11. R. Sorjamaa *et al.*, *Atmos. Chem. Phys.* **4**, 2107–2117 (2004).
12. S. Ekström, B. Nozière, H. C. Hansson, *Atmos. Chem. Phys.* **9**, 973–980 (2009).
13. M. C. Facchini, M. Mircea, S. Fuzzi, R. J. Charlson, *Nature* **401**, 257–259 (1999).
14. B. Nozière, C. Baduel, J.-L. Jaffrezo, *Nat. Commun.* **5**, 3335 (2014).
15. Materials and methods and supporting analysis of the experimental data are available as supplementary materials on *Science* Online.
16. M. L. Shulman, M. C. Jacobson, R. J. Carlson, R. E. Synovec, T. E. Young, *Geophys. Res. Lett.* **23**, 277–280 (1996).
17. S. Henning *et al.*, *Atmos. Chem. Phys.* **5**, 575–582 (2005).
18. G. Jura, W. D. Harkins, *J. Am. Chem. Soc.* **68**, 1941–1952 (1946).
19. R. H. Moore *et al.*, *J. Geophys. Res. Atmos.* **117**, D00V12 (2012).
20. E. Hammer *et al.*, *Atmos. Chem. Phys.* **14**, 10517–10533 (2014).
21. K. Broekhuizen, P. P. Kumar, J. P. D. Abbatt, *Geophys. Res. Lett.* **31**, L01107 (2004).
22. K. E. H. Hartz *et al.*, *Atmos. Environ.* **40**, 605–617 (2006).
23. M. Hori, S. Ohta, N. Murao, S. Yamagata, *J. Aerosol Sci.* **34**, 419–448 (2003).
24. A. Kristensson, T. Rosenørn, M. Bilde, *J. Phys. Chem. A* **114**, 379–386 (2010).
25. M. Kuwata, W. Shao, R. Lebouteiller, S. T. Martin, *Atmos. Chem. Phys.* **13**, 5309–5324 (2013).
26. M. D. Petters *et al.*, *Tellus B Chem. Phys. Meterol.* **58**, 196–205 (2006).
27. H. Wex *et al.*, *Atmos. Chem. Phys.* **9**, 3987–3997 (2009).
28. C. R. Ruehl, K. R. Wilson, *J. Phys. Chem. A* **118**, 3952–3966 (2014).
29. G. C. Roberts, A. Nenes, *Aerosol Sci. Technol.* **39**, 206–221 (2005).
30. K. C. Young, A. J. Warren, *J. Atmos. Sci.* **49**, 1138–1143 (1992).

ACKNOWLEDGMENTS

This work is supported by the Office of Science Early Career Research Program, through the Office of Energy Research, Office of Basic Energy Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. The continuous-flow streamwise thermal gradient chamber was originally developed by Patrick Chuang and Anthanasios Nenes with support from NASA's Atmospheric Radiation Measurement program.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/351/6280/1447/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S7
Tables S1 to S3
References (*31*–*39*)

18 September 2015; accepted 18 February 2016
10.1126/science.aad4889

PROTEIN EVOLUTION

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

Luis A. Barrera,[1,2,3,4] Anastasia Vedenko,[1]* Jesse V. Kurland,[1]* Julia M. Rogers,[1,2] Stephen S. Gisselbrecht,[1] Elizabeth J. Rossin,[3,5,6] Jaie Woodard,[1,2] Luca Mariani,[1] Kian Hong Kock,[1,7] Sachi Inukai,[1] Trevor Siggers,[1]† Leila Shokri,[1] Raluca Gordân,[1]‡ Nidhi Sahni,[8,9,10]§ Chris Cotsapas,[5,6]|| Tong Hao,[8,9,10] Song Yi,[8,9,10] Manolis Kellis,[4,6] Mark J. Daly,[5,6,11] Marc Vidal,[8,9,10] David E. Hill,[8,9,10] Martha L. Bulyk[1,2,3,6,7,8,12]¶

Sequencing of exomes and genomes has revealed abundant genetic variation affecting the coding sequences of human transcription factors (TFs), but the consequences of such variation remain largely unexplored. We developed a computational, structure-based approach to evaluate TF variants for their impact on DNA binding activity and used universal protein-binding microarrays to assay sequence-specific DNA binding activity across 41 reference and 117 variant alleles found in individuals of diverse ancestries and families with Mendelian diseases. We found 77 variants in 28 genes that affect DNA binding affinity or specificity and identified thousands of rare alleles likely to alter the DNA binding activity of human sequence-specific TFs. Our results suggest that most individuals have unique repertoires of TF DNA binding activities, which may contribute to phenotypic variation.

Exome sequencing studies have identified many nonsynonymous single-nucleotide polymorphisms (nsSNPs) in transcription factors (TFs) (*1*). Genetic variants that alter transcript expression levels have been associated with human disease risk and are widespread in human populations (*2*, *3*). Numerous Mendelian diseases are attributable to mutations in TFs (*4*). Missense SNPs that change the amino acid sequence of TF DNA binding domains (DBDs) might disrupt their DNA binding activities and thus have detrimental effects on their gene regulatory functions. Despite their medical importance, the consequences of coding variation in DBDs for TF function have remained largely unexplored.

We identified 53,384 unique DBD polymorphisms (DBDPs) (table S1) (here, defined as missense variants) in a curated, high-confidence set of 1254 sequence-specific human TFs (*5*, *6*) (table S2) from genotype data for 64,706 individuals encompassing African, Asian, and European ancestries (Fig. 1A) (*1*, *2*, *7*). We also identified 4552 unique nonsense mutations that result in partial or full DBD truncation (table S3).

We found a median of 60 heterozygous and 20 homozygous DBDPs (Fig. 1B) per genome. We found a significant depletion (odds ratio = 3.7, $P = 0.005$, Fisher's exact test) of DBDPs among TFs with known Mendelian disease mutations (*6*, *8*), suggesting that DBDPs in disease-associated TFs have phenotypic consequences.

We developed a computational approach (*6*) to evaluate missense substitutions in TF DBDs for their impact on DNA binding activity. Existing methods for predicting the impact of missense mutations (*9*, *10*) do not adequately consider the roles of residues in protein-DNA interactions, which we reasoned should improve predictions. We first focused on homeodomain DBDs, as most known Mendelian disease mutations in TFs occur in homeodomain proteins. We analyzed homeodomain-DNA cocrystal structures in the Protein Data Bank to assemble a composite protein-DNA "contact map" (fig. S1). As anticipated, residues that contact DNA bases or phosphate backbone, or that immediately neighbor base-contacting residues, are enriched among Mendelian disease mutations ($P < 0.005$, permutation test). In contrast, individuals in the population are depleted for variants at base- or backbone-contacting positions ($P = 0.0134$ or 0.0312, respectively, permutation test) (Fig. 1C). This highlights the value of considering protein-DNA contacts in predicting the impact of variants.

On the basis of these results, we expanded our approach to other TF families. For each variant we considered multiple criteria, including (i) position of the residue relative to the protein-DNA interface in homologous cocrystal structures (fig.

[1]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA. [2]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02138, USA. [3]Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA. [4]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [5]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA. [6]Broad Institute of Harvard and MIT, Cambridge, MA 02139, USA. [7]Program in Biological and Biomedical Sciences, Harvard University, Cambridge, MA 02138, USA. [8]Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, MA 02215, USA. [9]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [10]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [11]Center for Human Genetics Research and Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA 02114, USA. [12]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA.

*These authors contributed equally to this work. †Present address: Department of Biology, Boston University, Boston, MA 02215, USA. ‡Present address: Departments of Biostatistics and Bioinformatics, Computer Science, and Molecular Genetics and Microbiology, Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708, USA. §Present address: Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. ||Present address: Department of Neurology and Department of Genetics, Yale School of Medicine, New Haven, CT 06520, USA. **¶Corresponding author. E-mail: mlbulyk@receptor.med.harvard.edu**

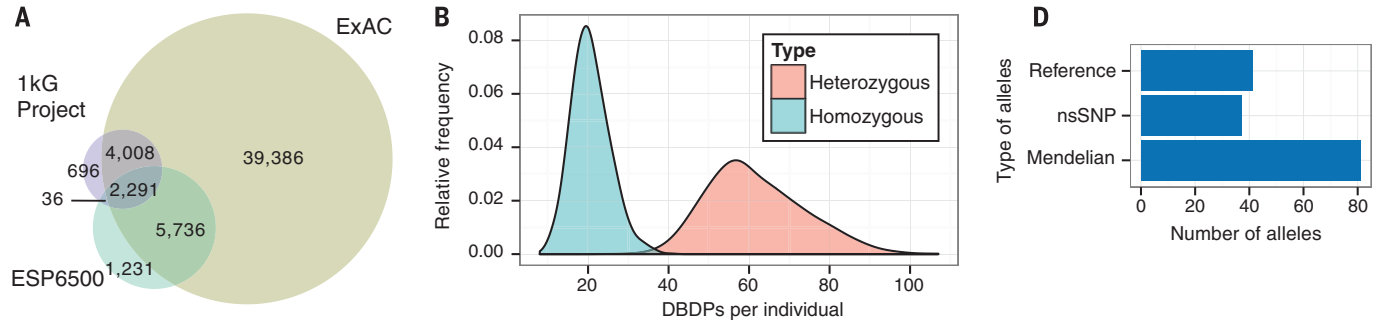
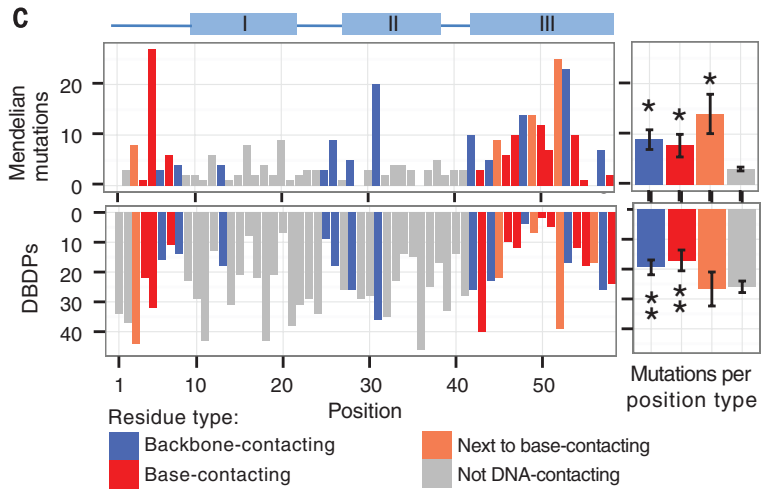
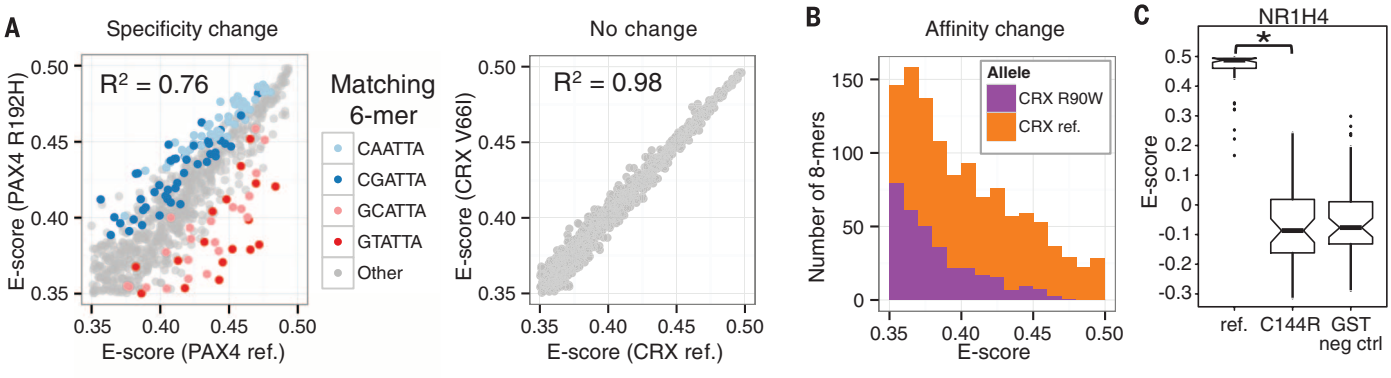


Fig. 1. Evaluation of coding variation in TF DBDs. (**A**) Number of unique DBDPs in 1kG (Phase 3), ESP 6500, or ExAC v0.2 individuals. (**B**) Histograms of unique DBDPs per individual in either homozygous or heterozygous states. (**C**) Number of Mendelian mutations, and nsSNPs found in ExAC, across all homeodomain TFs annotated by their position and type of DNA contact associated with each position. "I," "II," and "III" refer to α helices; III is the DNA-recognition helix. Adjacent bar graphs depict mean number of variants for each position type; enrichment (*) or depletion (**) relative to non–DNA-contacting residues ($P < 0.05$, permutation test); error bars represent 1 SEM, $N = 332$ Mendelian mutations, 1300 nsSNPs, and 11 base-contacting, 12 phosphate-backbone–contacting, 5 neighboring-DNA–contacting, and 30 non–DNA–contacting positions. (**D**) Allele types assayed by PBMs.





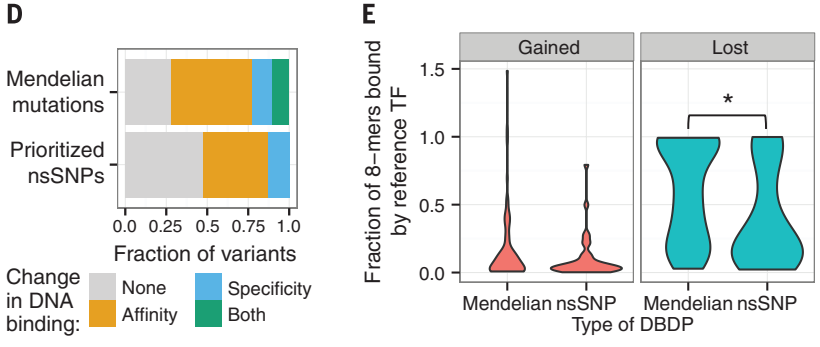


Fig. 2. Perturbed DNA binding caused by nsSNPs or Mendelian disease mutations. (**A**) Specificity change in PAX4 R192H allele (left) compared to no change in CRX V66I allele (right). Colored 6-mers are allele-preferred ($Q < 0.05$, intersection-union test with Benjamini-Hochberg correction). (**B**) Altered E-score distribution of CRX R90W allele relative to the reference allele indicates altered DNA binding affinity. (**C**) Box plots depict E-scores of NR1H4 reference and C144R alleles and glutathione *S*-transferase (GST) negative controls (6) for the top 50 8-mers bound by NR1H4 reference allele. C144R abolished binding specificity (*$P < 2.2 \times 10^{-16}$, Wilcoxon rank-sum test), resulting in E-scores indistinguishable from GST negative controls (table S7). (**D**) Fraction of alleles with observed changes in DNA binding affinity, specificity, both, or neither as determined from PBM binding profiles. Prioritized nsSNPs exclude those predicted as benign by both PolyPhen-2 and SIFT. (**E**) Violin plots depicting fraction of 8-mer binding sites gained or lost by variants relative to the number of 8-mers bound by the reference allele. Gains or losses were defined as $E \geq 0.4$ for one allele and $E < 0.4$ for the other allele. *$P = 0.0044$, Wilcoxon rank-sum test.

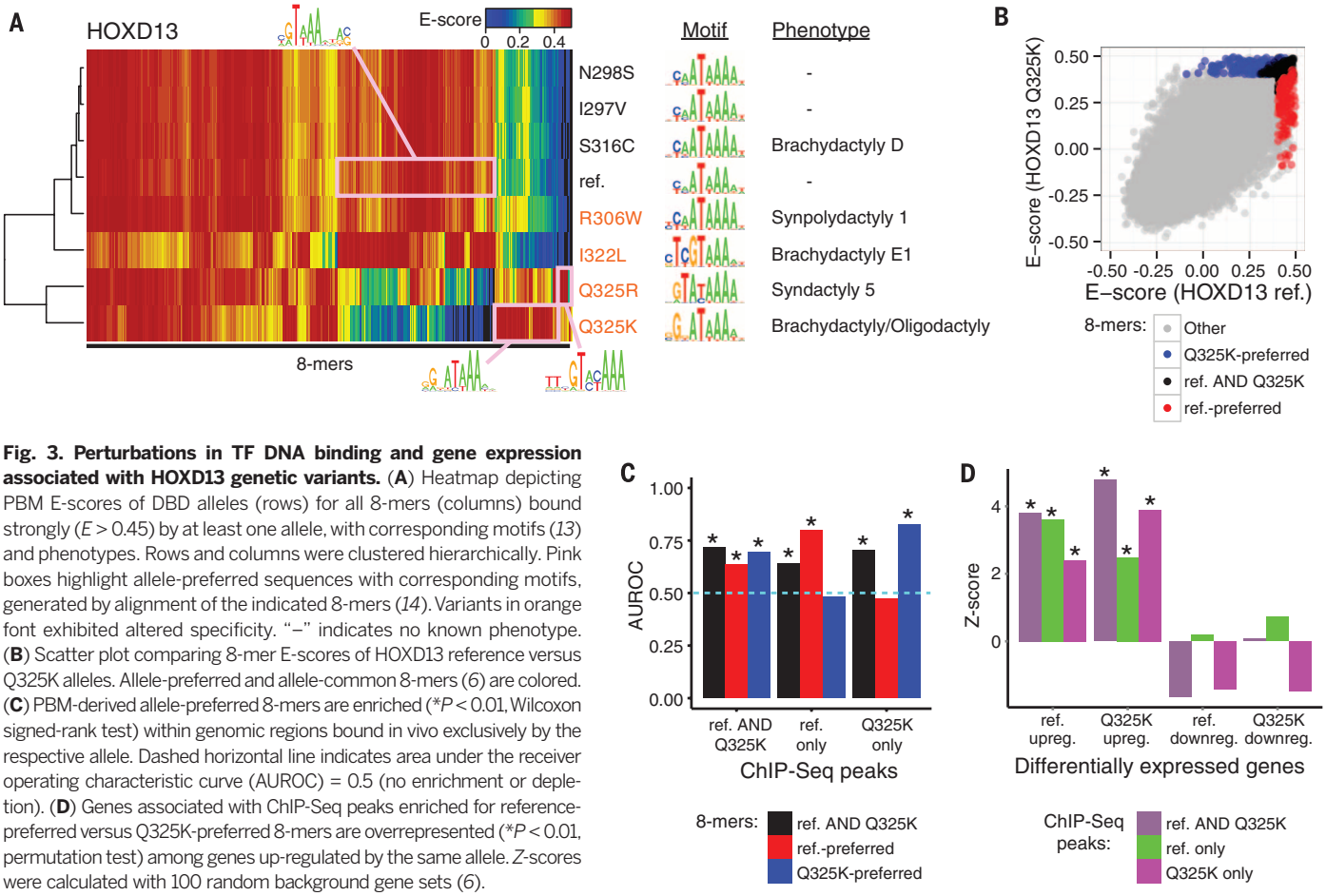


Fig. 3. Perturbations in TF DNA binding and gene expression associated with HOXD13 genetic variants. (**A**) Heatmap depicting PBM E-scores of DBD alleles (rows) for all 8-mers (columns) bound strongly ($E > 0.45$) by at least one allele, with corresponding motifs (*13*) and phenotypes. Rows and columns were clustered hierarchically. Pink boxes highlight allele-preferred sequences with corresponding motifs, generated by alignment of the indicated 8-mers (*14*). Variants in orange font exhibited altered specificity. "–" indicates no known phenotype. (**B**) Scatter plot comparing 8-mer E-scores of HOXD13 reference versus Q325K alleles. Allele-preferred and allele-common 8-mers (*6*) are colored. (**C**) PBM-derived allele-preferred 8-mers are enriched (*$P < 0.01$, Wilcoxon signed-rank test) within genomic regions bound in vivo exclusively by the respective allele. Dashed horizontal line indicates area under the receiver operating characteristic curve (AUROC) = 0.5 (no enrichment or depletion). (**D**) Genes associated with ChIP-Seq peaks enriched for reference-preferred versus Q325K-preferred 8-mers are overrepresented (*$P < 0.01$, permutation test) among genes up-regulated by the same allele. *Z*-scores were calculated with 100 random background gene sets (*6*).

S1); (ii) DNA binding specificity–determining residues for particular DBD classes (fig. S2); (iii) scores from tools that predict mutation pathogenicity (*9*, *10*); (iv) minor allele frequencies; and (v) phenotypic associations from genome-wide association studies (*11*) or known Mendelian disease mutations (*8*).

Using these criteria, we selected 37 TF DBDPs (*6*) to assay for direct, sequence-specific DNA binding activity (fig. S3). These DBDPs were obtained from the 1000 Genomes Project (1kG) Phase 2, the Exome Sequencing Project (ESP 6500), and the Exome Aggregation Consortium (ExAC). To calibrate the effects of these nsSNPs, we selected 80 Mendelian disease mutations, which are known or believed to be pathogenic (Fig. 1D) (*8*, *12*). The 117 variant DBD alleles span six major structural classes, representing 41 distinct TF allelic series (fig. S4). We assayed these 158 DBD alleles using universal protein-binding microarrays (PBMs) (*6*), on which each nonpalindromic 8–base pair sequence (8-mer) occurs on at least 32 spots (*13*) (table S4).

We identified variant-induced changes in DNA binding specificity (*14*) (Fig. 2A) or affinity (Fig. 2B) by comparing the enrichment (E) scores of each of 32,768 nonredundant, ungapped 8-mers represented on universal PBMs to those of the corresponding reference allele (*6*, *13*). DNA binding changes were reproducible across replicate PBM experiments and support previously reported DNA binding affinity differences (table S5 and fig. S5). We categorized all 117 variant alleles as having altered DNA binding specificity, affinity, both, or neither (table S6). Three nsSNPs completely abrogated sequence-specific DNA binding (Fig. 2C and fig. S6). In total, 77 variants altered DNA binding affinity and/or specificity (Fig. 2D). Several nsSNPs predicted to be damaging but not scored here as having altered DNA binding might cause subtle changes beyond the sensitivity of our approach or alternatively affect protein-protein interactions.

Compared to DBDPs, Mendelian disease mutants lost a larger fraction of 8-mers bound by the corresponding reference alleles ($P = 0.0044$, Wilcoxon rank-sum test), consistent with more extreme phenotypes being associated with more drastic in vitro binding changes. The overall difference in gained 8-mers was not significantly different between these two sets of variants ($P = 0.32$, Wilcoxon rank-sum test; Fig. 2E).

PBM binding profiles within an allelic series differed for variants associated with distinct disease phenotypes (fig. S7), supporting results from a yeast one-hybrid screen of Mendelian disease TF mutants (*15*). They also provided molecular insights into the molecular basis of clinical heterogeneity of disease mutations affecting the same genes. For example, *CRX* is associated with Mendelian diseases of retinal degeneration (*16*). The R90W allele, associated with the severe disease Leber congenital amaurosis 7 (*17*), lost the ability to bind most 8-mers bound by wild-type CRX. In contrast, the R41W allele, associated with cone-rod dystrophy 2 (*18*), resulted in a moderate specificity change (fig. S7B).

The 8-mer binding profiles of HOXD13 alleles displayed a range of effects; several of these alleles are associated with various limb malformations (*19*) (Fig. 3A). The I297V and N298S variants, predicted to be benign, did not alter DNA binding activity. The Q325K and Q325R alleles gained recognition of novel motifs, consistent with those learned from chromatin immunoprecipitation with high-throughput sequencing (ChIP-Seq) data (*12*). Allele-preferred 8-mers (Fig. 3B and fig. S8A) are enriched within ChIP-Seq peaks bound exclusively by the respective allele (Fig. 3C and figs. S8B and S9) ($P < 0.01$, Wilcoxon signed-rank test). Putative target genes, associated with ChIP-Seq peaks enriched ($P < 2.2 \times 10^{-16}$, one-tailed Wilcoxon signed-rank test) for Q325K- or Q325R-preferred versus reference-preferred 8-mers (fig. S10) (*6*), are overrepresented among genes up-regulated by the corresponding allele ($P < 0.01$, permutation test) (Fig. 3D and figs. S8C and S11), consistent with HOXD13 acting as a transcriptional activator (*20*). These results suggest that these variants' changes in binding specificity alter genomic

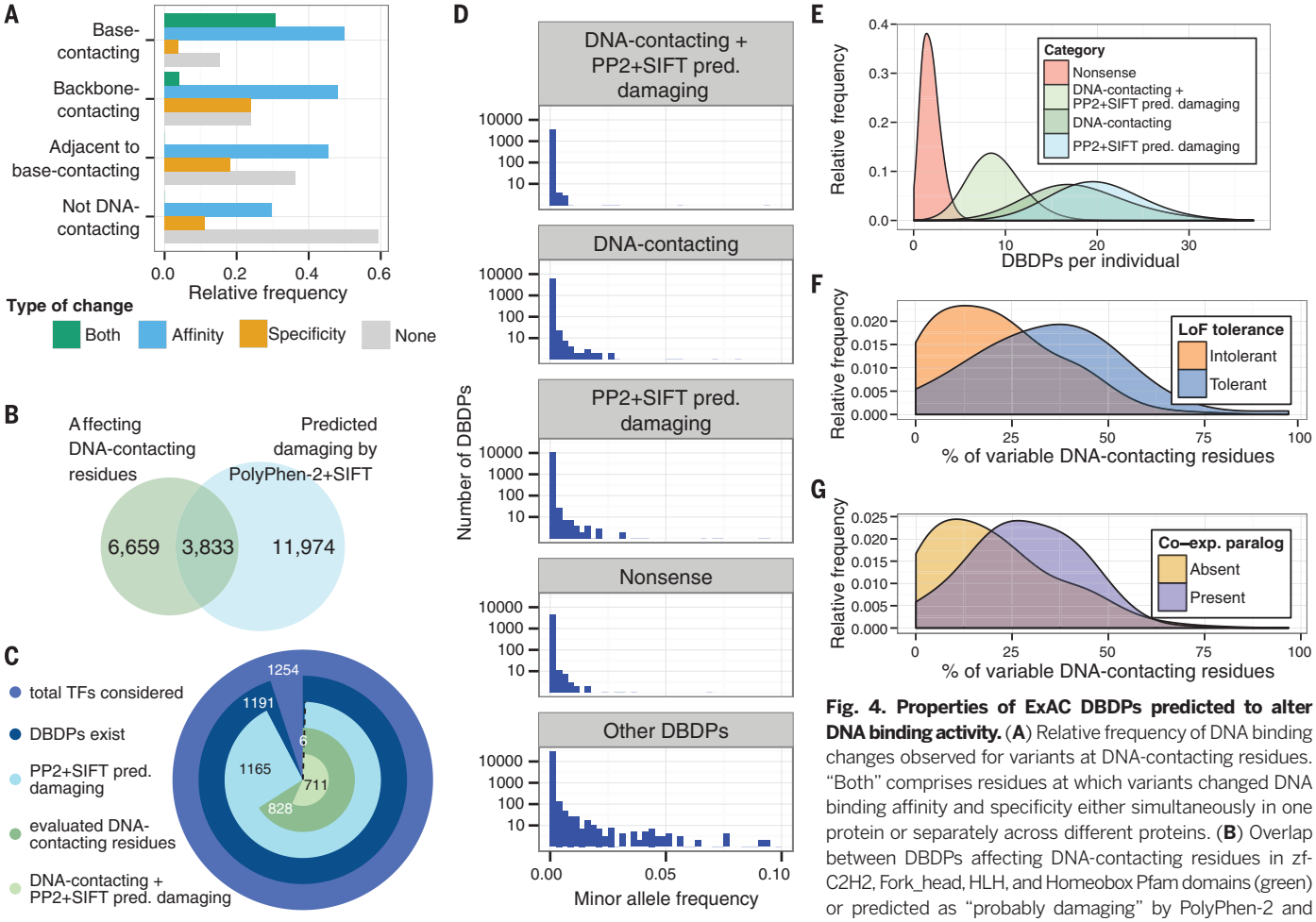


Fig. 4. Properties of ExAC DBDPs predicted to alter DNA binding activity. (**A**) Relative frequency of DNA binding changes observed for variants at DNA-contacting residues. "Both" comprises residues at which variants changed DNA binding affinity and specificity either simultaneously in one protein or separately across different proteins. (**B**) Overlap between DBDPs affecting DNA-contacting residues in zf-C2H2, Fork_head, HLH, and Homeobox Pfam domains (green) or predicted as "probably damaging" by PolyPhen-2 and "damaging" by SIFT (blue). (**C**) Number of sequence-specific TFs for which DBDPs were identified and their evaluation, as in (B). (**D**) Minor allele frequencies (ExAC v0.2) of nsSNPs in DBDs. (**E**) Histogram of DBD variants per individual (1000 Genomes Project Phase 3), annotated as in (C). (**F**) Fraction of DNA-contacting residues per TF altered by at least one nsSNP (ExAC), for genes tolerant of homozygous or compound heterozygous LoF mutations versus genes for which LoF tolerance was not observed (*21*). (**G**) Fraction of variable DNA-contacting residues (ExAC) in TFs with versus without at least one coexpressed paralog.

occupancy, leading to inappropriate gene expression through gained binding sites.

As expected, mutations in residues that either contact DNA or neighbor a base-contacting residue were enriched (odds ratio = 4.3, $P = 0.003$, Fisher's exact test) among DBDPs with altered DNA binding affinity or specificity (Fig. 4A). We also found variants at non–DNA-contacting positions that altered DNA binding, potentially by affecting protein conformation or stability. We identified 3833 unique missense variants that are predicted to be damaging by both PolyPhen-2 (*9*) and SIFT (*10*) and occur at DNA-contacting residues (Fig. 4B). These values are likely an underestimate of damaging DBDPs across all human TFs (Fig. 4C). These damaging nsSNPs occur at lower frequencies in the ExAC population than do nsSNPs for which no change in DNA binding is predicted ($P < 0.05$, permutation test) (Fig. 4D), suggesting that they are more likely to be deleterious.

Per individual, there were very few (median = 2) nonsense DBD variants but a wide range in the number of putatively damaging missense variants (median = 9, DBDPs at DNA-contacting residues and predicted as damaging by PolyPhen-2 and SIFT) (Fig. 4E and fig. S12). Hence, we investigated what mechanisms might allow damaged DBDPs to be tolerated. TFs reported to tolerate homozygous loss-of-function (LoF) mutations in Icelanders (*21*) had a significantly higher fraction of DNA-contacting residues altered by our identified nsSNPs ($P = 6.63 \times 10^{-8}$, permutation test) (Fig. 4F). TFs with a coexpressed paralog (*22*) had a significantly higher fraction of variable DNA-contacting residues ($P = 6.11 \times 10^{-8}$, permutation test) (Fig. 4G); this enrichment was significant independent of LoF-tolerance status ($P < 0.005$, t test) (*6*). Additional compensation could arise from epistasis with cis-regulatory variants (*23*). Damaged DBDPs might be associated with undiagnosed or subclinical phenotypes, variably penetrant phenotypes due to epistatic or gene-environment interactions, or phenotypes that present in later life.

Our results highlight the utility of PBM profiling to reveal changes in the DNA binding activities of variant DBDs. PBM profiling of DBDPs identified through additional sequencing studies may elucidate disease pathologies by revealing alterations in DNA binding that result in transcriptional dysregulation.

Our analyses suggest that most unrelated individuals have a unique repertoire of TF alleles with a distinct landscape of DNA binding activities. Variants with subtle changes in DNA binding activities may confer reduced deleteriousness and thus have greater potential for giving rise to phenotypic variation. Analysis of genetic interactions among TFs, TF variants, and noncoding regulatory variation likely will provide insights into the structure of genetic variation that leads to phenotypic differences among people.

REFERENCES AND NOTES

1. Exome Aggregation Consortium, *bioRxiv* (2015); http://dx.doi.org/10.1101/030338.
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. UniProt Consortium, *Nucleic Acids Res.* **43**, D204–D212 (2015).

9. I. A. Adzhubei *et al.*, *Nat. Methods* **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–D1006 (2014).
12. D. M. Ibrahim *et al.*, *Genome Res.* **23**, 2091–2102 (2013).
13. M. F. Berger *et al.*, *Nat. Biotechnol.* **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).

ACKNOWLEDGMENTS

We thank M. Hume, Y.-H. Hsu, Y. Shen, and D. Balcha for technical assistance and A. Gimelbrant for helpful discussions. We are grateful to the Exome Aggregation Consortium for making its data publicly available prior to publication. This work was supported by the National Institutes of Health (grants NHGRI R01 HG003985 to M.L.B. and T.H. and P50 HG004233 to M.V. and D.E.H.), an A*STAR National Science Scholarship to K.H.K., and National Science Foundation Graduate Research Fellowships to L.A.B. and J.M.R. TF PBM data have been deposited into UniPROBE (publication data set accession BAR15A). GST negative control PBM 8-mer data are provided in table S7. M.L.B. is a coinventor on U.S. patents no. 6,548,021 and no. 8,530,638 on PBM technology and corresponding universal sequence designs, respectively. Universal PBM array designs used in this study are available via a materials transfer agreement with The Brigham and Women's Hospital. A.V., J.V.K., J.M.R., N.S., T.H., and S.Y. performed experiments; L.A.B., J.V.K., J.M.R., S.S.G., E.J.R., J.W., L.M., K.H.K., S.I., T.S., L.S., R.G., and C.C. performed data analysis; M.K., M.J.D., M.V., D.E.H., and M.L.B. supervised research; L.A.B., L.M., K.H.K., D.E.H., and M.L.B. designed the study and wrote the manuscript; and L.A.B., J.V.K., J.M.R., S.S.G., L.M., K.H.K., S.I., and M.L.B. prepared figures and tables.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/351/6280/1450/suppl/DC1
Materials and Methods
Figs. S1 to S12
Tables S1 to S7
References (*24*–*55*)

21 September 2015; accepted 18 February 2016
10.1126/science.aad2257

CANCER

Activation of proto-oncogenes by disruption of chromosome neighborhoods

Denes Hnisz,[1]* Abraham S. Weintraub,[1,2]* Daniel S. Day,[1] Anne-Laure Valton,[3] Rasmus O. Bak,[4] Charles H. Li,[1,2] Johanna Goldmann,[1] Bryan R. Lajoie,[3] Zi Peng Fan,[1,5] Alla A. Sigova,[1] Jessica Reddy,[1,2] Diego Borges-Rivera,[1,2] Tong Ihn Lee,[1] Rudolf Jaenisch,[1,2] Matthew H. Porteus,[4] Job Dekker,[3,6] Richard A. Young[1,2]†

Oncogenes are activated through well-known chromosomal alterations such as gene fusion, translocation, and focal amplification. In light of recent evidence that the control of key genes depends on chromosome structures called insulated neighborhoods, we investigated whether proto-oncogenes occur within these structures and whether oncogene activation can occur via disruption of insulated neighborhood boundaries in cancer cells. We mapped insulated neighborhoods in T cell acute lymphoblastic leukemia (T-ALL) and found that tumor cell genomes contain recurrent microdeletions that eliminate the boundary sites of insulated neighborhoods containing prominent T-ALL proto-oncogenes. Perturbation of such boundaries in nonmalignant cells was sufficient to activate proto-oncogenes. Mutations affecting chromosome neighborhood boundaries were found in many types of cancer. Thus, oncogene activation can occur via genetic alterations that disrupt insulated neighborhoods in malignant cells.

Tumor cell gene expression programs are typically driven by somatic mutations that alter the coding sequence or expression of proto-oncogenes (*1*) (Fig. 1A), and identifying such mutations in patient genomes is a major goal of cancer genomics (*2*, *3*). Dysregulation of proto-oncogenes frequently involves mutations that bring transcriptional enhancers into proximity of these genes (*4*). Transcriptional enhancers normally interact with their target genes through the formation of DNA loops (*5*–*7*), which typically are constrained within larger CCCTC-binding factor (CTCF) cohesin–mediated loops called insulated neighborhoods (*8*–*10*), which in turn can form clusters that contribute to topologically associating domains (TADs) (*11*, *12*) (fig. S1A). This recent understanding of chromosome structure led us to hypothesize that silent proto-oncogenes located within insulated neighborhoods might be activated in cancer cells via loss of an insulated neighborhood boundary, with consequent aberrant activation by enhancers that are normally located outside the neighborhood (Fig. 1A, lowest panel).

To test this hypothesis, we used chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) to map neighborhoods and other cis-regulatory interactions in a cancer cell genome (Fig. 1B and table S1). A T cell acute lymphoblastic leukemia (T-ALL) Jurkat cell line was selected for these studies because key T-ALL oncogenes and genetic alterations are well known (*13*, *14*). The ChIA-PET technique generates a high-resolution (~5 kb) chromatin interaction map of sites in the genome bound by a specific protein factor (*8*, *15*, *16*). Cohesin was selected as the target protein because it is involved in both CTCF-CTCF interactions and enhancer-promoter interactions (*5*–*7*) and has proven useful for identifying insulated neighborhoods (*8*, *10*) (fig. S1, A and B). The cohesin ChIA-PET data were processed using multiple analytical approaches (figs. S1 to S4 and table S2), and their analysis identified 9757 high-confidence interactions, including 9038 CTCF-CTCF interactions and 379 enhancer-promoter interactions (fig. S4C). The CTCF-CTCF loops had a median length of 270 kb, contained on average two or three genes, and covered ~52% of the genome (table S2). Such CTCF-CTCF loops have been called insulated neighborhoods because disruption of either CTCF boundary causes dysregulation of local genes due to inappropriate enhancer-promoter interactions (*8*, *10*). Consistent with this, the Jurkat chromosome structure data showed that the majority of cohesin-associated enhancer-promoter interactions had end points that occurred within the CTCF-CTCF loops (Fig. 1C and fig. S2H). These results provide an initial map of the three-dimensional (3D) regulatory landscape of a tumor cell genome.

We next investigated the relationship between genes that have been implicated in T-ALL pathogenesis and the insulated neighborhoods. The majority of genes (40 of 55) implicated in T-ALL pathogenesis, as curated from the Cancer Gene Census and individual studies (table S3), were located within the insulated neighborhoods identified in Jurkat cells (Fig. 2A and fig. S5); 27 of these genes were transcriptionally active and 13 were silent, as determined by RNA sequencing (RNA-seq) (Fig. 2A and table S4). Active oncogenes are often associated with super-enhancers (*17*, *18*), and we found that 13 of the 27 active T-ALL pathogenesis genes were associated with superenhancers (Fig. 2, A and B, and fig. S5A). Silent genes have also been shown to be protected by insulated neighborhoods from active enhancers located outside the neighborhood, and we found multiple instances of silent proto-oncogenes located within CTCF-CTCF loop structures in the Jurkat genome (Fig. 2, A and C, and fig. S5B). Thus, both active oncogenes and

[1]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA. [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [3]Program in Systems Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605, USA. [4]Department of Pediatrics, Stanford University, Stanford, CA, USA. [5]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [6]Howard Hughes Medical Institute.
*These authors contributed equally to this work. †Corresponding author. E-mail: young@wi.mit.edu

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 and Martha L. Bulyk

Science **351** (6280), 1450-1454.
DOI: 10.1126/science.aad2257

Variation and transcription factor binding

Little is known about the phenotypic and functional effects of genetic variants that result in amino acid changes within functional proteins. Barrera *et al.* investigated whether amino acid variants changed the DNA binding specificity or affinity of transcription factors. Predictive analyses identified changes in the proteins, and protein-binding microarrays verified changes that affected transcription factor function, including those associated with disease. Thus, within-human protein sequence variation can affect transcriptional regulatory networks, which, depending on the genetic variant, may confer robustness and buffer against amino acid changes and could explain phenotypic variation among individuals.
Science, this issue p. 1450