Cell

Leading Edge
Review

The Genetics of Transcription Factor DNA Binding Variation

Bart Deplancke,[1,*] Daniel Alpern,[1] and Vincent Gardeux[1]
[1]Laboratory of Systems Biology and Genetics, Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne and Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
*Correspondence: bart.deplancke@epfl.ch
http://dx.doi.org/10.1016/j.cell.2016.07.012

Most complex trait-associated variants are located in non-coding regulatory regions of the genome, where they have been shown to disrupt transcription factor (TF)-DNA binding motifs. Variable TF-DNA interactions are therefore increasingly considered as key drivers of phenotypic variation. However, recent genome-wide studies revealed that the majority of variable TF-DNA binding events are not driven by sequence alterations in the motif of the studied TF. This observation implies that the molecular mechanisms underlying TF-DNA binding variation and, by extrapolation, inter-individual phenotypic variation are more complex than originally anticipated. Here, we summarize the findings that led to this important paradigm shift and review proposed mechanisms for local, proximal, or distal genetic variation-driven variable TF-DNA binding. In addition, we discuss the biomedical implications of these findings for our ability to dissect the molecular role(s) of non-coding genetic variants in complex traits, including disease susceptibility.

Introduction

Analysis of genomic variation in humans (Auton et al., 2015) as well as in model species such as the mouse (Keane et al., 2011; Yalcin et al., 2011) and fruit fly (Huang et al., 2014; Massouras et al., 2012) is providing unprecedented opportunities to understand the genetic basis of complex traits, including disease susceptibility. An important insight that emerged from genome-wide association studies (GWAS) is that the vast majority of significantly associated genetic variants is located in non-coding regions and may thus impact gene regulation. For example, of 465 unique trait/disease-associated single nucleotide polymorphisms (SNPs) derived from 151 GWAS studies, only 12% are located in protein-coding regions, while 40% fall within introns and another 40% in intergenic regions (Hindorff et al., 2009). In addition, genome-wide profiling of accessible chromatin regions using DNase I hypersensitivity (DHS) mapping revealed that almost 60% of non-coding GWAS SNPs and other variants are located within DHS sites, with another 20% being in complete linkage disequilibrium (LD) with variants that lie in a proximate DHS site (Maurano et al., 2012). Since DHS sites reflect the occupancy of DNA binding proteins such as transcription factors (TFs), these data indicate that GWAS loci may alter the binding of TFs and, as such, induce variation in gene expression and ultimately in complex organismal phenotypes. In this Review, we summarize the findings that led to this increasingly accepted notion of the importance of variation in TF-DNA binding in mediating phenotypic diversity. In addition, we strive to clarify why, for the majority of studied traits or diseases, establishing a mechanistic link between regulatory and phenotypic variation is still very challenging.

For this purpose, we explore the molecular mechanisms mediating TF-DNA binding variation and address a major question in the field—namely, why the majority of variable TF-DNA binding events appear to be driven by mechanisms other than nucleotide variation in the cognate motifs. We thereby focus on human inter-individual, molecular variation and restrict this Review to discussing mechanisms underlying variable TF-DNA binding. Consequently, we will only briefly cover the functional consequences of this variation or other modes of regulatory variation, which have been extensively detailed elsewhere both for humans and model organisms (Albert and Kruglyak, 2015; Lehner, 2013; Lowe and Reddy, 2015; Mackay et al., 2009; Pai et al., 2015).

A Brief Historical Perspective on Variable TF-DNA Interactions as Key Drivers of Inter-individual, Phenotypic Diversity

The discovery of regulatory sequences (or "operators") in bacteria by Jacob and Monod initiated the debate of whether variation in "regulator-operator" interactions could drive phenotypic diversity (Jacob and Monod, 1961). It was proposed that this variation could arise either through mutations in the regulator itself or through mutations in the operator that would "alter or abolish its specific affinity for the repressor (i.e., regulator)" (Jacob and Monod, 1961). This fundamental prediction proved to be accurate across species, and multiple examples have since been revealed that support both scenarios (Barrera et al., 2016; Hoekstra and Coyne, 2007; Lynch and Wagner, 2008; Wray, 2007). The first concrete evidence supporting the importance of such non-coding or regulatory variation for human traits or diseases started to emerge in the early 1980s, when the molecular mechanisms underlying thalassemias were investigated. These heritable blood disorders, characterized by an abnormal form of hemoglobin, made it intuitive to explore the globin gene locus for disease-causing genetic variants. Numerous variants were

detected, including several polymorphisms in the β globin gene (HBB) promoter that correlated with reduced *HBB* expression (Orkin et al., 1982; Poncz et al., 1982). For example, a single nucleotide substitution (C to G) at position −87 of *HBB*'s transcription start site (Orkin et al., 1982) was hypothesized to affect the recruitment of a transcriptional activator. However, it was only 11 years later, when the erythroid Krüppel-like factor (KLF1) was cloned, that the affected site (CA(C→G)CC) was matched with a TF (Miller and Bieker, 1993) (Table 1). This groundbreaking example (as well as several others listed in Table 1) support the idea of TF-DNA interactions being key drivers of phenotypic variation. However, the fact that the underlying molecular mechanisms were uncovered for these diseases is still more the exception than the rule.

Assessing the Impact of Genetic Variation on TF-DNA Binding: A Complex Affair

The ability to elucidate the molecular mechanisms underlying thalassemia, haemophila B, or malaria resistance was made possible because several critical pieces of information were available that are often missing in other genotype-phenotype relationship studies: (1) knowledge of the affected gene, which facilitated the identification of the causal mutation(s); (2) availability of DNA binding specificity data for the implicated TF; and (3) relatively straightforward imputation of the effect of the causal mutation(s) on TF-DNA binding. Below, we will discuss each of these three items in more detail and explain why they collectively complicate studies that investigate the impact of genetic variation on molecular or organismal variation.

Identification of Causal Mutation(s) and Affected Gene(s)

In contrast to cases like the thalassemia mutations discussed above, GWAS studies identify genetic variants linked to particular traits, but not necessarily those actually causing the disease (Manolio, 2013). In addition, such studies do generally have little prior knowledge regarding which genes will be uncovered. Therefore, by simply matching a GWAS SNP with a TF binding site, one risks wrongly inferring that a TF must affect the expression of the gene that is most proximal to a particular binding site. However, the actual culprit could be another genetically linked but unprobed variant such as an indel or a rare SNP that impacts a different binding site and thus a distinct TF and/or target gene. This potential misidentification is why significant efforts are currently undertaken to fine-map complex traits using statistical arguments (Figure 1; see also the "Imputing DNA binding variation" section) and/or integrative genomic approaches to identify causal variants and their target genes at nucleotide-level resolution. A striking recent example involves variants that have been consistently associated with an elevated body mass index in both children and adults (Dina et al., 2007; Frayling et al., 2007). These variants are located in the first and second introns of a large (>250 kb) gene named *fatso*, or *FTO*, because its deletion causes a fused toes phenotype in mouse (Peters et al., 1999). Given its association with obesity, it was subsequently rephrased as fat mass and obesity-associated gene and was widely mechanistically studied for its role in energy homeostasis (Fischer et al., 2009). However, recent studies revealed that the focal variants are, in fact, located in a regulatory element that controls the expression of the TF-coding genes *IRX3* and *IRX5* more than 1 Mbp away (Claussnitzer et al., 2015; Smemo et al., 2014). Thus, these variants appear to have little impact on the *FTO* gene, even though they are positioned within its introns. Rather, one variant disrupts the binding site of ARID5B (AA(T→C)ATT), resulting in elevated *IRX3* and *IRX5* expression. This, in turn, increases the formation of white fat cells, possibly leading to excessive fat accumulation (Claussnitzer et al., 2015). Significant efforts involving a battery of advanced computational (Claussnitzer et al., 2014) and experimental approaches were required to study the molecular function of the "FTO variants" and their relationship with body mass index. This complexity illustrates why the number of mechanistically well-studied relationships between regulatory and phenotypic variation is still relatively low. It also explains why the majority of such studies focused on gene proximal variants (Table 1), especially prior to 2005, when the importance of chromosome conformation in gene regulation was still less established.

Incomplete TF Motif Catalog

In the early 1990s, DNA binding variation of the TFs CEBPα and HNF4A, as well as GATA1, was linked to, respectively, haemophilia B and malaria resistance (Table 1). The identification of these TFs is intuitive since they were among the relatively few TFs whose DNA binding properties had been described at the time when these genetic studies were carried out (Faisst and Meyer, 1992). Several other studies have since established regulatory variant-phenotype relationships on the basis of GATA1 (Table 1), illustrating how such studies tend to be restricted to investigating phenotypes that involve well-characterized TFs.

While the development of several large-scale in vitro DNA binding characterization technologies, such as protein-binding microarrays (PBM) (Berger et al., 2006), bacterial one-hybrid (B1H) screening (Meng et al., 2005), and high-throughput (HT)-SELEX (Jolma et al., 2010), has enabled a significant expansion of the TF motif catalog, it is worth noting that at least one-third of human TFs remains uncharacterized (Box 1). In other words, several hundred TFs are still devoid of DNA binding specificity models such as the most routinely used position weight matrix (PWM) (Stormo and Zhao, 2010) (Figure 1). This lack of information severely limits the ability to analyze the effects of genetic variation on TF-DNA binding, as a large fraction of human TFs can simply not be taken into account in such studies.

This problem becomes even larger when considering that many TFs do not bind DNA as single entities but, rather, in the form of obligate heterodimers such as TFs containing bZIP, bHLH, MADS box, or Rel DNA binding domains. Since the focus of DNA binding specificity determination studies has largely been on single protein-DNA interactions, DNA binding motifs for such heterodimers are underrepresented in current regulatory lexicons. Moreover, many TFs also participate in facultative heterodimers since they can bind to DNA both in monomeric or dimeric context. It is difficult to know how many of such heterodimers routinely form in cells, but predictions range from 3,000 (Ravasi et al., 2010) to >25,000 (Jolma et al., 2015). Interestingly, these cooperative TF pairs often show distinct binding site preferences compared to the respective, individual TFs, as the heterodimer core motif typically consists of closely packed individual motifs that overlap at their flanks. Consequently, individual

Table 1. Examples Linking Variable TF-DNA Binding to Phenotypic Variation Arranged by Date of Characterization

Phenotype	Affected Gene	Causal Variant Position Relative to TSS	Affected Binding Site	TFBS Outcome	Reference(s)
Hereditary persistance of fetal haemoglobin	*HBG*	−175 bp	GATA1; TAL1	Gain	(Martin et al., 1989; Wienert et al., 2015)
Haemophilia B Leyden	*F9*	−20 bp; 10 bp; −6 bp	HNF4a; C/EBPa; OC1/OC2	Loss	(Reijnen et al., 1992; Crossley and Brownlee, 1990; Funnell et al., 2013)
Haemophilia B Brandenburg	*F9*	−26 bp	AR	Loss	(Crossley et al., 1992)
Delta-thalassemia	*HBD*	−77 bp	GATA1	Loss	(Matsuda et al., 1992)
Duffy blood antigen/chemokine receptor expression	*DARC*	−46 bp	GATA1	Loss	(Tournamille et al., 1995)
Familial combined hyperlipidemia	*LPL*	−39 bp	OCT1	Loss	(Yang et al., 1995)
Bernard-Soulier syndrome	*GP1BB*	−133 bp	GATA1	Loss	(Ludlow et al., 1996)
Osteoporosis	*COLIAI*	+2 kb	Sp1	Gain	(Grant et al., 1996)
Maturity-onset diabetes of the young	*HNF1A*	−58 bp	HNF4A	Loss	(Gragnoli et al., 1997)
Asthma	*IL10*	−509 bp	YY1	Gain	(Hobbs et al., 1998)
Pyruvate kinase deficiency	*PKLR*	−72 bp	GATA1	Loss	(Manco et al., 2000)
Congenital erythropoietic porphyria	*UROS*	−70 bp; −90 bp	GATA1; CP2	Loss	(Solis et al., 2001)
Psoriasis	*SLC9A3R1*	−237 bp	RUNX1	Loss	(Helms et al., 2003)
Systemic lupus erythematosus	*FASLG*	−844 bp	CEBPB	Loss	(Wu et al., 2003)
Esophageal cancer	*COX-2*	−1195 bp	c-MYB	Gain	(Zhang et al., 2005)
Treacher Collins syndrome	*TCOF1*	−346 bp	YY1	Loss	(Masotti et al., 2005)
Alpha-thalassemia	*HBA*	−13 bp	GATA1	Gain	(De Gobbi et al., 2006)
Holoprosencephaly	*SHH*	−460 kb	SIX3	Loss	(Jeong et al., 2008)
Various cancers	*TERT*	−187 bp	ETS2	Loss	(Xu et al., 2008)
Nonsyndromic cleft lip	*IRF6*	−14 kb	AP2	Loss	(Rahimov et al., 2008)
Pierre Robin syndrome	*SOX9*	−1.44 Mb	MSX1	Loss	(Benko et al., 2009)
Prostate cancer	*MYC*	−200 kb	FOXA1	Gain	(Jia et al., 2009)
Colorectal cancer	*MYC*	−300 kb	TCF7L2	Gain	(Tuupanen et al., 2009)
Asthma and autoimmune diseases	*ZPBP2; GSDMB; ORMDL3*	−5 kb; +44 kb; +54 kb	CTCF	Loss	(Verlaan et al., 2009)
Myocardial infarction	*SORT1*	−44 kb	CEBPA	Loss	(Musunuru et al., 2010)
Beta-thalassemia	*HBB*	−71 bp	GATA1	Loss	(Al Zadjali et al., 2011)
Coagulant factor VII deficiency	*F7*	−60 bp	HNF4A	Loss	(Zheng et al., 2011)
Osteoarthritis	*GDF5*	−41 bp	YY1	Loss	(Dodd et al., 2013)
Breast cancer	*CCND1*	−127 kb; −76 kb	ELK4; GATA3	Loss; Gain	(French et al., 2013)
Melanoma, various cancers	*TERT*	+2bp; −66 bp; −88 bp	ETS2	Gain	(Horn et al., 2013) (Huang et al., 2013)
Increased cancer susceptibility	*KITLG*	+20 kb	P53	Loss	(Zeron-Medina et al., 2013)
Hirschsprung disease	*SOX10*	−30 kb	AP2; SOX10	Loss	(Lecerf et al., 2014)
Insulin resistance	*PPARG2*	−6 kb	PRRX1	Loss	(Claussnitzer et al., 2014)
Type 2 diabetes and proinsulin-decrease	*ARAP1*	+418 bp	PAX6/PAX4	Loss	(Kulzer et al., 2014)
Melanoma	*SDHD*	−25 bp; −7 bp; −4 bp	EHF, ELF1 & ETS1	Loss	(Weinhold et al., 2014)
Pancreatic agenesis	*PTF1A*	−25 kb	FOXA2, PDX1	Loss	(Weedon et al., 2014)
Acute lymphoblastic leukemia	*TAL1*	−7.5 kb	MYB	Gain	(Mansour et al., 2014)
Obesity and Type 2 diabetes	*IRX3; IRX5*	−0.5 Mb; −1.2 Mb	ARID5B	Loss	(Claussnitzer et al., 2015)
Colorectal cancer	*FASLG*	−1377 bp; −670 bp	SP1; STAT1	Loss	(Wang et al., 2016)

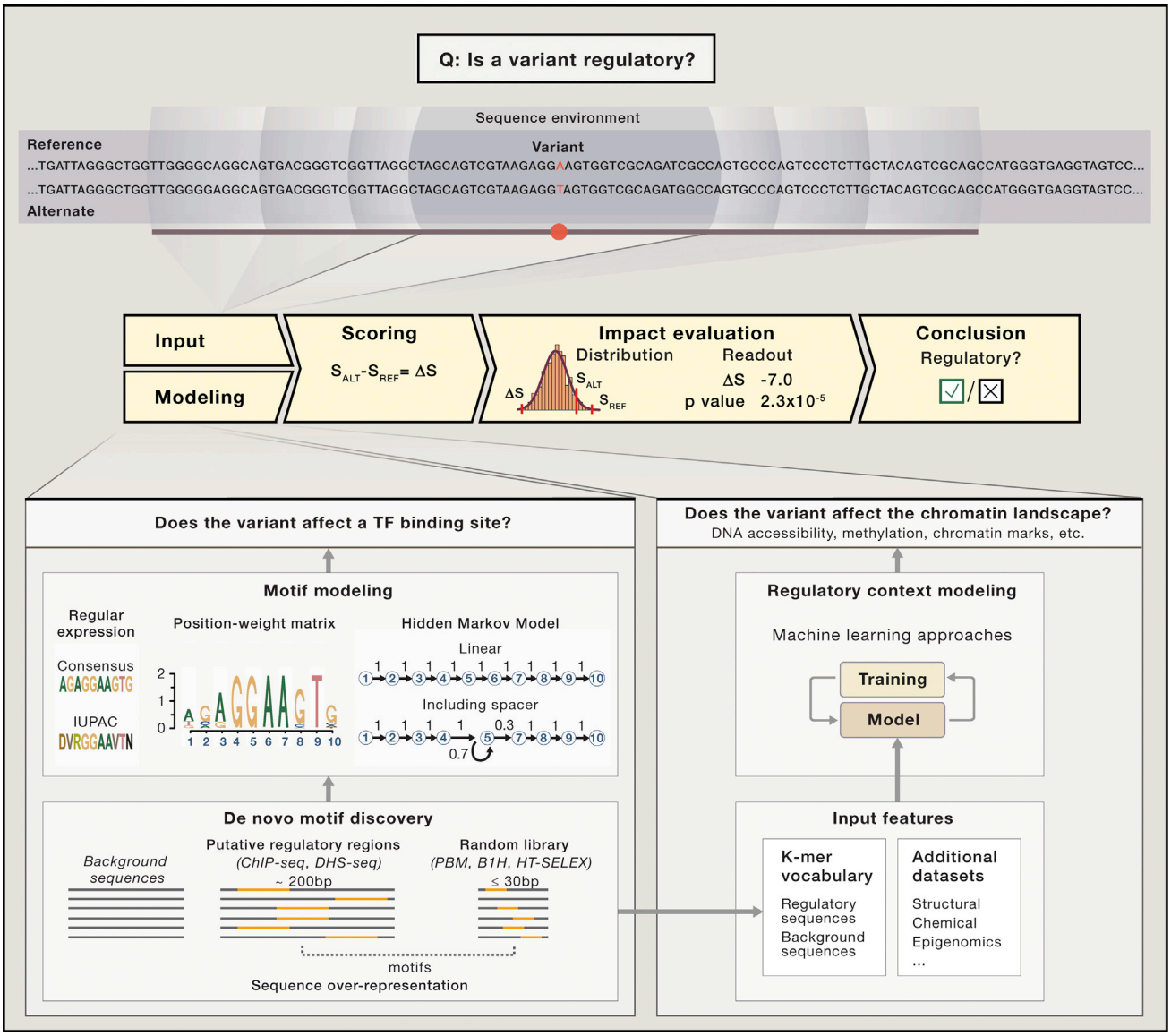

Figure 1. A Methodological Workflow for Identifying Regulatory Variants

Sequence-based, computational methodologies that evaluate the impact of potential regulatory variants on TF-DNA binding and downstream regulatory processes are schematically presented. For every putative variant (SNV, as in this example, or indel), a reference and alternate (containing the variant) sequence of pre-defined length (illustrated by the distinct shades of gray) is extracted. The chosen length defines the "sequence environment" and varies according to the type of model that is used. The middle yellow panel shows the common workflow, where both sequences are scored (S_{ALT} and S_{REF}) according to a specific model representation to obtain a differential score (ΔS) that may indicate a change in DNA binding or more generally in chromatin state. As shown, ΔS supports a model in which the variant impacts a gene regulatory process. The bottom part of the figure illustrates the two main strategies that are employed for modeling the regulatory effect of a variant. The choice of the strategy depends on the posed question: does the variant impact (1) the binding of a TF (left) or (2) the local chromatin landscape (right)? In the first scenario, computational methods are used that depend on the availability of a comprehensive catalog of TF binding sequences or motifs (Box 1). The "de novo motif discovery" part schematizes the procedure that is required to obtain such a catalog, illustrating the use of sequence over-representation strategies that are applied on both in vivo (ChIP-seq, DHS-seq, etc.) and in vitro (e.g., PBM, HT-SELEX, or B1H) derived datasets. These strategies then produce TF motifs that can be represented either in regular expression format or using PWM- or HMM-based models. In this example, the linear HMM model is a generic representation of the PWM motif, with each node (state) of the HMM representing the position of a base in the motif. Additionally, a second HMM model is depicted, which inherently takes a variable space within the motif into account, for accurate representation of more complex binding scenarios (e.g., TF dimers). To answer the second question (lower-right), computational methods mainly rely on machine learning models that are trained on a wide variety of features such as a k-mer vocabulary built on regulatory versus background sequences or additional (epi)genomic datasets. These more elaborate models can also be used to score the two input sequences. The pipeline then evaluates the regulatory nature of the variant by directly assessing the differential score ΔS or by calculating a p value based on the distribution of the scores. Of note, this pipeline can be applied multiple times on different variants, after which the results can be aggregated and compared to prioritize variants.

Box 1. How Many Human TFs Have Assigned Motifs?

While seemingly straightforward, it turns out to be difficult to precisely enumerate the number of TFs with defined motifs. There is no consensus on the number of TF-coding genes in the human genome. A comprehensive, manual curation primarily based on the presence of sequence-specific DNA binding domains revealed 1,391 high-confidence genes, with another 216 listed as plausible (Vaquerizas et al., 2009). However, this list may not be exhaustive as a protein microarray-based survey of DNA binding capacity revealed that hundreds of proteins among ~3,000 that were not annotated as TFs were able to bind to DNA in a site-specific manner (Hu et al., 2009). Thus, additional experimental efforts will be required to derive a more precise estimate of the number of TFs that the human genome encodes. Therefore, we need to simplify the question to how many of the ~1,400 high-confidence TFs have annotated motifs. A very recent, expansive analysis involving almost 1,000 high-quality ChIP-seq and 542 HT-SELEX datasets produced binding site models for 601 human TFs that are retrievable from the HOCOMOCO database (Kulakovskiy et al., 2016). Why only ~40% of human TFs feature experimentally derived motifs despite the development of powerful DNA binding characterization technologies such as PBM (Berger et al., 2006), bacterial one-hybrid (B1H) (Meng et al., 2005), or HT-SELEX (Jolma et al., 2010) is unclear but may largely be due to technical limitations, including loss of DNA binding properties or weak in vitro expression of full-length proteins. This is why most in vitro DNA binding assays rely on analysis of the DNA binding domains (DBDs) of TFs, as these are easier to work with in terms of cloning and expression while exhibiting DNA binding properties that appear largely comparable to the respective full-length protein versions (Jolma et al., 2013).

The largest TF families that still resist a comprehensive characterization are the high-mobility group (HMG) TFs and C2H2 zinc finger proteins (Jolma et al., 2013), of which the human genome encodes more than 700 (Weirauch and Hughes, 2011). Progress is being made though, as illustrated by a recent study that combined PBMs and B1H assays to probe thousands of individual C2H2 zinc finger domains with the aim of inferring a specific DNA recognition code (Najafabadi et al., 2015). The resulting motifs proved to be highly diverse in terms of nucleotide composition and exhibited extensive degeneracy, which means that these motifs can be represented by many different sequences and that small internal perturbations in these motifs tend to have little impact on DNA binding. Consequently, there is still ample room for alternative approaches or technologies that will enable the further expansion or fine-tuning of the current catalog of human TF PWMs, also named the "human regulatory lexicon."

Nevertheless, it may not be necessary to gather experimental data for all TFs, given that many have nearly identical DNA binding properties because their DBDs are highly similar. Indeed, TFs (independent of organism) whose DBDs share >87.5% of their amino acids were found to bind to motifs that were almost indistinguishable from one another (Weirauch et al., 2014). Applying this principle to human TFs adds another 200 inferred motifs to the current catalog, which can be found in the Cis-BP database (Weirauch et al., 2014). In sum, the DNA binding properties of a significant fraction of human TFs remain uncharacterized without even taking into account heterodimer or higher-order complex formation (see main text).

TFs may still be able to bind to this core motif, albeit with much lower affinity. This may, in part, explain the observed discrepancy between in vivo DNA occupancy levels and in-vitro-derived DNA binding affinities (Biggin, 2011), since these in vivo binding events may reflect binding by interacting TF pairs and not individual TFs. It is therefore clear that a large portion of motifs remain to be characterized, emphasizing the need for new technologies or efforts to close this gap.

Imputing DNA Binding Variation

It has often proven difficult to infer whether a specific polymorphism will significantly change TF-DNA binding and act as a regulatory variant, even if the PWM model of the TF is available. This complication stems from difficulties in capturing the DNA binding complexity of a TF in a robust binding model either to confidently detect a genuine binding site within a given sequence or to accurately infer the impact of a variant on detected motifs.

The Accuracy of Binding Models and Robustness of Motif Detection. The majority of motif detection methodologies rely on PWM representation since PWMs perform relatively well with respect to capturing the overall binding affinity. This is because PWMs can be modeled as a numerical matrix, which enables the scoring of a given sequence according to its similarity to a motif (Figure 1). Nevertheless, it is important to acknowledge that this model also has several limitations, which may impede the discovery of the correct binding patterns. For example, PWM models assume that the nucleotide binding energies are independent (Stormo and Zhao, 2010), which proved not to be generally valid (Bulyk et al., 2002; Jolma et al., 2013; Maerkl and Quake, 2009; Nutiu et al., 2011), and are also suboptimal to represent the binding of TF dimers, since many of these bind to two sequences that are separated by a spacer with variable length. These caveats have spurred the development of different models for representing TF motifs, such as hidden Markov models (HMMs) (Gelfond et al., 2009; Zhao et al., 2005) and more advanced machine learning models, stimulated by the increasing availability of multiple layers of genomic, transcriptomic, and epigenomic information. Among these are support vector machine (SVM) or neural network (NN) approaches that are trained on datasets containing both known regulatory and random sequences, with the goal of recognizing and scoring new putative regulatory sequences (Gao and Ruan, 2015) (Figure 1 and Table S1). Such representations have many advantages over conventional models because they are highly flexible. In addition, they are not limited to the DNA sequence recognized by the TF and can incorporate additional features that are also important to model TF-DNA binding. These features include the 3D structural conformation of DNA and its steric characteristics (Levo and Segal, 2014; Rohs et al., 2009; Zhou et al., 2015), the chemical properties used to model TF amino acid-nucleotide contacts at the atomic level (Bauer et al., 2010; Maienschein-Cline et al., 2012), protein concentration (Djordjevic et al., 2003; Wang and Batmanov, 2015) that allows for a more accurate estimation of DNA occupancy and thus intrinsic DNA binding affinity (Biggin, 2011; Simicevic et al., 2013), and, finally, the nucleotide composition of motif-neighboring sequences.

Indeed, recent work revealed that the sequence environment of a genuine binding site tends to be distinct from that of unbound sequences. In particular, it was shown to exhibit specific sequence features such as high GC content (White et al., 2013) or a higher similarity to the core motif (Dror et al., 2015) that may guide TFs to their cognate binding sites. These findings

have important consequences in terms of predicting DNA binding events, since motif-scanning tools typically penalize for local nucleotide composition biases. Instead, a better practice may now involve rewarding motifs that are surrounded by established DNA binding-promoting features, such as a high GC fraction or lower-scoring and thus weaker homotypic (i.e., similar) motifs. Together, these studies illustrate that the formulation of DNA binding models and computational detection of genuine binding sites is far from trivial and that further efforts aimed at integrating a wide range of genomic datasets will be required to increase the robustness of motif definition and mapping approaches.

The Complexity of Correctly Inferring the Effect of Motif Variation on TF-DNA Binding. Genetic variants that change a TF motif often affect the binding ability of a TF to that site because of an altered DNA binding affinity (Table 1 and Figure 1). Initial efforts to computationally predict relevant regulatory variants simply revolved around the consideration of all SNPs that overlap with TF binding sites (Ameur et al., 2009; Chorley et al., 2008; Ponomarenko et al., 2001). However, given the degenerate nature of binding motifs (i.e., binding is not binary but is variable depending on different sequences), these kinds of analyses tend not to provide good sensitivity. A more refined approach in this regard is to analyze the difference in DNA binding affinity (for example, scored using a PWM) between two alleles, i.e., the reference and the alternate impacted by the variant (Figure 1 and Table S1). The greater this difference, the greater the predicted impact of the variant on binding of the respective TF and thus also the greater the likelihood of it being causal.

More recent machine learning methods no longer depend on the use of a strict motif database and directly infer regulatory effects from k-mer vocabularies trained on ChIP-seq or other experimental data. These vocabularies consist of all possible DNA sequences of length k that collectively capture specific sequence properties of certain regulatory elements such as cell-type-specific enhancers. This methodological development stems from the general appreciation in the field that the motif alone cannot accurately predict differential DNA binding and thus should be complemented (or even replaced) with information on the sequence environment around the focal variant, as well as on other DNA or chromatin features that enhance the model's overall predictive power (as already covered in the previous section). Indeed, it is now well accepted that only a minority of motif-disrupting variants effectively result in altered DNA binding of the respective TF (Heinz et al., 2013; Kilpinen et al., 2013; Maurano et al., 2015; Spivakov et al., 2012). One possible explanation is based on the finding that, across the genome, TF motifs appear to occur in clusters with some built-in redundancy (Gotea et al., 2010), in line with the observation that the sequence environment of relevant TF binding sites tends to have a certain similarity to the core motif (Dror et al., 2015). These clustered sites may buffer genetic perturbations that affect one of the motifs. Indeed, the greater the number of such homotypic motifs, the greater the buffering effect (Kilpinen et al., 2013). Given the pervasive nature of this buffering phenomenon (Maurano et al., 2015), the failure to take such neighboring homotypic motifs into account may result in false TF-DNA binding event predictions.

More generally, epigenomic properties such as nucleosome location (Soufi et al., 2015) or density (Barozzi et al., 2014) or DNA methylation (Domcke et al., 2015) may impact the ability of a TF to bind to a certain DNA sequence. Upon screening ~1,300 human TFs for their ability to bind to one of 150 distinct CpG-containing motifs, 47 were found to bind to DNA with several exhibiting methylation-specific DNA specificities (Hu et al., 2013). This is consistent with an earlier study demonstrating that the TF Kaiso is capable of binding not only to an unmethylated motif, TCCTGCNA, but also to a methylated, clearly distinct palindromic motif, TCTmCGmCGAGA, with even greater affinity (Raghav et al., 2012). How frequently such methylation-dependent changes in DNA binding occur and the extent to which other DNA modifications affect DNA binding specificities is still a matter of debate. Nevertheless, it is clear that it adds another complexity in linking DNA variation to variable TF binding.

Ongoing computational studies are attempting to take these complexities into account by implementing "big data" analyses that are creating extended machine learning models that rely on multilayered information of different types of genomic data, including TF motifs, DNase hypersensitivity sites (DHS), chromatin marks, etc. (see, for example, Table S1). As such, they can recognize regulatory regions based not only on pure sequence information, but also on the chromatin state of the DNA both at the variant locus as well as at neighboring regions. Once correctly trained, these approaches can be very precise and predict causal variants and their effects at distinct molecular levels (Alipanahi et al., 2015; Zhou and Troyanskaya, 2015) (Figure 1).

However, it is important to emphasize that their performance depends not only on the diversity of input data, but also on the correct selection of relevant features. For example, it has been repeatedly shown that it is crucial to gather data that are specific to the variant-linked trait or disease in terms of cell type, differentiation stage, tissue, or species since regulatory activity is variable and context dependent (Consortium, 2012; Maurano et al., 2015). Another limitation of these extended representation models that may dampen their widespread implementation is their inherent "black box" nature. Indeed, most of the binding patterns that were unveiled by these techniques are difficult to interpret, especially when no visual representation is provided. However, despite these caveats, advanced models have the potential of uncovering completely novel and potentially unexpected cross-mechanisms that more standard methodologies may fail to grasp.

TF-DNA Binding Is Itself a Complex, Molecular Trait

We are currently limited in our ability to predict TF binding as well as in our understanding of how genetic variation impacts on this process. Nevertheless, there is general consensus that differential, regulatory control by TFs is a major driver of phenotypic variation. A key aspect of this regulatory variation is variable TF-DNA binding. It is in this regard intriguing that only a minority of variable TF binding events are driven by nucleotide changes in the motifs of the studied TFs. For example, upon assessing binding variation of the TF NFκB in ten distinct human lymphoblastoid cell lines (LCLs), only 79 out of >1,100 variable TF-DNA binding

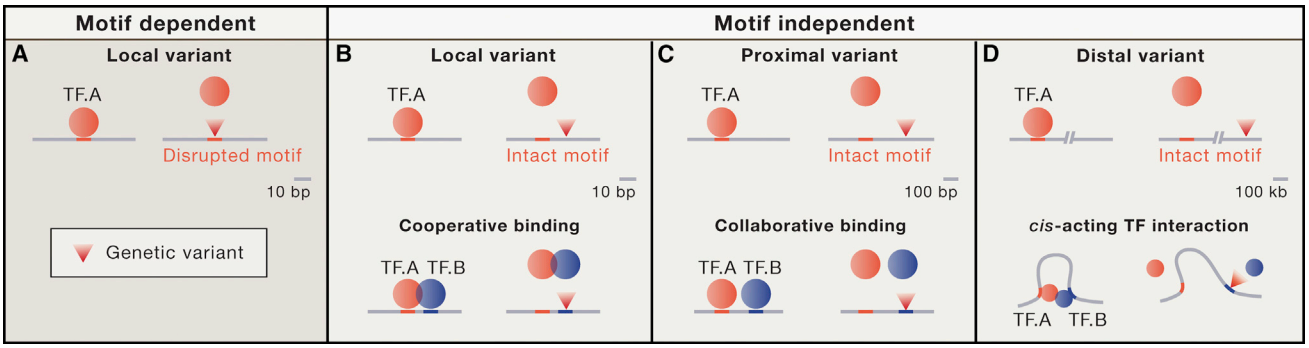

Figure 2. Distinct Modes of Genetic Variation-Mediated Changes in TF-DNA Binding

(A) Only a minority of variable TF-DNA binding events are caused by DNA variants disrupting the cognate TF recognition motif.

(B–D) The majority of variably binding events are motif variation independent, signifying that a variant located either proximally (<200 bp, B and C) or distally (D) to the focal motif affects the binding of the respective TF. Proximal variants can affect local cooperative DNA binding (B), which involves physical protein-protein interactions that require overlapping or very closely located (a few bp) motifs, or collaborative DNA binding (C), which reflects TF interdependencies needed, for example, to compete with nucleosomes and thus to access DNA. In contrast, distal variants (D) may alter chromatin state or conformation (e.g., DNA loops), which could affect the stability of interactions with DNA and between TFs.

events had a SNP directly located in the NFκB motif and induced a binding difference that was consistent with its perceived impact on motif quality (i.e., reduced binding was linked to a SNP that lowered the PWM binding score and vice versa) (Kasowski et al., 2010). One of the possible reasons that were listed (next to LD or putative epigenomic variation) involved *trans*-effects. However, ChIP-seq analyses of >20 TFs revealed extensive, allele-specific DNA binding (in a constant *trans* environment), effectively refuting this hypothesis (Reddy et al., 2012). Subsequent studies in human LCLs and in cells or tissues derived from distinct mouse strains observed a similar pattern (Heinz et al., 2013; Kilpinen et al., 2013; Soccio et al., 2015; Stefflova et al., 2013), collectively emphasizing the importance of *cis*-regulatory variation. Importantly, only a minority of differential allelic occupancy events involved nucleotide changes in the respective motifs (Reddy et al., 2012). However, this does not mean that variation in the motifs of other TFs should also be dispensed as a possible molecular mechanism for these observations—quite the contrary, in fact, as we will clarify in greater detail in the next paragraphs (see also Figure 2).

If a particular genetic variant does not affect the motif of the studied TF, what then causes the respective TF to exhibit differential DNA binding? It appears that an important fraction (at least 7.5% according to our own estimate [Kilpinen et al., 2013]) of variable TF-DNA binding events can be explained by alterations of proximal motifs (Reddy et al., 2012). Thus, at some genomic sites, TFs appear to be dependent on the proximal presence of other TFs to bind to DNA. Qualitative motif analysis combined with prior knowledge about the biological process in which the focal TF is operational lends credibility to this notion. For example, in mouse white adipose tissue, PPARγ binding sites that vary between strains and do not harbor an altered PPARγ motif were analyzed for enriched, polymorphic motifs. The top-scoring motifs corresponded to the TFs CEBPα and glucocorticoid receptor (Soccio et al., 2015) that exhibit extensive co- localization with PPARγ in mature white fat cells (Siersbæk et al., 2014). Similarly, differential PU.1 binding correlated with alterations in the motifs for the TFs CEBP and AP-1, which modulate macrophage activity (Heinz et al., 2013). However, this correlation appears to differ according to macrophage subtype. Indeed, a follow-up study in mouse microglia revealed that other TF motifs correlate better with variable PU.1 DNA binding, emphasizing the importance of cellular context in determining this type of TF interactions (Gosselin et al., 2014). Together, these studies strongly support the notion of pervasive DNA binding whose occurrence is dependent on the presence of other TFs. Since it is well appreciated that regulatory regions tend to harbor binding sites for multiple TFs, this notion may not be entirely surprising. Nevertheless, it is worthwhile in the current context of genetic variation to briefly revisit this mode of DNA binding, which is interchangeably called cooperative or collaborative DNA binding (Gosselin et al., 2014; Mirny, 2010; Slattery et al., 2014; Waszak et al., 2015). We would thereby like to argue that, for the sake of discussion and molecular understanding, it might be valuable to differentiate between these two terms (Figure 2).

Local, Cooperative TF-DNA Binding

In the context of protein-DNA interactions, cooperativity was initially used in describing the assembly of *E. coli* lambda repressors on DNA (Ptashne et al., 1980). Binding of a lambda dimer on a first operator site facilitates binding of another lambda dimer on the second operator site, given that physical interactions between the first and second dimer increase the affinity of the latter for DNA, which explains why "cooperative DNA binding" is evoked to define this process. Consequently, the term cooperativity may be especially suited for DNA binding processes that involve TFs whose physical interactions at the protein level may increase the affinity of the entire complex to specific sites in the genome. For example, binding of the winged HTH DNA binding domain-containing TF IRF4 is cooperatively enhanced by the TF PU.1 (Escalante et al., 2002). This is because binding of the two TFs contorts the DNA in a peculiar S shape, placing the TFs in an optimal position for electrostatic and hydrophobic interactions and thus stabilizing the entire complex (Escalante et al., 2002). Consequently, individual

nucleotide alterations in one of the two binding sites may alter the extent of cooperativity between two heterodimerizing TFs, as has recently been quantified for the PPARγ-RXRα heterodimer (Isakova et al., 2016). This, in turn, illuminates why the disruption of either of the two TF motifs tends to affect binding of the respective heterodimer.

Proximal, Collaborative TF-DNA Binding

For TFs to physically interact on DNA and thus for cooperative DNA binding to occur, one would intuitively expect that the respective motifs would be located very close to one another or would even overlap. However, DNA binding relationships exist between TFs whose motifs are separated tens, hundreds, or even thousands of base pairs from one another. For example, upon examining which TFs (based on motif matches) associated with NFκB binding enrichment (based on ChIP-seq data), EBF1 and STAT1 were among the most correlated TFs (Karczewski et al., 2011). Interestingly, this covariation signal of EBF1 and STAT1 motifs within variable binding regions of the TF NFκB remained significant up to 500 bp from the NFκB binding peak center, suggesting that DNA binding dependencies between these TFs were maintained over a relatively long distance. It is now increasingly appreciated that many such dependencies do not require direct contacts but instead reflect a relatively well-understood phenomenon termed collaborative DNA binding in which two or more TFs compete with a nucleosome to access DNA (Biggin, 2011; Mirny, 2010; Spitz and Furlong, 2012) (Figure 2). Given that the intrinsic affinity of a nucleosome for DNA is much greater than that of a TF alone (Polach and Widom, 1996), it may often require two or more collaborating TFs to displace the nucleosome. In this scenario, TFs would be mutually dependent, and this is indeed what is observed. For example, HNF4A and CEBPα functioning in mouse liver exhibit a mutual dependency, given that loss of HNF4A affected CEBPα DNA binding and vice versa, whereas the absence of HNF4A did not impact on the DNA binding dependency between CEBPα and FOXA1 (Stefflova et al., 2013). A similar DNA binding interdependency was found between PU.1 and CEBPα in primary macrophages (Heinz et al., 2013). It is worth noting that such nucleosome-mediated, collaborative DNA binding could still be regarded as a form of indirect, cooperative DNA binding, as modeling has revealed an analogy between this process and the one involving cooperative binding of oxygen to hemoglobin (Mirny, 2010). Nevertheless, to avoid confusion, it may be best to continue to define this process as collaborative DNA binding.

Interestingly, several of these collaborating TFs have previously been defined as pioneer TFs that are uniquely able to access and open silent or compacted chromatin (Iwafuchi-Doi and Zaret, 2014). The fact that, at a wide range of loci, they are nevertheless dependent on other TFs to access DNA constitutes in this regard an intriguing paradox. For example, FOXA1 is defined as an archetypical pioneer TF (Mancini and West, 2015), given its ability to open closed chromatin by binding to DNA with its core DNA binding domain and to core histones with a binding motif that is located in its C terminus (Cirillo et al., 2002). However, its DNA binding interdependency with CEBPα or potentially other TFs suggests that FOXA1's "pioneering" ability may often not be sufficient to allow DNA binding, implying that FOXA1 requires the cumulative contribution of other TFs to displace nucleosomes and successfully unlock chromatin. This model may be consistent with the dispensability of FOXA1 and the related TF FOXA2 in maintaining the chromatin state in liver cells (Li et al., 2012). Similarly, PU.1 is also recognized as a pioneer factor for its ability to promote nucleosome depletion (Barozzi et al., 2014; Heinz et al., 2010), yet it depends on CEBP TFs to bind DNA at many genomic sites. What emerges is that, at some loci, these TFs may act as true pioneer TFs, whereas at others, they may require collaborations with other TFs to open chromatin.

Consequently, it is of interest to better understand what distinguishes genomic sites with pioneer activity from collaborative ones. An interesting observation in this regard is that regions with high PU.1 occupancy in primary macrophages had, in general, similar motif scores to those with lower PU.1 binding, but the two types of regions differed in nucleosome organization (Barozzi et al., 2014). Specifically, the latter regions were surrounded by two nucleosomes (in contrast to sites with high PU.1 binding) and showed enrichment for the NFκB motif, suggesting that, at those regions, PU.1 and NFκB need to collaborate to outcompete nucleosomes and thus to achieve high DNA occupancy. As such, TFs may have locus-dependent TF interdependencies reflecting both nucleosome structure and the presence of distinct TF motif clusters, consistent with the dependency of PU.1 on NFκB at some sites or either OCT2, BLIMP1, or STAT2 at others in human LCLs (Kilpinen et al., 2013). This view is also consistent with the flexible binding site grammar that is typically observed in enhancers in that the position of individual TF motifs within enhancers tends to be of secondary importance to their simple presence (Arnosti and Kulkarni, 2005). In other words, since collaborative DNA binding does not require physical contacts, the spacing and orientation of motifs can be flexible with respect to preserving enhancer activity, as long as the motifs are intact. This also implies that, at collaborative genomic sites, TFs should in principle bind to DNA in seemingly joint fashion since loss of one TF-DNA interaction (either in *cis* [e.g., because of a DNA mutation] or in *trans* [because of TF dysfunction]) would reduce the binding capacity of all other TFs at this locus. Such "collective" DNA binding behavior has indeed been observed for TFs mediating heart development in *Drosophila melanogaster* (Junion et al., 2012), and evidence for simultaneous, collaborative TF-DNA binding is also available for mammals (Adam et al., 2015; Siersbæk et al., 2014; Tijssen et al., 2011). A final illustration of this important notion is the dependency of the pioneer TF NRF1 (Sherwood et al., 2014) on other TFs to keep a specific set of its target sites from being methylated, which otherwise would block NRF1 DNA binding to these sites (Domcke et al., 2015). This example again illuminates how sequence context may affect the ability of a TF to bind independently to DNA, even if this TF may normally act as a pioneer factor. In sum, based on the currently available data, care needs to be taken when classifying TFs into specific categories without considering sequence and chromatin context.

Variable Chromatin Modules Mediate Long-Range TF-DNA Binding Interdependence

The previous sections highlighted that proximal variants can affect DNA binding through cooperative or collaborative mechanisms. However, many of the variants that drive TF-DNA binding

variation are located beyond the sequence span that is required for the formation of local TF-TF interactions or for competition with local nucleosomes, respectively. One possibility is that proteins overcome this distance restraint by inducing DNA looping through physical interactions. Even though this is an energetically costly process (Saiz and Vilar, 2006), both short- and long-range looping have now been extensively documented (de Wit et al., 2013; Gheldof et al., 2010; Lieberman-Aiden et al., 2009; Rao et al., 2014; Saiz and Vilar, 2006) and may play an important yet poorly understood role in mediating long-distance TF-DNA binding interdependencies.

It is therefore valuable to explore approaches to study the molecular origin of both short- and especially long-range TF-DNA binding variation. One such approach is the identification of genetic polymorphisms that significantly correlate with changes in DNA occupancy. Genomic regions in which such variants are located are interchangeably termed TF or binding quantitative trait loci (tfQTLs or bQTLs), as their detection suggests that a polymorphism within this locus causally affects the ability of a TF to bind to DNA. One study adopting this approach aimed to identify variants that affect DNA binding of the insulator protein CTCF by profiling its binding landscape in human LCLs using ChIP-seq, after which tfQTLs were explored within a 50 kb region centered around the CTCF binding region (Ding et al., 2014). Only a minority of detected tfQTLs overlapped the CTCF motif, even when the local LD structure was taken into account. A similar picture emerged from a comparable study on PU.1 DNA binding variation also in human LCLs, since PU.1 tfQTLs exhibited a bimodal log-normal distribution in terms of their distance to the PU.1 binding region (Waszak et al., 2015). The first mode represented tfQTLs that were located close to or at the PU.1 binding site and, consistent with the CTCF study, encompassed only a minority of the significantly associated variants. The second mode featured tfQTLs that were located distally to the PU.1 binding region with a median distance between 20 and 30 kb. Together, these findings suggest that many variable CTCF or PU.1 binding events are driven by long-distance mechanisms, which renders TF-DNA binding a complex molecular trait by itself. Consequently, and even though the effect size of distal tfQTLs tends to be inferior to that of proximal ones (Waszak et al., 2015), it will be valuable to decipher how these distal variants affect TF-DNA binding, given that they constitute the majority of DNA binding QTLs.

In the LCLs, PU.1 binding variation often correlated with variation in active chromatin marks such as H3K4me1 or H3K27ac—not only locally, but often over extended distances (Waszak et al., 2015). That is, high PU.1 DNA occupancy coincided with both high proximal and distal H3K4me1 and H3K27ac enrichment and vice versa. Such regions with a high level of molecular coordination between TF and chromatin marks have recently been termed "variable chromatin modules" (VCMs; Waszak et al., 2015; Figure 3A). Each VCM is thus composed of molecular phenotypes (e.g., the level of DNA occupancy by a TF or enrichment for a specific chromatin mark) that are highly coordinated, often over multiple kbp of DNA. More than 14,000 distinct VCMs were discovered in human LCLs, covering about 5% of the genome (Waszak et al., 2015). The majority of these were "totem" VCMs—so named because they were composed of stacked or overlapping molecular phenotypes that did not correlate with other neighboring molecular phenotypes. Thus, a totem VCM represents local chromatin state variation (Figure 3A). The remaining "multi-VCMs" are more interesting since, while a minority, they typically cover two or more distinct regulatory elements—hence the term "multi"—and capture the majority of all detected molecular phenotypes (Figure 3A). The origin of a "multi- VCM" is less intuitive than that of a totem-VCM. Its structure suggests, however, a higher-order chromatin organization that is reminiscent of the modular genomic structure that has been uncovered in the form of topologically associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012).

These TADs constitute distinct, three-dimensional genomic structures in which sequences are more likely to interact with one another than with those located outside the respective TAD. However, VCMs and TADs constitute different molecular entities because VCMs tend to be embedded within TADs and thus tend to be smaller (Waszak et al., 2015) (Figure 3B). In addition, TADs are relatively stable across cell types and during development and are even conserved across species (Dixon et al., 2012; Vietri Rudan et al., 2015), whereas VCMs are by definition variable. As such, multi-VCMs correspond conceptually better to sub-TADs, which are more fine-grained (sub-Mb), genomic topologies that have been shown to be dynamic across cellular differentiation (Dixon et al., 2015; Phillips-Cremins et al., 2013) and to even differ between individual cells (Giorgetti et al., 2014). In addition, sub-TADs have been suggested to define *cis*-regulatory networks (Berlivet et al., 2013), with their internal conformational dynamics being directly related to embedded transcriptional activity (Giorgetti et al., 2014; Tang et al., 2015). In parallel, the vast majority of gene-associated multi-VCMs exhibited a molecular activity state that significantly correlated with the transcriptional activity of the included gene(s) (Waszak et al., 2015) (Figure 3B). Moreover, the more regulatory elements encompassed in a VCM, the more likely it was to associate with variable gene expression. Together, the conceptual similarities between sub-TADs and multi-VCMs suggest that the latter also reflect fine-grained configurations of interacting regulatory elements with one or a few target genes whose collective, molecular activity is highly coordinated. As such, VCMs may provide substantial insights into the structural and thus modular organization of the chromatin landscape, including TF-DNA interactions.

Which mechanisms lie at the origin of multi-VCMs? Since the long-range molecular coordination that typifies multi-VCMs has been observed at the allelic level (Kasowski et al., 2013; Kilpinen et al., 2013; McVicker et al., 2013) and since recent chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) data has also provided evidence for allele-specific chromatin topologies (Tang et al., 2015), it is reasonable to assume that the observed molecular variation is largely driven by genetic factors. Moreover, most of the molecular variation within each VCM could be captured by a single, quantitative phenotype (Waszak et al., 2015), which suggests that the activity state of a VCM can be attributed to relatively few but strong causal variants. QTL mapping using the activity state of each VCM as input yielded vcmQTLs that were highly enriched in TF-occupied regions (Waszak et al., 2015) (Figure 3A). Together with previous

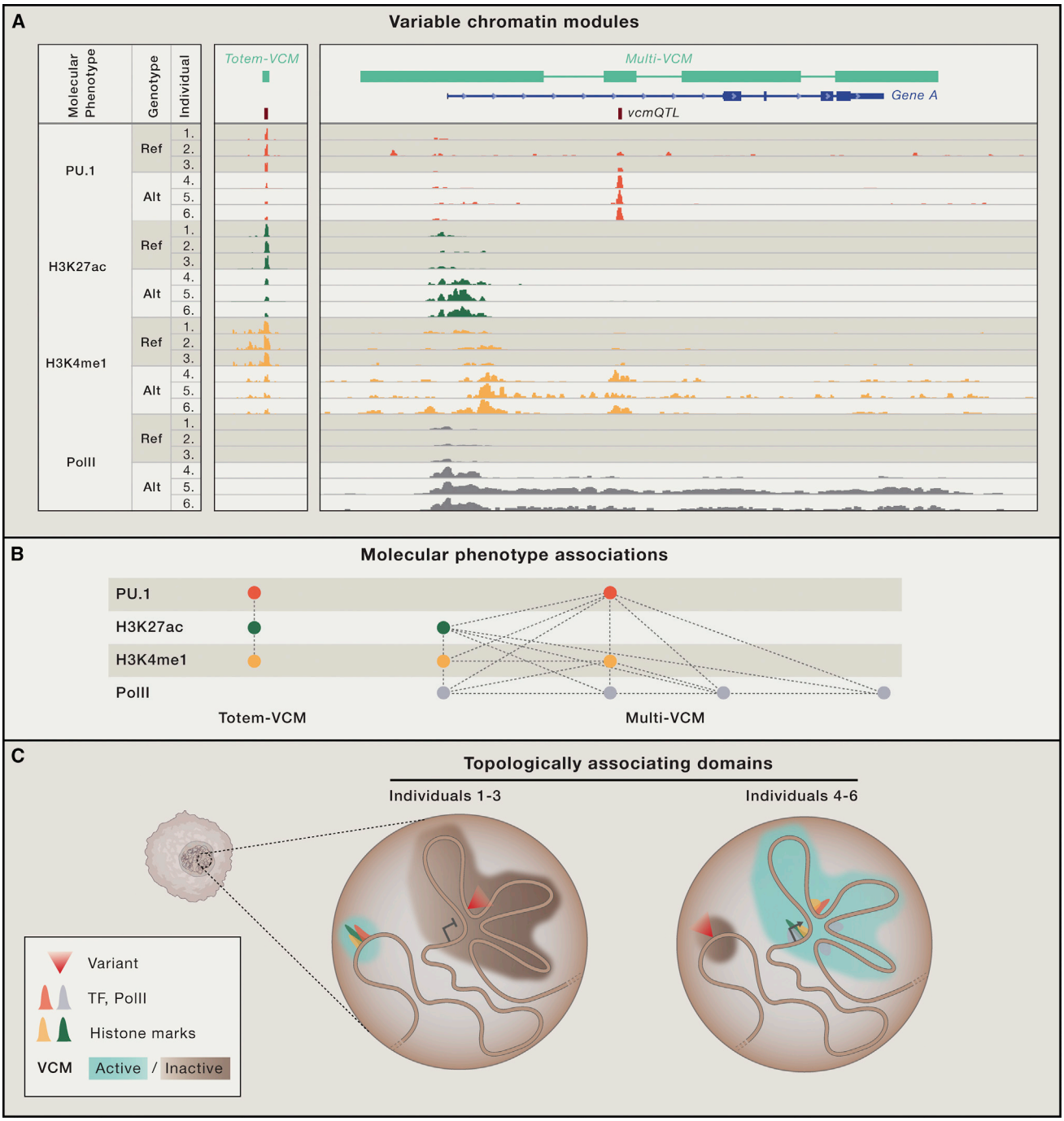

Figure 3. Variable Chromatin Modules

(A) Correlated TF (e.g., PU.1 or RNA polymerase II [PolII]) binding and chromatin mark (e.g., H3K27Ac, H3K4me1, H3K4me3) enrichment analyses across individuals allows the mapping of "variable chromatin modules" (VCMs) (shown in light green in the upper panel and in network format in the panel below). VCMs thus embody variable regions with highly coordinated, molecular phenotypes.

(B) The majority of VCMs have a "totem" structure of stacked molecular phenotypes that do not correlate with other neighboring molecular phenotypes and, as such, reflect local chromatin state variation. Multi-VCMs encompass sub-Mb regions involving distinct regulatory elements whose activity is highly coordinated and driven by a single or a few highly penetrating variants ("vcmQTL") with enrichment in TF-bound regions.

(C) VCMs constitute functional entities of higher-order chromatin organization embedded within topologically associating domains (TADs) and provide a molecular rationale as to how TF-DNA binding can be affected by distal genetic variation.

observations that TF-DNA binding perturbations are initiating drivers of downstream changes in chromatin state and gene expression (Kasowski et al., 2013; Kilpinen et al., 2013; McVicker et al., 2013) (Table 1), these findings support a model in which the alteration of one or a few TF binding events affects all molecular phenotypes in the respective VCM, including other embedded TF-DNA interactions (Figure 3B). Thus, the ability of a TF to bind to a VCM-associated genomic region appears to be a function of the respective VCM's activity state, which itself seems determined by one or few key TFs. As such, VCMs provide a conceptual framework to rationalize how distal genetic variation can affect TF-DNA binding.

Defining these key TFs remains a work in progress, since only few TFs reached significant enrichment in terms of their overlap with vcmQTLs (Waszak et al., 2015). This suggests that each VCM may have its own set of activity-determining TFs. Interestingly, distinct pairs of these same TFs were also enriched at pairs of regulatory elements that belonged to the same VCM (Waszak et al., 2015), suggesting that the functional interactions between these TFs (or among themselves) may be instrumental for forming VCMs. Together, the presented findings support a scenario in which the activity of each VCM is driven by a set of cell- and chromatin-context-specific TFs. This would be consistent with TF-DNA binding being highly dependent on proximal sequence environment and chromatin organization, which may differ from one VCM to the next. In addition, it would be compatible with the "multiple enhancer variant" hypothesis (Corradin et al., 2014), which dictates that linked variants in distinct regulatory elements often jointly contribute to gene expression variation. The VCM landscape may, as such, also be compatible with the LD structure of the genome.

What is the molecular nature of VCM-embedded TF-TF interactions? Based on the conceptual similarity between VCMs and sub-TADs and on how canonical enhancer-promoter interactions are established (Ciabrelli and Cavalli, 2015), it is conceivable that they are mediated by either direct physical contacts or by indirect protein-protein interactions involving more generic factors such as mediator, CTCF, and cohesins (Dekker and Mirny, 2016). The latter proteins may function to stabilize the interactions both with DNA and between TFs such that distal DNA binding interdependencies arise. However, other interaction-independent mechanisms could also underwrite such interdependencies, including long-range, transcription-, or repression-coupled chromatin remodeling processes (Hathaway et al., 2012; Smolle and Workman, 2013). Further experimentation will be required to elucidate the involvement and contributions of key individual TFs or TF pairs in VCM formation.

From Causal Variant to Complex Phenotype

While the identification and characterization of a trait- or disease-associated variant that causally disrupts TF-DNA binding is difficult, elucidating how it impacts on other potentially downstream molecular and biological processes may be equally if not more challenging (Edwards et al., 2013). One intuitive strategy to expand on the relatively few cases so far in which a causal relationship between molecular and phenotypic variation was established (Table 1) is the integration of other genetic or molecular data to infer the functional consequences of the focal variant. For example, distinct QTL datasets can be used to determine whether the variant impacts not only TF-DNA binding, but also the chromatin landscape, gene expression, or even other molecular phenotypes (Pai et al., 2015). The most common molecular QTL analysis, involving the identification of variants that associate with gene expression changes (i.e., eQTLs), is highly informative in this regard. For six distinct human populations, the most significant eQTLs were consistently found to overlap TF binding sites (Auton et al., 2015), thus providing direct insights into the identity of genes whose expression may be affected by variable TF-DNA binding.

However, other layers of molecular phenotypes—more associated with regulatory functions and therefore often defined as regulatory QTLs—can also be associated with genotypes. These include: (1) DNase I sensitivity (ds)QTLs that are strongly enriched in predicted TF binding sites in addition to being major determinants of gene expression variation (Degner et al., 2012); (2) chromatin (c)QTLs or histone marks (hm)QTLs that are largely concordant with TF-DNA binding and transcription (Grubert et al., 2015; Waszak et al., 2015), and (3) methylation (m)QTLs that also often exhibit a functional link with the other regulatory QTLs (Banovich et al., 2014; Domcke et al., 2015; Gutierrez-Arcelus et al., 2013; Heyn et al., 2013; McClay et al., 2015). Since regulatory QTLs as well as eQTLs were found to be enriched in complex trait or disease susceptibility variants (Albert and Kruglyak, 2015; Grubert et al., 2015; Nicolae et al., 2010; Waszak et al., 2015), their joint analysis may reveal how specific perturbations triggered by causal genetic variants in a certain condition or environment may first spread through transcriptional and other molecular networks before affecting the cellular, tissue, and finally organismal networks (Lehner, 2013; Mackay et al., 2009).

An intriguing observation that emerged from such analyses is that many regulatory QTLs do not overlap eQTLs, even if they overlap other types of regulatory QTLs (Degner et al., 2012; Grubert et al., 2015; Waszak et al., 2015). This is consistent with the well-established finding that many changes in TF-DNA binding have no measurable effect on gene expression (Cusanovich et al., 2014; Farnham, 2009). Thus, regulatory QTL analyses suffer from the same limitations as complex trait or disease susceptibility GWAS studies, i.e., difficulties in uncovering leading causal variants among LD blocks or in reaching statistical significance without a high number of samples (Veyrieras et al., 2008). Approaches that link chromatin organization to transcriptional function such as ChIA-PET (Dowen et al., 2014; Tang et al., 2015) or VCM mapping (Waszak et al., 2015) may in this regard prove valuable, as they can provide a structural framework for interpreting regulatory variation. Indeed, as regulatory variants tend to impact different layers of molecular phenotypes, it is intrinsically valuable to know how these layers are coordinated across distinct genomic domains. For example, many VCMs were identified that consisted of active chromatin marks as well as TF binding sites, even though an important portion of such VCMs did not vary along with the expression of neighboring genes. These VCMs, termed "island VCMs" (Waszak et al., 2015), thus represent coordinated changes in TF binding and chromatin state without measurable impact on gene expression. Accordingly, QTLs for such island VCMs tend to overlap with

tfQTL and cQTLs, but not with eQTLs. There are several complementary hypotheses that could explain the existence of such island VCMs, including (1) "futile" regulatory activity without transcriptional consequences (Cusanovich et al., 2014; Farnham, 2009; Wasserman and Sandelin, 2004); (2) regulatory redundancy, which prevents a gene-specific regulatory network from collapsing even if one node or edge is impacted (Pai et al., 2015), consistent with the shadow enhancer concept (Hong et al., 2008); (3) regulatory regions that are not transcriptionally operational, at least in the studied condition/cellular environment, which implies that the activity of these regions is tissue specific. Indeed, if, in a hypothetical study, a complex trait-associated regulatory variant would be linked to an island VCM, it might indicate that an incorrect system or context is being studied, as its disconnection with gene expression is unlikely to yield a cellular or organismal phenotype. This reasoning is consistent with the observation across several studies that GWAS variants tend to be most enriched for eQTLs in tissues that are relevant to the phenotype (Emilsson et al., 2008; Nica et al., 2010; Torres et al., 2014). However, in most cases, the causal variants are obviously unknown a priori. To identify them, it may prove valuable to, similar to eQTLs, map VCMs in as many distinct cell types/tissues as possible. The resulting set of VCMs may then provide guidance to both variant identification and characterization. Indeed, the most interesting candidates among the set of associated GWAS variants would be those that impact not only on the chromatin topology (e.g., vcmQTLs) and state of the respective locus (e.g., cQTLs or tfQTLs), but also on expression of the embedded gene. Once identified, it should be relatively straightforward to detangle the underlying molecular mechanisms since the coordinated, molecular phenotypes that make up the focal VCM should provide clear insights into the flow of regulatory information, i.e., from causal nucleotide over gene to ultimately cellular or organismal phenotype.

Conclusions

The fundamental discovery that most complex trait-associated variants are located in non-coding, putatively regulatory regions of the genome has focused the spotlight on TF-DNA interactions as important mediators of phenotypic variation. Yet, to date, relatively few examples are available in which a clear mechanistic relationship between TF-DNA binding variation and phenotypic variation was established (Table 1). To clarify why this is such a challenging task, we focused in this Review on elucidating how the impact of genetic variation on TF-DNA binding can be assessed and why, contrary to expectations, this is itself already inherently complex. There are several current limitations that will have to be addressed to improve our ability to identify and interpret regulatory variation, including the need for new experimental or computational approaches that will enable us to expand the TF motif catalog, to better predict genuine TF binding sites, and to evaluate how motif variation affects TF-DNA binding. Promising research avenues in this regard include the development of new technologies to characterize monomeric and higher complex TF-DNA binding properties and the incorporation of additional DNA binding features such as the sequence environment and the conformational and chemical nature of DNA in machine learning approaches. In addition, it is increasingly appreciated that the chromatin context needs to be accounted for when searching for causal, regulatory variants and that, in general, the use of cell types or systems that are most relevant for the studied trait or disease will yield the best results. It is also important to recognize that only a small fraction of all variable TF-DNA binding events is actually driven by variation within the motif of the studied TF. Thus, similar to gene expression, TF-DNA binding is a complex molecular trait by itself, which has profound implications for our understanding of how regulatory variation arises.

Well-established concepts in the gene regulation field provide an intuitive molecular foundation for local or proximal variant-driven DNA binding variation. Specifically, the former involves cooperative DNA binding that is mediated by direct, physical interactions between TFs, while the latter appears to be driven by collaborative DNA binding that is likely reflective of sequence- or chromatin-context conditioned TF interdependencies to displace nucleosomes and open chromatin. However, the mechanisms that underlie distal variant-driven DNA binding changes are much less well understood (Figure 2). The identification of 3C-, ChIA-PET-, or VCM-based chromatin entities that link structural information to transcriptional function is important in this regard since they offer a molecular rationale to explain these prevalent, long-range DNA binding dependencies. Sustained efforts will therefore be required to unravel the modular structure of the variable (epi)genome across a wide range of cells or tissues. Thus, although many challenges remain, exciting progress is being made in elucidating the genetic basis of TF-DNA binding variation that will undoubtedly improve our ability to achieve a nucleotide-level understanding of the molecular mechanisms underlying many complex traits, including disease susceptibility.

SUPPLEMENTAL INFORMATION

Supplemental Information includes one table and is available with this article online at http://dx.doi.org/10.1016/j.cell.2016.07.012.

ACKNOWLEDGMENTS

We thank Richard Benton (University of Lausanne), Sebastian Waszak (EMBL), Alina Isakova, Antonio Meireles-Filho, Petra Schwalie, and other members of the Deplancke Laboratory, as well as the anonymous reviewers for useful comments on the manuscript. We also would like to acknowledge scientific discussions with all members of the "Effect of sequence variation on chromatin structure and transcription" Sinergia Consortium (i.e., the Reymond and Hernandez Laboratories [UNIL] and the Dermitzakis Laboratory [University of Geneva]). This work was supported by the Swiss National Science Foundation grant CRSI33_130326, by SystemsX.ch (AgingX, 51RTP0_151019), and by institutional support from the Swiss Federal Institute of Technology in Lausanne (EPFL).

REFERENCES

Adam, R.C., Yang, H., Rockowitz, S., Larsen, S.B., Nikolova, M., Oristian, D.S., Polak, L., Kadaja, M., Asare, A., Zheng, D., and Fuchs, E. (2015). Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. Nature *521*, 366–370.

Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. *16*, 197–212.

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

Al Zadjali, S., Wali, Y., Al Lawatiya, F., Gravell, D., Alkindi, S., Al Falahi, K., Krishnamoorthy, R., and Daar, S. (2011). The β-globin promoter -71 C>T mutation is a β+ thalassemic allele. Eur. J. Haematol. *87*, 457–460.

Ameur, A., Rada-Iglesias, A., Komorowski, J., and Wadelius, C. (2009). Identification of candidate regulatory SNPs by combination of transcription-factor-binding site prediction, SNP genotyping and haploChIP. Nucleic Acids Res. *37*, e85.

Arnosti, D.N., and Kulkarni, M.M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J. Cell. Biochem. *94*, 890–898.

Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Banovich, N.E., Lan, X., McVicker, G., van de Geijn, B., Degner, J.F., Blischak, J.D., Roux, J., Pritchard, J.K., and Gilad, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS Genet. *10*, e1004663.

Barozzi, I., Simonatto, M., Bonifacio, S., Yang, L., Rohs, R., Ghisletti, S., and Natoli, G. (2014). Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. Mol. Cell *54*, 844–857.

Barrera, L.A., Vedenko, A., Kurland, J.V., Rogers, J.M., Gisselbrecht, S.S., Rossin, E.J., Woodard, J., Mariani, L., Kock, K.H., Inukai, S., et al. (2016). Survey of variation in human transcription factors reveals prevalent DNA binding changes. Science *351*, 1450–1454.

Bauer, A.L., Hlavacek, W.S., Unkefer, P.J., and Mu, F. (2010). Using sequence-specific chemical and structural properties of DNA to predict transcription factor binding sites. PLoS Comput. Biol. *6*, e1001007.

Benko, S., Fantes, J.A., Amiel, J., Kleinjan, D.-J., Thomas, S., Ramsay, J., Jamshidi, N., Essafi, A., Heaney, S., Gordon, C.T., et al. (2009). Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. Nat. Genet. *41*, 359–364.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. *24*, 1429–1435.

Berlivet, S., Paquette, D., Dumouchel, A., Langlais, D., Dostie, J., and Kmita, M. (2013). Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs. PLoS Genet. *9*, e1004018.

Biggin, M.D. (2011). Animal transcription networks as highly connected, quantitative continua. Dev. Cell *21*, 611–626.

Bulyk, M.L., Johnson, P.L., and Church, G.M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res. *30*, 1255–1261.

Chorley, B.N., Wang, X., Campbell, M.R., Pittman, G.S., Noureddine, M.A., and Bell, D.A. (2008). Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. Mutat. Res. *659*, 147–157.

Ciabrelli, F., and Cavalli, G. (2015). Chromatin-driven behavior of topologically associating domains. J. Mol. Biol. *427*, 608–625.

Cirillo, L.A., Lin, F.R., Cuesta, I., Friedman, D., Jarnik, M., and Zaret, K.S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. Mol. Cell *9*, 279–289.

Claussnitzer, M., Dankel, S.N., Klocke, B., Grallert, H., Glunk, V., Berulava, T., Lee, H., Oskolkov, N., Fadista, J., Ehlers, K., et al.; DIAGRAM+Consortium (2014). Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. Cell *156*, 343–358.

Claussnitzer, M., Dankel, S.N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N. Engl. J. Med. *373*, 895–907.

Consortium, T.E.; ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal lari, R., Lupien, M., Markowitz, S., and Scacheri, P.C. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. *24*, 1–13.

Crossley, M., and Brownlee, G.G. (1990). Disruption of a C/EBP binding site in the factor IX promoter is associated with haemophilia B. Nature *345*, 444–446.

Crossley, M., Ludwig, M., Stowell, K.M., De Vos, P., Olek, K., and Brownlee, G.G. (1992). Recovery from hemophilia B Leyden: an androgen-responsive element in the factor IX promoter. Science *257*, 377–379.

Cusanovich, D.A., Pavlovic, B., Pritchard, J.K., and Gilad, Y. (2014). The functional consequences of variation in transcription factor binding. PLoS Genet. *10*, e1004226.

de Wit, E., Bouwman, B.A.M., Zhu, Y., Klous, P., Splinter, E., Verstegen, M.J.A.M., Krijger, P.H.L., Festuccia, N., Nora, E.P., Welling, M., et al. (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature *501*, 227–231.

Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390–394.

De Gobbi, M., Viprakasit, V., Hughes, J.R., Fisher, C., Buckle, V.J., Ayyub, H., Gibbons, R.J., Vernimmen, D., Yoshinaga, Y., de Jong, P., et al. (2006). A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science *312*, 1215–1217.

Dekker, J., and Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. Cell *164*, 1110–1121.

Dina, C., Meyre, D., Gallina, S., Durand, E., Körner, A., Jacobson, P., Carlsson, L.M., Kiess, W., Vatin, V., Lecoeur, C., et al. (2007). Variation in FTO contributes to childhood obesity and severe adult obesity. Nat. Genet. *39*, 724–726.

Ding, Z., Ni, Y., Timmer, S.W., Lee, B.-K., Battenhouse, A., Louzada, S., Yang, F., Dunham, I., Crawford, G.E., Lieb, J.D., et al. (2014). Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. PLoS Genet. *10*, e1004798.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

Djordjevic, M., Sengupta, A.M., and Shraiman, B.I. (2003). A biophysical approach to transcription factor binding site discovery. Genome Res. *13*, 2381–2390.

Dodd, A.W., Syddall, C.M., and Loughlin, J. (2013). A rare variant in the osteoarthritis-associated locus GDF5 is functional and reveals a site that can be manipulated to modulate GDF5 expression. Eur. J. Hum. Genet. *21*, 517–521.

Domcke, S., Bardet, A.F., Adrian Ginno, P., Hartl, D., Burger, L., and Schübeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. Nature *528*, 575–579.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., and Young, R.A. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374–387.

Dror, I., Golan, T., Levy, C., Rohs, R., and Mandel-Gutfreund, Y. (2015). A widespread role of the motif environment in transcription factor binding across diverse protein families. Genome Res. *25*, 1268–1280.

Edwards, S.L., Beesley, J., French, J.D., and Dunning, A.M. (2013). Beyond GWASs: illuminating the dark road from association to function. Am. J. Hum. Genet. *93*, 779–797.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., et al. (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423–428.

Escalante, C.R., Brass, A.L., Pongubala, J.M.R., Shatova, E., Shen, L., Singh, H., and Aggarwal, A.K. (2002). Crystal structure of PU.1/IRF-4/DNA ternary complex. Mol. Cell *10*, 1097–1105.

Faisst, S., and Meyer, S. (1992). Compilation of vertebrate-encoded transcription factors. Nucleic Acids Res. *20*, 3–26.

Farnham, P.J. (2009). Insights from genomic profiling of transcription factors. Nat. Rev. Genet. *10*, 605–616.

Fischer, J., Koch, L., Emmerling, C., Vierkotten, J., Peters, T., Brüning, J.C., and Rüther, U. (2009). Inactivation of the Fto gene protects from obesity. Nature *458*, 894–898.

Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R., Elliott, K.S., Lango, H., Rayner, N.W., et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science *316*, 889–894.

French, J.D., Ghoussaini, M., Edwards, S.L., Meyer, K.B., Michailidou, K., Ahmed, S., Khan, S., Maranian, M.J., O'Reilly, M., Hillman, K.M., et al.; GENICA Network; kConFab Investigators (2013). Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. Am. J. Hum. Genet. *92*, 489–503.

Funnell, A.P., Wilson, M.D., Ballester, B., Mak, K.S., Burdach, J., Magan, N., Pearson, R.C., Lemaigre, F.P., Stowell, K.M., Odom, D.T., et al. (2013). A CpG mutational hotspot in a ONECUT binding site accounts for the prevalent variant of hemophilia B Leyden. Am. J. Hum. Genet. *92*, 460–467.

Gao, Z., and Ruan, J. (2015). A structure-based Multiple-Instance Learning approach to predicting in vitro transcription factor-DNA interaction. BMC Genomics *16* (*Suppl 4*), S3.

Gelfond, J.A., Gupta, M., and Ibrahim, J.G. (2009). A Bayesian hidden Markov model for motif discovery through joint modeling of genomic sequence and ChIP-chip data. Biometrics *65*, 1087–1095.

Gheldof, N., Smith, E.M., Tabuchi, T.M., Koch, C.M., Dunham, I., Stamatoyannopoulos, J.A., and Dekker, J. (2010). Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. Nucleic Acids Res. *38*, 4325–4336.

Giorgetti, L., Galupa, R., Nora, E.P., Piolot, T., Lam, F., Dekker, J., Tiana, G., and Heard, E. (2014). Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell *157*, 950–963.

Gosselin, D., Link, V.M., Romanoski, C.E., Fonseca, G.J., Eichenfield, D.Z., Spann, N.J., Stender, J.D., Chun, H.B., Garner, H., Geissmann, F., and Glass, C.K. (2014). Environment drives selection and function of enhancers controlling tissue-specific macrophage identities. Cell *159*, 1327–1340.

Gotea, V., Visel, A., Westlund, J.M., Nobrega, M.A., Pennacchio, L.A., and Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res. *20*, 565–577.

Gragnoli, C., Lindner, T., Cockburn, B.N., Kaisaki, P.J., Gragnoli, F., Marozzi, G., and Bell, G.I. (1997). Maturity-onset diabetes of the young due to a mutation in the hepatocyte nuclear factor-4 alpha binding site in the promoter of the hepatocyte nuclear factor-1 alpha gene. Diabetes *46*, 1648–1651.

Grant, S.F., Reid, D.M., Blake, G., Herd, R., Fogelman, I., and Ralston, S.H. (1996). Reduced bone density and osteoporosis associated with a polymorphic Sp1 binding site in the collagen type I alpha 1 gene. Nat. Genet. *14*, 203–205.

Grubert, F., Zaugg, J.B., Kasowski, M., Ursu, O., Spacek, D.V., Martin, A.R., Greenside, P., Srivas, R., Phanstiel, D.H., Pekowska, A., et al. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell *162*, 1051–1065.

Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S.B., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., Giger, T., Romano, L., Planchon, A., et al. (2013). Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife *2*, e00523.

Hathaway, N.A., Bell, O., Hodges, C., Miller, E.L., Neel, D.S., and Crabtree, G.R. (2012). Dynamics and memory of heterochromatin in living cells. Cell *149*, 1447–1460.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Heinz, S., Romanoski, C.E., Benner, C., Allison, K.A., Kaikkonen, M.U., Orozco, L.D., and Glass, C.K. (2013). Effect of natural genetic variation on enhancer selection and function. Nature *503*, 487–492.

Helms, C., Cao, L., Krueger, J.G., Wijsman, E.M., Chamian, F., Gordon, D., Heffernan, M., Daw, J.A., Robarge, J., Ott, J., et al. (2003). A putative RUNX1 binding site variant between SLC9A3R1 and NAT9 is associated with susceptibility to psoriasis. Nat. Genet. *35*, 349–356.

Heyn, H., Moran, S., Hernando-Herraez, I., Sayols, S., Gomez, A., Sandoval, J., Monk, D., Hata, K., Marques-Bonet, T., Wang, L., and Esteller, M. (2013). DNA methylation contributes to natural human variation. Genome Res. *23*, 1363–1372.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.

Hobbs, K., Negri, J., Klinnert, M., Rosenwasser, L.J., and Borish, L. (1998). Interleukin-10 and transforming growth factor-beta promoter polymorphisms in allergies and asthma. Am. J. Respir. Crit. Care Med. *158*, 1958–1962.

Hoekstra, H.E., and Coyne, J.A. (2007). The locus of evolution: evo devo and the genetics of adaptation. Evolution *61*, 995–1016.

Hong, J.-W., Hendrix, D.A., and Levine, M.S. (2008). Shadow enhancers as a source of evolutionary novelty. Science *321*, 1314.

Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., et al. (2013). TERT promoter mutations in familial and sporadic melanoma. Science *339*, 959–961.

Hu, S., Xie, Z., Onishi, A., Yu, X., Jiang, L., Lin, J., Rho, H.S., Woodard, C., Wang, H., Jeong, J.-S., et al. (2009). Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. Cell *139*, 610–622.

Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H.N., Shin, J., Cox, E., Rho, H.S., Woodard, C., et al. (2013). DNA methylation presents distinct binding sites for human transcription factors. eLife *2*, e00726.

Huang, F.W., Hodis, E., Xu, M., Kryukov, G.V., Chin, L., and Garraway, L.A. (2013). Highly recurrent TERT promoter mutations in human melanoma. Science *339*, 957–959.

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A.M., Turlapati, L., Zichner, T., Zhu, D., Lyman, R.F., et al. (2014). Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Res. *24*, 1193–1208.

Isakova, A., Berset, Y., Hatzimanikatis, V., and Deplancke, B. (2016). Quantification of Cooperativity in Heterodimer-DNA Binding Improves the Accuracy of Binding Specificity Models. J. Biol. Chem. *291*, 10293–10306.

Iwafuchi-Doi, M., and Zaret, K.S. (2014). Pioneer transcription factors in cell reprogramming. Genes Dev. *28*, 2679–2692.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. *3*, 318–356.

Jeong, Y., Leskow, F.C., El-Jaick, K., Roessler, E., Muenke, M., Yocum, A., Dubourg, C., Li, X., Geng, X., Oliver, G., and Epstein, D.J. (2008). Regulation of a remote Shh forebrain enhancer by the Six3 homeoprotein. Nat. Genet. *40*, 1348–1353.

Jia, L., Landan, G., Pomerantz, M., Jaschek, R., Herman, P., Reich, D., Yan, C., Khalid, O., Kantoff, P., Oh, W., et al. (2009). Functional enhancers at the gene-poor 8q24 cancer-linked locus. PLoS Genet. *5*, e1000597.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpää, M.J., et al. (2010). Multiplexed massively

parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. *20*, 861–873.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Jolma, A., Yin, Y., Nitta, K.R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E., and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384–388.

Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E.H., Birney, E., and Furlong, E.E. (2012). A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell *148*, 473–486.

Karczewski, K.J., Tatonetti, N.P., Landt, S.G., Yang, X., Slifer, T., Altman, R.B., and Snyder, M. (2011). Cooperative transcription factor associations discovered using regulatory variation. Proc. Natl. Acad. Sci. USA *108*, 13353–13358.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M., Habegger, L., Rozowsky, J., Shi, M., Urban, A.E., et al. (2010). Variation in transcription factor binding among humans. Science *328*, 232–235.

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., et al. (2013). Extensive variation in chromatin states across humans. Science *342*, 750–752.

Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature *477*, 289–294.

Kilpinen, H., Waszak, S.M., Gschwind, A.R., Raghav, S.K., Witwicki, R.M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N.I., et al. (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science *342*, 744–747.

Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Soboleva, A.V., Kasianov, A.S., Ashoor, H., Ba-Alawi, W., Bajic, V.B., Medvedeva, Y.A., Kolpakov, F.A., and Makeev, V.J. (2016). HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Res. *44* (D1), D116–D125.

Kulzer, J.R., Stitzel, M.L., Morken, M.A., Huyghe, J.R., Fuchsberger, C., Kuusisto, J., Laakso, M., Boehnke, M., Collins, F.S., and Mohlke, K.L. (2014). A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. Am. J. Hum. Genet. *94*, 186–197.

Lecerf, L., Kavo, A., Ruiz-Ferrer, M., Baral, V., Watanabe, Y., Chaoui, A., Pingault, V., Borrego, S., and Bondurand, N. (2014). An impairment of long distance SOX10 regulatory elements underlies isolated Hirschsprung disease. Hum. Mutat. *35*, 303–307.

Lehner, B. (2013). Genotype to phenotype: lessons from model organisms for human genetics. Nat. Rev. Genet. *14*, 168–178.

Levo, M., and Segal, E. (2014). In pursuit of design principles of regulatory sequences. Nat. Rev. Genet. *15*, 453–468.

Li, Z., Gadue, P., Chen, K., Jiao, Y., Tuteja, G., Schug, J., Li, W., and Kaestner, K.H. (2012). Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. Cell *151*, 1608–1616.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lowe, W.L., Jr., and Reddy, T.E. (2015). Genomic approaches for understanding the genetics of complex disease. Genome Res. *25*, 1432–1441.

Ludlow, L.B., Schick, B.P., Budarf, M.L., Driscoll, D.A., Zackai, E.H., Cohen, A., and Konkle, B.A. (1996). Identification of a mutation in a GATA binding site of the platelet glycoprotein Ibbeta promoter resulting in the Bernard-Soulier syndrome. J. Biol. Chem. *271*, 22076–22080.

Lynch, V.J., and Wagner, G.P. (2008). Resurrecting the role of transcription factor change in developmental evolution. Evolution *62*, 2131–2154.

Mackay, T.F., Stone, E.A., and Ayroles, J.F. (2009). The genetics of quantitative traits: challenges and prospects. Nat. Rev. Genet. *10*, 565–577.

Maerkl, S.J., and Quake, S.R. (2009). Experimental determination of the evolvability of a transcription factor. Proc. Natl. Acad. Sci. USA *106*, 18650–18655.

Maienschein-Cline, M., Dinner, A.R., Hlavacek, W.S., and Mu, F. (2012). Improved predictions of transcription factor binding sites using physicochemical features of DNA. Nucleic Acids Res. *40*, e175.

Mancini, E.J., and West, M.J. (2015). How to Be a Pioneer: A One-Sided View. Trends Biochem. Sci. *40*, 547–548.

Manco, L., Ribeiro, M.L., Máximo, V., Almeida, H., Costa, A., Freitas, O., Barbot, J., Abade, A., and Tamagnini, G. (2000). A new PKLR gene mutation in the R-type promoter region affects the gene transcription causing pyruvate kinase deficiency. Br. J. Haematol. *110*, 993–997.

Manolio, T.A. (2013). Bringing genome-wide association findings into clinical use. Nat. Rev. Genet. *14*, 549–558.

Mansour, M.R., Abraham, B.J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A.D., Etchin, J., Lawton, L., Sallan, S.E., Silverman, L.B., et al. (2014). Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science *346*, 1373–1377.

Martin, D.I., Tsai, S.F., and Orkin, S.H. (1989). Increased gamma-globin expression in a nondeletion HPFH mediated by an erythroid-specific DNA-binding factor. Nature *338*, 435–438.

Masotti, C., Armelin-Correa, L.M., Splendore, A., Lin, C.J., Barbosa, A., Sogayar, M.C., and Passos-Bueno, M.R. (2005). A functional SNP in the promoter region of TCOF1 is associated with reduced gene expression and YY1 DNA-protein interaction. Gene *359*, 44–52.

Massouras, A., Waszak, S.M., Albarca-Aguilera, M., Hens, K., Holcombe, W., Ayroles, J.F., Dermitzakis, E.T., Stone, E.A., Jensen, J.D., Mackay, T.F., and Deplancke, B. (2012). Genomic variation and its impact on gene expression in Drosophila melanogaster. PLoS Genet. *8*, e1003055.

Matsuda, M., Sakamoto, N., and Fukumaki, Y. (1992). Delta-thalassemia caused by disruption of the site for an erythroid-specific transcription factor, GATA-1, in the delta-globin gene promoter. Blood *80*, 1347–1351.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

Maurano, M.T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R., and Stamatoyannopoulos, J.A. (2015). Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. Nat. Genet. *47*, 1393–1401.

McClay, J.L., Shabalin, A.A., Dozmorov, M.G., Adkins, D.E., Kumar, G., Nerella, S., Clark, S.L., Bergen, S.E., Hultman, C.M., Magnusson, P.K., et al.; Swedish Schizophrenia Consortium (2015). High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. Genome Biol. *16*, 291.

McVicker, G., van de Geijn, B., Degner, J.F., Cain, C.E., Banovich, N.E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J.K. (2013). Identification of genetic variants that affect histone modifications in human cells. Science *342*, 747–749.

Meng, X., Brodsky, M.H., and Wolfe, S.A. (2005). A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat. Biotechnol. *23*, 988–994.

Miller, I.J., and Bieker, J.J. (1993). A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the Krüppel family of nuclear proteins. Mol. Cell. Biol. *13*, 2776–2786.

Mirny, L.A. (2010). Nucleosome-mediated cooperativity between transcription factors. Proc. Natl. Acad. Sci. USA *107*, 22534–22539.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature *466*, 714–719.

Najafabadi, H.S., Mnaimneh, S., Schmitges, F.W., Garton, M., Lam, K.N., Yang, A., Albu, M., Weirauch, M.T., Radovani, E., Kim, P.M., et al. (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat. Biotechnol. *33*, 555–562.

Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. *6*, e1000895.

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. *6*, e1000888.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Nutiu, R., Friedman, R.C., Luo, S., Khrebtukova, I., Silva, D., Li, R., Zhang, L., Schroth, G.P., and Burge, C.B. (2011). Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nat. Biotechnol. *29*, 659–664.

Orkin, S.H., Kazazian, H.H., Jr., Antonarakis, S.E., Goff, S.C., Boehm, C.D., Sexton, J.P., Waber, P.G., and Giardina, P.J. (1982). Linkage of beta-thalassaemia mutations and beta-globin gene polymorphisms with DNA polymorphisms in human beta-globin gene cluster. Nature *296*, 627–631.

Pai, A.A., Pritchard, J.K., and Gilad, Y. (2015). The genetic and mechanistic basis for variation in gene regulation. PLoS Genet. *11*, e1004857.

Peters, T., Ausmeier, K., and Rüther, U. (1999). Cloning of Fatso (Fto), a novel gene deleted by the Fused toes (Ft) mouse mutation. Mamm. Genome *10*, 983–986.

Phillips-Cremins, J.E., Sauria, M.E., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S., Ong, C.T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295.

Polach, K.J., and Widom, J. (1996). A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. J. Mol. Biol. *258*, 800–812.

Poncz, M., Ballantine, M., Solowiejczyk, D., Barak, I., Schwartz, E., and Surrey, S. (1982). beta-Thalassemia in a Kurdish Jew. Single base changes in the T-A-T-A box. J. Biol. Chem. *257*, 5994–5996.

Ponomarenko, J.V., Merkulova, T.I., Vasiliev, G.V., Levashova, Z.B., Orlova, G.V., Lavryushev, S.V., Fokin, O.N., Ponomarenko, M.P., Frolov, A.S., and Sarai, A. (2001). rSNP_Guide, a database system for analysis of transcription factor binding to target sequences: application to SNPs and site-directed mutations. Nucleic Acids Res. *29*, 312–316.

Ptashne, M., Jeffrey, A., Johnson, A.D., Maurer, R., Meyer, B.J., Pabo, C.O., Roberts, T.M., and Sauer, R.T. (1980). How the λ repressor and cro work. Cell *19*, 1–11.

Raghav, S.K., Waszak, S.M., Krier, I., Gubelmann, C., Isakova, A., Mikkelsen, T.S., and Deplancke, B. (2012). Integrative genomics identifies the corepressor SMRT as a gatekeeper of adipogenesis through the transcription factors C/EBPβ and KAISO. Mol. Cell *46*, 335–350.

Rahimov, F., Marazita, M.L., Visel, A., Cooper, M.E., Hitchler, M.J., Rubini, M., Domann, F.E., Govil, M., Christensen, K., Bille, C., et al. (2008). Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. Nat. Genet. *40*, 1341–1347.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., Bajic, V.B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. Cell *140*, 744–752.

Reddy, T.E., Gertz, J., Pauli, F., Kucera, K.S., Varley, K.E., Newberry, K.M., Marinov, G.K., Mortazavi, A., Williams, B.A., Song, L., et al. (2012). Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. Genome Res. *22*, 860–869.

Reijnen, M.J., Sladek, F.M., Bertina, R.M., and Reitsma, P.H. (1992). Disruption of a binding site for hepatocyte nuclear factor 4 results in hemophilia B Leyden. Proc. Natl. Acad. Sci. USA *89*, 6300–6303.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248–1253.

Saiz, L., and Vilar, J.M.G. (2006). DNA looping: the consequences and its control. Curr. Opin. Struct. Biol. *16*, 344–350.

Sherwood, R.I., Hashimoto, T., O'Donnell, C.W., Lewis, S., Barkal, A.A., van Hoff, J.P., Karun, V., Jaakkola, T., and Gifford, D.K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat. Biotechnol. *32*, 171–178.

Siersbæk, R., Rabiee, A., Nielsen, R., Sidoli, S., Traynor, S., Loft, A., La Cour Poulsen, L., Rogowska-Wrzesinska, A., Jensen, O.N., and Mandrup, S. (2014). Transcription factor cooperativity in early adipogenic hotspots and super-enhancers. Cell Rep. *7*, 1443–1455.

Simicevic, J., Schmid, A.W., Gilardoni, P.A., Zoller, B., Raghav, S.K., Krier, I., Gubelmann, C., Lisacek, F., Naef, F., Moniatte, M., and Deplancke, B. (2013). Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics. Nat. Methods *10*, 570–576.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends Biochem. Sci. *39*, 381–399.

Smemo, S., Tena, J.J., Kim, K.H., Gamazon, E.R., Sakabe, N.J., Gómez-Marín, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature *507*, 371–375.

Smolle, M., and Workman, J.L. (2013). Transcription-associated histone modifications and cryptic transcription. Biochim. Biophys. Acta *1829*, 84–97.

Soccio, R.E., Chen, E.R., Rajapurkar, S.R., Safabakhsh, P., Marinis, J.M., Dispirito, J.R., Emmett, M.J., Briggs, E.R., Fang, B., Everett, L.J., et al. (2015). Genetic Variation Determines PPARγ Function and Anti-diabetic Drug Response In Vivo. Cell *162*, 33–44.

Solis, C., Aizencang, G.I., Astrin, K.H., Bishop, D.F., and Desnick, R.J. (2001). Uroporphyrinogen III synthase erythroid promoter mutations in adjacent GATA1 and CP2 elements cause congenital erythropoietic porphyria. J. Clin. Invest. *107*, 753–762.

Soufi, A., Garcia, M.F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K.S. (2015). Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell *161*, 555–568.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Spivakov, M., Akhtar, J., Kheradpour, P., Beal, K., Girardot, C., Koscielny, G., Herrero, J., Kellis, M., Furlong, E.E., and Birney, E. (2012). Analysis of variation at transcription factor binding sites in Drosophila and humans. Genome Biol. *13*, R49.

Stefflova, K., Thybert, D., Wilson, M.D., Streeter, I., Aleksic, J., Karagianni, P., Brazma, A., Adams, D.J., Talianidis, I., Marioni, J.C., et al. (2013). Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell *154*, 530–540.

Stormo, G.D., and Zhao, Y. (2010). Determining the specificity of protein-DNA interactions. Nat. Rev. Genet. *11*, 751–760.

Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., et al. (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell *163*, 1611–1627.

Tijssen, M.R., Cvejic, A., Joshi, A., Hannah, R.L., Ferreira, R., Forrai, A., Bellissimo, D.C., Oram, S.H., Smethurst, P.A., Wilson, N.K., et al. (2011). Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. Dev. Cell *20*, 597–609.

Torres, J.M., Gamazon, E.R., Parra, E.J., Below, J.E., Valladares-Salgado, A., Wacher, N., Cruz, M., Hanis, C.L., and Cox, N.J. (2014). Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. Am. J. Hum. Genet. *95*, 521–534.

Tournamille, C., Colin, Y., Cartron, J.P., and Le Van Kim, C. (1995). Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. Nat. Genet. *10*, 224–228.

Tuupanen, S., Turunen, M., Lehtonen, R., Hallikas, O., Vanharanta, S., Kivioja, T., Björklund, M., Wei, G., Yan, J., Niittymäki, I., et al. (2009). The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat. Genet. *41*, 885–890.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. *10*, 252–263.

Verlaan, D.J., Berlivet, S., Hunninghake, G.M., Madore, A.-M., Larivière, M., Moussette, S., Grundberg, E., Kwan, T., Ouimet, M., Ge, B., et al. (2009). Allele-specific chromatin remodeling in the ZPBP2/GSDMB/ORMDL3 locus associated with the risk of asthma and autoimmune disease. Am. J. Hum. Genet. *85*, 377–393.

Veyrieras, J.-B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M., and Pritchard, J.K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. *4*, e1000214.

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D.T., Tanay, A., and Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. *10*, 1297–1309.

Wang, J., and Batmanov, K. (2015). BayesPI-BAR: a new biophysical model for characterization of regulatory sequence variations. Nucleic Acids Res. *43*, e147.

Wang, S., Wu, S., Meng, Q., Li, X., Zhang, J., Chen, R., and Wang, M. (2016). FAS rs2234767 and rs1800682 polymorphisms jointly contributed to risk of colorectal cancer by affecting SP1/STAT1 complex recruitment to chromatin. Sci. Rep. *6*, 19229.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. *5*, 276–287.

Waszak, S.M., Delaneau, O., Gschwind, A.R., Kilpinen, H., Raghav, S.K., Witwicki, R.M., Orioli, A., Wiederkehr, M., Panousis, N.I., Yurovsky, A., et al. (2015). Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. Cell *162*, 1039–1050.

Weedon, M.N., Cebola, I., Patch, A.M., Flanagan, S.E., De Franco, E., Caswell, R., Rodríguez-Seguí, S.A., Shaw-Smith, C., Cho, C.H., Lango Allen, H., et al.; International Pancreatic Agenesis Consortium (2014). Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. Nat. Genet. *46*, 61–64.

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. Nat. Genet. *46*, 1160–1165.

Weirauch, M., and Hughes, T.R. (2011). A Catalogue of Eukaryotic Transcription Factor Types, Their Evolutionary Origin, and Species Distribution. In A Handbook of Transcription Factors, T.R. Hughes, ed. (Springer Netherlands), pp. 25–73.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443.

White, M.A., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc. Natl. Acad. Sci. USA *110*, 11952–11957.

Wienert, B., Funnell, A.P.W., Norton, L.J., Pearson, R.C.M., Wilkinson-White, L.E., Lester, K., Vadolas, J., Porteus, M.H., Matthews, J.M., Quinlan, K.G.R., and Crossley, M. (2015). Editing the genome to introduce a beneficial naturally occurring mutation associated with increased fetal globin. Nat. Commun. *6*, 7085.

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. *8*, 206–216.

Wu, J., Metz, C., Xu, X., Abe, R., Gibson, A.W., Edberg, J.C., Cooke, J., Xie, F., Cooper, G.S., and Kimberly, R.P. (2003). A novel polymorphic CAAT/enhancer-binding protein β element in the FasL gene promoter alters Fas ligand expression: a candidate background gene in African American systemic lupus erythematosus patients. J. Immunol. *170*, 132–138.

Xu, D., Dwyer, J., Li, H., Duan, W., and Liu, J.-P. (2008). Ets2 maintains hTERT gene expression and breast cancer cell proliferation by interacting with c-Myc. J. Biol. Chem. *283*, 23567–23580.

Yalcin, B., Wong, K., Agam, A., Goodson, M., Keane, T.M., Gan, X., Nellåker, C., Goodstadt, L., Nicod, J., Bhomra, A., et al. (2011). Sequence-based characterization of structural variation in the mouse genome. Nature *477*, 326–329.

Yang, W.S., Nevin, D.N., Peng, R., Brunzell, J.D., and Deeb, S.S. (1995). A mutation in the promoter of the lipoprotein lipase (LPL) gene in a patient with familial combined hyperlipidemia and low LPL activity. Proc. Natl. Acad. Sci. USA *92*, 4462–4466.

Zeron-Medina, J., Wang, X., Repapi, E., Campbell, M.R., Su, D., Castro-Giner, F., Davies, B., Peterse, E.F., Sacilotto, N., Walker, G.J., et al. (2013). A polymorphic p53 response element in KIT ligand influences cancer risk and has undergone natural selection. Cell *155*, 410–422.

Zhang, X., Miao, X., Tan, W., Ning, B., Liu, Z., Hong, Y., Song, W., Guo, Y., Zhang, X., Shen, Y., et al. (2005). Identification of functional genetic variants in cyclooxygenase-2 and their association with risk of esophageal cancer. Gastroenterology *129*, 565–576.

Zhao, X., Huang, H., and Speed, T.P. (2005). Finding short DNA motifs using permuted Markov models. J. Comput. Biol. *12*, 894–906.

Zheng, X.-W.W., Kudaravalli, R., Russell, T.T., DiMichele, D.M., Gibb, C., Russell, J.E., Margaritis, P., and Pollak, E.S. (2011). Mutation in the factor VII hepatocyte nuclear factor 4α-binding site contributes to factor VII deficiency. Blood Coagul. Fibrinolysis *22*, 624–627.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934.

Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R.S., Bussemaker, H.J., Gordân, R., and Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. Proc. Natl. Acad. Sci. USA *112*, 4654–4659.