



Published in final edited form as:

Genomics. 2015 September ; 106(3): 165–170. doi:10.1016/j.ygeno.2015.06.003.

Understanding how *cis*-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences

Michael A. White

Center for Genome Sciences and Systems Biology, Department of Genetics, Washington University in St. Louis School of Medicine, St. Louis, MO 63108

Abstract

Genome-scale methods have identified thousands of candidate *cis*-regulatory elements, but methods to directly assay the regulatory function of these elements on a comparably large scale have not been available. The inability to directly test and perturb the regulatory activity of large numbers of DNA sequences has hindered efforts to discover how *cis*-regulatory function is encoded in genomic sequence. Recently developed massively parallel reporter gene assays combine next generation sequencing with high-throughput oligonucleotide synthesis to offer the capacity to test and mutationally perturb thousands of specifically chosen or designed *cis*-regulatory sequences in a single experiment. These assays are the basis of recent studies that include large-scale functional validation of genomic CREs, exhaustive mutational analyses of individual regulatory sequences, and tests of large libraries of synthetic CREs. The results demonstrate how massively parallel reporter assays with libraries of designed sequences provide the statistical power required to address previously intractable questions about *cis*-regulatory function.

Keywords

cis-regulation; massively parallel reporter assays; genomics; enhancers

1. Introduction

A major function of the genome is to specify patterns of gene expression through the action of *cis*-regulatory elements (CREs), but how that function is encoded in genomic sequence is still poorly understood. While many aspects of gene regulation can currently be measured on a genomic scale, including transcription factor binding, transcript levels, chromatin state, and the three-dimensional association of distal elements with proximal promoters [1], the output of these methods are tens of thousands of candidate CREs, very few of which have been directly tested for regulatory function. The lack of technologies to validate and

Correspondence: mwhite@genetics.wustl.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

functionally characterize CREs on the same scale at which we discover them constitutes a major bottleneck. To relieve this bottleneck and advance our understanding of how *cis*-regulatory function is encoded in DNA sequence, we need methods to directly assay the regulatory activity of large numbers of genomic, mutant, and synthetic sequences.

Recently developed massively parallel reporter gene assays make it possible to directly test the function of up to tens of thousands of CREs in a single experiment. Several versions of massively parallel reporter assay technology exist that make use of different approaches to construct and measure libraries of reporter genes for different experimental purposes [2–14]. This review focuses on methods to assay specifically designed sequences that are synthesized on programmable microarrays. By combining high-throughput, custom oligonucleotide synthesis technology with next generation sequencing, these methods overcome two major technical challenges that limit the throughput of traditional reporter gene assays: 1) the challenge of physically obtaining a large number of specifically chosen or designed CREs to construct libraries of reporter genes, and 2) the challenge of reading out the activity of thousands of individual reporter genes in a single experiment. With the ability to synthesize and test thousands of designed sequences, it is possible to achieve sufficient statistical power to address three critical questions about *cis*-regulatory function that were previously difficult to study on a genomic scale: First, how well do indirect measures of *cis*-regulatory function such as transcription factor binding or chromatin state predict actual function? Second, what is the nucleotide-level functional architecture of CREs and how is it affected by genetic variants? Finally, what rules govern how different arrangements of transcription factor binding sites produce distinct levels of gene expression?

2. Massively Parallel Reporter Assay Technology

2.1 Methods to detect reporter gene expression

The key innovation that makes massively parallel reporter assays possible is the use of next generation sequencing to uniquely match CREs in a pooled library of reporter gene constructs with their level of reporter expression. Pooled reporter libraries are either transiently transfected or, rarely, genomically integrated [15,16] into a population of cells, after which reporter gene activity is measured using one of two approaches. In the first approach, each *cis*-regulatory element is identified by a short sequence barcode that is placed in the 3' untranslated region (UTR) of the reporter construct, where it is co-transcribed with the reporter gene (Fig. 1A). By performing RNA-seq on the transcribed barcodes, the activity of thousands of CREs can be measured simultaneously [2,6,7,9]. To control for variable DNA representation of different reporter constructs in the library, barcode DNA is also sequenced and used to normalize the RNA measurement (Fig. 1B). An important advantage of this approach is that replicate measurements can be obtained by tagging each *cis*-regulatory element in the library with multiple, independent barcodes. In the second approach, a population of cells carrying a reporter gene library is sorted by flow cytometry into different bins based on the expression of a fluorescent reporter, thereby avoiding the need for a co-transcribed barcode. DNA sequencing is performed to identify the reporter constructs in each bin. The activity of each *cis*-regulatory element is thus measured by its distribution among bins representing different expression levels [3,8]. These

discrete expression bins are coarse-grained measures of expression, in contrast to the continuous measurements obtained by RNA-seq. However, due to the large number of sequences assayed, such coarse-grained measurements are sufficient to train quantitative models of *cis*-regulation [3]. Because flow cytometry measurements are made on single cells, it is also possible to determine variability in reporter expression, making this method applicable to large-scale studies of gene expression noise [8,17].

2.2 Synthesis of libraries of designed CREs

With the challenge of measuring reporter activity on a large scale met, the challenge of obtaining large numbers of specifically designed *cis*-regulatory sequences is overcome using oligonucleotide synthesis on programmable microarrays [18]. Programmable microarrays are used to create libraries of designed CREs as barcoded oligonucleotides, which are then incorporated into reporter gene constructs. This approach was first demonstrated by Patwardhan and colleagues [2], who used programmable microarrays to synthesize barcoded oligonucleotide libraries containing every possible single-nucleotide substitution and deletion in three bacteriophage and three mammalian core promoters. The oligonucleotides were directly subjected to *in vitro* transcription, after which the barcode transcript levels were measured by RNA-seq. This method was subsequently improved and adapted for use in living cells by cloning the designed CREs into a plasmid backbone. For RNA-seq based reporter assays, this is followed by placement of a minimal promoter and a reporter gene between the *cis*-regulatory sequence and the barcode (Fig. 1C) [6,9]. For flow cytometry reporter assays, oligonucleotides are cloned into a plasmid that already contains a minimal promoter and fluorescent reporter gene [8].

In certain applications, barcoded custom oligonucleotides can be replaced by randomly generated oligonucleotides, such as random mutagenesis of individual promoters [3,7], or random ligation of transcription factor binding sites to create synthetic CREs [15]. To barcode these reporter constructs for RNA-seq based assays, barcodes are cloned into the plasmid backbone independently. The barcodes are then matched with their corresponding CREs by an additional sequencing step (Fig. 1D).

2.3 Biological context

Massively parallel reporter assays with libraries of designed oligonucleotides have been shown to be reproducible, quantitative, and able to recapitulate the results of traditional reporter gene assays in variety of cellular contexts, including human cell lines [6,19,20], whole retina [9,21], *in vivo* murine liver [7,22,23], and yeast [8,15,17]. There are however two primary drawbacks of these assays: 1) Currently only short (<200 bp) sequences can be synthesized on programmable microarrays, and 2) *cis*-regulatory activity is nearly always measured on plasmids. Thus, massively parallel reporter assays of designed sequences are presently limited to assessing the ectopic activity of relatively short CREs. Nevertheless, as the studies discussed below demonstrate, these assays still capture important functional properties of CREs, such as cell type specificity. These assays sacrifice some biological complexity in exchange for the substantially increased statistical power that comes from the capacity to assay thousands of wild-type, mutant, and synthetic sequences. Such statistical

power is necessary to detect subtle patterns of functional DNA sequence features that would be impossible to discover in smaller datasets.

Other versions of massively parallel reporter assays exist that do not use libraries of designed CREs, but these are beyond the scope of this review. Rather than precisely quantify the activity of a defined set of CREs, these assays are designed to efficiently screen entire genomes [10,24,25], regions within accessible chromatin [13], and large candidate *cis*-regulatory modules [5,11,12,14] for active regulatory DNA. Because these assays are conducted with naturally-occurring sequences only, they are not appropriate for experiments that require designed sequences such as specific mutations, particular negative controls, or synthetic CREs.

3. Large-scale validation and characterization of genomic CREs

An important application of massively parallel reporter assays is to address the question of how well indirect measures of *cis*-regulatory function such as transcription factor binding or chromatin state predict actual function. Major efforts in genomics are focused on the discovery and characterization of functional elements in the non-coding genome [26–28], yet identifying genuinely functional CREs, particularly enhancers, remains challenging. In contrast to protein-coding genes, *cis*-regulatory function is not specified by a single, broadly applicable genetic code [29–31], and the functional significance of specific genomic features such as active chromatin states or transcription factor binding motifs — present in millions of copies in mammalian genomes [21,32] — is unclear. Of the millions of transcription factor-bound sites and regions of open or active chromatin identified by genome-scale methods, few have been directly tested for *cis*-regulatory activity or subjected to mutational analysis to identify their salient sequence features.

Massively parallel reporter assays with designed sequences offer the ability to test the regulatory potential of thousands of candidate CREs in a single experiment. An important advantage of these assays is the ability to include large sets of designed negative controls and specific mutational perturbations in the reporter gene library. Thousands of randomized, negative control sequences are used to establish background levels of assay activity [20,21]. These negative control sequences are critical for establishing a measure of significant *cis*-regulatory function, given that most non-coding sequence exhibits some biochemical activity associated with gene regulation [1]. Cell type specificity of CREs is assessed by including sequences expected to be active in different cell types, and the reporter gene library can then be assayed in multiple cell types [19,20]. Mutational disruptions of transcription factor binding sites are used to determine how *cis*-regulatory activity depends on the presence of specific binding motifs [19,21]. These various controls make it possible to identify sequences that show specific *cis*-regulatory activity with high confidence.

Several recent studies have used massively parallel reporter assays with designed sequences to test thousands of candidate CREs, including transcription factor-bound sites and regions of cell type-specific, active chromatin. We conducted a large-scale functional validation of transcription-factor bound genomic regions by assaying 1,300 sequences centered on sites bound by the photoreceptor transcription factor Crx, as identified by ChIP-seq [21]. The

activity of these Crx-bound sequences was compared to several negative controls, including randomized DNA, mutant sequences in which all Crx binding motifs were abolished, and nearly 900 genomic regions that contained high-quality Crx motifs but which were *not* bound by Crx in ChIP-seq experiments. The results showed that the activity of a majority of Crx ChIP-seq peaks depended on the presence of intact Crx motifs, and that a large number of Crx-bound sequences drove high levels of reporter activity, while unbound sequences with high quality Crx motifs did not. This suggests that high quality but unbound transcription factor motifs, present in millions of copies in large genomes, are intrinsically non-functional and not merely occluded by inactive chromatin. Strikingly, the Crx-bound regions that were most active in the assay contained higher GC nucleotide content, compared to both unbound Crx motifs and Crx-bound regions that did not drive high reporter activity. This finding is consistent with the proposal that functional transcription factor binding sites are distinguished from non-functional sites by the content of the sequence flanking the motif [31].

Two studies used RNA-seq based massively parallel reporter assays to test putative enhancers predicted by the presence of an active chromatin state. Kheradpour and colleagues [19] tested over 2,000 sequences with conserved regulatory motifs and active chromatin states in either the HepG2 liver carcinoma cell line or the K562 erythroleukemia cell line, while Kwasnieski and colleagues [20] tested 2,000 genomic regions predicted by the ENCODE consortium to be either enhancers or repressed regions in the K562 erythroleukemia cell line [33–35]. Both studies demonstrated that predictions of cell type specificity based on chromatin state were accurate. Additionally, Kheradpour and colleagues found that disruption of repressor binding sites sometimes resulted in reporter activity in the wrong cell type, suggesting that the cell type specificity in many cases requires the presence of a repressor. They also found that chromatin accessibility, active chromatin marks and the presence of conserved transcription factor motifs were the strongest predictors of reporter activity. Kwasnieski and colleagues showed that the ENCODE predictions of enhancers and repressed genomic regions were broadly accurate: predicted enhancer sequences often drove more activity than randomized negative control sequences, while predicted repressive regions did not. The validation rate for enhancers was moderate however, with approximately 33% of predicted enhancers driving more reporter activity than the negative controls. Surprisingly, regions predicted to be only weak enhancers in K562 cells, due primarily to weaker chromatin ChIP-seq signal for H3K27 acetylation, were more active than regions predicted to be stronger enhancers. This suggests that we do not yet fully understand the functional significance of the different chromatin modifications used to make predictions of *cis*-regulatory function.

Taken together, these studies suggest that many candidate CREs identified in genomic assays do have genuine *cis*-regulatory potential, though a significant fraction do not. The results of these plasmid-based studies also indicate that much of the function of CREs is encoded by highly local sequence features that act independently of the broader genomic context, a hypothesis that can now be tested in the genome with DNA editing technologies.

4. Functional architecture of CREs at single-nucleotide resolution

A second major application of massively parallel reporter assays is exhaustive mutational analysis of regulatory elements to discover their functional architecture at single-nucleotide resolution. Single nucleotide polymorphisms in regulatory regions are important contributors to phenotypic variation [36,37] and they are likely to account for significant differences in human disease risk [38]. Yet we lack the ability to reliably predict the functional consequences of genetic variants in regulatory elements. Given the sometimes rapid evolution of regulatory sequence [39], it is not clear what types of mutations, such as mutations within transcription factor binding sites, are most likely to affect *cis*-regulatory function [40,41]. Massively parallel reporter assays offer the ability to test a greater portion of the mutational spectrum of regulatory elements and thereby address several critical questions about the functional architecture of CREs: What is the relative functional importance of conserved bases compared to non-conserved positions? How likely are mutations outside of transcription factor binding sites to affect expression?

Several exhaustive mutational analyses of viral, bacterial, and mammalian enhancers have been conducted using massively parallel reporter assays. Two early studies found that mutations in core promoter elements have the greatest impact on *cis*-regulatory activity. Using an *in vitro* transcription system, Patwardhan and colleagues [2] found that mutations with the largest effects occurred immediately upstream of the transcription start site of bacteriophage promoters, and within the TATA box and initiator elements of mammalian promoters. In an extensive mutational analysis of the *E. coli lac* promoter, Kinney and colleagues [3] used massively parallel reporter data to recover the known binding sites for RNA polymerase and the transcription factor CRP. They demonstrated the statistical power of this large dataset by using it to obtain an estimate of the interaction energy between RNA polymerase and CRP that matched the previously reported value.

Several later studies reported conflicting estimates of the fraction of positions within a regulatory element that have important effects in *cis*-regulatory activity. Melnikov and colleagues [6] tested all possible single nucleotide substitutions and a large number of multiple substitutions in short regions of two enhancers, a synthetic cAMP-regulated enhancer and the interferon- β enhancer, assayed in a human embryonic kidney cell line. They found that nearly 60 percent of single-nucleotide mutations affected the activity of the synthetic enhancer, while only 32 percent affected the function of the interferon- β enhancer, with the most consequential mutations falling within or near transcription factor binding sites. The results of these mutational analyses were successfully used to design new versions of the enhancers that responded more strongly to induction. Kwasnieski and colleagues [9], assayed ~1000 mutants of the *Rhodopsin* proximal promoter in whole murine retina and found that 86 percent of mutations affected *cis*-regulatory activity. This included mutations outside of known transcription factor binding sites. In some cases, these mutations restored a binding site that has been lost in the mouse genome, but which is present in other mammals. In contrast with these findings, two studies of three intergenic [7] and four exonic [23] liver enhancers, assayed in mouse liver *in vivo*, found that fewer than 25 percent of mutations affected reporter activity. Mutations in transcription factor binding sites and evolutionarily conserved regions were most likely to affect enhancer activity.

The differences between these studies in the reported fraction of mutations with *cis*-regulatory are likely due to several factors, including the biological system used, the specific CREs tested, and differences in reporter assay technology among the different studies. Notably, Birnbaum and colleagues [23] demonstrated that the mutational profile of a *cis*-regulatory element can sensitively depend on biological context: they found that mutations in the exonic liver enhancers had significantly different effects when assayed in HeLa cells rather than the liver.

Overall, the results of these studies suggest that mutations in conserved positions and transcription factor binding sites are most likely to affect *cis*-regulatory activity. However, it is also important to consider turnover of transcription factor binding sites, as shown in the *Rhodopsin* promoter, where a recently lost site was recreated with single nucleotide mutations. These studies also demonstrate the power of massively parallel assays to test specific hypotheses about the sequence basis of *cis*-regulatory function by using the data to design new enhancers with specific properties.

5. Deciphering combinatorial *cis*-regulatory logic with synthetic promoters

A third major application of massively parallel reporter assays with designed sequences is to discover the rules that govern how different combinations of transcription factor binding sites produce distinct transcriptional outcomes, using large libraries of synthetic CREs. A key rationale for studying synthetic CREs is that a much larger and more unbiased fraction of “sequence space” can be explored by testing combinations of transcription factor binding sites that do not occur in the genome [42,43].

With massively parallel reporter assay technology, libraries of synthetic regulatory elements can be made much larger than previous collections of synthetic regulatory elements, and thus provide greater power to discover *cis*-regulatory rules. Sharon and colleagues [8] assayed over 6,000 different synthetic yeast promoters, in which key variables such as spacing, binding site affinity, and binding site identity were varied within different promoter contexts. From the data, they were able to infer several *cis*-regulatory rules, including the incremental effect of adding additional binding sites for the same transcription factor, and a ~10 bp periodic effect between binding site location and promoter activity. By using fluorescence-activated cell sorting to measure reporter activity, they also measured expression variance for each reporter construct, enabling a follow-up analysis examining the effect of promoter sequence on gene expression noise [17]. Mogno and colleagues [15] constructed a library of ~2,500 synthetic CREs, consisting of combinations of binding sites with different predicted affinities for four yeast transcription factors. By fitting a quantitative model to the data, cooperative effects between different transcription factor binding sites were identified. Their results also suggested that some binding sites act as a transcriptional amplifier, amplifying the effects of both activators and repressors. Finally, Smith and colleagues [22] assayed 5,000 synthetic mammalian CREs containing combinations of twelve liver-specific transcription factor binding sites. They found that heterotypic synthetic CREs, consisting of multiple binding sites for different transcription factors, drove much stronger expression than homotypic elements, comprised of multiple binding sites for the same factor. Similar findings have been reported in other systems [30], which suggests that

the stronger activity of heterotypic clusters of binding sites relative to homotypic ones may be a general *cis*-regulatory rule.

6. Future developments

Two major limitations of current massively parallel reporter assay technology are that CREs are assayed on transiently transfected plasmids rather than being genomically integrated (except in yeast [15]), and that the CREs included in reporter libraries are very short, typically less than 200 bp. However, new approaches are being developed to overcome these limitations. With lentiviral technology it is now feasible to integrate large barcoded libraries of reporter constructs into the genome [16], although integrations occur randomly and reporter activity is thus subject to uncontrolled genomic position effects. This issue could be addressed by averaging the activity of CREs over a large number of integrations or by targeting integrations to specific sites in the genome. While massively parallel assay technology itself is not limited to short sequences [4,5,7,10–12,15,25], assays conducted with libraries of specifically designed *cis*-regulatory sequences are currently subject to the limits of oligonucleotide synthesis technology. As the length of synthesized oligonucleotides increases [44], it will be possible to include designed versions of full-size enhancers in massively parallel reporter libraries. Along with these efforts to improve the current technology, the massively parallel reporter method is being adapted to study other sequence-associated aspects of transcriptional regulation, including the role of different 3' UTR sequences [45,46], RNA splicing [47], and the sequence specificity of DNA methylation [48].

Despite the current technical limitations of massively parallel reporter assays, the statistical power achieved by testing thousands of specifically chosen or designed *cis*-regulatory sequences with these assays is essential to address fundamental questions about how *cis*-regulatory function is encoded in genomic sequence, and about the mechanisms by which genetic variation in regulatory sequence contributes to phenotypic diversity and human disease.

Acknowledgments

The author thanks Barak Cohen, Hemangi Chaudhari, Chris Fiore, Dana King, Jamie Kwaneski, Brett Maricque, and Devjane Swain Lenz for helpful discussions.

References

1. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74.10.1038/nature11247 [PubMed: 22955616]
2. Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol*. 2009; 27:1173–1175.10.1038/nbt.1589 [PubMed: 19915551]
3. Kinney JB, Murugan A, Callan CG, Cox EC. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci USA*. 2010; 107:9158–9163.10.1073/pnas.1004290107 [PubMed: 20439748]
4. Nam J, Dong P, Tarpine R, Istrail S, Davidson EH. Functional *cis*-regulatory genomics for systems biology. *Proc Natl Acad Sci USA*. 2010; 107:3930–3935.10.1073/pnas.1000147107 [PubMed: 20142491]

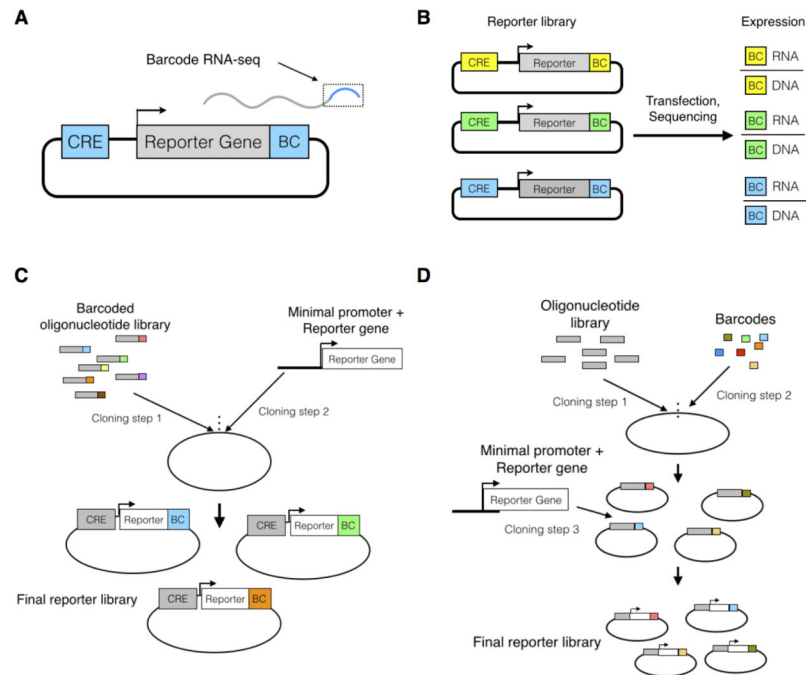
5. Nam J, Davidson EH. Barcoded DNA-Tag Reporters for Multiplex Cis-Regulatory Analysis. *PLoS ONE*. 2012; 7:e35934.10.1371/journal.pone.0035934 [PubMed: 22563420]
6. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol*. 2012; 30:271–277.10.1038/nbt.2137 [PubMed: 22371084]
7. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, et al. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol*. 2012; 30:265–270.10.1038/nbt.2136 [PubMed: 22371081]
8. Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol*. 2012; 30:521–530.10.1038/nbt.2205 [PubMed: 22609971]
9. Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc Natl Acad Sci USA*. 2012; 109:19498–19503.10.1073/pnas.1210678109 [PubMed: 23129659]
10. Arnold CD, Gerlach D, Stelzer C, Bory LM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*. 2013; 339:1074–1077.10.1126/science.1232542 [PubMed: 23328393]
11. Gisselbrecht SS, Barrera LA, Porsch M, Aboukhalil A, Estep PW, Vedenko A, et al. Highly parallel assays of tissue-specific enhancers in whole *Drosophila* embryos. *Nat Meth*. 2013; 10:774–780.10.1038/nmeth.2558
12. Dickel DE, Zhu Y, Nord AS, Wylie JN, Akiyama JA, Afzal V, et al. Function-based identification of mammalian enhancers using site-specific integration. *Nat Meth*. 2014; 11:566–571.10.1038/nmeth.2886
13. Murtha M, Tokcaer-Keskin Z, Tang Z, Strino F, Chen X, Wang Y, et al. FIREWACH: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nat Meth*. 2014; 11:559–565.10.1038/nmeth.2885
14. Vanhille L, Griffon A, Maqbool MA, Zacarias-Cabeza J, Dao LTM, Fernandez N, et al. High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nat Comms*. 2015; 6:6905.10.1038/ncomms7905
15. Mogno I, Kwasnieski JC, Cohen BA. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res*. 2013; 23:1908–1915.10.1101/gr.157891.113 [PubMed: 23921661]
16. Akhtar W, de Jong J, Pindyurin AV, Pagie L, Meuleman W, de Ridder J, et al. Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell*. 2013; 154:914–927.10.1016/j.cell.2013.07.018 [PubMed: 23953119]
17. Sharon E, van Dijk D, Kalma Y, Keren L, Manor O, Yakhini Z, et al. Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Res*. 2014; 24:1698–1706.10.1101/gr.168773.113 [PubMed: 25030889]
18. Leproust EM, Peck BJ, Spirin K, McCuen HB, Moore B, Namsaraev E, et al. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res*. 2010; 38:2522–2540.10.1093/nar/gkq163 [PubMed: 20308161]
19. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res*. 2013; 23:800–811.10.1101/gr.144899.112 [PubMed: 23512712]
20. Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res*. 2014; 24:1595–1602.10.1101/gr.173518.114 [PubMed: 25035418]
21. White MA, Myers CA, Corbo JC, Cohen BA. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc Natl Acad Sci USA*. 2013.10.1073/pnas.1307449110
22. Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, et al. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet*. 2013; 45:1021–1028.10.1038/ng.2713 [PubMed: 23892608]

23. Birnbaum RY, Patwardhan RP, Kim MJ, Findlay GM, Martin B, Zhao J, et al. Systematic dissection of coding exons at single nucleotide resolution supports an additional role in cell-specific transcriptional regulation. *PLoS Genet.* 2014; 10:e1004592.10.1371/journal.pgen.1004592 [PubMed: 25340400]
24. Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, et al. Quantitative genome-wide enhancer activity maps for five *Drosophila* species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat Genet.* 2014; 46:685–692.10.1038/ng.3009 [PubMed: 24908250]
25. Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, et al. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature.* 2015; 518:556–559.10.1038/nature13994 [PubMed: 25517091]
26. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature.* 2012; 489:83–90.10.1038/nature11212 [PubMed: 22955618]
27. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. *Nature.* 2012; 489:75–82.10.1038/nature11232 [PubMed: 22955617]
28. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. *Nature.* 2014; 507:455–461.10.1038/nature12787 [PubMed: 24670763]
29. Yáñez-Cuna JO, Kvon EZ, Stark A. Deciphering the transcriptional cis-regulatory code. *Trends Genet.* 2013; 29:11–22.10.1016/j.tig.2012.09.007 [PubMed: 23102583]
30. Levo M, Segal E. In pursuit of design principles of regulatory sequences. *Nat Rev Genet.* 2014; 15:453–468.10.1038/nrg3684 [PubMed: 24913666]
31. Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordân R, Rohs R. Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci.* 2014; 39:381–399.10.1016/j.tibs.2014.07.002 [PubMed: 25129887]
32. Cao Y, Yao Z, Sarkar D, Lawrence M, Sanchez GJ, Parker MH, et al. Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. *Dev Cell.* 2010; 18:662–674.10.1016/j.devcel.2010.02.014 [PubMed: 20412780]
33. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, et al. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* 2013; 41:827–841.10.1093/nar/gks1284 [PubMed: 23221638]
34. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Meth.* 2012; 9:473–476.10.1038/nmeth.1937
35. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Meth.* 2012; 9:215–216.10.1038/nmeth.1906
36. Gerke J, Lorenz K, Cohen B. Genetic interactions between transcription factors cause natural variation in yeast. *Science.* 2009; 323:498–501.10.1126/science.1166426 [PubMed: 19164747]
37. Pai AA, Pritchard JK, Gilad Y. The Genetic and Mechanistic Basis for Variation in Gene Regulation. *PLoS Genet.* 2015; 11:e1004857.10.1371/journal.pgen.1004857 [PubMed: 25569255]
38. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science.* 2012; 337:1190–1195.10.1126/science.1222794 [PubMed: 22955828]
39. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, et al. Enhancer Evolution across 20 Mammalian Species. *Cell.* 2015; 160:554–566.10.1016/j.cell.2015.01.006 [PubMed: 25635462]
40. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. The functional consequences of variation in transcription factor binding. *PLoS Genet.* 2014; 10:e1004226.10.1371/journal.pgen.1004226 [PubMed: 24603674]
41. Yun Y, Adesanya TMA, Mitra RD. A systematic study of gene expression variation at single-nucleotide resolution reveals widespread regulatory roles for uAUGs. *Genome Res.* 2012; 22:1089–1097.10.1101/gr.117366.110 [PubMed: 22454232]

42. Gertz J, Siggia ED, Cohen BA. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*. 2009; 457:215–218.10.1038/nature07521 [PubMed: 19029883]
43. Gertz J, Cohen BA. Environment-specific combinatorial cis-regulation in synthetic promoters. *Mol Syst Biol*. 2009; 5:244.10.1038/msb.2009.1 [PubMed: 19225457]
44. Kosuri S, Church GM. Large-scale de novo DNA synthesis: technologies and applications. *Nat Meth*. 2014; 11:499–507.10.1038/nmeth.2918
45. Shalem O, Carey L, Zeevi D, Sharon E, Keren L, Weinberger A, et al. Measurements of the impact of 3' end sequences on gene expression reveal wide range and sequence dependent effects. *PLoS Comp Biol*. 2013; 9:e1002934.10.1371/journal.pcbi.1002934
46. Oikonomou P, Goodarzi H, Tavazoie S. Systematic identification of regulatory elements in conserved 3' UTRs of human transcripts. *Cell Reports*. 2014; 7:281–292.10.1016/j.celrep.2014.03.001 [PubMed: 24656821]
47. Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. Saturation editing of genomic regions by multiplex homology-directed repair. *Nature*. 2014; 513:120–123.10.1038/nature13695 [PubMed: 25141179]
48. Krebs AR, Dessus-Babus S, Burger L, Schübeler D, Ferguson-Smith AC. High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *eLife*. 2014; 3:e04094.10.7554/eLife.04094 [PubMed: 25259795]

Highlights

- New methods offer high-throughput testing of designed cis-regulatory elements.
- These assays are used to validate genomic predictions of cis-regulatory function.
- Designed reporter libraries offer improved statistical power for hypothesis testing.

**Fig. 1.**

Constructing massively parallel reporter libraries with co-transcribed barcodes. **A.** Co-transcribed sequence barcode (BC) in the 3' UTR of the reporter gene uniquely identifies the *cis*-regulatory element (CRE) driving reporter expression. In massively parallel reporter assays, reporter gene activity is detected by performing RNA-seq on the transcribed barcodes. The reporter gene itself is not measured in this version of the assay. **B.** Quantification of reporter activity. A pooled reporter library is transfected into a population of cells, followed by barcode sequencing of the RNA and DNA fractions. Reporter expression is determined by the number RNA sequence reads per barcode. To account for variable representation of different reporter constructs in the library, RNA barcode reads are normalized by DNA barcode reads. **C.** Construction of a reporter library using barcoded oligonucleotides synthesized on programmable microarrays. Barcoded oligonucleotides are cloned into a plasmid backbone (cloning step 1), following which a minimal promoter and reporter gene are cloned between the barcode (BC) and the *cis*-regulatory element (CRE). **D.** Construction of a reporter library using non-barcoded, randomly generated oligonucleotides. Oligonucleotides are cloned into a plasmid backbone (cloning step 1), followed by random cloning of barcodes into the plasmid constructs (cloning step 2). Barcode-oligonucleotide pairings are determined by sequencing, following which a minimal promoter and reporter gene are cloned between the *cis*-regulatory sequence and the barcode (cloning step 3).