A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity

Fumitaka Inoue, Martin Kircher, Beth Martin, et al.

Genome Res. published online November 9, 2016
Access the most recent version at doi:10.1101/gr.212092.116

A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity

Running title: **Comparing chromosomal and episomal reporter assays**

Fumitaka Inoue[1]†, Martin Kircher[2]†, Beth Martin[2], Gregory M. Cooper[3], Daniela M. Witten[4], Michael T. McManus[5], Nadav Ahituv[1]*, Jay Shendure[2,6]*

Affiliations:

[1]Department of Bioengineering and Therapeutic Sciences, Institute for Human Genetics, University of California San Francisco, San Francisco CA, USA

[2]Department of Genome Sciences, University of Washington, Seattle WA, USA

[3]HudsonAlpha Institute for Biotechnology, Huntsville AL, USA.

[4]Departments of Statistics and Biostatistics, University of Washington, Seattle WA, USA

[5]Department of Microbiology and Immunology, UCSF Diabetes Center, Keck Center for Noncoding RNA, University of California, San Francisco, San Francisco, CA, USA

[6]Howard Hughes Medical Institute, Seattle WA, USA

†These authors contributed equally to this work

*Correspondence to nadav.ahituv@ucsf.edu and shendure@uw.edu

Abstract

Candidate enhancers can be identified on the basis of chromatin modifications, the binding of chromatin modifiers and transcription factors and cofactors, or chromatin accessibility. However, validating such candidates as bona fide enhancers requires functional characterization, typically achieved through reporter assays that test whether a sequence can increase expression of a transcriptional reporter via a minimal promoter. A longstanding concern is that reporter assays are mainly implemented on episomes, which are thought to lack physiological chromatin. However, the magnitude and determinants of differences in *cis*-regulation for regulatory sequences residing in episomes versus chromosomes remain almost completely unknown. To address this question in a systematic manner, we developed and applied a novel lentivirus-based massively parallel reporter assay (lentiMPRA) to directly compare the functional activities of 2,236 candidate liver enhancers in an episomal versus a chromosomally integrated context. We find that the activities of chromosomally integrated sequences are substantially different from the activities of the identical sequences assayed on episomes, and furthermore are correlated with different subsets of ENCODE annotations. The results of chromosomally-based reporter assays are also more reproducible and more strongly predictable by both ENCODE annotations and sequence-based models. With a linear model that combines chromatin annotations and sequence information, we achieve a Pearson's R^2 of 0.362 for predicting the results of chromosomally integrated reporter assays. This level of prediction is better than with either chromatin annotations or sequence information alone and also outperforms predictive models of episomal assays. Our results have broad implications for how *cis*-regulatory elements are identified, prioritized and functionally validated.

Introduction

An enhancer is defined as a short region of DNA that can increase the expression of a gene, independent of its orientation and flexible with respect to its position relative to the transcriptional start site (Banerji et al. 1981; Moreau et al. 1981). Enhancers are thought to be modular, in the sense that they are active in heterologous sequence contexts and in that multiple enhancers may additively dictate the overall expression pattern of a gene (Shlyueva et al. 2014). They act through the binding of transcription factors, which recruit histone modifying factors, such as histone acetyltransferase (HAT) or histone methyltransferase (HMT). Enhancers are also associated with chromatin remodeling factors (e.g. SWI/SNF) and the cohesin complex, which are involved in regulating chromatin structure and accessibility (Schmidt et al. 2010; Euskirchen et al. 2011; Faure et al. 2012).

Antibodies against specific transcription factors (TFs), histone modifications or transcriptional co-activators are commonly used for chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) to identify candidate enhancers in a genome-wide manner. For example, the ENCODE Consortium and other efforts have identified thousands of candidate enhancers in mammalian genomes on the basis of such marks or their correlates (e.g. EP300 ChIP-seq; H3K27ac ChIP-seq; DNase I hypersensitivity) in diverse cell lines and tissues (Visel et al. 2009; The ENCODE Project Consortium 2012). However, a major limitation of such assays is that they reflect biochemical marks that are correlated with enhancer activity, rather than directly showing that any particular sequence actually functions as an enhancer. In other words, while such assays yield genome-wide catalogs of potential enhancers, they do not definitively predict bona fide enhancers nor precisely define their boundaries.

For decades, the primary means of functionally validating enhancers has been the episomal reporter assay. The standard approach is to relocate the candidate enhancer sequence to an episomal vector, adjacent to a minimal promoter driving expression of a reporter gene, e.g. luciferase or others. More recently, massively parallel reporter assays (MPRAs) have enabled the functional characterization of *cis*-regulatory elements including enhancers in a high-throughput manner. MPRAs use sequencing-based quantification of reporter barcodes to enable multiplexing of the reporter assay (Patwardhan et al. 2009). MPRAs have been used primarily in an episomal manner for the saturation mutagenesis of promoters and enhancers (Patwardhan et al. 2009; Kinney et al. 2010; Melnikov et al. 2012; Patwardhan et al. 2012), for exploring the grammatical rules of promoters and enhancers (Smith et al. 2013; Sharon et al. 2014), and for testing of thousands of enhancer candidates in different cells or tissues (Kwasnieski et al. 2012; Arnold et al. 2013; Kheradpour et al. 2013; Arnold et al. 2014; Shlyueva et al. 2014; Savic et al. 2015; White 2015). Adeno-associated virus (AAV) MPRAs have also been developed, allowing these assays to be carried out *in vivo* and to perform reporter assays within target cells and tissues that

are difficult to transduce, such as the brain (Shen et al. 2016), although these do not involve genomic integration.

Despite their widespread use to validate enhancers and other *cis*-regulatory elements, a longstanding concern about reporter assays is that they are almost always carried out via transient transfection of non-integrating episomes. It is unknown whether transiently transfected sequences are chromatinized in a way that makes them appropriate models for endogenous gene expression from chromosomes (Smith and Hager 1997); but to the extent that this question has been explored, there are differences. For example, work from Archer, Hager and colleagues using the mouse mammary tumor virus (*MMTV*) promoter as a model shows differences in histone H1 stoichiometry and nucleosome positioning resulting in an inability of episomal assays to reliably assay cooperative TF binding (Archer et al. 1992; Smith and Hager 1997; Hebbar and Archer 2007; Hebbar and Archer 2008). In another work, the chromatin structure of transiently transfected non-replicating plasmid DNA was observed to be differently fragmented than endogenous chromatin by micrococcal nuclease, and along with other data supports a model in which atypical chromatin might be induced by association of episomes with nuclear structures (Jeong and Stein 1994). However, the extent to which these factors operate to confound the results of reporter assays performed for enhancer validation, whether categorical (i.e. is a particular sequence an enhancer?), qualitative (i.e. in what tissues is an element an enhancer?), or quantitative (i.e. what level of activation does a particular sequence confer?), has yet to be systematically investigated.

To address these questions, we developed lentiviral MPRA (lentiMPRA), a technology that uses lentivirus to integrate enhancer MPRA libraries into the genome. To overcome the substantial position-effect variegation observed by others in attempting to use lentiviral infection for MPRA (Murtha et al. 2014), we employed a flanking antirepressor element (#40) and a scaffold-attached region (SAR) (Klehr et al. 1991; Kwaks et al. 2003) on either side of our construct. In addition, we relied on as many as 100 independent reporter barcode sequences per assayed candidate enhancer sequence, integrated at diverse sites. The resulting system allows for high-throughput, highly reproducible and quantitative measurement of the regulatory potential of candidate enhancers in a chromosomally integrated context. Furthermore, the cell-type range of lentivirus transduction is much broader than transfection, e.g. permitting MPRAs to be conducted in neurons, primary cells or organoids.

Results

Construction and validation of the lentiMPRA vector

The potential for confounding of lentiviral assays by site-of-integration effects was demonstrated by a recent MPRA study that used lentiviral infection and found that 26% of positive controls did not show activated GFP expression, while other measures estimated a false positive rate of 22% (Murtha et al. 2014). We therefore constructed a lentiviral vector (pLS-mP) that contains a minimal promoter (mP) and the enhanced green fluorescent protein (EGFP) gene flanked on one side by the antirepressor element #40 and the other by a SAR (Fig. 1A, Supplementary File 1) (Klehr et al. 1991; Kwaks et al. 2003; Kissler et al. 2006). In experiments involving chromosomal integration of this enhancer reporter, we confirmed that EGFP is not expressed in the absence of an enhancer, while abundantly expressed under the control SV40 enhancer across a panel of cell lines representing diverse tissues-of-origin. These include: K562 (lymphoblasts), H1-ESC (embryonic stem cells), HeLa-S3 (cervix), HepG2 (hepatocytes), T-47D (epithelial) and Sk-n-sh retinoic acid treated (neuronal) cells (Fig. S1). Furthermore, when SV40 and the *Ltv1* liver enhancer (Patwardhan et al. 2012) are tested without the flanking antirepressor sequences, we observed much lower levels of EGFP expression in HepG2 cells (Fig. 1B). FACS analysis showed that the inclusion of antirepressors increased the proportion of the cells that strongly express GFP (Fig. 1C). This result was consistent with our expectation that the antirepressors facilitate robust enhancer-mediated expression from the integrated reporter.

Design and construction of a library of candidate liver enhancers

To evaluate lentiMPRA, we designed a liver enhancer library that comprises 2,236 candidate sequences and 204 control sequences (Fig. 1D, Supplementary File 2), each 171 bp in length. All enhancer candidate sequences were chosen on the basis of having ENCODE HepG2 ChIP-seq peaks for EP300 and H3K27ac, which are generally indicative of enhancer function (Heintzman et al. 2007; Visel et al. 2009). A subset of candidates ("type 1") are centered at ChIP-seq peaks for forkhead box A1 (FOXA1) or FOXA2, known liver pioneer transcription factors (Lupien et al. 2008) or hepatocyte nuclear factor 4 alpha (HNF4A), a nuclear receptor involved in lipid metabolism and gluconeogenesis (Watt et al. 2003), while also overlapping with ENCODE-derived ChIP-seq peaks for the cohesin complex (RAD21 and SMC3) or chromodomain helicase DNA binding protein 2 (CHD2), a chromatin remodeler that is part of the SWI/SNF complex. Other subsets of candidates were required to overlap only a liver transcription factor peak ("type 2"), only a chromatin remodeler peak ("type 3"), or neither ("type 4"). The 204 control sequences comprised 200 synthetically designed controls from a previous study (synthetic regulatory element sequences (SRESs); 100 positive & 100 negative) (Smith et al. 2013) and an additional 2 positive (pos1 and pos2) and 2 negative endogenous controls (neg1 and neg2). We confirmed by standard

luciferase reporter assay that pos1 and pos2 showed weak and strong enhancer activity, respectively, while neg1 and neg2 showed no activity (Fig. S2).

Each of the 2,440 enhancer candidates or controls was synthesized in *cis* with 100 unique reporter barcodes on a 244,000-feature microarray (Agilent OLS; 15 bp primer + 171 bp enhancer candidate or control + 14 bp spacer + 15 bp barcode + 15 bp primer = 230-mers). The purpose of encoding a large number of barcodes per assayed sequence was to facilitate reproducible and quantitative measurements of regulatory activity, as well as to mitigate against non-uniformity in oligonucleotide synthesis. We cloned these oligonucleotides to a version of the lentiMPRA vector that lacked mP and EGFP reporter. Subsequently, a restriction site in the spacer was used to reinsert the mP + EGFP cassette between the candidate enhancer and barcode, thus positioning the barcode in the 3' UTR of EGFP (Fig. S3)

To evaluate the quality of the designed oligonucleotides and the representation of individual barcodes, we sequenced the cloned oligonucleotide library (i.e. prior to reinsertion of the mP + EGFP cassette) to a depth of 19.2 million paired-end consensus sequences, 52.6% of which had the expected length. Analysis of these data showed that most molecular copies of a given oligonucleotide are correct, that synthesis errors are distributed evenly along the designed insert sequence, and that single base deletions dominate the observed errors (Fig. S4A). Nonetheless, there was substantial non-uniformity in the library (Fig. S4B). While 90.5% of the 244,000 designed barcodes were observed at least once amongst 11.0 million full-length barcodes sequenced, their abundance is sufficiently dispersed that we estimated that a subset of 56-67% of the designed oligonucleotides would be propagated when maintaining a library complexity of 350,000-600,000 clones.

Chromosomally integrated versus episomal lentiMPRA

We next sought to directly compare the functional activities of the 2,236 candidate liver enhancer sequences in a chromosomally integrated versus an episomal context. To this end, we packaged the lentiMPRA library with either a wild-type integrase (WT-IN) or a mutant integrase (MT-IN), with the latter allowing for the production of non-integrating lentivirus and transient transgene expression from non-integrated DNA (Leavitt et al. 1996; Nightingale et al. 2006) (Fig. 1A). Because the integrase is not encoded by the lentiMPRA library, this experimental design allows us to test the same exact library in both integrated and non-integrated contexts.

To reduce background of unintegrated lentivirus in the integrating lentivirus prep, we obtained DNA/RNA from the cells with the WT-IN liver enhancer library at day 4 when they have an estimated 50 viral particles/cell and the MT-IN library at day 3 when they had an estimated 100 viral particles/cell (Fig. S5A, see Methods for details). The total copy number of viral DNA in the cells infected with the liver enhancer libraries was validated by qPCR (Fig. S5B). During human

immunodeficiency virus (HIV) infection, non-integrating virus represents a major portion of the virus at early infection time points and includes linear DNA that is rapidly degraded along with circular DNA containing terminal repeats (1-LTRc and 2-LTRc) (Munir et al. 2013). We further confirmed the copy number of non-integrated virus at our assayed time points by carrying out a qPCR on 2-LTRc, observing the expected low and high amounts of non-integrated virus with WT-IN and MT-IN, respectively (Fig. S5B).

lentiMPRA on 2,236 candidate liver enhancer sequences

We recovered RNA and DNA from both WT-IN and MT-IN infections (three replicates each consisting of independent infections with the same library), amplified barcodes, and performed sequencing (Illumina NextSeq). We used both the forward and reverse reads to sequence the 15 bp reporter barcodes and obtain consensus sequences. We obtained an average of ~4.1 million raw barcode counts for DNA and an average of ~26 million raw barcode counts for RNA. Across replicates and sample types, ~97% of barcodes were the correct length of 15 bp.

We matched the observed barcodes against the designed barcodes and normalized RNA and DNA for different sequencing depths in each sample by dividing counts by the sum of all observed counts and reporting them as counts per million. Only barcodes observed at least once in both RNA and DNA of the same sample were considered. Subsequently, RNA/DNA ratios were calculated. The average Spearman's rho for DNA counts of the three integrase mutant (MT) experiments was 0.907, and for RNA counts of the MT experiments was 0.982. The average Spearman's rho values for the wild-type integrase (WT) experiments were 0.864 and 0.979 for DNA and RNA, respectively. These correlations were determined for barcodes observed in pairs of replicates. Scatter plots for the MT and WT experiments are shown in Fig. S6 and Fig. S7, respectively.

While the DNA and RNA counts for individual barcodes are highly correlated between experiments, the noise of each measure results in a poor correlation of RNA/DNA ratios (Fig. S6, Fig. S7). However, there are on average 59-62 barcodes per candidate enhancer sequence (insert) in each replicate (out of 100 barcodes programmed on the array, with ~40% lost during cloning as discussed above) (Fig. S8). To reduce noise, we summed up the RNA or DNA counts across all associated barcodes for each insert observed in a given experiment and recalculated RNA/DNA ratios (Fig. S9). After this step, pairwise-correlations of DNA and RNA counts of replicates are very high (average Spearman's rho MT-RNA 0.996, MT-DNA 0.994, WT-RNA 0.997 and WT-DNA 0.991). Fig. 2 shows scatter plots and correlation values for per-insert RNA/DNA ratios for the MT and WT experiments. RNA/DNA ratios show markedly improved reproducibility after summing across barcodes, with an average Spearman's rho of 0.908 (MT) and 0.944 (WT). In all pairwise comparisons of replicates, the integrated (WT) MPRA experiments exhibit a broader dynamic range and greater reproducibility than the episomal (MT) MPRA experiments. We also

explored how stable the correlation of RNA/DNA ratios is between replicates by down-sampling the number of barcodes per insert or specifying an exact number of barcodes per insert (Fig. S10). Again, the WT experiments show greater reproducibility, especially for inserts represented by fewer independent barcodes.

To combine replicates, we normalized the RNA/DNA ratios for inserts observed in each replicate by dividing by their median, and then averaged this normalized RNA/DNA ratio for each insert across replicates (red box in Fig. 2; Fig. 3A). Fig. 3A shows scatter plots of the resulting MT and WT RNA/DNA ratios colored by the type of insert and/or transcription factors considered in the design (Fig. S11 shows RNA/DNA ratio ranges by type of insert). As noted above, we observe a broader dynamic range in the WT experiment. Furthermore, the Spearman correlation between MT and WT is 0.785, which is considerably lower than the correlation observed when correlating replicates of the same experimental type (Spearman correlation of 0.908 (MT) and 0.944 (WT)). This is also the case in pairwise comparisons of MT versus WT replicates (i.e. prior to combining replicates) (yellow boxes in Fig. 2). Overall, these results show that there are substantial differences in regulatory activity between identical sequences assayed in an integrated vs. episomal context.

Importantly, we can see clear separation of positive and negative controls. Fig. 3B and 3C display RNA/DNA ratios obtained for the highest and lowest SRESs in the MT and WT experiments compared to their previously measured effects in HepG2. While the highest and lowest SRESs are well separated in both experiments (Kolmogorov-Smirnov and Wilcoxon Rank Sum p-values below 2.2×10^{-16}), the WT experiment separates the highest and lowest SRE controls slightly better than the MT experiment (Kolmogorov-Smirnov test D 0.97 vs 0.95, Wilcoxon Rank Sum test W 9951 vs 9937).

We next sought to assess whether any of our design categories (i.e. types 1-4 defined above, reflecting subsets of candidate enhancers with coincident liver TF and/or chromatin remodeler/cohesin complex ChIP-seq peaks) might underlie the observed differences (Figures 3A and S11). We did not observe differences in expression range between type 1 and type 2 designs, which differ with respect to whether they overlap with ChIP-seq peaks for chromatin remodeling factors or members of the cohesin complex (Table S1; Fig. S16). Unexpectedly, type 3 and 4 designs, which unlike type 1 and 2 designs lack liver TF ChIP-seq peaks for FOXA1/2 and HNF4A, were more active. To attempt to explain overall differences in activity both for the different category types and in general, we next explored a broader set of genomic annotations and whether those are predictive of enhancer activity in the lentiMPRA assay.

ENCODE and other genomic annotations that predict enhancer activity

We evaluated whether genomic annotations, some numerical and other categorical (Supplementary File 3), were predictive of our results in HepG2 cells. The performance of individual numerical annotations for predicting the observed activity of candidate enhancer sequences are shown in Fig. 4. We use Kendall's tau, a non-parametric rank correlation that is more conservative than Spearman's rho, because of the large number of zero-values in our annotations which can result in artifacts from ties with Spearman's rho. In contrast with our design bins, many genomic annotations are observed to predict enhancer activity in both the WT and MT experiments. Across the board, annotations correlate better with the WT than the MT results, suggesting that integrated activity read-outs (WT) correlate better with endogenous functional genomic signals (e.g. ChIP-seq data) than do episomal activity read-outs (MT).

The most highly predictive numerical annotations, in both types of experiments, are HepG2 ChIP-seq datasets of JUND (Transcription Factor Jun-D) and FOSL2 (FOS-Like Antigen 2), consistent with a previous MPRA study which also highlighted the role of these transcription factors in HepG2 cells (Savic et al. 2015). For chromosomally based MPRA (WT), the number of overlapping ENCODE ChIP-seq peaks (TFBS) and the average ENCODE ChIP-seq signal (TFBSPeaks) as measured across different cell lines also rank amongst the more highly predictive annotations. However, these same features are the most discrepant with MT; that is, substantially less predictive of episomal MPRA. Of note, the highest observed T^2 for an individual annotation is only 0.033 (MT) and 0.059 (WT), highlighting the need for a model combining annotations and other available information (see below).

We also evaluated whether categorical annotations were predictive of our results (Fig. S12-S15). Most of these annotations were derived for HepG2 cells by the ENCODE project (ChromHMM (Ernst and Kellis 2012), SegWay (Hoffman et al. 2012) and Open Chromatin annotation). Although none of these were strongly predictive of the measured RNA/DNA ratios, the best performance was observed with the 25-state multi-cell-type SegWay or 15-state HepG2 ChromHMM chromatin segmentations. For SegWay, sequences annotated as TSS (transcription start sites) exhibited the highest expression while sequences annotated as D (genomic death zones) exhibited the lowest expression.

Sequence-based predictors of functional activity

Because sequences are removed from their native genomic context in the lentiMPRA assay, the relative ability of the elements that we test to increase transcription must be inherent to the 171 bp sequences themselves. It is therefore natural to ask whether it is possible to build models for predicting functional activity that are based on primary sequence, rather than on genomic annotations. Although such models have historically been challenging to develop for mammalian genomes, Ghandi, Lee *et al.* recently introduced a "gapped *k*-mer" approach (gkm-SVM), with promising results for discriminating ENCODE ChIP-seq peaks vs. matched control sequences

(Ghandi et al. 2014). In brief, gkm-SVM trains a binary classifier based on the sets of *k*-mers that the training data contains (but allowing for some non-informative positions within *k*-mers, *i.e.* gaps). We collected all training data that Ghandi, Lee *et al.* used for HepG2, obtaining ~225,000 unique peak sequences as well as controls, and trained a combined, sequence-based model for predicting ChIP-seq peaks in HepG2 cells (see Methods). Based on a set-aside test dataset, the gkm-SVM model had a specificity of 71.8%, a sensitivity of 88.8% and a precision of 75.9% for separating ChIP-seq peak sequences from random control sequences.

We asked how well the resulting gkm-SVM scores correlated with the RNA/DNA ratios obtained for the MT and WT experiments (Fig. 5A-B). The combined gkm-SVM HepG2 model results in a Spearman's R^2 of 0.080 and 0.128, for MT and WT respectively. However, this correlation is at least partially driven by the synthetic control sequences, which can be scored with the sequence-based model but not with the genomic annotations. When excluding all control sequences, Spearman's R^2 values drop from 0.080 to 0.039 and from 0.128 to 0.076 for MT and WT, respectively. As such, there are a few ENCODE-based annotations which outperform the sequence-based gkm-SVM model, namely summaries of JUND/FOSL2 HepG2 ChIP-seq peaks, the number of overlapping ChIP-seq peaks (TFBS) or the average ChIP-seq signal (TFBSPeaks) measured across multiple ENCODE cell-types.

Combining annotations and sequence information to predict enhancer activity

We next sought to combine information across multiple annotations to better predict enhancer activity. We fit Lasso linear models and selected the Lasso tuning parameter value by cross-validation (CV). Scatter plots as well as correlation coefficients were also obtained in a CV setup (see Methods). We built models with all the genomic annotations described above (including the categorical annotations as binary features) as well as with and without the sequence-based gkm-SVM score from scaled and centered annotation matrixes (Fig. S17-19). SRESs and other controls were naturally excluded, as they are largely synthetic sequences and therefore missing genomic annotations. The resulting linear models were considerably more predictive of WT ratios than MT ratios (e.g. CV Spearman R^2 of 0.271 WT vs. 0.148 MT; CV Pearson R^2 of 0.314 WT vs. 0.194 MT). Including gkm-SVM scores in the models improved performance further (CV Spearman R^2 of 0.302 WT vs. 0.156 MT; CV Pearson R^2 of 0.341 WT vs. 0.203 MT). We noticed that gkm-SVM scores were assigned the largest model coefficients in both WT and MT models when they were included (Fig. S19). Thus, while reasonably performing models are obtained from genomic annotations, the sequence-based gkm-SVM scores appear to capture independently predictive information.

We therefore decided to further explore sequence-based models with an improved implementation of gkm-SVM ('LS-GKM') (Lee 2016). We trained models from each of the 64 narrow-peak ChIP-seq datasets for which we had included summary statistics for the annotation matrix above (see

Methods). We then asked how well the LS-GKM scores generated by each of these 64 models predicted the results of the lentiMPRA experiments. Although the scores now correspond to sequence-based models of ChIP-seq peaks rather than the ChIP-seq peaks themselves, we once again observed the highest Spearman R^2 values for the individual factors JUND (0.120 WT/0.050 MT) and FOSL2 (0.108 WT/0.049 MT), and these are also the factors that show the largest differences in predictive value for WT vs. MT (Fig. S19). As such, sequence-based models of binding by these two factors as well as other individual factors exceed the performance of the pooled gkm-SVM sequence model.

We next fit Lasso linear models from the TF-specific LS-GKM SVM scores in order to predict the measured activities. The combined MT model (using 37 individual scores) achieves a CV Spearman R^2 of 0.128 (CV Pearson R^2 of 0.162), and the combined WT model (using 34 individual scores) of 0.227 (CV Pearson R^2 of 0.261) (Fig. S21). This still falls short of models obtained purely from genomic annotations as described above. To test whether multiple ChIP-seq datasets should be combined in a sequence model rather than combining individual model scores in a linear model to improve prediction, we also trained LS-GKM models based on the peak sequences of the 37 (MT) and 34 (WT) scores selected by Lasso models as well as the top 2, 3, 5, 10 and top 15 coefficients in the Lasso models for MT and WT. We included coefficients independent of their direction (sign), but also trained only with sequences corresponding to a positive coefficient. Model performance increased for combining small numbers of peak sets and when limiting it to positive coefficients (Table S2). Combining all or too many peaks in one sequence model reduced overall performance. From the combinations of ChIP-seq datasets tested here, the best performing sequence-based models achieved Spearman R^2 values of 0.054 (MT) and 0.133 (WT).

Finally, when we used both genomic annotations and the individual LS-GKM scores in a single linear model to predict the measured activities, performance increased to a CV Spearman R^2 of 0.168 (MT; Pearson R^2 of 0.206) and 0.322 (WT; CV Pearson R^2 of 0.362). These are our highest performing models predicting the activities of candidate enhancer sequences for both the episomally and chromosomally encoded MPRA experiments (Fig. 5C-D).

Discussion

In this work, we report the first systematic comparison of episomal and chromosomally integrated reporter assays. Key aspects of our approach include: (1) Lentivirus-based MPRA or lentiMPRA, which can be used to in an episomal or integrated context by toggling whether a mutant vs. wild-type integrase is used, and can furthermore be used in a wide variety of cell types, including neurons; (2) The use of numerous barcodes per candidate enhancer sequence, which results in highly reproducible measurements of transcriptional activation; and (3) Extensive predictive modeling of our results, with the implicit assumption that a reasonable measure of a reporter

assay's biological relevance is the extent to which it is correlated with endogenous genomic annotations.

We find that the results of integrated reporter assays are more reproducible, robust and biologically relevant than episomal reporter assays. These conclusions are supported by the following observations: (1) We observed consistently greater reproducibility and dynamic range for the WT replicates as compared with the MT replicates. (2) The correlation of WT vs. MT replicates (Spearman correlation of 0.785) was substantially lower than for WT vs. WT (0.944) or MT vs. MT (0.908), with clear systematic differences between the integrated and episomal contexts that exceed technical noise (Fig. 2, Fig. S9). (3) The WT experiments were consistently more correlated with and more predictable by genomic annotations, which are based on biochemical marks measured in these sequences' native genomic contexts. (4) Many genomic annotations significantly predict the results of the WT but not the MT experiments.

Of note, we observed generally higher levels of expression with integrated reporters (Fig. 3A and Table S1), consistent with previous findings that showed higher reporter gene levels for integrating relative to non-integrating HIV-1 (Gelderblom et al. 2008; Thierry et al. 2016). However, it is worth noting that we used a lentivirus (not HIV-1) with a self-inactivating (SIN) LTR, which lacks viral promoters or enhancers, potentially influencing these expression differences. We also observed a larger variability between replicates for the MT condition. This variability could be explained by differences in histone structure, our assay being geared more for the WT condition (i.e. high MOI, many barcodes per sequence, anitrepressors), different time points (3 days for MT and 4 for WT), suppression of transient plasmids (Qiu et al. 2011) and other factors. Results from hydrodynamic tail vein assays (which delivers reporter constructs into the mouse liver) also show that when chromatinized plasmid DNA leads to higher expression levels than naked plasmid DNA (Kamiya et al. 2013). For HIV-1, both integrating and non-integrating HIV-1 viral DNA are associated with histones (Kantor et al. 2009), which is probably also the case for our lentiviral vector. However, even if the lentiviral episome is chromatinized, there remain myriad potential causes for the observed differences in expression, including differences in H1 stoichiometry, nucleosome positioning, cooperative TF binding (Hebbar and Archer 2007; Hebbar and Archer 2008), and/or nuclear location (Jeong and Stein 1994).

The number of overlapping ENCODE ChIP-seq peaks was one of the most strongly predictive annotations for our integrated sequences (Fig. 4, left). Interestingly, in experiments previously performed on the *MMTV* promoter it was observed that non-integrating constructs could not adequately assess cooperative TF binding due to differences in H1 stoichiometry and nucleosome positioning (Hebbar and Archer 2007), which may relate to the fact that this multiple TF binding was also one of the most differentiating annotations between the WT vs. MT experiments (Fig. 4, right). Specific TF ChIP-seq-based sequence models that are similarly differentiating (Fig. S19) include JUND, FOSL2, ATF3 and ELF1, which are known to interact and form complexes. Jun and Fos family members form the heterodimeric protein complex AP-1, which regulates gene expression in response to various stimuli including stress (Hess et al. 2004) and in the liver has

known roles in hepatogenesis (Hilberg et al. 1993) and hepatocyte proliferation (Alcorn et al. 1990). AP-1 is known to form complexes with several additional protein partners (Hess et al. 2004), including ATF proteins such as ATF3 (Hai and Curran 1991) and ELF1, an ETS transcription factor (Bassuk and Leiden 1995). Of note, while lentivirus infection can induce stress potentially leading to increased expression of AP-1 and related factors, these same TFs were less predictive in MT-infected sequences, and furthermore the ChIP-seq datasets were generated on cells in normal physiological conditions. Combined, these findings suggest that differences in cooperative TF binding, possibly involving TFs including JUND, FOSL2, ELF1 and ATF3, might drive differences in the results of integrated vs. episomal reporter assays. Interestingly, we observed a few histone related predictive annotations for the MT high expressing signals, including the SIN3 transcriptional regulator family member B (SIN3B), a scaffold protein that recruits chromatin modifying proteins (Kadamb et al. 2013) and histone variant H2A.Z, which has been shown to have important transcriptional activation roles (Subramanian et al. 2015). While these annotations are based on the respective genomic DNA regions (e.g. ChIP-seq results) and only H1 stoichiometric differences were previously observed for the transient versus integrated MMTV promoter (Hebbar and Archer 2008), it would be interesting to test whether other histone differences exist for non-integrating constructs on a larger scale.

Using single-feature models, we also systematically evaluated more than 400 genomic annotations and sequence models to explore which are significantly predictive of expression in the integrated and/or episomal lentiMPRA experiments (Figure S22). Consistent with our other analyses, there are many more annotations that are significantly predictive of the integrated (WT) assay but not the episomal (MT) assay. These include several annotations related to histone acetylation (HDAC2, EP300, and ZBTB7A) as well as a factor (CEBPB) with increased liver expression and which is associated with adipogenesis, gluconeogenic and hematopoiesis (Tsukada et al. 2011). Interestingly, there are also a number of annotations which are only significantly predictive of the episomal assay, including BRCA1 ChIP-seq peaks and SIN3 transcriptional regulator family member B (SIN3B) binding motifs.

A limitation of this study and most contemporary MPRA experiments is that the length of the sequences tested are less than 200 bases. This may be addressable in the future through improvements to array-based oligonucleotide synthesis, multiplex DNA assembly protocols (Klein et al. 2015), or multiplex DNA capture protocols (Vockley et al. 2015). Furthermore, it remains uncertain what the true length distribution of enhancers is, although the size distribution of distal DNase hypersensitive sites seems to provide a reasonable proxy of approximately 300bp (Natarajan et al. 2012).

A contemporary challenge for our field is how to best identify, prioritize, and functionally validate *cis*-regulatory elements, especially enhancers. To address this, we envision a virtuous cycle, in which annotation and/or sequence-based models are used to nominate candidate enhancer sequences for validation, these candidates are tested in massively parallel reporter assays, and then the results are used to improve the models which in turn results in higher quality nominations. Eventually, this will lead to not only a catalog of validated enhancers but also a deeper mechanistic

understanding of the relationship between primary sequence, transcription factor binding, and quantitative enhancer activity. In this study, our best performing model achieves a Pearson's R^2 of 0.362 in predicting the results of the integrated lentiMPRA, with both genomic annotations and sequence-based models providing independent information. The result is encouraging, but the majority of variation remains unexplained. Of note, these are quantitative predictions of activity, a more challenging task than categorizing enhancers vs. non-enhancers. Nonetheless, all the information underlying the differences is contained within the short sequences that we are testing, and in principle should be learnable. Possible avenues for improving our predictive models include increasing the size of the regions tested so as to better align with the genomic annotations that we are using for prediction, additional annotations or alternative ways of summarizing available annotations, alternative sequence representations to gapped *k*-mers, allowing for interaction terms between model features, and markedly expanding the amount of training data. Indeed, with sufficient training data, we hope that it will eventually be possible to accurately predict MPRA results from sequence-based models alone.

As our field scales MPRAs to characterize very large numbers of candidate enhancers, it is obviously critical that the reporter assays are as reflective as possible of endogenous biology. Our results directly test a longstanding concern about episomal reporter assays, and suggest that there are substantial differences between the integrated and episomal contexts. Furthermore, based on the fact that their output is more correlated with genomic annotations, we infer that integrated reporter assays are more reflective of endogenous enhancer activity. This fits with our expectation, as both the integrated reporter and endogenous enhancers reside within chromosomes as opposed to episomes.

Of course, an equally valid perspective on our observations is that reporter assays in the integrated vs. episomal contexts are reasonably well correlated, and therefore the results of episomal assays remain informative of what activity would be in the integrated context (e.g. sequences observed to be strongly active in an episomal assay are likely to also be strongly active in an integrated assay). Whether the differences that we observe here, with respect to activity, reproducibility, and correlation with genomic features, impact the interpretation of an episomal assay depends on the particulars of what was done and why (e.g. these differences may be more relevant for enhancer prediction and for quantitative modeling than for the binary classification of individual elements as activating vs. inert). Furthermore, it is important to acknowledge that even integrated reporter assays are limited in important ways (e.g. by removing each tested element from its native sequence and epigenetic context). We therefore urge caution in the interpretation of the results of all reporter assays, but also that integrated reporter assays such as lentiMPRA be used where possible as an alternative to episomal reporter assays.

Methods

Lentivirus enhancer construct generation

To generate the lentivirus vector (pLS-mP), a minimal promoter sequence, which originates from pGL4.23 (Promega), including an *Sbf*I site was obtained by annealing of oligonucleotides (Sense: 5'- CTAGACCTGCAGGCACTAGAGGGTATATAATGGAAGCTCGACTTCCAGCTTGG CAATCCGGTACTGTA-3', Antisense: 5'- CCGGTACAGTACCGGATTGCCAAGCTGGAA GTCGAGCTTCCATTATATACCCTCTAGTGCCTGCAGGT-3'; *Sbf*I site is underlined), and subcloned into *Xba*I and *Age*I sites in the pLB vector (Addgene 11619; (Kissler et al. 2006)) replacing the U6 promoter and CMV enhancer/promoter sequence in the vector. To generate pLS-mP-SV40, the SV40 enhancer sequence was amplified from pGL4.13 (Promega) using primers (Forward: 5'- CAGGGCCCGCTCTAGAGCGCAGCACCATGGCCTGAA-3', Reverse: 5'-TGCCTGCAGGTCTAGACAGCCATGGGGCGGAGAATG-3') and inserted into *Xba*I site in the vector using In-Fusion (Clontech). pos1, pos2, neg1, and neg2 sequences were amplified from human (pos1, neg2, pos2) or mouse (neg1) genome, and inserted into *Eco*RV and *Hin*dIII site in pGL4.23 (Promega). Primers used are shown in Table S3 and the annotated plasmid sequence file is available as Supplementary File 1.

Library sequence design

We picked 171bp candidate enhancer sequences based on ChIP-seq peaks calls for HepG2. We used narrow peak calls for DNA binding proteins/transcription factors (FOXA1, FOXA2, HNF4A, RAD21, CHD2, SMC3 and EP300) and wide peak calls for histone marks (H3K27ac). We downloaded the call sets from the ENCODE portal (Sloan et al. 2016) (https://www.encodeproject.org/) with the following identifiers: ENCFF001SWK, ENCFF002CKI, ENCFF002CKJ, ENCFF002CKK, ENCFF002CKN, ENCFF002CKY, ENCFF002CUS, ENCFF002CTX, ENCFF002CUU, ENCFF002CKV, and ENCFF002CUN. We defined four classes of sites: 1) Regions centered over peak calls of ≤171bp for FOXA1, FOXA2 or HNF4A that overlap H3K27ac and EP300 calls as well as at least one of three factors: components of the Cohesin complex (RAD21, SMC3) or chromatin remodeling factor CHD2 . 2) Regions like in 1, but with no CHD2, RAD21 or SMC3 overlap. 3) Regions of 171 bp centered in an EP300 peak overlapping H3K27ac as well as at least one of three factors RAD21, CHD2 or SMC3, but without peaks in FOXA1, FOXA2 or HNF4A. 4) Regions like in 3 but with no CHD2, RAD21 or SMC3 overlap. Sites of type 1 and 2 involving HNF4 were the most abundant sites and we used those to fill-up our design after exhausting other target sequences.

Potential 171bp target sequences were inserted into a 230bp-oligo backbone with a 5'-flanking sequence (15bp, AGGACCGGATCAACT), 14bp-spacer sequence (CCTGCAGGGAATTC), 15bp designed tag sequences (see below) and a 3'-flanking sequence (15bp, CATTGCGTGAACCGA) (Supplementary Fig. Y). Sequences were checked for *Sbf*I and *Eco*RI restriction sites after joining the target sequence with the 5'-flanking sequence and the spacer sequence. Such potential target sequences were discarded.

In our final array design, we included 2,440 different target sequences each with 100 different barcodes (i.e. a total of 244,000 oligos). These included the highest 100 and lowest 100 synthetic regulatory element (SRE) sequences identified by Smith RP et al. (Smith et al. 2013), 4 control sequences (neg1 MGSCv37 chr19 35,531,983-35,532,154, neg2 GRCh37 chr5 172,177,151-172,177,323, pos1 GRCh37 chr3 197,439,136-197,439,306, pos2 GRCh37 chr19 35,531,984-35,532,154) which we tested using Luciferase assays in the HepG2 cell line (Fig. S2), 1,029 type 1 inserts (202 FOXA1, 180 FOXA2, 464 HNF4A, 120 FOXA1&FOXA2, 33 FOXA1&HNF4A, 17 FOXA2&HNF4A, 13 FOXA1&FOXA2&HNF4A), 1,030 type 2 inserts (195 FOXA1, 174 FOXA2, 470 HNF4A, 126 FOXA1&FOXA2, 31 FOXA1&HNF4A, 20 FOXA2&HNF4A, 14 FOXA1&FOXA2&HNF4A), 90 type 3 inserts and 87 type 4 inserts.

Tag sequences of 15bp length were designed to have at least two substitutions and one 1bp-insertion distance to each other. Homopolymers of length 3 bp and longer were excluded in the design of these sequences, and so were ACA/CAC and GTG/TGT trinucleotides (bases excited with the same laser during Illumina sequencing). More than 556,000 such barcodes were designed using a greedy approach. The barcodes were then checked for the creation of *Sbf*I and *EcoR*I restriction sites when adding the spacer and 3'-flanking sequences. From the remaining sequences, a random subset of 244,000 barcodes was chosen for the design. The final designed oligo sequences are available in Supplementary File 2.

Generation of MPRA libraries

The lentiviral vector pLS-mP was cut with *Sbf*I and *Eco*RI to temporarily liberate the minimal promoter and EGFP reporter gene. Array-synthesized 230bp oligos (Agilent Technologies) containing an enhancer, spacer, and barcode (Fig. S3) were amplified with adaptor primers (pLSmP-AG-f and pLSmP-AG-r) that have overhangs complementary to the cut vector backbone (Table S3), and the products were cloned using NEBuilder HiFi DNA Assembly mix (E2621). The adaptors were chosen to disrupt the original *Sbf*I and *Eco*RI sites in the vector. The cloning reaction was transformed into electrocompetent cells (NEB C3020). Multiple transformations were pooled and midi-prepped (Chargeswitch Pro Filter Plasmid Midi Kit, Invitrogen CS31104). This library of cloned enhancers and barcodes was then cut using *Sbf*I and *Eco*RI sites contained within the spacer, and the minimal promoter and *EGFP* that were removed earlier were reintroduced via a sticky end ligation (T4 DNA Ligase, NEB M0202) between the enhancer and its barcode. These finished vectors were transformed and midi-prepped as previously mentioned.

Quality control of designed array oligos

Before inserting the minimal promoter and EGFP reporter gene, the plasmid library was sampled by high-throughput sequencing on an Illumina MiSeq (206/200+6 cycles) to check for the quality

of the designed oligos and the representation of individual barcodes (sequencing primers are pLSmP-AG-seqR1, pLSmP-AG-seqIndx, and pLSmP-AG-seqR2; Table S3). We sequenced the target, spacer and tag sequences from both read ends and called a consensus sequence from the two reads. We obtained 19.2 million paired-end consensus sequences from this sequencing experiment. 52.6% of those sequences had the expected length, 26.1% of sequences were 1bp short and 8.9% were 2bp short (summing up to 87.6%). Only 1.6% of sequences showed an insertion of 1bp. These results are in line with expected dominance of small deletion errors in oligo synthesis. We aligned all consensus sequences back to all designed sequences using BWA MEM (Li and Durbin 2009) with parameters penalizing soft-clipping of alignment ends (-L 80). We consensus called reads aligning with the same outer alignment coordinates and SAM-format CIGAR string to reduce the effects of sequencing errors. We analyzed all those consensus sequences based on at least three sequences with a mapping quality above 0. We note that substitutions are removed in the consensus calling process if the correct sequence is the most abundant sequence. Among these 992,513 consensus sequences, we observe instances of 91% designed oligos and 78% of oligos with one instance matching the designed oligo perfectly. Across all consensus sequences the proportion of perfect oligos is only 19%, however the proportion vastly increases with the number of observations (69% at 20 counts, 99% at 40 counts; Table S4). These observations are in agreement with most molecular copies of an oligo being correct, in combination with high representation differences in the library. Supplementary Fig. S4A shows the distribution of alignment differences (as a proxy for synthesis errors) along the designed oligo sequences. Errors are distributed evenly along the designed insert sequence, with deletions dominating the observed differences. We observe that at some positions the deletion rate is reduced while the insertion rate is increased. We speculate that this might be due to certain sequence contexts.

Limited coverage of designed oligos in MPRA libraries

From the analysis of oligo quality and oligo abundance above, we saw a first indication of the existence of a wide range of oligo abundance and that more frequent sequences tend to match the designed sequences perfectly (see Table S4). We characterized the abundance of oligos further and looked at the consequences that this has for generating libraries of lentivirus constructs with limited complexity (due to the transformation of a limited number of bacteria). Rather than looking at full length oligos, we focused only on the tag sequences. Tag sequences were identified from the respective alignment positions of the alignments created above. To match the RNA/DNA count data analysis (see below), we only considered barcodes of 15bp length (10.96M/57.0%, similar to the proportion of correct length sequences above). Of those 10.96M barcodes, 345,247 different sequences are observed. We clustered (dnaclust (Ghodsi et al. 2011)) the remaining sequences allowing for one substitution and selecting the designed or most abundant sequence (reducing to 238,206 different sequences). The clustered sequences were matched against the designed barcodes (217,176 sequences, 99.2% of counts). The distribution of the abundance of these barcodes is available in Supplementary Fig. S4B. We used those counts to simulate sampling from

this over dispersed pool of sequences, as done when taking a sample of plasmids infusing the reporter gene and minimal promoter and again transforming the resulting plasmids. We sampled 10 times and averaged the number of unique designed barcodes: 150k clones – 87,944 unique barcodes, 250k clones – 116,297 unique barcodes, 350k clones – 135,222 unique barcodes, 500k clones – 154,090 unique barcodes, 600k clones – 163,831 unique barcodes, 750k clones – 172,770 unique barcodes and 1M clones – 183,685 unique barcodes. Thus, even for high sampling depth only a subset of barcodes will be captured in the final library. We observe on average 145,876 different barcodes which is concordant with more than 430k clones going into the lenti construction.

Cell culture and GFP / luciferase assays

HepG2 cells were cultured as previously described (Smith et al. 2013). K562, H1-ESC, HeLa-S3, T-47D and Sk-n-sh cells were culture as previously described (Consortium et al. 2012). Sk-n-sh cells were treated with 24 μM *all trans*-retinoic acid (Sigma) to induce neuronal differentiation. K562, H1-ESC, HeLa-S3, T-47D and Sk-n-sh were treated with retinoic acid and infected with pLS-mP or pLS-SV40-mP lentivirus along with 8 μg/ml polybrene, and incubated for 2 days, when they have an estimated 30, 60, 90, 90, and 90 viral particles/cell, respectively. The number of viral particles/cell was measured as described below. For the four control sequences (two negatives and two positives) luciferase assay, we amplified the controls from the designed oligo pool (primer sequences available in Table S5) and inserted those into the pGL4.23 (Promega) reporter plasmid. $2x10^4$ HepG2 cells/well were seeded in a 96-well plate. 24 hour later, the cells were transfected with 90 ng of reporter plasmids (pGL4.23-neg1, pGL4.23-neg2, pGL4.23-pos1, and pGL4.23-pos2) and 10 ng of pGL4.74 (Promega), which constitutively expresses Renilla luciferase, using X-tremeGENE HP (Roche) according to the manufacturer's protocol. The X-tremeGENE:DNA ratio was 2:1. Three independent replicate cultures were transfected. Firefly and Renilla luciferase activities were measured as previously described (Smith et al. 2013).

Lentivirus packaging, titration and infection

To optimize conditions and reduce background of unintegrated lentivirus in the integrating lentivirus prep, we utilized our positive control virus (pLS-SV40-mP) that was packaged with WT-IN and MT-IN, and examined the viral titer by qPCR for three different volumes (1, 5 and 25 μl per well of a 24-well plate) at four different time points (2-5 days post infection). For the lower volumes (1 and 5 μl), we observed a substantial reduction in total virus amounts at day 4 for both MT-IN and WT-IN that stabilized in the WT-IN only (Fig. S5A). This suggests that the non-integrated virus declines at this time point, similar to what was previously reported (Butler et al. 2001). For the high volume (25 μl), we did not observe a substantial reduction or stabilization for MT-IN and WT-IN respectively until day 5 (Fig. S5A), suggesting that high amounts of virus would make it difficult to distinguish between integrated and non-integrated virus.

Twelve million HEK293T cells were plated in 15 cm dish and cultured for 24 hours. The cells were co-transfected with 8 μg of the liver enhancer library and 4 μg of packaging vectors using jetPRIME (Polyplus-transfections). psPAX2 that encodes wild-type *pol* was used to generate integrating lentivirus, while pLV-HELP (InvivoGen) that encodes a mutant *pol* was used to generate non-integrating lentivirus. pMD2.G was used for both. The transfected cells were cultured for 3 days and lentivirus were harvested and concentrated as previously described (Wang and McManus 2009).

To measure DNA titer for the integrating and non-integrating lentivirus library, HepG2 cells were plated at $2x10^5$ cells/well in 12-well plates and incubated for 24 hours. Serial volume (0, 1, 5, 25 μL) of the lentivirus was added with 8 μg/ml polybrene, to increase infection efficiency. The infected cells were cultured for 2-5 days and then washed with PBS three times. Genomic DNA was extracted using the Wizard SV genomic DNA purification kit (Promega). Copy number of viral particle per cell was measured as relative amount of viral DNA (WPRE region) over that of genomic DNA (intronic region of *LIPC* gene) by qPCR using SsoFast EvaGreen Supermix (BioRad), according to manufacturer's protocol. PCR primer sequences are shown in Table S3. For the lentiMPRA, 2.4 million HepG2 cells were plated on a 10 cm dish and cultured for 24 hours. The cells were infected with integrating or non-integrating lentivirus libraries along with 8 μg/ml polybrene, and incubated for 4 and 3 days, when they have an estimated 50 and 100 viral particles/cell, respectively. Three independent replicate cultures were infected. The cells were washed with PBS three times, and genomic DNA and total RNA was extracted using AllPrep DNA/RNA mini kit (Qiagen). Copy number of viral particle per cell was confirmed by qPCR and shown in Supplementary Fig. S5B. Messenger RNA (mRNA) was purified from the total RNA using Oligotex mRNA mini kit (Qiagen) and treated with Turbo DNase to remove contaminating DNA.

RT-PCR, amplification and sequencing of RNA/DNA

For each replicate, 3x500ng was reverse transcribed with SuperscriptII (Invitrogen 18064-014) using a primer downstream of the barcode (pLSmP-ass-R-i#, Table S3), which contained a sample index and a P7 Illumina adaptor sequence. The resulting cDNA was pooled and split into 24 reactions, amplified with Kapa Robust polymerase for 3 cycles using this same reverse primer paired with a forward primer complementary to the 3' end of EGFP with a P5 adaptor sequence (BARCODE_lentiF_v4.1, Table S3). The implemented two-round PCR set-up is supposed to reduce PCR jack-potting effects and allows for incorporating unique molecular identifiers (UMIs), which could be used to correct for other PCR biases in future experiments. PCR products are then cleaned up with AMPure XP beads (Beckman Coulter) to remove the primers and concentrate the products. These products underwent a second round of amplification in 8 reactions per replicate for 15 cycles, with a reverse primer containing only P7. All reactions were pooled at this point,

run on an agarose gel for size-selection, and submitted for sequencing. For the DNA, 16x500ng of each replicate was amplified for 3 cycles just as the RNA. First round products were cleaned up with AMPure XP beads, and amplified for another 16 reactions, each for 20 cycles. Reactions were pooled, gel-purified, and sequenced. Sequencing primers are BARCODE-SEQ-R1-V4, pLSmP-AG-seqIndx, and BARCODE-SEQ-R2-V4 for both RNA and DNA barcodes (Table S3).

RNA and DNA for each of three replicates was sequenced on an Illumina NextSeq instrument (2x26 + 10bp index). The forward and reverse reads on this run each sequenced the designed 15bp barcodes as well adjacent sequence to correctly trim and consensus call barcodes. We obtained a minimum of 2.9M and a maximum of 5.9M raw counts for DNA (average 4.1M) and a minimum of 20.0M and a maximum of 32.3M raw counts for RNA (average 25.6M). Across replicates and sample type, 97% of barcodes were of the correct length of 15bp.

The number of unique sequences was on average 446k for DNA and 1.2M for RNA. When clustering sequences with one substitution (dnaclust; (Ghodsi et al. 2011)), the average number of unique sequences reduced to 280k for DNA and 697k for RNA. We speculate that our RNA readouts are impacted by sequence errors to a greater extent due to the reverse transcriptase (RT) step. When overlapping the observed with the designed sequences, clustering keeps more counts but reduces the total number of observed barcodes (93.1% vs. 90.3%, 145k vs 151k). We believe this is due to too many errors in barcodes which are sufficiently similar to cause clusters to merge across different designed tag sequences. We therefore dismissed the clustered data and only matched against the designed barcodes. This is further supported by counts being more highly correlated between replicates when using the non-clustered data (Spearman's rho without clustering: DNA replicates 88.6%, RNA replicates 98.0%; with clustering: DNA replicates 85.0%, RNA replicates 94.3%).

Replicates, normalization and RNA/DNA ratios

To normalize RNA and DNA for different sequencing depths in each sample, we divided reads by the sum of observed counts and reported them as counts per million. Only barcodes observed in RNA and DNA of the same sample were considered. Subsequently, RNA/DNA ratios were calculated. We observe that the dynamic range observed in the WT experiments is larger and that the average Spearman's rho is also higher for the WT experiments (44.3% vs. 39.0%). To determine the RNA/DNA ratios per insert, we summed up the counts of all barcodes contributing and determined the ratio of the average normalized counts. We explored how stable the correlation of RNA/DNA ratios is between replicates when limiting the number of barcodes per insert (Fig S10). We limited the maximum number of barcodes considered by (1) randomly down sampling and (2) requiring an exact number of barcodes per insert (i.e. down sampling those with more and excluding those inserts with less barcodes).

Even though normalized individually, the three replicates of each experiment do not seem to be on the exact same scale (Figure 2 & S9). We therefore chose to divide the RNA/DNA ratios by the median across the technical replicate value before averaging them.

Predictors of sequence effects

To correlate available annotations with the observed sequence activity in HepG2 cells, we downloaded additional narrow peak calls for DNA binding proteins/transcription factors in HepG2 from ENCODE data. We obtained call-sets for the following 64 factors: ARID3A, ATF3, BHLHE40, BRCA1, CBX1, CEBPB, CEBPD, CHD2, CTCF, ELF1, EP300, EZH2, FOSL2, FOXA1, FOXA2, FOXK2, GABPA, GATA4, HCFC1, HDAC2, HNF4A, HNF4G, IRF3, JUN, JUND, MAFF, MAFK, MAX, MAZ, MBD4, MXI1, MYBL2, MYC, NFIC, NR2C2, NRF1, POLR2A, POLR2AphosphoS2, POLR2AphosphoS5, RAD21, RCOR1, REST, RFX5, RXRA, SIN3A, SIN3B, SMC3, SP1, SP2, SRF, TAF1, TBP, TCF12, TCF7L2, TEAD4, TFAP4, USF1, USF2, YY1, ZBTB33, ZBTB7A, ZHX2, ZKSCAN1, and ZNF274. Additionally, we downloaded ChromHMM segmentations for HepG2, Open Chromatin State, SegWay, and DHS call-sets from the ENCODE portal (Sloan et al. 2016). From the NIH Roadmap Epigenomics Consortium we obtained RNAseq, DNA methylation, DNase, CAGE, H2A.Z, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K79me2, H4K20me1, and ChromHMM segmentations tracks (Roadmap Epigenomics Consortium et al. 2015). We also downloaded the Fantom5 Robust Enhancer annotations (Andersson et al. 2014), Fantom5 CAGE data for HepG2 (Forrest et al. 2014), GenoSTAN Enhancer and promoter predictions (http://i12g-gagneurweb.in.tum.de/public/paper/GenoSTAN/), enhancerFinder predictions (Erwin et al. 2014), as well as motif scan results and annotated regulatory features from the Ensembl Regulatory Build (Zerbino et al. 2015). For further genome-wide and organismal metrics, we turned to the CADD v1.3 annotation file (Kircher et al. 2014) and extracted local GC and CpG content, SegWay, chromHMM state across the NIH RoadMap cell types, priPhCons, mamPhCons, verPhCons, priPhyloP, mamPhyloP, verPhyloP, GerpN, GerpS, GerpRS, bStatistic, tOverlapMotifs, motifECount, motifEHIPos, TFBS, TFBSPeaks, TFBSPeaksMax, distance to TSS, and the actual CADD score column. We included the number of bases covered by peak calls as well as the average and maximum values across the designed sequences for those metrics. Supplementary File 3 outlines all annotations used.

Gapped-*k*-mer SVM (gkm-SVM) model of HepG2 activity

We collected training data of individual ChIP-seq binding factors described by M Ghandi & D Lee et al. (Ghandi et al. 2014) for HepG2 (5k specific ChIP-seq peak regions and the same number of random controls, http://www.beerlab.org/gkmsvm/) and removed duplicate sequences, obtaining 225k peak sequences as well as matched random controls. Attempting to train a classification model with the gkm-SVM software based on all peak sequences exceeded reasonable memory

requirements (>1TB). Therefore, we iteratively reduced the number of training examples and ended up sampling each 50k peak and 50k control sequences for a combined HepG2 sequence model. Based on a test data set (2k sampled from the unused training data set), the obtained model has a specificity of 71.8%, a sensitivity of 88.8% and precision of 75.9% for separating ChIP-seq peak from the random control sequences.

Linear models integrating individual annotations

We used the R glmnet package to fit Lasso-penalized linear models to predict RNA/DNA ratios. We used 10-fold cross-validation (cv.glmnet) to determine the Lasso tuning parameter lambda resulting in the minimum squared error. The Lasso forces small coefficients to zero, and thereby performs regression and feature selection simultaneously. Otherwise missing annotation values were mostly in count features (70.1%) or absence of the conserved block annotation "GerpRS" (27.5%) and thus all these values were imputed to zero. All annotation features were scaled and centered. Categorical features with K levels were included as K-1 binary columns. We excluded ZNF274 and EZH2 annotations from the model as none of the inserts overlapped with these ChIP-seq tracks. To report unbiased correlation values and scatter plots between the true and predicted RNA/DNA ratios, we randomly split up our data into 10 folds, trained models using 9 folds and the above identified tuning parameter and then extracted the fitted values after applying the model to the remaining fold.

Sequence-based LS-GKM models

LS-GKM (Lee 2016) is a faster and lower memory profile version of gkm-SVM. Its default settings are different from gkm-SVM (e.g. using 11 bases with 7 informative positions rather than 10 bases with 6 informative positions). We applied LS-GKM using parameters corresponding to gkm-SVM (-l 10 -k 6 -d 3 -t 2 -T 4 -e 0.01) as well as default parameters (-T 4 -e 0.01) on the HepG2 training data described for gkm-SVM above (225,327 positive/negative sequences each, 10,000 kept set aside for validation). We also compared performance for using the negative sequences as described for gkm-SVM (Ghandi et al. 2014) versus obtaining negative sequences by permutation of the real sequences maintaining dinucleotide content (Jiang et al. 2008). We found that best results were obtained for LS-GKM defaults in combination with selected negative sequences rather than permuted sequences (Table S6). However, permuted sequences as negative set produced a higher true positive rate and substantially simplify computation. We therefore used permuted sequence sets and ran LS-GKM with default parameters for all models. We extracted genomic sequences (GRCh37) below the 64 ChIP-seq peak sets by concatenating multiple call-sets for the same factor and merging overlapping peak regions using BEDTools (Quinlan and Hall 2010). We extracted up to 1kb of sequence for each peak, or centered 1kb fragments on the peak for larger peak calls. We chose the model convergence parameter *e* based on the number of positive training sequences

(mean 16600, min. 186, max. 63948) multiplied with 1E-07; investing more training iterations for smaller training data sets.

We then used Lasso regression (as described above) to create combined models and we also trained LS-GKM models from pooled peak data sets (Table S2). For this purpose, we pooled sequences using the peak data sets underlying the top 2, 3, 5, 10 and top 15 and all sequence models selected using Lasso regression.

Individual feature models

To explore whether certain annotations are more strongly predictive for either the non-integrated (MT) or integrated (WT) expression measurements (despite the correlations among the annotations), we used the R glm (Generalized Linear Models) implementation to fit 430 linear single coefficient plus intercept models for predicting log2 RNA/DNA ratios for MT and WT experiments. We report the two-sided p-value for the t-statistic corresponding to the coefficient in the linear model, and used a significance criterion of 0.05 after Bonferroni correction (Figure S22, Table S7).

Data access

The sequencing data, designed oligo sequences and processed count and RNA/DNA ratio data including annotations have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE83894.

Acknowledgements

We thank members of the Ahituv, McManus and Shendure laboratories for helpful discussions and suggestions. Our work was supported by the National Human Genome Research Institute (NHGRI) grant number 1R01HG006768 (NA & JS), NHGRI and National Cancer Institute grant number 1R01CA197139 (GC, DW, NA, JS), National Institute of Mental Health grant number 1R01MH109907 (NA) and the Uehara Memorial Foundation (FI). JS is an investigator of the Howard Hughes Medical Institute.

Author's contributions

FI, MK, BM, MTM, NA and JS designed experiments; FI and BM performed all wet lab experiments; MK, JS, NA, DMW and GMC outlined data analysis; MK performed data analysis; FI, MK, NA, JS interpreted the experimental results; FI, MK, BM, NA and JS wrote the manuscript.

References

Alcorn JA, Feitelberg SP, Brenner DA. 1990. Transient induction of c-jun during hepatic regeneration. *Hepatology* **11**(6): 909-915.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**(7493): 455-461. doi: 410.1038/nature12787.

Archer TK, Lefebvre P, Wolford RG, Hager GL. 1992. Transcription factor loading on the MMTV promoter: a bimodal mechanism for promoter activation. *Science* **255**(5051): 1573-1576.

Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, Lau NC, Stark A. 2014. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat Genet* **46**(7): 685-692. doi: 610.1038/ng.3009. Epub 2014 Jun 1038.

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**(6123): 1074-1077. doi: 1010.1126/science.1232542. Epub 1232013 Jan 1232517.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**(2 Pt 1): 299-308.

Bassuk AG, Leiden JM. 1995. A direct physical association between ETS and AP-1 transcription factors in normal human T cells. *Immunity* **3**(2): 223-237.

Butler SL, Hansen MS, Bushman FD. 2001. A quantitative assay for HIV DNA integration in vivo. *Nat Med* **7**(5): 631-634.

Consortium EP Dunham I Kundaje A Aldred SF Collins PJ Davis CA Doyle F Epstein CB Frietze S Harrow J et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.

Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**(3): 215-216. doi: 210.1038/nmeth.1906.

Erwin GD, Oksenberg N, Truty RM, Kostka D, Murphy KK, Ahituv N, Pollard KS, Capra JA. 2014. Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput Biol* **10**(6): e1003677. doi: 1003610.1001371/journal.pcbi.1003677. eCollection 1002014 Jun.

Euskirchen GM, Auerbach RK, Davidov E, Gianoulis TA, Zhong G, Rozowsky J, Bhardwaj N, Gerstein MB, Snyder M. 2011. Diverse roles and interactions of the SWI/SNF chromatin remodeling complex revealed using global approaches. *PLoS Genet* **7**(3): e1002008. doi: 1002010.1001371/journal.pgen.1002008. Epub 1002011 Mar 1002003.

Faure AJ, Schmidt D, Watt S, Schwalie PC, Wilson MD, Xu H, Ramsay RG, Odom DT, Flicek P. 2012. Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. *Genome Res* **22**(11): 2163-2175. doi: 2110.1101/gr.136507.136111. Epub 132012 Jul 136510.

Forrest AR Kawaji H Rehli M Baillie JK de Hoon MJ Haberle V Lassmann T Kulakovskiy IV Lizio M Itoh M et al. 2014. A promoter-level mammalian expression atlas. *Nature* **507**(7493): 462-470. doi: 410.1038/nature13182.

Gelderblom HC, Vatakis DN, Burke SA, Lawrie SD, Bristol GC, Levy DN. 2008. Viral complementation allows HIV-1 replication without integration. *Retrovirology* **5:60.**(doi): 10.1186/1742-4690-1185-1160.

Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* **10**(7): e1003711. doi: 1003710.1001371/journal.pcbi.1003711. eCollection 1002014 Jul.

Ghodsi M, Liu B, Pop M. 2011. DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics* **12:271.**(doi): 10.1186/1471-2105-1112-1271.

Hai T, Curran T. 1991. Cross-family dimerization of transcription factors Fos/Jun and ATF/CREB alters DNA binding specificity. *Proc Natl Acad Sci U S A* **88**(9): 3720-3724.

Hebbar PB, Archer TK. 2007. Chromatin-dependent cooperativity between site-specific transcription factors in vivo. *J Biol Chem* **282**(11): 8284-8291.

Hebbar PB, Archer TK. 2008. Altered histone H1 stoichiometry and an absence of nucleosome positioning on transfected DNA. *J Biol Chem* **283**(8): 4595-4601.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**(3): 311-318.

Hess J, Angel P, Schorpp-Kistner M. 2004. AP-1 subunits: quarrel and harmony among siblings. *J Cell Sci* **117**(Pt 25): 5965-5973.

Hilberg F, Aguzzi A, Howells N, Wagner EF. 1993. c-jun is essential for normal mouse development and hepatogenesis. *Nature* **365**(6442): 179-181.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**(5): 473-476. doi: 410.1038/nmeth.1937.

Jeong S, Stein A. 1994. Micrococcal nuclease digestion of nuclei reveals extended nucleosome ladders having anomalous DNA lengths for chromatin assembled on non-replicating plasmids in transfected cells. *Nucleic Acids Res* **22**(3): 370-375.

Jiang M, Anderson J, Gillespie J, Mayne M. 2008. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics* **9:192.**(doi): 10.1186/1471-2105-1189-1192.

Kadamb R, Mittal S, Bansal N, Batra H, Saluja D. 2013. Sin3: insight into its transcription regulatory functions. *Eur J Cell Biol* **92**(8-9): 237-246. doi: 210.1016/j.ejcb.2013.1009.1001. Epub 2013 Oct 1019.

Kamiya H, Miyamoto S, Goto H, Kanda GN, Kobayashi M, Matsuoka I, Harashima H. 2013. Enhanced transgene expression from chromatinized plasmid DNA in mouse liver. *Int J Pharm* **441**(1-2): 146-150. doi: 110.1016/j.ijpharm.2012.1012.1004. Epub 2012 Dec 1012.

Kantor B, Ma H, Webster-Cyriaque J, Monahan PE, Kafri T. 2009. Epigenetic activation of unintegrated HIV-1 genomes by gut-associated short chain fatty acids and its implications for HIV infection. *Proc Natl Acad Sci U S A* **106**(44): 18786-18791. doi: 18710.11073/pnas.0905859106. Epub 0905852009 Oct 0905859120.

Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M. 2013. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23**(5): 800-811. doi: 810.1101/gr.144899.144112. Epub 142013 Mar 144819.

Kinney JB, Murugan A, Callan CG, Jr., Cox EC. 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci U S A* **107**(20): 9158-9163. doi: 9110.1073/pnas.1004290107. Epub 1004292010 May 1004290103.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**(3): 310-315.

Kissler S, Stern P, Takahashi K, Hunter K, Peterson LB, Wicker LS. 2006. In vivo RNA interference demonstrates a role for Nramp1 in modifying susceptibility to type 1 diabetes. *Nat Genet* **38**(4): 479-483. Epub 2006 Mar 2019.

Klehr D, Maass K, Bode J. 1991. Scaffold-attached regions from the human interferon beta domain can be used to enhance the stable expression of genes under the control of various promoters. *Biochemistry* **30**(5): 1264-1270.

Klein JC, Lajoie MJ, Schwartz JJ, Strauch EM, Nelson J, Baker D, Shendure J. 2015. Multiplex pairwise assembly of array-derived DNA oligonucleotides. *Nucleic Acids Res* **8**.

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**(7539): 317-330. doi: 310.1038/nature14248.

Kwaks TH, Barnett P, Hemrika W, Siersma T, Sewalt RG, Satijn DP, Brons JF, van Blokland R, Kwakman P, Kruckeberg AL et al. 2003. Identification of anti-repressor elements that confer high and stable protein production in mammalian cells. *Nat Biotechnol* **21**(5): 553-558. Epub 2003 Apr 2007.

Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. 2012. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc Natl Acad Sci U S A* **109**(47): 19498-19503. doi: 19410.11073/pnas.1210678109. Epub 1210672012 Nov 1210678105.

Leavitt AD, Robles G, Alesandro N, Varmus HE. 1996. Human immunodeficiency virus type 1 integrase mutants retain in vitro integrase activity yet fail to integrate viral DNA efficiently during infection. *J Virol* **70**(2): 721-728.

Lee D. 2016. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **15**.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Lupien M, Eeckhoute J, Meyer CA, Wang Q, Zhang Y, Li W, Carroll JS, Liu XS, Brown M. 2008. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**(6): 958-970. doi: 910.1016/j.cell.2008.1001.1018.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Jr., Kinney JB et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30**(3): 271-277. doi: 210.1038/nbt.2137.

Moreau P, Hen R, Wasylyk B, Everett R, Gaub MP, Chambon P. 1981. The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res* **9**(22): 6047-6068.

Munir S, Thierry S, Subra F, Deprez E, Delelis O. 2013. Quantitative analysis of the time-course of viral DNA forms during the HIV-1 life cycle. *Retrovirology* **10:87.**(doi): 10.1186/1742-4690-1110-1187.

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**(5): 559-565. doi: 510.1038/nmeth.2885. Epub 2014 Mar 1023.

Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. 2012. Predicting cell-type-specific gene expression from regions of open chromatin. Genome Res 22(9):1711-22. doi: 10.1101/gr.135129.111.

Nightingale SJ, Hollis RP, Pepper KA, Petersen D, Yu XJ, Yang C, Bahner I, Kohn DB. 2006. Transient gene expression by nonintegrating lentiviral vectors. *Mol Ther* **13**(6): 1121-1132. Epub 2006 Mar 1123.

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* **30**(3): 265-270.

Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**(12): 1173-1175.

Qiu GH, Leung CH, Yun T, Xie X, Laban M, Hooi SC. 2011. Recognition and suppression of transfected plasmids by protein ZNF511-PRAP1, a potential molecular barrier to transgene expression. *Mol Ther* **19**(8): 1478-1486. doi: 1410.1038/mt.2011.1480. Epub 2011 May 1473.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6): 841-842.

Savic D, Roberts BS, Carleton JB, Partridge EC, White MA, Cohen BA, Cooper GM, Gertz J, Myers RM. 2015. Promoter-distal RNA polymerase II binding discriminates active from inactive CCAAT/ enhancer-binding protein beta binding sites. *Genome Res* **25**(12): 1791-1800. doi: 1710.1101/gr.191593.191115. Epub 192015 Oct 191520.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**(5981): 1036-1040. doi: 1010.1126/science.1186176. Epub 1182010 Apr 1186178.

Sharon E, van Dijk D, Kalma Y, Keren L, Manor O, Yakhini Z, Segal E. 2014. Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Res* **24**(10): 1698-1706. doi: 1610.1101/gr.168773.168113. Epub 162014 Jul 168716.

Shen SQ, Myers CA, Hughes AE, Byrne LC, Flannery JG, Corbo JC. 2016. Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res* **26**(2): 238-255. doi: 210.1101/gr.193789.193115. Epub 192015 Nov 193717.

Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**(4): 272-286. doi: 210.1038/nrg3682. Epub 2014 Mar 1011.

Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT et al. 2016. ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44**(D1): D726-732. doi: 710.1093/nar/gkv1160. Epub 2015 Nov 1092.

Smith CL, Hager GL. 1997. Transcriptional regulation of mammalian genes in vivo. A tale of two templates. *J Biol Chem* **272**(44): 27493-27496.

Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. 2013. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* **45**(9): 1021-1028.

Subramanian V, Fields PA, Boyer LA. 2015. H2A.Z: a molecular rheostat for transcriptional control. *1000Prime* **7:01.**(doi): 10.12703/P12707-12701. eCollection 12015.

Thierry S, Thierry E, Subra F, Deprez E, Leh H, Bury-Mone S, Delelis O. 2016. Opposite transcriptional regulation of integrated vs unintegrated HIV genomes by the NF-kappaB pathway. *Sci Rep* **6:25678.**(doi): 10.1038/srep25678.

Tsukada J, Yoshida Y, Kominato Y, Auron PE. 2011. The CCAAT/enhancer (C/EBP) family of basic-leucine zipper (bZIP) transcription factors is a multifaceted highly-regulated system for gene regulation. *Cytokine* **54**(1): 6-19. doi: 10.1016/j.cyto.2010.1012.1019. Epub 2011 Jan 1022.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, 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**(7231): 854-858.

Vockley CM, Guo C, Majoros WH, Nodzenski M, Scholtens DM, Hayes MG, Lowe WL, Jr., Reddy TE. 2015. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. *Genome Res* **25**(8): 1206-1214. doi: 1210.1101/gr.190090.190115. Epub 192015 Jun 190017.

Wang X, McManus M. 2009. Lentivirus production. *J Vis Exp* **(32).**(pii): 1499. doi: 1410.3791/1499.

Watt AJ, Garrison WD, Duncan SA. 2003. HNF4: a central regulator of hepatocyte differentiation and function. *Hepatology* **37**(6): 1249-1253.

White MA. 2015. Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. *Genomics* **106**(3): 165-170. doi: 110.1016/j.ygeno.2015.1006.1003. Epub 2015 Jun 1010.

Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. 2015. The ensembl regulatory build. *Genome Biol* **16:56.**(doi): 10.1186/s13059-13015-10621-13055.

Figure 1. Study design for LentiMPRA. (*A*) Schematic diagram of lentiMPRA. Candidate enhancers and barcode tags were synthesized in tandem as a microarray-derived oligonucleotide library, and cloned into the pLS-mP vector, followed by cloning of a minimal promoter (mP) and reporter (EGFP) between them. The resulting lentiMPRA library was packaged with either wildtype or mutant integrase, and infected into HepG2 cells. Both DNA and mRNA were extracted, and barcode tags were sequenced to test their enhancer activities in an episomal vs. genome integrating manner. (*B*) HepG2 cells infected with lentiviral reporter construct bearing no enhancer (pLS-mP), an SV40 enhancer (pLS-SV40-mP), and LTV1 (pLS-LTV1-mP), a known liver enhancer (Patwardhan et al. 2012), with or without antirepressors. The inclusion of antirepressors results in stronger and more consistent expression, but is still dependent on the presence of an enhancer. (*C*) FACS analyses quantifying the fluorescence intensity of pLS-mP, pLS-SV40-mP, and pLS-LTV1-mP with (blue lines) or without (red lines) antirepressors following infection into HepG2 with 1 copy of viral molecule per cell. Analysis of all GFP expressing cells (fluorescence >500 intensity units) shows a higher proportion of cells that strongly express GFP (>2,000 units) when antirepressors are included. Specifically, 54.8% vs. 45.1% for SV40, and 35.6% versus 29.0% for Ltv1, of cells with and without antirepressors, respectively. (*D*) Venn diagram showing the composition of the lentiMPRA library. 2,236 enhancer candidate sequences were chosen on the basis of having ENCODE HepG2 ChIP-seq peaks for EP300 and H3K27ac marks. The candidates overlapped with or without ChIP-seq peaks for FOXA1, FOXA2, or HNF4A. Half of the candidates overlapped with ChIP-seq peaks for RAD21, SMC3, and CHD2. The library included 102 positive and 102 negative controls.

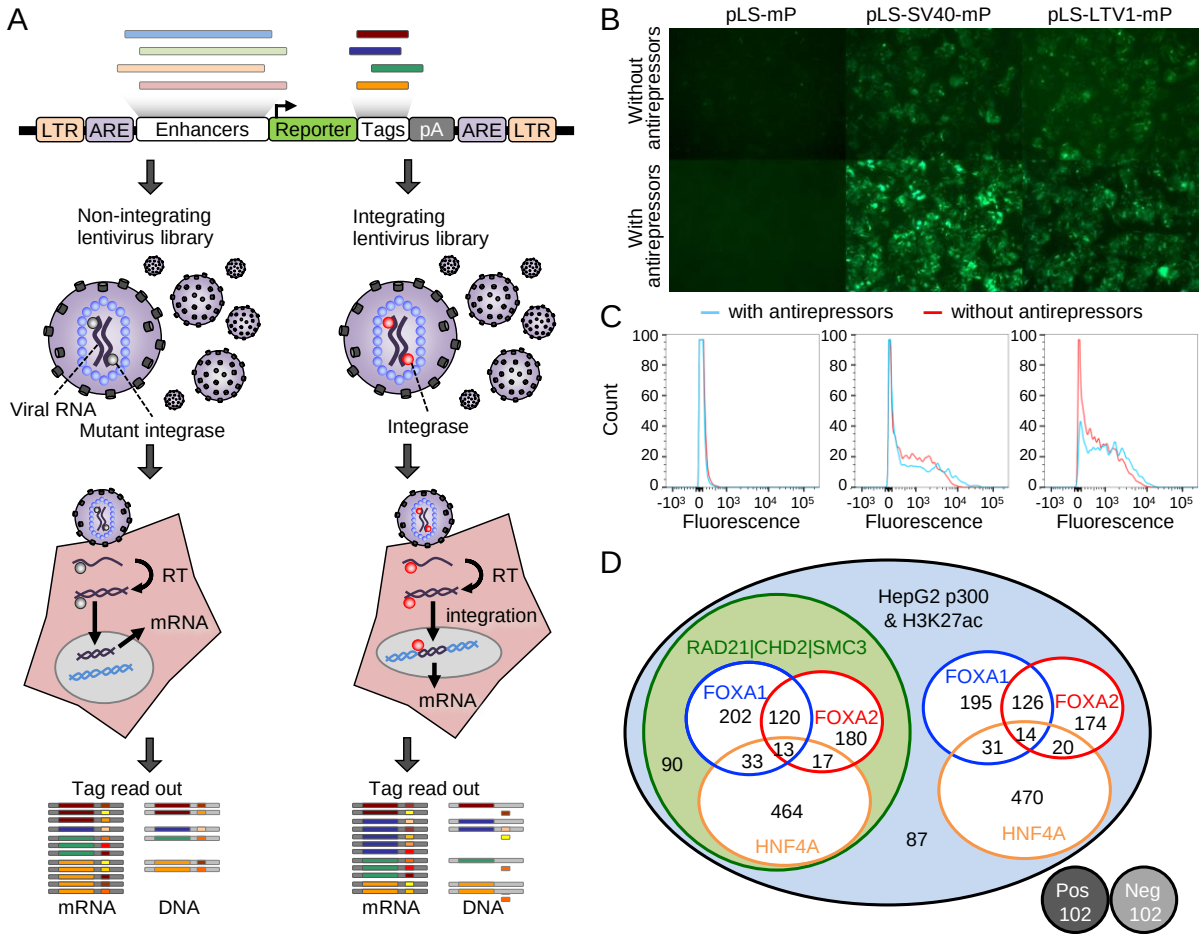

Figure 2. Pairwise correlation of per-insert RNA/DNA ratios between replicates, within and between MT vs. WT experiments. The lower left triangle shows pair-wise scatter plots. The diagonal provides replicate names and the respective histogram of the RNA/DNA ratios for that replicate. The upper triangle provides Pearson (p) and Spearman (s) correlation coefficients. MT vs. MT (green box) or WT vs. WT (blue box) comparisons are substantially more correlated than MT vs. WT (yellow boxes) comparisons, consistent with systematic differences between the episomal vs. integrated contexts for reporter assays that exceed technical noise. The two right-most columns and two bottom-most rows correspond to MT and WT after combining across the three replicates, with the combined MT vs. the combined WT comparison in the red box.

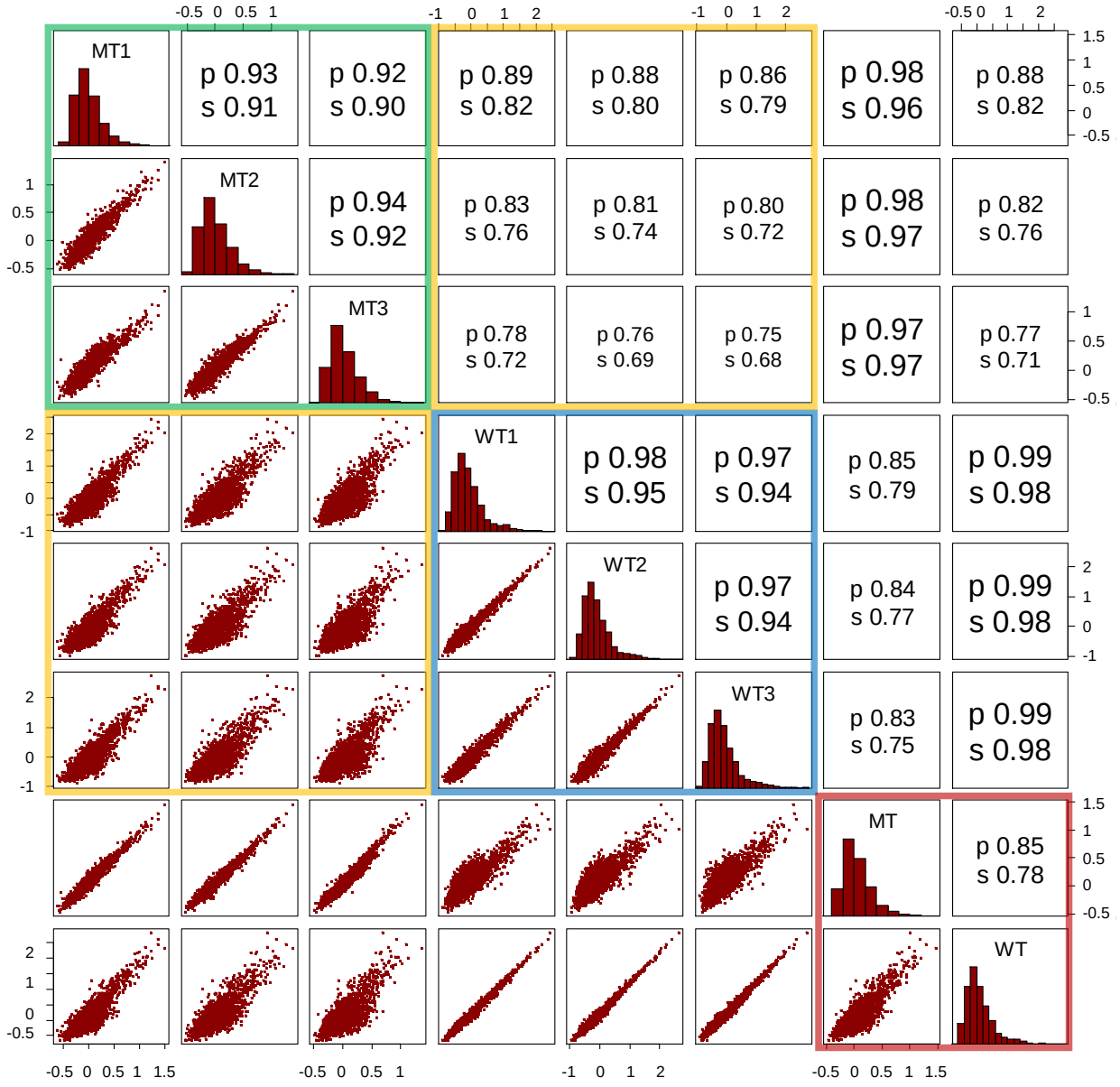

Figure 3. Comparisons between the non-integrating (MT) and integrating (WT) libraries (*A*) Scatter plot of combined MT vs. WT RNA/DNA ratios. MT ratios show a smaller dynamic range and thus seem compressed compared to WT results. Data points are colored by the type of insert sequence, including two types of controls: a total of four positive and negative controls (black) as well as the highest 100 and lowest 100 synthetic regulatory element sequences (SRES, red) identified by Smith *et al.* (Smith et al. 2013). The four classes of putative enhancer elements are: Regions of FOXA1, FOXA2 or HNF4A binding that overlap H3K27ac and EP300 calls as well as at least one of three factors RAD21, CHD2 or SMC3 (type 1); Regions like in 1 but with no RAD21, CHD2 or SMC3 overlapping (type 2); EP300 peak regions overlapping H3K27ac as well as at least one overlap with RAD21, CHD2 or SMC3, but without peaks in FOXA1, FOXA2 or HNF4A (type 3); Regions like in 3 but with no RAD21, CHD2 or SMC3 overlapping (type 4). As shown here and in Fig. S11, we do not observe major differences between the four design types, either with respect to activity or MT vs. WT. (*B*+*C*) Enhancer activity of 200 synthetic regulatory element sequences (SRES) in the MT (*B*) and WT experiments (*C*). Scatter plot of RNA/DNA ratios for the top 100 positive and top 100 negative synthetic regulatory element (SRE) sequences in HepG2 experiments by Smith *et al.* (Smith et al. 2013). Plots show the combined RNA/DNA ratios on the y-axis and measurements by Smith *et al.* on the x-axis. Intervals indicate the mean, minimum and maximum values observed for three replicates performed with each experiment.

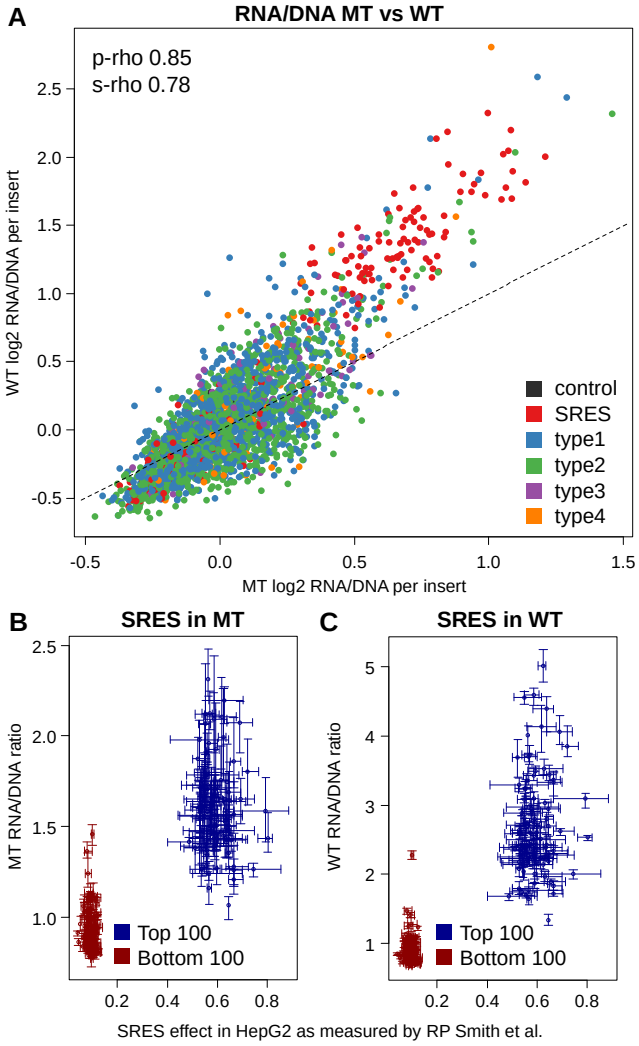

Figure 4. Squared Kendall's tau (T^2) values for available genome annotations for predicting the activity of candidate enhancer sequences in the non-integrating (MT) and integrating (WT) experiments. (*A*) WT RNA/DNA ratios correlate better with annotations than the respective MT values. The left panel highlights the top correlated annotations for WT and MT ratios. The right panel highlights annotations with the largest difference in T^2 values between the MT and WT experiments. (*B*) Same analysis for the 20% most active elements (Suppl. Table 1).

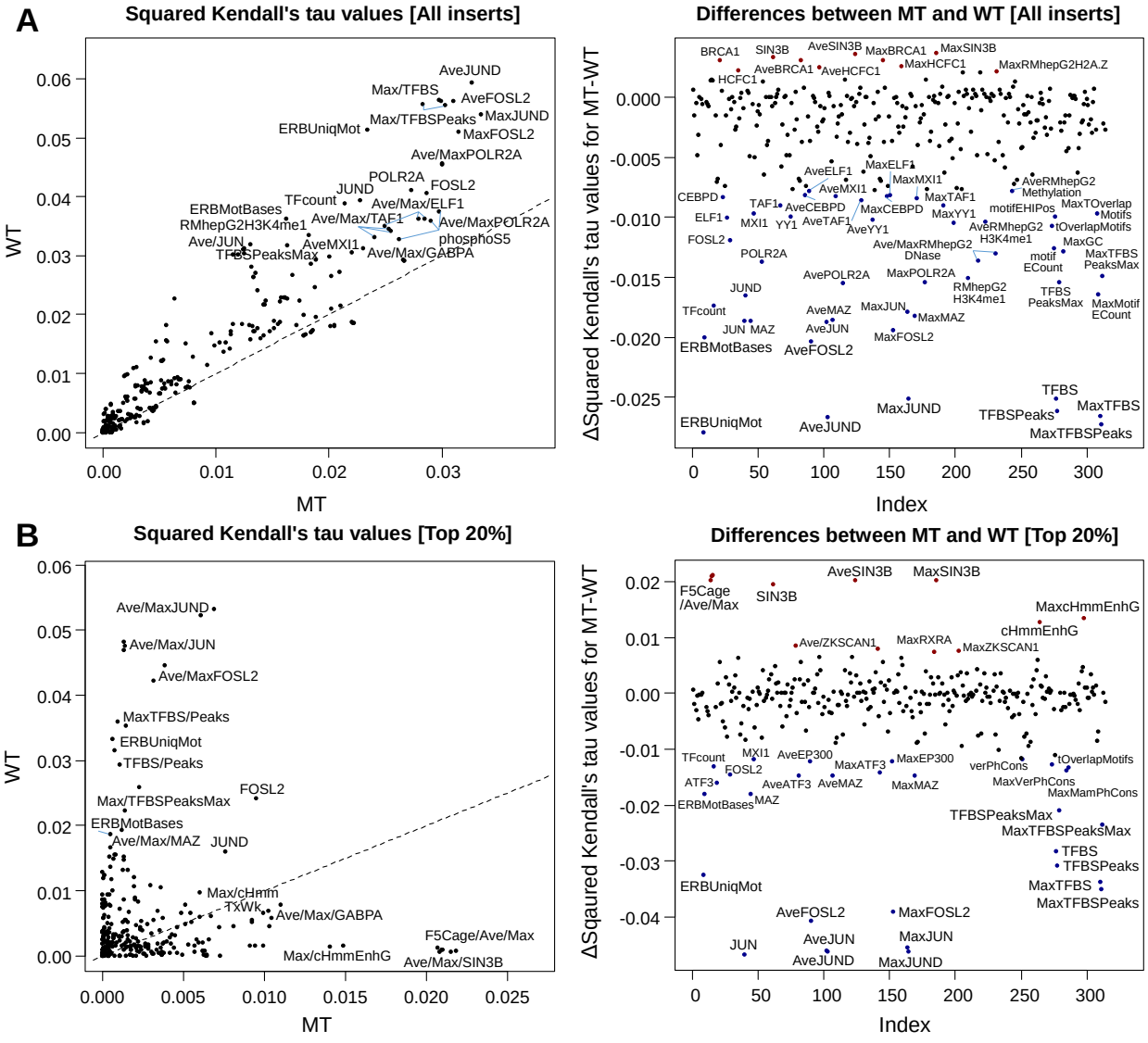

Figure 5. Prediction models. (*A+B*) Correlation of gkm-SVM scores obtained for a combined HepG2 model with RNA/DNA ratios obtained from the mutant (MT) and wild-type integrase (WT) experiments. Data points are colored by the type of insert sequence, including two types of controls, 200 synthetic regulatory element sequences (SRES, red) identified by Smith *et al.* (Smith et al. 2013) and four other control sequences (dark gray). The four classes of putative enhancer elements are: (type 1) regions of FOXA1, FOXA2 or HNF4A binding that overlap H3K27ac and EP300 calls as well as at least one of three factors RAD21, CHD2 or SMC3; (type 2) regions like in 1 but with RAD21, CHD2 or SMC3; (type 3) EP300 peak regions overlapping H3K27ac as well as at least one overlap with RAD21, CHD2 or SMC3, but without peaks in FOXA1, FOXA2 or HNF4A; (type 4) regions like in 3 but with no remodeling factor overlapping. Correlations are partially driven by the SRES; when excluding all controls, Spearman's R^2 values drop from 0.080 to 0.039 and from 0.128 to 0.076 for MT and WT, respectively. (*C+D*) Scatter plots of measured RNA/DNA ratios with predicted activity from linear Lasso models using annotations (numerical and categorical) as well as sequence-based (individual LS-GKM scores) information. Correlation coefficients are 0.45 Pearson / 0.40 Spearman for the non-integrated experiment (MT) and 0.60 Pearson / 0.57 Spearman for the integrated constructs (WT). The models selected 110 (MT) and 133 (WT) out of a total of 384 annotation features. Based on Pearson R^2 values, these combined models explain 20.6% (MT) and 36.2% (WT) of the variance observed in these experiments.