Directing an artificial zinc finger protein to new targets by fusion to a non-DNA-binding domain

Wooi F. Lim, Jon Burdach, Alister P.W. Funnell, Richard C.M. Pearson, Kate G.R. Quinlan and Merlin Crossley[*]

School of Biotechnology and Biomolecular Sciences, University of New South Wales, NSW 2052, Australia

ABSTRACT

Transcription factors are often regarded as having two separable components: a DNA-binding domain (DBD) and a functional domain (FD), with the DBD thought to determine target gene recognition. While this holds true for DNA binding *in vitro*, it appears that *in vivo* FDs can also influence genomic targeting. We fused the FD from the well-characterized transcription factor Krüppel-like Factor 3 (KLF3) to an artificial zinc finger (AZF) protein originally designed to target the Vascular Endothelial Growth Factor-A (*VEGF-A*) gene promoter. We compared genome-wide occupancy of the KLF3FD-AZF fusion to that observed with AZF. AZF bound to the *VEGF-A* promoter as predicted, but was also found to occupy approximately 25 000 other sites, a large number of which contained the expected AZF recognition sequence, GCTGGGGGC. Interestingly, addition of the KLF3 FD re-distributes the fusion protein to new sites, with total DNA occupancy detected at around 50 000 sites. A portion of these sites correspond to known KLF3-bound regions, while others contained sequences similar but not identical to the expected AZF recognition sequence. These results show that FDs can influence and may be useful in directing AZF DNA-binding proteins to specific targets and provide insights into how natural transcription factors operate.

INTRODUCTION

Transcription factors are sequence specific DNA-binding proteins that play a dominant role in the regulation of gene expression. They are typically thought of as being composed of independent and separable DNA-binding domains (DBDs) and functional domains (FDs). Recognizing the capability of the two distinct domains to function autonomously has allowed the development of important methodologies such as the yeast two-hybrid system that is used to detect protein–protein interactions (1) and has also facilitated the generation of sequence-specific nucleases, such as zinc finger (ZF) and transcription activator-like effector (TALE) nucleases, that are becoming invaluable in genome editing (2–4).

Nevertheless, the fact that two domains can work autonomously, does not imply that *in vivo* the domains of most or all transcription factors have independent and distinct functions. There is increasing evidence that natural transcription factors localize to their many target genes via the combined functions of both their DBDs and FDs (5–8). We have previously examined an archetypal ZF transcription factor, Krüppel-like factor 3 (KLF3) (9), a member of the Sp/KLF family of transcription factors (10,11). KLF3 has a C-terminal ZF DBD that recognizes CACCC boxes and GC-rich sequences in DNA, and an N-terminal FD known to recruit co-repressors, such as C-terminal Binding Protein (CtBP) to silence gene expression (12,13). Using a loss-of-function approach, we recently showed that unexpectedly the DBD was not the sole determinant of DNA-binding specificity. We found that deletion of the entire FD of KLF3 reduced DNA occupancy across the genome, as assessed by Chromatin Immunoprecipitation followed by genomic sequencing (ChIP-Seq). This result highlighted the importance of the FD for proper *in vivo* DNA-binding specificity (14).

In the current study, we have extended this investigation by performing gain-of-function experiments. We fused the KLF3 FD onto an unrelated, but well characterized artificial zinc finger (AZF) protein. We chose one of the first synthetic ZF proteins developed, the AZF that was originally designed to target a model target gene, Vascular Endothelial Growth Factor-A (*VEGF-A*) (15). The *VEGF-A* gene encodes an angiogenic and neuroprotective factor that may be effective in the treatment of heart ischemia, diabetic neuropathy and cancer (16–18). Given that three alternatively spliced isoforms of *VEGF-A* are required for its maximal biological activity, AZFs that target the *VEGF-A* promoter and modulate the expression of all three isoforms of *VEGF-A* represent a promising therapeutic strategy (15,19–24).

Several different AZFs have now been reported and shown to bind DNA robustly *in vitro* and, when fused to

---

[*]To whom correspondence should be addressed. Tel: +61 2 9385 7916; Fax: +61 2 9385 7920; Email: m.crossley@unsw.edu.au

an activation or a repressor domain, to be effective in modulating expression from the *VEGF-A* promoter both *in vitro* and in reporter systems (15,23). Some of these proteins are also effective *in vivo* where the *VEGF-A* targeting AZF has been used to modulate angiogenesis in normal and diseased mouse models (21,22,24) and to induce neuroprotection in a diabetic mouse (19,20).

Perhaps surprisingly, however, to our knowledge no *in vivo* DNA-binding specificity study has yet been reported with these *VEGF-A* targeting AZF proteins. Our first step therefore was to assess the *in vivo* genome-wide DNA-binding specificity of one potentially therapeutically relevant AZF. We show that the AZF does indeed bind to the *VEGF-A* promoter as expected but we also detected an additional 25 322 binding sites. We then went on to test whether the binding pattern of the AZF protein was affected by the addition of a heterologous FD.

When we compared the sites bound by the AZF alone with the set of peaks generated by the KLF3FD-AZF fusion protein we found that the fusion protein bound many additional sites. Peaks represent protein bound genomic DNA regions identified using a peak calling program Homer. A proportion of these new binding sites corresponded to our previously-identified endogenous KLF3 targets (14). Most importantly we noted that many of the sites bound by the KLF3FD-AZF fusion, but not by AZF alone, corresponded to genomic regions that in loss-of-function experiments were bound by full length KLF3 but not the KLF3 ZF domain missing its N-terminus (14). That is, a proportion of the new sites observed in gain-of-function experiments with the KLF3 FD corresponded to peaks that were lost in the KLF3 FD loss-of-function experiments. This is further supported by a KLF3 FD-only ChIP-Seq experiment that demonstrated, KLF3 FD, in the absence of a DBD, retains a low level of *in vivo* chromatin binding activity, with approximately a quarter of these DBD-deficient KLF3 FD bound regions corresponding to KLF3 FD-AZF bound sites. This suggests that the KLF3 FD is both necessary and sufficient for targeting its cargo to a subset of sites, and that the KLF3 FD is thus important in target gene recognition.

This finding contrasts with the conventional view that DBDs are the primary determinants that direct transcription factor target gene selection. Taken together, our results suggest that, *in vivo*, natural transcription factors use a number of distinct mechanisms, including both DNA–protein interactions mediated by DBDs as well as additional interactions mediated by their non-DNA-binding FDs to localize to the full set of their target genes. Moreover, we suggest that the addition of novel FDs may alter the DNA-binding repertoire of AZF proteins.

MATERIALS AND METHODS

Constructs and generation of stable cell lines

The AZF used in this study was designed and functionally validated in HEK293 cells by Liu *et al.* 2001 (15), previously referred to as construct VZ+42/+530. This AZF was designed to target two GCTGGGGGC sites within the DNase I hypersensitive regions on the human *VEGF-A* locus, 42 bases and 530 bases, respectively, downstream

of the human *VEGF-A* transcriptional start site (+1 TSS). KLF3FD-AZF was made by fusing KLF3 FD amino acid 1–262 to the N-terminus of the AZF. The third construct, termed KLF3 FD, lacking a DBD, consists of KLF3 FD amino acids 1–262 alone. All the constructs contain an NLS from SV40 large T antigen, Pro-Lys-Lys-Lys-Arg-Lys-Val, N-terminal to the AZF or C-terminal to the KLF3 FD, and a C-terminal glycine-serine linker followed by a V5 tag for immunoprecipitation with an anti-V5 antibody. Human embryonic kidney (HEK293) cells were cultured in Dulbecco's modified Eagle's medium (DMEM) (Cat# 11320-082, Life Technologies) supplemented with 10% fetal bovine serum (Cat# 16000044, Life Technologies) and 1% Penicillin-Streptomycin-Glutamate (Life Technologies #10378016). DNA sequences encoding AZF, KLF3FD-AZF or KLF3 FD were cloned into a mammalian expression vector with an EF1α promoter (pEF.IRES.puro) for transient studies and into a retroviral expression system pMSCVpuro (Clontech Laboratories, Mountain View, CA, USA) to generate stable HEK293 cell lines expressing AZF, KLF3FD-AZF or KLF3 FD. Phoenix Ampho (Phoenix A) cells were used as the packaging cell line. Puromycin antibiotic selection (2.5 µg/ml) was initiated 48 h post transduction and was maintained for at least 2 weeks to allow stable expression of the transgene. Single stable clones expressing each transgene were isolated and transcript and protein expression were assessed via quantitative Real Time PCR (RT-PCR) and western Blot using an anti-V5 antibody (Cat# R960CUS, Life Technologies), respectively, as described previously (25). The *in vitro* DNA binding properties of the proteins were assessed via Electrophoretic Mobility Shift Assay (EMSA) using ^{32}P radiolabeled probes, as previously described (9). Oligonucleotides used for RT-PCR and EMSA are available in Supplementary Table S1.

Chromatin immunoprecipitation (ChIP)

ChIP was conducted in six stable HEK293 clones, two each expressing equivalent levels of AZF, KLF3FD-AZF or KLF3 FD, representing two biological replicates. Approximately 7×10^7 cells were used for each ChIP and the experiments were conducted as described (26) using 14 µg of anti-V5 antibody (Cat# R960CUS, Life Technologies). DNA samples obtained were used for RT PCR and high throughput DNA sequencing. Oligonucleotides used for RT PCR are available in Supplementary Table S1.

DNA library preparation and next-generation sequencing

Libraries were prepared using the TruSeq ChIP Sample Prep Kit (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. The libraries were multiplexed into two lanes using sample specific adapters such that there were four samples per lane. A total of 50 bp single reads or 100 bp paired end reads were sequenced on the HiSeq 2500 or HiSeq 2000 (Illumina, San Diego, CA, USA). For KLF3 FD samples, 75 bp single reads were sequenced on the NextSeq 500. Library preparation and sequencing were performed by the Ramaciotti Centre for Genomics, University of New South Wales, NSW, Australia. Quality control was performed using FastQC v0.10.1

available from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Reads were quality filtered, trimmed and adapter sequences were removed using Trimmomatic v0.3.2 (27).

Alignment

Reads were aligned to the hg19/GRCh37 *Homo sapiens* genome using Bowtie v2.2.1 (28) set to –very-sensitive. Resulting alignments were sorted and indexed using Samtools v0.1.18 (29).

Peaking calling and IDR analysis

Pseudoreplicates were created using homer-idr v0.1 (available from https://zenodo.org/record/11619#) for individual and combined IP samples. Peaks were called using Homer v4.7.2 (30) using the permissive settings (-P .1 -LP .1 -poisson .1) on individual replicates, combined replicates, individual pseudoreplicates and combined pseudoreplicates against the combined input control. Peaks lists were then supplied to homer-idr to determine the irreproducible discovery rate (IDR) statistic for each peak generating a final peak list satisfying the thresholds set by homer-idr. Peaks were merged using mergePeaks using the switch -d meaning that peaks had to literally overlap in genomic space to be considered overlapping. Venn diagram summarizing the total number of the AZF only peaks, KLF3FD-AZF only peaks and AZF and KLF3FD-AZF common peaks was generated.

Quantification of ChIP tags

HOMER was used to quantify ChIP tag density at peak locations across the genome. Unless otherwise noted, tags were counted within 162 bp (for KLF3 FD) and 214 bp (for AZF and KLF3FD–AZF) around the peak center (as peak widths could vary across the different samples). All tag counts were normalized to 100 M reads, and were thus expressed as reads/100 M reads to allow comparison across samples.

Differential binding analysis

Peaks were called on each replicate against its corresponding input control using Homer v4.7.2 (30) with the default settings except with (-style factor -F 10). Peaks from each replicate along with alignments were piped to DiffBind v1.10.2 (31) and the contrast was set based on the factor that was immunoprecipitated (AZF/KLF3FD-AZF). DiffBind was used to calculate differential binding statistics for the KLF3FD-AZF and AZF groups using edgeR-based analysis. PCA and MA plots were also generated using DiffBind on the factor contrast. Peaks were considered differentially bound if they showed FDR < 0.1, $P < 0.05$ and a log2 fold change of either > 2 or < -2 between the groups.

Genomic annotation and visualization

Peak lists from IDR and differential binding analysis were annotated using annotatePeaks.pl using the HOMER annotation set for hg19/GRCh37. HOMER was used to create bedgraph files using the makeUCSCfile program. These were viewed using Integrative Genomics Viewer IGV v2.2 (32).

De novo motif analysis

De novo motif discovery was performed on the top 600 peaks, ranked by normalized tag counts, from the AZF total peaks, KLF3FD-AZF total peaks and AZF and KLF3FD-AZF common peaks obtained from the HOMER-IDR analysis, and on the top 600 peaks, ranked by log2 fold-change contrasting AZF/KLF3FD-AZF normalized tag counts, from the KLF3FD-AZF differential bound peaks obtained from the differential binding analysis. Sequence databases in fasta format consisting of the 100 bp surrounding peak centers were created using an open source, web-based platform, Galaxy (available from https://usegalaxy.org/) and piped to MEME-ChIP (33) using the default settings to identify the most significantly bound motif from each sample group.

For motif analysis using the complete ChIP-Seq datasets, alternative web-based motif finding tools, RSAT peak motifs (available from http://floresta.eead.csic.es/rsat/peak-motifs_form.cgi) (34) and DREME (available from http://meme-suite.org/tools/dreme) (35), were used.

Common KLF3FD-AZF and KLF3 promoter binding

A list was generated consisting of all promoter peaks (2025 peaks) bound by KLF3FD-AZF but not by the variant lacking the KLF3 FD (from the differential binding analysis). This list (referred to as KLF3FD-AZF HEK293 list) was compared to a list consisting of all the promoter peaks (4212 peaks) bound by unmodified KLF3 protein but not by the variant lacking the KLF3 FD, referred to as KLF3 mouse embryonic fibroblasts (MEFs) list. These two lists represent promoters that are differentially bound by KLF3FD-AZF and KLF3, respectively, that require the presence of the KLF3 FD for DNA occupancy. The latter was obtained from published ChIP-Seq experiments performed in MEFs (14). Promoter binding was defined as DNA occupancy observed between −1000 bp to +100 bp relative to the +1 TSS of a coding gene. The comparison was made based on common downstream gene names. To address the issue of inter-species gene nomenclature difference, mouse gene names on the KLF3 MEFs list was converted to the corresponding human orthologs using HCOP: Orthology Prediction Search (available from http://www.genenames.org/cgi-bin/hcop). The KLF3FD-AZF HEK293 list was refined to contain only the human/mouse orthologous genes. These two lists (containing 3525 and 1720 genes, respectively) were overlapped based on common downstream gene names to generate a list of common promoters bound by KLF3FD-AZF and KLF3. The expected number of common promoters found by chance was calculated based on the total number of promoter peaks from each dataset, that were used for comparison, as percentages of the total number of protein coding genes (36) (as an estimation to the total number of promoters). The product of these percentages gives an estimation of the expected common promoter binding that would occur by chance.

Statistical test

Chi-squared (X^2) test was performed to determine whether the difference between the observed and the expected number of common promoter binding is significant. A chi-squared value was computed yielding a *P*-value using GraphPad software available from http://graphpad.com/quickcalcs/.

RESULTS

Both AZF and KLF3FD-AZF bind the *VEGF-A* recognition sequence *in vitro* and *in vivo*

One of the first AZF constructs ever generated was a protein consisting of three classical C_2H_2 ZFs designed to target two GCTGGGGGC sites within DNase I hypersensitive regions in the human *VEGF-A* locus. One of these sites lies within the *VEGF-A* promoter, 42 bases downstream of the transcriptional start site (15) (Figure 1A).

We generated human embryonic kidney cell lines (HEK293) stably expressing this AZF protein using the MSCV retroviral transduction system. In addition, a second construct designated KLF3FD-AZF, was made by fusing the KLF3 FD (amino acids 1–262 from full length KLF3 protein) upstream of the AZF domain. Both constructs also contained a nuclear localization signal (NLS) and were tagged with a V5 epitope via a glycine-serine linker to enable consistent immunoprecipitation and comparison between samples (Figure 1B).

Two independent clonal cell lines stably expressing the AZF and two lines expressing the KLF3FD-AZF were selected for further analysis. Western blot and quantitative real-time PCR experiments showed robust expression in all cases. While there was some variation in mRNA levels, the protein levels were comparable in Western blotting using an antibody to the shared V5 epitope tag (Figure 1C and D).

We investigated *in vitro* binding with EMSA using a previously validated target site containing the canonical motif GCTGGGGGC (15). We observed strong and comparable binding of both AZF and KLF3FD-AZF, respectively, in all four lines (Figure 1E). Supershift experiments with an anti-V5 antibody confirmed the identity of retarded species. These validation experiments suggested that these cell lines were appropriately expressing the proteins of interest at equivalent levels.

We next performed ChIP-seq on the four HEK293 lines. Across the four samples, a total of more than 150 M reads were mapped to the human genome using Bowtie2. IDR analysis was used to identify consensus peaks across the biological replicates. IDR analysis was performed using the HOMER-IDR package that provides thresholds based on reproducibility. These peaks were then further analyzed. An annotated table containing ChIP peaks across all samples can be found in Supplementary Tables S2 and S3.

We first interrogated the ChIP-Seq data to assess binding at the *VEGF-A* locus. Satisfyingly, peaks were observed at the expected and previously validated AZF domain binding site in the human *VEGF-A* locus in both the AZF and KLF3FD-AZF expressing cell lines (Supplementary Figure S1A). This result was also confirmed by ChIP PCR analysis. ChIP assays were performed on each of the four stable clonal cell lines and the recovered DNA was subjected to amplification by quantitative real-time PCR using primers specific for human *VEGF-A* promoter region in comparison to a negative control region (Primer sequences are available in Supplementary Table S1). The ChIP-PCR confirmed the presence of the expected fragments with marked enrichment (Supplementary Figure S1B).

The AZF binds to a large number of sites within the genome

We next analyzed genome-wide binding profiles as reflected in peak profiles. We found that the *VEGF-A* promoter peak was among the top 30% of peaks bound by the AZF as ranked by normalized tag count. However, the AZF also binds to many other sites in the genome *in vivo*—in total, we identified 25 322 peaks as significant and consistent across the two independent biological replicate sets of AZF samples (Supplementary Table S2).

To elucidate the widespread binding of AZF, we further interrogated the top 600 AZF peaks using tools including the *de novo* motif discovery algorithm MEME and central motif enrichment analysis CENTRIMO (33) to generate a consensus motif bound by AZF *in vivo*. Interestingly, 95% of these peak regions contain a centrally enriched consensus motif that conformed to the GCTGGGGGC 9 nt target sequence bound by the AZF *in vitro* (Figure 2). The tolerance of mismatches varies across this 9 nt consensus sequence. At position 4, a G is present in all the 571 sequences containing this consensus motif; therefore it has the maximum height and information content of 2 bits. This suggests that essentially a G is required at position 4. On the other hand, any of the 4 possible nucleotides could be present at position 9, although with different frequencies. This represents an increase in uncertainty, thus lowering the information content available at this position. While most of the bound sequences show high conservation to the 9 nt target sequence the AZF was designed to recognize, nucleotide identity at position 3 and 9 of the GCTGGGGGC target sequence, the T and C, respectively, seems to be more forgiving to mismatches. This was further confirmed using RSAT peak motifs and DREME, alternative motif finding tools that allow extraction of binding motifs from the complete AZF ChIP-Seq peaks (25 322 peaks), where similar outputs were obtained (Supplementary Figure S2A).

In addition, we also observed AZF binding to DNA in the absence of the consensus motif in 5% of the top 600 AZF peaks investigated (two examples are shown in Supplementary Figure S3). At this stage it is not clear whether these additional sites represent cases where the AZF is localizing to particular genes via protein–protein interactions, or is binding to highly divergent motifs, or whether these non-canonical sites represent secondary long or short range interactions possibly resulting from enhancer looping to the promoter, thus, representing binding sites that were not directly bound by the AZF.

Taken together, it is important to note that the AZF binds to a large number of sites in the genome given it was initially designed to regulate the expression of one gene, human *VEGF-A*. Statistically, a 9 base pair sequence should occur approximately 23 000 times in a three billion base pair human genome, however, the complexity of the chromatin
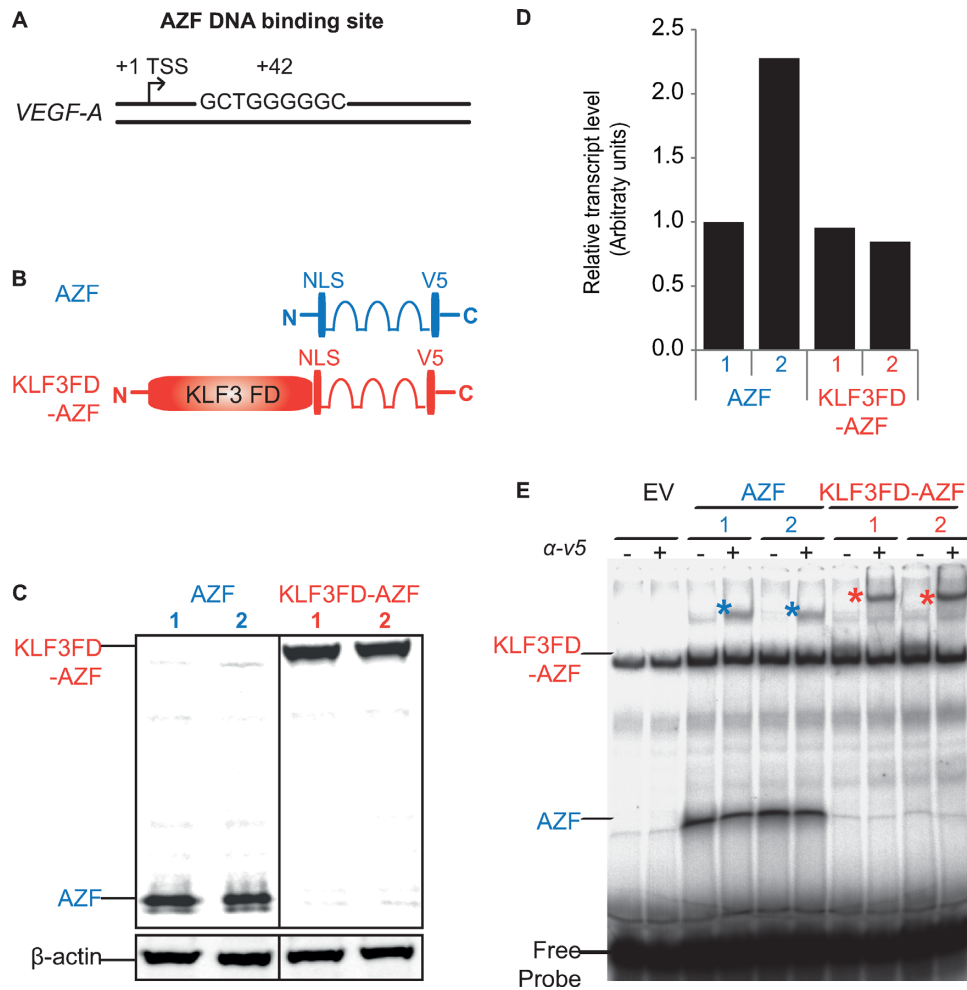
Figure 1. Construction and verification of AZF and KLF3FD-AZF protein. (**A**) Human *VEGF-A* locus showing the location of the 9 nt target sequence GCTGGGGGC 42 bases downstream of the +1 TSS that the AZF was designed to recognize and bind (15). (**B**) Schematic representation of AZF and KLF3FD-AZF with an extra KLF3 functional domain (KLF3 FD) fused to the N-terminus of the artificial zinc fingers (AZFs). Western blot (**C**) and quantitative real time PCR (**D**) showing equivalent protein and mRNA transcript expression, respectively, for the four selected HEK293 clones stably expressing AZF or KLF3FD-AZF. For western blot, β-actin was included as the loading control and for real-time PCR, transcript expression was normalized to *18S* rRNA level and is shown relative to the expression for AZF clone 1 (first bar), which was set to an arbitrary value of 1. (**E**) Electrophoretic mobility shift assay showing equivalent *in vitro* AZF and KLF3FD-AZF binding to a ^{32}P radiolabeled EMSA probe containing the 9 nt target sequence GCTGGGGGC. Asterisks indicate the supershift of the protein-DNA probe complex by an anti-V5 antibody confirming the identity of the V5-tagged AZF proteins.

organization may make a substantial portion of those perfect sites less accessible to the AZF. In this study we found AZF binds to about 25 000 sites. The degeneracies of the AZF DNA binding preference observed in this study mean that the 3-ZF protein effectively recognizes a 7 or 8 nt sequence, rather than the extended 9 nt sequence, may partly explain the binding pattern observed for this AZF.

KLF3FD-AZF shows increased DNA occupancy across the genome

We next sought to investigate if the fusion of an additional KLF3 FD from a natural ZF transcription factor to an AZF protein has an effect on the *in vivo* DNA binding specificity by comparing the genomic DNA binding profile of KLF3FD-AZF to that of the AZF. ChIP-Seq analysis on HEK293 cells expressing KLF3FD-AZF revealed 48 003 peaks across the human genome (Supple-

mentary Table S3). Compared with the 25 322 peaks observed with AZF, KLF3FD-AZF thus generates approximately twice as many peaks. We further examined and compared the regions where peaks were observed with AZF and/or KLF3FD-AZF. As expected, with both AZF and KLF3FD-AZF sharing the same ZF DBD, most of the peaks (more than 80%) observed with AZF were also observed with KLF3FD-AZF. But strikingly KLF3FD–AZF also generates 27 610 additional peaks (Figure 3A).

Differential binding analysis was performed using DIFFBIND to further define peaks that were more than 2-fold different in normalized read counts in the KLF3FD-AZF and the AZF samples. A list consisting of 4357 peaks that were preferentially or uniquely bound by KLF3FD-AZF was produced using stringent cut-offs of FDR < 0.1, P-value < 0.05 and requiring at least a 2-fold increase in the normalized read counts (based on evidence of binding affin-

%	1	2	3	4	5	6	7	8	9
A	-	2	23	-	9	-	-	-	31
C	3	98	1	-	-	-	-	-	59
G	97	-	19	100	91	100	100	100	1
T	-	-	58	-	-	-	-	-	9
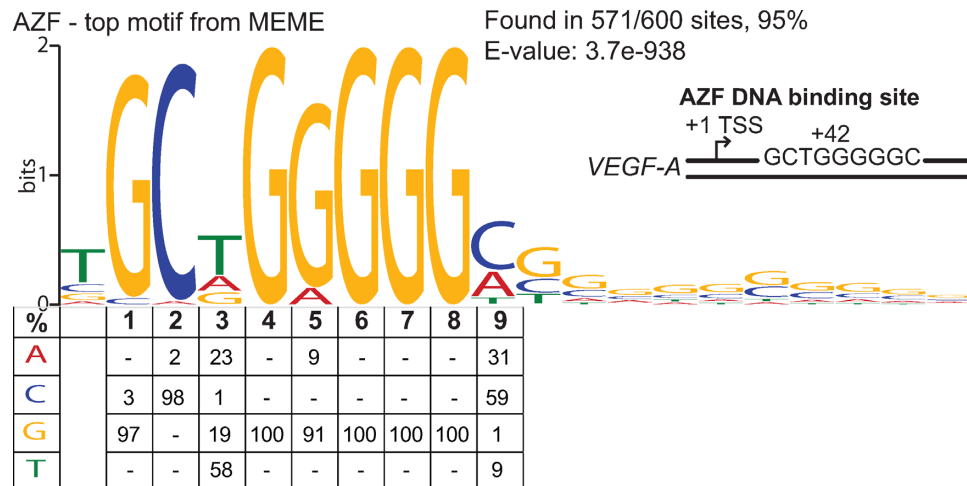
Figure 2. AZF binds the 9 nt target sequence GCTGGGGGC *in vivo*. MEME a *de novo* motif discovery tool identified a consensus motif bound by the AZF that conforms to the predicted target sequence GCTGGGGGC that the AZF was designed to bind. The height of the letters in the position weight matrix (PWM) represents information content (in bits) at each position that is related to the degree of certainty of the particular nucleotide at a given position and the table below shows the relative frequency in percentage of each A, C, G, T nucleotide observed at a given position.

ity) for KLF3FD-AZF compared to that of AZF (Figure 3B and Supplementary Table S4). ChIP-Seq traces showing examples of common peaks and KLF3FD-AZF unique peaks or preferentially bound peaks are shown (Figure 3C and D, respectively). As part of the output from DIFF-BIND, we also found a small number of peaks, that is 4620 sites, were preferentially bound by the AZF (Supplementary Figure S4).

KLF3FD-AZF peaks are abundant in promoter regions

Previously we reported that ChIP-Seq analysis of KLF3 occupancy reveals a disproportionate number of peaks in promoter regions consistent with its role as a transcription regulator (14). Interestingly, a deletion mutant lacking the KLF3 FD is not similarly enriched at promoters but generates peaks more broadly across the genome. Thus, our previous loss-of-function experiment suggested that the KLF3 FD may play a role in localizing KLF3 to target gene promoters.

To investigate if the KLF3 FD exhibits similar behavior in gain-of-function experiments we compared the peaks generated by AZF to KLF3FD-AZF and determined which were in promoters, introns, intergenic and intragenic regions. To achieve this, the peaks were analyzed based on RefSeq annotations, and were categorized according to their genomic localization across the *Homo sapiens* (hg19/GRCh37) genome into the four main categories. In this analysis promoters were defined as the regions −1 to +0.1 kb from the RefSeq TSS and a fifth category 'other' was used to encompass peaks that fell into coding exons, 5′UTR and 3′UTR exons or were close to the transcription termination sites (−100 bp to +1 kb).
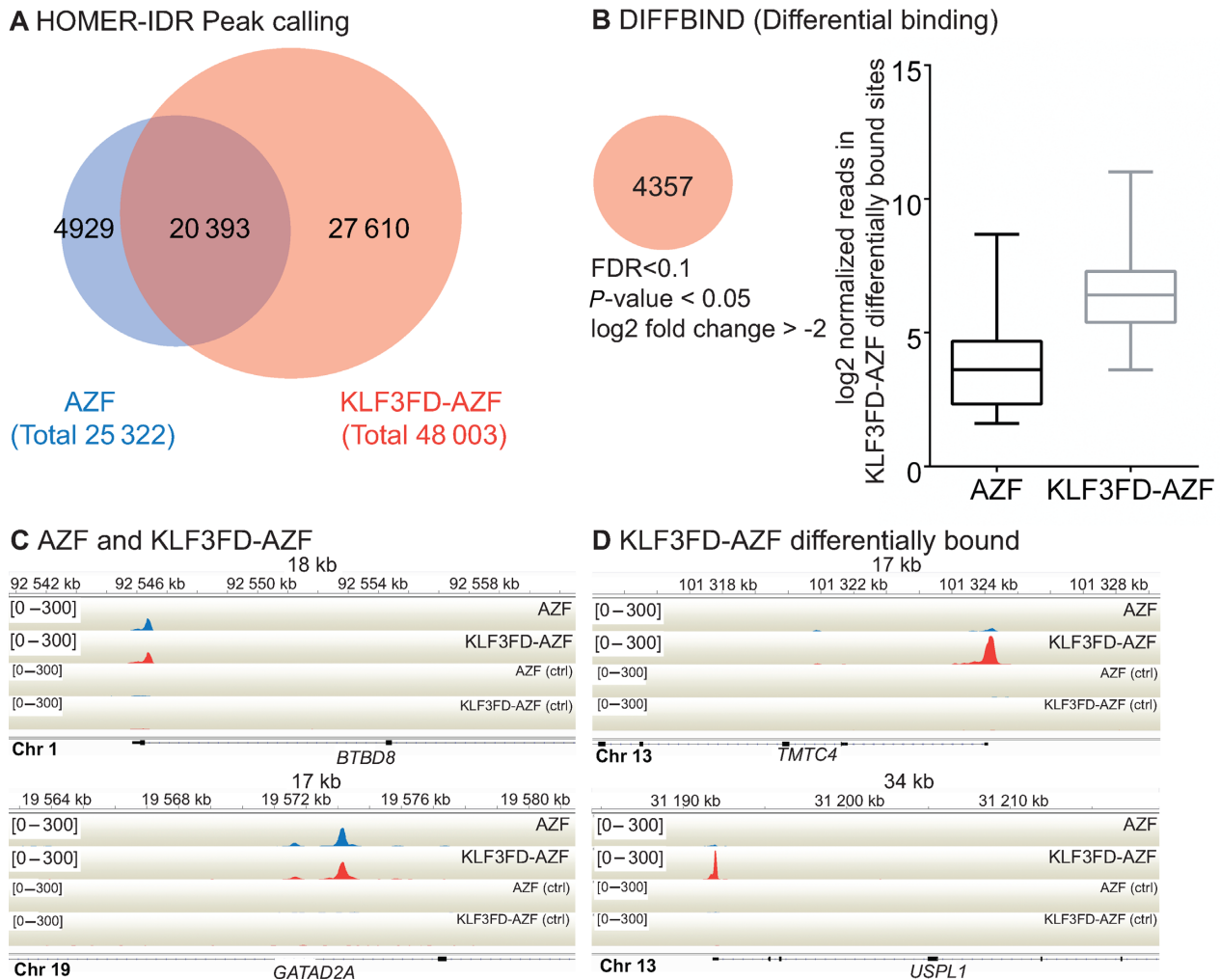
We found that 18 and 20% of the AZF and KLF3FD-AZF peaks, respectively, lie within promoters. Intriguingly, when we examined peaks that were differentially bound, that is, the category of peaks that were 2-fold or more higher with KLF3FD-AZF, we observed that 46% of these resided in promoter regions (Figure 4). This observation suggests

that the addition of the KLF3 FD does indeed help to target the fusion protein to promoter regions that are not significantly occupied by AZF alone.

Peaks preferentially bound by KLF3FD-AZF contain an imperfect AZF DNA binding site

To further characterize the new binding sites acquired in the presence of the KLF3 FD, we generated and compared consensus motifs from the top 600 peaks bound by AZF and/or KLF3FD-AZF using the MEME algorithm (Figure 5 and Supplementary Figure S2). As expected, about 95% of the regions bound by AZF and KLF3FD-AZF contain the 9 nt AZF target sequence (Supplementary Figure S2C). This was confirmed by RSAT peak motifs and DREME with the whole ChIP-Seq datasets consisting of 48 003 peaks bound by KLF3FD-AZF and 20 393 peaks commonly bound by AZF and KLF3FD-AZF (Supplementary Figures S2B and S2C). In contrast, examination of the peaks generated either uniquely or preferentially by KLF3FD-AZF revealed a related motif, consisting of a string of G and C bases (Figure 5A and Supplementary Figure S2D). This consensus resembles an imperfect version of the AZF target sequence that is more forgiving to mismatches at nucleotide position 3, 5, 8 and 9. Particularly at position 3, a G base has become more prominent than the original T base. Additionally, tolerance of a C base and an A or C base at position 5 and 8, respectively, at frequencies of 8 to 28%, is acquired. The G base at position 8 occurs at a frequency of 52% compared to 99% in peaks shared by both AZF and KLF3FD-AZF.

To evaluate this further, we performed EMSA experiments to assess AZF and KLF3FD-AZF binding affinity to probes containing a G to C point mutation at nucleotide position 5. As a positive control we used a previously validated probe containing a perfect 9 nt AZF target sequence, referred to as the 'wild type' AZF probe. As expected, the G to C point mutation compromises AZF binding (Figure 5B). It similarly reduced the binding of KLF3FD-AZF *in vitro*.

A HOMER-IDR Peak calling

B DIFFBIND (Differential binding)

AZF
(Total 25 322)

KLF3FD-AZF
(Total 48 003)

FDR<0.1
P-value < 0.05
log2 fold change > -2

C AZF and KLF3FD-AZF

D KLF3FD-AZF differentially bound

Figure 3. Regions that are differentially bound by KLF3FD-AZF. (**A**) A proportional Venn diagram showing the regions bound by AZF and/or KLF3FD-AZF. (**B**) Differential binding study DIFFBIND identified 4357 sites that are differentially bound by KLF3FD-AZF and box plot showing log2 normalized read counts at these regions in AZF and KLF3FD-AZF samples. (**C** + **D**) An illustrative range of examples showing regions that are bound equivalently by both AZF and KLF3FD-AZF (C) and regions that show greater binding by KLF3FD-AZF than by AZF as identified by DIFFBIND (D). Binding profiles of the respective input samples are included as control and shown in the third and fourth track.

This result highlights the difference between the *in vitro* and *in vivo* DNA-binding behaviors. While *in vitro* both AZF and KLF3FD-AZF show weak binding to the 5G > C sequence, *in vivo* KLF3FD–AZF, but not AZF, generates a peak at many such sequences (Figure 5C). This suggests that in the *in vivo* setting where the complexity of the chromatin comes into play, the additional KLF3 FD may be facilitating binding to more degenerate sites that the AZF alone cannot bind.

KLF3 FD, in the absence of a DBD, is capable of chromatin binding *in vivo*

We next assessed whether KLF3 FD, in the absence of a DBD, is recruited to genomic sites. We performed ChIP-Seq on two HEK293 lines stably expressing KLF3 FD only protein tagged with V5 epitope for immunoprecipitation to investigate genome-wide DNA occupancy of this protein (Figure 6A). We identified 1439 sites bound by KLF3 FD across the genome that were consistent across the two bi-

ological replicates (Supplementary Table S5). In general, we observed lower normalized tag counts for these binding sites compared to those that are bound by AZF or KLF3FD-AZF, an indication of weaker binding by the protein lacking a DBD. This is not surprising based on the previous evidence implying that KLF3 FD is involved in facilitating recruitment of a protein to DNA, with DBD playing the primary role in DNA binding and recognition.

We then compared the sites bound by KLF3 FD to those that are occupied by KLF3FD-AZF and found 264 sites commonly bound by both KLF3 FD and KLF3FD-AZF. This represents 23% of the genomic sites occupied by KLF3 FD (Figure 6B and Supplementary Table S6). The correlation of normalized tag counts at individual KLF3 FD and KLF3FD-AZF common bound peaks shows different chromatin binding efficiencies between the two proteins, with the DBD deficient protein, KLF3 FD, generating lower peaks (Figure 6C). Representative peak image of KLF3 FD and KLF3FD-AZF binding to one of the com-
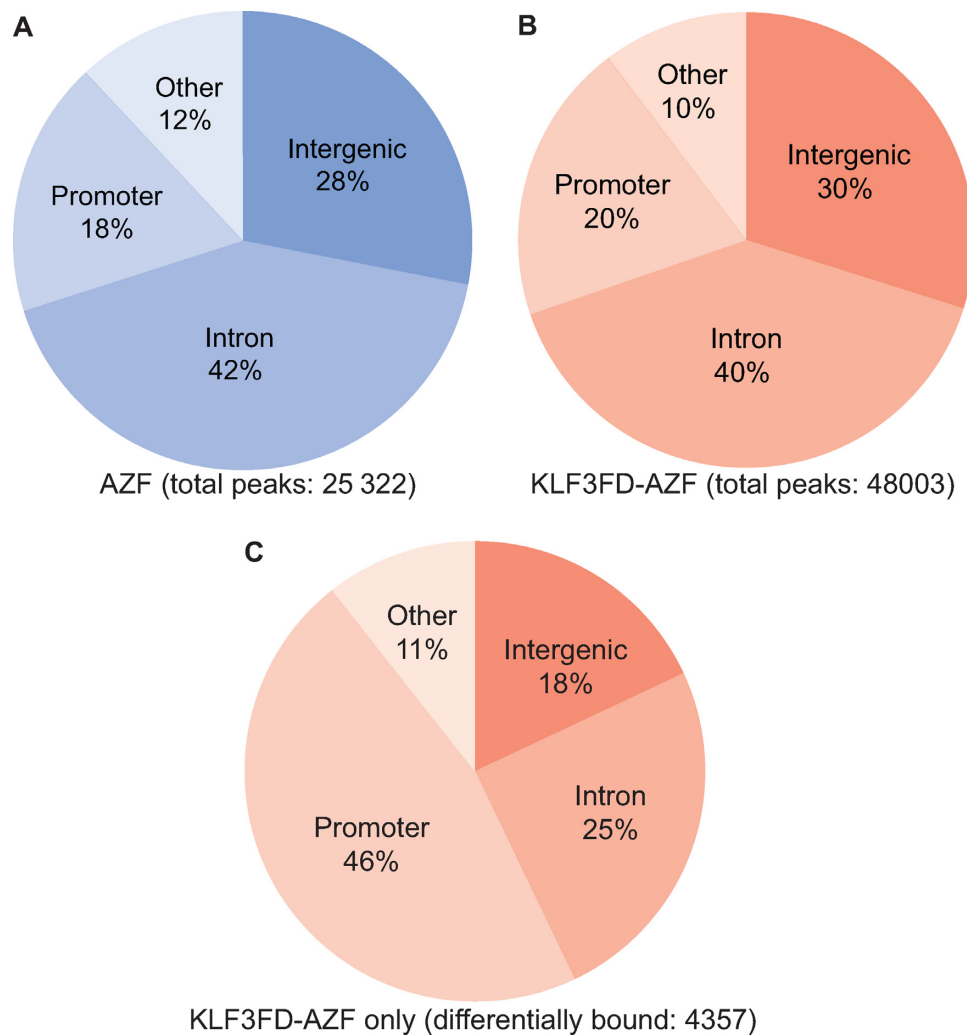
Figure 4. Genomic localization of regions bound by AZF or KLF3FD-AZF showing enrichment of KLF3FD-AZF differential binding at the promoter regions. Genomic localization of (**A**) total regions bound by AZF, (**B**) total regions bound by KLF3FD-AZF and (**C**) regions differentially bound by KLF3FD-AZF as identified by DIFFBIND. Promoters are defined as the region −1000 bp to +100 bp around the +1 TSS of Refseq genes. Peaks that fell into CDS exons, non-coding, 5′ and 3′ UTR exons and transcription termination sites are all labeled as 'other.' Percentages lying in each region are given and the total regions sampled are shown in parenthesis.

mon target sites, *ARID3A*, is shown in Figure 6D, and the binding was validated with an independent ChIP-PCR experiment (Figure 6E).

Taken together, this evidence suggests that KLF3 FD, despite lacking a DBD, is still capable of *in vivo* chromatin binding, albeit with a lower efficiency. We next sought to investigate whether this protein binds to naked DNA *in vitro* via EMSA. We assessed KLF3 FD binding to AZF target site, natural KLF3 target site and a 50 nt DNA sequence derived from *ARID3B* intron region, shown to be bound by KLF3 FD *in vivo*. As shown in Supplementary Figure S5, there is no detectable KLF3 FD binding to any these naked DNA probes. Consistent with our results for AZF and KLF3FD-AZF chromatin binding, this indicates that the FD plays a role in the *in vivo* DNA binding but not *in vitro*.

KLF3FD-AZF generates peaks at known KLF3 target sites

To further analyze the nature of peaks that were dependent on the KLF3 FD we compared the genomic DNA occupancy of KLF3FD-AZF to that of full length KLF3 protein (containing the same FD but this time in its native configuration upstream of the KLF3 ZF DBD rather than the AZF domain). KLF3 is known to bind CACCC boxes and GC-rich elements in the regulatory regions of its target genes. We examined our previously published KLF3 ChIP-Seq dataset which was produced by experiments in immortalized MEFs (14). We refined our search to focus on peaks that were generated by full length KLF3 but not by a KLF3 deletion mutant that was missing the FD. Thus this list represents regions that require the presence of the KLF3 FD for DNA binding, as determined by loss-of-function experiments.

We compared this dataset to the list of peaks bound strongly by KLF3FD-AZF but not by AZF, that is, to the
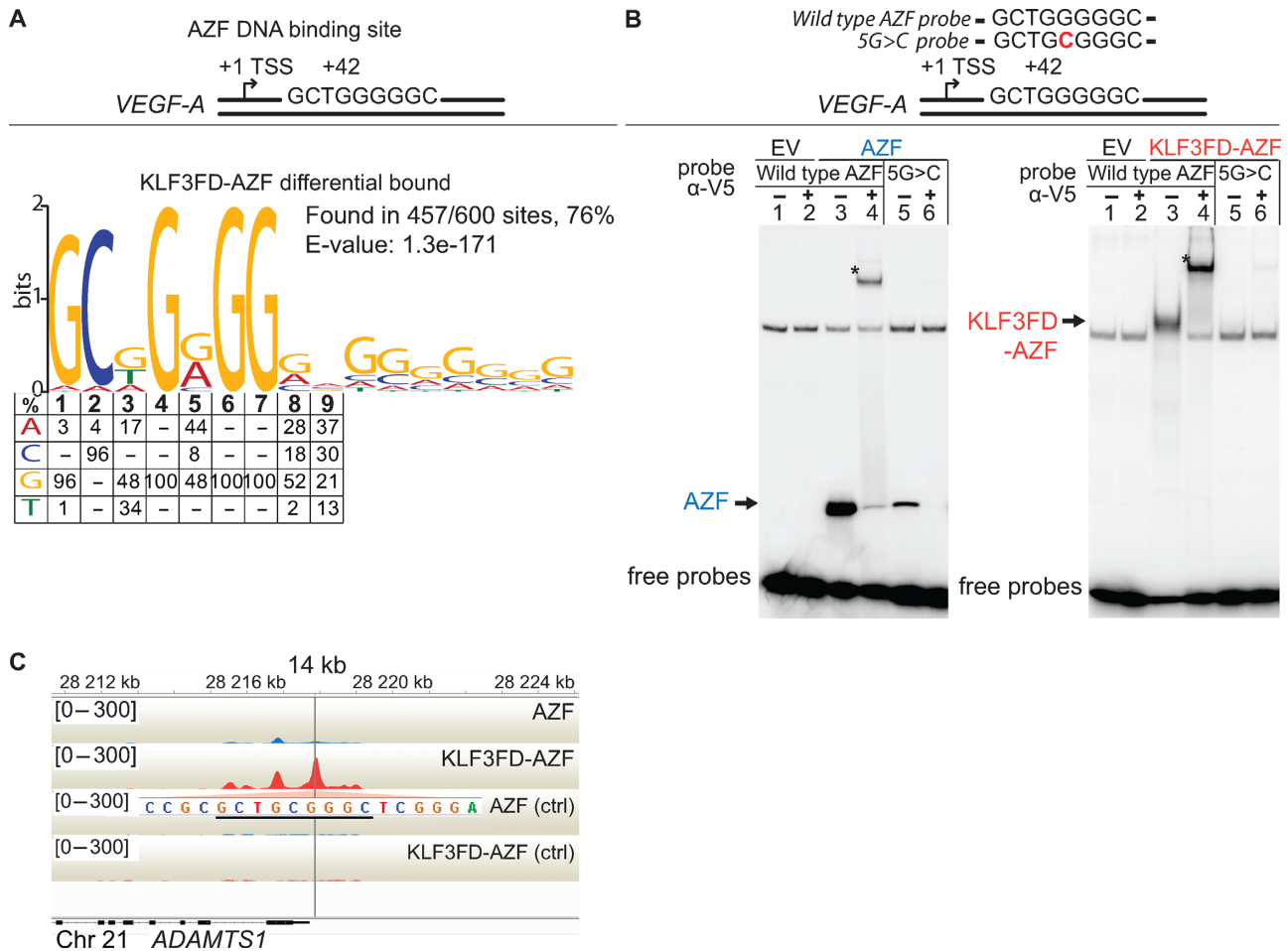
Figure 5. KLF3FD-AZF binds to a degenerate AZF target site *in vivo* but not *in vitro*. (**A**) Top panel shows the AZF target sequence GCTGGGGGC. Bottom panel shows a consensus motif found in 76% of regions differentially bound by KLF3FD-AZF identified using MEME, a *de novo* motif discovery tool and a table contains the relative frequency in percentage of each A, C, G, T nucleotide observed at a given position. (**B**) EMSA experiments showing AZF (left) and KLF3FD-AZF (right) binding to the AZF target sequence, labeled as 'wild type AZF', (lane 3) and upon addition of anti-V5 antibody, supershift was observed identifying the V5-tagged AZF proteins, marked with asterisk. Radiolabeled probes containing a mutation at position 5 from a G base to a C base (labeled as '5G > C') was included to investigate AZF and KLF3FD-AZF binding to a degenerate AZF target sequence that was able to be bound *in vivo* by KLF3FD-AZF (lane 5 and lane 6 with addition of anti-V5 antibody). EV denotes empty vector that serves as a negative control. (**C**) Peak image from the ChIP-Seq experiments show differential binding by AZF and by KLF3FD-AZF *in vivo* at region containing the corresponding degenerate AZF target site (underline).

list of unique peaks that were generated in the gain-of-function experiments. The ChIP-Seq experiments used for comparison were performed in two different cell lines from two different species and thus cannot be directly overlaid. Nevertheless by comparing only the promoter occupancies (by overlapping downstream gene names from the two lists), which are more highly conserved between mouse and human rather than focussing on less well conserved non-coding regions in the genome, clear similarities emerged. The issue of interspecies gene nomenclature difference was addressed by converting all the downstream gene names to the corresponding human orthologs using HCOP: Orthology Prediction Search. This resulted in two refined lists containing KLF3 FD dependent promoter occupancies, 3525 and 1720, respectively, from the two ChIP-Seq datasets (Supplementary Figure S6A). Common promoter binding events were calculated based on the occurrence of common downstream gene names in the two lists. Based on a proba-

bilistic calculation, we expect to find approximately 0.8% or approximately 223 common promoter peaks from the two datasets, taking into the consideration the number of promoter peaks found in each of the datasets as a percentage of the total number of protein coding genes (as an estimation to the total number of promoters) (36). Interestingly, we found at least 578 peaks (more than twice the expected occurrence of common promoter binding expected by chance) that were generated by KLF3FD-AZF and not by AZF alone, are similarly bound by full length KLF3 but not by its counterpart lacking the KLF3 FD, with a *P*-value of <0.0001 from a Chi-squared test assessing the significance of the difference between the observed and the expected occasions of the common promoter binding (Supplementary Figures S6A and S6B and Supplementary Table S7). This result is surprising because the AZF is hypothesized to bind GCTGGGGGC sequences which are not expected to be enriched in this proportion of KLF3 target genes. A similar
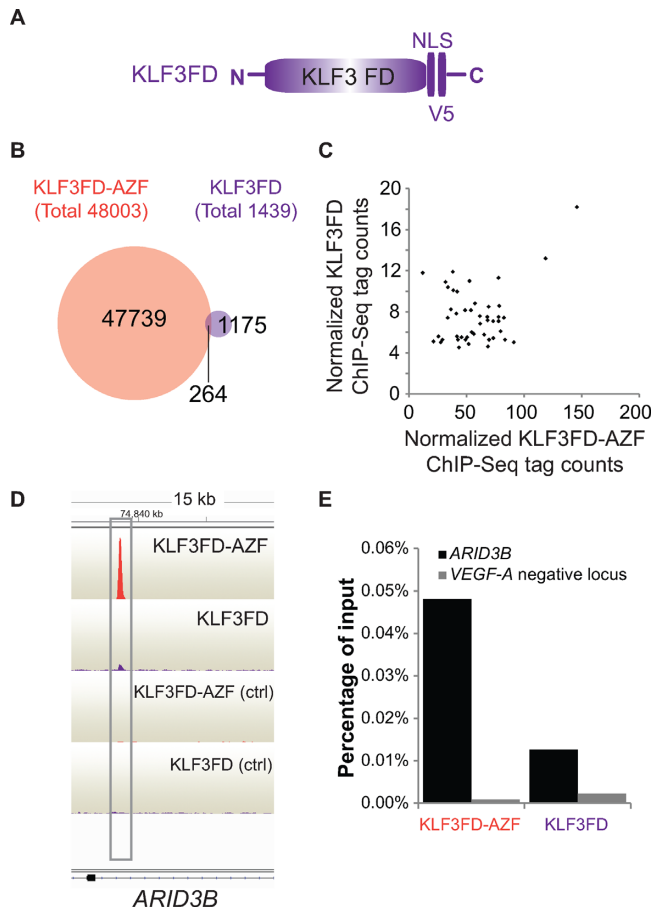
Figure 6. ChIP-Seq and ChIP-PCR demonstrate KLF3 FD *in vivo* chromatin binding. (**A**) Schematic representation of KLF3 FD protein with V5-epitope tag at C-terminus for immunoprecipitation. (**B**) A proportional Venn diagram showing 264 genomic regions bound by both KLF3FD-AZF and KLF3 FD. (**C**) Chromatin binding efficiencies of KLF3 FD is lower than KLF3FD-AZF as illustrated by normalized tag counts of the two proteins at the commonly bound regions. A representative ChIP-Seq track (**D**) and ChIP-PCR (**E**) showing KLF3 FD genomic occupancy. ChIP-Seq tracks for KLF3FD-AZF and KLF3 FD are on scale (0–500) and (0–100), respectively.

analysis procedure was also applied to compare KLF3 FD dependent promoter peaks from the previous KLF3 loss-of-function experiment to KLF3 FD promoter peaks, consisting of 107 promoter regions bound by the KLF3 FD protein lacking a DBD. We found 45 common promoter binding events overlapping the two lists based on common gene name. This represents 42% of the total promoter peaks bound by the DBD-deficient KLF3 FD protein (Supplementary Figure S6C and Supplementary Table S8). Despite the inevitable caveats associated with this comparative analysis method as the results of human and mouse genome sequence variations, these observations remain striking.

One potential explanation for this finding is that the KLF3FD-AZF fusion is simply dimerizing with endogenous KLF3 in HEK293 cells, and thereby localizing to KLF3's normal target genes. To explore this possibility we tested whether endogenous KLF3 protein was expressed in HEK293 cells. No signal corresponding to full length KLF3 protein could be detected in western blotting (Supplemen-

tary Figure S7). We therefore conclude that the KLF3 FD is not likely to be dimerizing with endogenous KLF3 to facilitate genomic localization but, more interestingly, may be contacting another molecule, perhaps another transcription factor or histone mark that identifies KLF3 target promoters.

In summary, our results suggest that the KLF3 FD is both required to localize KLF3 to certain target genes in loss-of-function experiments (14), and is also capable of redirecting an artificial AZF to known target genes in gain-of-function experiments.

DISCUSSION

AZF-based DNA-binding platforms are currently being developed for therapeutic purposes. Throughout the short history of the development of the artificial DNA-binding platform, the field has moved from the early use of three ZF proteins based on the natural ZF protein ZNF268 (15,22) to adopting six ZF proteins that were expected to be more specific in targeting sites within the whole human genome (37,38).

However, to date, there have been no direct comparisons between genome-wide DNA binding specificity of three ZF and six ZF AZF proteins. One recent study by Grimmer and colleagues assessed genome-wide DNA occupancy of two six ZF proteins targeting two different 18 nt sequences on human *SOX2* promoter (39). While statistically one would expect an 18 nt sequence to appear once or at most a few times in the whole human genome, the study found these six finger proteins occupy thousands of sites across the genome. One explanation for this is the idea that instead of requiring a perfect 18 nucleotide site, they can use a subset of their available ZF domains for genomic interactions. In short, they possibly have more available sites than the corresponding three ZF proteins targeting similar sites.

In the current study assessing genome-wide DNA occupancy of a three ZF protein targeting 9 nt sequence on human *VEGF-A* promoter, we found that, in contrast to the six ZF proteins that bind many more sites compared to the statistically predicted number of sites, our three ZF protein binds about 25 000 sites across the genome, that is close to a statistically predicted number of occurrence, approximately 23 000 sites, for a 9 nt sequence, genome-wide, assuming equal distribution of the four different DNA bases. Although the two studies were based on ZF proteins targeting two different sites and thus not directly comparable, these findings do imply that adding more ZFs may not necessarily improve DNA binding specificity of the AZF proteins, in the way one might have expected. Therefore, in the current study, we investigated the effect of fusing a non-DBD that is implicated in the *in vivo* DNA binding specificity (14) to a typical three ZF artificial protein.

In this study, we have carried out conventional ChIP-Seq experiments to define the genomic binding sites of a typical AZF protein. We selected the AZF directed against the *VEGF-A* gene promoter because this was one of the first synthetic DNA-binding proteins generated and to date no genomic profiling of its target sites has been reported. This AZF is a 'first generation' synthetic DNA-binding protein that contains three classical C_2H_2 ZFs designed to recog-
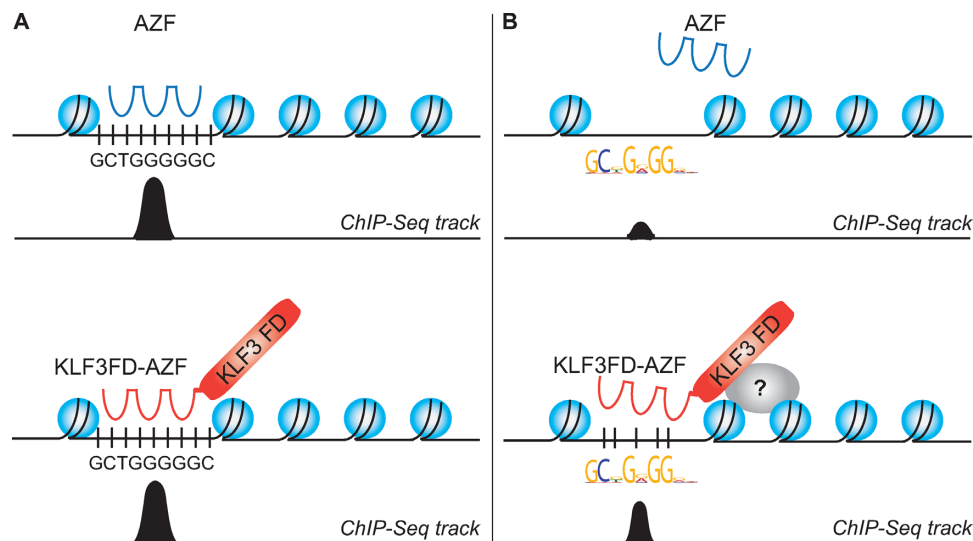
Figure 7. Models illustrating possible mechanisms of AZF and KLF3FD-AZF binding to the 9 nt target sequence and KLF3FD-AZF but not AZF to a more degenerate sequence *in vivo*. This mechanism outlines the involvement of DNA-binding domain of a transcription factor, in this case, zinc fingers (ZFs) of KLF3 or artificial ZFs of KLF3FD-AZF, to establish a protein–DNA interaction, and also of the FD of the transcription factor, in this case, KLF3 FD to interact with a component possibly via protein–protein interaction to facilitate and specify transcription factor binding to DNA. (**A**) Both AZF and KLF3FD-AZF bind to the 9 nt target sequence GCTGGGGGC *in vivo*, which mainly rely on ZFs–DNA interactions. This represents regions that are bound equivalently by both AZF and KLF3FD-AZF, as depicted by the ChIP-Seq track. (**B**) KLF3FD-AZF but not AZF binds to a degenerate site via weak ZFs–DNA interaction and the binding is facilitated further by specific interaction between the KLF3 FD and another molecule, which could be another protein, histone or histone mark, or RNA.

nize the 9 nt sequence GCTGGGGGC (15). We found that it did bind the *VEGF-A* locus as expected.

We also observed peaks at around 25 000 additional genomic locations. In addition, we observed binding to related motifs, where 8 or fewer of the 9 nts were conserved, as well as some sites where no clear motif could be identified. These results are similar to those obtained with natural transcription factors. Natural transcription factors are also found to generate ChIP-Seq peaks at multiple genes, including genes that do not contain recognizable canonical binding sites, and genes that are not functionally regulated by the transcription factor (7,14,40).

In the case of natural transcription factors, the possibility that the protein is docking to a biologically relevant partner protein and thereby localizing to its target gene indirectly via protein–protein interactions is often evoked to explain binding to non-canonical sites (7,8,41). This may also be occurring with the AZF, however, as a minimal ZF protein, it is uncertain whether it has many high-affinity partners. Overall we favor the hypothesis that *in vivo* synthetic AZF proteins are simply more promiscuous in their binding than has previously been suspected.

It should be noted that developments in this area have led to the use of more artificial DNA-binding proteins to either work synergistically (42,43), or, in the case of genomic nucleases, dimerise to a heterologous partner so that full binding is specified by more ZFs (44). It remains to be definitively determined how specific such systems are.

In addition to the results on the minimal AZF we also studied a fusion protein containing the N-terminal FD from the well-studied transcription factor KLF3. KLF3 itself has three classical C_2H_2 ZFs, so by fusing the FD upstream of the three ZF AZFs, we are essentially producing a protein

that is related to KLF3, but has different ZFs. Accordingly, we found that the KLF3FD-AZF fusion protein was stable, detectable in Western blotting and bound the expected 9 nt AZF recognition element in EMSA experiments *in vitro*.

When we tested the genomic binding profile of the KLF3FD-AZF fusion protein by ChIP-Seq experiments we detected around 50 000 peaks, that is, about twice as many peaks as observed with the AZF alone. In many cases the KLF3FD-AZF generated peaks reside at the same location as the AZF alone. There were also what appeared to be new peaks. Although it is impossible to distinguish whether the binding was unique to the KLF3FD-AZF protein or whether some level of residual peak was generated by AZF alone but at levels that were below the background noise in our ChIP-Seq experiment. Irrespective of which interpretation one takes it is clear that addition of the KLF3 FD altered the behavior of the AZF protein.

We examined the elements where we observed strong peaks with the KLF3FD-AZF fusion protein and weaker or no peaks with the AZF alone. We chose a cut-off where the KLF3FD-AZF peaks were at least twice as high. These peaks had a number of interesting features. Firstly they were enriched in promoters (46% were in promoter regions). Second, they tended to contain a sequence that resembled but was not identical to the 9 nt recognition element of the AZF. Third, approximately 34% of peaks that are mapped to known KLF3 bound gene promoters are the genes where binding had previously been shown to depend upon the presence of the KLF3 FD in loss-of-function experiments. That is, both loss-of-function and gain-of-function experiments implicated the KLF3 FD in targeting its cargo to a shared set of gene promoters.

These results provide evidence that the KLF3 FD, the N-terminal, non-DBD, that is thought to function primarily in turning genes off by recruiting the co-repressor CtBP, has a second function in helping localize the protein to its target genes. Such a situation is not unprecedented and aligns with recent work on other Sp/KLF family transcription factors, namely Sp2 and Sp3 (8). Sp3 was found to function as a conventional DNA-binding protein with its 3 ZF domain being necessary and sufficient to localize the protein to CACCC and GC rich elements in the genome. On the other hand the ZF domain of Sp2 was entirely dispensable for such activity and the N-terminal FD of Sp2 was necessary and sufficient for promoter targeting.

These results are of considerable interest because they may help explain why different family members, for example, different members of the KLF family, bind and regulate different sets of target genes *in vivo* – despite having near identical DBDs (11,14,45,46). One might expect all KLFs bind to all CACCC or GC-containing regulatory elements, but it now seems more likely that gene targeting is orchestrated by the combined action of the shared ZF DBD and the N-terminal FDs, and since these FDs differ dramatically between KLF family members, the sets of target genes also differ. This is also supported by our results on the DBD-deficient KLF3 FD, which is capable of *in vivo* chromatin binding, in the absence of a DBD. The reduced chromatin binding affinity observed compared to the KLF3FD-AZF is likely to be associated with protein–protein interactions that are secondary to the DNA–protein interactions of the DBD, and in this case, the AZF domain.

The final question concerns how the FDs achieve gene targeting. Several possibilities could be envisaged. They could bind directly to DNA, to a gene-localized non-coding RNA, to another DNA-bound transcription factor or cofactor, or to a specific histone or transcription factor mark. We are currently investigating these possibilities. To date we have found no good evidence that the FD of KLF3 binds to DNA or RNA, nor does it contain any known histone or histone mark binding domain (such as a bromodomain or chromodomain). It is known to bind the co-repressor CtBP (13). CtBP can dimerize and is resident at many different promoters so this is an attractive possibility. Previous experiments using a mutant version of KLF3 unable to bind CtBP showed that CtBP may be important for proper KLF3 localization to the promoters. However, this mutant showed largely similar genomic binding profiles to that of the wild-type KLF3 protein at the common KLF3 FD dependent promoter binding sites (the 578 KLF3 FD dependent promoter bindings resulted from overlapping the two related KLF3 and KLF3 FD-AZF ChIP-Seq), and thus it seems unlikely that binding to CtBP completely explains the genomic binding patterns (14). In the case of Sp2 the FD localized to a duplicated CCAAT element that is known to be bound by NFY (8) but no direct binding to NFY was detected, so in both cases the precise mechanism behind the gene targeting remains unknown.

Our results on the AZF and KLF3FD-AZF fusion are illustrated in a model in Figure 7. As shown, AZF binds to genes containing its canonical binding site or a near fit to this site. On the other hand it is unable to bind to sites that diverge too far from the ideal consensus element. When

the KLF3 FD is fused to the AZF, the fusion protein can also recognize optimal binding sites where the sequence is conserved. In addition, some property of the KLF3 FD generally allows localization to weak, divergent consensus elements. As explained above, this property may relate to protein–protein interactions made by the KLF3 FD as illustrated.

It is possible that many natural transcription factors have evolved to use both protein–DNA contacts made through their DBDs, as well as protein–protein interactions via their FDs in order to localize to the full repertoire of their target genes. Recognizing this possibility and analyzing more FDs for DNA-binding activity or specificity functions, may reveal insights that will be useful in designing AZF-FD fusion proteins with increased or even more restricted AZF specificity in the future.

ACCESSION NUMBER

The ChIP-Seq data from this article are available from the Gene Expression Omnibus (GEO) under accession number GSE69739.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Fields,S. and Song,O. (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245–246.
2. Klug,A. (2010) The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu. Rev. Biochem.*, **79**, 213–231.
3. Joung,J.K. and Sander,J.D. (2013) TALENs: a widely applicable technology for targeted genome editing. *Nat. Rev. Mol. Cell. Biol.*, **14**, 49–55.
4. Bogdanove,A.J. and Voytas,D.F. (2011) TAL effectors: customizable proteins for DNA targeting. *Science*, **333**, 1843–1846.
5. Tsang,A.P., Visvader,J.E., Turner,C.A., Fujiwara,Y., Yu,C., Weiss,M.J., Crossley,M. and Orkin,S.H. (1997) FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. *Cell*, **90**, 109–119.
6. Merika,M. and Orkin,S.H. (1995) Functional synergy and physical interactions of the erythroid transcription factor GATA-1 with the Kruppel family proteins Sp1 and EKLF. *Mol. Cell. Biol.*, **15**, 2437–2447.
7. Kassouf,M.T., Hughes,J.R., Taylor,S., McGowan,S.J., Soneji,S., Green,A.L., Vyas,P. and Porcher,C. (2010) Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. *Genome Res.*, **20**, 1064–1083.

8. Volkel,S., Stielow,B., Finkernagel,F., Stiewe,T., Nist,A. and Suske,G. (2015) Zinc Finger Independent Genome-Wide Binding of Sp2 Potentiates Recruitment of Histone-Fold Protein Nf-y Distinguishing It from Sp1 and Sp3. *PLoS Genet*, **11**, e1005102.

9. Crossley,M., Whitelaw,E., Perkins,A., Williams,G., Fujiwara,Y. and Orkin,S.H. (1996) Isolation and characterization of the cDNA encoding BKLF/TEF-2, a major CACCC-box-binding protein in erythroid cells and selected other cells. *Mol. Cell. Biol.*, **16**, 1695–1705.

10. Pearson,R., Fleetwood,J., Eaton,S., Crossley,M. and Bao,S. (2008) Kruppel-like transcription factors: a functional family. *Int. J. Biochem. Cell. Biol.*, **40**, 1996–2001.

11. Kaczynski,J., Cook,T. and Urrutia,R. (2003) Sp1- and Kruppel-like transcription factors. *Genome Biol.*, **4**, 206.

12. Pearson,R.C., Funnell,A.P. and Crossley,M. (2011) The mammalian zinc finger transcription factor Kruppel-like factor 3 (KLF3/BKLF). *IUBMB Life*, **63**, 86–93.

13. Turner,J. and Crossley,M. (1998) Cloning and characterization of mCtBP2, a co-repressor that associates with basic Kruppel-like factor and other mammalian transcriptional regulators. *EMBO J.*, **17**, 5129–5140.

14. Burdach,J., Funnell,A.P., Mak,K.S., Artuz,C.M., Wienert,B., Lim,W.F., Tan,L.Y., Pearson,R.C. and Crossley,M. (2014) Regions outside the DNA-binding domain are critical for proper in vivo specificity of an archetypal zinc finger transcription factor. *Nucleic Acids Res.*, **42**, 276–289.

15. Liu,P.Q., Rebar,E.J., Zhang,L., Liu,Q., Jamieson,A.C., Liang,Y., Qi,H., Li,P.X., Chen,B., Mendel,M.C. *et al.* (2001) Regulation of an endogenous locus using a panel of designed zinc finger proteins targeted to accessible chromatin regions. Activation of vascular endothelial growth factor A. *J. Biol. Chem.*, **276**, 11323–11334.

16. Nagy,J.A., Dvorak,A.M. and Dvorak,H.F. (2007) VEGF-A and the induction of pathological angiogenesis. *Annu. Rev. Pathol.*, **2**, 251–275.

17. Benjamin,L.E. (2001) Glucose, VEGF-A, and diabetic complications. *Am. J. Pathol.*, **158**, 1181–1184.

18. Whitlock,P.R., Hackett,N.R., Leopold,P.L., Rosengart,T.K. and Crystal,R.G. (2004) Adenovirus-mediated transfer of a minigene expressing multiple isoforms of VEGF is more effective at inducing angiogenesis than comparable vectors expressing individual VEGF cDNAs. *Mol. Ther.*, **9**, 67–75.

19. Pawson,E.J., Duran-Jimenez,B., Surosky,R., Brooke,H.E., Spratt,S.K., Tomlinson,D.R. and Gardiner,N.J. (2010) Engineered zinc finger protein-mediated VEGF-a activation restores deficient VEGF-a in sensory neurons in experimental diabetes. *Diabetes*, **59**, 509–518.

20. Price,S.A., Dent,C., Duran-Jimenez,B., Liang,Y., Zhang,L., Rebar,E.J., Case,C.C., Gregory,P.D., Martin,T.J., Spratt,S.K. *et al.* (2006) Gene transfer of an engineered transcription factor promoting expression of VEGF-A protects against experimental diabetic neuropathy. *Diabetes*, **55**, 1847–1854.

21. Rebar,E.J., Huang,Y., Hickey,R., Nath,A.K., Meoli,D., Nath,S., Chen,B., Xu,L., Liang,Y., Jamieson,A.C. *et al.* (2002) Induction of angiogenesis in a mouse model using engineered transcription factors. *Nat. Med.*, **8**, 1427–1432.

22. Snowden,A.W., Zhang,L., Urnov,F., Dent,C., Jouvenot,Y., Zhong,X., Rebar,E.J., Jamieson,A.C., Zhang,H.S., Tan,S. *et al.* (2003) Repression of vascular endothelial growth factor A in glioblastoma cells using engineered zinc finger transcription factors. *Cancer Res.*, **63**, 8968–8976.

23. Kwon,H.S., Shin,H.C. and Kim,J.S. (2005) Suppression of vascular endothelial growth factor expression at the transcriptional and post-transcriptional levels. *Nucleic Acids Res.*, **33**, e74.

24. Kang,Y.A., Shin,H.C., Yoo,J.Y., Kim,J.H., Kim,J.S. and Yun,C.O. (2008) Novel cancer antiangiotherapy using the VEGF promoter-targeted artificial zinc-finger protein and oncolytic adenovirus. *Mol. Ther.*, **16**, 1033–1040.

25. Perdomo,J., Verger,A., Turner,J. and Crossley,M. (2005) Role for SUMO modification in facilitating transcriptional repression by BKLF. *Mol. Cell. Biol.*, **25**, 1549–1559.

26. Schmidt,D., Wilson,M.D., Spyrou,C., Brown,G.D., Hadfield,J. and Odom,D.T. (2009) ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods*, **48**, 240–248.

27. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

28. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

29. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

30. 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.

31. Ross-Innes,C.S., Stark,R., Teschendorff,A.E., Holmes,K.A., Ali,H.R., Dunning,M.J., Brown,G.D., Gojis,O., Ellis,I.O., Green,A.R. *et al.* (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389–393.

32. Robinson,J.T., Thorvaldsdottir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

33. Machanick,P. and Bailey,T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**, 1696–1697.

34. Thomas-Chollier,M., Darbo,E., Herrmann,C., Defrance,M., Thieffry,D. and van Helden,J. (2012) A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nat. Protoc.*, **7**, 1551–1568.

35. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.

36. Xuan,Z., Zhao,F., Wang,J., Chen,G. and Zhang,M.Q. (2005) Genome-wide promoter extraction and analysis in human, mouse, and rat. *Genome Biol.*, **6**, R72.

37. Onori,A., Pisani,C., Strimpakos,G., Monaco,L., Mattei,E., Passananti,C. and Corbi,N. (2013) UtroUp is a novel six zinc finger artificial transcription factor that recognises 18 base pairs of the utrophin promoter and efficiently drives utrophin upregulation. *BMC Mol. Biol.*, **14**, 3.

38. Stolzenburg,S., Rots,M.G., Beltran,A.S., Rivenbark,A.G., Yuan,X., Qian,H., Strahl,B.D. and Blancafort,P. (2012) Targeted silencing of the oncogenic transcription factor SOX2 in breast cancer. *Nucleic Acids Res.*, **40**, 6725–6740.

39. Grimmer,M.R., Stolzenburg,S., Ford,E., Lister,R., Blancafort,P. and Farnham,P.J. (2014) Analysis of an artificial zinc finger epigenetic modulator: widespread binding but limited regulation. *Nucleic Acids Res.*, **42**, 10856–10868.

40. Yu,M., Riva,L., Xie,H., Schindler,Y., Moran,T.B., Cheng,Y., Yu,D., Hardison,R., Weiss,M.J., Orkin,S.H. *et al.* (2009) Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. *Mol. Cell*, **36**, 682–695.

41. Chlon,T.M., Dore,L.C. and Crispino,J.D. (2012) Cofactor-mediated restriction of GATA-1 chromatin occupancy coordinates lineage-specific gene expression. *Mol. Cell*, **47**, 608–621.

42. Garriga-Canut,M., Agustin-Pavon,C., Herrmann,F., Sanchez,A., Dierssen,M., Fillat,C. and Isalan,M. (2012) Synthetic zinc finger repressors reduce mutant huntingtin expression in the brain of R6/2 mice. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E3136–E3145.

43. Perez-Pinera,P., Ousterout,D.G., Brunger,J.M., Farin,A.M., Glass,K.A., Guilak,F., Crawford,G.E., Hartemink,A.J. and Gersbach,C.A. (2013) Synergistic and tunable human gene activation by combinations of synthetic transcription factors. *Nat. Methods*, **10**, 239–242.

44. Urnov,F.D., Miller,J.C., Lee,Y.L., Beausejour,C.M., Rock,J.M., Augustus,S., Jamieson,A.C., Porteus,M.H., Gregory,P.D. and Holmes,M.C. (2005) Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature*, **435**, 646–651.

45. Tallack,M.R., Whitington,T., Yuen,W.S., Wainwright,E.N., Keys,J.R., Gardiner,B.B., Nourbakhsh,E., Cloonan,N., Grimmond,S.M., Bailey,T.L. *et al.* (2010) A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res.*, **20**, 1052–1063.

46. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.