HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2015 July 01.

Published in final edited form as:

Nat Methods. 2014 May ; 11(5): 491–492. doi:10.1038/nmeth.2933.

Discovering enhancers directly by activity

Ross C. Hardison

Center for Comparative Genomics and Bioinformatics, Huck Institutes for Life Sciences; Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802, phone: 814-863-0113

Ross C. Hardison: rch8@psu.edu

DNA sequences that regulate the timing, tissue-specificity, and level of gene expression are critical determinants of normal organismal development and differentiation[1]. Changes in their function likely underly biological processes ranging from disease susceptibility[2-4] to speciation[5]. However, our ignorance of a grammar by which to identify and decipher such regulatory sequences has impeded discovery and manipulation of these important genetic elements. Two papers in this issue of **Nature Methods** report novel methods for discovering regulatory sequences by directly monitoring their activity during the initial screening assay: **f**unctional **i**dentification of **r**egulatory **e**lements **w**ithin **a**ctive **ch**romatin (FIREWACh[6]) and **s**ite-specific **i**ntegration **f**luoresence activated cell sorting followed by **seq**uencing (SIF-seq[7]). These methods open new avenues for discovery of regulatory sequences.

The critical roles of regulatory sequences fostered decades of research into their structures and mechanisms of action. Most regulatory regions are modular, comprised of multiple binding sites for transcription factors (TFs). The TF binding site motifs direct binding by the TFs, but such short (frequently 6-8 bp) sequences do not provide sufficient discriminatory information to explain specific TF binding genome-wide. Regulatory regions control genes on the same chromosome (in *cis*), but they can be quite far away and need not target the nearest gene. The limited information in motifs and large distances over which enhancers act have impeded the formulation of a regulatory "grammar", *i.e.* rules for interpreting regulatory information in DNA sequences of complex organisms. Thus discovery of such *cis*-regulatory modules (CRMs) based solely on DNA sequence information is daunting[8]. Phylogenetic conservation or epigenetic features such as chromatin accessibility, histone modifications, and binding by TFs and co-activators can be used to improve CRM prediction. However, many of the DNA segments marked by these features are not independently active in enhancer assays[8]. These apparent "false positives" can be identified by experimental tests, and several groups have introduced massively parallel reporter assays (MPRAs) for high throughput testing[e.g.,9].

The new methods reported in this issue take a different approach to identifying CRMs, using their ability to boost gene expression as part of the assay to identify the candidates. Thus they are not biased toward evolutionarily constrained regions or to DNA sequences in chromatin with known activating epigenetic marks. FIREWACh builds upon the fact that

Conflict of interest statement: The author declares no conflict of interest.

almost all CRMs are found in accessible chromatin. Digestion of nuclei with frequently cutting restriction endonucleases releases at least a portion of the accessible chromatin, thereby enriching substantially for regulatory sequences (Figure). Fragments of the released DNA are then cloned into a lentiviral vector upstream of a green fluorescent protein (GFP) reporter gene, and a library of viruses carrying different DNA segments is transduced into an appropriate cell line[6]. In SIF-seq, DNA fragments are cloned upstream of a yellow fluorescent protein (YFP) gene and inserted into a vector for homologous recombination into the *Hprt* locus of mouse embryonic stem (ES) cells[7]. In both methods, cells carrying an active enhancer upstream of the fluorescent reporter gene are isolated by fluorescence activated cell sorting (FACS). The positive cells from each technique contain a single integrant carrying a candidate enhancer. Candidate enhancers can be located by sequencing the integrated DNA from the pool of positive cells and mapping the reads to the genome or target locus. The SIF-seq approach was effective not only in ES cells, but it also was used to discover enhancers active in cardiomyocytes or neural progenitor cells after *in vitro* differentiation of the ES cells.

Given that the candidate enhancers were discovered by an increased expression of a reporter gene, one expects these new methods to have a very high success rate in identifying active enhancers. This expectation was met by both approaches. Subsequent, independent enhancer assays validated the function of candidate enhancers in 78% of the tested FIREWACh positives and all of the tested SIF-seq positives. This is substantially higher than the results reported when using a MPRA approach[10] for enhancer discovery based on histone modifications and motif instances (25% to 41%) or the roughly 50% positive rate of predicted enhancers in moderate throughput assays[8]. Importantly, several DNA segments associated with epigenetic features indicative of enhancement (such as binding by EP300 or acetylation of histone H3K27) that were inactive in SIF-seq were confirmed to be inactive in an independent assay. While these inactive regions could reflect “opportunistic” binding by TFs and recruitment of chromatin modifiers that does not impact gene regulation, they could also be DNA segments that cooperate with other CRMs in gene regulation but are not independently active.

The new methods do have limitations, e.g. they were not designed to be comprehensive. SIF-seq was developed to interrogate in detail regulatory regions around specific loci, using as the input genomic clones in bacterial artificial chromosomes. FIREWACh was not targeted to specific loci, but coverage of all accessible chromatin would require lentiviral libraries larger than is practical. Each method was successful in achieving the goals for which it was designed. In contrast, comprehensive prediction of CRMs still relies on genome-wide maps of epigenetic features associated with regulation, but those candidate CRMs require functional assays. Perhaps future developments will reveal ways to use these activity-based assays in series with the epigenetic maps to accomplish more comprehensive coverage while maintaining high specificity.

Acknowledgments

The author is supported by NIH grants R01DK065806, R56DK065806, and U54HG006998.

References

1. Davidson EH, Erwin DH. Gene regulatory networks and the evolution of animal body plans. Science. 2006; 311:796–800. [PubMed: 16469913]
2. Forrester WC, et al. A deletion of the human β-globin locus activation region causes a major alteration in chromatin structure and replication across the entire β-globin locus. Genes & Devel. 1990; 4:1637–1649. [PubMed: 2249769]
3. The_ENCODE_Project_Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
4. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337:1190–1195. [PubMed: 22955828]
5. Chan YF, et al. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx1* enhancer. Science. 2010; 327:302–305. [PubMed: 20007865]
6. Murtha M, et al. FIREWACh: High-throughput Functional Detection of Transcriptional Regulatory Modules in Mammalian Cells. Nature Methods. 2014 in press.
7. Dickel DE, et al. Function-based Identification of Mammalian Enhancers Using Site-Specific Integration. Nature Methods. 2014 in press.
8. Hardison RC, Taylor J. Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet. 2012; 13:469–483. [PubMed: 22705667]
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 U S A. 2012; 109:19498–19503. [PubMed: 23129659]
10. Kheradpour P, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. 2013; 23:800–811. [PubMed: 23512712]

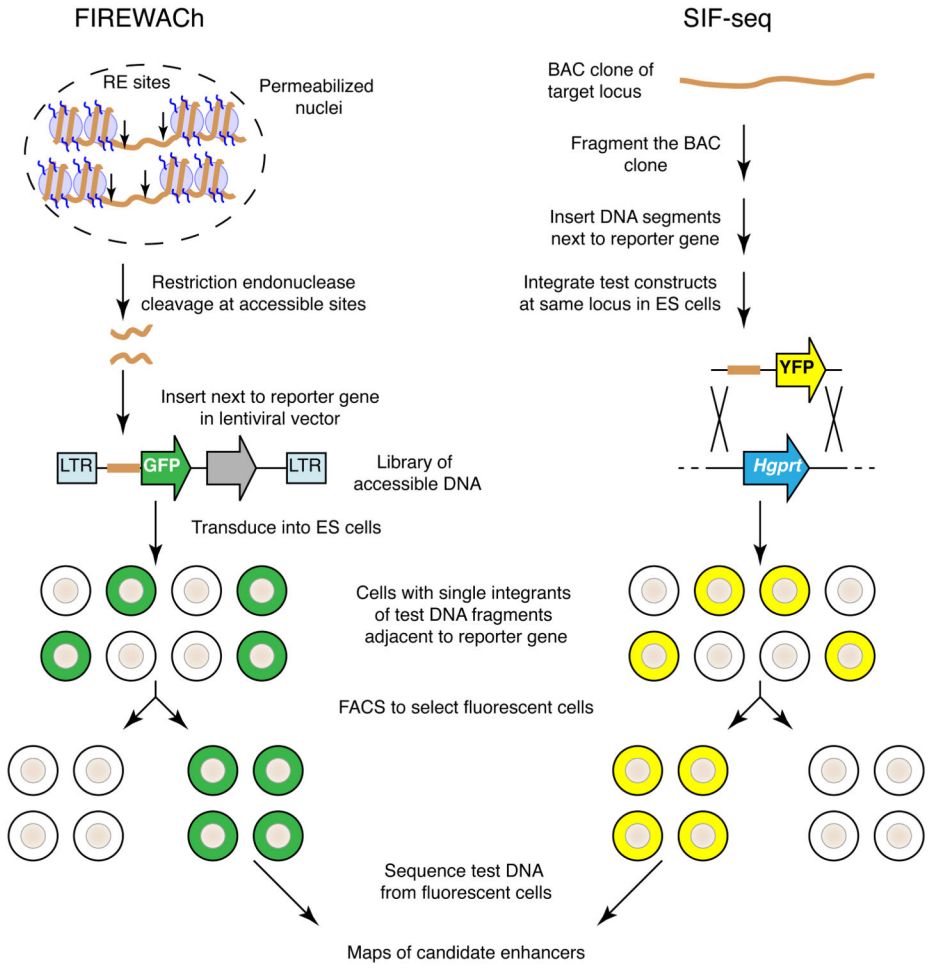

Figure.

Two methods for identifying enhancers directly by their activity. FIREWACh starts with DNA fragments cleaved from accessible chromatin, whereas SIF-seq begins with DNA segments from a locus containing a gene of interest. In both methods, the isolated DNA fragments are inserted adjacent to a reporter gene encoding a fluorescent protein and then introduced as single copy integrants into ES cells. Cells carrying constructs with active candidate enhancers are isolated by FACS. DNA isolated from the pools of positive cells is sequenced and the reads are mapped back to the genome or locus of interest, generating maps of candidate enhancers.