Published online 23 October 2016 *Nucleic Acids Research, 2017, Vol. 45, No. 4* ***e16***
doi: 10.1093/nar/gkw942

A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of *cis*-regulatory activity in neural cells

Brett B. Maricque[1,2], Joseph D. Dougherty[2,3] and Barak A. Cohen[1,2,*]

[1]Center for Genome Sciences and Systems Biology, Washington University School of Medicine, Saint Louis, MO 63108, USA, [2]Department of Genetics, Washington University School of Medicine, Saint Louis, MO 63108, USA and [3]Department of Psychiatry, Washington University School of Medicine, Saint Louis, MO 63108, USA

Received May 02, 2016; Revised October 05, 2016; Editorial Decision October 08, 2016; Accepted October 11, 2016

ABSTRACT

Recent large-scale genomics efforts to characterize the *cis*-regulatory sequences that orchestrate genome-wide expression patterns have produced impressive catalogues of putative regulatory elements. Most of these sequences have not been functionally tested, and our limited understanding of the non-coding genome prevents us from predicting which sequences are *bona fide cis*-regulatory elements. Recently, massively parallel reporter assays (MPRAs) have been deployed to measure the activity of putative *cis*-regulatory sequences in several biological contexts, each with specific advantages and distinct limitations. We developed LV-MPRA, a novel lentiviral-based, massively parallel reporter gene assay, to study the function of genome-integrated regulatory elements in any mammalian cell type; thus, making it possible to apply MPRAs in more biologically relevant contexts. We measured the activity of 2,600 sequences in U87 glioblastoma cells and human neural progenitor cells (hNPCs) and explored how regulatory activity is encoded in DNA sequence. We demonstrate that LV-MPRA can be applied to estimate the effects of local DNA sequence and regional chromatin on regulatory activity. Our data reveal that primary DNA sequence features, such as GC content and dinucleotide composition, accurately distinguish sequences with high activity from sequences with low activity in a full chromosomal context, and may also function in combination with different transcription factor binding sites to determine cell type specificity. We conclude that LV-MPRA will be an important tool for identifying *cis*-regulatory elements and stimulating new understanding about how the non-coding genome encodes information.

INTRODUCTION

Gene expression is controlled by short DNA sequences known as *cis*-regulatory elements, which collectively activate defined target genes to produce temporally constrained and cell type-specific expression patterns (1). Despite this high-level understanding of how gene expression works, two fundamental challenges in gene regulation remain. First, we toil to identify which non-coding DNA sequences are *bona fide* regulatory elements; and second, we lack understanding about which sequence features within a *cis*-regulatory element drive its activity.

Traditionally, *cis*-regulatory elements are identified and dissected using reporter genes that are measured one-at-a-time (2,3), limiting the number of unique sequences that can be screened. However, recent genome-scale studies have identified thousands of putative regulatory sequences across dozens of cell types (4), primarily based on indirect measurements of activity, such as chromatin modification status or transcriptional co-activator binding. To date, most predicted regulatory elements remain untested for *cis*-regulatory activity. Massively parallel reporter gene assays (MPRAs) (5–18) have facilitated large scale studies to probe regulatory element function, resulting in activity measurements for thousands of biochemically defined genomic regions (12,19). Studies using MPRA methods have also provided new insights about the local encoding of *cis*-regulatory information (10,12,20), and about complex interactions between nucleotides in regulatory elements (5). Still, validating current predictions and extending the impact of predictive models to more biologically relevant systems will require a high-throughput assay for *cis*-regulatory activity that works with high efficiency in many cell types.

Most implementations of MPRAs have been limited to transient, plasmid-based experiments in cultured cells that can be transfected or electroporated. We sought to further improve our ability to identify genomic regulatory elements and understand the DNA sequence basis of their activity by expanding MPRA applications to more biologically rel-

[*]To whom correspondence should be addressed. Tel: +1 314 362 3674; Email: cohen@genetics.wustl.edu

evant systems and by measuring regulatory activity in the context of genomic chromatin and higher order genomic structure. To this end, we developed a novel lentivirally-integrated massively parallel reporter assay (LV-MPRA), a quantitative platform for measuring the *cis*-regulatory activity of thousands of genome-integrated sequences, which can in principle be applied to any mammalian cell type.

Some recent studies describe methods that can assay the activity of genome-integrated reporter genes (15–17). LV-MPRA addresses several key limitations to extend the applications of genome-integrated assays and gain greater insight into the factors controlling gene expression. First, methods that rely on transient transfection to introduce integrating reporter gene libraries (15,17) are limited to *in vitro* studies of transfectable cell types. The use of a lentiviral delivery system in LV-MPRA greatly expands the range of cell types available for MPRA assays, including primary cells and cells in live organisms. Second, we explicitly measure the reproducibility of our expression measurements. This allows us to compare expression measurements between cell types and estimate the effects of regional chromatin domains on activity, neither of which can be done with methods that do not quantify reproducibility. Third, some mammalian assays use Fluorescence Activated Cell Sorting (FACS) to separate libraries of integrated fluorescent reporter genes into qualitative bins of expression (16,17) (e.g. 'no expression' versus 'high expression'). In contrast, we use barcode sequencing as a digital measure of expression, which provides a continuous and quantitative scale on which to measure the activity of regulatory sequences. Finally, the combination of a lentiviral delivery system coupled to direct barcode sequencing also allows us to increase throughput relative to methods that rely on targeted integration of reporter genes into a single locus per cell (17).

To establish LV-MPRA as a robust method for exploring regulatory element function in biologically relevant systems, we screened thousands of putative regulatory elements in human U87 glioblastoma cells and human neural progenitor cells (hNPCs) and assessed the DNA sequence features underlying activity. We determined that LV-MPRA is reproducible and sensitive to small changes in gene expression. We demonstrate the usefulness of measuring *cis*-regulatory element activity in fully integrated chromatin by quantifying the relative contributions of local DNA sequence features and regional chromatin domains. We discovered that simple DNA sequence features such as GC content and dinucleotide composition are important determinants of *cis*-regulatory activity and contribute to cell type-specificity in these cells. Moreover, we show that regulatory elements with high activity in U87 cells are enriched for different transcription factor binding sites than elements with high activity in hNPCs. Our approach expands the applications of MPRA technology to make quantitative activity measurements for genome-integrated *cis*-regulatory elements in a broad array of mammalian cell types.

MATERIALS AND METHODS

Lentiviral library construction

Pools of custom made 200-mer sequences were ordered through a limited licensing agreement with Agilent Technologies. Individual oligonucleotides were designed and synthesized according to the following template: 5′ primer annealing sequence (GTAGCGTCTGTCCCCTGCAG)/ SbfI site/*cis*-regulatory element/XhoI site/BamHI site/bar code/KpnI site/3′ primer annealing sequence (tggtaccctactactacag). Putative regulatory sequences containing SbfI, KpnI, XhoI, and BamHI were excluded from the library. See Supplemental Figure S3 for detail.

We created a streamlined lentiviral construct (pBM_01) by modifying a self-inactivating, replication-deficient FCIV backbone for use in LV-MPRA library preparation. We removed extraneous sequence from the construct while maintaining the elements required for viral packaging, transduction, and insertion into mammalian genomes, thereby reducing the size of the construct by 3.5kb. Primers BM155 and BM156 (Supplemental Table S2) were used to amplify array-synthesized oligonucleotides as previously described with an annealing temperature of 51C. PAGE purified oligos were cloned into pBM_01 using SbfI and KpnI. Plasmid DNA was prepared from ∼95,000 colonies to create library BM_101. The *Hspa1b* minimal promoter and dsRed reporter gene were amplified using primers BM130 and BM107 (Supplemental Table S2) and cloned into library BM_101 using BamHI and XhoI creating library BM_102. The *Hspa1b* minimal promoter was originally amplified from *Hspa1b* LacZ (kind gift of M. de Bruijn, Oxford Stem Cell Institute, Oxford, UK) and cloned upstream of dsRed in a pGL 4.23 plasmid backbone (Promega). Library BM_102 was linearized using PvuI in order to remove BM_101 library members that did not receive the *Hspa1b* -dsRed cassette, and subsequently recircularized to form library BM_103. Library BM_103 was submitted to the Hope Center Viral Vectors Core at Washington University School of Medicine for production of high-titer lentivirus. One-hundred percent of the *cis*-regulatory element-BC pairs present in the lentiviral plasmid pool were detected in the viral preparation.

U87 cell culture and transduction

U87 human glioblastoma cells were maintained in Dulbeco's modified Eagle's medium with 10% fetal bovine serum (DMEM + FBS). 7.5×10^5 cells were seeded into a T-25 flask with 5 ml of DMEM + FBS and cultured for 24 h prior to transduction at standard conditions. The lentiviral library of putative regulatory elements was applied to each well such that the multiplicity of infection (MOI) was equal to three. Based on U87 cell transduction by others, this should result in ∼85% transduction efficiency with each cell getting on average one lentiviral insertion (21). In total we targeted $\sim 1.3 \times 10^6$ cells for lentiviral insertion, resulting in approximately 100 insertions per unique regulatory element (13 000 in total) on average. Cells were cultured for 24 h in the presence of lentivirus before replating into a T-75 flask with 10 ml of fresh DMEM + FBS and expanded for 24 h. Cells were replated a second time into a T-150 flask with 20mL of fresh DMEM + FBS and expanded for an additional 48 h. Cells were counted and 1.5 × 10^7 cells were used in total RNA collection for each replicate.

Human neural progenitor cell culture and transduction

Human neural progenitor cells (derived from human induced pluripotent stem cells courtesy of the Induced Pluripotent Stem Cell Core at Washington University School of Medicine) were maintained in STEMdiff™ Neural Progenitor Medium (Stemcell Technologies). Cells were seeded into a T-25 flask with 5 ml STEMdiff™ media and cultured for 24 h prior to transduction (37°C, 6% CO_2). The lentiviral library of putative regulatory elements was applied to each well such that the multiplicity of infection (MOI) was equal to three. In total we targeted ∼1.3 × 10^6 cells for lentiviral insertion, resulting in approximately 100 insertions per barcoded regulatory element (13 000 in total) on average. Cells were cultured for 8 h in the presence of lentivirus before washing cells twice with STEMdiff™ media and replenishing the cultures with 5 ml fresh STEMdiff™ media. Cells were expanded for 24 h before being replated into two T-75 flasks per biological replicate. Cells were expanded for an additional 72 h. Cells were counted and 1.5 × 10^7 cells were used in total RNA collection for each replicate.

Measuring regulatory activity with RNA-seq

Total RNA was extracted from U87 cells and hNPCs 96 h after transduction using the ZR-Duet RNA/DNA MiniPrep kit (Zymo Research). DNA contamination was removed using the TURBO DNA-free kit (Applied Biosystems) following the standard manufacturer's protocol, and cDNA was generated using SuperScript II Reverse Transcriptase (Life Technologies) off of 800 ng input RNA per replicate. cDNA was prepared for Illumina sequencing according to methods previously described (5). Barcoded regions of cDNA molecules were amplified using PCR primers BM159 and BM160, which flank the barcoded region. PCR products were digested with BamHI and EcoRI and indexed Illumina adapters were ligated to the samples (U87 cells: P1_BM_BamHI_6, P1_BM_BamHI_7, P1_BM_BamHI_8, PE2_EcoRI_Ind70; hNPCs: P1_BM_BamHI_4, P1_BM_BamHI_5, P1_BM_BamHI_6, PE2_EcoRI_Ind70, Supplemental Table S2). Lentiviral genomic RNA was extracted from lentiviral particles using the QIAmp Viral RNA Mini Kit (Qiagen). cDNA was generated using SuperScript II Reverse Transcriptase and prepared for Illumina sequencing as described above. Indexed Illumina adapters were ligated to the viral cDNA PCR products (P1_BM_BamHI_1, PE2_EcoRI_Ind70, Supplemental Table S2).

We sequenced barcode amplicons from hNPC cDNA and hNPC gDNA using the Illumina NextSeq technology. Reads that perfectly matched the first 14 nucleotides of the amplion were included in subsequent analysis. We generated 90.4 million reads from the hNPC cDNA and 115.8 million reads from the hNPC gDNA of three biological replicates. We also generated 19.2 million reads from the viral cDNA. Expression in hNPCs is calculated as $\log_2$(RNA cDNA reads/gDNA reads) and the expression for each regulatory element in each replicate experiment is the mean expression of the barcodes associated with it. Viral cDNA reads are highly correlated with hNPC gDNA reads (R = 0.89), allowing the viral cDNA to be used for normalization. For U87 cells, two lanes of the Illumina HiSeq 2000 machine were used to sequence barcode amplicons from the U87 cDNA, and reads that perfectly matched the first 14 nucleotides of the amplicon were included in subsequent analysis. The expression of each barcode is calculated as $\log_2$(RNA cDNA reads/viral cDNA reads) and the expression for each regulatory element in each replicate experiment is the mean expression of the barcodes associated with it. In both cell types, expression was calculated for barcodes with at least 100 reads in the gDNA or viral cDNA and at least 10 reads in the RNA cDNA, and the final expression is the mean expression for the regulatory element across biological replicates. More than 80% of regulatory element-BC pairs present in the lentiviral preparation were detected in the RNA pool. Elements represented by at least three barcodes in at least two replicates were used for subsequent analysis. The standard error of the mean was calculated for each element as previously described (5).

Luciferase validation in U87 cells

Ten individual *cis*-regulatory elements determined to have high activity in U87 cells using LV-MPRA were selected for validation by luciferase assays. Each element was cloned into a pGL4.23-minP-luciferase expression vector upstream of the minP minimal promoter using Gibson assembly to create pGL-CRE-luciferase plasmids. Each construct was individually electroporated in triplicate into U87 cells using U87-specific conditions on the Neon transfection system. Each transfection consisted of 500 ng of a single pGL-CRE-luciferase construct with 100 ng of *Renilla* control plasmid and 5 × 10^4 cells. Transfected cells were grown in 75 ul of growth media in 96-well plates for 20 h, at which point luciferase assays were conducted according to standard protocols (Dual-Glo Luciferase Assay System, Promega). Firefly luciferase measurements were first normalized to Renilla Luciferase measurements within each replicate, averaged across the three replicates, and then normalized by the expression driven by the empty minP construct (Supplemental Figure S4). Two-tailed Student's t-tests were performed between pGL-minP-luciferase replicates and pGL-CRE-luciferase replicates to identify *cis*-regulatory elements with significant activity.

Integration site number on expression estimates

TRIP data (15) were downloaded from the Gene Expression Omnibus (GSE 48608), and used for simulations to determine how the number of integration sites for a reporter gene affects the estimate of its expression. We computed the expression driven by the mouse PGK promoter at each integrated region (IR), as described previously (15). Next, we computed the mean expression across all IRs measured (n = 17 857), which we refer to as the 'Overall mean' in Supplemental Figure S1A. Then we randomly sampled a collection of IRs (n = 10, 50, 100, 200, 300, 400, 500, 750, 1000 and

10000), and computed the mean expression across that random sample. For each value of *n,* we sampled 1000 times to create a distribution of mean expression estimates for each sample size. The spreads of these distributions reflect the accuracy to which a random sample of *n* IRs estimates the mean expression.

Sensitivity analysis

Regulatory elements for which all five barcodes were measured in all three replicates were used to determine the sensitivity of LV-MPRA. For each of these elements, the fold difference in activity relative to all other elements was determined. Previous work demonstrated that most expression changes produced by single nucleotide variants in a mammalian enhancer are 3.5-fold or greater (5). For each change that was 3.5-fold or greater (2484 in total), a Mann–Whitney U test was performed to determine whether the difference was statistically significantly different. *P*-value adjustments after Bonferroni multiple test corrections are presented in the text.

Logistic regression models

Logistic regression models were developed to distinguish *cis*-regulatory elements with high and low activities in our assay. High- and low-activity elements were defined as two sets of elements with non-overlapping standard errors. Model inputs were selected from a set of primary DNA sequence features based on a significant difference between regulatory elements with high and low activities (Mann–Whitney U tests two-tailed, $P < 0.01$, Bonferroni corrected). Our models incorporate the frequencies of all 16 dinucleotides. Only additive terms were used in the logistic regression models. Receiver operating characteristic (ROC) curves were created to assess predictive power and the area under the curve (AUC) was calculated for each model. Additionally, five-fold cross validation was used to ensure our models were not over-fit. The regulatory elements were split into five training groups, and the model was trained on the data holding out each group in turn and tested on the group held out. AUC was calculated for each of these sets (Supplemental Table S1).

RESULTS

Viral packaging and genome integration preserve LV-MPRA library complexity

Our motivation for developing LV-MPRA was 2-fold: first, we aimed to extend MPRA applications to a wider range of biologically relevant cell types, and second, we sought to quantitatively measure *cis*-regulatory activity in the context of the mammalian genome. To this end, we re-engineered an episome-based MPRA (5,12,19) system into a lentiviral-based, genome-integrated reporter assay (LV-MPRA). LV-MPRA builds on established MPRA technology, leveraging the efficiencies of array-based oligonucleotide synthesis (22) and molecular barcoding to construct libraries of regulatory elements and measure their activities *en masse*. Importantly, lentiviruses are a streamlined platform for inserting reporter genes into the genomes of diverse mammalian cell types, including primary cells (23–25). As such, LV-MPRA enables us to measure the activity of thousands of genome-integrated regulatory sequences in a single lentiviral transduction.

To evaluate LV-MPRA as a robust method for exploring the DNA sequence features underlying regulatory element activity, we selected 1800 regions from across the genome to serve as potential regulatory elements. We also created 800 synthetic DNA sequences based on the distribution of dinucleotide frequencies of the selected genomic regions. These elements will contain many of the same basic properties of genomic sequences, aiding our ability to identify sequence features sufficient to drive regulatory activity. Each element was cloned into a modified lentiviral construct (26) upstream of a dsRed reporter gene and a unique set of co-transcribed sequence barcodes in its 3′UTR. The plasmid library was then used to produce a high-titer lentiviral library composed of thousands of regulatory sequences (Figure 1A). The resulting lentiviral library was transduced into U87 cells, a glial derived cell line, and human neural progenitor cells (hNPCs), a relatively undifferentiated neural cell type. In total, we aimed to measure the activities of 13,000 unique genome-integrated reporter genes.

To ensure that lentiviral production and genomic integration did not diminish our library's complexity, we examined the relationship between barcode abundance in the lentiviral plasmid pool and barcode abundance in the lentiviral genomic RNA or the genomic DNA from transduced cells. We find strong correlation between barcodes well represented (>10 reads) in the plasmid pool and the viral genomic RNA ($R = 0.92$) or genomic DNA from transduced cells ($R = 0.85$), respectively. Moreover, 100% of barcodes in the lentiviral plasmid pool were packaged into virus, and we detected 87% of those barcodes in RNA from transduced cells, on average. These results indicate that LV-MPRA library construction and experimentation is robust to library complexity, and suggests that LV-MPRA will be an effective tool to study the function of genome-integrated regulatory sequences in biologically relevant cell types.

LV-MPRA quantitatively measures genome-integrated regulatory element activity

Lentiviruses integrate into widespread locations in mammalian genomes, often times near or within active transcriptional units (27–29); thus, different integration sites can exert different effects on reporter gene expression. Recent expression measurements for a transgene delivered with the *piggyBac* transposon system vary up to 1000-fold, depending on the chromatin context surrounding the insertion (15,30–32). Consequently, activity measurements for genome-integrated regulatory elements are a combination of the activity of the regulatory element itself and the activity of its genomic integration site. Quantitative activity measurements must be precise to be useful in determining how regulatory information is encoded in DNA sequence. Such genomic position effects are a direct threat to quantitatively measuring the activity of genome-integrated regulatory sequences.

In the current experiment, we hypothesized that measuring the activity of regulatory elements integrated at a

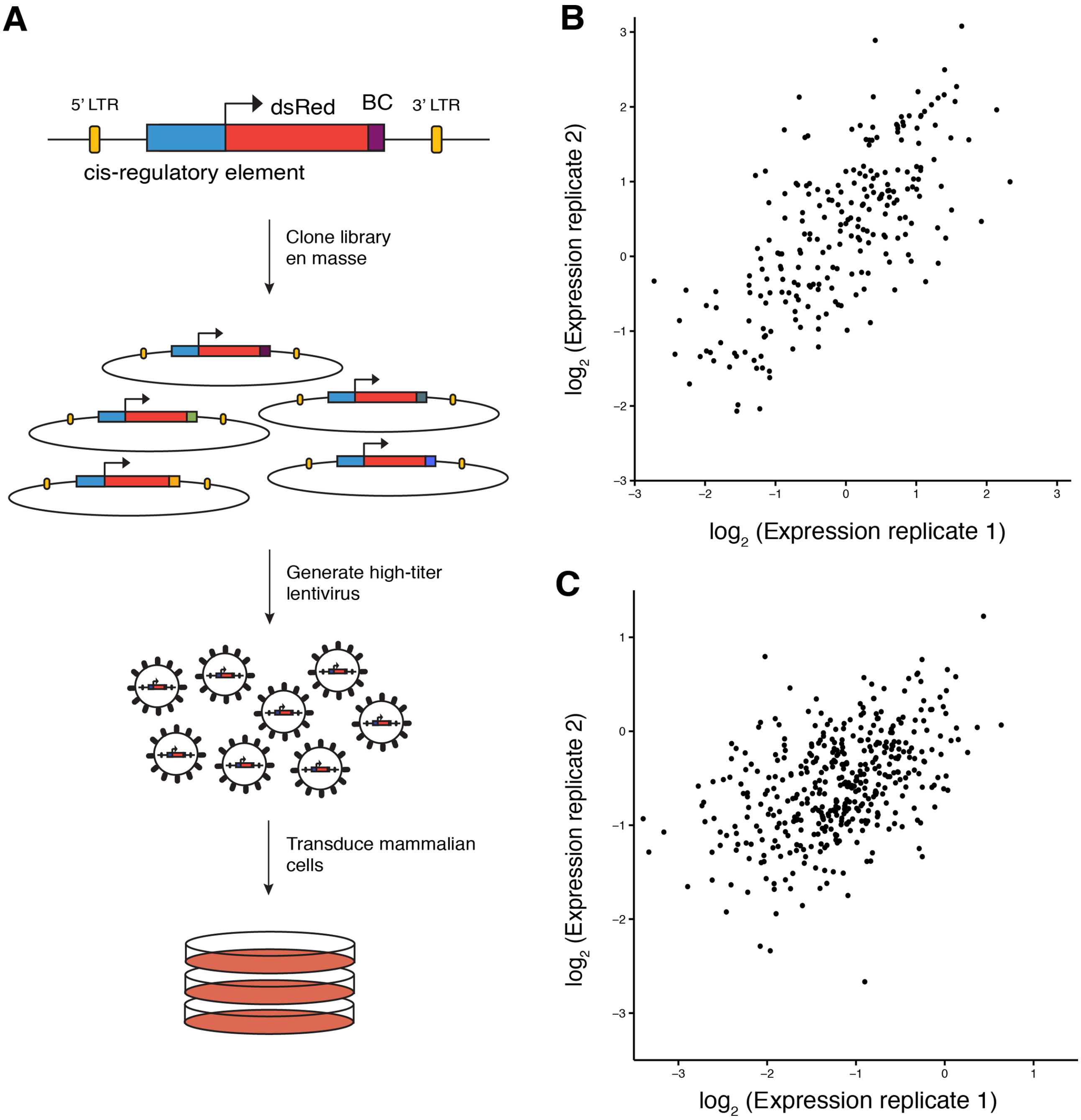

Figure 1. Quantitative LV-MPRA expression measurements are reproducible. (**A**) Putative *cis*-regulatory elements cloned upstream of an *Hspa1b* minimal promoter drive expression of dsRed containing DNA sequence barcodes in the 3′ UTR. 2600 unique reporter gene constructs were cloned in parallel and the complex lentiviral plasmid library was used to produce lentivirus for transduction. U87 cells and human neural progenitor cells were transduced in triplicate and RNA/DNA were extracted after 96 h in culture. (**B**) Scatter plot showing expression measurements for *cis*-regulatory elements from two biological replicates in U87 cells. Expression in each replicate is plotted as $\log_2$(RNA read count/DNA read count). Correlation between expression values of regulatory elements with measurements for all five barcodes is 0.70. (**C**) Scatter plot from two biological replicates in human neural progenitor cells. Expression in each replicate is plotted as $\log_2$(RNA read count/DNA read count). Correlation between expression values of regulatory elements with measurements for all five barcodes is 0.63.

large number of genomic loci would improve reproducibility by providing a robust estimate of their activity in the average chromatin environment. Averaging across many genomic integration sites will diminish the effects of individual chromatin environments, and should result in reproducible activity measurements. Indeed, our data is reproducible in both U87 cells and hNPCs. Regulatory elements with measurements for all five barcodes are reproducible between U87 cell replicate experiments with an $R = 0.70$ (Figure 1B) and hNPC replicate experiments with an $R = 0.63$ (Figure 1C). These results demonstrate that measuring the activity of regulatory elements in the average genomic context is an effective approach to minimizing genomic position effect.

Furthermore, if averaging measurements from more independent genomic integration sites improves reproducibility, then regulatory elements with many insertions should be more reproducible than elements with fewer insertions. This hypothesis is supported by data from Akhtar and colleagues, which demonstrates that the variance in the mean expression estimates for an integrated regulatory element decreases as the number of integration sites measured increases (Supplemental Figure S1A). To assess the effect of integration site number on reproducibility in our data we evaluated regulatory elements for which we measured different numbers of barcodes. We expect that elements for which we have measured all five barcodes to be more reproducible than elements for which we have measured only four barcodes or three barcodes because those measurements are averaged across more integration sites. As expected, elements with activity measurements for five barcodes are more reproducible ($R = 0.70$) than elements with measurements for four barcodes ($R = 0.65$) or three barcodes ($R = 0.59$) in U87 cells (Supplemental Figure S1B). Expression reproducibility follows a similar trend in hNPC replicate experiments (Supplemental Figure S1C). These results confirm that greater numbers of genomic integration sites leads to better reproducibility through averaging of chromatin contexts, and demonstrate that LV-MPRA measures regulatory activity in the context of genomic chromatin with good precision.

The quantitative nature of LV-MPRA allows us address an important question in regulatory genomics: What is the relative contribution to gene expression of local DNA sequence features versus regional chromatin position effects? Our reproducibility measures reveal that 40–50% of the variance in *cis*-regulatory activity in our experiments can be accounted for by differences in local DNA sequence. Using this, we can put an upper bound on the extent to which genomic position contributes to activity at 50–60%. This is the first demonstration of MPRA being used to estimate the effects of local DNA sequence and regional chromatin on regulatory activity; thus, LV-MPRA provides a novel framework for a new line of inquiry about how local DNA sequence and regional chromatin domains interact to control gene expression.

Importantly, measuring average regulatory activity across hundreds of integration sites in multiple replicates also provides the statistical power to detect small differences in activity. Previous work demonstrates that the majority of single nucleotide variants in an enhancer known to regulate the mouse *Rhodopsin* gene produce 3.5-fold changes in expression or greater (5). In this experiment, we find that 100% of the 3.5-fold differences between elements are significant (uncorrected P-values < 0.05, Mann–Whitney U test). In most applications of MPRA assays, experimental designs will result in hundreds of comparisons between elements, requiring correction for multiple tests. For example, in saturation mutagenesis experiments the fraction of expression differences among sequence variants that remain significant following multiple test correction depends entirely on the number of variants tested, which is determined by the length of the element. In our assay, 81% of 3.5-fold changes would remain significant in a saturation mutagenesis experiment for a 50 bp element (Bonferroni corrected P-values < 0.05, Mann–Whitney U test, $N = 150$), 76% for a 75 bp element ($N = 225$) and 74% for a 100 bp element ($N = 300$). We conclude that LV-MPRA is a sensitive method for quantitatively measuring the activity of thousands of genome integrated reporter genes allowing us to explore how regulatory elements encode information for gene expression control. Furthermore, it will be a powerful tool for measuring the effect sizes associated with single nucleotide variants in regulatory elements in the future.

Basic DNA sequence features encode regulatory activity and cell type-specificity

We explored the sequence features that distinguish regulatory elements with high activity (High elements) from elements with low activity (Low elements) within the genomes of U87 cells and hNPCs. We defined High elements as those above the 85th percentile in the expression distributions, and Low elements as those below the 15th percentile (Figure 2A and D). In this analysis, we included all regulatory elements with measurements for at least three barcodes in replicate experiments in our analysis, and averaged regulatory activity measurements across replicates resulting in a single expression dataset for each cell type. With respect to comparing U87 to hNPC cells, our experimental design allows us to compare each regulatory element's positions in the expression distributions from two different cell types, but not to compare absolute expression levels. Before starting our bioinformatic analyses we validated our LV-MPRA approach and confirmed that high elements have significant *cis*-regulatory activity using luciferase reporter gene assays. Nine of ten elements tested drove significant expression, and ranged in activity from 1.3- to 34-fold induction of the minimal promoter alone (median: 3.76-fold induction, Supplementary Figure S4). With two sets of sequences with distinct expression distributions in hand (Figure 2A and D), we went on to identify primary DNA sequence features that encode regulatory activity.

We found that High Elements in U87 cells have significantly higher GC content than Low Elements in U87 cells (Figure 2B, $P < 0.005$), a result previously observed for active regulatory sequences in the mouse retina (12), and that 11 of the 16 possible dinucleotides are represented significantly differently in the two classes of sequences (Figure 2C, $P < 0.005$). GC content and dinucleotide composition have the potential to affect regulatory activity by modulating structural features of the DNA sequence including nu-

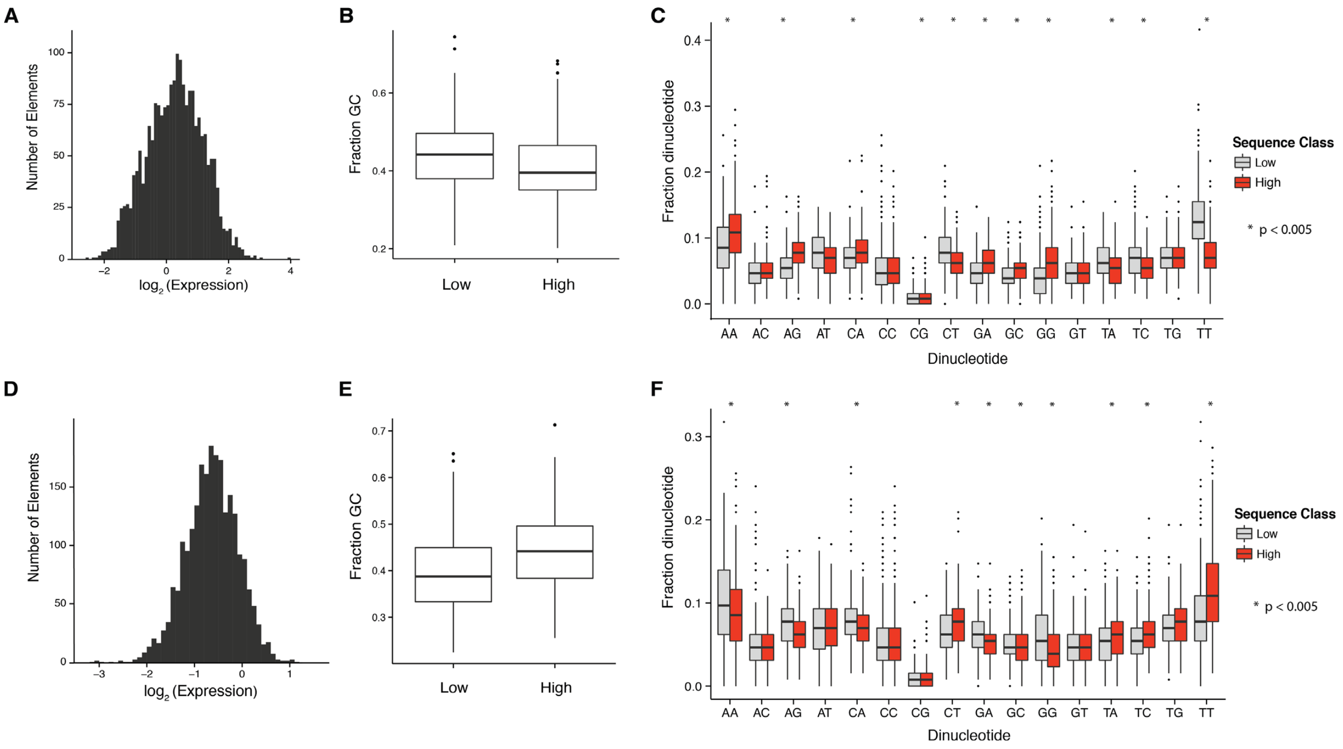

Figure 2. GC content and dinucleotide composition are associated with regulatory activity in U87 cells and hNPCs. (**A**, **D**) Distributions of expression values for all *cis*-regulatory elements in U87 cells (A) and hNPCs (D). Expression values are averaged across the three biological replicates and plotted as $\log_2$(RNA read count/DNA read count). (**B**, **E**) Box plots displaying GC content for high and low activities *cis*-regulatory elements in U87 cells (B) and hNPCs (E). GC content is significantly different between high- and low-activity elements in both cell types ($P < 0.01$, Mann–Whitney U test). (**C**, **F**) Box plots displaying dinucleotide composition for high and low activities elements in U87 cells (C) and hNPCs (F). Dinucleotide compositions are significantly different between high and low expressing regulatory elements in both cell types ($P < 0.005$, Mann–Whitney U test).

cleosome occupancy (33–36) and minor groove width (37). Nucleosome occupancy can promote TF binding or compete for TF binding sites (38) depending on which transcription factors are involved. Dinucleotides are also critical determinants of nucleosome positioning on genomic sequences (34), and the overall dinucleotide composition of regulatory sequences appears to be an important modulator of their activity (20). Minor groove width has been previously shown to affect Hox protein binding (39) and likely impacts nucleosome phasing (40). We find that High elements in U87 cells have wider minor grooves (Supplemental Figure S2A, $P < 0.01$) than Low elements in U87 cells, consistent with high-expressing regulatory sequences in previous experiments (12).

To assess the predictive power of these features, we developed logistic regression models to distinguish elements with high and low activities. A model incorporating frequencies for the sixteen dinucleotides performed very well (Figure 3A, AUC = 0.85) while GC content alone had modest power to distinguish High elements from Low elements (AUC = 0.62) in U87 cells. Five-fold cross validation was performed without significant loss of predictive power indicating that these models are not over fit (Supplemental Table S1). These results confirm that LV-MPRA is an efficient tool for studying the sequence features that comprise regulatory elements, and suggest that dinucleotide composition is an important contributor to regulatory activity.

In hNPCs, High Elements have significantly lower GC content than Low Elements (Figure 2E, $P < 0.005$), consistent with recent work measuring the activity of ENCODE segments (19), which shows that highly active regulatory sequences tend to have lower GC content (19). Similarly, AT-rich sequences are high-expressing in the mouse retina and GC-rich sequences are high-expressing in the mouse cortex (18). These findings suggest that the relationship between GC content and regulatory activity depends on cellular context. We also find that ten dinucleotides are represented significantly differently in High and Low elements in hNPCs (Figure 2F, $P < 0.005$), but the directionality of most differences is reversed relative to U87 cells. High elements in hNPCs also have narrower minor grooves relative to Low elements (Supplemental Figure S2B, $P < 0.01$), consistent with our findings about GC content in hNPCs. A logistic regression model based on dinucleotide composition had power to distinguish high- and low-activity elements reasonably well (Figure 3A, AUC = 0.74), suggesting that a substantial part of *cis*-regulatory activity is captured in both cell types.

Our data revealed a significant inverse correlation ($R = -0.326$, $P = 5.02e{-}53$, Pearson's product-moment correlation) between regulatory activity in U87 cells and regulatory activity in hNPCs suggesting that many elements display cell type-specific activity in our assay (Figure 4A). Indeed, only 5.4% of High Elements in U87 cells (<1% of all assayed elements) are also high expressing in hNPCs (4.8%

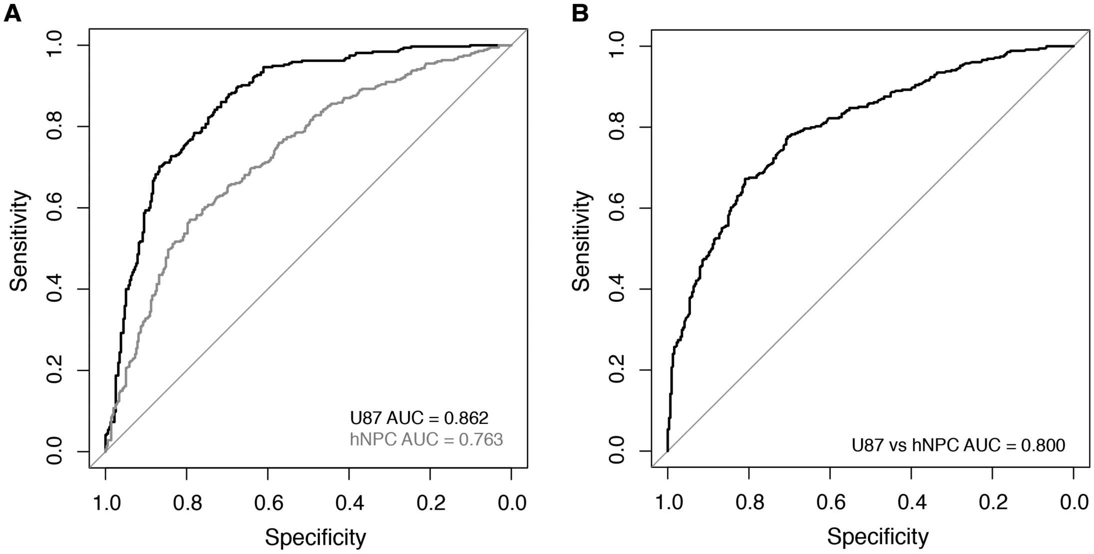

Figure 3. Dinucleotide composition predicts *cis*-regulatory element activity and cell type-specificity. (**A**) Receiver operating characteristic (ROC) curves for logistic regression models used to distinguish High and Low activity regulatory elements in U87 cells (AUC = 0.862) and hNPCs (AUC = 0.763). (**B**) ROC curve for a logistic regression model used to partition regulatory elements with high activity in specific cell types (AUC = 0.800). Models were trained and tested with 5-fold cross validation (Supplemental Table S1).

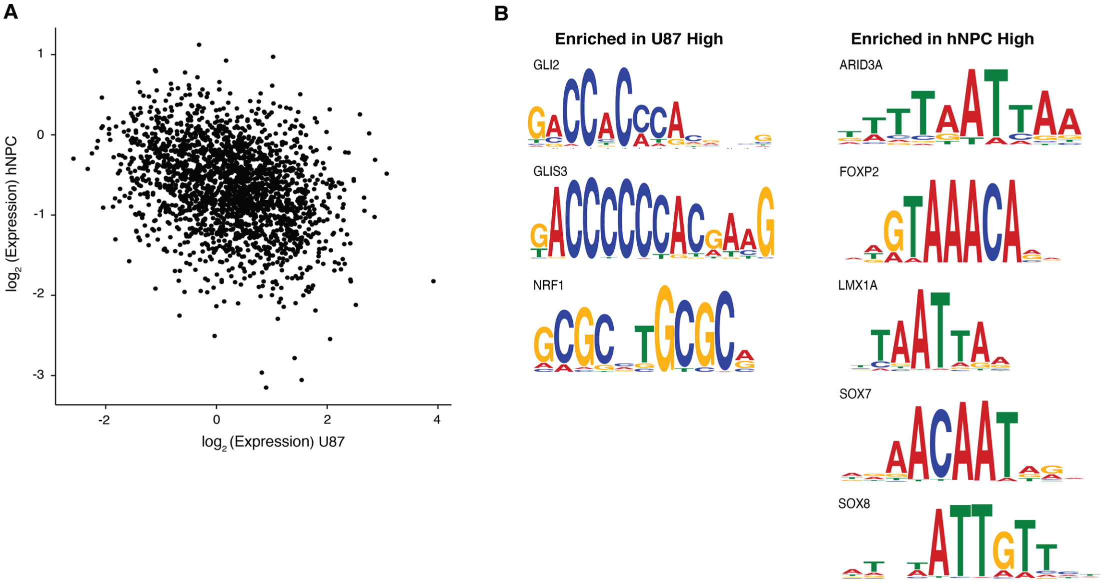

Figure 4. Different DNA motifs are enriched in high expressing *cis*-regulatory elements in U87 cells and hNPCs. (**A**) Scatter plot showing average expression measurements for regulatory elements in U87 cells and hNPCs. Expression in each cell type is plotted as $\log_2$(RNA read count/DNA read count). Activity in the two cell types is significantly negatively correlated ($R = -0.326$, $P = 5.02e{-}53$, Pearson's product-moment correlation). (**B**) AME analysis identified motifs enriched in regulatory elements with high activity in U87 cells relative to elements with high activity in hNPCs. (**C**) Motif enrichment in hNPC High Elements relative to U87 High Elements. Motifs enriched in U87 cells ($P < 0.05$) are GC rich and match TFBSs for broadly expressed TFs and motifs enriched in hNPCs ($P < 0.05$) are AT rich and match TFBSs for TFs involved in brain development and patterning and stem cell maintenance in the CNS.

of High Elements in hNPC are also High in U87). We asked whether the DNA sequence features important for regulatory activity also encoded information about cell type specificity. Again, we trained a logistic regression model based on dinucleotide compositions to classify sequences as having activity specific to one cell type or the other. The model performs well (Figure 3B, AUC = 0.80) indicating that basic DNA sequence features underlie a significant fraction of regulatory activity and contribute to cell type specificity.

We next tried to better understand the molecular basis for the difference in GC content and dinucleotide frequencies between High elements in U87 cells or hNPCs, respectively. One explanation is that differences in transcription factor binding site content among cell type-specific regulatory elements result in differences in overall GC content and dinucleotide frequencies. Thus, we examined the transcription factor binding sites present in both sets of sequences, hypothesizing that high expressing elements would be enriched for either GC-rich or AT-rich motifs important for specifying each cell type. Using AME (41), we searched for sequence motifs enriched in High elements from U87 cells relative to High elements from hNPCs, and for motifs enriched in hNPC High elements relative to U87 High elements. As, expected, we find that different types of motifs are enriched in these two sets of high-activity sequences (Figure 4B). In U87 cells, we find enrichment of GC-rich motifs matching binding sites for three broadly functional transcription factors, *GLIS3, GLI2* and *NRF1*. *De novo* motif finding using DREME (42) identified a short motif matching the core motif for GCM1, the prototypical transcription factor bearing a 'glial cell missing' motif, in U87 cell High Elements. In hNPC High elements, we find enrichment of multiple AT-rich motifs matching binding sites for transcription factors important for embryonic development and patterning (*SOX7, SOX15, ARID3A*) (43,44), brain development (*SOX8, DBX1, LMX1A, GLIS2, FOXP2*) (45–48) and stem cell maintenance in the central nervous system (*SOX2*) (49). These results suggest that some fraction of the differences in favorable GC content and dinucleotide compositions between U87 cells and hNPCs is due to different transcription factor binding site content. However, motif occurrence frequencies alone have marginal power to distinguish sequences with high activity in one cell type or the other. These data are consistent with recent work in *Drosophila* indicating that dinucleotide composition works in combination with motifs to define cell type-specific gene expression (20), and inform additional studies designed to decouple the effects of dinucleotide composition and motifs on *cis*- activity.

DISCUSSION

Here, we describe LV-MPRA, a genome-integrated, massively parallel approach for quantitatively measuring *cis*-regulatory element activity in the context of mammalian genomes. We used LV-MPRA to quantitatively measure regulatory activity in a full chromosomal context, estimate the contributions of local sequence features and regional chromatin to regulatory activity, and discover that simple DNA sequence features encode a significant fraction of *cis*-regulatory function. We identify regulatory elements with cell type-specific activity and report corresponding differences in dinucleotide composition and sequence motif enrichment. Because of the widespread tropism of lentiviruses, LV-MPRA can be extended to more biologically relevant cell types than plasmid-based MPRAs, as we demonstrate with experiments measuring regulatory activity in human neural progenitor cells.

Our understanding of how the non-coding genome encodes regulatory information has been limited because we lack functional assays to identify genuinely active regulatory sequences. The ideal functional assay to identify *cis*-regulatory sequences and understand how their regulatory information is encoded would be high-throughput and measure regulatory activity at native genomic loci in biologically relevant cell types or intact organisms. Zinc finger nuclease, TALEN, and CRISPR technologies have made it possible to specifically edit genomes, but their current throughput is not sufficient to screen all non-coding sequences or large collections of predicted *cis*-regulatory elements for activity. In the meantime, multiple groups have developed high-throughput techniques to study *cis*-regulatory sequence activity (5–13,16–18), which assay putative regulatory sequences in different contexts. Each technique resembles the ideal functional assay to a different degree, provides unique advantages, and has distinct limitations.

To date, most MPRA implementations assay sequences on episomal vectors in cultured cells. The episomal context enables experiments requiring very high throughput or high sensitivity since thousands of plasmids can be delivered into individual cells. However, these systems lack chromatin context and are limited to transfectable cell types, though recent work uses adeno-associated virus to extend the applications of episome-based methods to non-transfectable cells (18). Other methods assay the activity of genome-integrated regulatory elements (15–17). These approaches make an important trade-off between throughput and control over genomic integration site, and are not quantitative. SIF-seq was developed (17) to measure regulatory activity using fluorescence activated cell sorting (FACS), and assays regulatory sequences at a single genomic locus, thereby controlling for genomic position effects but severely limiting throughput. Upward of 10^9 cells must be transfected to measure the activity of 10 000 elements using SIF-seq. In contrast, LV-MPRA can measure the activity of more than 10 000 elements in fewer than 10^6 cells, drastically improving throughput without the use of intense selection. FIREWACh (16), a lentiviral-based enhancer-screening assay, also deploys FACS to make genome-integrated activity measurements. FIREWACh is high-throughput and applicable to diverse mammalian cell types, but each measurement is subject to the position effect of a single lentiviral integration site. LV-MPRA is high throughput and accounts for the effects of genomic integration site by measuring regulatory activity of each element integrated at hundreds (∼300–500) of different locations, thereby averaging the element's regulatory activity across different chromatin contexts. SIF-seq and FIREWACh are qualitative measures of regulatory activity based on arbitrary FACS cutoffs; accordingly, the resolution and dynamic range of these techniques are unknown. In contrast, our use of barcode sequencing enables quantitative activity measurements, giving LV-MPRA finer resolution and substantially greater dynamic range than other genome-integrated methods. Thus, LV-MPRA advances beyond existing methods by increasing the throughput of genome-integrated approaches while accounting for random genomic integration sites.

We note several additional features of LV-MPRA that provide advantages for different experimental designs. First, lentiviral constructs can accommodate large inserts (up to 10 kb without effecting viral titer), making it possible to assay libraries of long regulatory sequences. Importantly, though the scale of DNA synthesis is on an upward curve enabling larger experiments with longer oligos, LV-MPRA does not rely on the DNA synthesis methods used in this study. Rather, LV-MPRA can accommodate DNA sequences from diverse sources, including randomly cloned genomic DNA, DNA isolated from capture arrays, or targeted PCR amplicons. Second, lentivirally-delivered reporter genes display relatively fast expression dynamics (robust expression <48 h post-transduction), enabling studies in primary cells that are difficult to culture for extended periods of time. Third, barcode sequencing eliminates the need for time- and labor-intensive protocols that are necessary to deploy FACS-based assays on *in vivo* tissues. Last, genome-integration of regulatory elements permits experiments involving differentiation, prolonged drug exposure, or other environmental manipulations or screens.

Our data demonstrate that local primary DNA sequence features such as GC content and dinucleotide composition determine a significant fraction of *cis*-regulatory activity and cell type-specificity in U87 cells and hNPCs. GC content has been associated with regulatory activity in multiple biological systems (12,18,19,50) and could function by modulating nucleosome occupancy and/or DNA shape. We also identify different sets of sequence motifs enriched in highly active sequences from U87 cells and hNPCs, respectively. A recent study in *Drosophila* indicates that dinucleotide composition and specific motifs combine to encode cell type-specific regulatory information (20). Our data are consistent with such a model and suggest that dinucleotide composition might poise a *cis*-regulatory element for activity, while transcription factor binding sites provide information for cell type specificity. Finally, we used our data to estimate the relative contributions of local DNA sequence features and regional chromatin position effects to *cis*-regulatory activity in the genome. We show that local DNA sequence accounts for 40–50% of the variance in regulatory activity. This is the first attempt to quantify the contribution of genomic position effects to the total variance in gene expression across the genome. Our work shows that short DNA sequences can often drive reproducible expression despite the strong influence of regional chromatin effects.

Improving our functional understanding of the noncoding genome remains a top priority in biomedical research. As such, several groups have developed useful MPRA technologies to assay the function of putative regulatory elements in different contexts. We extend the applications of MPRAs by developing LV-MPRA, a flexible platform for quantitatively measuring the activity of genome-integrated regulatory sequences in a wide range of biologically relevant cell types. Our study demonstrates that LV-MPRA can be used to explore how regulatory information is encoded in DNA sequence, and will be an important tool for dissecting the functional potential of diverse non-coding sequences.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank members of the Dougherty Lab for helpful advice on viral transduction and cell culture. We thank members of the Cohen Lab for their feedback on experimental design and analysis, and for their critical reading of the manuscript.

FUNDING

National Institutes of Health [R01GM092910, R01HG006790, 5R21NS083052-02]; Hope Center Viral Vectors Core at Washington University School of Medicine; NSF graduate research fellowship [DGE-1143954 to B.B.M.]. Funding for open access charge: R01HG008687.

Conflict of interest statement. None declared.

REFERENCES

1. Levine,M. (2010) Transcriptional enhancers in animal development and evolution. *Curr. Biol.*, **20**, R754–R763.
2. Visel,A., Blow,M.J., Li,Z., Zhang,T., Akiyama,J.A., Holt,A., Plajzer-Frick,I., Shoukry,M., Wright,C., Chen,F. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.
3. Swanson,C.I., Evans,N.C. and Barolo,S. (2010) Structural rules and complex regulatory circuitry constrain expression of a notch- and EGFR-regulated eye enhancer. *Dev. Cell*, **18**, 359–370.
4. ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
5. Kwasnieski,J.C., Mogno,I., Myers,C.A., Corbo,J.C. and Cohen,B.A. (2012) Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 19498–19503.
6. Kheradpour,P., Ernst,J., Melnikov,A., Rogov,P., Wang,L., Zhang,X., Alston,J., Mikkelsen,T.S. and Kellis,M. (2013) Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.*, **23**, 800–811.
7. Patwardhan,R.P., Hiatt,J.B., Witten,D.M., Kim,M.J., Smith,R.P., May,D., Lee,C., Andrie,J.M., Lee,S.I., Cooper,G.M. *et al.* (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.*, **30**, 265–270.
8. Melnikov,A., Murugan,A., Zhang,X., Tesileanu,T., Wang,L., Rogov,P., Feizi,S., Gnirke,A., Callan,C.G. Jr, Kinney,J.B. *et al.* (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.*, **30**, 271–277.
9. Arnold,C.D., Gerlach,D., Stelzer,C., Boryń,Ł.M., Rath,M. and Stark,A. (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, **339**, 1074–1077.
10. Smith,R.P., Taher,L., Patwardhan,R.P., Kim,M.J., Inoue,F., Shendure,J., Ovcharenko,I. and Ahituv,N. (2013) Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.*, **45**, 1021–1028.
11. Kinney,J.B., Murugan,A., Callan,C.G. and Cox,E.C. (2010) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 9158–9163.
12. White,M.A., Myers,C.A., Corbo,J.C. and Cohen,B.A. (2013) Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 11952–11957.
13. Sharon,E., Kalma,Y., Sharp,A., Raveh-Sadka,T., Levo,M., Zeevi,D., Keren,L., Yakhini,Z., Weinberger,A. and Segal,E. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.*, **30**, 521–530.
14. Mogno,I., Kwasnieski,J.C. and Cohen,B.A. (2013) Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res.*, **23**, 1908–1915.
15. Akhtar,W., de Jong,J., Pindyurin,A.V., Pagie,L., Meuleman,W., de Ridder,J., Berns,A., Wessels,L.F., van Lohuizen,M. and van Steensel,B. (2013) Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell*, **154**, 914–927.
16. Murtha,M., Tokcaer-Keskin,Z., Tang,Z., Strino,F., Chen,X., Wang,Y., Xi,X., Basilico,C., Brown,S., Bonneau,R. *et al.* (2014) FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nat. Methods*, **11**, 559–565.

17. Dickel,D.E., Zhu,Y., Nord,A.S., Wylie,J.N., Akiyama,J.A., Afzal,V., Plajzer-Frick,I., Kirkpatrick,A., Göttgens,B., Bruneau,B.G. *et al.* (2014) Function-based identification of mammalian enhancers using site-specific integration. *Nat. Methods*, **11**, 566–571.
18. Shen,S.Q., Myers,C.A., Hughes,A.E., Byrne,L.C., Flannery,J.G. and Corbo,J.C. (2015) Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res.*, **26**, 238–255.
19. Kwasnieski,J.C., Fiore,C., Chaudhari,H.G. and Cohen,B.A. (2014) High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.*, **24**, 1595–1602.
20. Yáñez-Cuna,J.O., Arnold,C.D., Stampfel,G., Boryń,Ł.M., Gerlach,D., Rath,M. and Stark,A. (2014) Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. *Genome Res.*, **24**, 1147–1156.
21. Feng,S.Y., Dong,C.G., Wu,W.K.K., Wang,X.J., Qiao,J. and Shao,J.F. (2012) Lentiviral expression of anti-microRNAs targeting miR-27a inhibits proliferation and invasiveness of U87 glioma cells. *Mol. Med. Rep.*, **6**, 275–281.
22. LeProust,E.M., Peck,B.J., Spirin,K., McCuen,H.B., Moore,B., Namsaraev,E. and Caruthers,M.H. (2010) Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.*, **38**, 2522–2540.
23. Cartier,N., Hacein-Bey-Abina,S., Bartholomae,C.C., Veres,G., Schmidt,M., Kutschera,I., Vidaud,M., Abel,U., Dal-Cortivo,L., Caccavelli,L. *et al.* (2009) Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science*, **326**, 818–823.
24. Modlich,U., Navarro,S., Zychlinski,D., Maetzig,T., Knoess,S., Brugman,M.H., Schambach,A., Charrier,S., Galy,A., Thrasher,A.J. *et al.* (2009) Insertional transformation of hematopoietic cells by self-inactivating lentiviral and gammaretroviral vectors. *Mol. Ther.*, **17**, 1919–1928.
25. Lois,C., Hong,E.J., Pease,S., Brown,E.J. and Baltimore,D. (2002) Germline transmission and tissue-specific expression of transgenes delivered by lentiviral vectors. *Science*, **295**, 868–872.
26. Araki,T., Sasaki,Y. and Milbrandt,J. (2004) Increased nuclear NAD biosynthesis and SIRT1 activation prevent axonal degeneration. *Science*, **305**, 1010–1013.
27. Ciuffi,A., Mitchell,R.S., Hoffmann,C., Leipzig,J., Shinn,P., Ecker,J.R. and Bushman,F.D. (2006) Integration site selection by HIV-based vectors in dividing and growth-arrested IMR-90 lung fibroblasts. *Mol. Ther.*, **13**, 366–373.
28. Yang,S.-H., Cheng,P.-H., Sullivan,R.T., Thomas,J.W. and Chan,A.W.S. (2008) Lentiviral Integration Preferences in Transgenic Mice. *Genes*, **46**, 711–718.
29. Ustek,D., Sirma,S., Gumus,E., Arikan,M., Cakiris,A., Abaci,N., Mathew,J., Emrence,Z., Azakli,H., Cosan,F. *et al.* (2012) A genome-wide analysis of lentivector integration sites using targeted sequence capture and next generation sequencing technology. *Infect. Genet. Evol.*, **12**, 1349–1354.
30. Grewal,S.I.S. and Jia,S. (2007) Heterochromatin revisited. *Nat. Rev. Genet.*, **8**, 35–46.
31. Girton,J.R. and Johansen,K.M. (2008) In: Genetics,B-A (ed). Academic Press. **61**, 1–43.
32. Walters,M.C., Fiering,S., Eidemiller,J., Magis,W., Groudine,M. and Martin,D.I. (1995) Enhancers increase the probability but not the level of gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 7125–7129.
33. Kaplan,N., Moore,I.K., Fondufe-Mittendorf,Y., Gossett,A.J., Tillo,D., Field,Y., LeProust,E.M., Hughes,T.R., Lieb,J.D., Widom,J. *et al.* (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.
34. Struhl,K. and Segal,E. (2013) Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.*, **20**, 267–273.
35. Tillo,D. and Hughes,T.R. (2009) G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics*, **10**, 442.
36. Raveh-Sadka,T., Levo,M., Shabi,U., Shany,B., Keren,L., Lotan-Pompan,M., Zeevi,D., Sharon,E., Weinberger,A. and Segal,E. (2012) Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.*, **44**, 743–750.
37. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein-DNA recognition. *Nature*, **461**, 1248–1253.
38. Lidor Nili,E., Field,Y., Lubling,Y., Widom,J., Oren,M. and Segal,E. (2010) p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy. *Genome Res.*, **20**, 1361–1368.
39. Dror,I., Zhou,T., Mandel-Gutfreund,Y. and Rohs,R. (2013) Covariation between homeodomain transcription factors and the shape of their DNA binding sites. *Nucleic Acids Res.*, doi:10.1093/nar/gkt862.
40. West,S.M., Rohs,R., Mann,R.S. and Honig,B. (2010) Electrostatic interactions between arginines and the minor groove in the nucleosome. *J. Biomol. Struct. Dyn.*, **27**, 861–866.
41. McLeay,R.C. and Bailey,T.L. (2010) Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, **11**, 165.
42. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.
43. Stovall,D.B., Cao,P. and Sui,G. (2014) SOX7: from a developmental regulator to an emerging tumor suppressor. *Histol. Histopathol.*, **29**, 439–445.
44. Maruyama,M., Ichisaka,T., Nakagawa,M. and Yamanaka,S. (2005) Differential roles for Sox15 and Sox2 in transcriptional control in mouse embryonic stem cells. *J. Biol. Chem.*, **280**, 24371–24379.
45. Cheng,Y.-C., Lee,C.-J., Badge,R.M., Orme,A.T. and Scotting,P.J. (2001) Sox8 gene expression identifies immature glial cells in developing cerebellum and cerebellar tumours. *Mol. Brain Res.*, **92**, 193–200.
46. Causeret,F., Ensini,M., Teissier,A., Kessaris,N., Richardson,W.D., de Couville,T.L. and Pierani,A. (2011) Dbx1-expressing cells are necessary for the survival of the mammalian anterior neural and craniofacial structures. *PLoS ONE*, **6**, e19367.
47. Sánchez-Danés,A., Consiglio,A., Richaud,Y., Rodríguez-Pizà,I., Dehay,B., Edel,M., Bové,J., Memo,M., Vila,M., Raya,A. *et al.* (2012) Efficient generation of A9 midbrain dopaminergic neurons by lentiviral delivery of LMX1A in human embryonic stem cells and induced pluripotent stem cells. *Hum. Gene Ther.*, **23**, 56–69.
48. Lai,C.S.L., Gerrelli,D., Monaco,A.P., Fisher,S.E. and Copp,A.J. (2003) FOXP2 expression during brain development coincides with adult sites of pathology in a severe speech and language disorder. *Brain*, **126**, 2455–2462.
49. Zappone,M.V., Galli,R., Catena,R., Meani,N., De Biasi,S., Mattei,E., Tiveron,C., Vescovi,A.L., Lovell-Badge,R., Ottolenghi,S. *et al.* (2000) Sox2 regulatory sequences direct expression of a (beta)-geo transgene to telencephalic neural stem cells and precursors of the mouse embryo, revealing regionalization of gene expression in CNS stem cells. *Dev. Camb. Engl.*, **127**, 2367–2382.
50. Manor,O. and Segal,E. (2013) Robust prediction of expression differences among human individuals using only genotype information. *PLoS Genet.*, **9**, e1003396.