Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq

Cosmas D. Arnold, Daniel Gerlach, Christoph Stelzer, Łukasz M. Boryń, Martina Rath, Alexander Stark*

Genomic enhancers are important regulators of gene expression, but their identification is a challenge, and methods depend on indirect measures of activity. We developed a method termed STARR-seq to directly and quantitatively assess enhancer activity for millions of candidates from arbitrary sources of DNA, which enables screens across entire genomes. When applied to the *Drosophila* genome, STARR-seq identifies thousands of cell type–specific enhancers across a broad continuum of strengths, links differential gene expression to differences in enhancer activity, and creates a genome-wide quantitative enhancer map. This map reveals the highly complex regulation of transcription, with several independent enhancers for both developmental regulators and ubiquitously expressed genes. STARR-seq can be used to identify and quantify enhancer activity in other eukaryotes, including humans.

Enhancers (*1*) are DNA sequences that recruit transcription factors (TFs) to regulate the transcription of target genes in a cell type–specific manner, thereby governing development and physiology (*2*). Despite their importance, enhancer discovery has remained challenging, and rather few enhancers have been described and functionally characterized (*3*). Two types of recently developed methods enable genome-wide enhancer predictions via enhancer-associated chromatin features (*4*). Deep sequencing of deoxyribonuclease I–hypersensitive sites [DHS-seq (*5*)] or formaldehyde-assisted isolation of regulatory elements sequencing [FAIRE-seq (*6*)] allows the mapping of open chromatin, and chromatin immunoprecipitation followed by deep sequencing [ChIP-seq (*7*, *8*)] enables the detection of regulator (e.g., TF or cofactor) binding sites and enhancer-associated histone modifications [e.g., histone 3 (H3) Lys^4 monomethylation (H3K4me1) or H3 Lys^{27} acetylation (H3K27ac)]. These methods, however, do not provide a direct functional or quantitative readout of enhancer activity, which requires reporter assays that infer enhancer strength from the abundance of reporter transcripts [e.g., by means of luciferase or barcodes (*9*, *10*)]. Such assays, however, do not scale to the millions of tests required for genome-wide enhancer identification.

To comprehensively identify sequences that function as transcriptional enhancers in a direct, quantitative, and genome-wide manner, we developed STARR-seq (self-transcribing active regulatory region sequencing). Because enhancers can function independently of their relative positions (*1*), we placed candidate sequences downstream of a minimal promoter (Fig. 1A), such that active enhancers transcribe themselves, and each enhancer's strength is reflected by its abundance among cellular RNAs. This direct coupling of candidate sequences to enhancer activity allows the parallel assessment of millions of DNA fragments from arbitrary sources.

We cloned a genome-wide reporter library from randomly sheared genomic DNA of the *Drosophila melanogaster* reference strain (*11*) (fig. S1). This library contained at least 11.3 million independent candidate fragments with a median length of ~600 base pairs (bp) as revealed by paired-end sequencing (fig. S2A). It covered 96% of the non-repetitive genome at least 10-fold and was sufficiently complex to represent the entire 169 Mb *D. melanogaster* genome (fig. S2, B to E). We transfected the library (fig. S2F) into *Drosophila* S2 cells, isolated polyadenylated RNA, and selectively reverse-transcribed, polymerase chain reaction (PCR)–amplified, and paired-end sequenced the candidate fragments. After mapping the sequences to the genome, we quantified the enrichment over input for each position (Fig. 1B and fig. S3).

This yielded 5499 regions that were significantly enriched ["peaks"; $P \leq 0.001$, binomial test; empirical false discovery rate (FDR) = 1.8%; Fig. 1B and fig. S4A] at various genomic positions, including weak and strong (1953 with ≥threefold enrichment) enhancers over a wide dynamic range (fig. S5A). A biological replicate showed a Pearson correlation coefficient (r) of 0.92 for the peak summits, which indicated that STARR-seq is highly reproducible (figs. S6, A, B, and E, and S7, A, B, and G).

To validate STARR-seq, we tested 77 peaks chosen across a wide range of enrichments and 65 negative controls by luciferase assays. Our assays showed that 81% (62 out of 77) of the peaks but only 14% (9 out of 65) of the negative controls were at least two times the baseline ($P < 0.05$, t test; see detailed sensitivity and specificity analyses in figs. S8 to S11). STARR-seq enrichment and

Research Institute of Molecular Pathology (IMP), 1030 Vienna, Austria.

*To whom correspondence should be addressed. E-mail: stark@starklab.org

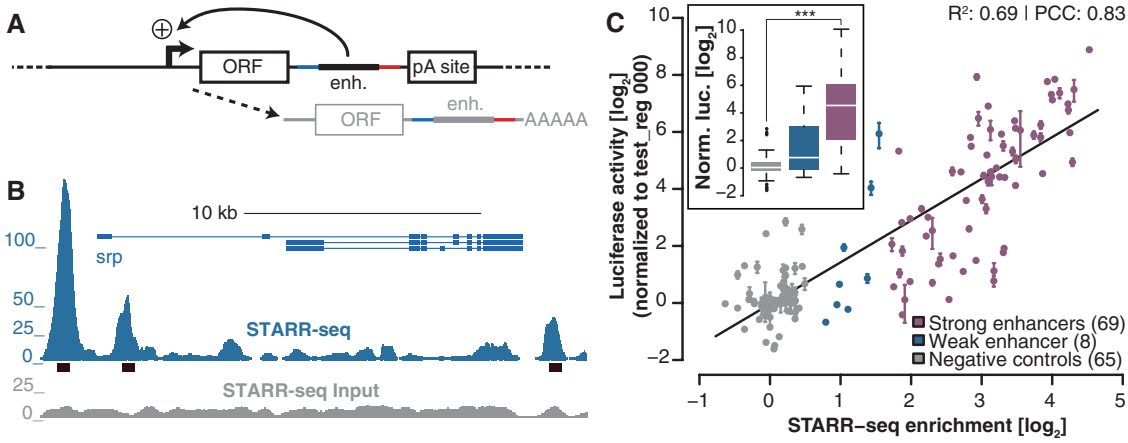

Fig. 1. STARR-seq genome-wide quantitative enhancer discovery. (**A**) STARR-seq reporter setup [enh., enhancer candidate; ORF, open-reading frame (here: GFP); pA site, polyadenylation site; +, transcriptional activation]. (**B**) STARR-seq (blue) and input (gray) fragment densities in the *srp* locus. Black boxes denote predicted enhancers ("peaks"). (**C**) STARR-seq and luciferase signals are linearly correlated: R^2, coefficient of determination and Pearson correlation coefficient (PCC or r). [Error bars indicate two independent biological replicates; (Inset) the same data as a boxplot; ***$P \leq 0.001$, Wilcoxon rank-sum test; n = 65, 8, and 69.]

luciferase activity were strongly linearly related over the entire range of enrichment values ($r = 0.83$; Fig. 1C), which established STARR-seq as a quantitative assay applicable to millions of fragments for the genome-wide identification of enhancers that substantially outperforms methods based on chromatin features (figs. S9 to S11). We did not find any evidence that the location of candidates inside the transcript would bias their assessment:

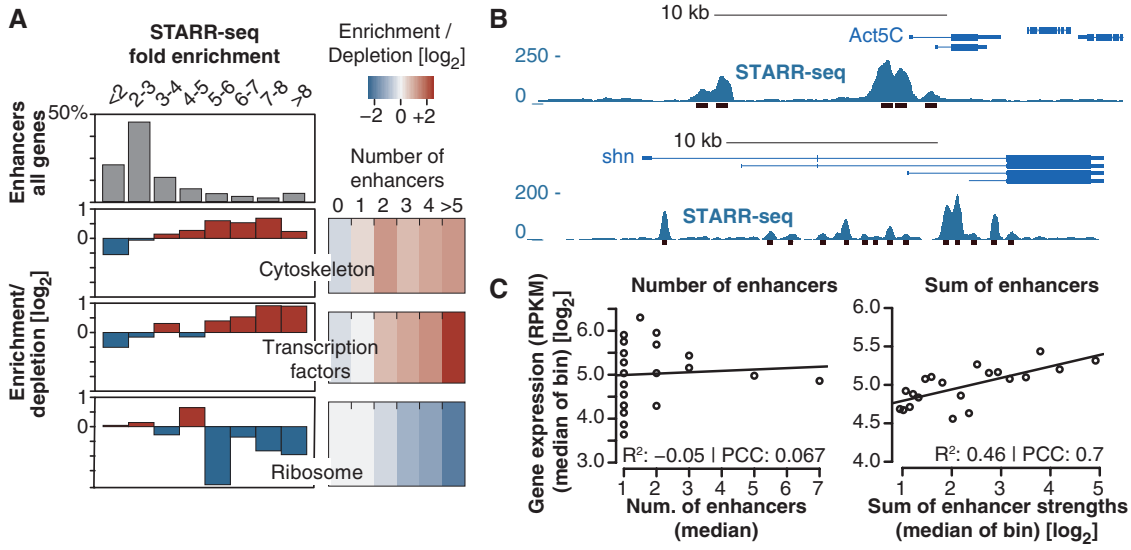

Fig. 2. Complexity of gene regulation. (**A**) Enrichment and depletion of different enhancer strengths (left) and enhancer number (right) in specific gene categories versus all genes. (**B**) STARR-seq enhancers and fragment densities in the *Act5C* and *shn* loci. (**C**) Gene expression versus the number and the combined strengths of enhancers per gene locus (bins of 100 genes).

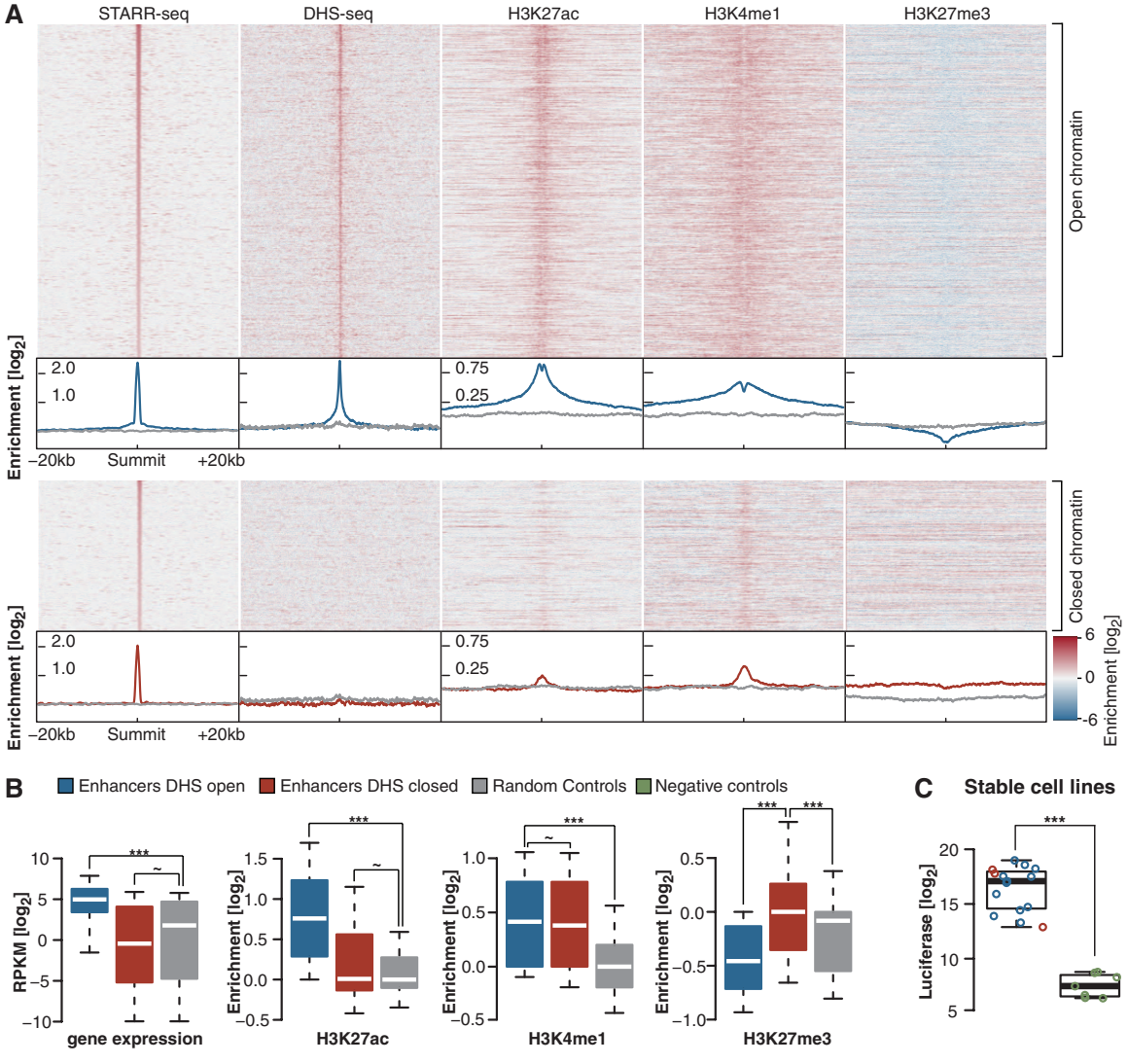

Fig. 3. Regulation of enhancer activity at the chromatin level. (**A**) Histone modifications (modEncode) at STARR-seq enhancers separated into open ($P \leq 0.05$, binomial test) and closed ($P > 0.05$) classes by DHS-seq (rows within each class sorted by STARR-seq P value). (**B**) Boxplots for STARR-seq enhancers in open (blue) and closed (red) chromatin indicating flanking gene expression (RPKM) and H3K27ac, H3K4me1, and H3K27me3 enrichments compared with a random control (gray) (***$P \leq 0.001$, ~$P > 0.05$, Wilcoxon rank-sum test; n = 1349, 604, and 500). (**C**) Luciferase assays of stable cell lines with genomically integrated reporter constructs for open (blue) and closed (red) S2 STARR-seq enhancers and negative controls (green; ***$P \leq 0.001$, Wilcoxon rank-sum test; n = 15, 7; see fig. S21 for details).

STARR-seq and luciferase assays agreed for sequences that, in their endogenous genomic contexts, occur upstream or within transcribed regions (fig. S12A). Moreover, more than 99% of all enhancers were supported similarly by fragments from both strands (figs. S12C and S13F), and even sequences that contained transcript-destabilizing elements were not substantially depleted during STARR-seq (fig. S12B).

The majority (55.6%) of identified enhancers were located within introns, especially in the first intron (37.2%), and in intergenic regions (22.6%) (fig. S13, A and B), consistent with predictions based on chromatin features (*12*, *13*) (figs. S13, A and B, and S14). To our surprise, 4.5% overlapped annotated transcription start sites (TSSs), which suggested that these sequences can both initiate transcription and enhance transcription from a remote TSS (fig. S13E). Indeed, as expected for bona fide enhancers, all tested (7 out of 7) TSS proximal regions activated luciferase expression irrespective of their relative orientation toward the luciferase TSS (fig. S13F).

The strongest enhancers were next to housekeeping genes such as enzymes (e.g., *Cct1*; enhancer rank no. 3) or constituents of the cytoskeleton (e.g., *Actin5C*; no. 31; Fig. 2, A and B) and also developmental regulators such as the TFs *luna* (no. 37), *shn* (no. 46), *pnt* (no. 49), or the fly fibroblast growth factor receptor *htl* (no. 45) (Fig. 2B). In fact, the strongest enhancer was located in the intron of the TF *zfh1* (fig. S15), and 18 of the top 100 and 364 of all strong enhancers were in TF gene loci. The only prominent class of genes with poorly ranking enhancers was the ribosomal protein genes (e.g., *RpS3*, rank 760) (Fig. 2A), presumably because their enhancers require a TCT motif–containing core promoter (*14*).

Even in a single cell type, many genes appeared to be regulated by several independently functioning enhancers (e.g., *shn*): 203 gene loci (see supplementary materials) contained five or more enhancers and 26 loci, 10 or more (Fig. 2A). Moreover, 434 genes had at least two and 56 had three or more enhancers within 2 kb of the TSS, including 14 TFs but also 30 housekeeping genes such as *Actin5C* (Fig. 2A). The sum of the enhancer strengths per gene correlated well with gene expression levels on average ($r = 0.7$; Fig. 2C), which directly linked gene expression to enhancer activity. Together, these findings suggest that transcription of a large number of genes, including ubiquitously expressed housekeeping genes, is complex and controlled by numerous enhancers per gene, even in a single cell type. These enhancers presumably function additively or redundantly to ensure robustness (*15*, *16*).

STARR-seq assesses the ability of DNA sequences to enhance transcription in a heterologous context, whereas the complementary DHS-seq and ChIP-seq determine enhancer-associated characteristics in the endogenous genomic context. We performed DHS-seq in S2 cells (figs. S16, A, C, and D, and S17, A and B) and found that the majority (69%) of strong STARR-seq enhancers were accessible (DHS enrichment $P \leq 0.05$, binomial test; Fig. 3A), and all weak enhancers showed above-random DHS enrichment on average (fig. S18), which suggests that they are active in their endogenous contexts. However, 604 (31%) strong STARR-seq enhancers were not accessible ("closed chromatin"; Fig. 3A) and occurred next to genes (e.g., the Hox TFs; fig. S19) expressed at significantly lower levels than genes next to open STARR-seq enhancers (25-fold difference; $P < 2.2 \times 10^{-16}$, Wilcoxon rank-sum test; Fig. 3B). Open and closed enhancers both function in luciferase assays and show the same linear correlation between STARR-seq and luciferase signals (fig. S20), which suggests that sequences with enhancer potential can be silenced in their endogenous contexts, presumably at the chromatin level. Indeed, in contrast to open STARR-seq enhancers, closed enhancers are not marked by H3K27ac, a histone modification associated with active enhancers, but lie in broad domains of repressive H3K27me3 (Fig. 3A and fig. S19), suggestive of Polycomb-mediated repression (*17*) or a poised enhancer state (*18*). It is noteworthy that both open and closed enhancers are marked to similar extents by H3K4me1, which labels enhancers irrespective of their activity (*19*, *20*) (Fig. 3A). The precise labeling of closed enhancers by H3K4me1 is particularly evident for Hox genes (fig. S19) and holds genome-wide (Fig. 3, A and B), which suggests that

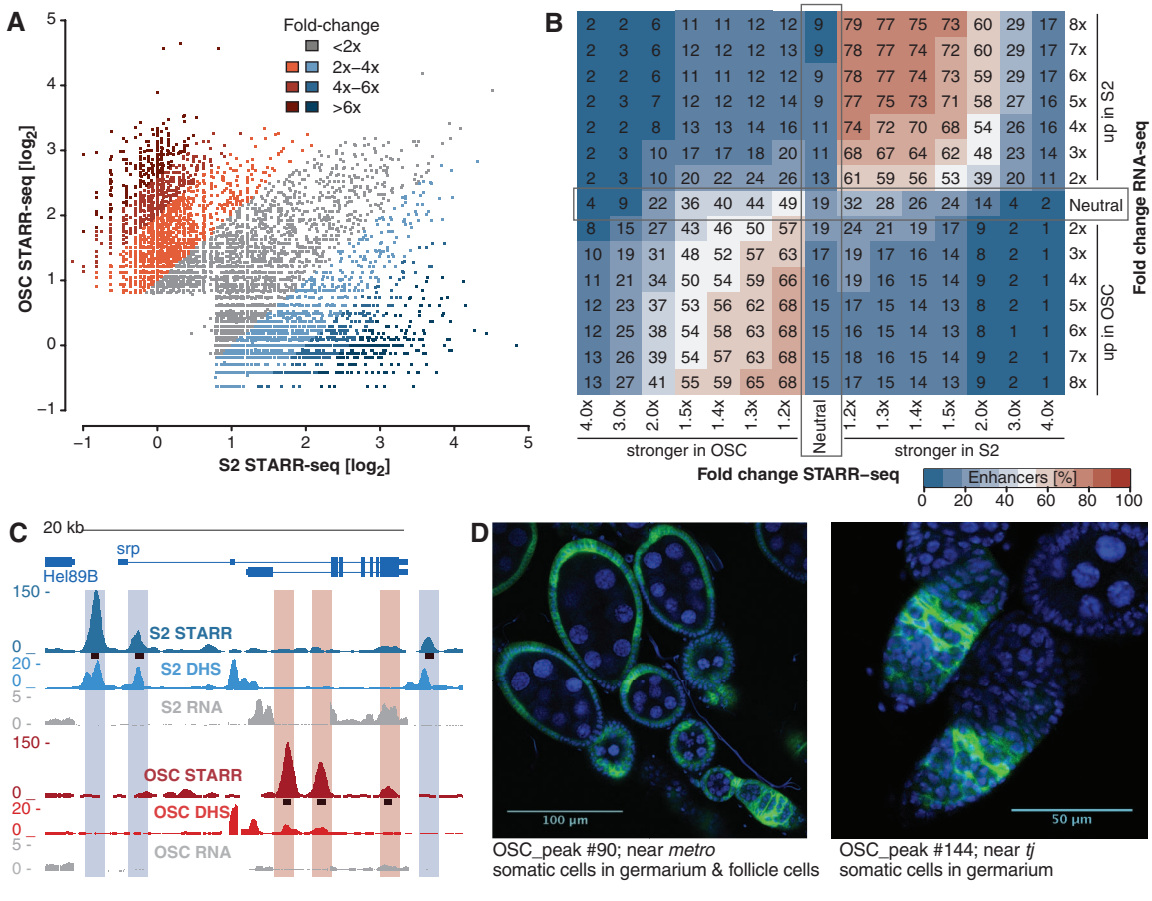

Fig. 4. Identification of cell type–specific enhancers. (**A**) STARR-seq enrichments in S2 cells versus OSCs. (**B**) Enhancer activities and gene expression levels change consistently between cell types (fraction of enhancers for each gene expression class; details in fig. S26). (**C**) STARR-, DHS-, and RNA-seq tracks for S2 cells and OSCs at the *srp* locus (black bars, STARR-seq enhancers). (**D**) In vivo validation of predicted OSC enhancers in ovaries of transgenic flies (OSC enhancers→Gal4; UAS→CD8-GFP; green, GFP; blue for DNA, 4′,6′-diamidino-2-phenylindole). See fig. S29 for details and all 13 tested candidates.

these sequences are recognized as functional enhancers, yet actively repressed.

To test the functionality of STARR-seq enhancers when integrated into the genome, we created 22 stable S2 cell lines each carrying stably integrated luciferase reporter constructs with an enhancer (15 lines) or a negative fragment (fig. S21). All enhancers, including three out of three closed enhancers, showed strong luciferase activity, whereas none of the negative controls did (Fig. 3C and fig. S21). For all, the luciferase activity was constant over a course of 4 weeks, measured 3, 5, and 7 weeks after integration (fig. S21). The activity of the three closed enhancers suggests that their endogenous inactive state might depend on the genomic context and/or on regulatory activities in S2 precursor cells. This shows that enhancers identified by ectopic assays, such as STARR-seq, can function when integrated into a chromosomal context, even if they are silenced endogenously.

We next applied STARR-seq to *Drosophila* adult ovarian somatic cells [OSCs (*21*)] and identified a comparable number of enhancers (4682; $P \leq 0.001$, binomial test; FDR = 0.2%) with similar characteristics (figs. S4B; S5B; S6, C to E; S12D; S13, C and D; and S22, A to C). Out of 8659 enhancers found in S2 cells or OSCs, 5404 (62.4%) changed at least twofold and 2138 (24.7%) at least fourfold between both cell types (Fig. 4A, and figs. S23 and S24, A and D), and luciferase assays confirmed these differences quantitatively ($r = 0.85$; fig. S24, B and C). Changes in enhancer strengths between the two cell types were reflected in the differential mRNA abundance (fig. S25) of the flanking genes (Fig. 4, B and C, and fig. S26): 74% of all enhancers near genes that are fourfold up-regulated in S2 cells appear stronger in S2 cells, whereas only 16% appear stronger in OSCs and vice versa (66% versus 19%). This establishes a direct link between quantitative differences in genome-wide enhancer strengths and differential gene expression. Up to 19% of cell type–specific enhancers were accessible in the cell-type in which they were not active (fig. S27). We also observed 514 genes for which individual enhancers changed more than twofold between cell types, whereas the sum of enhancer activities and the gene expression levels remained constant (<twofold change; fig. S28).

As OSCs have been derived from adult *Drosophila* ovaries and retained marker gene expression and other functional aspects of their in vivo counterparts (*21*), we assessed the activity of 13 OSC STARR-seq enhancers in ovaries of transgenic flies with site-specifically integrated transcriptional reporter constructs. In these flies, 85% (11 out of 13) of the enhancers but none of five control regions were active (Fig. 4D and fig. S29).

Here, we present STARR-seq, which complements ChIP-seq and DHS-seq as the third principal method to study transcriptional regulatory elements in entire genomes. It is unique in its ability to assess enhancer strengths quantitatively and to discover regulatory elements directly based on their ability to enhance transcription, even when silenced endogenously. Applied to two *Drosophila* cell types, it revealed thousands of cell type–specific enhancers with a broad range of strengths and provided the first genome-wide quantitative enhancer activity maps in any organism. STARR-seq is widely applicable to screening arbitrary sources of DNA in any cell type or tissue that allow the efficient introduction of reporter constructs (e.g., by plasmid transfection). This includes human HeLa cells, for which we confirm the quantitative nature of STARR-seq and its ability to identify enhancers that function in luciferase assays independent of their chromatin states and, thus, more reliably than previous methods (figs. S30 and S31). STARR-seq should be widely applied to many cell types across organisms to annotate cell type–specific gene regulatory elements and functionally assess noncoding mutations.

References and Notes

1. J. Banerji, S. Rusconi, W. Schaffner, *Cell* **27**, 299 (1981).
2. M. Levine, *Curr. Biol.* **20**, R754 (2010).
3. J. O. Yáñez-Cuna, E. Z. Kvon, A. Stark, *Trends Genet.* **29**, 11 (2013).
4. N. D. Heintzman *et al.*, *Nature* **459**, 108 (2009).
5. A. P. Boyle *et al.*, *Cell* **132**, 311 (2008).
6. K. J. Gaulton *et al.*, *Nat. Genet.* **42**, 255 (2010).
7. D. S. Johnson, A. Mortazavi, R. M. Myers, B. Wold, *Science* **316**, 1497 (2007).
8. G. Robertson *et al.*, *Nat. Methods* **4**, 651 (2007).
9. A. Melnikov *et al.*, *Nat. Biotechnol.* **30**, 271 (2012).
10. R. P. Patwardhan *et al.*, *Nat. Biotechnol.* **30**, 265 (2012).
11. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
12. P. V. Kharchenko *et al.*, *Nature* **471**, 480 (2011).
13. modENCODE Consortium *et al.*, *Science* **330**, 1787 (2010).
14. T. J. Parry *et al.*, *Genes Dev.* **24**, 2013 (2010).
15. M. W. Perry, A. N. Boettiger, M. Levine, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 13570 (2011).
16. N. Frankel *et al.*, *Nature* **466**, 490 (2010).
17. L. A. Boyer *et al.*, *Nature* **441**, 349 (2006).
18. A. Rada-Iglesias *et al.*, *Nature* **470**, 279 (2011).
19. M. P. Creyghton *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 21931 (2010).
20. S. Bonn *et al.*, *Nat. Genet.* **44**, 148 (2012).
21. K. Saito *et al.*, *Nature* **461**, 1296 (2009).

Acknowledgments: We thank H. K. Akyüz, M. Pagani, K. Schernhuber, D. Spies, H. Tagoh, M. Busslinger, S. Westermann, J. M. Peters, J. Zuber (IMP), J. Brennecke and group, U. Elling [Institute of Molecular Biotechnology (IMBA)], and the IMP/IMBA BioOptics and Graphics Services for help. Deep sequencing was performed at the CSF Next-Generation Sequencing Unit (http://csf.ac.at); Vienna Tile (VT) lines were obtained from the Dickson laboratory via the Vienna *Drosophila* RNAi Center (http://stockcenter.vdrc.at). C.D.A., L.M.B., and M.R. are supported by a European Research Council (ERC) Starting Grant (no. 242922) awarded to A.S. Basic research at the IMP is supported by Boehringer Ingelheim GmbH. C.D.A. and A.S. are authors on a patent application for STARR-seq (EP 12 004 520.8) filed by Boehringer Ingelheim International GmbH. All constructs are available from the authors subject to a Material Transfer Agreement. All deep sequencing data are available at www.starklab.org and GEO (series accession number GSE40739).

Supplementary Materials

www.sciencemag.org/cgi/content/full/science.1232542/DC1
Materials and Methods
Figs. S1 to S31
Tables S1 to S12
References (*22*–*48*)

9 November 2012; accepted 8 January 2013
Published online 17 January 2013;
10.1126/science.1232542