Functional *cis*-regulatory genomics for systems biology

Jongmin Nam[a], Ping Dong[a], Ryan Tarpine[b], Sorin Istrail[b], and Eric H. Davidson[a,1]

[a]Division of Biology, California Institute of Technology, Pasadena, CA 91125; and [b]Center for Computational Molecular Biology and Department of Computer Science, Brown University, Providence, RI 02912

Contributed by Eric H. Davidson, January 7, 2010 (sent for review November 19, 2009)

Gene expression is controlled by interactions between *trans*-regulatory factors and *cis*-regulatory DNA sequences, and these interactions constitute the essential functional linkages of gene regulatory networks (GRNs). Validation of GRN models requires experimental *cis*-regulatory tests of predicted linkages to authenticate their identities and proposed functions. However, *cis*-regulatory analysis is, at present, at a severe bottleneck in genomic system biology because of the demanding experimental methodologies currently in use for discovering *cis*-regulatory modules (CRMs), in the genome, and for measuring their activities. Here we demonstrate a high-throughput approach to both discovery and quantitative characterization of CRMs. The unique aspect is use of DNA sequence tags to "barcode" CRM expression constructs, which can then be mixed, injected together into sea urchin eggs, and subsequently deconvolved. This method has increased the rate of *cis*-regulatory analysis by >100-fold compared with conventional one-by-one reporter assays. The utility of the DNA-tag reporters was demonstrated by the rapid discovery of 81 active CRMs from 37 previously unexplored sea urchin genes. We then obtained simultaneous high-resolution temporal characterization of the regulatory activities of more than 80 CRMs. On average 2–3 CRMs were discovered per gene. Comparison of endogenous gene expression profiles with those of the CRMs recovered from each gene showed that, for most cases, at least one CRM is active in each phase of endogenous expression, suggesting that CRM recovery was comprehensive. This approach will qualitatively alter the practice of GRN construction as well as validation, and will impact many additional areas of regulatory system biology.

high-throughput discovery | sea urchin gene regulation

Genomic regulatory systems control development of the body plan, morphogenesis, differentiation, and physiological response. The primary mechanism in gene regulation is interaction of transcription factors with *cis*-regulatory modules (CRM). For this basic reason, experimental functional analysis of *cis*-regulatory interactions provides the primary validation of predictive, system level models of gene regulatory networks (1). However, *cis*-regulatory examination on the scale of the large networks now coming on line is a formidable proposition, given the present technological limitations; our object in the present work has been to surmount these limitations. The sea urchin embryo gene regulatory networks (GRNs) are the most comprehensive developmental GRNs currently available (2–5). Published sea urchin embryo ectoderm, and endomesoderm GRNs contain >80 regulatory genes, as well as various signaling and other genes, together with the inputs, outputs, and predicted regulatory interactions of all of these genes. Recent progress in the area of experimental and computational tools has substantially accelerated the construction of experimentally based GRN models. In the sea urchin and other model systems, GRNs are solved by prediction of regulatory inputs and outputs on the basis of matrices of gene perturbation results, together with spatial and temporal gene expression data, superimposed on background knowledge of the developmental biology of the system. The sea urchin GRNs have successfully provided mechanistic causal explanations of the developmental process. They have generated insights into the principles of GRN organization, leading to formulation of more advanced hypotheses on developmental regulatory systems; and they have been instrumental in generating new theories of evolutionary process (1, 4, 6, 7).

To validate network topology, predicted *trans*-regulatory inputs must be authenticated by isolation of the relevant CRMs, followed by test of the functionality of the predicted inputs by mutation of the transcription factor target sites in an appropriate CRM expression construct (reviewed in ref. 8). At the cost of substantial effort over the last several years, regulatory inputs have been authenticated at many of the key nodes of the sea urchin GRNs, gene by gene (for current status accessible at http://sugp.caltech.edu/endomes/). However, many predicted *trans*-regulatory inputs still remain to be investigated. Furthermore, as technical advances accelerate acquisition and analysis of system-wide perturbation results and measurement of gene expression as well (9, 10), the number of predicted regulatory inputs requiring validation will grow at an increased rate. It is inevitable that the magnitude of this challenge will increasingly exceed the capacity of current gene-by-gene methods of *cis*-regulatory analysis to handle. *Cis*-regulatory analysis technology clearly represents a major bottleneck of great significance for the future of GRN bioscience. More broadly, there are many other kinds of genomically oriented studies that also require experimental tests of in vivo function for large sets of candidate CRMs (11, 12–14).

Here, using sea urchin eggs, we show that use of a DNA-tag system enables simultaneous introduction and subsequent deconvolution of large numbers of CRM expression constructs. Remarkably, these do not interfere with each other's activity. We demonstrate the usefulness of the tag system by rapid recovery of >80 active CRMs responsible for temporal expression profiles of 34 regulatory genes that had never before been studied at the *cis*-regulatory level. We then simultaneously obtained the expression kinetics of all of these CRMs, which made possible a systematic comparison of the endogenous expression profiles with those of the set of CRMs isolated for each gene.

Results

Theory and Overview of Approach. Conventionally, *cis*-regulatory analysis requires (*i*) that the regulatory DNA sequence be placed in an expression vector in which it causes expression of a reporter gene; (*ii*) that the construct be introduced by some method of gene transfer into a living system, preferably one in which it is incorporated into the genome; (*iii*) that its function can be assessed by measurement of reporter expression in vivo, often after mutation of selected target sites. *Cis*-regulatory constructs can be efficiently assembled in sets by fusion PCR (15, 16) (*Methods*), and the major bottleneck is in the gene transfer procedure and measurement of construct output, which have always been done only one or two constructs at a time. In the sea urchin embryo, the method of gene transfer is injection of linear constructs anywhere into egg cytoplasm, after which the exogenous DNA is immediately concatenated and taken up into one of the first few blastomeres where it is stably integrated and replicated together with the host chromosome (17–21). We had found earlier that pairs of constructs

Author contributions: J.N. and E.H.D. designed research; J.N. and P.D. performed research; R.T. and S.I. provided new reagents/analytical tools; J.N. analyzed data; and J.N. and E.H.D. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: davidson@caltech.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/1000147107/DCSupplemental.

injected together, e.g., a control and a mutated CRM, do not interfere with each others' expression (22). Capitalizing on this observation, to begin this work we conducted experiments in which multiple different constructs were injected together to the same total mass of exogenous DNA as conventionally used for one or two constructs (always including about a 7-molar excess of total sea urchin DNA fragments of a nominal 10-kb length as carrier). Greater amounts of total exogenous DNA per egg can produce nonspecific toxic effects. RNA and genomic DNA were extracted from the embryos at various stages. To identify positive regulatory activities of each of the individual constructs, unique seqence tags designed to facilitate quantitative PCR (QPCR) detection were used to mark each vector (the "barcode"). These tags were positioned in the vectors so that the amount of QPCR product would provide the quantity of transcript driven off the *cis*-regulatory module being tested in each vector. Later, the number of different constructs cointroduced was increased to >100, and this required an additional feature, i.e, a universal amplification primer site, the same for every tagged vector. The amount of total DNA introduced was sufficient to include 200–400 loaded vectors per egg, and so there is a small but significant chance that any given egg will not receive any given construct. However this is of no consequence, as eggs were injected and analyzed in batches of ~100, and each transcript product was normalized to the number of molecules of incorporated construct present in the genomic DNA, also measured by use of its specific QPCR tag (21).

Figure 1*A* shows the structure of the DNA-tag reporters developed in this study. Each reporter construct contains a pair of unique DNA tags, TagF, and TagR, used as QPCR priming sites, and a GFP coding sequence. The amount of GFP mRNA read off this sequence gives the quantitative activity of the construct, after normalization to the incorporated vector quantity, and the translated GFP fluorescence also permits microscopic spatial observation if desired. Two multiplexed *cis*-regulatory systems were used in the work reported here, one consisting of 13 vectors, and the other of 129 vectors. The universal primer sites included in the 129-vector set (orange in Fig. 1*A*) were used to amplify the entire pool of DNA-tag reporters, plus a part of the GFP sequence (magnified region in Fig. 1*A*). Both cDNA prepared from injected embryo transcript, and genomic DNA isolated from the same embryos, were used as template for QPCR measurements (21). This signal amplification step significantly reduces the number of embryos required for reliable QPCR results (for tag reporter vector sequences and primers in *SI Appendix*). The 129 tags were chosen from an initial set of 150 tags culled to remove any tags that produced observable intrinsic regulatory activity in empty vectors, and when loaded with the same active *cis*-regulatory module their activities vary <2-fold (*SI Appendix*).

In a proof-of-principle experiment, regulatory activities of three known active CRMs were successfully recapitulated using the tag vector system (*SI Appendix*). The expression levels of the DNA-tag reporters driven by these active CRMs were easily distinguished from those of those produced by DNA sequence known to have no positive regulatory activity. We then proceeded to a test of the real-life usefulness of the multiplexed tag vectors for high-speed isolation of a large set of previously unknown *cis*-regulatory modules.

High-Throughput Recovery of Active CRMs. Thirty-seven sea urchin (*S. purpuratus*) genes, which are expressed in at least one embryonic territory within the first 2 days of embryogenesis (23–25) were chosen to test the tag system as a method for CRM discovery. These genes were as follows: *atbf1*, *chordin*, *dach*, *dlx*, *dri*, *ecr*, *elk*, *ese*, *ficolin*, *foxb*, *foxg*, *foxj1*, *foxk*, *foxn2/3*, *foxo*, *gatac*, *gsc*, *hnf1*, *id*, *irxa*, *lim1*, *myc*, *nfe2*, *nk1*, *nk2.2*, *not*, *prox*, *shr2*, *sip1*, *six1/2*, *soxb1*, *tbx2/3*, *unc4.1*, *z13*, *z188*, *bmp2/4*, and *univin*. None had ever been characterized at the *cis*-regulatory level, but most had been included, if only tentatively, in GRN models (3, 23–26, 27). For 35 of the 37 genes, we were able to use interspecific sequence comparison to narrow the search space. We had found earlier that conservation of noncoding sequence between *S. purpuratus* (*Sp*) and *Lytechinus variegatus* (*Lv*) is an exceedingly efficient guide to finding putative CRMs in *S. purpuratus* (28, 29). The lineages leading to these two species diverged ~50 million years ago (30).

Previous comparisons of *Sp* and *Lv* genomic sequence required ordered and oriented *Lv* BAC sequences containing the gene of interest plus complete flanking sequence (i.e., the BACs are of average length 140 kb, whereas the average sea urchin intergenic space is ~30 kb) (31), as the genome of *Lv* is not yet sequenced. Here, however, we used a much faster and much less expensive strategy, capitalizing on the Ilumina DNA sequencer. DNA from up to six *Lv* BAC clones was mixed per sequencing lane and 32 base (b) reads obtained (the first 25 b were used for the analysis). Coverage of useable reads was greater than ×20. The pooled reads were directly mapped in a genome browser ("*cis*-Browser," available on request from S. Istrail, Brown University, Providence, RI) onto the corresponding scaffolds of the *Sp* genome, allowing up to four mismatches within the 25 b window (screenshot in *SI Appendix*). On the browser those locations where the *Sp* and *Lv* sequences were conserved within these limits (i.e., exons and putative CRMs) the Illumina reads piled up, thus identifying the

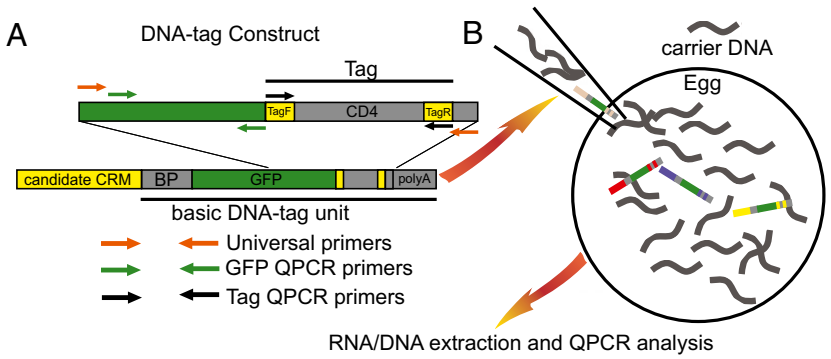

Fig. 1. Parallel measurement of *in vivo cis*-regulatory activities of many DNA sequences by using the DNA-tag reporters. *(A)* The structure of a DNA-tag reporter construct. Each basic unit of DNA-tag reporter construct is composed of a basal promoter (BP) from the sea urchin gatae gene (35), a GFP ORF (ORF), a pair of DNA-tags flanking a 145-bp-long fragment of human CD4 cDNA, and a core poly-adenylation (*A*) signal (39). For a later set of 129 DNA-tag reporters, a pair of primer sites (orange coded regions) were introduced to amplify the entire pool of DNA-tag reporters and a part of the GFP ORF (magnified region) either from cDNA or genomic DNA isolated from injected embryos. *(B)* A pool of many DNA constructs is injected with seven molar excess amount of randomly sheared genomic DNA as carrier/spacer (40). Coinjected linear DNA molecules form random concatenates and are incorporated into chromosomes in a mosaic fashion (17). Expression of each tag reporter is measured by QPCR following the method developed by Revilla-i-Domingo et al. (21).

map locations in the *Sp* sequence of the regions to be tested for function. This heuristic version of phylogenetic footprinting is an efficient method of examining sequence conservation when genome sequence for only one species of interest is available.

In this way, several hundred candidate CRMs were identified as conserved sequence patches from 35 genes, isolated by PCR and fused into the tag vectors. The incorporated fragments ranged from ~400-bp single patches to ~4 kb, for sequences that contained multiple, closely positioned conserved patches. Their sequences and their genomic coordinates are provided in *SI Appendix*. For two additional genes, *univin* and *bmp2/4*, *Lv* BACs were not available, and only a truncated BAC was recovered for the *lim1* gene. In these cases, the intergenic and intronic regions were blindly divided into fragments ~3 kb in length and fused into the tag vectors. A total of 390 constructs containing candidate CRMs were generated by fusion PCR, using the 13-tag vector system, and divided into 13-construct pools. These were injected into ~1,200 eggs per pool over a 3-day period, using three different females, several hours of injection per day. Normalized expression levels were measured for each tag vector in the presence of the others in the same pool at four different time points: 12, 24, 36, and 48 h postfertilization (hpf). The background activity level was established for each batch of eggs as described in detail, as are all other procedures, in *SI Appendix*, and any candidate CRM generating reporter expression at least 2.5-fold above background was considered active in this study (equivalent to a *P*-value cutoff of 0.01).

Results for the *foxn2/3* locus are shown as an example in Fig. 2. In Fig. 2*A* the gene structure and the pattern of interspecific sequence conservation are summarized; this gene has many regions upstream of the gene, in its introns, and downstream, that display possibly meaningful sequence conservation (blue and green bars in Fig. 2 legend). A total of 17 candidate CRMs were examined for this locus, as indicated below the browser view, four upstream (U_01 and U_03 - U_05), 11 within the introns of the gene (I_01–I_11), and two downstream (D_01 and 02). As can be seen in Fig. 2*B*, three of these constructs displayed significant activity at different times. Construct U_04 was active at 12, 24, 36, and 48 h; U_01 only in the 24, 36, and 48 h measurements; and I_09 only in the 12-h dataset. The remainder of the constructs showed insignificant activity (statistical criterion of significance in *SI Appendix*). Note that the values on the ordinate represent the relative numbers of transcript molecules per molecule of incorporated construct DNA to the background expression. In a subsequent measurement described below, detailed activity time courses were obtained for all of the active modules discovered in this study, and the time courses for the three active *foxn2/3* modules, U_04, U_01, and I_09, are reproduced in Fig. 2*C* together with the endogenous *foxn2/3* time course. The high-resolution time course is entirely consistent with the results of the high-throughput, four-point activity screen. Because every DNA construct encodes GFP, those that are active can be used without further modification for examination of spatial expression. The entire set of results for the 13 DNA-tag experiments are provided in *SI Appendix*.

Overall results of the multiplex CRM discovery experiment for the 37 genes are summarized in Fig. 3 and Table 1. Of the 37 genes examined, we found at least one active CRM for 34 genes (Fig. 3). Two or more active CRMs were detected for 20 genes, and the largest number of CRMs detected for one gene was five. Collectively, 81 of 390 candidate CRMs tested were found to be active. All but a few of the fragments tested included some interspecifically conserved sequence elements, and statistical data for these are shown in Table 1. As mentioned above, it took 3 days for the microinjections of 390 candidate CRMs from 37 genes; we estimate that it would have required at least a month of continuous egg injection, had the same fragments been examined by our usual one-by-one methods. The 13-tag system thus yielded at least a 10-fold increase in CRM discovery rate, at

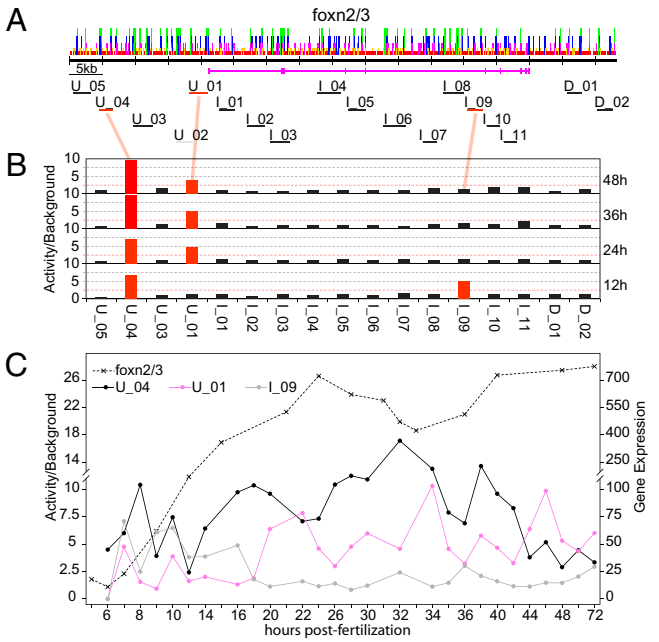

Fig. 2. *Cis*-regulatory modules for the *foxn2/3* locus. *(A)* Conserved DNA patches and candidate CRMs selected for the *foxn2/3* locus. Black horizontal line with ticks represents genomic sequence of the locus. Interval between ticks on black line is 5 kb. Color-coded vertical lines above the black horizontal line indicate sequence conservation between *Sp* and *Lv* for each 25-bp-long window: green, 0 mismatch; blue, 1 mismatch; pink, 2 mismatches; yellow, 3 mismatches; red, 4 mismatches. Thick and thin pink horizontal lines below black line indicate exons and introns of *foxn2/3* from 5′ to 3′, respectively. Horizontal lines with labels indicate candidate CRMs: red, active; yellow, marginally active; black, inactive; gray, cloning failed. *(B) Cis*-regulatory activities of 17 candidate CRMs from the *foxn2/3* locus. Normalized expression level after background correction is shown for each time point measured. The names of candidate CRMs are shown at the bottom. Expression level significantly higher (≥2.5) than the background level for each time point is red coded. Note that expression levels >10 are shown as 10. *(C)* High-resolution time course activities of CRMs measured using the 129 DNA-tag reporters over 27 different time points. CRM activities are corrected for background activities, and the relative expression level of *foxn2/3* is shown (*SI Appendix*).

the same time permitting quantitative activity measurement across developmental time. Parallel processing of samples for QPCR analysis was another significant improvement compared with one-by-one experimental methodologies (*Methods*).

Table 1 displays the frequency with which active CRMs were recovered as a function of position relative to the gene. These data are of limited quantitative significance because no systematic

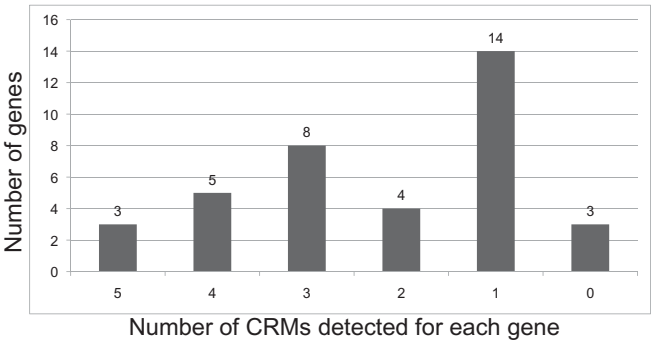

Fig. 3. Distribution of the numbers of CRMs discovered for each gene. Genes are categorized by the number of CRMs discovered, and the number of genes in each category is presented as a bar.

Table 1. Summary of CRM discovery by using the 13 DNA-tag system

Location	Active/tested	Proximal or distal	Active/tested
Upstream	42 / 134 (31%)	Proximal (U_01)	26*/36† (72%)
		Distal (≥U_02)	16/98 (16%)
Transcriptional unit	32 / 166 (19%)	Proximal (I_01)	9/32 (28%)
		Distal (≥I_02)	23/134 (17%)
Downstream	7 / 90 (8%)	Proximal (D_01)	3/33 (9%)
		Distal (≥D_02)	4/57 (7%)
Total	81 / 390 (21%) for 37 genes		

*5′-Proximal fragment of *unc4.1*, which was active only in antisense orientation, was not counted.
†5′-Proximal fragment of *nfe2* was not counted, as this fragment was not tested because of cloning failure.

attempt to sample the upstream, intron, or downstream regions was made, nor were the genomic fragments confined to conserved sequence but instead usually included both conserved and flanking nonconserved sequence in various proportions. The general qualitative conclusion is that active CRM are likely to occur in introns and far upstream of the transcription start site, or downstream of the poly(A) addition site, confirming for this relatively large sample what has been seen over and over in studies of specific genes (reviewed in ref. 1). One interesting feature is that the most proximal 5′ regions very frequently (72%) scored as active, although less than half of all active CRM recovered were from these regions. As the test was only for expression in the first 2 days of the life cycle of a long-lived animal, and as every regulatory gene is likely to be used multiple times during the life cycle (32), this could mean that the proximal regions participate in many different phases of expression, together with more specific distal enhancers. To determine whether these proximal regions must be oriented appropriately with respect to the nearby gene, we recloned them in antisense orientation and tested them for activity. However, 18 of 21 active proximal sequence fragments were also active in the reverse orientation; and so, in this respect, these fragments do not differ from distal enhancers.

Simultaneous, High-Resolution Temporal Measurement of CRM Activities with the 129 DNA-Tag Reporter System. The time course of CRM output is a biologically important parameter of regulatory function. Because multiple CRMs often contribute to the overall pattern of expression of a given gene, it is advantageous to examine activities of the relevant CRMs simultaneously instead of in different experiments. Furthermore, a typical system-wide GRN may contain >50 regulatory genes and might involve the activities of more than this number of relevant CRMs. To accelerate system-wide temporal *cis*-regulatory analysis on this scale, we developed the 129 DNA-tag reporter system. This permits simultaneous analysis of up to this number of CRMs in a single experiment. This system was used to generate high-resolution output time courses for all of the active CRMs discovered in the foregoing experiments, i.e., >80 active CRMs. Each was fused to one of the 129 DNA-tag reporters, and the constructs were mixed and coinjected into eggs, which were then allowed to develop. Measurements of CRM regulatory activities were carried out at 27 successive time points, up to 72 hpf. Although we now know that much less would have been sufficient, ~300 embryos were sampled for each of the 27 time points. Most of the CRMs identified as active using the 13 DNA-tag reporter system were also found to be significantly active in the 129 DNA-tag reporter system as well, and the few that were not were CRMs that had originally displayed marginal levels of activity. In only one exceptional case (that of the *ficolin* gene) were the two results significantly different. Such artifacts appear to be due to an occasional combinatorial interaction between a given CRM and a given DNA-tag pair. The entire set of results for the 129 DNA-tag experiments are provided in *SI Appendix*.

As an example, Fig. 2*C* illustrates data extracted from the large-scale experiment for the *foxn2/3* gene. The three CRMs found to be active with the 13 DNA-tag reporters (Fig. 2*B*) were also active here: U_04 was the most active throughout; U_01 activity achieved significant levels only at 20 h, *et seq*.; I_09 was functional only before 16 h. For comparison, the endogenous activity of the *foxn2/3* gene, as measured with NanoString nCounter system (9) in the same experiment, is also plotted. The three active CRMs can be seen qualitatively to account for the whole temporal range of expression, except perhaps for the latest period, 72 h. This last is not surprising, as the CRM discovery project identified CRMs active only up to 48 h.

Preliminary Measurement of Sufficiency of Recovered CRMs. How complete was CRM recovery for the 34 genes that produced active CRMs? This question cannot be answered with finality in the absence of spatial expression data. However, as a preliminary assessment, we used the high-resolution time course data to ask, for each gene, whether there had been recovered at least one CRM which produced regulatory output at each time point the endogenous gene is active. A "temporal sufficiency score" (S_t) was thus calculated as the proportion of time points when the gene is active, and CRM function was also observed, compared to the total number of time points when the gene is expressed. This is a strictly qualitative measure; although our data provide quantitative expression levels for the active constructs in terms of transcript molecules per incorporated construct molecule, we cannot directly compare these levels to the endogenous gene because we do not know what fraction of expressing cells contain the exogenous constructs, nor whether there is ectopic expression. In the event, none the less, the S_t values that we obtained are informative. If a complete set of CRMs were obtained, the S_t value would be 1. This is what happens, for example, if the output time course for a BAC construct including the complete regulatory system is compared with the endogenous gene output (33) or if the outputs of all of the relevant *cis*-regulatory modules in a well-studied gene are summed (e.g., the nodal gene) (22, 27).

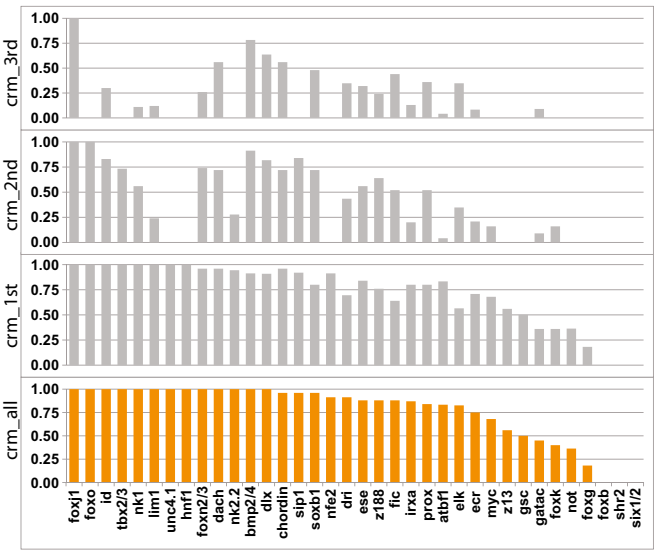

Fig. 4. Temporal sufficiency scores (S_t) of CRMs. The S_t value of the entire set of CRMs discovered for each gene (crm_all; orange coded) and the S_{t_crm} value of each CRM (crm_1st, crm_2nd, and crm_3rd; gray coded) are presented as bar graphs. Genes are ordered by the S_t values. When there are more than three CRMs discovered for a gene, only the top three CRMs ranked by S_{t_crm} value are shown.

SYSTEMS BIOLOGY

This preliminary analysis is shown in Fig. 4 (crm_all). Including the three genes for which no CRM was found, for which $S_t = 0$, 36 genes are represented (the S_t value for *univin* was not computed, as the temporal expression of endogenous *univin* has not been measured). We can see at a glance that for the great majority, $S_t > 0.8$, and for half, $S_t > 0.9$. Thus, with the caveats addressed below, these data suggest that recovery of CRMs was fairly comprehensive for most of the genes that we examined. For a minority, the recovered CRMs fail to account for some phase of the temporal output of the endogenous genes.

Figure 4 also shows an estimate, by the same metric, of overlap in temporal expression driven by the multiple CRMs of given genes. The same score can be computed for each CRM, S_{t_CRM}. Where there are entirely nonoverlapping, temporal outputs from the various CRM of each gene, $\Sigma\, S_{t\_CRM}/S_t = 1$; but in the case that there is overlap, i.e., more than one CRM operates at the same time, $\Sigma\, S_{t\_CRM}/S_t > 1$. The S_{t_CRM} values were computed for all of the CRMs, and the results are presented for the top three CRMs of each gene in Fig. 4 (crm_1st, crm_2nd and crm_3rd; gray bars). At least one CRM with $S_{t\_CRM} \geq 0.8$ was observed for most genes, and for others the sum of two $S_{t\_CRM} \geq 0.8$. The main import is that for most genes for which more than one CRM was recovered, $\Sigma\, S_{t\_CRM}/S_t > 1$, i.e., there appears to be some overlap in temporal output. This could indicate that overlap of CRM activity is a common phenomenon, as reported for Drosophila (34); but, as we discuss below, it may also quantitatively illuminate an important aspect of the regulatory behavior of short expression constructs such as those used in our multiplex tag system.

Discussion

In this work, we demonstrate the utility of the DNA-tag reporter systems for high-throughput, quantitative measurements of positive *cis*-regulatory activity: first, by identification de novo of 81 active CRMs for 34 genes; and second, by simultaneous high-resolution measurement of the temporal outputs of >80 active CRMs. There is no obvious reason why DNA-tag systems designed according to the same principles as demonstrated here should not work in other model systems. This approach may be directly transferrable to other animals or plants with only minor modifications, such as change of the basal promoter to a promiscuously active endogenous one.

Strengths and Weaknesses of High-Throughput *cis*-Regulatory Analysis Using Tag Vectors. The main advantage this advance confers is obvious: it provides a major increase in the efficiency and rate of the most important operation in experimental functional genomics, *cis*-regulatory analysis. The use of the 13 DNA-tag system literally afforded a greater than 10-fold increase in analysis rate, in addition presenting an opportunity for comparative measurements of multiple CRMs in the same experiment. The experiments using the 129 DNA-tag system could not even have been practically conceived using traditional methods, so although it may be calculated that the analysis rate is improved >100-fold, the real improvement is qualitative: we can now consider kinds of experiments that were before wholly out of bounds. Furthermore, many other aspects of *cis*-regulatory analysis than those that we chose to investigate in this study are amenable to acceleration by the same methods. For example, multiple site–specific CRM mutations could be examined at once in single experiments for their effects on regulatory output; and, again, such comparisons carried out in single experiments with a common set of controls offer superior experimental designs.

Several important caveats need to be considered. First, there are two essential aspects of *cis*-regulatory analysis that this initial development does not address. It does not provide spatial expression information; and it does not detect CRMs, the function of which is repression. Both of these objectives can, in principle, be attained by high-throughput procedures as well, using similar principles, and we are now engaged in just these projects. More serious perhaps are the inherent functional features of short expression constructs, which are the usual workhorses of *cis*-regulatory analysis. In recent studies in our laboratory, we have directly compared the functionality of short constructs containing given *cis*-regulatory modules with the activity of the same modules in context of the complete multi-module regulatory system, using recombineered BAC expression vectors (33, 35). A not-uncommon observation in such comparisons is that, in context, mechanisms of module choice mediate exclusive use of one module at a given time and exclusion of the others. However, in short constructs, where there is no choice, the basal promoter will use whatever it is provided with, and so the observed range of activity of short constructs exceeds that of the same module in context. This is certainly not always true, and sometimes the difference between expression of a BAC from which a module has been deleted and expression of the control BAC exactly equals prediction, based on the behavior of a short construct driven by that module. However, because the temporal readout of short constructs may exceed their normal function, the results of the analysis in Fig. 4 must be regarded only as indicating the outside possible limits of module overlap rather than as a measurement thereof. This will not affect the cases in which the modules of a given gene operate at different times, as this simply means that they respond to regulatory inputs presented in different developmental stages. Similarly, short constructs that express in various specific locations are perfectly reliable indicators of the spatial inputs to which they respond.

Multiple CRMs per Gene. The genes in this study are mostly regulatory genes, and the high fraction of these that have multiple CRMs is thus no surprise. However, the ease with which we were quickly able to recover multiple modules per gene deserves remark. So does the result of the S_t analysis, which, even given the above caveat, suggests that it is usually not difficult to recover CRMs that encompass all phases of a gene's activity. The technology is sufficiently powerful that, even in the absence of interspecific conservation information, or irrespective of it, a large intergenic region can be divided into 3–5 kb (or larger) pieces blind, and all of them could be assayed in a single experiment. Although we did not systematically attempt to do this, the way is now open to recover a priori all positive CRM for any gene of interest, by examining all possible sequence space in which that gene's CRMs might exist. Of course the problem of determining which gene in the vicinity of a CRM is being regulated by it requires further information; because we are mainly interested in specifically expressed genes, spatial expression data for newly discovered CRMs is here critical. Evidence of the potential functions of all CRMs constituting the overall control system of a gene will be invaluable when carried out in conjunction with in-context *cis*-regulatory analysis of that gene. This evidence provides the baseline of potential individual CRM function against which the operation of the system when it is whole can be compared. Thus we believe that the high-throughput CRM discovery methodology should materially contribute to progress in a largely unexplored, but conceptually very important, area of regulatory molecular biology, control of use of alternate CRMs.

Implications of High-Throughput *cis*-Regulatory Methodology for Solving Gene Regulatory Networks. Gene regulatory networks are at present solved by "top down" methods, in which system-wide perturbation and expression data are used to construct the network model. Following this a crucial step is validation of the predicted linkages in the network, by isolation of the relevant *cis*-regulatory modules and test of the functionality of the target sites mediating responses to the regulatory inputs. Now, however, we can consider an approach based on the use of the high-throughput *cis*-regulatory systems ab initio, in which the analysis

of modular *cis*-regulatory responses to perturbations of gene expression is assayed simultaneously with analysis of the effects of the perturbation on the endogenous genes. This will indicate which *cis*-regulatory module is relevant to the network in question, and will greatly aid in distinguishing direct from indirect linkages even as the network is being formulated. In addition, the quantitative and internally controlled data acquisition procedures that the tag system makes possible should facilitate subsequent computational and statistical analyses. The outcome will be to revolutionize GRN analysis where these methods can be applied, producing draft GRNs that from the beginning use *cis-trans* interactions to determine network architecture.

Methods

Design and Generation of DNA-Tag Reporter Vectors. More than 500 sequences of potential PCR primers were computationally selected from randomly generated sequences using the Primer3 program (36) or the FastPCR program (37). To filter tag sequences for potential matches to the genome or transcriptome, these sequences were BLASTed (38) against the genome and annotated genes of *S. purpuratus* (31) with an E-value cutoff of 1. A similar test was performed among the tag sequences to avoid primer dimerization or nonspecific PCR amplication. To generate reporter vectors shown in Fig. 1*A*, each pair of tags flanking a common fragment of human CD4 cDNA was cloned into a vector containing an *Sp-gatae* basal promoter (35), a GFP ORF, and a core polyA signal (39). Details of these computational and experimental procedures are provided in *SI Appendix.* Sequences of reporter vectors are also provided in *SI Appendix.*

Amplification of Candidate CRMs and Generation of Reporter Constructs. The sequences of candidate CRMs were extracted from the genome sequence and, when possible, the flanking 50 bp were used for designing PCR primers using the Primer3 program. In general, two forward primers and one reverse primer were designed for each candidate CRM. Candidate CRMs were then connected to one of the 13 or 129 basic units of tag reporters, which were amplified by PCR and column purified before fusion PCR. Details of these experimental procedures are provided in *SI Appendix.* Sequences of candidate CRMs and sequences of primers for each candidate CRM are provided in *SI Appendix.*

Microinjection and QPCR Analysis. Microinjection was performed as previously described (40) with a slight variation to account for the number of constructs in each injection (22). Details of microinjection are provided in *SI Appendix.*

Approximately 200–300 microinjected embryos were collected for each time point. AllPrep DNA/RNA micro kit (Qiagen) was used to simultaneously extract genomic DNA and total RNA. The number of expressed tag reporters was normalized to the number of DNA copies incorporated (21). Background normalization was also applied to correct for batch and developmental stage–specific background expression level. The details of these experimental and analytical procedures are provided in *SI Appendix*.

ACKNOWLEDGMENTS. The authors appreciate the following individuals for their contributions to this work: Miki Yun for sequencing; Julie Hahn for providing some *Sp*BAC DNAs for templates; Lydia Dennis and Andy Cameron for some of *Lv*BAC screening; Ali Mortazavi, Lorian Schaeffer, and Barbara Wold for Illumina sequencing; and Julius Barsi, Michael Collins, Sagar Damle, Smadar Ben-Tabou DeLeon, David McClay, and Joel Smith for their criticisms and suggestions on an earlier version of the manuscript. This work was supported by National Institutes of Health Grants GM061005 and HG00533201, the Caltech Beckman Institute (E.H.D.), and National Science Foundation Grant 0645955 (to S.I.).

1. Davidson EH (2006) *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution* (Academic Press/Elsevier, San Diego).
2. Davidson EH, et al. (2002) A genomic regulatory network for development. *Science* 295:1669–1678.
3. Su YH, et al. (2009) A perturbation model of the gene regulatory network for oral and aboral ectoderm specification in the sea urchin embryo. *Dev Biol* 329:410–421.
4. Peter IS, Davidson EH (2009) Genomic control of patterning. *Int J Dev Biol* 53:707–716.
5. Oliveri P, Tu Q, Davidson EH (2008) Global regulatory logic for specification of an embryonic cell lineage. *Proc Natl Acad Sci USA* 105:5955–5962.
6. Davidson EH, Erwin DH (2006) Gene regulatory networks and the evolution of animal body plans. *Science* 311:796–800.
7. Erwin DH, Davidson EH (2009) The evolution of hierarchical gene regulatory networks. *Nat Rev Genet* 10:141–148.
8. Smith J (2008) A protocol describing the principles of cis-regulatory analysis in the sea urchin. *Nat Protoc* 3:710–718.
9. Geiss GK, et al. (2008) Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol* 26:317–325.
10. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628.
11. Birney E, et al.; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799–816.
12. Dimas AS, et al. (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325:1246–1250.
13. Visel A, et al. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457:854–858.
14. Katzman S, et al. (2007) Human genome ultraconserved elements are ultraselected. *Science* 317:915.
15. Yon J, Fried M (1989) Precise gene fusion by PCR. *Nucleic Acids Res* 17:4895–4895.
16. Hobert O (2002) PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic C. elegans. *Biotechniques* 32:728–730.
17. McMahon AP, et al. (1985) Introduction of cloned DNA into sea urchin egg cytoplasm: Replication and persistence during embryogenesis. *Dev Biol* 108:420–430.
18. Flytzanis CN, Britten RJ, Davidson EH (1987) Ontogenic activation of a fusion gene introduced into sea urchin eggs. *Proc Natl Acad Sci USA* 84:151–155.
19. Franks RR, Hough-Evans BR, Britten RJ, Davidson EH (1988) Direct introduction of cloned DNA into the sea urchin zygote nucleus, and fate of injected DNA. *Development* 102: 287–299.
20. Livant DL, Hough-Evans BR, Moore JG, Britten RJ, Davidson EH (1991) Differential stability of expression of similarly specified endogenous and exogenous genes in the sea urchin embryo. *Development* 113:385–398.
21. Revilla-i-Domingo R, Minokawa T, Davidson EH (2004) R11: A cis-regulatory node of the sea urchin embryo gene network that controls early expression of SpDelta in micromeres. *Dev Biol* 274:438–451.
22. Nam J, et al. (2007) Cis-regulatory control of the nodal gene, initiator of the sea urchin oral ectoderm gene network. *Dev Biol* 306:860–869.
23. Howard-Ashby M, et al. (2006) Identification and characterization of homeobox transcription factor genes in Strongylocentrotus purpuratus, and their expression in embryonic development. *Dev Biol* 300:74–89.
24. Howard-Ashby M, et al. (2006) Gene families encoding transcription factors expressed in early development of Strongylocentrotus purpuratus. *Dev Biol* 300:90–107.
25. Materna SC, Howard-Ashby M, Gray RF, Davidson EH (2006) The C2H2 zinc finger genes of Strongylocentrotus purpuratus and their expression in embryonic development. *Dev Biol* 300:108–120.
26. Pancer Z, Rast JP, Davidson EH (1999) Origins of immunity: Transcription factors and homologues of effector genes of the vertebrate immune system expressed in sea urchin coelomocytes. *Immunogenetics* 49:773–786.
27. Range R, et al. (2007) Cis-regulatory analysis of nodal and maternal control of dorsal-ventral axis formation by Univin, a TGF-beta related to Vg1. *Development* 134: 3649–3664.
28. Yuh CH, et al. (2002) Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin. *Dev Biol* 246:148–161.
29. Brown CT, Xie Y, Davidson EH, Cameron RA (2005) Paircomp, FamilyRelationsII and Cartwheel: Tools for interspecific sequence comparison. *BMC Bioinformatics* 6(70):1–7.
30. Smith AB, et al. (2006) Testing the molecular clock: Molecular and paleontological estimates of divergence times in the Echinoidea (Echinodermata). *Mol Biol Evol* 23: 1832–1851.
31. Sodergren E, et al.; Sea Urchin Genome Sequencing Consortium (2006) The genome of the sea urchin Strongylocentrotus purpuratus. *Science* 314:941–952.
32. Howard-Ashby M, et al.; (2006) High regulatory gene use in sea urchin embryogenesis: Implications for bilaterian development and evolution. *Dev Biol* 300: 27–34.
33. Wahl ME, Hahn J, Gora K, Davidson EH, Oliveri P (2009) The cis-regulatory system of the tbrain gene: Alternative use of multiple modules to promote skeletogenic expression in the sea urchin embryo. *Dev Biol* 335:428–441.
34. Hong JW, Hendrix DA, Levine MS (2008) Shadow enhancers as a source of evolutionary novelty. *Science* 321:1314.
35. Lee PY, Nam J, Davidson EH (2007) Exclusive developmental functions of gatae cis-regulatory modules in the Strongylocentrorus purpuratus embryo. *Dev Biol* 307: 434–445.
36. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132:365–386.
37. Kalendar R, Lee D, Schulman AH (2009) FastPCR Software for PCR Primer and Probe Design and Repeat Search. *Genes. Genomes Genomics* 3(1):1–14.
38. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
39. Nag A, Narsinh K, Kazerouninia A, Martinson HG (2006) The conserved AAUAAA hexamer of the poly(A) signal can act alone to trigger a stable decrease in RNA polymerase II transcription velocity. *RNA* 12:1534–1544.
40. Arnone MI, Dmochowski IJ, Gache C (2004) Using reporter genes to study cis-regulatory elements. *Development of Sea Urchins, Ascidians, and Other Invertebrate Deuterostomes: Experimental Approaches.* Methods in Cell Biology, eds Ettensohn CA, Wessel GM, Wray GA (Elsevier Academic Press, San Diego), Vol 74, pp 621–652.