nature biotechnology

High-throughput mapping of regulatory DNA

Nisha Rajagopal[1], Sharanya Srinivasan[1,2], Kameron Kooshesh[2,3], Yuchun Guo[1], Matthew D Edwards[1], Budhaditya Banerjee[2], Tahin Syed[1], Bart J M Emons[2,4], David K Gifford[1] & Richard I Sherwood[2]

Quantifying the effects of *cis*-regulatory DNA on gene expression is a major challenge. Here, we present the multiplexed editing regulatory assay (MERA), a high-throughput CRISPR-Cas9–based approach that analyzes the functional impact of the regulatory genome in its native context. MERA tiles thousands of mutations across ~40 kb of *cis*-regulatory genomic space and uses knock-in green fluorescent protein (GFP) reporters to read out gene activity. Using this approach, we obtain quantitative information on the contribution of *cis*-regulatory regions to gene expression. We identify proximal and distal regulatory elements necessary for expression of four embryonic stem cell–specific genes. We show a consistent contribution of neighboring gene promoters to gene expression and identify unmarked regulatory elements (UREs) that control gene expression but do not have typical enhancer epigenetic or chromatin features. We compare thousands of functional and nonfunctional genotypes at a genomic location and identify the base pair–resolution functional motifs of regulatory elements.

Gene regulation provides the basis for cell type–specific function. Although differences in *cis*-regulatory DNA are known to underlie human variation and disease, predicting the effects of *cis*-regulatory variants on gene expression remains challenging.

Important strides have been made over the past decade in cataloging gene regulatory elements. A histone modification code has been found to correlate with *cis*-regulatory elements, such as enhancers and promoters, and chromatin states, such as active and poised[1–5]. Gene-expression reporter assays, which can now be done in high-throughput formats[6–8], have confirmed elements that are sufficient to activate gene expression in heterologous contexts. Additionally, techniques to identify distal DNA interactions have begun to associate enhancers with their cognate promoters[9–12], which often are not in close proximity and can at times be megabases apart.

However, existing techniques for identifying gene regulatory regions have several shortcomings. Reporter assays focus on elements that are sufficient to activate gene expression in a heterologous context and therefore cannot characterize elements that are necessary but not sufficient for gene expression or whose activity does not transfer to a non-native context. Additionally, genes can have many regulatory elements, and there is no high-throughput approach capable of determining the relative importance of each element in influencing native gene expression levels. Efforts to systematically test enhancers, predicted using histone modification data from reporter assays, have found that the majority of predicted enhancers do not activate gene expression as expected[13]. This suggests that additional assays are required to decipher native gene regulation.

CRISPR-Cas9 has been used in genome-wide mutation screens to identify genes required for survival, drug resistance and tumor metastasis[14–18]. In these screens, guide RNAs (gRNAs) targeting tens of thousands of sites within genes are cloned into lentiviral vectors and delivered as a pool into target cells along with Cas9. By identifying gRNAs that are enriched or depleted in the cells after selection for a desired phenotype, genes that are required for this phenotype can be systematically identified.

Here we develop CRISPR-Cas9–based MERA to analyze the regulatory genome at single-base resolution in its native context. MERA employs Cas9, which has been shown to cleave DNA when paired with a single gRNA[19–22]. In MERA, Cas9-induced double-strand breaks (DSBs) are repaired in an error-prone fashion by cellular non-homologous end joining (NHEJ), inducing a wide range of mutations initiated at the cleavage site that typically are small (<10-bp) insertion or deletions (indels) but can include larger (>100-bp) indels[20,21,23] and altered individual bases.

The MERA assay first carries out a high-throughput screen that maps the effects of genomic variation on gene expression. Selected elements can then be characterized by functional motif discovery and validated. We map elements that are required for gene expression by expressing gRNAs that tile a gene's *cis*-regulatory region and measuring how likely each gRNA is to diminish gene expression. We then perform deep sequencing of the gRNA-induced mutations in targeted regions to reveal thousands of genotypes that either did or did not lose gene expression. This enables us to characterize the functional importance of each base. Finally, we validate the results of the MERA screen through the replacement of selected genomic elements by homologous recombination.

RESULTS

Developing the MERA assay

There are two distinctions between MERA and previous gene mutation screening approaches that spurred us to alter the CRISPR-Cas9–based mutation screening technique. First, the targeted sites

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [3]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. [4]Program in Cancer, Stem Cells, and Developmental Biology, University of Utrecht, Utrecht, the Netherlands. Correspondence should be addressed to D.K.G. (gifford@mit.edu) or R.I.S. (rsherwood@partners.org).

Received 11 May 2015; accepted 24 December 2015; published online 25 January 2016; doi:10.1038/nbt.3468

in our screen are often close together, so cells receiving more than one gRNA may undergo deletion instead of mutation of a region, which would complicate downstream analysis. Although this issue can be addressed for lentiviral libraries by lowering the multiplicity of infection (MOI), we sought a more elegant approach to limit cells to a single gRNA. Second, each gene for which we perform MERA requires a different gRNA library. All high-throughput CRISPR-Cas9–based approaches to date have required cloning gRNA libraries into a lentiviral vector and producing a batch of virus, a time-consuming process that would have to be done separately for each library. We sought an approach that would allow a library to be used on the day it arrives.

To enable the efficient targeting of precisely one regulatory element per cell, we devised a strategy that ensures that only one gRNA can be expressed per cell and allows gRNA libraries to be used without any molecular cloning into a delivery vector. We integrated a single copy of the gRNA expression construct (a U6 promoter driving expression of a dummy gRNA hairpin) into the universally accessible *ROSA* locus of mouse embryonic stem cells (mESCs) using CRISPR-Cas9–mediated homologous recombination (**Fig. 1a**). We then use CRISPR-Cas9–mediated homologous recombination to replace the dummy gRNA with a gRNA from our library. We use PCR to add 79- to 90-bp homology arms to our gRNA library, as we found that longer homology arms increase background cutting of gRNAs transcribed from unintegrated PCR fragments (**Supplementary Fig. 1**). We then introduce the pool of gRNA homology fragments into cells along with Cas9 and a gRNA plasmid that induces a DSB at the dummy gRNA site. In a substantial fraction of cells (~30%), the dummy gRNA is repaired by homologous recombination, creating a functional gRNA expression construct targeting a single genomic site from the library (**Fig. 1a** and **Supplementary Fig. 2**). Only random chance dictates which gRNA is integrated in each cell, allowing a pooled screen in which each cell expresses only one gRNA.

Of note, the genomic integration–based gRNA screening platform used in MERA could also be applied to other CRISPR-based high-throughput screens as long as the cell line used undergoes homologous recombination at appreciable frequency, and it could be modified to achieve expression of any set number of gRNAs per cell for combinatorial screening. Although the integration-based approach is thus ill-suited to *in vivo* screens or screens in cells with limited homologous recombination, it provides an alternative to lentiviral screening that substantially reduces the time, effort and cost involved in CRISPR library screening for applicable cell lines such as embryonic stem cells.

We generated GFP knock-in lines for four mESC-specific genes, *Nanog*, *Rpp25*, *Tdgf1* and *Zfp42* (**Fig. 1b**, **Supplementary Fig. 3**, ref. 24), and synthesized corresponding gRNA libraries, each with 3,908 gRNAs tiling *cis*-regulatory regions. In the case of *Tdgf1*, the library targeted the 40-kb region proximal to the gene in an unbiased manner. In other cases, we selected regions proximal to the gene most likely to be involved in regulation based on enhancer-like features[25–31] that are a maximum of ~150 kb away from the gene, as well as distal regions up to 92 Mb away from the gene when ChIA-PET distal interaction data[2] suggested a possible interaction with the target gene promoter[3]. Among the 3,621 gRNAs found to be integrated in at least one of three replicates of the *Tdgf1* library, the mean distance between adjacent gRNAs was 11 bp. Of note, repetitive and unmappable genomic regions cannot be tiled with gRNAs, and gRNAs targeting regions whose sequences differ from those in the reference genome cannot be appropriately tiled without genome sequence data of the cell line.

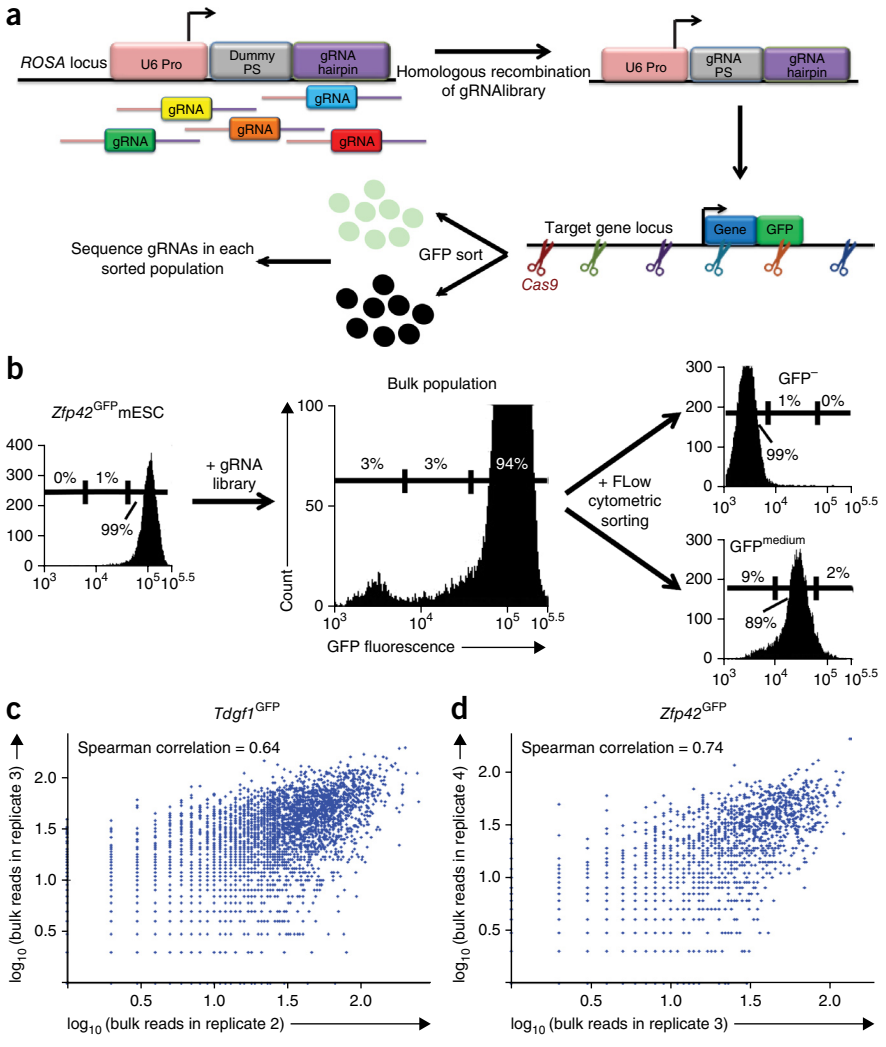

Figure 1 Multiplexed editing regulatory assay (MERA). (**a**) In MERA, a genomically integrated dummy gRNA is replaced with a pooled library of gRNAs through CRISPR-Cas9–based homologous recombination such that each cell receives a single gRNA. Guide RNAs are tiled across the *cis*-regulatory regions of a GFP-tagged gene locus, and cells are flow cytometrically sorted according to their GFP expression levels. Deep sequencing on each population is used to identify gRNAs preferentially associated with partial or complete loss of gene expression. (**b**) *Zfp42*^{GFP} mESCs show uniformly strong GFP expression. After bulk gRNA integration, a subpopulation of cells lose GFP expression partially or completely. These cells are flow cytometrically isolated for deep sequencing. (**c**,**d**) Bulk reads for gRNAs are highly correlated between replicates from the *Tdgf1* (**c**) or *Zfp42* libraries (**d**), indicating consistent and replicable integration rates.

Each library also contained ten positive control gRNAs targeting the *GFP* open reading frame that we expected would cause loss of GFP expression.

MERA screens identify required regulatory regions

We performed four biological replicate screens for *Zfp42* and *Tdgf1*, two replicates for *Nanog* and a single replicate for *Rpp25*. Selected screen hits were independently confirmed as described below. Starting 1 week after electroporation, we collected genomic DNA of the unsorted library-integrated cells to examine differences in gRNA integration. Over 90% of correctly synthesized gRNAs were detected in the genomic DNA for both *Tdgf1* and *Zfp42* libraries (**Supplementary Methods**). In addition, gRNA integration rates in the bulk populationshowed concordance between the biological replicates (**Fig. 1c**,**d** and **Supplementary Fig. 4a**). All of the regulatory regions that we surveyed had sufficient coverage of gRNAs to allow us to assay their detailed function (bulk density track, **Figs. 2** and **3**; **Supplementary Figs. 5** and **6**).

Library-integrated mESCs were then flow cytometrically sorted to identify gRNAs that induced loss of GFP expression. Separate GFP^{neg} and GFP^{medium} populations were sorted in the *Tdgf1*^{GFP} and *Zfp42*^{GFP} experiments, whereas GFP^{neg} and GFP^{medium} populations were combined in the *Nanog*^{GFP} and *Rpp25*^{GFP} experiments because of incomplete population separation (**Fig. 1b** and **Supplementary Fig. 3**).

The distribution of gRNA abundance in GFP^{neg} and GFP^{medium} populations in all screens clearly indicates that a subset of *cis*-regulatory genomic space is required for gene expression at all four gene loci (**Figs. 2a**,**b** and **3**, **Supplementary Tables 1**–**4**). We detected significant over-representation of nearly all integrated positive-control *GFP* coding region–targeting gRNAs in all replicates (**Figs. 2d** and **3c**, **Supplementary Fig. 4b**), suggesting that MERA robustly identifies gRNAs that induce loss of gene expression. Using the relative abundances of *GFP* coding region-targeting positive control gRNAs and the dummy gRNA as a negative control, we devised a method to detect gRNAs with statistically significant over-representation in GFP^{neg} and GFP^{medium} populations (Online Methods, **Supplementary Fig. 4b**,**c**, **Supplementary Table 5**).

In our MERA screen of *Tdgf1*, we observed differential enrichment of gRNAs in established functional categories of genomic elements associated with gene regulation[27–31] (**Fig. 2a**,**d** and **Supplementary Fig. 5**).

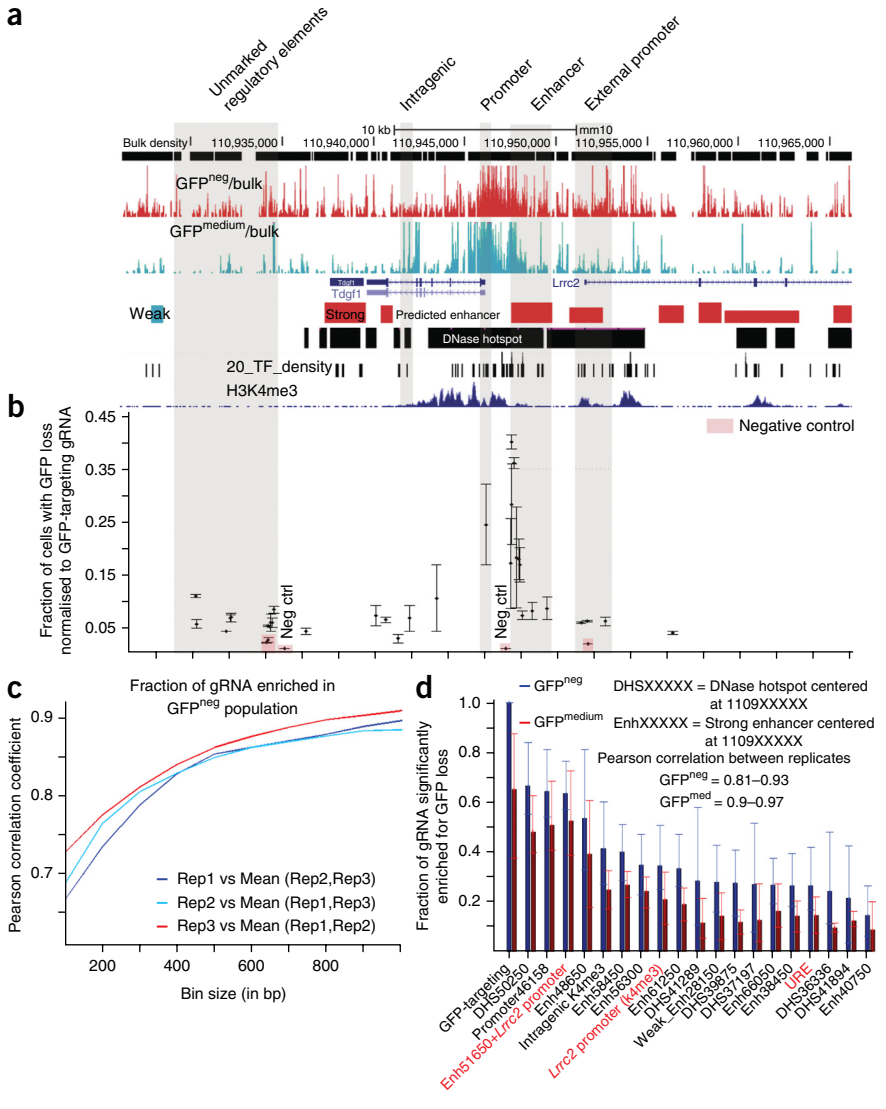

Figure 2 MERA enables systematic identification of required *cis*-regulatory elements for *Tdgf1*. (**a**) A genomic view of the 40-kb *Tdgf1* proximal regulatory region showing the following in track order from top to bottom. (i) The locations of all 3,621 integrated gRNAs in any one of three biological replicates. (ii) gRNAs enriched in GFP^{neg} cells (red) in any one of three replicates at $P < 10^{-10}$ using a binomial test as described in the methods; bar height is proportional to the mean log-ratio of GFP^{neg} to bulk reads across replicates. (iii) gRNAs enriched in GFP^{medium} cells (cyan) in any one of three replicates at $P < 10^{-10}$ using a binomial test as described in the Online Methods; bar height is proportional to the mean log-ratio of GFP^{neg} to bulk reads across replicates. (iv) Annotated genes. (v) Predicted enhancers (cyan, weak; red, strong). (vi) DNase I hotspot regions. (vii) Transcription factor binding density based on ChIP-seq data. (viii) H3K4me3 ChIP-seq data (blue). Several active regulatory elements coincide with dense clusters of overlapping gRNAs. Numerous gRNAs significantly enriched in the GFP^{neg} population are also observed in regions devoid of regulatory element features (UREs). Genomic regions of interest are shaded, annotated above the plot, and described in further detail in the text. (**b**) Individual validation of specific gRNAs detected as enriched in the GFP^{neg} population in the MERA assay using the self-cloning CRISPR system. The proportion of cells undergoing GFP loss upon incorporation of a particular gRNA divided by the proportion of cells undergoing GFP loss upon incorporation of GFP-targeting positive control gRNA is plotted against the actual genomic location of the gRNA. Negative controls defined as gRNAs showing no reads in either GFP^{neg} or GFP^{medium} populations but present in the bulk population are highlighted in red. Error bars indicate experimental variability in two replicates. (**c**) Correlation of gRNAs significantly enriched in the GFP^{neg} population in fixed-size bins varying from 100 bp to 1 kb for biological replicates in *Tdgf1* libraries. (**d**) Fraction of GFP^{neg}-enriched gRNAs among the different functional genomic categories surrounding the *Tdgf1* gene. Error bars show variability due to the three biological replicates.

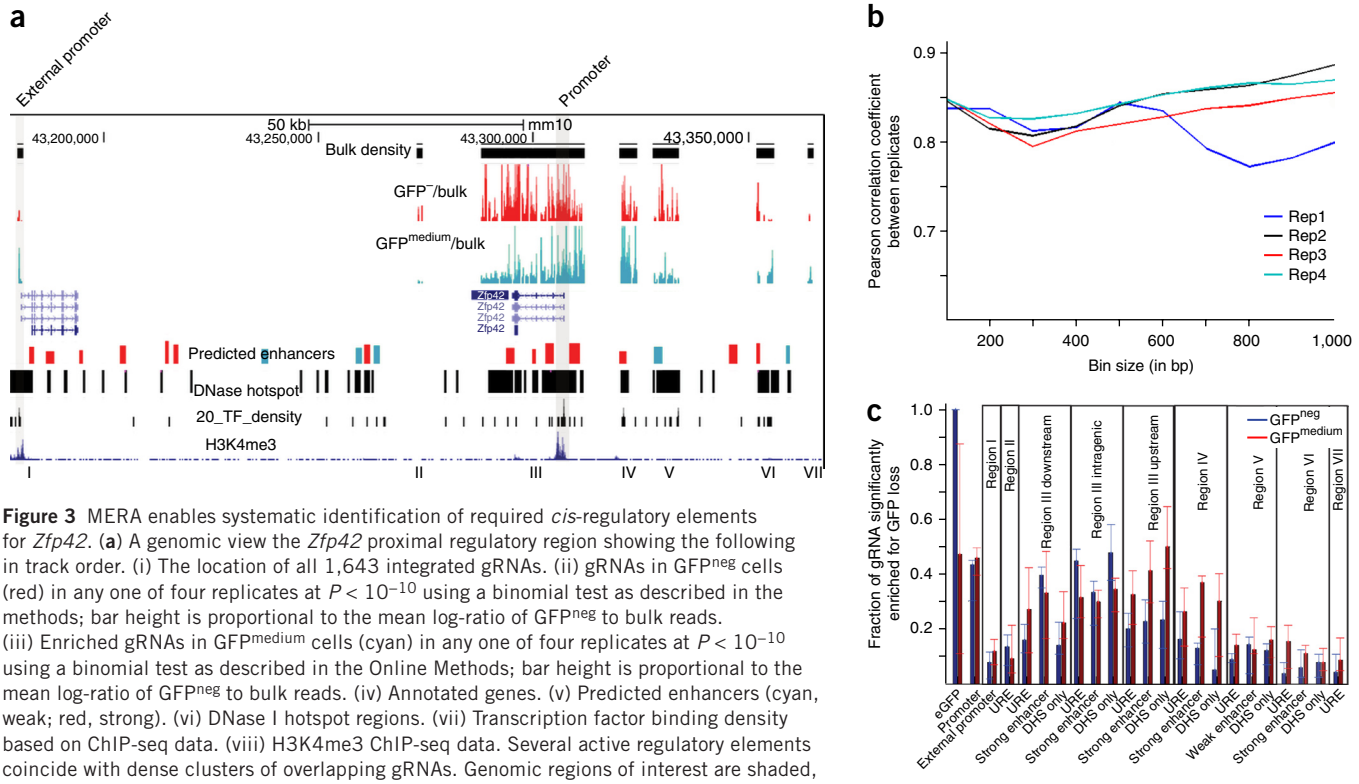

Figure 3 MERA enables systematic identification of required *cis*-regulatory elements for *Zfp42*. (**a**) A genomic view the *Zfp42* proximal regulatory region showing the following in track order. (i) The location of all 1,643 integrated gRNAs. (ii) gRNAs in GFP^{neg} cells (red) in any one of four replicates at $P < 10^{-10}$ using a binomial test as described in the methods; bar height is proportional to the mean log-ratio of GFP^{neg} to bulk reads. (iii) Enriched gRNAs in GFP^{medium} cells (cyan) in any one of four replicates at $P < 10^{-10}$ using a binomial test as described in the Online Methods; bar height is proportional to the mean log-ratio of GFP^{neg} to bulk reads. (iv) Annotated genes. (v) Predicted enhancers (cyan, weak; red, strong). (vi) DNase I hotspot regions. (vii) Transcription factor binding density based on ChIP-seq data. (viii) H3K4me3 ChIP-seq data. Several active regulatory elements coincide with dense clusters of overlapping gRNAs. Genomic regions of interest are shaded, annotated above the plot and described in further detail in the text. (**b**) Correlation of gRNAs significantly enriched in the GFP^{neg} population in fixed-size bins varying from 100 bp to 1 kb for biological replicates in *Zfp42*. (**c**) Fraction of GFP^{neg} enriched gRNA among the different functional genomic categories surrounding the *Zfp42* gene. Error bars show variability due to the four biological replicates.

The highest density of significant gRNAs in the genomic regions was observed at the promoter region for *Tdgf1*, the strong proximal enhancer 4 kb upstream of *Tdgf1* and the strong enhancer overlapping the *Lrrc2* promoter (**Fig. 2a**,**d**).

Surprisingly, we observed a novel class of genomic elements downstream of *Tdgf1* (**Fig. 2a**, highlighted in gray) that did not coincide with any known markers of regulatory activity, such as H3K27ac, H3K4me1, H3K4me3, known transcription factor (TF)-binding sites, DNase I hypersensitivity sites, predicted DNase I hotspots, or enhancers predicted from chromatin modifications. We designated such elements that do not contain any of these markers as unmarked regulatory elements (UREs). UREs were often over 1 kb in length and produced a loss of GFP comparable to that induced by some distant enhancers (**Fig. 2d**).

In our MERA screen of *Zfp42*, we also observed the strongest enrichment for GFP loss in the promoter and proximal enhancer regions (**Fig. 3a**,**c**). We observed enrichment of gRNAs in the GFP^{neg} and GFP^{medium} population at UREs in regions II, III, VI and VII (**Fig. 3a** and **Supplementary Fig. 6a**) and observed the participation of the neighboring *Triml2* promoter in regulating *Zfp42* (**Fig. 3a** and **Supplementary Fig. 6b**). We also note that regulatory regions upstream of *Zfp42* tended to cause intermediate rather than complete loss of GFP (GFP^{medium} in red versus GFP^{neg} in blue; **Fig. 3c**), suggesting that these enhancers are each responsible for only part of the overall *Zfp42* expression level in cells.

Validation of MERA hits

To determine the accuracy of the MERA screen in systematically determining required *cis*-regulatory regions, we first examined replicate consistency among our *Tdgf1*, *Zfp42* and *Nanog* MERA data. Spatial patterns of GFP^{neg} gRNA enrichment were largely conserved between replicates, with Pearson correlation values of 0.8 at a 300-bp bin size (**Figs. 2c** and **3b**, **Supplementary Fig. 6c**). At an individual level, the overlap between gRNAs enriched in GFP^{neg} populations between replicates was significant for all replicates (hypergeometric *P* value <0.001); however, it was not as high as for binned regions, likely because a single gRNA can cause thousands of distinct mutant genotypes with varying phenotypes.

To analyze false positives caused by off-target effects, we examined how putative off-target effects affect MERA results using a model based on GUIDE-Seq[32] (Online Methods, **Supplementary Fig. 7**). We found that when we eliminated gRNAs with potential off-target effects from our analysis, the global distribution of significantly enriched gRNAs along the regulatory landscape of the gene was unaltered and relative contributions of different functional categories were unaffected (**Supplementary Figs. 5a** and **6a**,**c**). Furthermore, several gRNAs with no predicted off-target effects supported the regulation of *Tdgf1* by the promoter of *Lrrc2* (**Supplementary Fig. 5b**), the promoter of *Triml2* and a URE region (**Supplementary Fig. 6a**–**c**), and none of these regions was more likely to contain off-target effects than other screened regions.

To analyze potential off-target effects with an independent method, we asked whether any gRNAs from the *Tdgf1* library would extinguish *Zfp42*^{GFP} activity and vice versa. We found that a much smaller percentage of cells lost GFP upon targeting by a mismatched gRNA library than upon targeting by the matched library (**Supplementary Fig. 8**). Sequencing revealed that the gRNAs enriched in GFP^{neg} mismatched library–targeted cells were predominantly GFP control gRNAs, with a small number of non-clustered gRNAs displaying off-target activity (**Supplementary Figs. 5** and **6**). Thus, the clustered

enrichment of GFP loss at enhancers, neighboring promoters and UREs in MERA is not replicated by computationally predicted or experimentally determined off-target effects, leading us to conclude that GFP loss in these regions is a result of on-target gRNA effects (**Supplementary Figs. 5a–c** and **6a**,**b**).

To determine the false-positive rate at the level of individual gRNAs, we introduced individual gRNAs to determine whether their rate of GFP loss correlated with their activity in the pooled MERA screen. These gRNAs fell within several functional categories, including UREs and neighboring promoters (**Fig. 2a**, highlighted in gray, and **Fig. 2b**). We confirmed significantly increased GFP loss in 29/30 gRNAs from these screens as compared to 5 similarly located control gRNAs (**Fig. 2b**). Altogether, we conclude that MERA has a low false-positive rate.

We next sought to determine the false-negative rate of MERA. As opposed to ORF-targeting screens, in which all gRNAs are assumed to be equivalently likely to induce frameshift mutations that inactivate gene function, we found that regulatory mutations induce more variable phenotypes with regard to gene expression (see **Supplementary Discussion**). In our individual follow-up assays, we found that gRNAs targeting the *GFP* ORF induced GFP loss in >90% of cells, those targeting promoter regions induced GFP loss in 20–40% of cells and those targeting distal regulatory elements induced GFP loss in 5–40% of cells, while negative controls induce GFP loss in <2% of cells (**Fig. 2b**). We assert that this phenotypic diversity results from the wide spectrum of mutations at target sites, which are differentially likely to disrupt functional regulatory elements such as transcription factor–binding sites. We confirm this hypothesis in several cases by performing functional motif discovery, described later in the text.

To assess the false-negative rate of MERA gRNAs, we examined regions in our data with strong likelihood of inducing GFP loss. We found that 10/10 GFP-targeting gRNAs in all four GFP lines were highly enriched in GFP^{neg} cells (**Figs. 2d** and **3c**). Additionally, 67/83 (81%) gRNAs that target the first 700 bp of the *Rpp25* ORF were highly enriched in GFP^{neg} cells. In 500 bp around the *Tdgf1* promoter region, 48/59 (81%) of gRNAs induce GFP loss in multiple replicates (**Supplementary Fig. 4f**). Thus, a high percentage of gRNAs expected to have an effect on gene expression were enriched in GFP^{neg} cells. It is unclear whether the 20% of gRNAs in these regions that do not induce GFP loss are false negatives or true negatives, as their mechanism of inducing GFP loss is not as direct as when the GFP ORF itself is targeted. However, even if this appreciable percentage of individual gRNAs are false negatives, it does not impair the ability of MERA to determine required regulatory regions, as the high density of gRNAs in a region (~1 per 8 bp) allows highly reproducible resolution at the level of 100-1,000 bp (**Figs. 2c** and **3b**). We then asked whether annotated regulatory regions are necessary for gene function. An appreciable percentage of gRNAs induced significant GFP loss at 9/9 of *Tdgf1* predicted enhancers (±20 kb around *Tdgf1*) and 6/7 of predicted *Zfp42* enhancers (−21 to +45 kb around *Zfp42*) (**Supplementary Tables 6** and **7**). However, there was substantial heterogeneity in the percentage of gRNAs within an enhancer that induce GFP loss, and some DNase-hypersensitive sites without enhancer histone modifications contain a high fraction of GFP loss-inducing gRNAs (**Supplementary Tables 6** and **7**), indicating that enhancer histone modifications do not entirely predict required regulatory regions. We cannot rule out the possibility that certain regions may suffer from systematic inefficiencies in gRNA targeting.

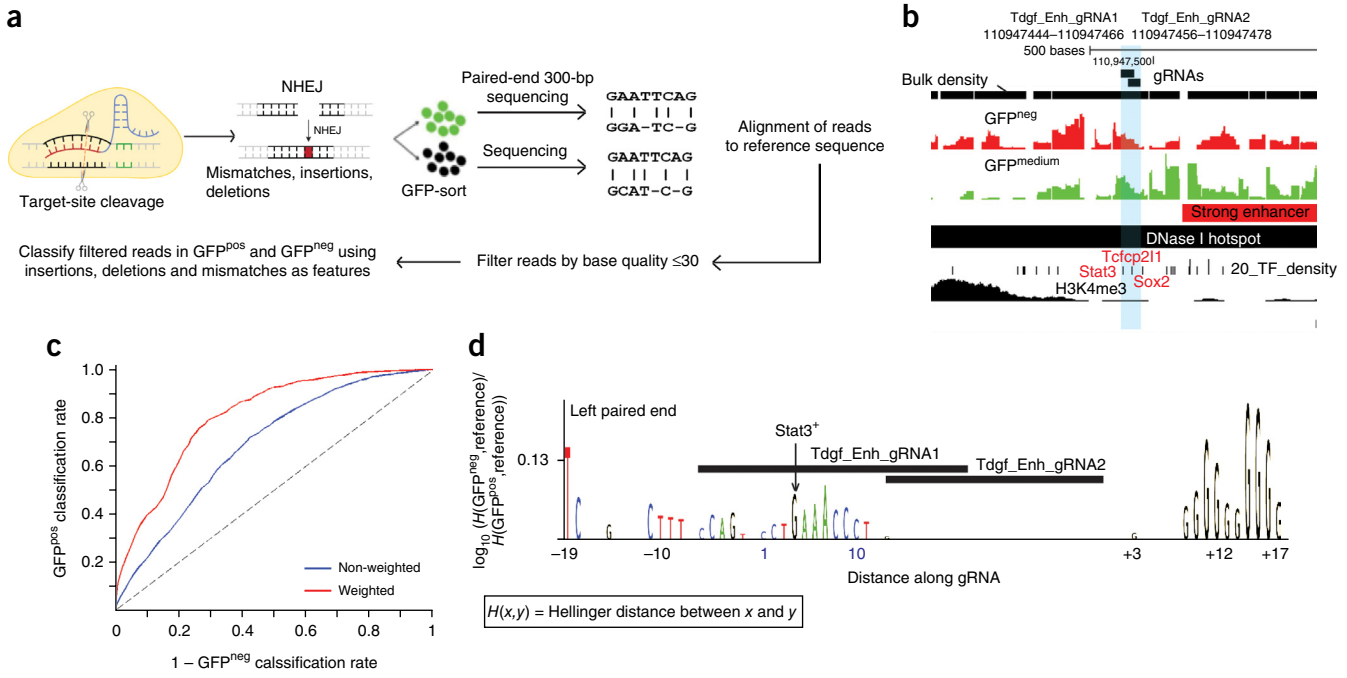

Figure 4 Functional motif discovery analysis of region-specific mutant genotypes at enhancers reveals required regulatory motifs. (**a**) A schematic of the procedure involved in finding mutations induced by a particular gRNA. (**b**) Plot showing the genomic regions surrounding two gRNAs at a proximal *Tdgf1* enhancer region (gRNAs are shaded) showing overlap with DNase I hotspot and predicted enhancer regions, and transcription factor binding sites for Stat3, Tcfcp2l1 and Sox2. (**c**) ROC curve for fivefold classification of GFP^{neg} and GFP^{pos} genotypes using mutations within −20 to +20 bp of the gRNA along left and right paired-end reads as features. (**d**) Motif logo for region mutated by gRNAs with base scores computed as log-ratios of the Hellinger distance of the GFP^{neg} genotypes at a base to the reference base to the Hellinger distance of the GFP^{pos} genotypes at a base to the reference base, caused by Tdgf_gRNA_1 and Tdgf_gRNA_2 along the left paired end read. The location of the Stat3 binding site with its positive-strand motif is shown along the length of the gRNA.

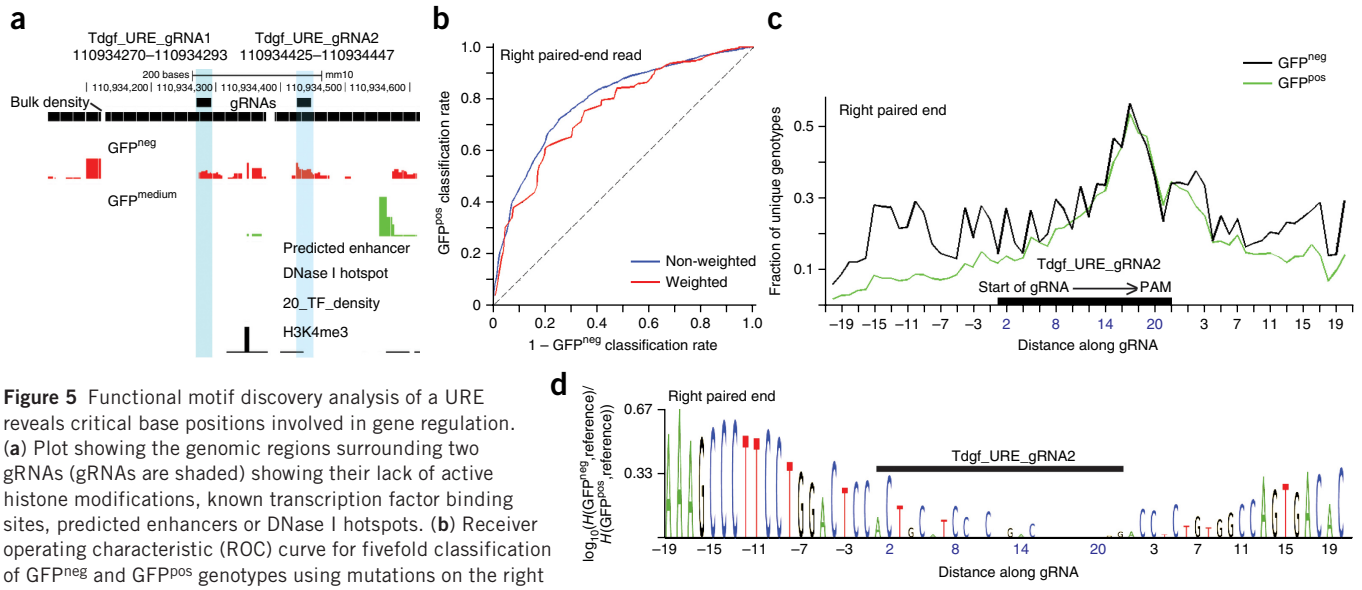

Figure 5 Functional motif discovery analysis of a URE reveals critical base positions involved in gene regulation. (**a**) Plot showing the genomic regions surrounding two gRNAs (gRNAs are shaded) showing their lack of active histone modifications, known transcription factor binding sites, predicted enhancers or DNase I hotspots. (**b**) Receiver operating characteristic (ROC) curve for fivefold classification of GFP^{neg} and GFP^{pos} genotypes using mutations on the right paired-end read within –20 to +20 bp of Tdgf_URE_gRNA2. Unweighted classification (in blue) counts each unique genotype in the test set only once, while weighted classification (red) counts each unique genotype in the test set as many times as the number of reads assigned to it, for calculating sensitivity and specificity. (**c**) Fraction of unique genotypes in GFP^{neg} and GFP^{pos} populations with mutations at bases along the right paired-end read reveals pattern of cleavage around Tdgf_URE_gRNA2. (**d**) Motif logo for the region mutated by Tdgf_URE_gRNA2 along the right paired-end read with base scores computed as log-ratios of the Hellinger distance of the GFP^{neg} genotypes at a base to the reference base to the Hellinger distance of the GFP^{pos} genotypes at a base to the reference base.

Gene regulatory trends emerging from MERA screens

Our MERA results revealed that *Tdgf1*, *Nanog*, *Rpp25* and *Zfp42* have different regulatory architectures (**Figs. 2** and **3**, **Supplementary Figs. 5**, **6** and **9**). All regulatory regions within ±20 kb of the *Nanog* promoter were associated with clusters of highly enriched gRNAs, and 20–40% of the tested gRNAs in predicted enhancers and DNase I hotspots proximal to *Nanog* resulted in GFP^{neg} cells (**Supplementary Fig. 9c**). In contrast, the *Rpp25* gene shows a dense concentration of significant gRNAs at its promoter and short ORF region. Other proximal regulatory regions of *Rpp25* had 12% of tested gRNAs resulting in GFP^{neg} cells (**Supplementary Fig. 9d**). *Tdgf1* shows a similar trend to *Nanog*, with dense clusters of significant gRNAs in the proximal regulatory regions (**Fig. 2a**,**d**). UREs were also seen in *cis*-regulatory regions near *Rpp25* (**Supplementary Fig. 9b**). In *Nanog*, a distal ChIA-PET region >92 Mb away showed several strongly enriched gRNAs, whereas three other distal ChIA-PET regions showed no strongly enriched gRNAs (**Supplementary Fig. 9a**), indicating that MERA is capable of measuring the functionality of long-distance chromatin interactions.

One observation common to all genes is the participation of the promoters of other genes in regulation. In some cases these gene promoters are several million bases away. Examples of foreign promoter involvement can be seen in the cases of the *Lrrc2* promoter in *Tdgf1* (**Fig. 2a**,**d**), *Mirc35hg* in *Nanog* (**Supplementary Fig. 9a**) and *Scamp5* and *Cox5a* in *Rpp25* (**Supplementary Fig. 9b**). Previous studies have documented the existence of dual-property elements[33] that can act as either promoter or enhancer in different cellular contexts. Additionally, it is known that neighboring promoters often interact with each other[34] and that expression of neighboring genes is often coordinated[35]. Here we observe that active promoters may coordinate gene expression patterns of neighboring genes by functioning as enhancers within the same cellular context.

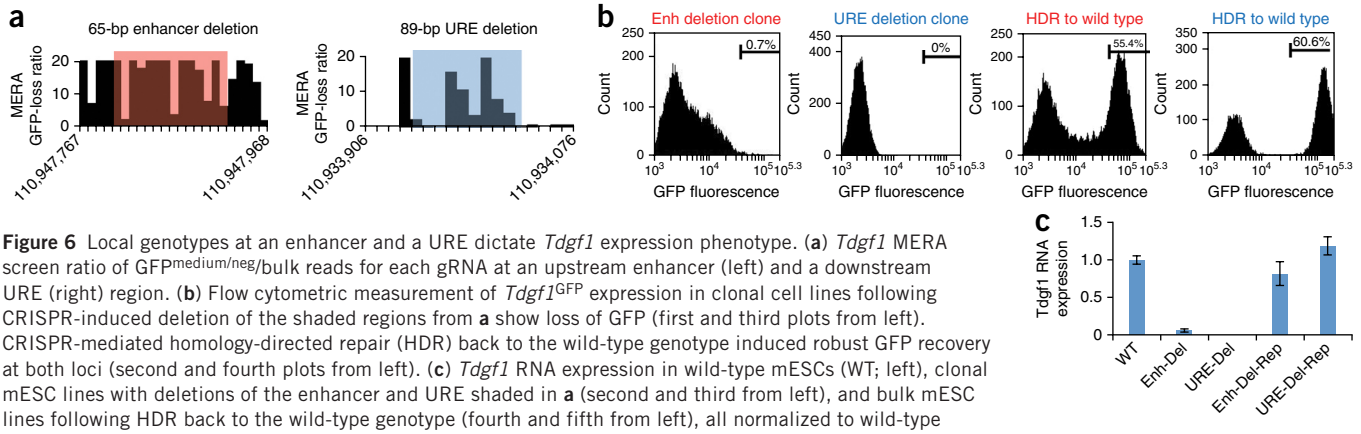

Figure 6 Local genotypes at an enhancer and a URE dictate *Tdgf1* expression phenotype. (**a**) *Tdgf1* MERA screen ratio of GFP$^{medium/neg}$/bulk reads for each gRNA at an upstream enhancer (left) and a downstream URE (right) region. (**b**) Flow cytometric measurement of *Tdgf1*^{GFP} expression in clonal cell lines following CRISPR-induced deletion of the shaded regions from **a** show loss of GFP (first and third plots from left). CRISPR-mediated homology-directed repair (HDR) back to the wild-type genotype induced robust GFP recovery at both loci (second and fourth plots from left). (**c**) *Tdgf1* RNA expression in wild-type mESCs (WT; left), clonal mESC lines with deletions of the enhancer and URE shaded in **a** (second and third from left), and bulk mESC lines following HDR back to the wild-type genotype (fourth and fifth from left), all normalized to wild-type expression level in two replicates.

Functional motif discovery to examine MERA-predicted regulatory regions

The second phase of MERA uses functional motif discovery to identify the causal elements governing expression at MERA screen hits. Because Cas9 induces random mutations, a pool of mESCs treated with Cas9 and a single gRNA will contain thousands of distinct mutant genotypes centered on the gRNA cleavage site. Recently, TAL effector nucleases have been used to derive functional footprints of regulatory DNA[36]. We hypothesized that we could pinpoint DNA sequence motif(s) that cause GFP loss by identifying sequence features that consistently differ between thousands of GFP^{pos} and GFP^{neg} genotypes at a given site (**Fig. 4a**). Functional motif discovery is achieved by performing individual Cas9-mediated mutation by a selected gRNA, obtaining thousands of genotypes from both GFP^{pos} and $GFP^{medium/neg}$ cells by high-throughput sequencing, and then summarizing the observed genotypes as motifs that reveal which bases are important for gene expression (**Fig. 4a**, Online Methods). Using the differences in fractions of genotypes at positions along the gRNA, we defined a base-level importance score that was independent of the cutting biases of the gRNA and built a random-forest[37] classifier to gauge the accuracy of distinguishing GFP^{neg} or GFP^{pos} genotypes using base-level features (Online Methods).

We first tested to see whether functional motif discovery in *Tdgf1* and *Zfp42* enhancer regions would permit us to classify genotypes held out of initial algorithmic training as GFP^{neg} or GFP^{pos}. We selected two overlapping gRNAs for functional motif discovery in a *Tdgf1* proximal enhancer that overlapped binding sites for the key mESC transcription factors Stat3, Sox2 and Tcfcp2lI, of which Stat3 is the only factor with a direct binding site. Using mutations constrained between −20 and +20 bp of the gRNA (**Supplementary Fig. 10**), we were able to classify held out genotypes with an area under the curve (AUC) of 0.81 (**Fig. 4c**), and we observed an enrichment of the bases for the Stat3 motif[25] in both the left and right paired-end reads (**Fig. 4d** and **Supplementary Fig. 11e**). We achieved similar success at *Zfp42* enhancer sites, identifying required bases around Nrf1 and p300 binding sites (**Supplementary Figs. 12** and **13**).

We next applied functional motif discovery to two gRNAs in a URE ~12 kb downstream of the *Tdgf1* transcript (**Fig. 5a**). We obtained high classification accuracy for held-out genotypes from both gRNAs (AUC 0.81 and 0.76, **Fig. 5b** and **Supplementary Fig. 14c**), and we observed blocks of consecutive bases whose deletion correlated with GFP loss (**Fig. 5c**,**d** and **Supplementary Fig. 15d**,**e**), suggesting focal regions of the genome that are required for URE function. Altogether, we conclude that functional motif discovery is a valuable method for ascertaining which bases at MERA-identified regulatory regions are required for gene expression. In enhancer regions, these bases correspond to known binding motifs, and in UREs, we identify blocks of bases that are required for gene expression.

We then used homologous recombination to confirm that the *Tdgf1* enhancer and URE regulatory elements are truly required for gene expression in the third phase of MERA. We used flanking gRNAs to induce short (>100-bp) deletions in two regions predicted to induce GFP loss by our MERA screen, one in the *Tdgf1* enhancer and one at a URE. As expected, a subset of cells lost GFP expression, and we obtained clonal GFP^{neg} lines containing the deletion genotype (**Fig. 6a**,**b**). We then used homology-directed repair to restore the wild-type genotype in these cells, finding at each site that a large percentage of cells reverted to a GFP^{pos} state (**Fig. 6b**). We replicated this experiment in wild-type cells without a $Tdgf1^{GFP}$ allele, finding that clonal deletion cells lost *Tdgf1* RNA expression, and clonal repaired lines restored *Tdgf1* expression (**Fig. 6c**). This robust and straightforward relationship between local genotype and GFP expression provides compelling evidence that the local DNA sequence at a URE is required for *Tdgf1* expression.

DISCUSSION

MERA offers an unbiased, high-resolution approach to directly interrogate the function of the regulatory genome. It not only provides a survey of required *cis*-regulatory elements but also enables functional motif discovery to dissect the precise nature of identified regulatory elements. We find evidence that neighboring gene promoters as well as unmarked regulatory elements (UREs) that are not associated with conventionally expected DNase hypersensitivity and histone mark features play unexpectedly large roles in controlling gene expression. This observation reinforces the importance of direct perturbation analysis to definitively characterize genome function, as we observe that correlative genome annotation does not fully predict regulatory requirement.

Although we do not yet have definitive data as to the function of UREs, we find that a URE downstream of the *Tdgf1* gene is highly sensitive to base substitution at a string of consecutive bases, suggesting that its DNA sequence is crucial to its regulatory activity. Furthermore, we find the first half of this URE to be highly conserved (phastcons score >0.85, **Supplementary Fig. 15e**), indicating potential functional significance of the genomic region. Consistent with these data, UREs may be RNA templates, elements bound by uncharacterized protein factors, or spacers whose precise base sequence is of secondary importance. We cannot exclude the possibility that UREs are active only in a cellular subpopulation and thus conventionally expected DNase hypersensitivity and histone mark features are not detected when the entire cellular population is assayed.

We designed our gRNA libraries to target a mix of previously annotated and unannotated *cis*-regulatory regions, and thus we did not uniformly tile the proximal regions of any of these genes. Therefore, we cannot estimate the frequency of UREs, and we expect that future MERA screens with even more extensive coverage at more loci will elucidate how pervasive UREs and neighboring gene promoters are in the regulatory architecture of the genome.

MERA is complementary to high-throughput reporter assays, and future experiments including both approaches should provide insight into the degree of concordance between necessary and sufficient gene regulatory elements. MERA also enables quantitative assessment of the relative effects of distinct *cis*-regulatory elements on gene expression and could potentially provide insights into how regulatory regions combine to achieve desired levels of expression. We note that lentiviral delivery can be used to expand the range of cell types that can be analyzed by MERA. Extending MERA to explore how changes in individual *cis*-regulatory elements alter gene networks will aid our understanding of how *cis*-regulatory variants lead to human disease. We expect that the direct interrogation of variant locations discovered in genome-wide association studies by MERA will provide a rapid way to screen such variants for function in relevant cell types.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Data have been submitted to the NCBI GEO database under accession code GSE76318.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank F. Zhang (Broad Institute of MIT and Harvard) for reagents, the MIT BiomicroCenter for high-throughput sequencing assistance, and Y. Qiu for flow cytometric assistance. The authors acknowledge funding from the US National Institutes of Health to D.K.G. (1U01HG007037) and to R.I.S. (1K01DK101684-01) and Harvard Stem Cell Institute's Sternlicht Director's Fund award, Brigham and Women's Hospital BRI Innovation Fund, and Human Frontier Science Program grants to R.I.S.

AUTHOR CONTRIBUTIONS

Experiments were designed by N.R., R.I.S. and D.K.G. MERA experiments were performed by R.S., S.S., K.K., B.B. and B.J.M.E. N.R. and D.K.G. performed the computational analysis. Y.G., T.S. and M.D.E. helped with the computational analysis.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Jenuwein, T. & Allis, C.D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
2. Bernstein, B.E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
3. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
4. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
5. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA* **107**, 21931–21936 (2010).
6. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
7. Arnold, C.D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
8. Patwardhan, R.P. *et al.* Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nat. Biotechnol.* **30**, 265–270 (2012).
9. Fullwood, M.J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
10. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
11. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006).
12. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006).
13. Kwasnieski, J.C., Fiore, C., Chaudhari, H.G. & Cohen, B.A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **24**, 1595–1602 (2014).
14. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
15. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
16. Zhou, Y. *et al.* High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* **509**, 487–491 (2014).
17. Koike-Yusa, H., Li, Y., Tan, E.P., Velasco-Herrera, Mdel.C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.* **32**, 267–273 (2014).
18. Chen, S. *et al.* Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. *Cell* **160**, 1246–1260 (2015).
19. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
20. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
21. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
22. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *eLife* **2**, e00471 (2013).
23. Cradick, T.J., Fine, E.J., Antico, C.J. & Bao, G. CRISPR/Cas9 systems targeting β-globin and CCR5 genes have substantial off-target activity. *Nucleic Acids Res.* **41**, 9584–9592 (2013).
24. Arbab, M., Srinivasan, S., Hashimoto, T., Geijsen, N. & Sherwood, R.I. Cloning-free CRISPR. *Stem Cell Reports* **5**, 908–917 (2015).
25. Young, R.A. Control of the embryonic stem cell state. *Cell* **144**, 940–954 (2011).
26. Yue, F. *et al.* Mouse ENCODE Consortium. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
27. Rajagopal, N. *et al.* RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput. Biol.* **9**, e1002968 (2013).
28. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
29. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).
30. Sherwood, R.I. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* **32**, 171–178 (2014).
31. Guo, Y., Mahony, S. & Gifford, D.K. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.* **8**, e1002638 (2012).
32. Tsai, S.Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
33. Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**, 350–354 (2015).
34. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
35. Woo, Y.H., Walker, M. & Churchill, G.A. Coordinated expression domains in mammalian genomes. *PLoS One* **5**, e12158 (2010).
36. Vierstra, J. *et al.* Functional footprinting of regulatory DNA. *Nat. Methods* **12**, 927–930 (2015).
37. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

ONLINE METHODS

Library design for MERA. In addition to 10 GFP-targeting gRNAs, we designed 3,908 gRNAs specific to each of the four libraries for TDGF, *Nanog*, *Zfp42* and *Rpp25*. For TDGF we selected a −20 kb to +20 kb proximal region around the TDGF promoter to profile 3908 gRNAs that were designed for this region. For *Nanog*, *Rpp25* and *Zfp42*, we prioritized the design of 3,908 gRNAs based on regions of strong DNase I enrichment going up to 100 kb on either side of the gene promoter. Further, we used PolII ChIA-PET data to find distal regions that are predicted to interact with the promoter. In case of a large number of ChIA-PET regions, we filtered interactions based on other enhancer features such as p300 binding, DNase I enrichment, active histone modifications etc. overlapping distal ChIA-PET regions.

We used the following algorithm to design gRNAs:

1. Determine region of interest for guide RNA design.
2. Find all GG sequences on both the forward and reverse strand.
3. Design guide RNA in the following format. Guide RNAs should have 19–20 bp of homology to the genome immediately preceding the NGG "PAM" sequence:
   a. If the genome sequence is GNNNNNNNNNNNNNNNNNN NGG (GN_{19}NGG), the guide RNA sequence should be GNNNNNNNNNNNNNNNNNN (GN_{19});
   b. If a is not satisfied but GNNNNNNNNNNNNNNNNNN NGG (GN_{18}NGG) is satisfied, the guide RNA sequence should be GNNNNNNNNNNNNNNNNNN (GN_{18});
   c. If a and b are not satisfied, the guide RNA sequence should be GNNNNNNNNNNNNNNNNNNN (GN_{20}) where the genomic sequence is NNNNNNNNNNNNNNNNNNNN NGG (N_{20}NGG)—it does not matter if the first G is in the genome.
4. We filtered adjacent gRNAs shifted by just 1 bp until we were able to achieve the 3908 gRNAs required to tile the region.
5. Libraries were ordered as 98- to 100-bp sequences containing a 19- to 20-bp protospacer targeting the genomic sequence of interest, an optional G if the protospacer does not already begin with one, and surrounding sequences homologous to the U6 promoter and gRNA hairpin. We ordered gRNA libraries of 3,918 members from LC Sciences.

TTATATATCTTGTGGAAAGGACGAAACACC[$GN_{18\text{-}20}$]GTTTAAGAG CTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGT

All libraries contained 10 gRNAs targeting the GFP open reading frame to serve as positive controls (**Supplementary Methods**).

Mapping of MERA reads. We mapped the sequence composing of sample barcode, primer and exact matches of the designed gRNA sequence to the sequenced reads. Counts for each gRNA for either GFP^{neg}, GFP^{medium} or bulk populations were obtained by counting the number of sequenced reads that showed exact matches to the gRNA.

The gRNA integration rate into cellular genomic DNA was found to be 93% for *Tdgf1* but appeared to be only 43% for *Zfp42*. In order to determine if this was caused by inefficient integration or due to synthesis errors, we sequenced the gRNA library for *Zfp42* and found that only 1,723 of the 3,919 guide RNAs in the *Zfp42* library were synthesized accurately. Among these, 1,718/1,723 were detected in the bulk library of at least one replicate. Hence, we estimate that the integration rate of gRNAs is >90% of those that are synthesized. Oligonucleotide library synthesis quality is unaffected by whether a gRNA integration approach such as MERA or a lentiviral cloning approach is taken, and thus MERA enables integration of the vast majority of available gRNAs.

Identification of gRNAs that are significantly enriched in GFP^{neg} and GFP^{medium} populations. In order to detect gRNAs with statistically significant overrepresentation in GFP^{neg} and GFP^{medium} populations, we perform a step-wise procedure.

Step 1. We normalize the gRNA sequence read counts, which can vary between sequencing runs of bulk, GFP^{medium} and GFP^{neg} populations due to differences in cell number and diversity of the respective populations (**Supplementary Fig. 4b**,**c**, *x*-axis versus *y*-axis limits) . In order to normalize these read ranges, we assume that the positive control gRNAs targeting the GFP coding region always induce loss of GFP expression, which is consistent with our previous results showing that over 99% of cells receiving a GFP-targeting gRNA lose GFP expression[24]. In addition, GFP^{neg} and to a lesser extent GFP^{medium} reads are always observed to be proportional to the bulk reads for the GFP targeting gRNAs, to a much greater extent than for all guide RNAs (**Supplementary Fig. 4c**,**d** and **Supplementary Table 5**). Hence, we predict the number of GFP^{neg} reads we would see for each gRNA given its bulk and GFP^{medium} count if it always caused GFP loss. In order to do this, we build two different kinds of linear models depending on the data available

I. In case of *Tdgf1*^{GFP} and *Zfp42*^{GFP}, we have a GFP^{medium} as well as GFP^{neg} population, along with 3 to 4 biological replicates per cell-line. We assume that for any GFP-targeting gRNA, the majority of bulk reads are derived from the GFP^{neg} population. However, each gRNA may also cause some intermediate loss of GFP due to variable mutations or imperfect sorting. In addition, there is a low gRNA-dependent intercept or GFP^{pos} population, which may be a small fraction of mutations induced by a particular gRNA that do not cause GFP-loss.

In order to transform the bulk reads to the GFP^{neg} scale, we model GFP^{neg} as the dependent variable, and GFP^{medium} and bulk reads as independent variables using a generalized linear model[38]. The intercept is modeled as being dependent on the gRNA but independent across replicates, while the slopes are considered as having a replicate-dependent component also.

The model is of the form

$$y \sim x1 + x2 + (z11 \mid g1) + (x1 \mid g2) + (x2 \mid g2)$$

where y = GFP^{neg}, x1 = bulk, x2 = GFP^{medium}, z11 = intercept, g1 = grouping by gRNA, g2 = grouping by replicate.

In order to transform the bulk reads to the GFP^{medium} scale, we use the same model but with y = GFP^{medium}, x2 = GFP^{neg}.

II. In case of *Nanog*^{GFP} and *Rpp25*^{GFP}, we have only a GFP^{neg} population and at most 2 replicates. In this case we build an independent linear regression model for each replicate of the form:

$$y \sim x1 + z11$$

where y = GFP^{neg}, x1 = bulk, z11 = intercept.

Using the linear regression models, we now transform all bulk reads to either GFP^{neg} or GFP^{medium} populations, depending on if we are interested in finding gRNAs enriched in GFP^{neg} or GFP^{medium} populations respectively.

Step 2. We now use the fact that since the dummy gRNA (negative control) should not occur in GFP^{neg}/GFP^{medium} cells any reads corresponding to this gRNA in the GFP^{neg}/GFP^{medium} population are due to random chance. Hence, we can obtain the null probability of observing reads in the GFP^{neg}/GFP^{medium} population by dividing the GFP^{neg}/GFP^{medium} reads for the dummy gRNA by the number of bulk reads for the dummy gRNA transformed to the GFP^{neg}/GFP^{medium} scale. We then use a binomial distribution to calculate significance for a gRNA based on this null probability, with the gRNA's observed number of GFP^{neg}/GFP^{medium} reads as the number of successes, and the number of bulk-transformed reads for the gRNA as the number of trials.

Data sets for comparison and visualization with enriched gRNA. The UCSC genome browser[39] was used to visualize the data and create genomic view snapshots for regulatory regions of various genes.

Enhancer predictions. The enhancer predictions were made using the RFECS method[27] using 6 histone modifications from ENCODE[28] trained on p300 binding site data from mouse embryonic stem cells. Enhancers were separated into "strong" and "weak" categories based on presence of H3K27ac at levels greater than input. Further boundaries of enhancers were called using a Sobel edge-detection algorithm implemented in MATLAB. Edges were identified for an input subtracted RPKM (reads per kilobase per million) – normalized H3K27ac reads[27] in the case of strong enhancers and RPKM-normalized H3K4me1 reads for weak enhancers.

DNase I hotspot. We used the DNase-seq data set previously generated[30] and called hotspots using a standard hotspot algorithm[29].

TF density. The GEM algorithm[31] was applied to transcription factor ChIP-seq data sets for the following transcription factor: Nanog, Oct4, Sox2, TCF3, p300, CTCF, Smc1, Smad3, c-Myc, Med12, Med1, CTCF, E2F1, Esrrb, Klf4, n-Myc, Nr5a2, Tcfcp2l1, Stat3, Zfx.

Analysis of deep sequencing data sets. Individual scCRISPR-mediated mutation by a selected gRNA was performed in a large pool of cells to create tens of

thousands of unique mutated genotypes at the site. We then flow cytometrically sorted GFP^{pos} and GFP$^{medium/neg}$ populations and performed 150-bp paired-end sequencing on regions surrounding each targeted site to obtain genotypic data on thousands of mutated regions that did and did not induce loss of GFP expression (**Fig. 4a**). Deep-sequencing data sets were filtered for sequence quality by using a minimum base quality filter of 30. After stripping barcodes, the length of each paired-end read was 145 bp. We aligned these 145-bp long genotypes to the reference genotype extended by 30 bp downstream (total of 175 bp). Alignment of sequenced reads to the reference genome was performed using the semi-global version of the Needleman-Wunsch algorithm[40] with a gap opening penalty of 8 and gap extension penalty of 4. The command in MATLAB used was:

```
nwalign(Reference_Seq,
Genotype_seq,'alphabet','NT','gapopen',8,'ExtendGap',4,'glocal','true');
```

Functional motif discovery. After globally aligning and filtering reads for sequence quality (per base quality ≥30), mismatches, deletions and insertions were counted with respect to the base position in the reference. We observed long stretches of mutations with combinations of mismatches and deletions . Hence, we defined a "length of disruption" as a continuous series of mutations with maximum intervening matches of <5 bases. We plotted the left and right ends of these disruptions and observed that the majority of disruptions originated within the gRNA, as expected, with very few short mutations lying outside that could be assumed to be one or two base sequencing errors (**Supplementary Figs. 10** and **14a**,**b**). While a majority of disruptions extending beyond the ends of the guide RNA were enriched for GFP^{neg} (**Supplementary Fig. 10a**–**d**, yellow-red versus blue), we also observed a mixed population of GFP^{neg} as well as GFP^{pos} deletions lying within −20/+20 bp of the gRNA. Since we wish to assess the local effect of the gRNA on GFP-loss, we limited further analysis to genotypes with disruptions that originate within the gRNA and do not go beyond 20 bp of the gRNA.

Restricting our analysis to these genotypes, we observed increased mutation around the gRNA cleavage site in both GFP^{pos} and GFP$^{medium/neg}$ populations (**Supplementary Figs. 11a**,**b** and **13a**,**b**). Mismatch, deletion, and insertion mutations were all observed, with deletions predominating in the GFP$^{medium/neg}$ genotypes (**Supplementary Figs. 11a**,**b** versus **11c**,**d**, **Supplementary Figs. 13a**,**b** versus **13c**,**d**).

In order to develop a base-level motif logo, we defined a base-level score representing the deviation of GFP^{neg} population from reference as compared to the deviation of the GFP^{pos} population from reference. In order to find the distance of a base from reference, we used the Hellinger measure[41] for finding the distance between two discrete distributions:

$$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}\left(\sqrt{p_i} - \sqrt{q_i -}\right)^2},$$

Here, we had five possible values per base which were the frequency of occurrence of each base type (A, C, T, G) and a fifth deletion (D). The motif score at any base was defined as:

$$\text{base score} = \log_{10}\left(H(\text{GFP}^{neg},\text{reference})/H(\text{GFP}^{pos},\text{reference})\right)$$

These base scores were plotted as a motif logo along −20/+20 bp of the gRNA to indicate relative importance of each base, independent of the cutting biases of the gRNA. It should be noted that since all mutations for GFP^{pos} as well as GFP^{neg} arise within the seed region of the gRNA, it is sometimes difficult to obtain a base-level importance score for these bases surrounding the cleavage site. However, due to the random lengths of stretches of mutations originating from the cleavage site, we can observe distinct sequence profiles emerging upstream and downstream of these bases.

Classification of GFP^{pos} and GFP^{neg} populations. We represented mismatches, insertions and deletions within −20/+20 bp of the gRNA as features. For all of the bases within the gRNA we represented 5 possibilities—A, C, T, G, and deletion. The feature for a base was one of four values for a particular base or the integer number of deleted bases starting at that base. Converting this categorical representation to a numeric format, we obtained 5×(length of gRNA + 40) features. Insertions were represented as the integer number of bases inserted immediately after each base of the gRNA and flanking boundaries. Hence, the total features were = 6×(length of gRNA + 40).

We performed fivefold classification of unique genotypes in GFP^{pos} and GFP^{neg} populations using a parallelized random forest implemented in MATLAB. We used 100 trees and ascertained that the out-of-bag classification error had reached convergence at this parameter value. Classification rate for a test-set genotype, was computed in an unweighted manner by counting each test-set genotype only once. In case of weighted accuracy measures, we weighted the accuracy of classification for each test-set genotype, by the number of reads assigned to it.

Conservation of bases. We examined the vertebrate phastcons score for every base in the gRNA at the URE to see if there was a correspondence with the importance of the base for regulation as determined above (**Supplementary Fig. 15e**).

Off-target effect analysis. To analyze potential false positives caused by off-target effects, we built a model of CRISPR off-target cutting using data from 13 gRNAs that were generated by GUIDE-Seq[32]. We found that in inverse proportion to the GC-content, gRNAs could tolerate a maximum of 3–6 total mismatches including the PAM region, with up to 3 mismatches in the seed region (9–20 bp) (**Supplementary Fig. 7a**–**d**).

Using our MERA data, we defined a true negative set as gRNAs that were tested in all replicates but did not cause a significant loss of GFP. Based on the positive and negative set, we created a set of rules of the following form to predict off-target effects:

No adjacent pairs of mismatches in the seed region (8–20 bp) allowed.

1. If total gRNA GC content ≤ 9: ≤3 total mismatches, ≤2 seed mismatches tolerated.
2. If total gRNA GC content ≥ 10 and ≤ 13 and seed GC content ≤ 7: ≤4 total mismatches, ≤2 seed mismatches tolerated.
3. If total GC ≥ 10 and ≤ 13 and seed GC > 7: ≤5 total mismatches, ≤3 seed mismatches tolerated.
4. If total GC ≥ 14 and ≤15: total mismatches ≤5;
   If seed GC ≤7: ≤2 seed mismatches;
   If seed GC>7: ≤3 seed mismatches.
5. ≥16 GC: ≤6 total mismatches, ≤3 seed mismatches tolerated.

Our off-target effect model predicted that none of these negative set gRNAs had a potential off-target effect overlapping a significant genomic site (**Supplementary Fig. 7e**).

In the *Tdgf1* library, 1,160/3,621 of the integrated gRNAs have potential off-target effects, and 150/925 of the gRNAs that were significantly enriched in GFP^{neg} populations have one or two potential off-target sites within the topological domain containing the *Tdgf1* gene as determined from mESC HiC data[42]. In the *Zfp42* library, 632/1,643 integrated guide RNAs have predicted off-target effects, and 34/332 of the gRNAs enriched in GFP^{neg} cells have predicted off-target effects in the topological domain containing the *Zfp42* gene.

Our off-target predictions are overly cautious, as off-target cutting is typically much rarer than on-target cutting[32] and the off-target sites predicted for MERA hits are often >200 kb away from the gene, reducing the likelihood of functional association. However, even this overestimate of off-target effects does not alter the patterns seen in MERA data.

Experimental methods. Experimental methods are described in the **Supplementary Methods**.

38. Nelder, J.A. & Wedderburn, R.W. Generalized linear models. *J. R. Stat. Soc. Ser. A* **135**, 370–384 (1972).
39. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
40. Needleman, S.B. & Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
41. Liese, F. & Miescke, K.J. *Statistical Decision Theory: Estimation, Testing, and Selection* (Springer, 2008).
42. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

doi:10.1038/nbt.3468