Developmental Biology 385 (2014) 160–167

Contents lists available at ScienceDirect

Developmental Biology

journal homepage: www.elsevier.com/locate/developmentalbiology

Resource

Quantitative developmental transcriptomes of the sea urchin *Strongylocentrotus purpuratus*

Qiang Tu, R. Andrew Cameron, Eric H. Davidson *

Division of Biology, California Institute of Technology, 1200 East California Boulevard, Pasadena, CA 91125, USA

ARTICLE INFO

Article history:
Received 15 October 2013
Received in revised form
13 November 2013
Accepted 18 November 2013
Available online 26 November 2013

Keywords:
Sea urchin
Developmental transcriptome
RNA sequencing

ABSTRACT

Development depends on the precise control of gene expression in time and space. A critical step towards understanding the global gene regulatory networks underlying development is to obtain comprehensive information on gene expression. In this study, we measured expression profiles for the entire expressed gene set during sea urchin embryonic development. We confirmed the reliability of these profiles by comparison with NanoString measurements for a subset of genes and with literature values. The data show that ~16,500 genes have been activated by the end of embryogenesis, and for half of them the transcript abundance changes more than 10-fold during development. From this genome scale expression survey, we show that complex patterns of expression by many genes underlie embryonic development, particularly during the early stages before gastrulation. An intuitive web application for data query and visualization is presented to facilitate use of this large dataset.

Introduction

Differential gene expression in time and space is the essence of the developmental process. With the availability of material on the scale of liters of eggs, sea urchin embryos, in particular those of *Strongylocentrotus purpuratus*, have long been used as an experimental model for study of gene expression. A rich earlier history of studies of cell specification in this species complements examinations of differential gene expression (Davidson, 1986). Current studies have led to one of the most complete gene regulatory networks (GRNs) for early development (Oliveri et al., 2008; Peter and Davidson, 2009; Peter et al., 2012). Genomically encoded *cis*-regulatory modules constitute the nodes of the GRN. Various inputs are integrated at these modules to control gene expression. The regulatory state of a particular embryonic cell is the sum of the various regulatory gene expressions in that cell. Subsequent regulatory states in descendant daughter cells follow through the activity of the regulatory state in the ancestor.

Hints of these regulatory processes had been revealed in mass culture experiments before the 1990s. First maternal RNA transcripts, then later zygotic ones coding for regulatory molecules, install the progressively more complex sets of regulatory states which accompany the increasing number of embryonic cells. The messenger RNA is stored in the unfertilized egg as a complex mixture of poly(A)- and non-poly(A)-RNA. It was estimated to amount to 30 pg per egg in mass and approximately 8500 message species of 2 kb length (Davidson, 1986). Transcription increases in the early divisions after fertilization, reaching a maximal rate by 4–5th cleavage (Wilt, 1970). Through nearly equivalent rates of synthesis and degradation, steady state mRNA content remains the same throughout embryonic development. From the kinetics of polysome assembly (Goustin and Wilt, 1981) and solution hybridization of genomic DNA with cytoplasmic RNA samples it is inferred that the loading of maternal RNA on polysomes to synthesize maternally encoded proteins is completed by approximately the 8-cell stage. Thereafter newly synthesized mRNAs are loaded.

Since these early mass measurements, our views on gene expression and developmental process have been refined through numerous measurements of single or small sets of genes in the embryo. New technologies make it possible to envision gene regulatory networks controlling the total of all genes expressed in development. Thus, a temporal profile of all of the genes expressed in the embryo would contribute an important piece of this puzzle. Using a largely complete gene set produced from deep sequencing of 22 stages and tissues of the purple sea urchin (Tu et al., 2012), we are now in a position to enumerate the temporal patterns of gene expression for the entire 21,000 genes at 10 time points in the embryonic development of this species. Estimates of temporal expression patterns are particularly informative during embryonic stages since the total mass of RNA remains approximately the same throughout this period.

Results and discussion

Quantification of developmental transcriptomes

We have analyzed transcriptomes throughout embryonic development of the *S. purpuratus* using the RNA-seq method. A set of gene

* Corresponding author.
E-mail addresses: jrigg@caltech.edu, davidson@caltech.edu (E.H. Davidson).

0012-1606/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ydbio.2013.11.019

models was generated from RNA-seq data in an earlier study (Tu et al., 2012). The transcript abundance was estimated for each gene at each of the 10 embryonic stages (Table S1) in FPKM (Fragments Per Kilobase of transcript per Million mapped reads) using Cufflinks (Trapnell et al., 2010). A set of internal standards comprising mRNAs present in known numbers of molecules was added to the sample (Mortazavi et al., 2008). The mRNA mass of each sequenced sample before adding the internal standards was 100 ng. Using the precise values for mRNA in the sea urchin embryo previously obtained, we calculated that this is equivalent to ~3300 embryos. Thus the relative FPKM values can be translated to the absolute abundance number in terms of transcripts per embryo. For example, an FPKM value of ~5 is equal to 300 transcripts per embryo.

We refined the measurement of transcript abundance in several ways. A key point is that we used the coding sequence (CDS) instead of the full-length gene models to calculate the number of reads mapped. As its name indicates, an FPKM value is the number of reads mapped to the given gene model normalized by the length of the model and by the total number of reads mapped to the genome. Since the full-length gene models were generated from 22 samples, including embryonic stages, larval stages, and adult tissues, they include the longest untranslated regions (UTRs). The UTRs of transcripts in individual samples might vary significantly. When the full-length gene models are used in the FPKM calculation, transcripts with shorter UTRs will still be normalized to the longest lengths, thus distorting the FPKM values. However, CDSs are the same across all samples. Thus we used CDS length for the abundance calculation, and we found that counting reads mapped only to the CDSs produces the most accurate FPKM values.

In order to confirm the reliability of the quantitation method, for 9 stages the abundance of a set of 173 regulatory genes in the same samples was measured using a NanoString, an instrument which counts mRNA directly without use of enzymatic reactions (Geiss et al., 2008). These NanoString measurements were used as a gold standard for validation of the values obtained by the RNA-seq method. The FPKM values derived from the RNA-seq data were compared with the NanoString counts of the control set of genes (Fig. 1A) and correlation coefficients for the 173 pairs of profiles were also calculated (Fig. 1B and C). The results from two methods matched very well. The median of the correlation coefficients is 0.923. A typical pair of expression profiles is shown in Fig. 1B, and all individual comparisons are shown in Fig. S1A.

General quantitative aspects of developmental transcriptomes

There are ~23,000 genes predicted to exist in the *S. purpuratus* genome (Sea Urchin Genome Sequencing Consortium et al., 2006) and the earlier transcriptome study compiled 21,092 gene models based on 22 transcriptomes (Tu et al., 2012). A cutoff of 300 transcripts per embryo (FPKM~5) was chosen as the lower limit for a functionally meaningful level of transcript representation. In an embryo at the mesenchyme blastula stage this criterion is equivalent to less than one transcript per cell for a ubiquitously expressed gene, or more for a cell-type-specifically expressed gene. Thus for example, even productively transcribed regulatory genes which produce low abundance mRNAs are expressed above this level, and technical variations in the quantification by the RNA-seq method are relatively low at this cutoff (Tu et al., 2012). Calculated in this way, ~16,500 genes (72% of the estimated total gene number) have been activated in at least one of the embryonic stages surveyed. These genes can be defined as an embryonic gene repertoire. They are shown as dots above the dashed ordinate cutoff line in Fig. 2A, as their highest transcript abundances during embryogenesis are >300. Similarly, embryonic housekeeping genes that are continuously expressed throughout embryogenesis can be identified if the lowest transcript abundance is over this cutoff (Fig. 2A, dots to the right of the dotted abscissa cutoff line). This class constitutes ~5700 genes (35% of the repertoire). A large-scale developmental transcriptome study of *Drosophila melanogaster* reported a similar percentage: 40% of expressed genes are constitutively expressed in all 27 stages studied, including embryonic, larval, and adult stages (Graveley et al., 2011). Considering the relative expression profiles, the transcript abundance levels of ~7900 genes (48% of the embryonic gene repertoire) undergo changes greater than 10-fold (Fig. 2A, red dots), while only ~1300 genes (8% of the embryonic gene repertoire) are expressed at relatively constant levels (less than 3-fold variation) (Fig. 2A, blue dots). These statistics reflect the highly dynamic use of the embryonic transcriptome, which is characteristic of the majority of genes used during this time.

At each stage of sea urchin embryogenesis, on average ~11,500 genes are actively transcribed, producing about 39 million transcripts per embryo. The histograms with linear and log scales show that the transcript abundance follows a log-normal distribution (Fig. 2B, C), and the cumulative histogram shows that 1% of genes contribute 39% of transcripts, 20% of genes contribute 85% of transcripts, and 50% of those genes that are considered as active genes contribute 97% of transcripts (Fig. 2D). From these plots, it is apparent that there is no definitive boundary between so called "active" and "non-active" genes; thus an empirical threshold activity (>300 transcripts per embryo) was used, as indicated above. However, defined in such a way the active genes produce the overwhelming majority of transcripts.

Complexity of an mRNA population is defined as the length of single copy DNA sequence represented in the RNA population.

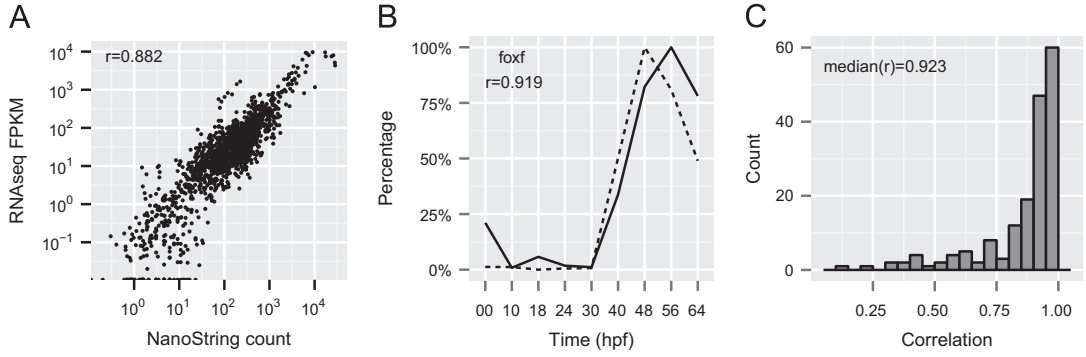

Fig. 1. Comparison between measurements by NanoString and RNA-seq. (A) Measurements of 173 genes in 9 samples of different embryonic developmental stages. (B) An example (FoxF) of time course profiles measured by NanoString and RNA-seq. The correlation coefficient between the two profiles is 0.919. Solid line: NanoString; dashed line: RNA-seq. (C) Distribution of correlation coefficients of the time course profiles measured by NanoString and RNAseq for each gene. The median value is 0.923. See also Fig. S1.