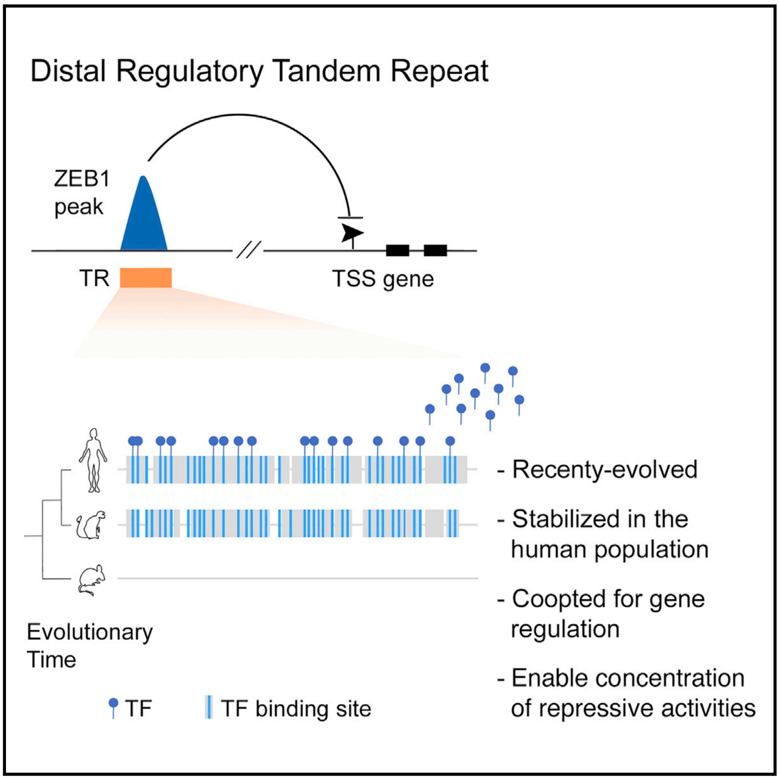
Cell

Article

Co-optation of Tandem DNA Repeats for the Maintenance of Mesenchymal Identity

Graphical Abstract


Distal Regulatory Tandem Repeat
ZEB1 peak
TR
TSS gene
Evolutionary Time
TF
TF binding site
- Recenty-evolved
- Stabilized in the human population
- Coopted for gene regulation
- Enable concentration of repressive activities

Highlights

- ZEB1, a master regulator of mesenchymal programs, binds clustered motifs in TRs
- A ZEB1-bound TR controls the expression of miR-200 family microRNAs
- Deletion of the miR-200-proximal TR causes partial loss of mesenchymal features
- Exapted TRs are integrated into pre-existing networks controlling mesenchymal identity

Authors

Chiara Balestrieri, Gabriele Alfarano,
Marta Milan, ..., Giorgio Scita,
Giuseppe R. Diaferia, Gioacchino Natoli

Correspondence

giuseppe.diaferia@humanitasresearch.it
(G.R.D.),
gioacchino.natoli@hunimed.eu (G.N.)

In Brief

Tandem repeats, iterated and unstable sequences generated by DNA replication errors, can be integrated into ancient gene regulatory networks controlling mesenchymal identity and stabilized in the human genome.

Balestrieri et al., 2018, Cell *173*, 1–15
May 31, 2018 © 2018 Elsevier Inc.
https://doi.org/10.1016/j.cell.2018.03.081

CellPress

Please cite this article in press as: Balestrieri et al., Co-optation of Tandem DNA Repeats for the Maintenance of Mesenchymal Identity, Cell (2018), https://doi.org/10.1016/j.cell.2018.03.081

Co-optation of Tandem DNA Repeats for the Maintenance of Mesenchymal Identity

Chiara Balestrieri,[1,8] Gabriele Alfarano,[1,6,8] Marta Milan,[1,6,8] Valentina Tosi,[1,6] Elena Prosperini,[2] Paola Nicoli,[1]
Andrea Palamidessi,[3] Giorgio Scita,[3,4] Giuseppe R. Diaferia,[1,7,8,*] and Gioacchino Natoli[2,5,7,9,*]
[1]Department of Experimental Oncology, European Institute of Oncology (IEO), Via Adamello 16, 20139 Milan, Italy
[2]Humanitas Clinical and Research Center, Via Manzoni 56, 20089 Rozzano, Milan, Italy
[3]IFOM, The FIRC Institute for Molecular Oncology, Via Adamello 16, 20139 Milan, Italy
[4]Department of Oncology and Hemato-Oncology, University of Milan, 20122 Milan, Italy
[5]Humanitas University, Via Rita Levi Montalcini 4, 20090 Pieve Emanuele, Milan, Italy
[6]These authors contributed equally
[7]These authors contributed equally
[8]Present address: Humanitas University, Via Rita Levi Montalcini 4, 20090 Pieve Emanuele, Milan, Italy
[9]Lead Contact
*Correspondence: giuseppe.diaferia@humanitasresearch.it (G.R.D.), gioacchino.natoli@hunimed.eu (G.N.)
https://doi.org/10.1016/j.cell.2018.03.081

SUMMARY

Tandem repeats (TRs) are generated by DNA replication errors and retain a high level of instability, which in principle would make them unsuitable for integration into gene regulatory networks. However, the appearance of DNA sequence motifs recognized by transcription factors may turn TRs into functional *cis*-regulatory elements, thus favoring their stabilization in genomes. Here, we show that, in human cells, the transcriptional repressor ZEB1, which promotes the maintenance of mesenchymal features largely by suppressing epithelial genes and microRNAs, occupies TRs harboring dozens of copies of its DNA-binding motif within genomic loci relevant for maintenance of epithelial identity. The deletion of one such TR caused *quasi*-mesenchymal cancer cells to reacquire epithelial features, partially recapitulating the effects of *ZEB1* gene deletion. These data demonstrate that the high density of identical motifs in TRs can make them suitable platforms for recruitment of transcriptional repressors, thus promoting their exaptation into pre-existing *cis*-regulatory networks.

INTRODUCTION

Due to their complexity as well as to experimental and analytical difficulties, repetitive DNA elements represent a major component of mammalian genomes that is still awaiting a complete molecular and functional understanding. While in most cases such elements represent non-functional evolutionary relics, in others they acquired a specific regulatory role and were thus incorporated into pre-existing *cis*-regulatory networks and stabilized in evolution, a process indicated as exaptation (Chuong et al., 2017). Analyses of the genomic distribution of mobile elements indicate that some networks were more prone than others to evolutionary innovations involving the cooptation of repetitive DNA. In particular, the molecular pathway controlling cell adhesion is associated with a remarkably high rate of recent exaptation of mobile elements, as indicated by the high frequency of *cis*-regulatory elements originating from transposons in the vicinity of genes encoding cell adhesion molecules (Lowe et al., 2007). The propensity of this pathway to undergo regulatory innovations may reflect the involvement of changes in cell adhesion in processes critical for evolution such as those involved in the wiring of synaptic connections in the brain (Yogev and Shen, 2014).

Changes in cell adhesion are also critical for epithelial-to-mesenchymal transition (EMT), a broad and heterogeneous spectrum of dynamic processes that occur during development and tissue repair, but also during transformation of cancer cells (Nieto et al., 2016). In EMT, cells partially or completely lose epithelial features, such as the ability to make stable cell-cell junctions, to maintain cell polarity or to express basal membrane components and epithelial secretory molecules. Concurrently, they acquire mesenchymal properties, notably the ability to migrate away from their original location in the tissue. The heterogeneity of EMT is due to the complexity of the transcriptional circuits that supervise the loss of epithelial features on the one hand and the acquisition of mesenchymal properties on the other. For example, in sea urchin, 13 different transcriptional regulators have been identified that participate in multiple regulatory modules, each one controlling different epithelial or mesenchymal programs (Saunders and McClay, 2014). In mammals, several transcriptional regulators and microRNAs (miRNAs) have been identified that enforce (e.g., ELF3/5, KLF5, GRHL2/3, miR-200) or suppress (e.g., SNAIL1/2, TWIST1/2, ZEB1) the maintenance of the epithelial state (De Craene and Berx, 2013; Lee et al., 2014; Sánchez-Tilló et al., 2012).

Among these, ZEB1 stands out as a potent inhibitor of epithelial identity whose increased expression is associated with the acquisition of mesenchymal features. ZEB1 is part of a transcriptional co-repressor complex (CtBP) containing histone H3K9 methyltransferases and histone deacetylases (Furusawa et al.,

Please cite this article in press as: Balestrieri et al., Co-optation of Tandem DNA Repeats for the Maintenance of Mesenchymal Identity, Cell (2018), https://doi.org/10.1016/j.cell.2018.03.081

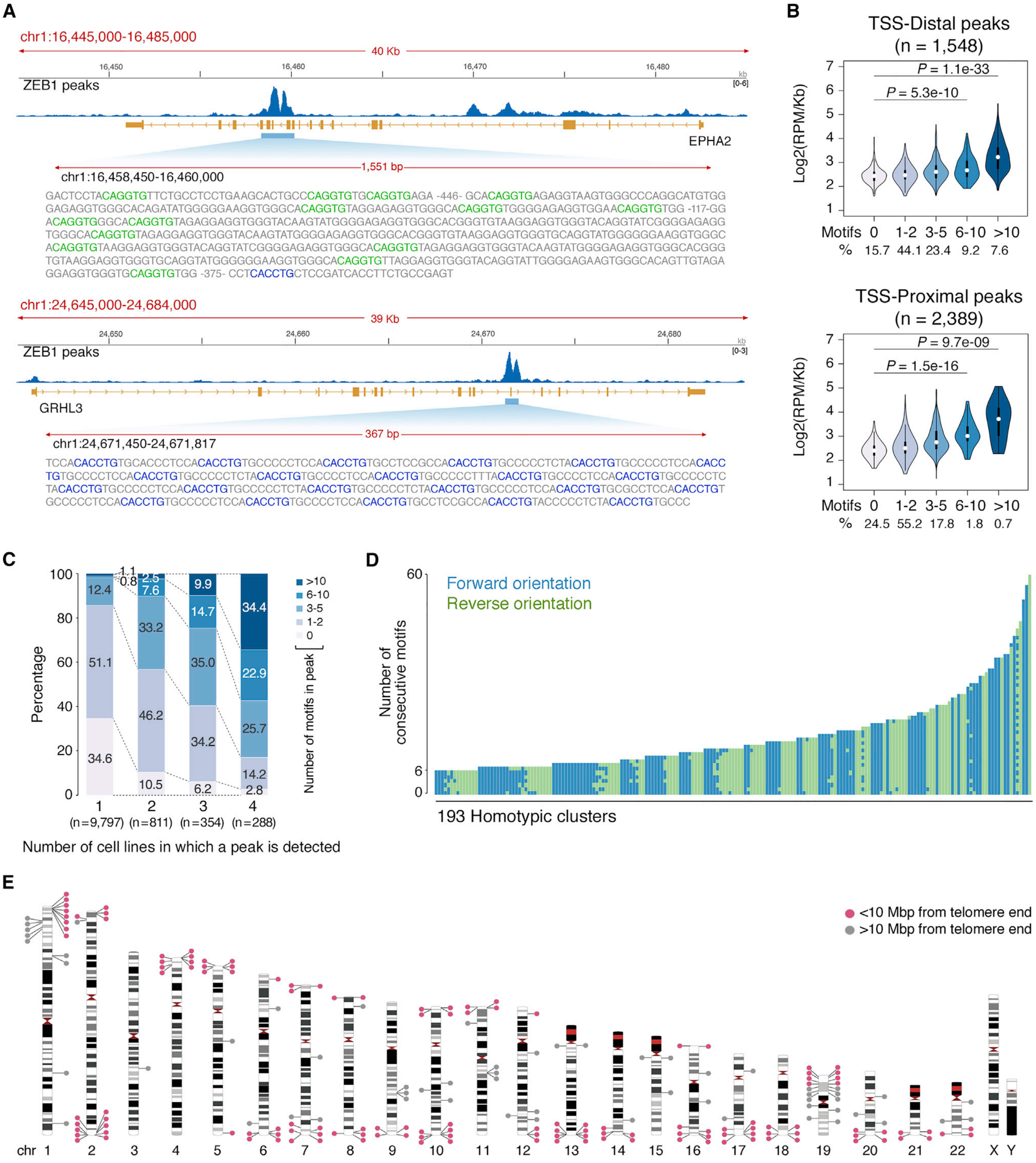


Figure 1. Tandem Repeats Containing Highly Clustered ZEB1 Motifs Promote Efficient ZEB1 Genomic Recruitment across Cell Types

(A) High-density clusters of motifs associated with ZEB1 ChIP-seq peaks. Two genomic regions containing epithelial identity genes are shown. The ZEB1 motifs in the underlying genomic sequences are indicated.

(B) Relationship between ZEB1 ChIP-seq peak intensity and number of ZEB1 motifs in the DNA sequence. Data refer to TSS-distal (top) and TSS-proximal (bottom) peaks in MiaPaCa2 cells. The percentage of peaks (%) for each class of motifs is indicated. White central dots represent the median. Statistical significance was calculated using a one-tailed Wilcoxon rank-sum test.

(C) Relationship between number of ZEB1 motifs and detection of ChIP-seq peaks across multiple cell lines. Peaks detected in all of the four cell lines tested were enriched for sequences containing more than ten motifs.

(legend continued on next page)

1999; Postigo and Dean, 1999; Shi et al., 2003) and is recruited to DNA via two zinc finger domains, each one recognizing an identical hexameric site (5′-CACCTG-3′) (Remacle et al., 1999). ZEB1 acts by suppressing expression of both epithelial genes such as *CDH1* (encoding E-cadherin) and miRNAs of the miR-200 family, which maintain epithelial identity by repressing a network of targets controlling cytoskeleton dynamics, invasion, and migration (Bracken et al., 2014). miR-200 family miRNAs also directly inhibit ZEB1 expression, thus generating a negative feedback loop (Bracken et al., 2008; Park et al., 2008). Whereas ZEB1 expression is dynamically regulated in response to the transient exposure of normal or neoplastic cells to EMT inducers such as transforming growth factor β (Chaffer et al., 2013), some tumor cells are trapped in a stable *quasi*-mesenchymal state characterized by constitutively high levels of ZEB1 (Diaferia et al., 2016).

In the context of a systematic analysis of regulatory circuits controlling maintenance of epithelial features in human pancreatic ductal adenocarcinoma (PDAC) cells (Diaferia et al., 2016), we found that ZEB1 is recruited not only to canonical *cis*-regulatory elements, but also to tandem repeats (TRs) containing a large number of clustered ZEB1 motifs. TRs are a family of repetitive elements that, in most cases, arise from and are highly prone to DNA replication errors, whereby a short nucleotide sequence (the TR unit) is duplicated in an iterative manner, thus resulting in the generation of a series of units in tandem with the same orientation (Ellegren, 2004; Gemayel et al., 2010). Such intrinsic instability determines variability in the number of units of individual TRs in the population, as well as polymorphisms in the sequence of each unit, with mutation rates up to 10^{-3} per cell division (Gemayel et al., 2010). Because of this high instability, cooptation of TRs for the control of gene expression is an unlikely occurrence in evolution. Indeed, while it is well established that inter-individual variability in the number of units of some TRs can impact the activity of adjacent *cis*-regulatory elements and therefore transcription of nearby genes (Gymrek et al., 2016; Martin et al., 2005; Vinces et al., 2009), the possibility that TRs can be exapted and integrated into normal transcriptional regulatory circuits is counterintuitive and is not supported by current data. In this study, we show that, analogously to other non-functional DNA sequences that acquired a *cis*-regulatory role in evolution (Villar et al., 2015), evolutionary recent TRs were integrated into the pre-existing transcriptional regulatory circuit controlling mesenchymal identity and were thus fixed and constrained in the human genome.

RESULTS

ZEB1 Binds Homotypic Clusters of DNA-Binding Motifs

To gain insight into the mechanism of action of ZEB1, we used ChIP-seq to analyze its genomic distribution in MiaPaCa2, a PDAC cell line with *quasi*-mesenchymal features and the ability to form poorly differentiated tumors in xenografted mice (Sipos et al., 2003). ZEB1 bound a relatively limited number of genomic sites (3,937) with a preference for transcription start site (TSS)-proximal regions (2,389 peaks, 60.7%) (Figure S1A and Table S1). Motif discovery analysis retrieved a top-scoring motif that matched the known ZEB1 site in the JASPAR database ($P = 8.47e\text{-}07$) in 483 out of the top 500 peaks (96.6%), thus confirming the specificity of the detected protein-DNA interactions (Figure S1B). Consistent with the ability of ZEB1 to suppress epithelial identity, a gene ontology (GO) analysis on the genes bound by ZEB1 at their promoter retrieved GO terms related to cell-cell junctions and migration (Figure S1C and Table S1).

Exploration of the DNA sequence underlying the ZEB1 peaks associated with the genes involved in epithelial identity showed the common occurrence of clusters of sequences with 100% identity to the ZEB1 motif (homotypic clusters, Table S1). For instance, a peak containing 15 motifs was found in the Ephrin A2 (*EPHA2*) locus, and one intronic cluster of 22 motifs was bound by ZEB1 in the *GRHL3* gene (Figure 1A). As additional examples, two peaks associated with clusters of dozens of motifs were found at ∼40 kb from the cadherin 4 (*CDH4*) TSS; multiple clusters, including two TSS-distal clusters of 16 and 14 sites, respectively, were associated with the *ARHGEF16* gene, encoding a guanine exchange factor for Rho GTPases (Table S1). Homotypic clusters of ZEB1 motifs were also present in the ±2.5 kb regions surrounding the TSS of the orthologs of these genes in other species and in various genes involved in epithelial specification (Table 1). Therefore, homotypic clusters of ZEB1 motifs in the TSS-proximal regions of genes controlling epithelial identity are a common occurrence in evolution.

Clusters of ZEB1 Motifs Coincide with Tandem Repeats

Binding of ZEB1 to homotypic clusters may increase the local concentration of co-suppressor activities, thus enabling efficient repression. We therefore analyzed the entire ZEB1 ChIP-seq data set for the occurrence of homotypic site clusters. We found 318 TSS-proximal or distal homotypic clusters consisting of six or more perfect matches to the ZEB1 motif (Table S1). A clear trend was observed whereby the number of motifs and the intensity of the ChIP-seq signal were correlated (Figure 1B), indicating that the repetition of individual motifs facilitated the local accumulation of ZEB1. A closer inspection of the homotypic clusters of ZEB1 motifs revealed that, in many cases, they were part of highly repeated sequences identified by computational analyses (Benson, 1999) as TRs (Table S2). Importantly, blacklisted genomic regions identified by ENCODE were filtered out, and ambiguously mapped reads were discarded. Differences in the DNA sequences separating individual ZEB1 motifs in TRs explain the mappability of such genomic regions. Because of the frequent association of ZEB1 with TRs, we explored the possibility that ZEB1-bound TRs may be involved in the transcriptional circuitry controlling mesenchymal identity.

(D) Orientation of ZEB1 motifs in genomic regions bound by ZEB1 in multiple cell lines. Homotypic clusters (columns) were sorted based on the number of consecutive motifs. Each motif was color coded based on its orientation.

E) Chromosomal distribution of TRs bound by ZEB1. 193 homotypic clusters are indicated by dots. Clusters within 10 Mbp from telomere ends are indicated as pink dots, all others as gray dots.

See also Figure S1 and Tables S1, S2, and S3.

Table 1. Homotypic Clusters of ZEB1 Motifs Associated with the Promoters of Epithelial Identity Genes

	Homo sapiens		*Mus musculus*	*Bos taurus*	*Rattus norvegicus*
	Nr	Peak intensity (RPM)	Nr	Nr	Nr
Epithelial polarity proteins					
Pard6b	4	5.49	11	5	9
Crb3	5	6.85	15	23	17
Inadl (Patj)	4	4.39	3	5	4
Plk5	21	16.04	11	nd	9
Rab17	6	6.58	7	nd	8
Rabep2	6	10.32	1	4	1
Plxnd1	12	17.03	11	nd	10
Flna	5	13.94	5	15	3
Tight and adherens junction proteins					
Tjp2	4	11.94	7	2	10
Cldn7	7	3.87	10	8	10
Ocln	4	28.52	9	10	12
Jup	2	5.28	9	10	7
F11r	5	16.26	3	6	6
Arhgef16	9	5.96	12	11	14
Pcdh1	4	11.41	5	7	nd
Epithelial transcription factors					
Elf3	2	10.67	4	5	6
Ovol2	3	8.25	8	6	6
Other epithelial genes					
Muc4	6	3.52	5	nd	nd
Adam6	11	5.09	4	nd	4
Scrib	5	5.17	12	nd	8
Grb7	7	8.96	11	10	16
Krt15	5	4.76	6	5	4

A selected set of epithelial identity genes associated with ZEB1 peaks in PANC1 cells is shown (complete list in Table S1). The intensity of the corresponding ChIP-seq peak (RPM, reads per million) is shown. Nr, number of perfect matches to the ZEB1 motif. The three columns on the right report the number of perfect ZEB1 motifs (in both orientations) in the TSS-proximal regions (±2.5 kb relative to mapped TSS) of the same genes in the *Mus musculus*, *Bos Taurus*, and *Rattus Norvegicus* genomes. nd, not determined (genes with no RefSeq annotation in the species).

We first determined whether the association between ZEB1 and clustered motifs in TRs could be observed in other cell types. We performed ZEB1 ChIP-seq in colon carcinoma RKO cells, and we analyzed our previously published data set in PANC1 cells, a *quasi*-mesenchymal pancreatic carcinoma cell line (Diaferia et al., 2016), and a publicly available data set generated in GM12878, a lymphoblastoid cell line. The analysis of these datasets confirmed the correlation between number of motifs and signal intensity (Figure S1D). We identified three features of ZEB1 motif clusters. First, while peaks containing one or two motifs were commonly detected only in a single cell line, those containing more than six motifs showed a strong tendency to be detected in all cell lines analyzed (Figure 1C and Table S3). It is likely that a high number of clustered motifs enabled efficient recruitment of ZEB1 in spite of the differences in the accessible *cis*-regulatory landscapes of individual cell types. Second, motifs within clusters in most cases showed an identical orientation (Figure 1D), which is consistent with the mechanisms leading to the addition of identical units in TRs. Finally, we noticed an overall strong bias of the highly clustered motifs toward chromosome ends ($P = 3.1e\text{-}28$ by Fisher's exact test), which may relate to both the intrinsic instability of subtelomeric regions and their propensity to favor a transcriptionally repressive environment (Figure 1E). From an evolutionary point of view, the genomic regions bound by ZEB1 that contained TRs were less conserved than the other ZEB1-bound regions ($p = 3.3e\text{–}19$ by one-tailed Wilcoxon rank-sum test) (Figure S1E).

A TR Bound by ZEB1 Upstream of the miR-200b Locus

One of the most prominent ZEB1 genomic peaks, located ~45 kb upstream of a cluster of three miRNAs of the miR-200 family (miR-200b, miR-200a, and miR-429), contained a high density of ZEB1 motifs, with 37 perfect matches in a 0.8 kb genomic region (Figure 2A). Additionally, this cluster was bound by ZEB1 in all cell lines tested (Figures S2A and S2B). The sequence underlying this peak contained 33 tandem repetitions (with a minimum 70% identity) of a degenerate 24 nt unit (Figure 2A). Since ZEB1 works in part by suppressing the expression of miR-200 family miRNAs, we considered the possibility that this TR was coopted for ZEB1-mediated regulation of the adjacent miR-200 cluster. We first analyzed this TR from an evolutionary point of view. Since low-complexity repeats are frequently excluded from genome assemblies, we focused on the four mammalian genomes (mouse, rat, cow, and dog) in which annotation of repeats is extensive and comparable to that in the human genome. The sequences flanking the TR, but not the TR itself, showed detectable conservation (Figure 2B). Moreover, no clustered ZEB1 motifs were found in a window of 100 kb around the orthologous miR-200b loci in these four species. Conversely, the same subtelomeric TR was readily detected in the Macaque genome. Therefore, this ZEB1-bound TR is an evolutionarily recent acquisition. Interestingly, the same miR-200 cluster has a similar subtelomeric location in the mouse chromosome 4 as well as in some others, but not in all species. While the mouse genome contains a TR in the vicinity of the miR-200b cluster, its basic unit contains a motif with a central nucleotide insertion that disrupts the integrity of the ZEB1 site.

While the TR in the miR-200b locus appears to be primate specific, the promoter upstream of the TSS of the single transcript encoding miR-200b, miR-200a, and miR-429 (Bracken et al., 2008) contains two ZEB1 motifs that are conserved across the human, mouse, rat, and dog genomes (Figures S2A and S2C). This promoter, which was previously shown to enable ZEB1-mediated repression of luciferase reporter plasmids (Bracken et al., 2008), is bound by ZEB1 in our datasets, although peak intensity is about ten times weaker than the upstream TR-associated ZEB1 peak (Figure 2A and Table S1). Overall, these data suggest that the TR upstream of the miR-200b locus is an evolutionary recent addition to an ancestral regulatory circuit based on a conserved promoter element.

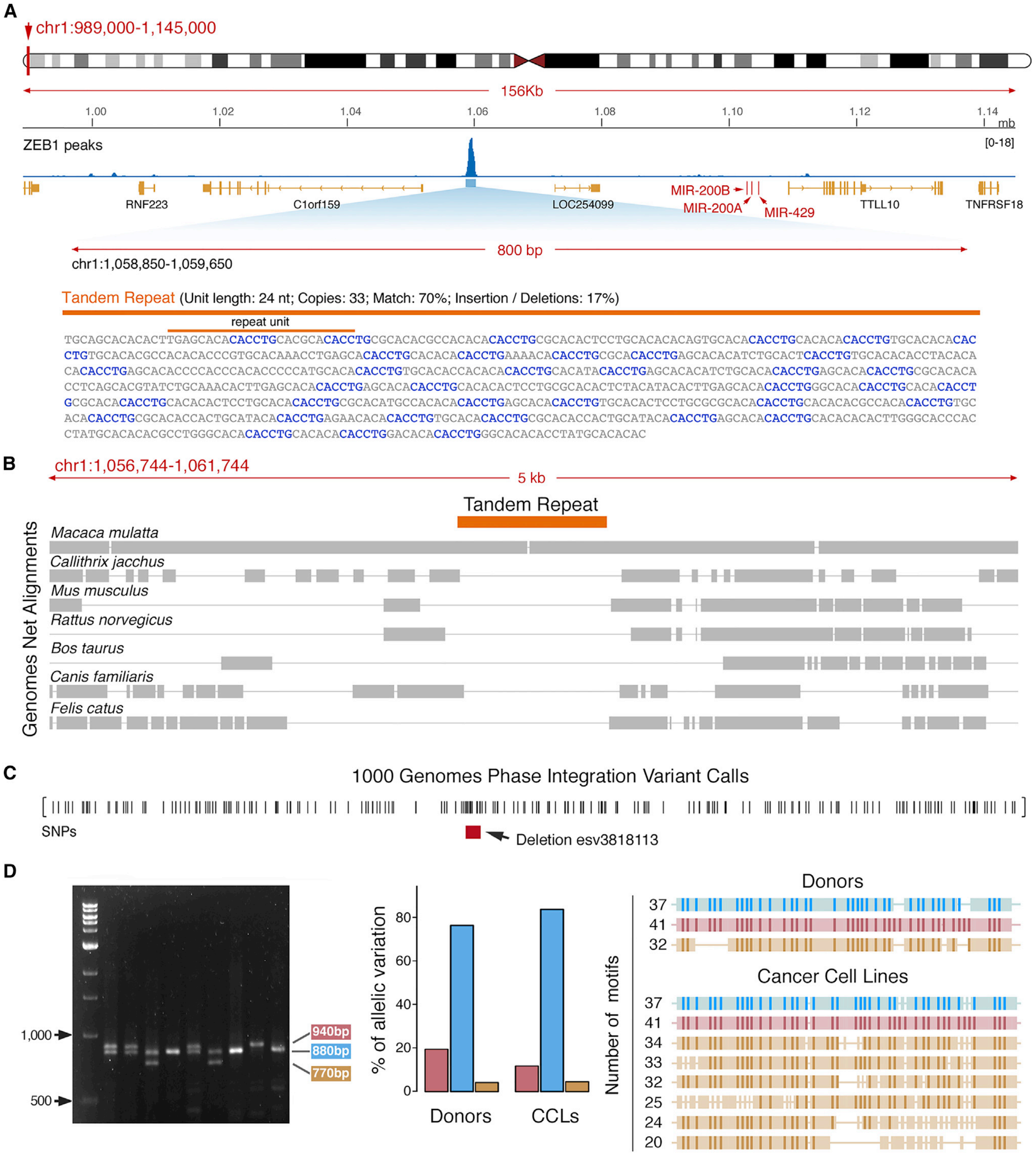


Figure 2. A TR in the miR-200b Locus

(A) Schematics of the miR-200b/miR200a/miR429 locus in the subtelomeric region of chr1, with the ZEB1 ChIP-seq peak and the TR with 37 ZEB1 motifs.

(B) Sequence conservation of the TR upstream of miR-200b locus. The alignments in selected species with completely annotated genomes are shown. Boxes, ungapped alignments; horizontal lines, gaps.

(C) Single-nucleotide polymorphisms (vertical lines) and structural variants (red box) from the 1000 Genomes Project are shown. The only structural variant reported (allele frequency = 6%, 299 out of 5,008 alleles) is indicated.

(legend continued on next page)

Since TRs are intrinsically unstable, albeit with differences in their mutation rates that span orders of magnitude (Gemayel et al., 2010), we first analyzed the TR in the miR-200b locus in the normal human population using data from 2,504 human genomes (Auton et al., 2015; Sudmant et al., 2015). The only reported change in the structure of this TR was the heterozygous loss at 299/5,008 alleles of an 80 nt sequence at the 5′ end, which resulted in a TR containing 33 instead of 37 motifs (Figure 2C). To corroborate these findings and eliminate concerns related to the mapping of repeated sequences by short-read sequencing, we analyzed the miR-200b locus by PCR followed by Sanger sequencing in 50 normal individuals and 78 cancer cell lines. Based on length and sequence, in normal individuals we identified three alleles: the most common allele (76.5%) matched the annotated genomic sequence and contained 37 copies of the ZEB1 motif, followed by a slightly expanded sequence containing 41 ZEB1 motifs (19.4%) and a less common variant with a contracted sequence with 32 ZEB1 sites (Figures 2D and S2D). In cancer cell lines, we observed a similar pattern, but contracted repeats were more heterogeneous in length (Figures 2D and S2E). Overall, this specific TR appears to be relatively stable in structure, undergoing limited expansions and contractions both across normal individuals and in cancer cells.

TRs Bound by ZEB1 Are Functional Silencers

We next determined the functional role of the TR upstream of the miR-200b cluster. In a miRNA-seq analysis, we found that, compared to the ZEB1-positive, poorly differentiated (high-grade), and *quasi*-mesenchymal lines (MiaPaCa2, PANC1 and PT45P1), well-differentiated (low-grade) pancreatic carcinoma cell lines, in which ZEB1 expression is low to undetectable, expressed significantly higher levels of five miRNAs belonging to the miR-200 family (Figure 3A and Table S4). In addition to the TR-proximal cluster on chr1 (that contains miR-200b, miR-200a, and miR-429), ZEB1 bound to two already described motifs just upstream of the miR-200c/miR-141 cluster on chr12 (Burk et al., 2008) (Table S1). We used the nickase variant of Cas9 and two sets of two single-guide RNAs (sgRNAs) each (Ran et al., 2013) to generate paired nicks upstream and downstream of the TR in the miR-200b/miR-200a/miR-429 locus, and we obtained multiple clones in which the intervening sequence was deleted (Δ-Repeat clones) (Figure 3B). ChIP-seq analysis in a pool of Δ-Repeat clones confirmed the selective loss of ZEB1 binding at the TR upstream of miR-200b, with minimal changes in ZEB1 occupancy elsewhere in the genome (Figure 3C). As a benchmark for the effects of the deletion of this TR in MiaPaCa2 cells, we engineered a deletion of the *ZEB1* gene in the same cell line (Figure S3A). *ZEB1* gene loss resulted in the de-repression of 126 genes (Figure S3B and Table S5) that were associated with GO terms related to EMT, such as cell adhesion and cell-cell junction organization (Figure S3C and Table S5). 51 out of these 126 genes (40.5%) were bound by ZEB1 at their promoter, indicating direct repression (Figure S3D).

The availability of both *ZEB1* knockout and Δ−Repeat clones allowed us to determine the contribution of the TR in the miR-200b locus to the phenotype caused by the absence of ZEB1. We initially focused on the expression of miR-200 family microRNAs. All five miR-200 family miRNAs were upregulated in ZEB1-KO cells (Figure 3D). Conversely, the deletion of the TR on chr1 resulted in the selective upregulation of the miRNAs in the adjacent miR-200b cluster. Deletion of the same TR in two other *quasi*-mesenchymal pancreatic cancer cell lines (PANC1 and PT45P1) also resulted in the upregulation of the adjacent miR-200 family miRNAs, indicating that the role of this TR is not restricted to a specific cell line (Figure 3E). To verify that the miR-200b cluster is in the same chromatin domain as the upstream ZEB1-bound TR, we used previous CTCF ChIA-PET datasets (Tang et al., 2015), which confirmed that the miR-200 cluster and the TR are in physical proximity (Figure 3F).

As an additional control, we sought to reinstate repression of miR-200b/miR-200a/miR429 in Δ-Repeat clones by targeting multiple copies of the KRAB repressor domain to sequences adjacent to the deleted TR (Figure S4A) (Tanenbaum et al., 2014). With this approach, we reduced the expression of the three microRNAs to levels comparable to those observed in control cells, indicating that the repressive activity of the TR can be replaced by targeting orthogonal repressive domains to adjacent regions. Finally, we determined the impact of the deletion of the TR on other genes in the same genomic region (Figure 3G). The two most TR-proximal genes, *C1ORF159* and *LOC254099*, were not expressed in wild-type cells, with the latter being upregulated in Δ-Repeat clones, although its expression remained close to the detection threshold. The other genes in the same domain were not significantly affected in Δ-Repeat clones. Overall, ZEB1 binding to the telomeric TR in chromosome 1 was required for transcriptional repression of the downstream miR-200b/miR-200a/miR-429 but was devoid of effects on the other two distally located and ZEB1-inhibited miR-200 family members.

The possibility that other TRs bound by ZEB1 may also act as functional silencers was investigated by acutely deleting six TRs and testing the expression of adjacent genes. In all cases, the deletion efficiency in the polyclonal population selected by puromycin treatment was higher than 90% (Figure S5). In some cases, we detected a significant increase in the expression of one or more adjacent gene(s), while in others no differences were observed, possibly due to redundancy in negative regulatory elements in the same region or to the lack of activators controlling the expression of the corresponding genes (Figure S5). For instance, in the locus containing the *ARHGEF16* gene, the deletion of the TR at the 5′ of the locus increased the expression of *ARHGEF16*, *MEGF6*, and *miR551a*, a microRNA previously described as a suppressor of metastases in colon cancer (Loo

(D) Analysis of the TR in the miR-200b locus in normal individuals and cancer cell lines. (Left) Agarose gel analysis showing a representative group of normal individuals carrying the three main allelic variants, whose sizes are indicated in the boxes. (Middle) Frequency of the three allelic variants in 50 normal human donors and 78 cancer cell lines (CCL). (Right) Schematic diagram of representative Sanger-sequenced samples showing the organization of the chr1 TR upstream of the miR-200 locus in normal donors and cancer cell lines. Each vertical bar represents a ZEB1 motif. Gaps, sequences lost in the TR.
See also Figure S2.

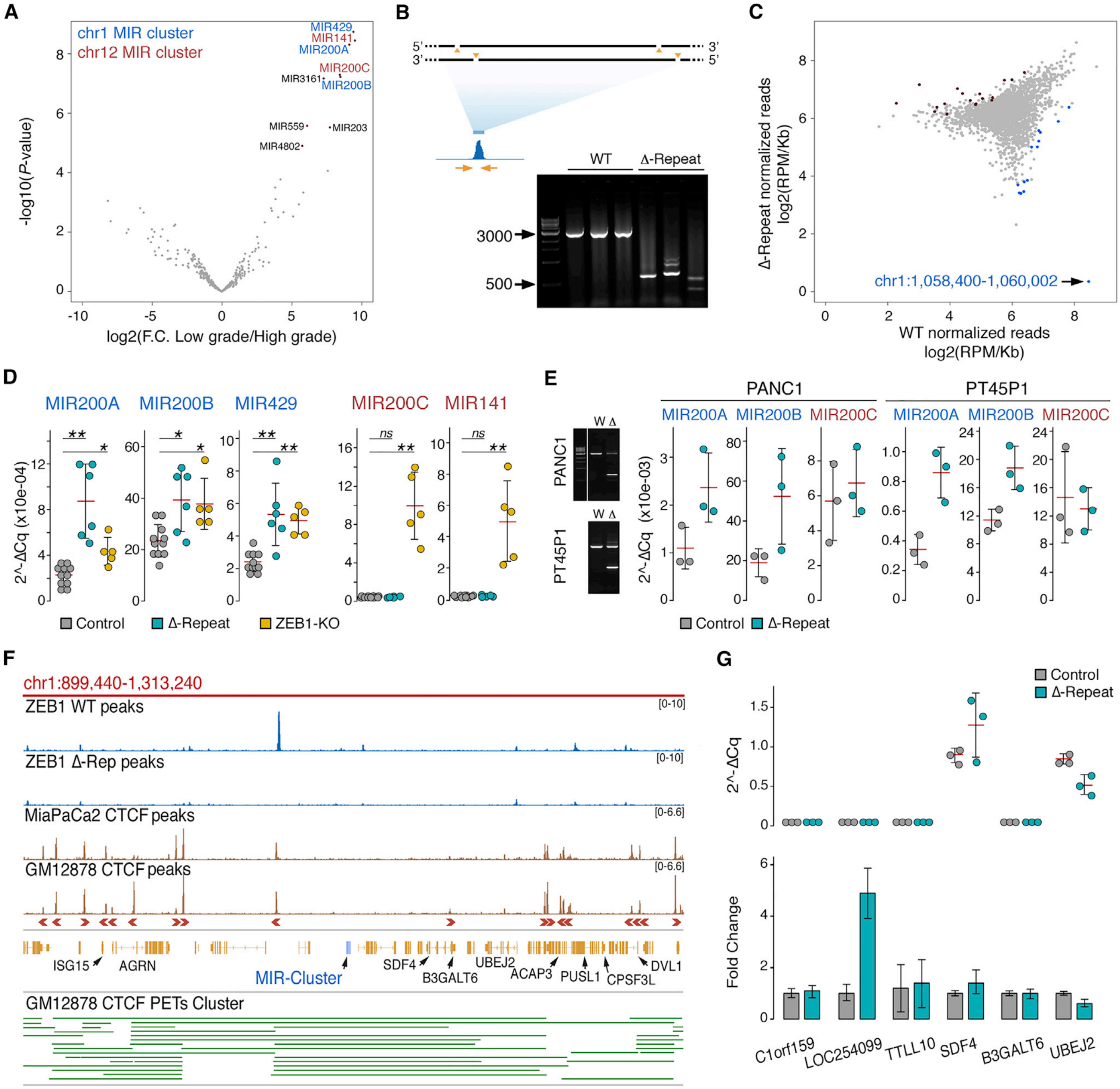


Figure 3. Deletion of the ZEB1-Bound TR in the miR-200b/a Locus Enhances the Expression of miR-200a, miR-200b, and miR-429

(A) Volcano plot showing differential expression of miRNAs (based on miRNA-seq data) in three high-grade and three low-grade PDAC cell lines. The log2 expression fold change (FC) is shown on the horizontal axis and the –log10 of the p value on the vertical axis. Upregulated miRNAs (FDR $\leq$ 0.001 and log2 FC > 1) are indicated by red dots.

(B) Schematics of Cas9 nickase-mediated deletion of the TR upstream of miR-200b locus in MiaPaCa2 cells. The gel on the right side shows the PCR-mediated amplification of the TR in three wild-type and three Δ-Repeat clones.

(C) ZEB1 genomic occupancy was investigated by ChIP-seq in wild-type MiaPaCa2 cells and in a pool of three Δ-Repeat clones. The chr1 peak corresponding to the deleted TR is indicated.

(D) Expression of the miR-200 family miRNAs was tested by RT-qPCR in wild-type, ZEB1 KO, and Δ-Repeat MiaPaCa2 cells. The miR-200 cluster on chr12 contains miR-200c and miR-141, while the cluster on chr1 contains miR-200b, miR-200a, and miR-429. Each dot represents a different wild-type or mutant clone (n = 11 wild-type, n = 6 ZEB1-KO and n=5 Δ-Repeat). Values represent the relative mRNA amount calculated as 2ˆ-ΔCq relative to *miR-103* as reference gene. Mean and SD are shown. *p < 0.01 and **p < 0.001 by two-tailed t test.

(E) Expression of miR-200 family miRNAs in wild-type and Δ-Repeat PANC1 and PT45P1 PDAC cells. Cas9-mediated deletion of the TR was carried out in bulk cell populations (n = 3 independent experiments) and deletion efficiency assessed by PCR (shown in the two agarose gels). W, wild-type cells; Δ, deleted cells. Molecular weight markers are shown on the left.

(legend continued on next page)

et al., 2015), while the other genes in the region were not affected. The deletion of the TR at the 3′ of the locus was instead devoid of effects on these genes but determined the upregulation of *miR551a*. Therefore, *miR551a* is under the negative control of two different ZEB1-bound TRs in the same locus. Consistent with these data, *miR551a* was expressed at higher levels in epithelial than *quasi*-mesenchymal PDAC cells (Table S4). The rather limited magnitude of the observed effects is in keeping both with the general notion that transcriptional repressors are fine-tuners of gene expression programs, rather than on/off switches (Reynolds et al., 2013), and with the combinatorial control exerted by multiple regulatory elements in a given genomic region. Overall, these data indicate that the TRs bound by ZEB1 are integrated into the regulatory circuitry controlling the transcriptional output of *quasi*-mesenchymal cells.

Loss of Mesenchymal Features upon Deletion of the TR in the miR-200b Locus

The observation that the TR in the miR-200b locus was stabilized in the human population and that it controlled the expression of the adjacent miR-200 cluster suggested that this sequence has a defined functional role. Therefore, we set out to dissect its contribution to the overall biological response controlled by ZEB1. Δ-Repeat clones showed an epithelial morphology similar to that of ZEB1-KO clones, with a doubling of their circularity index compared to controls (Figure 4A). Circularity index increases when cells shift from an elongated to a spherical or cuboidal shape. Furthermore, an obvious increase in compact cell clusters even at low plating densities was observed. Measurement of random cell migration by time-lapse microscopy showed that both the deletion of the TR and the loss of *ZEB1* robustly and significantly reduced motility as compared to control cells (Figure 4B and Movie S1). Notably, TR-deleted cells migrated at slightly higher speed than ZEB1 KO cells (Figure 4B) and were capable of extending multiple protrusions, although these were shorter lived and less persistent than those of control cells (Movie S1). ZEB1 KO cells nearly completely failed in extending migratory protrusions. The impaired cell locomotion was mirrored, and possibly caused, by altered cell adhesion to a panel of different substrates (Figure 4C). Loss of ZEB1 caused a robust and significant increase in adhesion as compared to controls, while Δ-Repeat clones adhered less efficiently than ZEB1-KO but more than control cells, which is in keeping with the relative degree of migratory defects. Similarly, Δ-Repeat clones lost invasive capacity in Matrigel assays (Figure 4D) and conversely increased the expression of the epithelial mucin *MUC1* (Figure 4E).

Additional distinguishing properties of epithelial cells include the presence of cadherin-dependent tight cell-cell adherens junctions, which drive the formation of compact colonies and a stereotypical architectural organization of filamentous actin in focused, linear arrays of fibers confined to junctions. As compared to the epithelial Capan2 cells, MiaPaCa2 cells were unable to form compact colonies and cell-cell adhesions, were devoid of E-cadherin, and displayed a prototypical mesenchymal organization of the actin cytoskeleton into parallel arrays of stress fibers running throughout the cell soma (Figure 4F). *ZEB1* gene loss restored cell compaction and the actin architecture typical of epithelial monolayers with cortical stress fibers primarily confined to cell junctions despite the persistent loss of E-cadherin (Figure 4F). Δ-Repeat clones displayed an intermediate phenotype characterized by a heterogeneous cell morphology. A fraction of Δ-Repeat cells formed tight epithelial-like colonies with cortically confined, junctional F-actin interspersed with more mesenchymal-like cells. Unbiased quantification of the number of cells with arrays of parallel actin stress fibers along the entire cell cytoplasm confirmed the progressive loss of mesenchymal features from WT to Δ-Repeat to ZEB1-KO cells (Figure 4G).

Reinstatement of repression in Δ-Repeat clones by targeting KRAB domains to the TR-adjacent regions reverted the phenotypes (Figures S4B and S4C and Movie S2). Hence, while most of the effects of ZEB1 on cell shape and motility appeared to be mediated by the TR upstream of miR-200b, other phenotypes determined by the loss of ZEB1 were only partially recapitulated in the Δ-Repeat clones.

EMT was previously linked to cancer cell stemness (Mani et al., 2008), and ZEB1 promotes tumorigenicity to a large extent by repressing the expression of the miR-200 family (Wellner et al., 2009). Therefore, we analyzed the impact of the deletion of the TR upstream of miR-200b on growth and tumorigenicity of MiaPaCa2 cells. *In vitro* growth of both ZEB1 and Δ-Repeat clones was indistinguishable from that of their wild-type counterparts (Figure 4H), indicating that the miR-200b cluster does not affect cell proliferation per se. However, when xenografted into nude mice, both ZEB1-KO and Δ-Repeat clones were unable to form tumors (Figure 4I), indicating that the upregulation of miR-200 family members in Δ-Repeat clones was sufficient to dampen the tumorigenic potential of the cells.

ZEB1 Binds to Clustered Sites in an Analog Manner

Highly clustered motifs may promote cooperative recruitment of ZEB1, thus resulting in an all-or-none occupancy of the TR. As opposed to this digital mode of regulation, individual sites may instead be bound separately, with the overall level of occupancy of the TR being determined in an analog fashion by the nuclear concentration of ZEB1. To discriminate between these two possibilities, we measured occupancy of clusters of multiple sites by electrophoretic mobility shift assay (EMSA) in the presence of increasing amounts of nuclear extracts from cells transfected with a ZEB1 expression vector. A DNA probe containing two ZEB1 motifs (2×) generated a single shifted complex when

(F) Genomic snapshot of ZEB1 ChIP-seq in wild-type and Δ-Repeat MiaPaCa2 cells, CTCF ChIP-seq data in MiaPaCa2 and GM12878 cells, and the 3D interactions from CTCF ChIA-PET experiments in GM12878 cells.

(G) Expression of the genes surrounding the TR in the miR-200b locus was tested by RT-qPCR in wild type and Δ-Repeat MiaPaCa2 cells (n = 3 clones each). Values represent relative mRNA amount (top) or fold change expression (bottom). mRNA amount was calculated as 2-ΔCq relative to *C1ORF43* as reference gene, and the fold change is calculated as 2ˆ-ΔΔCq based on the average of the ΔCq of the controls. Means ± SD are shown.

See also Figures S3, S4, and S5 and Tables S4 and S5.

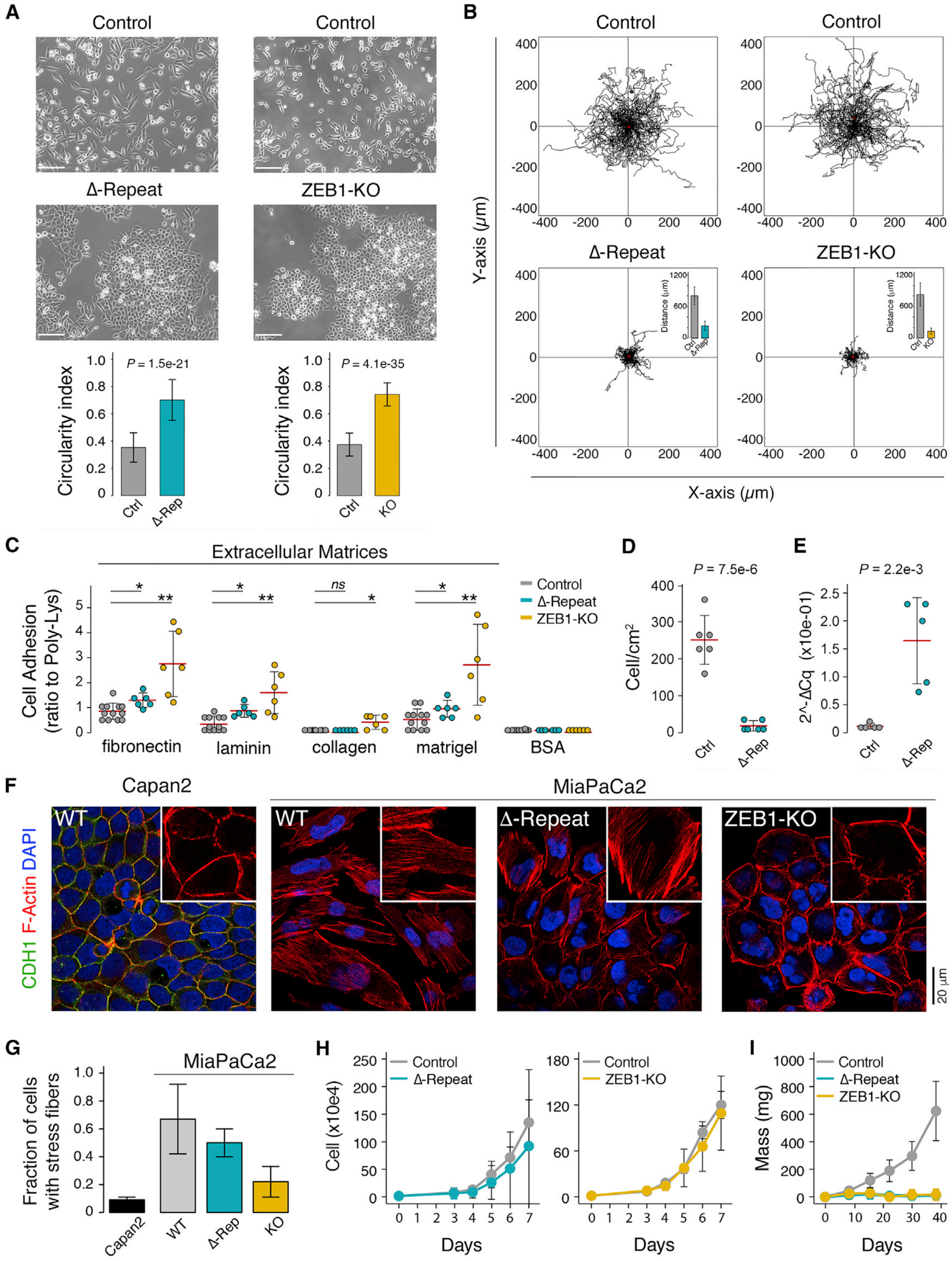


Figure 4. Functional Analysis of Δ-Repeat Clones

(A) Morphology and circularity index in Δ-Repeat, ZEB1-KO MiaPaCa2 cells and matched controls (n = 45 cells per group). Means ± SD are shown. p value by two-tailed t test. Scale bar, 100 μm.

(B) Trajectory plots and accumulated distance bar graph obtained by random migration assay. Δ-Repeat, ZEB1-KO MiaPaCa2 cells and matched controls were analyzed by time-lapse microscopy (n = 75 cells per group). Data were acquired every 5 min over a 24 hr time course.

(legend continued on next page)

incubated with ZEB1, but not mock lysates (Figure 5A, top), consistent with the notion that each ZEB1 molecule contacts one motif via the N-terminal Zn fingers (ZnFs) and a second motif with the ZnFs in the C terminus (Remacle et al., 1999). Using a 4× probe, we detected both a faster migrating complex corresponding to the occupancy of only two motifs and a slower migrating one where all four motifs were bound (Figure 5A, middle). Increasing ZEB1 concentrations resulted in an analog increase in occupancy of all four sites, without evidence for cooperativity. This result was confirmed using a 6× probe: high molecular weight ZEB1-DNA complexes appeared progressively without obvious cooperativity (Figure 5A, bottom). Quantification of the EMSA data is shown in Figure 5B. In keeping with an analog mode of function, transfection of a series of deletion mutants of the miR-200 TR cloned upstream of a constitutive promoter driving luciferase expression indicated a correlation between the number of ZEB1 motifs and repressive activity (Figure 5C).

Binding of TRs by Other Transcription Factors

Finally, we determined how commonly DNA sequence-specific transcription factors bind homotypic clusters of motifs contained within TRs. To this aim, we collected 49 high-quality ChIP-seq datasets deposited in public repositories. Datasets included TF ChIP-seq from the ENCODE consortium (Consortium, 2012) and our previous collection in CFPAC-1, a PDAC cell line (Diaferia et al., 2016). We first analyzed the percentage of ChIP-seq peaks overlapping TRs and then the percentage of those TRs that contained clusters of DNA binding motifs for the cognate transcription factor (Figure 6). Because of the high frequency of TRs in the human genome, we restricted our analysis to TRs with a period of ≥5 nt. The ChIP-seq peaks overlapping TRs ranged between 2% and 15%, depending on the transcription factor. However, the TRs underlying the ChIP-seq peaks contained the cognate transcription factor motif only in a fraction of cases. This suggested that, in most other instances, the TR was in the vicinity of the motif bound by the transcription factor, but it was not directly contacted. Nevertheless, in those specific cases, many of the transcription factors analyzed were found to bind TRs that contained multiple perfect matches to their cognate motif. Remarkably, ZEB1 and SNAI2—which recognize the same consensus, have similar repressive activity, and enforce mesenchymal identity—had the highest frequency of interactions with TRs containing clustered motifs. This property was particularly evident when considering TRs with clusters of more than six motifs (Figure 6). Overall, while these data indicate that motifs competent for binding of various transcription factors do occur in TRs, they also suggest a high propensity of transcriptional repressors controlling mesenchymal identity to use exapted TRs to bring about gene repression.

DISCUSSION

Novel *cis*-regulatory elements, particularly enhancers, are commonly generated within genomic DNA sequences that are non-functional in other species, while promoters tend to be relatively stable in evolution (Villar et al., 2015). The notion that TRs can be exapted to become part of normal gene regulatory circuits is a variant of this theme. TRs are generated by DNA replication errors and typically retain a high level of instability (Legendre et al., 2007). Expansion and contraction of TRs located nearby *cis*-regulatory regions can interfere with transcriptional control through various mechanisms, including alterations of nucleosomal organization and transcription factor binding (Martin et al., 2005; Vinces et al., 2009), and eventually contribute to determine variability in gene expression in the population (Bennett and Todd, 1996; Gymrek et al., 2016). Whereas such variability might enable the adaptation of bacteria and yeast to rapidly changing environments, thus leading to the positive selection of the TR (Ellegren, 2004; Stern et al., 1986; Verstrepen et al., 2005; Weiser et al., 1989), the exaptation of TRs to control critical regulatory networks in metazoans is counterintuitive because the intrinsic instability of TRs may be incompatible with the robustness needed for the control of essential biological processes. The data shown here indicate that, unexpectedly, evolutionarily recent TRs can be integrated into ancestral gene regulatory networks such as the one enforcing the maintenance of epithelial identity.

Specifically, a subtelomeric TR in human chromosome 1 enabled repression by ZEB1 of the adjacent miR-200 cluster, and its deletion was sufficient to convert cells towards a partial epithelial phenotype. The TR associated with the miR-200b locus is remarkably stable and poorly polymorphic both in the normal human population and in cancer cell lines, indicating

(C) Analysis of cell adhesion in ZEB1-KO, Δ-Repeat MiaPaCa2 cells and matched controls. The substrates used for adhesion assays are indicated. All data are expressed as ratio to adhesion to poly-lysine. Each dot represents a wild-type or mutant clone in the different experimental sets (two independent experimental sets performed with n = 6 wild type, n = 3 Δ-Repeat clones, and n = 3 ZEB1-KO clones each). Means ± SD are shown. *p < 0.05 and **p < 0.001 by two-tailed t test.

(D) Matrigel invasion assay of Δ-Repeat MiaPaCa2 cells and matched controls. Data are expressed as cell density, and each dot represents a wild-type or mutant clone in the different experimental sets (two independent experimental sets performed with n = 3 wild-type and Δ-Repeat clones each). Means ± SD are shown. p value by two-tailed t test.

(E) Expression of *MUC1* was tested by RT-qPCR in wild-type and Δ-Repeat MiaPaCa2 cells (n = 5 clones each). Values represent the relative mRNA amount calculated as 2ˆ-ΔCq relative to *C1ORF43* as reference gene. Means ± SD are shown. p value by two-tailed t test.

(F) Immunofluorescence analysis of CDH1 and F-Actin in wild-type, ZEB1-KO, and Δ-Repeat MiaPaCa2 cells. Capan2 cells (left) are representative of fully epithelial cells. The insets represent a magnified detail of the pictures showing only the red (F-Actin) channel. Scale bar, 20 μm.

(G) Quantification of the fraction of cells with a parallel array of stress fibers extending along the entire cell soma as depicted in (D), magnified insets (>200 cells were counted for each group). Means ± SD are shown.

(H) *In vitro* growth of ZEB1-KO (left) and Δ-Repeat clones (right) (n = 3). Means ± SD are shown.

(I) Wild-type, ZEB1-KO, and Δ-Repeat clones (1e7 cells) were xenografted subcutaneously in nude mice (n = 5 for each group). Tumor size was measured every 7±2 days for 5 weeks. Means ± SD are shown.

See also Figure S4 and Movies S1 and S2.

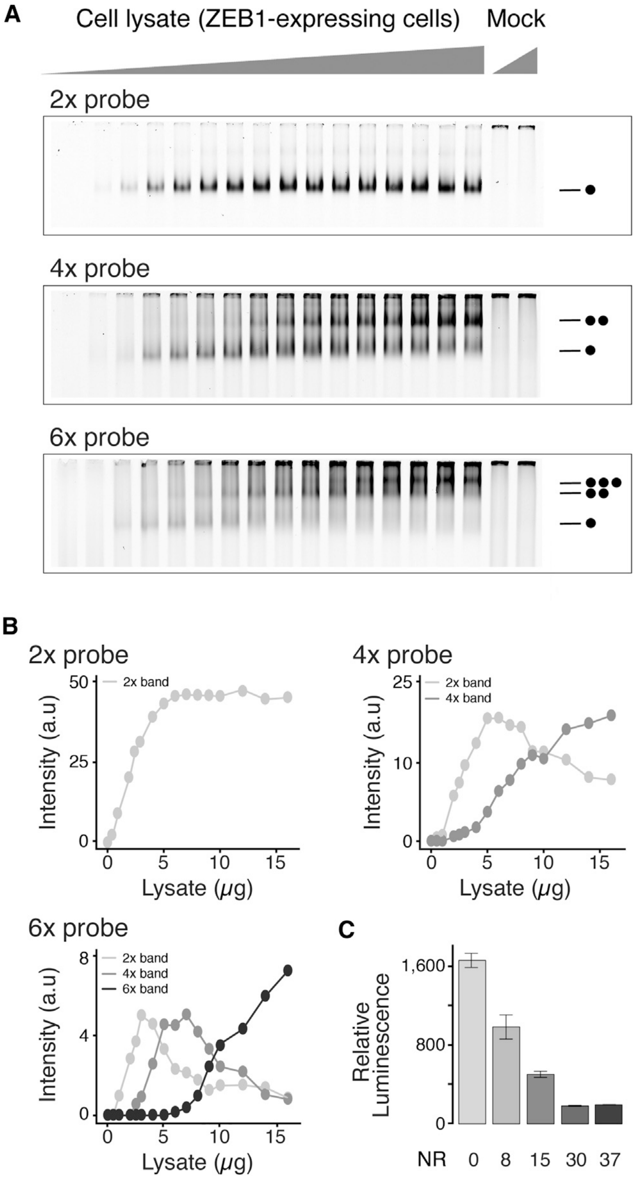


Figure 5. ZEB1 Motif Clustering and Non-Cooperative ZEB1 Binding *In Vitro*

(A) EMSA assays were carried out using increasing amounts of nuclear extracts of HEK-293 cells transfected with a ZEB1 expression vector or a mock control. Infrared dye-labeled probes used included a 2× ZEB1 motif (top), which is recognized by the two Zinc finger clusters of a single ZEB1 molecule; a 4× ZEB1 motif (middle); and a 6× ZEB1 motif (bottom).

(B) Band intensities of the EMSAs shown in (A) were quantified by ImageJ.

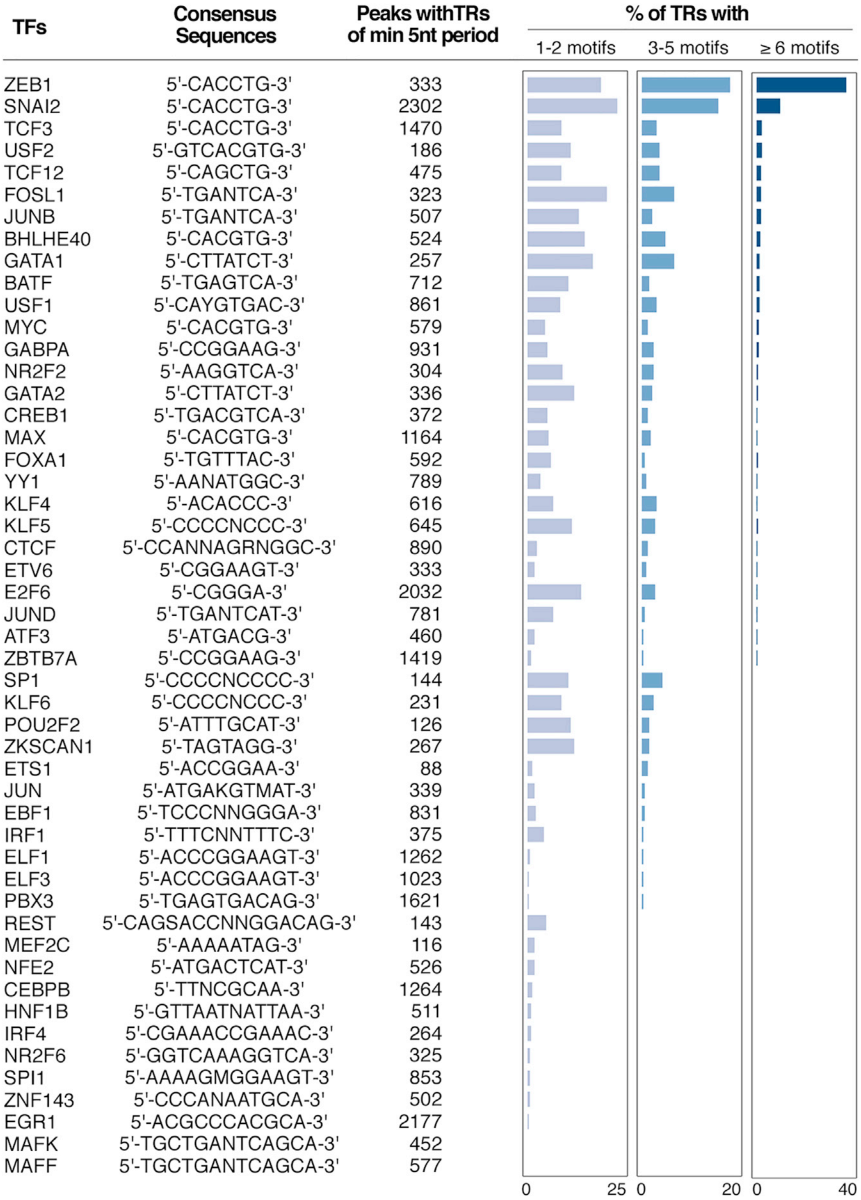
(C) Luciferase reporter assay. ZEB1 clustered motifs in TRs in miR200b/a locus were cloned upstream of the CMV promoter driving the NanoLuc reporter. NR, number of motifs occurring in the cloned genomic region. Data represent the means ± SD from four independent experiments normalized for Firefly luciferase activity.

that the exaptation of this TR was coupled to mechanisms enforcing the maintenance of its structure. The fact that individual TRs greatly differ in their mutation rates and tendency to expand and contract is well documented (Gemayel et al., 2010; Legendre et al., 2007), albeit unclear from a mechanistic point of view. One possible explanation comes from the demonstration that miR-200b and miR-429 are required for ovulation and fertility because of their involvement in a circuit enabling expression of luteinizing hormone in the pituitary gland (Hasuwa et al., 2013). Therefore, significant germline variations in the TR controlling the expression of miR-200b cluster may be counter-selected because of their impact on fertility.

Intriguingly, while miR-200 microRNAs are evolutionary ancient, we could document the presence of a TR bearing ZEB1 motifs in the miR-200b locus only in primates. miRNAs of the miR-200 family are present across all vertebrate classes as well as in invertebrates, in which a single ortholog exists (miR-8) (Trümbach and Prakash, 2015). The apparent primate-specific nature of the TR in the miR-200b locus thus raises an interesting conundrum. Indeed, ZEB1 controls the expression of the miR-200b cluster also in the mouse, but the synthenic region in the mouse genome does not contain a TR with functional ZEB1 motifs. Conversely, two ZEB1 motifs in the promoter of the miR-200b gene are conserved in evolution, although ZEB1 binding to these motifs in human cells was hardly detectable. These data hint at the possibility that, in primates, the TR in the miR-200b cluster has superseded an ancestral mechanism that is still operating in other mammalian lineages. This situation is reminiscent of the previously demonstrated contribution of transposable elements to primate-specific rewiring of regulatory networks (Chuong et al., 2016; Jacques et al., 2013).

On the speculative side, miRNAs of the miR-200 family are involved in neurogenesis both in invertebrates and in vertebrates (Trümbach and Prakash, 2015), and ZEB1 promotes neuronal differentiation through transcriptional repression of polarity and adhesion genes, thus hinting at unexpected analogies between the gene regulatory networks involved in EMT and those involved in neurogenesis (Singh et al., 2016). It is possible that the increased complexity of brain organization in primates imposed the reshaping of the existing regulatory mechanisms to fine-tune the expression of genes and miRNAs impacting neural development. This notion is consistent with the reported high frequency of exapted mobile elements, namely repeats distinct from TRs, in proximity of genes involved in cell adhesion (Lowe et al., 2007). This suggests that cell adhesion genes have been continuously modifying their *cis*-regulatory elements and their expression profiles over the last hundreds of millions of years of evolution.

Overall, our data suggest that gene repression by ZEB1 is favored by homotypic clusters of motifs contained in TRs. Because of the analog nature of ZEB1 recruitment to clustered sites, the level of occupancy and eventually the local concentration of transcriptional repressive activities can be adjusted over a broad range of ZEB1 nuclear concentrations, thus enabling an accurate tuning of repression of target genes. Moreover, homotypic clustering of binding sites, which can reach extreme levels in TRs, provides robustness in gene expression by attenuating the negative consequences of allelic variants that disrupt transcription factor binding (Kilpinen et al., 2013). This scenario likely explains the propensity to exaptation of TRs containing ZEB1 motifs. A similar regulatory logic may underlie exaptation of TRs in other gene regulatory networks.



TFs	Consensus Sequences	Peaks withTRs of min 5nt period
ZEB1	5'-CACCTG-3'	333
SNAI2	5'-CACCTG-3'	2302
TCF3	5'-CACCTG-3'	1470
USF2	5'-GTCACGTG-3'	186
TCF12	5'-CAGCTG-3'	475
FOSL1	5'-TGANTCA-3'	323
JUNB	5'-TGANTCA-3'	507
BHLHE40	5'-CACGTG-3'	524
GATA1	5'-CTTATCT-3'	257
BATF	5'-TGAGTCA-3'	712
USF1	5'-CAYGTGAC-3'	861
MYC	5'-CACGTG-3'	579
GABPA	5'-CCGGAAG-3'	931
NR2F2	5'-AAGGTCA-3'	304
GATA2	5'-CTTATCT-3'	336
CREB1	5'-TGACGTCA-3'	372
MAX	5'-CACGTG-3'	1164
FOXA1	5'-TGTTTAC-3'	592
YY1	5'-AANATGGC-3'	789
KLF4	5'-ACACCC-3'	616
KLF5	5'-CCCCNCCC-3'	645
CTCF	5'-CCANNAGRNGGC-3'	890
ETV6	5'-CGGAAGT-3'	333
E2F6	5'-CGGGA-3'	2032
JUND	5'-TGANTCAT-3'	781
ATF3	5'-ATGACG-3'	460
ZBTB7A	5'-CCGGAAG-3'	1419
SP1	5'-CCCCNCCCC-3'	144
KLF6	5'-CCCCNCCC-3'	231
POU2F2	5'-ATTTGCAT-3'	126
ZKSCAN1	5'-TAGTAGG-3'	267
ETS1	5'-ACCGGAA-3'	88
JUN	5'-ATGAKGTMAT-3'	339
EBF1	5'-TCCCNNGGGA-3'	831
IRF1	5'-TTTCNNTTTC-3'	375
ELF1	5'-ACCCGGAAGT-3'	1262
ELF3	5'-ACCCGGAAGT-3'	1023
PBX3	5'-TGAGTGACAG-3'	1621
REST	5'-CAGSACCNNGGACAG-3'	143
MEF2C	5'-AAAAATAG-3'	116
NFE2	5'-ATGACTCAT-3'	526
CEBPB	5'-TTNCGCAA-3'	1264
HNF1B	5'-GTTAATNATTAA-3'	511
IRF4	5'-CGAAACCGAAAC-3'	264
NR2F6	5'-GGTCAAAGGTCA-3'	325
SPI1	5'-AAAAGMGGAAGT-3'	853
ZNF143	5'-CCCANAATGCA-3'	502
EGR1	5'-ACGCCCACGCA-3'	2177
MAFK	5'-TGCTGANTCAGCA-3'	452
MAFF	5'-TGCTGANTCAGCA-3'	577

Figure 6. Binding of Other Transcription Factors to TRs

ChIP-seq data sets for 49 transcription factors were analyzed for the overlap with TRs containing one or more perfect matches to their cognate motif. Search was restricted to TRs with a period of >5 nt. The three series of histograms on the right show the percentage of transcription-factor-bound TRs containing 1–2 (left), 3–5 (middle), or more than 6 (right) motifs. Histograms are ordered by decreasing percentage of TRs containing >6 motifs.

See also Table S6.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENTS AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Human specimens
  - ○ Mice
  - ○ Cell lines
- METHOD DETAILS
  - ○ ChIP-seq, RNA-seq and miRNA-seq
  - ○ CRISPR/Cas9-mediated genome editing
  - ○ CRISPR interference (CRISPRi) for gene regulation
  - ○ Adhesion, Invasion, Random migration and Morphometric Analysis
  - ○ PCR, RT- and ChIP-qPCR and Western blots
  - ○ Immunofluorescence analysis
  - ○ EMSA (electrophoretic mobility shift assay)
  - ○ Luciferase assay
  - ○ Mouse xenografts
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ ChIP-seq data analysis
  - ○ Annotation and classification of ChIP-seq peaks
  - ○ Heatmap of ZEB1 ChIP-seq enrichment in MiaPaCa2 cell line
  - ○ *De novo* motif discovery
  - ○ Gene ontology analysis
  - ○ Smart-seq2 analysis

  - ○ Gene ontology analysis of genes de-repressed in ZEB1-KO clones
  - ○ Association of de-repressed genes to ZEB1 ChIP-seq peaks
  - ○ Identification of Tandem Repeats in ZEB1 ChIP-seq peaks
  - ○ Evolutionary analysis and genetic variability of the miR-200b/a/miR-429-proximal TR
  - ○ Correlation between ChIP-seq tag density and ZEB1 motif occurrences
  - ○ Construction of a catalog of homotypic clusters of motifs bound by ZEB1 in multiple cell lines
  - ○ Orientation of ZEB1 motifs in homotypic clusters
  - ○ PhyloP scores
  - ○ Analysis of sub-telomeric localization of homotypic clusters of ZEB1 motifs
  - ○ MicroRNA expression analysis in Low-grade and High-grade PDAC cell lines
  - ○ Genome browser tracks
  - ○ Datasets
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes six tables and two videos and can be found with this article online at https://doi.org/10.1016/j.cell.2018.03.081.

ACKNOWLEDGEMENTS

This project was supported by the Italian Association for Research on Cancer (AIRC IG grant 20251 and 11672 to G.N.). G.R.D. and M.M. received fellowships from the Umberto Veronesi Foundation (FUV). We thank B. Amati and S. Monticelli for critical comments on the manuscript; L. Rotta, T. Capra, and S. Bianchi (IEO and IIT Center for Genomic Sciences) for processing of NGS libraries; and I. Costa and F. Casagrande (Cogentech) for help in confocal and time-lapse experiments.

AUTHOR CONTRIBUTIONS

Conceptualization: G.N., C.B., and G.R.D. C.B. analyzed all sequencing data. G.R.D. generated all data and supervised work by G.A., M.M., V.T., and E.P. P.N. contributed to mouse xenografts. A.P. and G.S. contributed to planning, execution, and interpretation of cell biology data. G.N. designed and supervised the project and wrote the paper with contributions from all authors. Funding acquisition: G.N. and G.R.D.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 30, 2017
Revised: December 16, 2017
Accepted: March 29, 2018
Published: April 26, 2018

REFERENCES

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

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.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202–W208.

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.

Bennett, S.T., and Todd, J.A. (1996). Human type 1 diabetes and the insulin gene: principles of mapping polygenes. Annu. Rev. Genet. *30*, 343–370.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

Bracken, C.P., Gregory, P.A., Kolesnikoff, N., Bert, A.G., Wang, J., Shannon, M.F., and Goodall, G.J. (2008). A double-negative feedback loop between ZEB1-SIP1 and the microRNA-200 family regulates epithelial-mesenchymal transition. Cancer Res. *68*, 7846–7854.

Bracken, C.P., Li, X., Wright, J.A., Lawrence, D.M., Pillman, K.A., Salmanidis, M., Anderson, M.A., Dredge, B.K., Gregory, P.A., Tsykin, A., et al. (2014). Genome-wide identification of miR-200 targets reveals a regulatory network controlling cell invasion. EMBO J. *33*, 2040–2056.

Burk, U., Schubert, J., Wellner, U., Schmalhofer, O., Vincan, E., Spaderna, S., and Brabletz, T. (2008). A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells. EMBO Rep. *9*, 582–589.

Chaffer, C.L., Marjanovic, N.D., Lee, T., Bell, G., Kleer, C.G., Reinhardt, F., D'Alessio, A.C., Young, R.A., and Weinberg, R.A. (2013). Poised chromatin at the ZEB1 promoter enables breast cancer cell plasticity and enhances tumorigenicity. Cell *154*, 61–74.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science *351*, 1083–1087.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet. *18*, 71–86.

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

Curina, A., Termanini, A., Barozzi, I., Prosperini, E., Simonatto, M., Polletti, S., Silvola, A., Soldi, M., Austenaa, L., Bonaldi, T., et al. (2017). High constitutive activity of a broad panel of housekeeping and tissue-specific *cis*-regulatory elements depends on a subset of ETS proteins. Genes Dev. *31*, 399–412.

De Craene, B., and Berx, G. (2013). Regulatory networks defining EMT during cancer initiation and progression. Nat. Rev. Cancer *13*, 97–110.

Diaferia, G.R., Jimenez-Caliani, A.J., Ranjitkar, P., Yang, W., Hardiman, G., Rhodes, C.J., Crisa, L., and Cirulli, V. (2013). β1 integrin is a crucial regulator of pancreatic β-cell expansion. Development *140*, 3360–3372.

Diaferia, G.R., Balestrieri, C., Prosperini, E., Nicoli, P., Spaggiari, P., Zerbi, A., and Natoli, G. (2016). Dissection of transcriptional and cis-regulatory control of differentiation in human pancreatic cancer. EMBO J. *35*, 595–617.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet. *5*, 435–445.

Furusawa, T., Moribe, H., Kondoh, H., and Higashi, Y. (1999). Identification of CtBP1 and CtBP2 as corepressors of zinc finger-homeodomain factor deltaEF1. Mol. Cell. Biol. *19*, 8581–8590.

Gemayel, R., Vinces, M.D., Legendre, M., and Verstrepen, K.J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu. Rev. Genet. *44*, 445–477.

Gertz, J., Savic, D., Varley, K.E., Partridge, E.C., et al. (2013). Distinct properties of cell-type-specific and shared transcription factor binding sites. Mol. Cell *10*, 25–36.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biol. *8*, R24.

Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M.J., Price, A.L., Pritchard, J.K., Sharp, A.J., and Erlich, Y. (2016).

Abundant contribution of short tandem repeats to gene expression variation in humans. Nat. Genet. *48*, 22–29.

Hasuwa, H., Ueda, J., Ikawa, M., and Okabe, M. (2013). miR-200b and miR-429 function in mouse ovulation and are essential for female fertility. Science *341*, 71–73.

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.

Huang, D.W., Sherman, B.T., Tan, Q., Collins, J.R., Alvord, W.G., Roayaei, J., Stephens, R., Baseler, M.W., Lane, H.C., and Lempicki, R.A. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. *8*, R183.

Jacques, P.-É.É., Jeyakani, J., and Bourque, G. (2013). The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet. *9*, e1003504.

Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics *26*, 2204–2207.

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.

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

Lee, B., Villarreal-Ponce, A., Fallahi, M., Ovadia, J., Sun, P., Yu, Q.-C., Ito, S., Sinha, S., Nie, Q., and Dai, X. (2014). Transcriptional mechanisms link epithelial plasticity to adhesion and differentiation of epidermal progenitor cells. Dev. Cell *29*, 47–58.

Legendre, M., Pochet, N., Pak, T., and Verstrepen, K.J. (2007). Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res. *17*, 1787–1796.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liu, T. (2014). Use model-based analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells. Methods Mol. Biol. *1150*, 81–95.

Loo, J.M., Scherl, A., Nguyen, A., Man, F.Y., Weinberg, E., Zeng, Z., Saltz, L., Paty, P.B., and Tavazoie, S.F. (2015). Extracellular metabolic energetics can promote cancer progression. Cell *160*, 393–406.

Lowe, C.B., Bejerano, G., and Haussler, D. (2007). Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc. Natl. Acad. Sci. USA *104*, 8005–8010.

Mani, S.A., Guo, W., Liao, M.J., Eaton, E.N., Ayyanan, A., Zhou, A.Y., Brooks, M., Reinhard, F., Zhang, C.C., Shipitsin, M., et al. (2008). The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell *133*, 704–715.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12.

Martin, P., Makepeace, K., Hill, S.A., Hood, D.W., and Moxon, E.R. (2005). Microsatellite instability regulates transcription factor binding and gene expression. Proc. Natl. Acad. Sci. USA *102*, 3800–3804.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol *28*, 495–501.

Mistry, D.S., Chen, Y., Wang, Y., Zhang, K., and Sen, G.L. (2014). SNAI2 controls the undifferentiated state of human epidermal progenitor cells. Stem Cells *32*, 3209–3218.

Montagner, S., Leoni, C., Emming, S., Della Chiara, G., Balestrieri, C., Barozzi, I., Piccolo, V., Togher, S., Ko, M., Rao, A., Natoli, G., and Monticelli, S. (2016). TET2 regulates mast cell differentiation and proliferation through catalytic and non-catalytic activities. Cell Rep *15*, 1744.

Moreno-Mateos, M.A., Vejnar, C.E., Beaudoin, J.D., Fernandez, J.P., Mis, E.K., Khokha, M.K., and Giraldez, A.J. (2015). CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. Nat. Methods *12*, 982–988.

Nieto, M.A., Huang, R.Y., Jackson, R.A., and Thiery, J.P. (2016). Emt: 2016. Cell *166*, 21–45.

Park, S.-M.M., Gaur, A.B., Lengyel, E., and Peter, M.E. (2008). The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. Genes Dev. *22*, 894–907.

Pattyn, F., Speleman, F., De Paepe, A., and Vandesompele, J. (2003). RTPrimerDB: the real-time PCR primer and proble database. Nucleic Acids Res. *31*, 122–123.

Pham, H., Kearns, N.A., and Maehr, R. (2016). Transcriptional regulation with CRISPR/Cas9 effectors in mammalian cells. Methods Mol. Biol. *1358*, 43–57.

Picelli, S., Björklund, A.K., Reinius, B., Sagasser, S., Winberg, G., and Sandberg, R. (2014a). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. *24*, 2033–2040.

Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014b). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.

Postigo, A.A., and Dean, D.C. (1999). ZEB represses transcription through interaction with the corepressor CtBP. Proc. Natl. Acad. Sci. USA *96*, 6683–6688.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ran, F.A., Hsu, P.D., Lin, C.Y., Gootenberg, J.S., Konermann, S., Trevino, A.E., Scott, D.A., Inoue, A., Matoba, S., Zhang, Y., and Zhang, F. (2013). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. Cell *154*, 1380–1389.

Remacle, J.E., Kraft, H., Lerchner, W., Wuytens, G., Collart, C., Verschueren, K., Smith, J.C., and Huylebroeck, D. (1999). New mode of DNA binding of multi-zinc finger transcription factors: deltaEF1 family members bind with two hands to two target sites. EMBO J. *18*, 5073–5084.

Reynolds, N., O'Shaughnessy, A., and Hendrich, B. (2013). Transcriptional repressors: multifaceted regulators of gene expression. Development *140*, 505–512.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Sakuma, T., Nishikawa, A., Kume, S., Chayama, K., and Yamamoto, T. (2014). Multiplex genome engineering in human cells using all-in-one CRISPR/Cas9 vector system. Sci. Rep. *4*, 5400.

Sánchez-Tilló, E., Liu, Y., de Barrios, O., Siles, L., Fanlo, L., Cuatrecasas, M., Darling, D.S., Dean, D.C., Castells, A., and Postigo, A. (2012). EMT-activating transcription factors in cancer: beyond EMT and tumor invasiveness. Cell. Mol. Life Sci. *69*, 3429–3456.

Saunders, L.R., and McClay, D.R. (2014). Sub-circuits of a gene regulatory network control a developmental epithelial-mesenchymal transition. Development *141*, 1503–1513.

Shi, Y., Sawada, J., Sui, G., Affar, B., Whetstine, J.R., Lan, F., Ogawa, H., Luke, M.P., Nakatani, Y., and Shi, Y. (2003). Coordinated histone modifications mediated by a CtBP co-repressor complex. Nature *422*, 735–738.

Singh, S., Howell, D., Trivedi, N., Kessler, K., Ong, T., Rosmaninho, P., Raposo, A.A., Robinson, G., Roussel, M.F., Castro, D.S., and Solecki, D.J. (2016). Zeb1 controls neuron differentiation and germinal zone exit by a mesenchymal-epithelial-like transition. eLife *5*, 5.

Sipos, B., Möser, S., Kalthoff, H., Török, V., Löhr, M., and Klöppel, G. (2003). A comprehensive characterization of pancreatic ductal carcinoma cell lines: towards the establishment of an in vitro research platform. Virchows Arch. *442*, 444–452.

Spandidos, A., Wang, X., Wang, H., and Seed, B. (2010). PrimerBank: a resource of human and mouse PCR primer pairs for gene expression detection and quantification. Nucleic Acids Res. *38* (Database issue), D792–D799.

Stern, A., Brown, M., Nickel, P., and Meyer, T.F. (1986). Opacity genes in Neisseria gonorrhoeae: control of phase and antigenic variation. Cell *47*, 61–71.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H., et al.; 1000 Genomes Project Consortium (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE *6*, e21800.

Tanenbaum, M.E., Gilbert, L.A., Qi, L.S., Weissman, J.S., and Vale, R.D. (2014). A protein-tagging system for signal amplification in gene expression and fluorescence imaging. Cell *159*, 635–646.

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.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Trümbach, D., and Prakash, N. (2015). The conserved miR-8/miR-200 microRNA family and their role in invertebrate and vertebrate neurogenesis. Cell Tissue Res. *359*, 161–177.

Verstrepen, K.J., Jansen, A., Lewitter, F., and Fink, G.R. (2005). Intragenic tandem repeats generate functional variability. Nat. Genet. *37*, 986–990.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). Enhancer evolution across 20 mammalian species. Cell *160*, 554–566.

Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M., and Verstrepen, K.J. (2009). Unstable tandem repeats in promoters confer transcriptional evolvability. Science *324*, 1213–1216.

Weiser, J.N., Love, J.M., and Moxon, E.R. (1989). The molecular mechanism of phase variation of H. influenzae lipopolysaccharide. Cell *59*, 657–665.

Wellner, U., Schubert, J., Burk, U.C., Schmalhofer, O., Zhu, F., Sonntag, A., Waldvogel, B., Vannier, C., Darling, D., zur Hausen, A., et al. (2009). The EMT-activator ZEB1 promotes tumorigenicity by repressing stemness-inhibiting microRNAs. Nat. Cell Biol. *11*, 1487–1495.

Yogev, S., and Shen, K. (2014). Cellular and molecular mechanisms of synaptic specificity. Annu. Rev. Cell Dev. Biol. *30*, 417–437.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Please cite this article in press as: Balestrieri et al., Co-optation of Tandem DNA Repeats for the Maintenance of Mesenchymal Identity, Cell (2018), https://doi.org/10.1016/j.cell.2018.03.081

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit polyclonal anti-ZEB1 (H-102)	Santa Cruz Biotech	Cat# sc-25388, Lot D1511; RRID: AB_2217979
Mouse monoclonal anti-Tubulin (clone DM1A)	Santa Cruz Biotech	Cat#sc-32293; RRID: AB_628412
Rabbit monoclonal anti-E-Cadherin (CDH1) (clone 24E10)	Cell Signaling Technology	Cat# 3195; RRID: AB_2291471
Biological Samples		
Human cancer cell pellets	This paper; internal stock	N/A
Healthy human buccal swabs	This paper; internal stock	N/A
Chemicals, Peptides, and Recombinant Proteins		
Phalloidin-TRITC	Sigma-Aldrich	Cat# P1951; RRID: AB_2315148
Tn5	This paper (adapted from Picelli et al., 2014a)	N/A
Laminin	Roche	Cat# 11243217001
Fibronectin	Roche	Cat# 11080938001
Collagen	Sigma-Aldrich	Cat# C4243
Matrigel	BD Biosciences	Cat# 356231
Poly(dI:dC)	Sigma-Aldrich	Cat# P4929
Critical Commercial Assays		
TruSeq Small RNA Sample Prep Kit	Illumina	Cat# RS-200-0012, Set A
Nano-Glo Dual Luciferase reporter assay kit	Promega	Cat# N1610
T7E1 assay	New England Biolabs	Cat# M0302
Transwell Permeable supports	Costar	Cat# 3422
Deposited Data		
ChIP-Seq, miRNA-Seq and RNA-Seq data	This paper	GEO: GSE88738
Input ChIP-seq data in MiaPaCa2	Diaferia et al., 2016	GEO: GSM1574272
ZEB1 ChIP-seq data in PANC-1	Diaferia et al., 2016	GEO: GSM1574278
ZEB1 ChIP-seq data in GM12878	Gertz et al., 2013	GEO: GSM803411
Input ChIP-seq data in GM12878	Gertz et al., 2013; ENCODE	GEO: GSM803413; GEO: GSM733742
CTCF ChIA-PET data in GM12878	Tang et al., 2015	GEO: GSM1872886
Simple Tandem Repeats catalog	Benson, 1999	https://genome.ucsc.edu
PhyloP in 100 vertebrates	Pollard et al., 2010	https://genome.ucsc.edu
1000 Genomes Project	International Genome Sample Resource (IGSR).	http://www.internationalgenome.org/data
Structural Variants (study ID estd219)	Database of Genomic Variants Archive (DGVa)	https://www.ebi.ac.uk/dgva/
TF ChIP-seq data in CFPAC1	Diaferia et al., 2016	See Table S6
TF ChIP-seq data in GM12878 and	ENCODE	See Table S6
SNAI2 ChIP-seq data in Human keratinocytes	Mistry et al., 2014	See Table S6
Experimental Models: Cell Lines		
MiaPaCa-2	ATCC	CRL-1420; RRID: CVCL_0428
PANC1	ATCC	CRL-1469; RRID: CVCL_0480
RKO	ATCC	CRL-2577; RRID: CVCL_0504
U-118MG	ATCC	HTB-15; RRID: CVCL_0633

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
UM-UC-3	ATCC	CRL-1749; RRID: CVCL_1783
U2OS	ATCC	HTB-96; RRID: CVCL_0042
HS578T	ATCC	HTB-126; RRID: CVCL_0332
SHSY5Y	ATCC	CRL-2266; RRID: CVCL_0019
PT45P1	Paola Allavena, Humanitas Research Hospital, Milan	RRID: CVCL_8408
Capan2	DSMZ	ACC-245; RRID: CVCL_0026
Experimental Models: Organisms/Strains		
Mouse: CD-1 FONX1 NU/NU	Charles River	Strain code 086
Oligonucleotides		
Primers for CRISPR/Cas9 genome editing	See Table S6	N/A
Primers for CRISPRi	See Table S6	N/A
Primers for T7E1 assay	See Table S6	N/A
Primers for genotyping	See Table S6	N/A
Primers for qPCR	See Table S6	N/A
LNA primer sets	See Table S6	N/A
EMSA Probes	See Table S6	N/A
Primers for cloning in NanoLuc vector	See Table S6	N/A
Recombinant DNA		
pSpCas9(BB)-2A-GFP	Ran et al., 2013	Addgene cat#48138
multiplex CRISPR/Cas9 assembly system kit	Sakuma et al., 2014	Addgene cat#1000000055
SunTag plasmid: pHRdSV40-scFv-GCN4-sfGFP-VP64-GB1-NLS	Tanenbaum et al., 2014	Addgene cat#60904
pLenti-Guide Hygro	Pham et al., 2016	Addgene cat#62205
pLenti-Guide Hygro SV40-scFv-GCN4-sfGFP-KRAB-GB1-NLS	This Paper	N/A
SunTag plasmid: pHRdSV40-NLS-dCas9-24xGCN4_v4-NLS-P2A-BFP-dWPRE	Tanenbaum et al., 2014	Addgene cat#60910
pScalp_Puro		Addgene cat#99636
pScalps-dCas9-24xGCN4	This Paper	N/A
pCDNA3.1 CMV-NanoLuc	Bruno Amati Lab, IIT	N/A
pGL4.53[luc2/PGK] vector	Promega	Cat# E5011
Software and Algorithms		
BEDTools v2.19.1	Quinlan and Hall, 2010	http://bedtools.readthedocs.io/en/latest/
Bowtie2 v2.2.6	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
bedGraphToBigWig	Kent et al., 2010	http://hgdownload.soe.ucsc.edu/admin/exe/
Cutadapt v1.7.1	Martin, 2011	https://cutadapt.readthedocs.io/en/stable/index.html
DAVID tool v6.8	Huang et al., 2007	https://david.ncifcrf.gov
EdgeR	Robinson et al., 2010	http://bioconductor.org/packages/release/bioc/html/edgeR.html
GREAT Tool	McLean et al., 2010	http://bejerano.stanford.edu/great/public/html/
GSEA	Broad Institute	http://software.broadinstitute.org/gsea/index.jsp
HOMER v4.8	Heinz et al., 2010	http://homer.ucsd.edu/homer/
HTSeq v0.6.1	Anders et al., 2015	https://pypi.python.org/pypi/HTSeq/0.6.1
IGV	Broad Institute	http://software.broadinstitute.org/software/igv/
MACS2 v2.1.0.20150731	Liu, 2014	https://github.com/taoliu/MACS
MEME v4.10.1	Bailey et al., 2009	http://meme-suite.org/tools/meme

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
R v3.2.2	The R Project	https://www.r-project.org/
REVIGO	Supek et al., 2011	http://revigo.irb.hr/
SAMtools v1.2	Li et al., 2009	http://www.htslib.org
TomTom	Gupta et al., 2007	http://meme-suite.org/tools/tomtom
TopHat v2.1.0	Trapnell et al., 2012	http://ccb.jhu.edu/software/tophat/index.shtml
UCSC Genome Browser	UCSC	http://genome.ucsc.edu/
CRISPR design tool	MIT	http://crispr.mit.edu/
CRISPRscan	Giraldez lab, YALE	http://crisprscan.org
Chemotaxis tool	NIH ImageJ plugin	http://ibidi.com/software/chemotaxis_and_migration_tool
Primerbank	Spandidos et al., 2010	http://pga.mgh.harvard.edu/primerbank
RTPrimerDB	Pattyn et al., 2003	http://rtprimerdb.org

CONTACT FOR REAGENTS AND RESOURCE SHARING

Further information and requests for resources and reagents may be directed to and will be fulfilled by the Lead Contact, Gioacchino Natoli (gioacchino.natoli@hunimed.eu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human specimens

Buccal swabs were collected with informed consent and under the approval of the Humanitas IRCCS ethical committee by 50 healthy volunteers (19 males and 31 females) of Caucasian origin ranging from 24 to 54 years of age.

Mice

CD1-Nude mice were purchased from Charles River. Experiments involving animals have been done in accordance with the Italian Laws (D.lgs. 26/2014), which enforces Dir. 2010/63/EU (‘‘Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes’’). All animal procedures were approved by the OPBA (Organismo per il Benessere e Protezione Animale) of the Cogentech animal facility at the IFOM-IEO Campus, Milan. The project has been approved by the Italian Ministry of Health (1197/2016).

Cell lines

The following human cell lines were used: MiaPaCa-2 (primary high-grade pancreatic ductal carcinoma, ATCC CRL-1420), PANC1 (primary high-grade pancreatic ductal carcinoma, ATCC CRL-1469), PT45P1 (primary high-grade pancreatic ductal carcinoma obtained from Paola Allavena, Humanitas Research Hospital, Milan), Capan2 (primary low-grade pancreatic ductal adenocarcinoma, DSMZ ACC-245), RKO (poorly differentiated colon carcinoma, ATCC CRL-2577), U118MG (high-grade glioblastoma ATCC HTB-15), UMUC3 (transitional cell carcinoma of urinary bladder ATCC CRL-1749), U2OS (osteosarcoma ATCC HTB-96), HS578T (breast carcinoma ATCC HTB-126) and SHSY5Y (neuroblastoma from metastatic bone tumor ATCC CRL-2266). MiaPaCa-2, PANC-1, U118MG, U2OS and HS578T cells were maintained in DMEM (Lonza), PT45P1 and Capan2 in RPMI-1640 (Lonza), while RKO, UMUC3 and SHSY5Y cells were maintained in MEM (Sigma) with Earle’s Salts, 1mM Sodium Pyruvate (NaP) and 0.1mM Non-Essential Amino Acids (NEAA)(Sigma). Media were all supplemented with 10% FBS (Hyclone), 2 mM L-glutamine and 1% Pen/Strep with the addition of 0.01mg/ml insulin only for HS578T cells. All cell lines were authenticated by the Tissue Culture Facility of IEO using the GenePrint10 System (Promega) for the amplification of 10 short tandem repeat-containing loci, followed by Sanger sequencing.

METHOD DETAILS

ChIP-seq, RNA-seq and miRNA-seq

ChIP-seq was carried out using previously described protocols (Curina et al., 2017) on an Illumina HiSeq2000 platform. 20–40 x 10E6 fixed cells were lysed to prepare nuclear extracts. After chromatin shearing by sonication, lysates were incubated overnight at 4°C with protein G Dynabeads (Invitrogen) coupled with 10 μg of anti-ZEB1 antibody. After immunoprecipitation, beads were recovered using a magnet and washed; chromatin was eluted and cross-links reverted overnight at 65°C. DNA was either purified with QiaQuick columns (QIAGEN) or solid-phase reversible immobilization (SPRI) beads (Agencourt AMPure XP, Beckman Coulter), and then quantified with QuantiFluor (Promega). DNA libraries were prepared for HiSeq2000 sequencing as previously described (Curina et al., 2017). RNA-seq was carried out using the SMART-seq2 protocol (Picelli et al., 2014b) with minor modifications. Briefly, 2 x 10E3 cells were lysed and the poly-A containing mRNA molecules were copied into first strand cDNA by reverse transcription and template-switching using oligo(dT) primers and an LNA-containing template-switching oligo (TSO). The resulting cDNA was pre-amplified, purified and tagmented with Tn5 transposase produced in-house using a described protocol (Picelli et al., 2014a). cDNA fragments generated after tagmentation were gap-repaired, enriched by PCR and purified to create the final cDNA library.

miRNA-seq was carried out using the TruSeq Small RNA Sample Prep Kit (Illumina #RS-200-0012, Set A). Briefly, 1 μg of total RNA (QIAzol Lysis Reagent, QIAGEN) was used to produce miRNA sequencing libraries following manufacturer's instructions. The cDNAs generated by mature microRNA were separated by size on acrylamide gels and gel purified before proceeding to sequencing.

CRISPR/Cas9-mediated genome editing

Single-guide sequences specific to ZEB1 (exon 7) were designed using the CRISPR design tool (http://crispr.mit.edu) or CRISPRScan (Moreno-Mateos et al., 2015) and cloned into pSpCas9(BB)-2A-GFP (Addgene cat. #48138). The clones used in the manuscript were generated using the following sgRNA: 5'- GTTCTTGGTCGCCCATTCAC-3'. After transfection of MiaPaCa-2 cells, FACS sorted GFP$^+$ single cells were seeded in 96-well plates. Edited clones were screened using the T7E1 assay (New England Biolabs, M0302) with the following primers: ZEB1_F: 5'-AGTTCTGTCACAAGCATGCA-3', ZEB1_R: 5'- CTGAGGAGAACTGGTTGCCT-3'. Positive clones were validated by Sanger sequencing and western blot analysis. For the deletion of the TR in the miR-200b/a locus, we used the D10A mutant nickase version of Cas9 (Cas9n) with a pair of offset sgRNAs complementary to opposite DNA strands (Ran et al., 2013). Using the CRISPR design tool mentioned above we designed pairs of sgRNAs for two sites flanking the ZEB1-bound region upstream miR-200b/a locus on chromosome 1 (chr1:1,058,873-1,059,613) and cloned each pair in vectors provided in the multiplex CRISPR/Cas9 assembly system kit (Addgene cat. #1000000055)(Sakuma et al., 2014) following manufacturer's instructions. The sequence selected were the following: pair 1, 5'-TGAGGCACGGGGGCCGTCGG-3', 5'-TGCTGGCGACTCAGCGAGGT-3'; pair 2, 5'-TCCAGACCCCAAGAACCCCT-3', 5'-GAGACCCCGGAGCTGATGCG-3'. After transfection of MiaPaca-2 cells with equimolar amount of the two plasmids containing sgRNA pairs and Cas9n, single cells were seeded in 96-well plates by dilution. Clones were genotyped by PCR with the following primers flanking the edited region (chr1:1057946-1060104):

Δ-Repeats_F, 5'-GGGGATCTTCGGAGCTGATG-3' and

Δ-Repeats_R, 5'-CCATGGCCTTCCCTATCCTC-3'.

Clones with the correct deletion of the intervening region between the edited sites were further validated by Sanger sequencing.

For the deletion of the TR in PANC1 and PT45P1 cells we used Cas9 with a pair of sgRNAs (5'- TGCTGGCGACTCAGCGAGGT-3' and 5'- TCCAGACCCCAAGAACCCCT-3') flanking the ZEB1-bound region and clone them in a modified version of the vectors provided in the multiplex CRISPR/Cas9 assembly system kit in which we cloned the Puromycin resistance gene. After transfection and puromycin selection (1 μg/ml for 72h), cells were expanded for few more days and genotyped by PCR with the primers mentioned above.

For the deletion of the other TRs in MiaPaCa-2 cells we used the same strategy described above with the following pair of sgRNAs (for CRISPR/Cas9 assembly) and primers (for PCR genotyping): chr1:3300179-3300851, sgRNAs 5'- GGGGCAGGGGTGATGGATAA-3' and 5'- GGGACACCGTCTCTCCACAG-3', primers Δ-Repeats_2F 5'-ACAGGAAAGAGACTCGAGGC -3' and Δ-Repeats_2R 5'- GAAGGTGAAGGTGTTGCTGG -3'; chr1:3565508-3566731, sgRNAs 5'- GGAGGTCTTCGTTCCATGGG-3' and 5'- GGCCGGGGACGCTGGCACCG-3', primers Δ-Repeats_3F 5'- TAGGCACGTGTACAGCTGAA -3' and Δ-Repeats_3R 5'- AAGGTTCTCGGGTCGGAAAT -3'; chr19:11319348-11322795, sgRNAs 5'- GGGGGTACAGGTGAGGAGAT-3' and 5'- GGGACGGGTGATGGAGATGG-3', primers Δ-Repeats_4F 5'- AAGCAGATCTGATGGCACCT -3' and Δ-Repeats_4R 5'- CCTGCGACGTTCACTCAAAA -3'; chr9:138454738-138455764, sgRNAs 5'- GGCTGGGGCATGAGGGAGTG-3' and 5'- GGAGAAGTCACGTACGGGTG-3', primers Δ-Repeats_5F 5'- TCCCGTGGTGACCTGATTTT -3' and Δ-Repeats_5R 5'- AGATGAAGCCTCCGTCTCAG -3'; chr20:62446028-62447956, sgRNAs 5'- GGGCCTCCTGTTGAGGGGTG-3' and 5'- GGGCGGGCCCTCACCCAAT-3', primers Δ-Repeats_6F 5'- AAACAGTGGGAGAGAGTGGG -3' and Δ-Repeats_6R 5'- TAGCCTCCTTCTCCCAGACT -3'; chr1:24671145-24672019, sgRNAs 5'- GGGTGCACAGGTGTGGAGGG-3' and 5'- GGGCATGGTACAGAGGGCAC-3', primers Δ-Repeats_7F 5'- GGCGTGTCATCTCTAGAGCT -3' and Δ-Repeats_7R 5'- AACCCTGCTTGCCTAACTCT -3'.

CRISPR interference (CRISPRi) for gene regulation

To restore repression in the miR200b/a locus, we took advantage of a modified version of the SunTag system (Tanenbaum et al., 2014). We replaced the VP64 activator (from Addgene vector #60904) with the KRAB repressor gene fused in frame with GFP and the antibody that binds to the GCN4 epitope; we cloned the SV40-scFv-GCN4-sfGFP-KRAB-GB1-NLS cassette into the pLenti-Guide Hygro (Addgene #62205) to generate a KRAB expressing vector (pLenti-Guide KRAB) in which we cloned the sgRNA designed in a region adjacent to the deleted TR (5'-GCCACCCACCCAGCCCGGCG-3') under the U6 promoter. We moved the NLS-dCas9-24xGCN4_v4-NLS-P2A-BFP-dWPRE cassette from the Addgene vector #60910 into the pScalps vector (Montagner et al., 2016; kind gift from Silvia Monticelli, IRB, Switzerland) expressing the puromycin resistance gene under the CYPA promoter to generate pScalps-dCas9-24xGCN4. The two vectors (pLenti-Guide KRAB and pScalps-dCas9-24xGCN4) were used to produce lentiviral particles to transduce Δ-Repeat and control clones. The KRAB expressing vector containing no sgRNA was used on Δ-Repeat and wild type clones as negative controls, while KRAB expressing vector containing sgRNA complementary to a genomic region flanking the deleted TR upstream of the miR-200b locus was used in a duplicate set of Δ-Repeat clones. 36h after infection cells were selected with puromycin (1μg/ml) and hygromycin (0.5mg/ml) for 5 days before further manipulations.

Adhesion, Invasion, Random migration and Morphometric Analysis

For adhesion assay, MiaPaCa-2 control and genome-edited cells were plated onto different extracellular matrices as previously described (Diaferia et al., 2013). After 1 hour, cells were fixed in 4% PFA, stained with 1% Tolouidine Blue and counted under the microscope. Adhesion to specific matrices was normalized for the number of cells attaching to the non-integrin-dependent synthetic polymer poly-lysine. Invasion assay was performed with standard Boyden chamber technique using 8μm-pore polycarbonate membranes (Costar) coated with 0.2mg/ml of Matrigel (BD biosciences). 1 x 10E5 serum-starved cells were seeded on top of the matrigel and allowed to migrate and invade for 24 hours prior to further processing and counting as previously described. Time-lapse imaging of cell migration was performed on a Leica DMI 6000 B microscope equipped with an incubator chamber (OKOlab) maintained at 37C in a 5% CO2 atmosphere. Movies were acquired in phase contrast with Andor iXon DU-885 device camera using the objective HC PL Fluotar 10X/0,30. Leica LasX was the software used for both system and camera control. Tracking of cells was performed using the “Manual Tracking” plug-in distributed by ImageJ software. 5 random fields were acquired for each sample and 5 cells for each field were manually tracked resulting in 75 measurements for each group (*n=3*). Trajectories plots and accumulated distance bar-graph (μm run by tracked cells during the assay) were obtained by the “Chemotaxis” tool plugin for ImageJ (http://ibidi.com/software/chemotaxis_and_migration_tool). Morphometric measurements of cell shape from live cultures were performed with Image J by manually delineating the edges of randomly selected cells (*n=3* clones each group, *15* cells analyzed each clone for a total of 45 measurements per group) and recording the circularity value.

PCR, RT- and ChIP-qPCR and Western blots

Analysis of the TR in miR200b/a locus was performed by traditional PCR (primer F: 5’- GAGAAGCCCAGGAGCAAGTA-3’; primer R: 5’- AGGGTGGTGGTTTCTCAGAG-3’) on genomic DNA extracted from buccal swabs of 50 normal individuals and from cell pellets obtained from 78 publicly available human cancer cell lines. Ten representative PCR-amplified samples for each group (normal and tumor) were cloned into TA vectors and subjected to Sanger sequencing. Total RNA containing small RNA species was extracted from 0.5-1 x 10ˆ6 cells using Maxwell® 16 miRNA Tissue kit (Promega). For miRNA analysis, 0.5 μg of RNA was polyadenylated with *E. coli* Poly(A) Polymerase (New England Biolabs, M0276) and retrotranscribed with ImProm-II Reverse Transcription System (Promega) and oligo(dT) following manual instructions. qPCR reactions were assembled with Fast SYBR Green Master Mix using validated LNA primer sets (EXIQON): hsa-miR-200a-3p (no. 204707), hsa-miR-200b-3p (no. 206071), hsa-miR-200c-3p (no. 204482), hsa-miR-429 (no. 205901), hsa-miR-141-3p (no. 204504), hsa-miR-103a-3p (no. 204063). For mRNA analysis, 0.5 μg of RNA was reverse-transcribed with ImProm-II Reverse Transcription System (Promega) and random priming following manual instructions. qPCR reactions were assembled with Fast SYBR Green Master Mix using primer sets selected from the validated database suggested in the MIQE guidelines (PrimerBank, http://pga.mgh.harvard.edu/primerbank; ; RTPrimerDB, http://www.rtprimerdb.org) or manually designed with Primer3 (primer3.ut.ee): *C1ORF159* (ID 192447435c2), *TTLL10* (ID 194239677c1), *SDF4* (ID 170763489c1), *B3GALT6* (ID 116268096c1), *UBE2J2* (ID 37577129c2), *PRDM16* (ID 289547570c1), *ARHGEF16* (ID 163792207c1), *MEGF6* (ID 110347456c2), *WRAP73* (ID 224586776c2), *MRPS2* (ID 187167256c2), *SLC2A4RG* (ID 39777592c1), *TPD52L2* (ID 40805865b2), *UCKL1* (ID 301129206c2), *GRHL3* (ID 303324554c2), *NIPAL3* (ID 210032474c1), *RCAN3* (ID 354623075c1), *DOCK6* (ID 364023823c3), *TMEM205* (ID 224028277c1), *CCDC151* (ID 117553612c2), *MUC1* (ID 3226), *LOC254099* (5’ – TTGTTCAGGCACATGGTCAC - 3’; 5’ – GGACCTGGCATTTTCCGAAG - 3’) and *C1ORF43* (5’ – GGATGAAAGCTCTGGATGCC - 3’; 5’ – GCTTTGCGTACACCCTTGAA - 3’). The reactions were run on 7500HT ABI Prism machine (Applied Biosystems) and data analysed with SDS v2.0.6 software (Applied Biosystems) using hsa-miR-103a-3p or *C1ORF43* expression as reference (based on the analysis of data from the Human BodyMap, HBM 2.0 Project). ChIP-qPCR was performed with SYBR green Master mix as previously described using the following primers: Chr1-TR (5’-GAGAAGCCCAGGAGCAAGTA-3’; 5’-TGGGTGGGGTGTGCTCAG-3’) and Negative Ctrls (5’-AATGTTGGGCCTTGAAACAG-3’; 5’-CCAGTGTGGTCCAAAGAGGT-3’)**.** The complete list of oligonucleotides used in this study is shown in **Table S6**.

MiaPaCa-2 cells were lysed in RIPA buffer containing protease inhibitors, 1 mM PMSF, 1 mM EDTA, and 1 mM NaF, and 50 μg of clarified cell extracts was resolved on SDS–polyacrylamide gel, blotted onto nitrocellulose membranes, and probed with anti-ZEB1 (0.5 μg/ml) or anti-a-tubulin (0.2 μg/ml).

Immunofluorescence analysis

Three-color immunofluorescence and confocal analysis were performed on Capan-2 cells and MiaPaca-2 edited and control clones grown onto glass coverslips, as previously described (Diaferia et al., 2013). Briefly, PFA fixed cells where permeabilized with 0.1% Triton X-100, blocked and incubated with anti-CDH1 antibody (1:200). Alexa488 labeled anti-rabbit IgGs secondary antibody (Jackson ImmunoResearch) and Phalloidin-TRITC (1:1000) were used to detect E-Cadherin and F-Actin, respectively. Nuclei were counterstained with DAPI and samples mounted with Mowiol aqueous mounting medium supplemented with DABCO anti-fading agent (Sigma). Confocal microscopy was performed on a Leica TCS SP5 laser confocal scanner mounted on a Leica DMI 6000B inverted microscope equipped with motorized stage, HCX PL APO 63X/1.4NA oil immersion objective. Violet (405nm laser diode), blue (488nm argon laser), and yellow (561nm laser diode) laser lines have been used for excitation. Software used for all acquisitions was Leica LAS AF. Four random confocal images for each clone (n=3 clones for each group) were used to quantify the fraction of stress fibers-positive cells (> 200 cells were counted for each group) with ImageJ software.

EMSA (electrophoretic mobility shift assay)

Nuclear extracts from HEK-293 cells transfected with a ZEB1 expression vector and mock controls were prepared as previously described (Barozzi et al., 2014). IRDye 700 labeled synthetic oligonucleotides (Metabion) harboring 2 or 4 ZEB1 specific consensus sequences were designed from the ZEB1-bound region upstream of the miR-200b/a locus (chr1:1058873-1059613; the ZEB1 motif is underlined):

2x (chr1:1059589-1059607):
5’GCACACCTGGACACACACCTGCAC3’;
4x (chr1:1059147-1059198): 5’GCACACCTGCACATACACCTGAGCACACATCTGCACACACCTGAGCACACACCTGAGC3’;

The synthetic probe harboring 6 consensus sites was designed duplicating part of the 4x probe to generate 6 evenly distributed consensus sites: 5’GCACACCTGCACATACACCTGAGCACACATCTGCACACACCTGAGCACACACCTGAGCACACATCTGCACACACCTGAGCACACACCTGAGC3’.

Binding reactions were assembled in 10 mM Tris-HCl (pH 7.5), 50 mM NaCl, 0.1mM EDTA, 2 mM DTT, 0.25% Tween20, 4 mM MgCl2, 1 μg of poly dI:dC, and the desired amount of nuclear extract (ranging from 0.5 to 16 μg). The mixtures were incubated with 0.1 pmol of labeled probe for 20 min at room temperature, and complexes were resolved on 4% polyacrylamide Tris/Borate/EDTA (TBE) native gel in the dark using 0.5X TBE buffer (pH 8) and running for ~ 150 min at 0.01A at 4°C. The gel was scanned with the Li-Cor Odissey Infrared Imaging System and shifted bands were quantified.

Luciferase assay

ZEB1 clustered motifs in TRs in miR200b/a locus were PCR amplified with specific set of primers designed to obtained products harboring 37, 30, 15 and 9 ZEB1 binding motifs. The following reverse primers were designed progressively closer to a fixed forward primer (F: 5’-CGGCAATTGGAGAAGCCCAGGAGCAAGTA-3’) to obtained the desired number of motifs:

rev37 (5’-CGGACGCGTAGGGTGGTGGTTTCTCAGAG-3’),
rev30 (5’- CGGACGCGTTCAGGTGTGTATGCAGTGGT-3’),
rev15 (5’-CGGACGCGTTCAGGTGTATGTGCAGGTGT-3’),
rev9 (5’- CGGACGCGTGCAGGTGTGTTTTCAGGTGT-3’).

PCR products harboring MfeI and MluI restriction sites were cloned into pCDNA3.1 based vector upstream the CMV promoter driving NanoLuc expression (kind gift from Bruno Amati, IEO) and verified by Sanger sequencing. The day before transfection, 5 x 10E4 Hela cells were seeded into 24-well dish and co-transfected with 25ng NanoLuc vectors and 2.5ng PGK-Firefly luciferase (pGL4.53) using Lipofectamine2000 following manual instructions. 24 h post transfection luminescent activity was assessed with the Nano-Glo Dual Luciferase reporter assay kit and measured at GloMax Detection system (Promega). Data are presented as Relative Luminescence Unit (RLU) normalized to NanoLuc signal by Firefly luciferase activity.

Mouse xenografts

CD1-Nude mice obtained from Charles River (n = 5 for each group) were injected with 10 x 10E6 cells (MiaPaCa-2 ZEB1-KO, MiaPaCa-2 Δ-Repeats and their matched controls), resuspended in 100 μl of PBS under the skin of their hind flank. Subcutaneous tumor growth was assessed every 7±2 days for tumor growth for a total of ~5 weeks.

QUANTIFICATION AND STATISTICAL ANALYSIS

ChIP-seq data analysis

Short reads obtained from Illumina HiSeq 2000 were quality filtered according to the Illumina pipeline. Reads were then mapped to the human hg19 reference genome using Bowtie2 v2.2.6 (Langmead and Salzberg, 2012) with the “*–very-sensitive*” preset of parameters. Reads that did not align to the nuclear genome or aligned to the mitochondrial genome were removed. Moreover, duplicate reads were marked and removed using SAMtools (Li et al., 2009). Peak calling *vs.* the input genomic DNA was performed using MACS2 (version 2.1.0.20150731)(Zhang et al., 2008) using the “*–nomodel*”, *“–extsize 200*” and “*–qvalue 0.01*” flags and arguments. Peaks with a fold enrichment (FE) relative to input <5 (as determined by MACS2) and those blacklisted by the ENCODE consortium analysis of artifactual signals in human cells (https://sites.google.com/site/anshulkundaje/projects/blacklists) were removed using bedtools (Quinlan and Hall, 2010).

Annotation and classification of ChIP-seq peaks

To classify ChIP-seq peaks based on their genomic location and assign them to the nearest TSS, the September 2015 RefSeq annotation of the hg19 version of the human genome was given as input to the *annotatePeaks* script from HOMER package (Heinz et al., 2010). We classified each peak as either TSS-proximal or TSS-distal, depending on its distance (> or < 2.5 kb, respectively) from annotated transcription start sites (TSS).

Heatmap of ZEB1 ChIP-seq enrichment in MiaPaCa2 cell line

Reads Per Million (RPM) were measured in a window of 5 kb (500 bins of 10 bp) centered on the summits of ZEB1 peaks. To avoid any bias due to outliers, a saturation procedure was performed and values were then scaled to the range 0-1. Regions were sorted according to their intensity levels and visualized using heatmap.2 in R.

De novo motif discovery

Motif discovery was performed using MEME v4.10.1 (Bailey et al., 2009) with the options *"-dna -mod zoops -evt 2e-4 -nmotifs 6 -minw 6 -maxw 12 -revcomp -maxsize 10+7*" using a window of +/-100 bp centered on the summits of the 500 highest-scoring ZEB1 peaks. We next used TomTom (Gupta et al., 2007), with default parameters except for "*-dist ed*", in order to assess the similarity of the identified motifs to the ZEB1 consensus binding site obtained from the JASPAR database (http://jaspar.genereg.net).

Gene ontology analysis

Functional enrichment analyses were performed using the DAVID tool (version 6.8 Beta) (Huang et al., 2007). Gene Ontology (GO) terms were found by comparing the set of genes bound by ZEB1 at their TSS in MiaPaCa2 to a background corresponding to the complete list of annotated human genes. We restricted the analysis to GO terms with less than 100 annotated genes and with Fisher Exact P-value $\leq$ 0.01. Data visualization was carried out using REVIGO (Supek et al., 2011) with default parameters except for the resulting list that was set as "*small size*".

Smart-seq2 analysis

After quality filtering according to the Illumina pipeline, 50 bp single-end reads were aligned to the hg19 human reference genome and to the *Homo sapiens* transcriptome (NCBI build 37.2) using TopHat (version 2.1.0)(Trapnell et al., 2012) with the option "*–b2-very-sensitive*". Only uniquely mapped reads were retained. At the gene level, expression counts were estimated using HTSeq (version 0.6.1) (Anders et al., 2015), summarized across all exons as annotated in the NCBI build 37.2, with option "*union*" and "*no strand-specific assay*". Both coding and noncoding genes were retained for downstream analyses. Differentially expressed genes in biological triplicates of wild type and ZEB1-KO MiaPaCa2 clones were identified using EdgeR R-package (version 3.2.2)(Robinson et al., 2010). Prior to normalization using the Trimmed Mean of M (TMM) method, only genes with at least 1 CPM (Count Per Million) in at least half of the samples were retained. A common dispersion was estimated for all genes to measure the global biological variation (with option *robust = "TRUE"*). A negative binomial generalized log-linear model was fitted to each gene, and likelihood ratio tests were performed to assess differential expression. Genes were identified as differentially expressed when the following criteria were met: fold-changes (FC) $\geq |2|$ and false discovery rate (FDR) $\leq$ 0.01. Then, Transcript Per Million (TPM) values were used as expression unit.

Gene ontology analysis of genes de-repressed in ZEB1-KO clones

GO term enrichment was tested with the DAVID tool (version 6.8 Beta) (Huang et al., 2007), using derepressed genes in MiaPaCa2 ZEB1-KO as input and all the nonzero genes as background. We restricted the analysis to GO terms with Fisher Exact P-value $\leq$ 0.05. Data visualization was carried out using REVIGO (Supek et al., 2011) selecting default parameters except for the resulting list that was set to "*medium size*".

Association of de-repressed genes to ZEB1 ChIP-seq peaks

We considered windows encompassing 2.5 kb upstream and downstream of the RefSeq TSS of derepressed genes. These windows were intersected with the total set of ZEB1 ChIP-seq peaks, and the overlapping peaks were annotated to the corresponding genes. The number of perfect ZEB1 motifs in the underlying genomic DNA was then scored. When multiple peaks were associated with a single TSS, the number of motifs was summed.

Identification of Tandem Repeats in ZEB1 ChIP-seq peaks

To identify Tandem Repeats (TRs) overlapping with ZEB1 peaks, we first downloaded from the UCSC genome browser the Simple Tandem Repeats catalog generated by Tandem Repeats Finder (TRF)(Benson, 1999). TRs were then intersected with TSS-proximal and distal ZEB1 peaks. The results were filtered to remove low-quality TRs with an insert/delete percentage > 20% and TR sequences without a ZEB1 motif (one mismatch was allowed if in the first or last nucleotide).

Evolutionary analysis and genetic variability of the miR-200b/a/miR-429-proximal TR

The TR upstream of the miR-200b/a/miR-429 cluster on chromosome 1 (hg19 coordinates: chr1:1,056,744-1,061,743) was analyzed to determine interspecies alignment and inter-individual variation. For interspecies alignment, we downloaded from UCSC browser the following Net Alignments tracks: *Macaca mulatta* (Mac, BGI CR_1.0/rheMac3), *Callithrix jacchus* (Mar, WUGSC 3.2/calJac3), *Mus musculus* (Mm, GRCm38/mm10), *Rattus Norvegicus* (Rn, Baylor 3.4/rn4), *Bos taurus* (Bt, Btau 4.6.1/bosTau7), *Canis lupus familiaris* (Dog, CanFam3.1/canFam3), *Felis Catus* (Cat, Felis_catus 6.2/felCat5).

For the identification of genetic variants in the miR-200-associated TR, the phase 3 data of the 1000 Genomes Project (Auton et al., 2015); (Sudmant et al., 2015) were used. The VCF files for the Single Nucleotide Polymorphisms (SNPs) and the Single Nucleotide

Variants (SNVs) datasets (GRCh37 coordinates) were directly downloaded from the International Genome Sample Resource (IGSR). Moreover, the Structural Variants (study ID *estd219*) (Sudmant et al., 2015) were downloaded from the Database of Genomic Variants Archive (DGVa).

Correlation between ChIP-seq tag density and ZEB1 motif occurrences

Occurrences of perfect ZEB1 motifs (5’-*CACCTG*-3’ and its reverse 5’-*CAGGTG*-3’) at ZEB1 bound regions were identified for each cell line separately, in order to classify the binding regions by the presence of multiple motifs. Regions were sorted by the number of motif occurrences and the resulting ranked list was binned into the following five sets:

(I) no perfect match; (II) from 1 to 2 motifs; (III) from 3 to 5 motifs; (IV) from 6 to 10 motifs; and (V) more than 10 motifs.

The analysis was separately performed for TSS-proximal and distal peaks. The ZEB1 binding signal was evaluated as RPM/kb normalized read density. These values were log2-transformed (*pseudo-count* of 1) and the distributions across the 5 sets were displayed as a violin plot generated in the R statistical environment. The significance of distributions between sets ("(I) vs. (IV)" and "(I) set vs. (V)") was valuated using the one-tailed Wilcoxon rank-sum test.

Construction of a catalog of homotypic clusters of motifs bound by ZEB1 in multiple cell lines

Peaks from all cell lines were combined into a unified catalog using mergeBed (Quinlan and Hall, 2010) with option “*-d 50*”. The resulting regions were intersected with the original peaks in each cell line so that every region was annotated to the corresponding set of peaks. Subsequently, the regions were annotated based on the number of perfect ZEB1 motifs. This preliminary mapping was used to identify regions bound by ZEB1 across cell lines. Using the TSS-distal set, we removed those regions that were bound in less than 3 cell lines and/or that contained fewer than 6 ZEB1 motifs, obtaining an initial set of 254 regions. The remaining 12,011 regions represented a background set used in different comparison tests. To generate the final homotypic cluster catalog, we first detected the distance preferences between consecutive motif pairs in order to discard regions containing dispersed motifs. Specifically, we analyzed the distance arrangements between consecutive motifs in each repeat. To this aim, we computed the distance between the last nucleotide of the first motif and the first nucleotide of the second one. We observed that even if the motif-to-motif distance distribution was not equal across regions, 75% of consecutive motifs showed an interval shorter than 100 bp. Therefore, we used this cut-off to discard repeats with dispersed motifs. The final catalog (**Table S3**) included 193 TSS-distal homotypic clusters bound by ZEB1 in more than 3 cell lines. Every cluster was described by the number of motifs it contained and the phyloP scores. Genes in a 500kb window from the repeat were also annotated.

Orientation of ZEB1 motifs in homotypic clusters

We computed the orientation (forward or reverse complement) of motifs in the same homotypic cluster. For representation purposes, the motifs were lined up and the intervening sequences eliminated. Then, each motif was labeled 0 (forward) or 1 (reverse) and the resulting binary matrix shown as a bar plot.

PhyloP scores

PhyloP scores (Pollard et al., 2010) were calculated for each homotypic cluster using the 100 vertebrates Basewise Conservation by PhyloP (phyloP100way) track downloaded from the UCSC genome browser. Mean values in windows of 1 kb around the center of clusters were archived and significance relative to all other ZEB1-bound regions with the same size was calculated using a one-tailed Wilcoxon rank-sum test.

Analysis of sub-telomeric localization of homotypic clusters of ZEB1 motifs

We calculated the distance of the 193 ZEB1 homotypic clusters bound in at least three cell lines (see above) from the closest telomere. We classified as subtelomeric the clusters located within 10 Mbp from a telomere end. Fisher’s exact test was used to test the null hypothesis that proportion of homotypic clusters and background region associated with the sub-telomeric regions was equal.

MicroRNA expression analysis in Low-grade and High-grade PDAC cell lines

50 bp single end sequences were quality controlled using FastQC. Before alignment, raw sequences reads were filtered for low-quality reads, contaminating adapters and homopolymers and trimmed for 3′ adapters using Cutadapt (version 1.7.1) (Martin, 2011) with options “*-e 0.12 -O 5 -m 15*”. The preprocessed reads were aligned first to known miRNAs (GENCODE release 24, mapped to GRCh37) and then to the hg19 human reference genome using Bowtie2 (version 2.2.6)(Langmead and Salzberg, 2012) with the following parameters*:* “*–local*”, “*-D 20*”, “*-R 3*”, “*-N 0*”, “*-L 8*” and “*-i S,1,0.50*“. Only uniquely aligned reads were retained.

Expression of annotated miRNAs was estimated as raw read counts using bedtools multicov (Quinlan and Hall, 2010). Differential miRNA expression between Low-grade PDACs (CAPAN1, CAPAN2 and CFPAC1) and High-grade PDACs (MiaPaCa2, PANC1 and PT45P1) was calculated using EdgeR R-package (version 3.2.2)(Robinson et al., 2010). Prior to normalization using the Trimmed Mean of M (TMM) method, only miRNAs showing at least 1 CPM in at least half of the samples were kept for further analysis. In the absence of replicas, the options method = “*deviance*”*,* robust = ”*TRUE” and* subset = ”*NULL”* were used for estimating the common dispersion. Then, a negative binomial generalized log-linear model was fitted to each miRNA, and likelihood ratio tests were performed to assess differential expression. miRNAs were considered differentially expressed when both the following criteria were met: FC $\geq |2|$ and FDR ≤ 0.001.

Genome browser tracks

We applied RPM normalization to all datasets and tracks for visualization in the UCSC Genome Browser and Integrative Genomics Viewer (IGV) were generated using bedGraphToBigWig tool (Li et al., 2009; Quinlan and Hall, 2010).

Datasets

The ChIP-seq data sets used are reported in the Key Resource Table.

DATA AND SOFTWARE AVAILABILITY

Data sets are available in the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo) under the accession number GEO: GSE88738.

Supplemental Figures

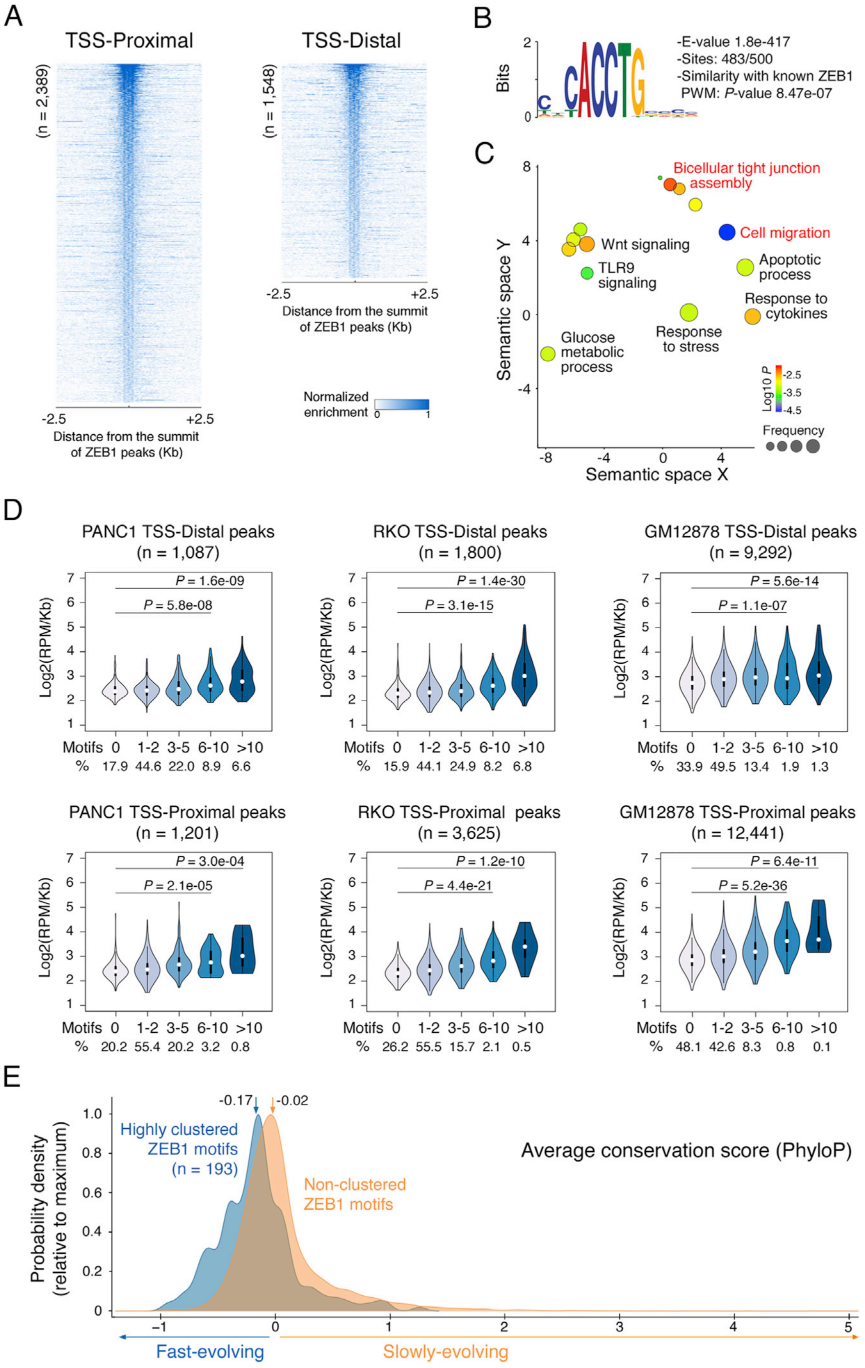


Figure S1. Binding of ZEB1 to Genomic Sites Containing Homotypic Clusters of Motifs, Related to Figure 1

(A) Heatmaps of z-normalized RPM values showing TSS-proximal (left) and distal ZEB1 peaks, (right) identified by ChIP-seq in MiaPaCa2 cells. Data were plotted for a 5kb window around the summit of peaks and ordered based on decreasing enrichment.

(B) *De novo* motif discovery was carried out on the top 500 ZEB1 peaks.

(C) GeneOntology (GO) categories associated with the genes directly bound at their promoter by ZEB1. Bubble color indicates the FDR while size indicates the frequency of the GO term in the GOA database (bubbles of more general GO terms are larger). The complete list is in Table S1.

(D) Relationship between ZEB1 ChIP-seq peak intensity and number of ZEB1 motifs in the underlying sequence in multiple cell lines. TSS-distal (top) and TSS-proximal (bottom) peaks are shown. The percentage of peaks (%) for each class of motifs is indicated. Median: white central dots. Significance of dissimilarity between distributions was calculated using a one-tailed Wilcoxon rank-sum test.

(E) Conservation of ZEB1-bound genomic regions. Density plot showing evolutionary conservation of ZEB1-bound genomic regions that contain homotypic clusters with high density of motifs, which include TRs (light blue), compared to the remaining genomic sequences bound by ZEB1 (orange).

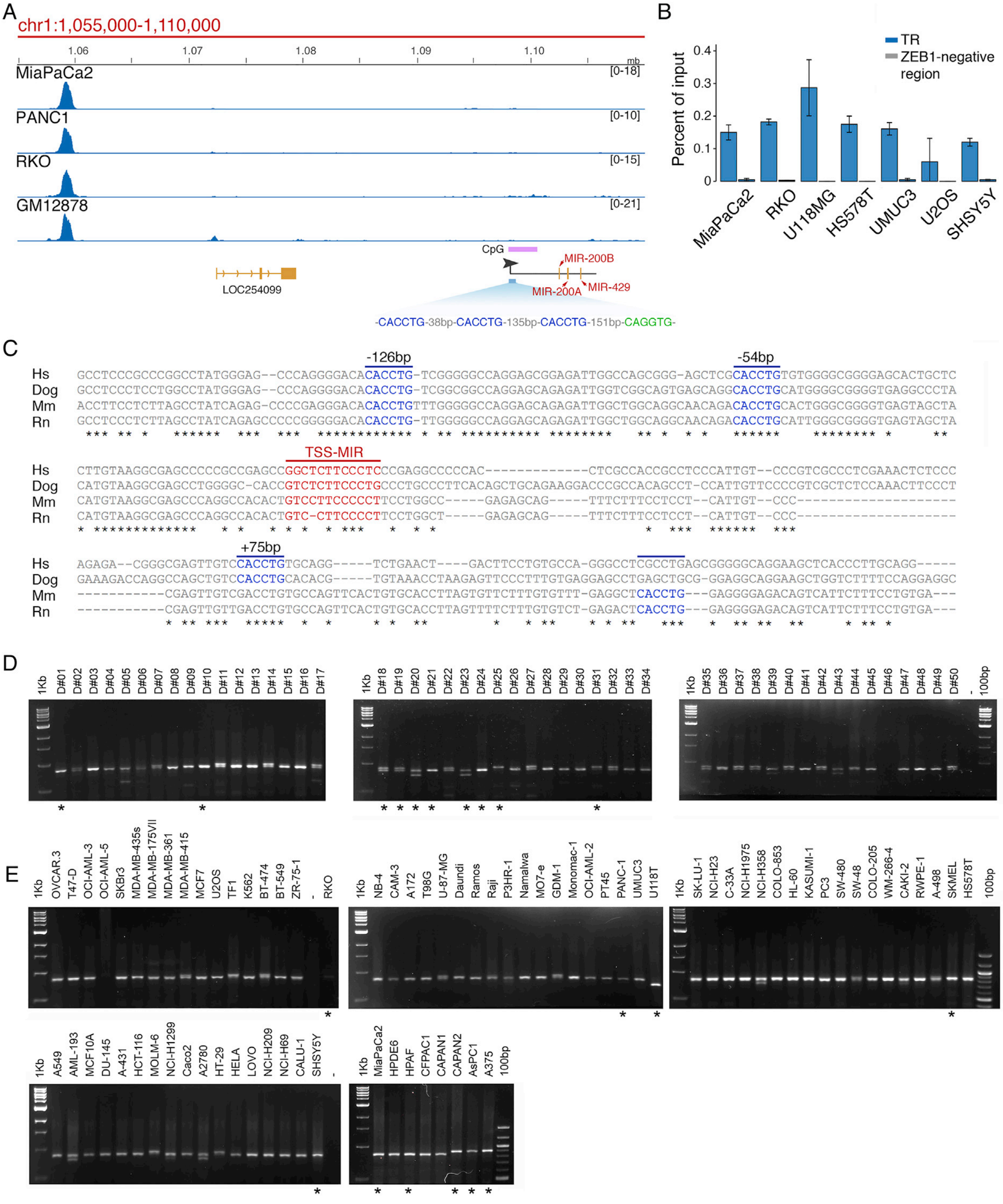


Figure S2. ZEB1 Binding to the chr1 TR Upstream of miR-200b/a across Multiple Cell Lines, Related to Figure 2

(A) ZEB1 ChIP-seq snapshot of the genomic region surrounding the miR200b-a/miR429 locus in four different cell lines.

(B) ZEB1 ChIP-qPCR data at the chromosome 1 TR were obtained from multiple cell lines. MiaPaCa2 and RKO cells are also shown as reference. Means ± SD are shown.

(C) Multi-species alignment of DNA sequences surrounding the TSS (TSS-MIR) of the transcript encoding miR-200b, miR-200a and miR-429. The conserved ZEB1 motifs are highlighted in blue. Asterisks indicate conserved nucleotides.

(D and E) Analysis of the TR upstream of the miR-200b locus in 50 normal individuals (D) and in 78 cancer cell lines (E). The region surrounding the chromosome 1 TR was amplified by PCR and analyzed by agarose gel electrophoresis. Asterisks indicate individuals whose amplified DNA was subjected to Sanger sequencing.

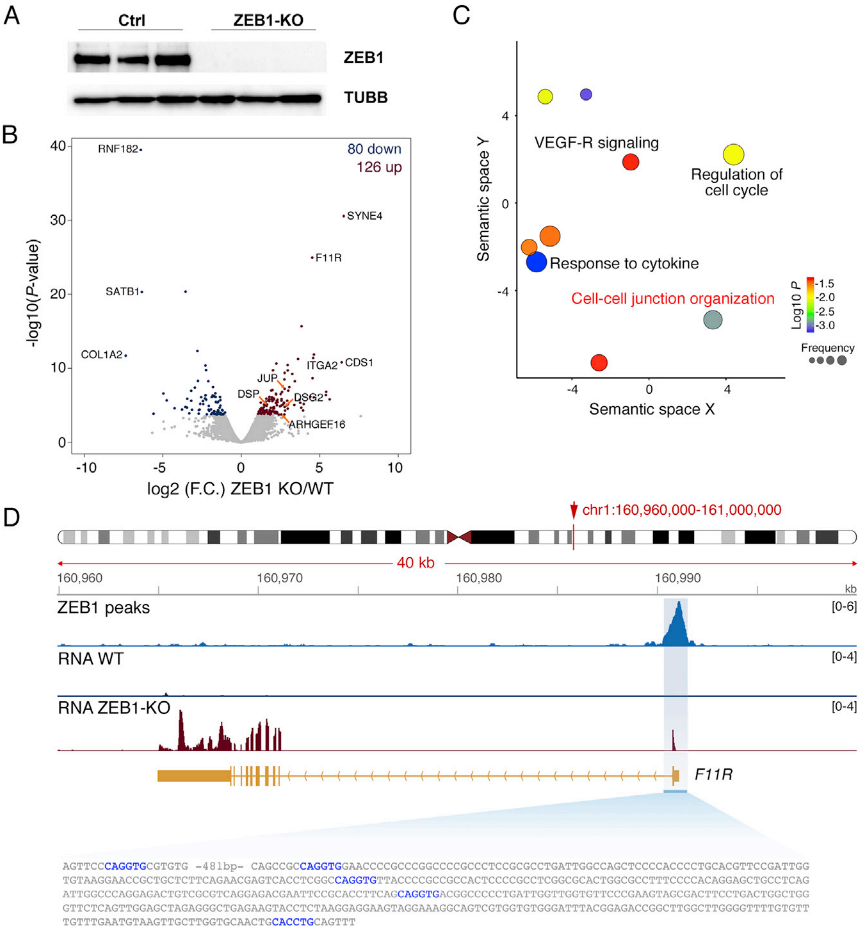


Figure S3. Effects of the Deletion of the ZEB1 Gene in MiaPaCa2 Cells, Related to Figure 3

(A) ZEB1 western blot in three control (Ctrl) clones and three ZEB1-KO clones generated by CRISPR/Cas9-mediated genome editing in MiaPaCa2 cells.

(B) Volcano plot showing differentially expressed genes in ZEB1-KO MiaPaCa2 cells. The log2 expression fold change (FC) is shown on the horizontal axis and -log10 of the *P*-value is shown on the vertical axis. Up-regulated genes (FDR≤0.01 and log2 FC≥1) are highlighted in red while down-regulated genes in blue (FDR≤0.01 and log2 FC≤-1).

(C) Representative set of GO categories associated with genes up-regulated in ZEB1-KO cells. The complete list is in Table S5.

(D) Snapshot showing ZEB1 binding to the promoter of the differentially expressed *F11R* gene. The cluster of ZEB1 motifs is shown below.

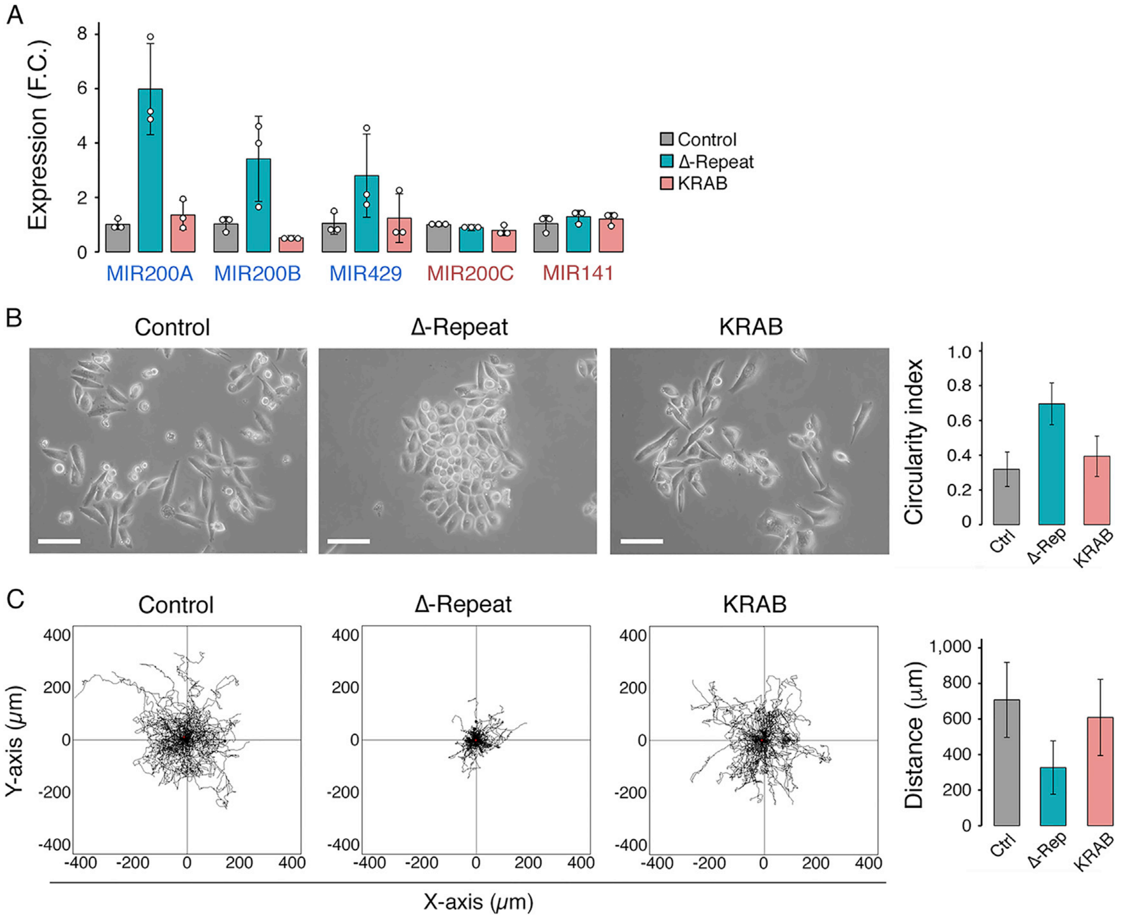


Figure S4. Reversion of Phenotypes in Δ-Repeat MiaPaCa2 Cells by Targeted Recruitment of KRAB-Repressive Domains, Related to Figure 3

(A) Δ-Repeat clones (n=3) were transduced with vectors for CRISPR/dCas9-mediated targeted recruitment of KRAB repressive domains to a region adjacent to the deleted TR (pink bars); wild type and an additional set of Δ-Repeat clones (*n=3* each) were transduced with the same vectors lacking the targeting sgRNA (grey and green bars, respectively). The indicated miRNAs were measured by qPCR. For each miRNA, data are expressed as fold change calculated as 2ˆ-ΔΔCq relative to the average of the ΔCq of the wild type control cells containing the TR normalized to *miR103* as reference gene. Means ± SD are shown.

(B) Morphology and circularity index of control, Δ-Repeat cells and Δ-Repeat-KRAB-targeted cells (*n=45* cells per group). Means ± SD are shown. Scale bar=50 μm.

(C) Trajectory plots and accumulated distance bar-graph obtained by random migration assay in the same cells as in B (*n=75* cells per group).

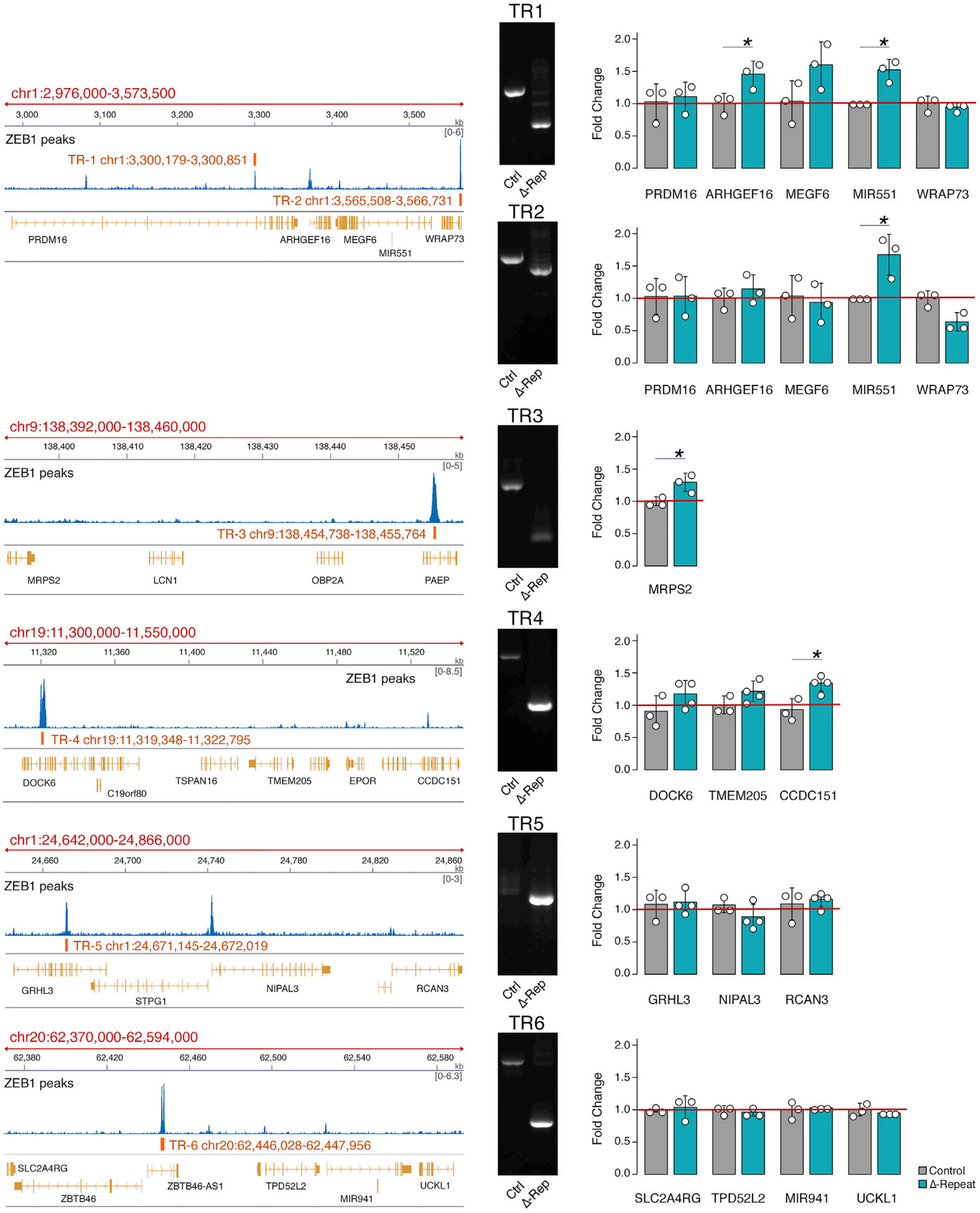


Figure S5. Effects of the Deletion of a Panel of TRs on the Transcription of Adjacent Genes, Related to Figure 3

(Left) The deletion efficiency of the six TRs analyzed was assessed by PCR on the genomic DNA of control and deleted (Δ-Rep) polyclonal populations of MiaPaCa2 cells.

(Right) Expression of the genes surrounding the TRs was tested by RT-qPCR on wild type (*n=3* independent experiments) and Δ-Rep cells (*n=3* for TR1, TR2, TR3, TR6; *n=4* for TR4 and TR5). Values represent fold changes calculated as 2ˆ-ΔΔCq relative to the average of the ΔCq of the wild type control cells normalized to *C1ORF43* as reference gene. Means ± SD are shown. * *P* < 0.05 by Two tailed t-test.