PROPOSAL FOR SEQUENCING FURTHER ECHINODERM GENOMES FOR THE PURPOSES OF GLOBAL GENE REGULATORY NETWORK ANALYSIS

Eric H. Davidson, R. Andrew Cameron,
California Institute of Technology. Pasadena CA 91125

and
Richard Gibbs
HGSC, Baylor University School of Medicine, Houston TX 77030

Gene regulatory networks (GRNs), when solved, provide immediate and predictive access to the specific functional termini at which genomic sequence determines the logic of the organismal regulatory system. GRN topology explains the fundamental processes of development (1), including dynamics (2), major evolutionary transitions (3), many aspects of physiological function, and normal and diseased regulatory states; and GRN structure also provides the only kind of blueprint on which to base rational strategies for re-engineering of developmental process (4,5). Thus in the next period of bioscience the elucidation of GRNs must become a major, targeted objective.

The most comprehensive examples of GRNs sufficiently mature to demonstrate their full range of predictive and explanatory power are at present those worked out experimentally for the embryo of the sea urchin (*Strongylocentrotus purpuratus, Sp*) (1,6). Acquisition of the genome sequence for *Sp* (7) was and continues to be indispensable for the solution and maturation of these GRNs. Using the current and near future *Sp* GRNs for validation, there now exists the opportunity to take a giant step forward if two additional echinoderm genome sequences were to become available. Specifically, acquisition of the genome sequence of the echinoid (sea urchin) species *Lytechinus variegatus* (*Lv*) could, as considered below, lead to the first algorithms (that work) for large scale prediction of overall GRN structure from genomic sequence. This would vastly accelerate experimental solution of GRNs, in that it would reverse the current protocols, in which the architecture of the GRN has first to be built up from very large scale perturbation analyses; experimental work would be oriented directly toward validation of predicted linkages. Acquisition of the genome sequence of the asteroid (sea star) *Asterina miniata* (*Am*) would produce direct evidence, and provide predictive principles, for assessing what types of GRN subcircuit are flexible in evolution and thus could be reorganized or altered, and which are virtually unchanging and inflexible (3). This kind of information, which is at present almost non-existent, will be essential in design of GRN re-engineering projects. An initial series of comparative GRN studies (8-10) have shown that *Am* is the right distance from *Sp* for this purpose: their last common ancestor was just post-Cambrian; their developmental processes are in many respects orthologous; but the distance is so great that everything that can change has changed.

Predictive Construction of GRN Models from Genomic Sequence Comparison: The *Lytechinus variegatus* Genome (840Mb)

The specific proposal that follows is designed to produce a priori predicted GRN structures for any given time-space domain of the post-gastrular sea urchin embryo. Our current approaches will soon yield GRNs of satisfactory levels of maturity for the pre-gastrular embryo, but to expand GRN analysis to the more complex set of structures that then arise thereafter will present a new level of challenge. Yet it is here, in GRNs for the specification of midgut, hindgut, foregut, neurogenic territories, etc., that the most direct comparative insight into vertebrate developmental GRNs is to be expected. The proposal that follows is based on two premises: (1) that for specification of any given time-space domain we can identify all relevant regulatory genes a priori; (2) that we can identify a sufficiently narrow genomic sequence search space that will include all relevant *cis*-regulatory modules of these genes. The proposal requires a reasonably high quality *Lv* genome sequence.

(1) Identification of regulatory players: In the course of the *Sp* genome project(11-15) and other ongoing studies (16), we and others identified all genes predicted to encode DNA-binding domains (i.e., of transcription factors). We then determined the developmental time courses of expression of these genes; and their spatial domains of activity. Thus we have on hand extensive lists of those regulatory genes expressed differentially in each given time-space domain of the embryo. Current knowledge of the spatial and temporal expression of signaling genes during *Sp* development is also extensive (7), but is less complete than for regulatory genes. Here we propose to ensure the completeness of these repertoires of specifically expressed regulatory and signaling genes (and also to obtain a sample of specific downstream differentiation genes) for each given spatial domain of the *Sp* embryo for the period of development up to the end of gastrulation. To do this we will exploit the currently extant libraries of *Sp cis*-regulatory modules: there are known genes the expression patterns of which define every developmental time-space domain in this embryo. We will inject into eggs GFP BAC knock-ins in genes expressed specifically in each given domain, e.g., *endo16* in the midgut (a broad set of these GFP knock-ins which all include the native regulatory sequences already exists; see http://sugp.caltech.edu/SpBase/recomb_bac/index.php). The embryos will be harvested, disaggregated, and the GFP positive cells separated out by FACS. Their mRNA will be isolated, amplified, subjected to deep sequencing, and the levels of expression of identified regulatory and signaling genes that are expressed particularly in these domains (compared to unselected mRNA) determined by analysis of the sequence reads. Extant knowledge of developmental expression categories for these types of genes will serve both as control, and as the platform for augmentation. In our experience in the sea urchin embryo ~20-50 regulatory genes are required for specification of any given developmental territory. Those selected genes not previously identified in given territorial contexts will of course be checked by in situ hybridization and QPCR.

(2) Identification of putative cis-regulatory sequence: For 10 years we have successfully utilized interspecific genomic sequence comparison against *Lv* BAC sequence to identify putative *cis*-regulatory modules in the *Sp* genome around selected genes of interest. The genome of *Lv* is appropriately diverged from that of *Sp* so that unselected sequence cannot be aligned at all due to intensity of occurrence of indels and SNPs, while exons and regulatory sequence patches are easily detected (17-19). The relevant *cis*-regulatory modules at most major nodes of the endomesoderm GRN of the *Sp* embryo have been discovered in this way (all nodes identified by thick lines in the BioTapestry network image at http://sugp.caltech.edu/endomes/#EndomesNetwork), as described in many publications. Though there are multiple sharply conserved sequence patches around any typical regulatory gene (80-90% identity over 10-50 bp windows for 100 to several hundred contiguous bp), the particular desired *cis*-regulatory module will always be found among them on experimental test. To expand GRN analysis globally, a *Lv* genome sequence is now required. It must be of sufficient quality that the complete intergenic sequence on either side of a gene of interest will be included in the contig containing that gene. The repertoire of conserved *Sp* sequence patches will then be available by appropriate computational comparison to the *Lv* sequence around each relevant gene for each territory or developmental domain.

(3) Computing a priori GRN architecture. In a specification GRN the inputs of each regulatory gene are outputs from another regulatory gene in the GRN. The inputs to the differentially expressed signaling genes activated in the course of specification are among the outputs of the GRN regulatory genes, and the termini of signal inputs into the GRN are also among the GRN regulatory genes. The differentiation genes activated at the periphery of the GRN use as drivers a subset of the same GRN regulatory genes. Target site (PWM) databases for transcription factors are rapidly improving and a great many more such sites will soon become available from the chip based assays of Martha Bulyk and associates. Thus for each space-time domain of the advanced embryo a finite set of 20-50 regulatory genes (from ***(1)*** above) interacting with a distinct set of target sites will be relevant; furthermore, the locations of these target sites will be clustered within the confined set of conserved sequence patches (from ***(2)*** above) surrounding this same set of regulatory genes. These constraints enable a computational prediction. Making use of the starting order of the regulatory genes, i.e., knowledge from the earlier GRNs and the expression time courses which are the initial genes, the conserved patches will be ranked as to the probability that each is nonrandomly likely to operate as a functional GRN *cis*-regulatory node. The probability will be based on the set of putative target sites for factors encoded by the selected set of regulatory genes for that domain. These predictions will constitute predicted nodes of the GRN and hence its linkages and hence its structure. The computation will be done hierarchically, beginning with the predicted initial genes, then their target genes, etc, and ending with the selected differentiation genes. Since developmentally active signals operate through canonical transcriptional regulatory inputs, e.g., Su(H) for Notch, Tcf for Wnt, Gli for HH, etc., the signal linkages into the GRN can also be predicted and the target genes placed thereby in the hierarchy. Meanwhile, since experimental check of putative *cis*-regulatory modules is easy in the sea urchin gene transfer system the top ranked predicted modules can be tested. But as indicated above, targeted checking of a predicted GRN structure would represent a vast saving of experimental labor as well as establishment of a new sequence based approach to GRN structure.

Modular Flexibility of GRN Subcircuits: The *Asterina miniata* Genome (690Mb)

The genomic sequence of the sea star *Am* would permit large scale test of the proposals that have been made (3) for which types of GRN subcircuit will turn out to be redeployed easily and frequently and which are refractory to redeployment (or re-design). For example it was argued that signaling inputs into GRNs are so flexible that they can be described by the term "plug-in", given their endless re-deployment in evolution of developmental regulatory systems. The sequence of biochemical steps in signal transmission and the immediate early response transcription factors are generally highly conserved in the metazoan world, but the regulatory apparatus into which the signaling cassettes are hooked in the GRN designs of different animals vary continuously, and this is in fact a major source of evolutionary novelty in development. On the other hand recursively wired, interlocking feedback loops are liable to live or die as a modular unit of architecture, and provide examples of unchanging circuit topology since the Cambrian (3,6). The *Am* genome is far too distant from that of *Sp* for "patchy" sequence conservation to have been retained, but orthologous *cis*-regulatory modules are detectable by searching for clusters of transcription factor target sites in the intergenic space of the orthologous gene (e.g., ref 20). As knowledge of *Sp* specification GRNs expands to more and more developmental domains, and achieves greater and greater levels of maturity, there will accumulate a deep collection of *Sp* GRN subcircuits of known structure/function relation. These could be searched for computationally by use of the predicted sets of input transcription factor target sites in the orthologous regions of the *Am* genome, were its sequence available. Like *Sp*, or perhaps even more so, *Am* is also a most accessible experimental system in which to check rapidly for predicted *cis*-regulatory modules. It is worth noting that it is difficult or impossible to identify as favorable a species pair as *Sp* and *Am* for the purpose of determining the rules of GRN flexibility, in any comparable set of animal model systems, in terms of their evolutionary distance; in terms of the relevant current basis of knowledge of GRN architecture, and in terms of easy experimental validation opportunities in both members of the pair. Thus an *Am* genome sequence would have a unique value in the quest to obtain a practical guide to understanding the flexibility of GRN structure. This type of information is very hard to come by at present but provides an invaluable guide to strategies for practical redesign of GRNs.

The Great Potential Usefulness of Additional Genomes

These themes could be powerfully extended in several dimensions were sequence available from additional genomes. This proposal is intended to be focused primarily on the specific pay-offs for understanding GRNs of sequencing the *Lv* and the *Am* genomes. But it is necessary to point out the specific scientific advantages along the same lines that would accrue from sequence of the genomes of three additional echinoderms and one basal ambulacrarian as an outgroup. These organisms are the euechinoid ("modern" sea urchin) *Arbacia punctulata* (*Ap*), the cidaroid ("primitive") sea urchin *Eucidaris tribuloides* (*Et*), the asteroid (sea star) *Dermasterias inbricata* (*Di*), and the hemichordate *Ptychodera flava* (*Pf*). Each of these animals is already the subject of various experimental studies, and we earlier constructed BAC libraries from each of these genomes, which are at present stored in our Caltech Genome Center.

Arbacia punctulata (700Mb): *Ap* is the most distantly related euechinoid with respect to *Sp* presently in frequent use as an experimental laboratory organism. Its last common ancestor with *Sp* is estimated to have existed about 150my ago (~3x the distance from *Lv*). The specific return to be expected from an *Ap* sequence pertains to deciphering the rules of *cis*-regulatory structure. Unlike *Sp-Lv*, for *Sp-Ap* comparison, the divergence is so great that no unselected sequence in a *cis*-regulatory module is conserved between it and *Sp* [in the *Sp-Lv* comparison, as we have earlier shown, change in sequence even between transcription factor target sites is slowed by suppression of one of the major sources of such change, the occurrence of insertions and deletions, resulting in long patches of high sequence conservation (18)]. Thus all that remains similar between *Sp* and *Ap* cis-regulatory modules are the transcription factor target sites themselves. About half of tested *Ap cis*-regulatory modules work appropriately on introduction into *Sp* eggs. Both those that do and those that do not provide an unmatched opportunity to capitalize on extensive *cis*-regulatory knowledge in *Sp*, to discover otherwise hidden rules of target site association and spacing. Unlike say mammals and fish, or *Sp* and *Am*, *Sp* and *Ap* are extremely similar in organismal structure and almost certainly developmental function, and genome wide *cis*-regulatory comparison will provide a perfect metric of allowed and unallowed change for similar regulatory function.

Eucidaris tribuloides (1800Mb)**:** *Et* is a representative of the sister group of all euechinoid sea urchins, which arose in the middle Triassic after the great extinction event at about 251 mya. Prior to the extinction only cidaroid sea urchins ("pencil urchins", after their thick spines) existed, and the euechinoids descend from a single Upper Paleozoic lineage of these which survived the extinction. They differ in a number of developmental respects from the euechinoid models, and because we know the paleontological history, we know that they represent the pleisiomorphic developmental program. In respect to how their genome could impact the issues surrounding decipherment of the genomic regulatory code, the key would be to obtain the GRNs for *Et* development orthologous to those we know for *Sp* . Knowing what linkages to look for in advance, the genome would allow a considerable fraction of this effort to be done computationally, thereby highlighting those parts of the GRN that are organized differently. Knowing the polarity of the differences, the comparison of GRN topologies would afford a priceless demonstration of pathways of GRN alteration that have functional consequences; and also of the modular differences in the antiquities and assembly histories of different elements of the Sp GRNs.

Dermasterias imbricate (490Mb)**:** *Di* is another easily available sea star, the importance of which for studies of genomic regulatory code is that its genome sequence would enable identification of *cis*-regulatory modules of the research model sea star *Am* by interspecific sequence comparison. This is a close to ideal method of finding *cis*-regulatory modules in echinoderms, as in many other forms, judging from our extensive experience in locating *Sp* regulatory modules by reference to *Lv* sequence. But at present we cannot apply this method to *Am*, a major impediment, as it is so far away from *Sp* that there is no sequence conservation in most *cis*-regulatory modules aside from individual target sites, and these are present in different arrangements. The *Di* sequence would cure this problem and greatly enhance regulatory GRN comparison between *Sp* and *Am* as above.

Ptychodera flava (~600Mb)**:** *Pf* is an indirectly developing hemichordate, which molecular phylogeny (21) shows to be basal to other hemichordate lineages such as that represented by *Saccoglossus*. We are aware that the genome of the latter is being sequenced, but unfortunately not only because of its highly derived character but also because *Saccoglossus* is a direct developer, its genome will not be very useful, if at all, for comparison of GRN topologies and subcircuits with *Sp*. The *Sp* GRNs underlie the development of the embryo/larva, and *Pf* produces an indirectly developing embryo/larva remarkably similar to that generated by echinoderms, including all those discussed in this document. As the reader will be aware, hemichordates and echinoderms are sister phyla sharing a common ancestor, and in basal forms, they share common embryonic process as well (22). Yet the adult body plans are bilateral in hemichordates and radial in modern echinoderms. However, Cambrian stem group echinoderms had a single left-right axis as do hemichordates and chordates (23). GRN subcircuits for embryonic development shared between *Sp* and *Pf* would represent modular regulatory apparatus of Precambrian origin and design. Also, ultimately GRN comparison between these organisms will reveal the circuitry underlying the radialization or angular multiplication of the left-right axis of the adult body plan, and its contrasts with the pleisiomorphic bilateral body plan circuitry that we have inherited.

Two general considerations should be mentioned. First, several of the organisms dealt with in this memo are laboratory workhorses, easily available, and if handled appropriately gametogenic at all times of the year; and the eggs of these species are easily amenable to gene transfer and gene expression knockout perturbation analysis (*Sp*, *Lv*, *Am*). While others are also easily available and have been used for laboratory research, some for decades, their gravid seasons are more confined (*Et*, *Pf*, *Ap*; and there is little experience with *Di*). However, for the specific purposes of these objectives the growing power of computational analysis of DNA sequence, once the GRNs of the *Sp* reference species are understood, much diminishes the importance of this problem. Second, with the establishment of SpBase, the public *Sp* genome database center at Caltech, we are in a position to ensure that the genome sequences here discussed will be maintained in a useful form.

Summary

We propose utilization of the newly available sequencing platforms to provide information that will lead directly to two enormously important objectives. These are first, development of means to derive GRN structure a priori, so that experimental requirements can be focused mainly on validation rather than GRN discovery de novo; and second, by rational comparison of GRN substructure, assessment of what aspects of GRN architecture are the most likely, and unlikely, targets for near future attempts at re-engineering GRNs. Additionally we note the very significant specific advantages for understanding the genomic regulatory code for development that would be obtained from four additional genome sequences, all building upon the benchmark *Sp* GRNs.

1. Oliveri, P., Tu, Q. and Davidson, E. H. Global regulatory logic for specification of an embryonic cell lineage. *Proc. Natl. Acad. Sci. USA* **105**, 5955-5962, 2008.
2. Smith, J., Theodoris, C. and Davidson, E. H. A gene regulatory network subcircuit drives a dynamic pattern of gene expression. *Science* **318**, 794-797, 2007.
3. Davidson, E. H. and Erwin, D. H. Gene regulatory networks and the evolution of animal body plans. *Science* **311**, 796-800, 2006.
4. Damle, S. and Davidson, E. H. *Trans*-specification of primary mesenchyme cells through genetic rewiring. In preparation, 2008.
5. Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B. E. And Jaenisch, R. *In vitro* reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* **448**, 318-324, 2007.
6. Davidson, E. H. The sea urchin genome: Where will it lead us? *Science* **314**, 939-940, 2006.
7. Sea Urchin Genome Sequencing Consortium, *et al.* The genome of the sea urchin *Strongylocentrotus purpuratus. Science* **314**, 941-952, 2006.
8. Hinman, V. F., Nguyen, A., Cameron, R. A. and Davidson, E. H. Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. *Proc. Natl. Acad. Sci. USA* **100**, 13356-13361, 2003.
9. Hinman, V. F. and Davidson, E. H. Evolutionary plasticity of developmental gene regulatory network architecture. *Proc. Natl. Acad. Sci. USA* **104**, 19404-19409, 2007.
10. Gao, F. and Davidson, E. H. Transfer of a large gene regulatory apparatus to a new developmental address in echinoid evolution. *Proc. Natl. Acad. Sci. USA* **105**, 6091-6096, 2008.
11. Rizzo, F., Fernandez-Serra, M., Squarzoni, P., Archimandritis, A., and Arnone, M., I. Identification and developmental expression of the ets gene family in the sea urchin (Strongylocentrotus purpuratus) *Dev. Biol.* **300**, 35–48, 2006.
12. Howard-Ashby, M., Materna, S. C., Brown, C. T., Chen, L., Cameron, A. and Davidson, E. H. Identification and characterization of homeobox transcription factor genes in *S. purpuratus*, and their expression in embryonic development. *Dev. Biol*. **300**, 74-89, 2006.
13. Howard-Ashby, M., Brown, C. T., Materna, S. C., Chen., L. and Davidson, E. H. Gene families encoding transcription factors expressed in early development of *Strongylocentrotus purpuratus. Dev. Biol.* **300,** 90-107, 2006.
14. Tu, Q., Brown, C. T., Davidson, E. H. and Oliveri, P. Sea urchin *forkhead* gene family: Phylogeny and embryonic expression. *Dev. Biol.* **300**, 49-62, 2006.
15. Materna, S. C., Howard-Ashby, M., Gray, R. F. and Davidson, E. H. The C_2H_2 zinc finger genes of *Strongylocentrotus purpuratus* and their expression in embryonic development. *Dev. Biol*. **300**, 108-120, 2006.
16. Su, Y.-H., Krämer, A., Li, E., Geiss. G. K., Longabaugh, W. J. R., and Davidson, E. H. A perturbation model of the gene regulatory network for oral and aboral ectoderm specification in the sea urchin embryo. In preparation, 2008
17. Brown, C. T., Rust, A. G., Clarke, P. J. C., Pan, Z., Schilstra, M. J., De Buysscher, T., Griffin, G., Wold, B. J., Cameron, R. A., Davidson, E. H. and Bolouri, H. New

computational approaches for analysis of *cis*-regulatory networks. *Dev. Biol.* **246**, 86-102, 2002.

18. Cameron, R. A., Chow, S. H., Berney, K., Chiu, T.-Y., Yuan, Q.-A., Krämer, A., Helguero, A., Ransick, A., Yun, M. and Davidson, E. H. An evolutionary constraint: Strongly disfavored class of change in DNA sequence during divergence of *cis*-regulatory modules. *Proc. Natl. Acad. Sci. USA* **102**, 11769-11774, 2005.
19. Yuh, C.-H., Brown, C. T., Livi, C. B., Rowen, L., Clarke, P. J. C. and Davidson, E. H. Patchy interspecific sequence similarities efficiently identify positive *cis*-regulatory elements in the sea urchin. *Dev. Biol.* **246**, 148-161, 2002.
20. Hinman, V. F., Nguyen, A. and Davidson, E. H. Caught in the evolutionary act: precise *cis*-regulatory basis of difference in organization of gene networks of sea stars and sea urchins. *Dev. Biol.* **312**, 584-595, 2007.
21. Swalla B. J.and Smith, A. B. Deciphering deuterostome phylogeny: molecular, morphological and palaeontological perspectives: *Phil. Trans. Royal Society B-Biol. Sci.* **363**, 1557-1568, 2008.
22. Peterson, K. J., Cameron, R. A. and Davidson, E. H. Set-aside cells in maximal indirect development: Evolutionary and developmental significance. *BioEssays* **19**, 623-631, 1997.
23. Bottjer, D. J., Davidson, E. H., Peterson, K. J. and Cameron, R. A. Paleogenomics of echinoderms. *Science* **314**, 956-960, 2006.