14**, 1188-1190 (2004).
50. M. T. Weirauch *et al.*, *Cell* **158**, 1431-1443 (2014).
51. Y. Benjamini, Y. Hochberg, *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).
52. M. J. Landrum *et al.*, *Nucleic Acids Res*, (2015).
53. M. T. Weirauch *et al.*, *Nat Biotech* **31**, 126-134 (2013).
54. C. P. Cabrera *et al.*, *G3 (Bethesda)* **2**, 1067-1075 (2012).
55. T. Obayashi *et al.*, *Nucleic Acids Res* **41**, D1014-1020 (2013).

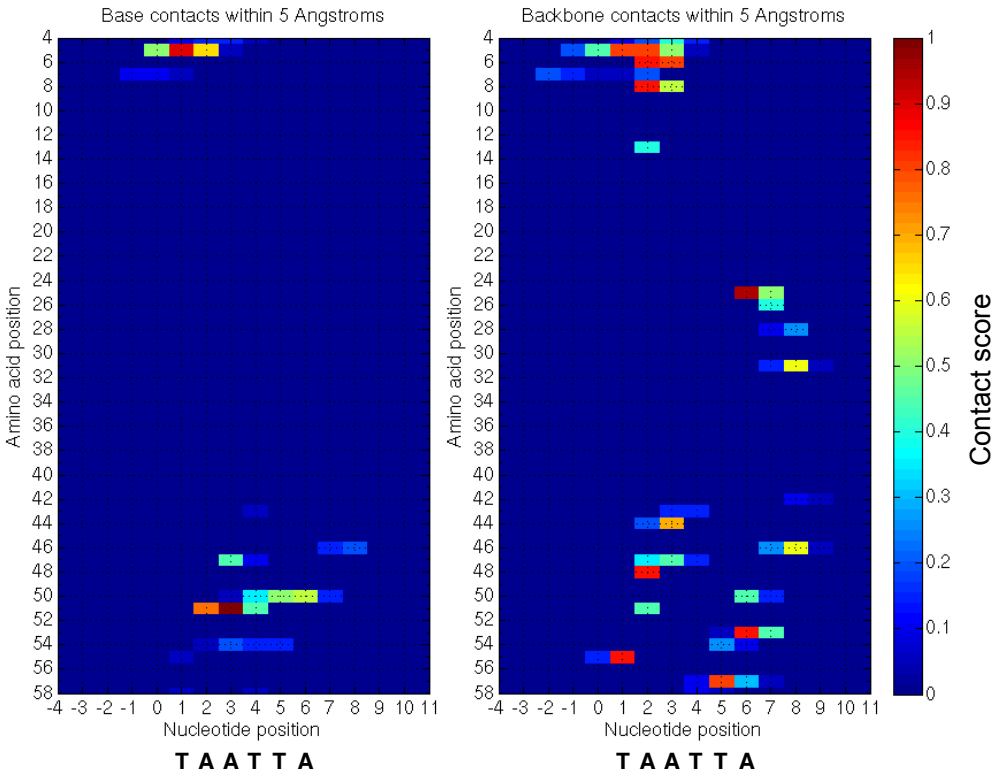

Fig. S1.

Homeodomain contact map derived from co-crystal structures.
Each heatmap shows the domain positions of amino acids identified as making base (*left*) or backbone (*right*) contacts based on a structural analysis of 10 representative homeodomain co-crystal structures. Larger values indicate close contacts (< 3.5 Å) identified with higher frequency in co-crystal structures (Materials and Methods). Amino acid positions were assigned in accordance with the canonical homeodomain numbering scheme (which differs from the Pfam domain numbering by one position; compare to Figure S2). A consensus sequence for all of the aligned DNA binding sites in the 10 co-crystal structures is shown below the heatmaps.

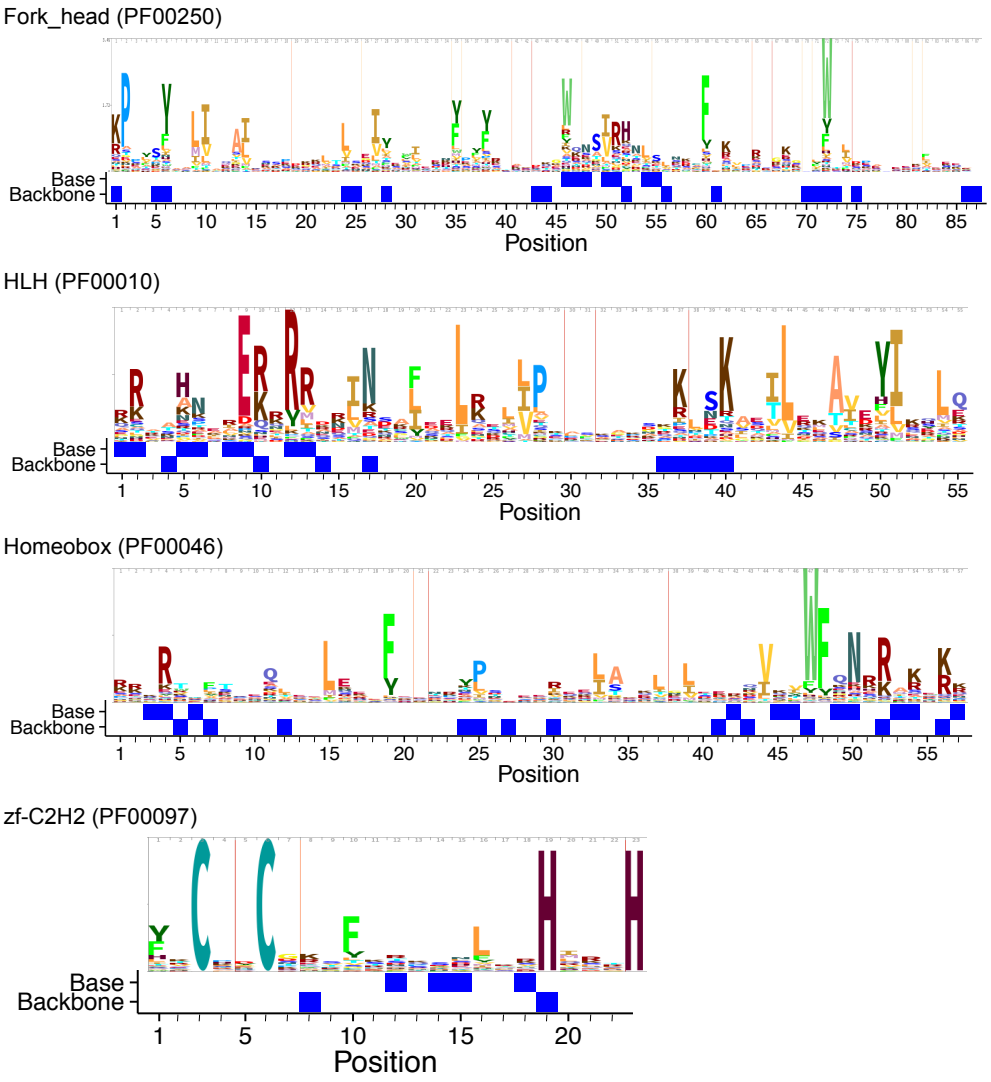

Fig. S2

Positions annotated in Pfam domains as DNA-contacting.
The domain positions annotated as either base- or backbone- contacting are shown for each of the four Pfam domains selected for detailed annotation (zf-C2H2, Homeobox, HLH and Fork_head). Each Pfam domain is represented by its HMM logo, which shows the amino acids that are overrepresented at specific positions within the domain. For details about how positions were annotated, see Materials and Methods.

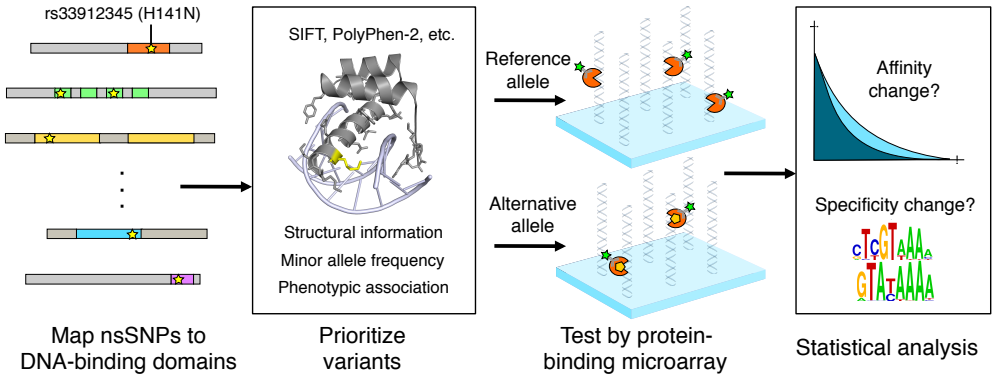

Fig. S3

Schema of study design.

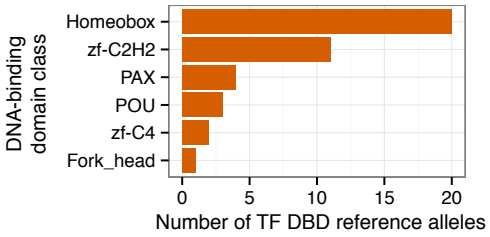

Fig. S4

DBD structural classes assayed by PBMs.

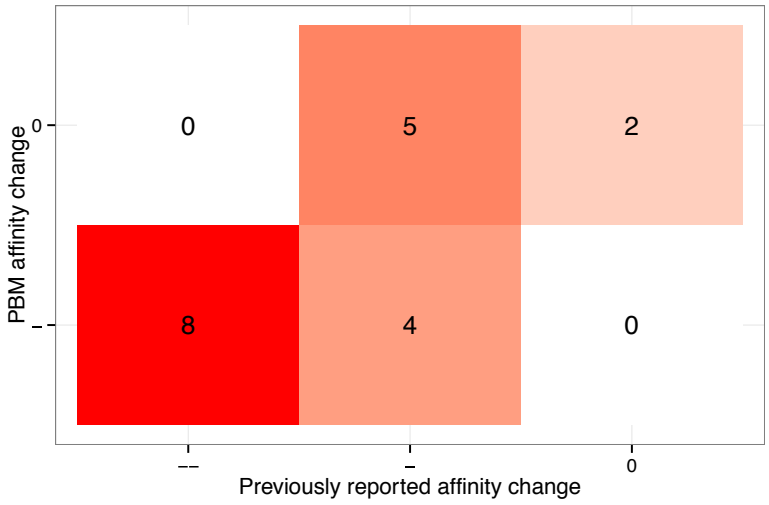

Fig. S5

Comparison with binding affinity changes reported in previous studies.
Numbers and shades of red depict the number of alleles that were identified as having altered affinity by the PBM-based method described here compared to how they were categorized in previous studies using experimental methods such as electrophoretic mobility shift assays. In the x-axis, "--" indicates a complete loss of binding, "-" indicates a partial loss of binding affinity, and "0" indicates no detectable change in binding affinity (Materials and Methods). In the y-axis, "-" indicates an affinity decrease was identified by the PBM-based approach used throughout the text and "0" indicates no affinity change was detected.

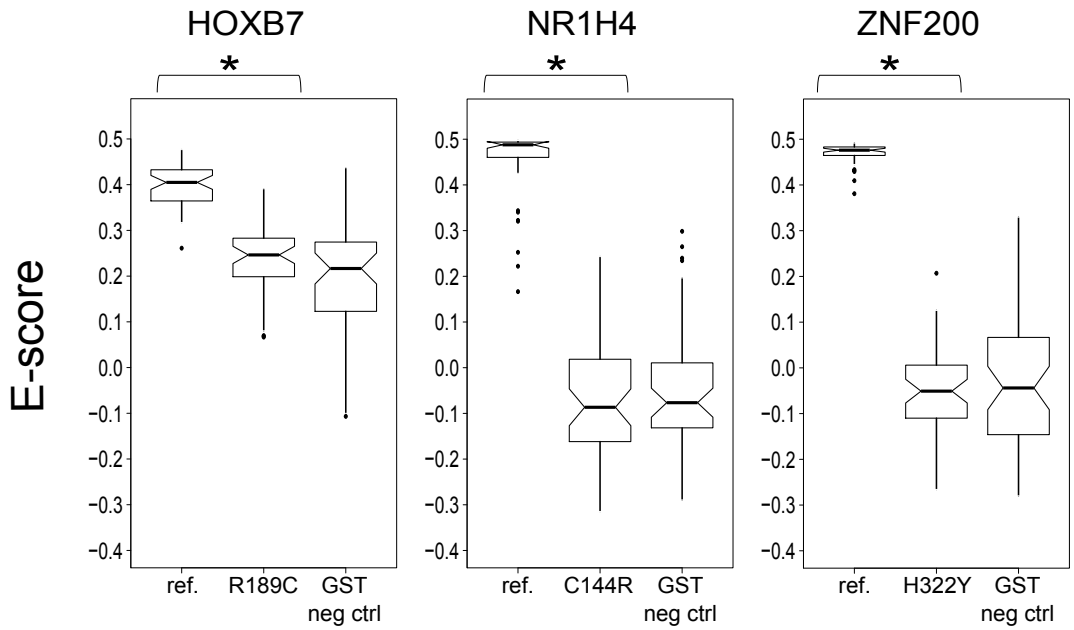

Fig. S6

E-score changes caused by loss-of-function variants.
Each box plot summarizes the differences in E-score distribution observed between reference and alternative alleles for each of the DBDPs identified as causing a complete loss of sequence-specific binding. Box plots are formatted and P-values were calculated as in Fig. 2C, except that the top 50 8-mers evaluated within each allelic series were for each corresponding reference allele. Indicated pairwise comparisons (*) yielded P-values $< 10^{-16}$ (Mann-Whitney U test).

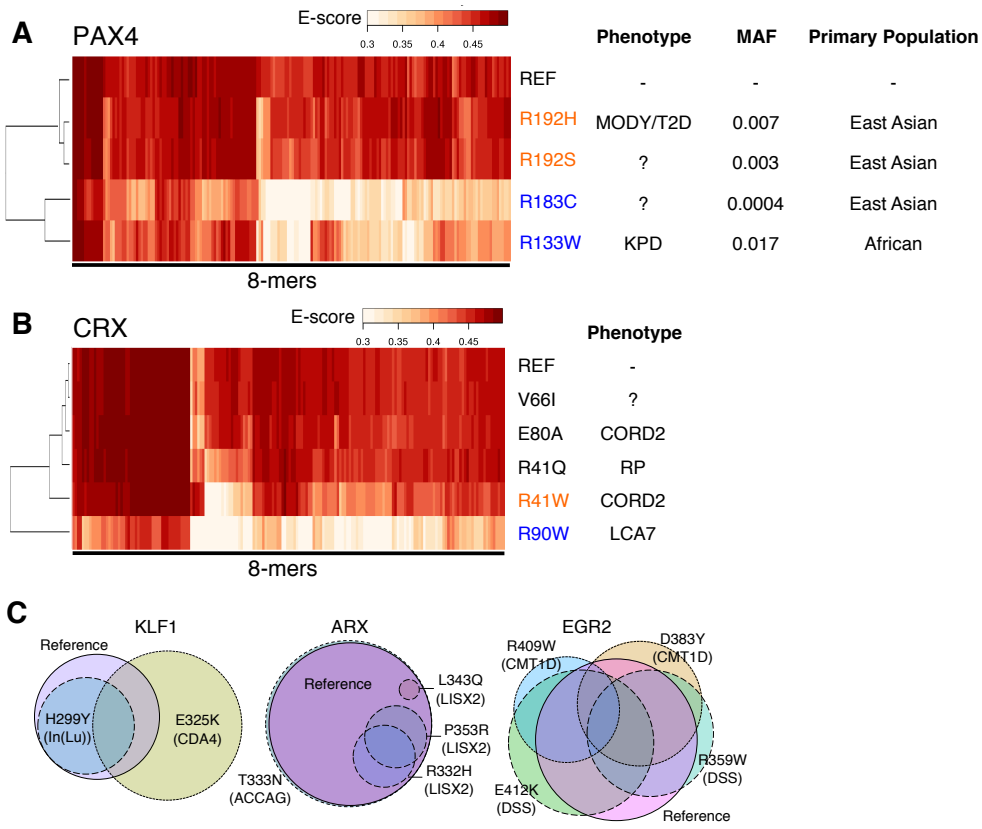

Fig. S7

PBM 8-mer binding profiles reveal a range of DNA-binding perturbations caused by different variants in TF allelic series.

(**A,B**) Heatmaps depicting PBM E-scores (columns) of DBD alleles (rows) for all 8-mers bound strongly (E > 0.45) by at least one allele within each allelic series. Rows and columns were clustered hierarchically. Variants in blue or orange font exhibited altered DNA binding affinity or specificity, respectively. The CRX V66I variant, predicted to be benign, did not alter affinity or specificity. (**A**) Binding profiles of PAX4 with corresponding phenotypes ("?" if unknown), ExAC minor allele frequencies (MAF) and population where allele is most prevalent. PAX4, a paired homeobox TF essential for formation of beta-cells during pancreatic islet development, has been associated with diabetes (Y. Shimajiri et al., Diabetes 50, 2864-2869 (2001)). Based on the severity of their effects on 8-mer binding profiles as compared to disease-associated variants (Fig. S7A), we propose that the R192S and R183C variants are pathogenic in diabetes. (**B**) CRX allelic series with corresponding phenotypes. (**C**) Venn diagrams depicting 8-mers shared across alleles within the KLF1, ARX, and EGR2 allelic series. Within each series, the circle size is proportional to number of 8-mers bound (E > 0.4) by each allele. Purple circles with solid outlines indicate reference alleles. Disease associated with each variant is abbreviated (see Table S6 for details).

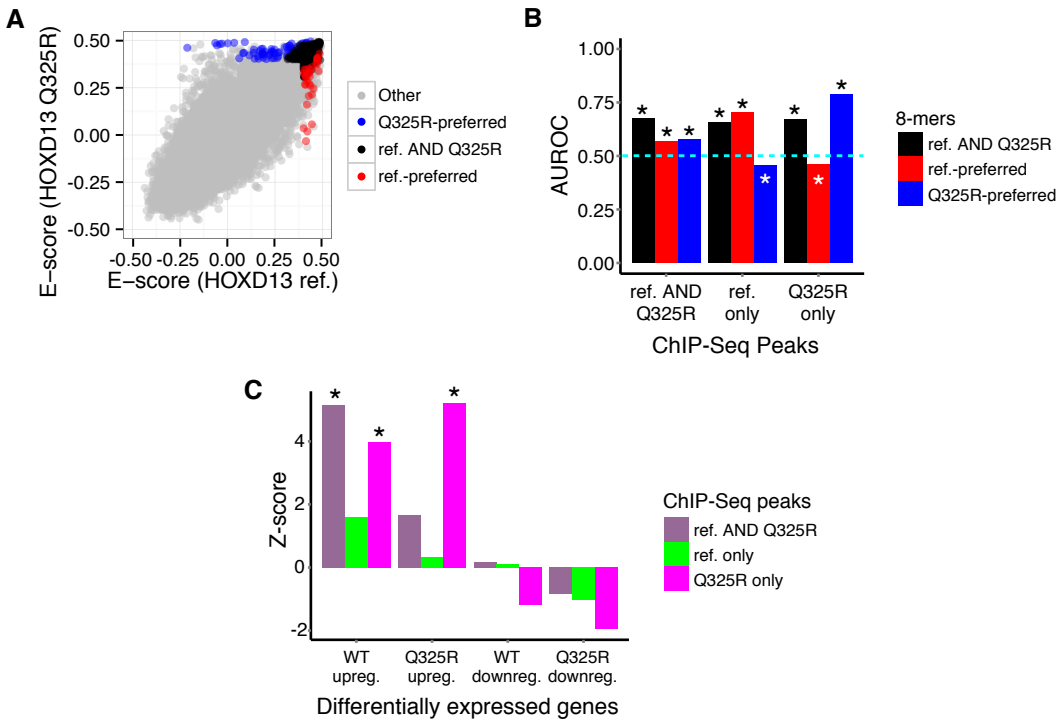

Fig. S8

Perturbations in TF DNA-binding and gene expression associated with HOXD13 Q325R allele.

(**A**) Scatter plot comparing 8-mer E-scores of HOXD13 reference versus Q325R alleles. Allele-preferred and allele-common 8-mers (see Materials and Methods) are colored. (**B**) PBM-derived allele-preferred 8-mers are enriched (black asterisks) or depleted (white asterisks) (* $P < 0.01$, Wilcoxon signed-rank test) within genomic regions bound in vivo exclusively by the respective allele. Dashed horizontal line indicates AUROC = 0.5 (no enrichment or depletion). (**C**) Genes associated with ChIP-Seq peaks enriched for reference- versus Q325R-preferred 8-mers are over-represented (* $P < 0.01$, permutation test) among genes up-regulated by the same allele. Z-scores were calculated using 100 random background gene sets (see Materials and Methods). ChIP-Seq and RNA-Seq data are from Ibrahim *et al*., *Genome Res*., 2013 (*12*).

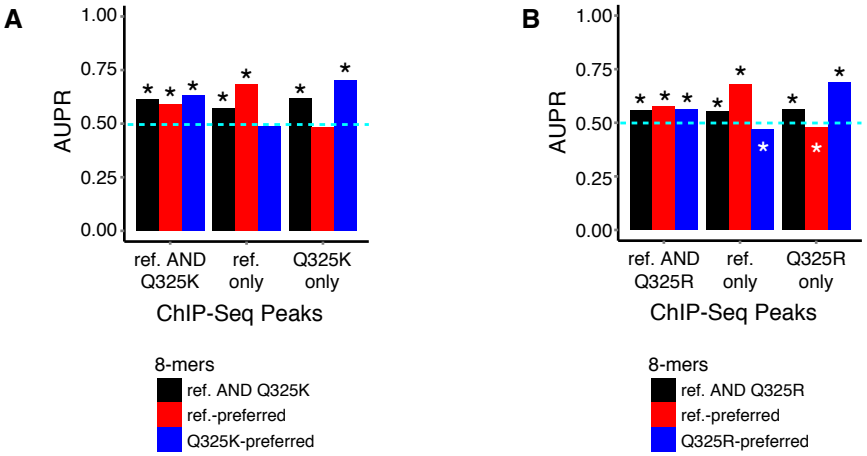

Fig. S9

The enrichment of allele-preferred 8-mers in the respective allele-specific ChIP-Seq peaks is robust to different evaluation metrics.
PBM-derived Q325K-preferred and Q325R-preferred 8-mers are enriched (black asterisks) (* $P < 0.01$, Wilcoxon signed-rank test) within genomic regions bound in vivo exclusively by HOXD13 Q325K (**A**) or Q325R (**B**) alleles, respectively, when area under the Precision-Recall curve (AUPR) is assessed. Dashed horizontal line indicates AUPR = 0.5 (no enrichment or depletion). White asterisks indicate depletion of allele-preferred 8-mers (* $P < 0.01$, Wilcoxon signed-rank test). ChIP-Seq data are from Ibrahim *et al*., *Genome Res*., 2013 (*12*).

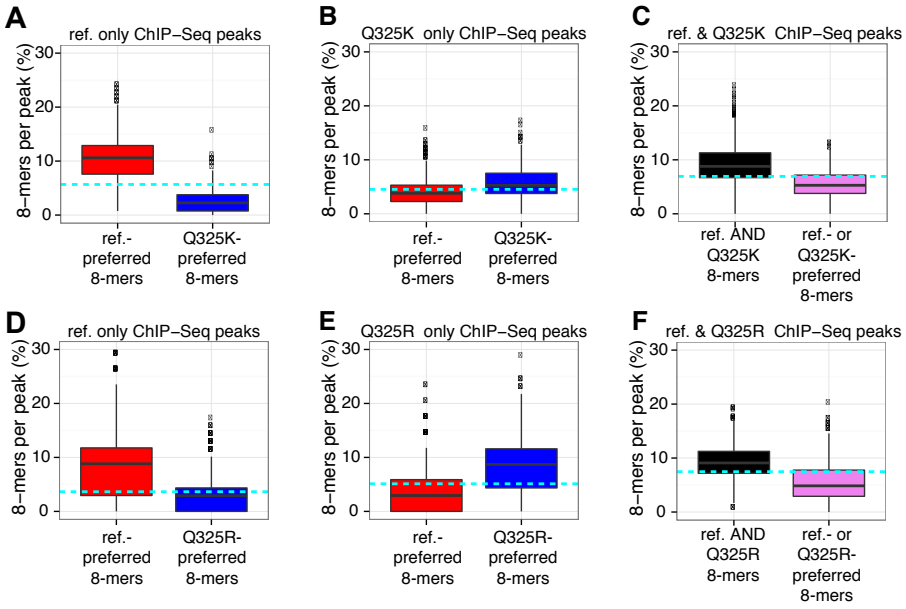

Fig. S10

Selection of thresholds for defining HOXD13 allele-preferred directly bound ChIP-Seq peaks.

Box plots showing the distributions of (**A**) reference-preferred 8-mers and Q325K-preferred 8-mers in ChIP-Seq peaks bound only by the reference HOXD13 protein; (**B**) reference-preferred 8-mers and Q325K-preferred 8-mers in ChIP-Seq peaks bound only by the HOXD13 Q325K variant; (**C**) 8-mers bound by both the reference and Q325K alleles, and the 8-mer set comprising reference-preferred or Q325K-preferred 8-mers, in ChIP-Seq peaks shared by both the HOXD13 wild-type and Q325K variant proteins; (**D**) reference-preferred 8-mers and Q325R-preferred 8-mers in ChIP-Seq peaks bound only by the reference HOXD13 protein; (**E**) reference-preferred 8-mers and Q325R-preferred 8-mers in ChIP-Seq peaks bound only by the HOXD13 Q325R variant; (**F**) 8-mers bound by both the reference and Q325R alleles, and the 8-mer set comprising reference-preferred or Q325R-preferred 8-mers, in ChIP-Seq peaks shared by both the HOXD13 wild-type and Q325R variant proteins. For each box plot we analyzed the top 1000 ChIP-Seq peaks, ranked by computed significance (*P*-values) of enrichment. The blue dashed line in each set of box plots indicates the mid-point between the bottom quartile of the distribution of PBM-derived allele-preferred 8-mers found in the associated allele-specific ChIP-Seq peaks (e.g., reference-preferred 8-mers found within the reference-only ChIP-Seq peaks in Fig. S10A) and the top quartile of the distribution of unique, PBM-derived allele-preferred 8-mers of the other allele found in the same ChIP-Seq peaks (e.g., Q325K-preferred 8-mers found within the reference-only ChIP-Seq peaks in Fig. S10A) We used these mid-points as thresholds for identifying ChIP-Seq peaks enriched for PBM-derived allele-preferred 8-mers of the same allele as compared to PBM-derived allele-preferred 8-mers of the other allele ("allele-preferred directly bound peaks") (see Materials and Methods). ChIP-Seq data are from Ibrahim *et al*., *Genome Res*., 2013 (*12*).

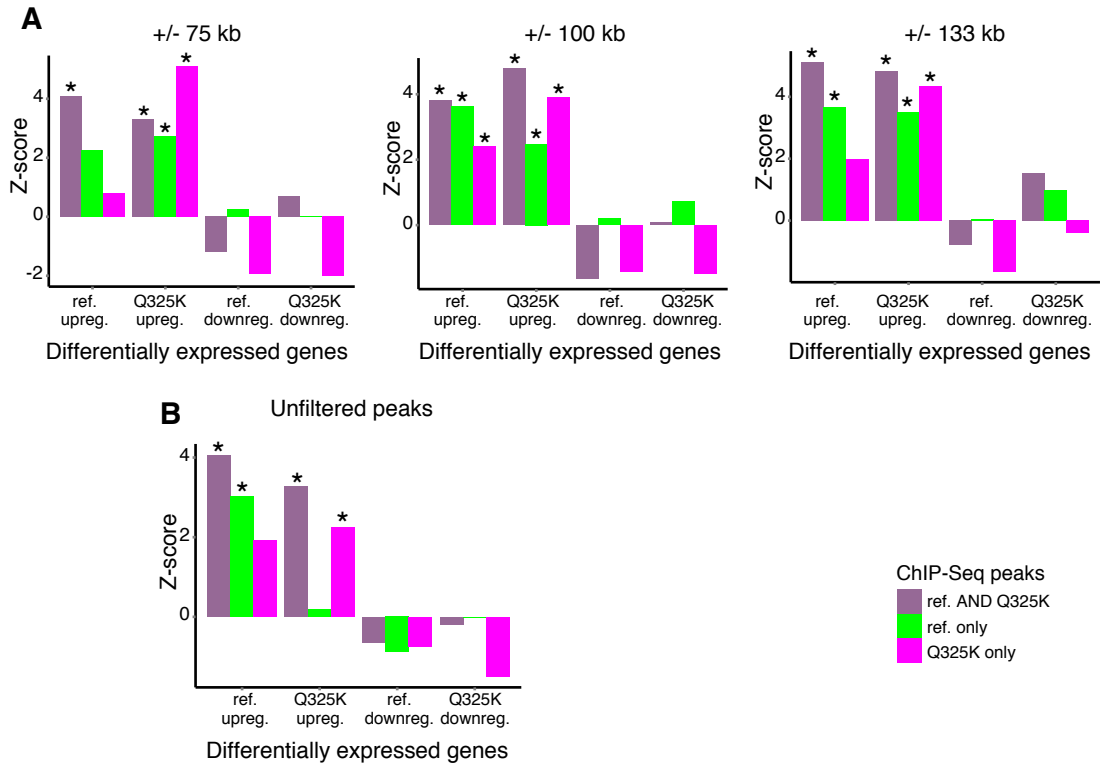

Fig. S11

Enrichment of genes associated with HOXD13 Q325K-preferred directly bound ChIP-Seq peaks, within the set of up-regulated genes in the HOXD13 mutant (Q325K) overexpression experiment, is robust across different ChIP-Seq peak parameter values. (**A**) Z-scores showing degree of enrichment (* $P < 0.01$, permutation test) among genes up-regulated by the same allele, of ChIP-Seq peak-associated genes within the up-regulated genes from wild-type HOXD13 and HOXD13 Q325K overexpression experiments. We analyzed all genes with transcription start sites within +/-75 kb (left panel), +/-100 kb (middle panel; same plot as Figure 3D in main body) and +/-133 kb (right panel) of each of the peak centers of the top 1000 HOXD13 allele-preferred directly bound ChIP-Seq peaks (ranked by computed significance [*P*-values] of enrichment) associated with their respective ChIP-Seq peaks. (**B**) Z-scores showing degree of enrichment of ChIP-Seq peak-associated genes without filtering peaks for allele-preferred or allele-common 8-mers. We analyzed all genes with transcription start sites within +/-100 kb of the peak centers of each of the top 1000 HOXD13 ChIP-Seq peaks (ranked by computed significance [*P*-values] of enrichment) associated with their respective ChIP-Seq peaks. For all analyses, we analyzed the enrichment of peak-associated genes within sets of differentially expressed genes in the HOXD13 overexpression experiments by computing Z-scores. Z-scores were calculated using 100 random background gene sets (see Materials and Methods). ChIP-Seq and RNA-Seq data were from Ibrahim *et al*., *Genome Res*., 2013 (*12*).

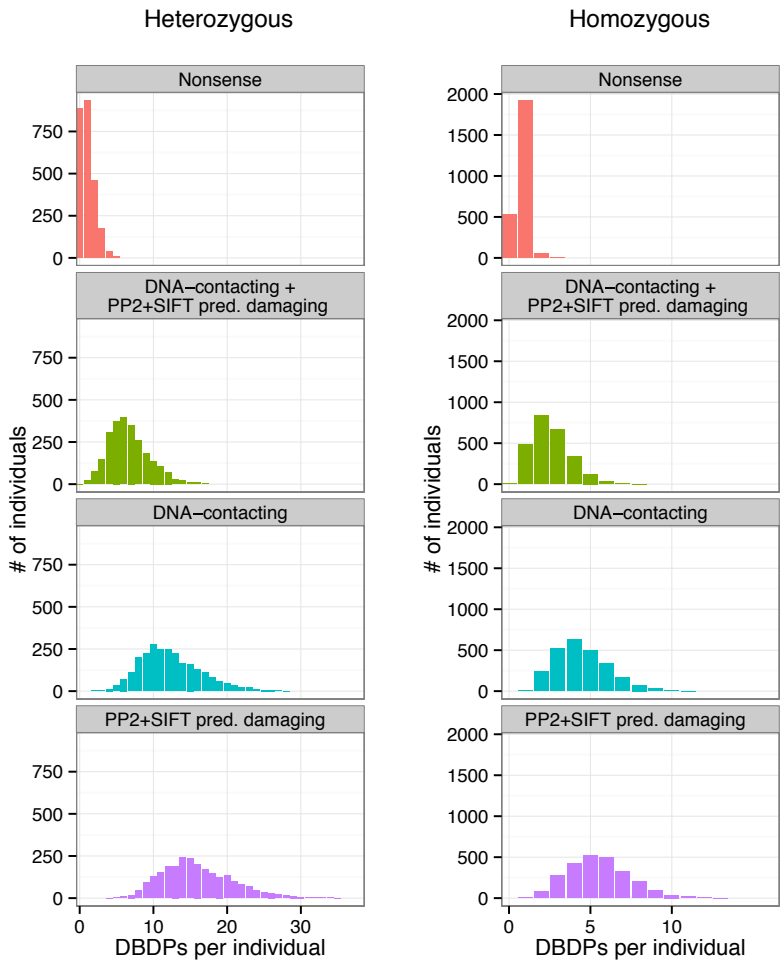

Fig. S12

Predicted damaging variants per individual in homozygous and heterozygous states. For each of the categories of DBD variants shown in Fig. 4B, the number of variant alleles observed per individual (1000 Genomes Project, Phase 3) is shown, separated by whether they were found in a heterozygous (left panels) or homozygous (right panels) state. “PP2+SIFT pred. damaging” indicates variants called “probably damaging” by PolyPhen-2 and “damaging” by SIFT.

Table S5.

Each row shows a comparison between a previously measured change in affinity for a TF allele and the PBM-based determination of affinity changes for the best replicate experiment performed for the same allele. The affinity change q-value derived from the PBM-based method is shown, as well as whether the difference was considered statistically significant ("-") or not ("0"). The last two columns show the magnitude of the previously reported effect as well as the study where the change was reported. A label of "--" indicates a complete loss of binding, "-" indicates a partial loss of binding affinity, and "0" indicates no detectable change in binding affinity (Materials and Methods).

Table S5. Comparison of PBM-based determination of affinity changes to affinity change data compiled from prior studies.

Each row shows a comparison between a previously measured change in affinity for a TF allele and the PBM-based determination of affinity changes for the best replicate experiment performed for the same allele. The affinity change q-value derived from the PBM-based method is shown, as well as whether the difference was considered statistically significant ("-") or not ("0"). The last two columns show the magnitude of the previously reported effect as well as the study where the change was reported. A label of "--" indicates a complete loss of binding, "-" indicates a partial loss of binding affinity, and "0" indicates no detectable change in binding affinity (Materials and Methods).

TF	Allele	Replicate	Affinity change	q-value	Reported effect	Source	Experiment type
ARX	L343Q	R1	-	1.49E-14	--	Cho et al., Neurogenetics (2012)	Luciferase assay
ARX	P353R	R1	-	1.49E-14	-	Cho et al., Neurogenetics (2012)	Luciferase assay
ARX	T333N	R1	0	1	-	Cho et al., Neurogenetics (2012)	Luciferase assay
CRX	R41W	R1	0	0.513703155	-	Swain et al., Neuron (1997)	EMSA
CRX	R90W	R1	-	1.49E-14	--	Swaroop et al., Hum. Mol. Genet. (1999)	EMSA
EGR2	R359W	R1	-	0.000824499	-	Warner et al., Human Molecular Genetics (1999)	EMSA
FOXC1	L130F	R1	-	1.96E-14	--	Ito et al., Arch. OphatImol. (2007)	EMSA
GFI1	N382S	R1	-	3.41E-15	--	Person et al., Nature Genetics (2003)	EMSA
HESX1	E149K	R1	0	1	0	McNay et al., J. Clinical Endocrinology & Metabolism (2007)	EMSA
HESX1	R160C	R1	-	6.61E-15	--	Dattani et al., Nature Genetics (1998)	EMSA
HOXD13	Q325R	R1	-	6.75E-06	-	Zhao et al., American Journal of Human Genetics (2007)	Luciferase assay
HOXD13	S316C	R1	0	1	0	Johnson et al., American Journal of Human Genetics (2003)	EMSA
MSX2	R172H	R1	0	1	-	Wilkie et al., Nature Genetics (2000)	EMSA
NKX2-5	R161P	R1	0	1	-	Dentice et al., J. Clinical Endocrinology & Metabolism (2006)	EMSA
NKX2-5	T178M	R1	-	1.05E-14	-	Kasahara et al., J. Clin. Invest. (2000)	EMSA
POU4F3	L289F	R1	-	2.33E-09	--	Collin et al., Human Mutation (2008)	EMSA
PROP1	R99Q	R1	0	0.376993783	-	Vieira et al., J. Clinical Endocrinology & Metabolism (2003)	EMSA
VSX2	R200P	R1	-	3.41E-15	--	Ferda Percin et al., Nature Genetics (2000)	EMSA
VSX2	R200Q	R1	-	3.23E-05	--	Ferda Percin et al., Nature Genetics (2000)	EMSA

Additional Data table S1 (separate file)

Nonsynonymous (missense) SNPs identified in TF DNA-binding domains.
Each of the nsSNPs identified in DNA-binding domains in the 1000 Genomes Project (Phase 3), Exome Sequencing Project 6500 or Exome Aggregation Consortium (ExAC v0.2) datasets is shown in this table. When a SNP causes amino acid substitutions in multiple Ensembl transcript models, it is repeated in subsequent lines, with its predicted effects shown for each. Additional information about variant annotations is provided in the sheet labeled "Notes." Amino acid changes that cannot be assigned to a position within the Pfam domain are labeled "NA".

Additional Data table S2 (separate file)

Transcription factors analyzed for the presence of variants in this study.

Additional Data table S3 (separate file)

Nonsense SNPs identified in TF DNA-binding domains.
Each of the nonsense SNPs identified as DBD-truncating in the 1000 Genomes Project (Phase 3), Exome Sequencing Project 6500 or Exome Aggregation Consortium (ExAC v0.2) datasets is shown in this table. When a nonsense SNP causes DBD truncations in multiple Ensembl transcript models, it is repeated in subsequent lines, with its predicted effects shown for each. The positions along the amino acid sequence and structural classes of truncated DBDs are shown in the last column, with each domain separated by a semicolon.

Additional Data table S4 (separate file)

PBM experimental conditions and clone sequences.
Summary of all experiments performed, including TF concentration, buffer used, and the amino acid sequence encoded by the PDEST15 vector used in each experiment (not including the GST tag). In buffers including “+Zn”, zinc acetate was added to achieve a final zinc ion concentration of 50 μM.

Additional Data table S6 (separate file)

Specificity and affinity changes identified by PBMs.
Summary of affinity and specificity changes observed for each allele. The first sheet contains Mendelian variants and the second contains nsSNPs along with the associated phenotypes, SNP IDs and rationale for their selection. In each row, the following information is shown: the number of 8-mers bound at an E-score > 0.45 by the reference and alternative alleles, whether the variant changed affinity ("+" if increased, "-" if decreased, "0" if no change), the q-value for the affinity change, whether the variant

altered specificity, and if so, how many 6-mers were identified as preferred by the reference or alternative allele, respectively.

Additional Data table S7 (separate file)

PBM 8-mer data from GST-only negative control duplicate PBM experiments using 100 nM and 600 nM GST.