Application of a systems approach to study developmental gene regulation

Joshua W. K. Ho

Received: 14 May 2012 / Accepted: 21 June 2012 / Published online: 19 July 2012
© International Union for Pure and Applied Biophysics (IUPAB) and Springer 2012

Abstract All cells in a multicellular organism contain the same genome, yet different cell types express different sets of genes. Recent advances in high throughput genomic technologies have opened up new opportunities to understand the gene regulatory network in diverse cell types in a genome-wide manner. Here, I discuss recent advances in experimental and computational approaches for the study of gene regulation in embryonic development from a systems perspective. This review is written for computational biologists who have an interest in studying developmental gene regulation through integrative analysis of gene expression, chromatin landscape, and signaling pathways. I highlight the utility of publicly available data and tools, as well as some common analysis approaches.

Keywords Developmental signaling · Gene regulatory network · Computational biophysics · ChIP-seq · Histone modification · Epigenomics

Introduction

Developmental biology is one of the most fascinating fields in biology. It encapsulates the study of growth, differentiation, self-renewal, morphogenesis, and regeneration of cells, tissues, and organs. All these developmental processes require intricate spatial and temporal regulation of gene expression. A remarkable diversity of cell types in a multicellular organism share the same genome, yet they manage to express the right

set of genes at the right time. In other words, the information encoded in the DNA sequence must be retrieved and processed differently in different developmental contexts (Fig. 1a). There are many mechanisms in which developmental genes are regulated. In this short article, I focus on three classes of mechanism: (1) coordinated expression of transcription factors, (2) control of chromatin structure and properties, and (3) intercellular communication via signaling pathway modulation.

Transcription factors are often expressed in a cell-type- and stage-specific manner. They often serve as good lineage markers for determining cell type identity (Fig. 1a). In fact, the expression of lineage specific transcription factors is so specific that it can be used to define the *regulatory state* of a cell (Davidson 2006). Transcription factors bind to specific locations of the DNA, such as promoters and enhancers, in a sequence-dependent manner to promote or inhibit expression of nearby target genes. A cluster of transcription factor binding sites may form a *cis-regulatory module* to collectively regulate the expression of one or more nearby genes (Davidson 2006). Transcription factors may also recruit other regulatory proteins, such as chromatin remodelers and histone modifiers, to regulate local chromatin structure. Therefore, gene expression is determined by binding of these transcription factors and other proteins that interact with the chromatin.

Besides DNA, eukaryotic chromatin also contains a battery of DNA packaging proteins—*histone* proteins. Along the chromosome, about 147 bp of DNA wrap around an octamer of histone protein to form a *nucleosome*, and neighboring nucleosomes are connected by about 80 bp of linker DNA. Inside a nucleosome, each histone octamer consists of two copies of each core histone protein: H2A, H2B, H3, and H4. The occupancy of nucleosomes along a genome reflects the accessibility of the local chromatin region for binding of regulatory proteins. Nucleosomes around regulatory regions

Special issue: Computational Biophysics

J. W. K. Ho (✉)
Division of Genetics, Department of Medicine, Brigham and
Women's Hospital, Harvard Medical School,
Boston, MA 02115, USA
e-mail: jwho@rics.bwh.harvard.edu

Fig. 1 An overview of the mechanisms of developmental gene regulation. **a** Cell type identity can, in many cases, be determined by expression of cell-type-specific transcription factors (TFs). Upon receiving different signals (orange arrow), a progenitor or stem cell can either self-renew to produce more of the same cell type (green arrow), or differentiate into another cell type (red arrow). Cell-cell interactions can take place between different cell types through the exchange of signaling molecules or other mechanisms. **b** Gene regulation also occurs at the chromatin level through regulation of chromatin accessibility, histone modifications, and binding of TFs. Chromatin at different regulatory regions carries different histone modifications. **c** Transcription factors, the chromatin landscape, and signaling pathways act together to regulate gene expression during cell development

(e.g., promoters, enhancers, and gene bodies) typically incorporate histone variants or covalently modified histone proteins (Fig. 1b). For example, H3K4me3 (histone 3 lysine 4 trimethylation) is commonly present near a promoter, H3K4me1 marks enhancers, H3K36me3 is enriched in actively transcribing gene bodies, and H3K27me3 is highly enriched in repressed heterochromatin. *Chromatin landscape* (including DNA accessibility and histone modification) may be a result of binding of transcription factors or chromatin remodeling factors. However, the chromatin landscape may also define the activity of genes in the vicinity of the local chromatin regions, and may affect further transcription factor binding. One proposal is that this change in chromatin landscape creates an epigenetic memory to store information about the developmental stage and guide subsequent protein binding events.

A developing cell often does not act alone. A cell constantly receives signals from neighboring cells and tissues (Fig. 1c). Almost a century ago, Hans Spemann and Hilde Mangold proposed the concept of *embryonic induction* after observing that a small group of cells transplanted from one embryo to another can lead to body of axis duplication (Spemann and Mangold 2001). Induction can therefore be defined as a tissue interaction in which one tissue initiates the differentiation of another. These inductive events are highly coordinated, and are mediated by multiple developmental signaling pathways—such as the Wnt pathway,

the Bone Morphogenetic Protein (BMP) pathway, the Hedgehog pathway, and the Fibroblast Growth Factor (FGF) pathway. These pathways are used iteratively in the development of many organs. Cells communicate through expression and reception of secreted signaling molecules of these pathways. Secreted molecules are also called morphogens if they act in a concentration-dependent manner. For example, in vertebrate organogenesis, many organs are developed by the signaling interactions between juxtaposed epithelial and mesenchymal tissues. Epithelial–mesenchymal signaling interactions are characterized by the reciprocal expression and secretion of signaling molecules in many tissues (Grobstein 1967; Saxén and Thesleff 1992).

Over the last half a century, the study of developmental biology has mostly taken a molecular biology approach. The basic premise of the use of a molecular approach is that cellular and developmental phenotype can be directly explained by the action of one or more genes or biomolecules. This rationale of molecular biology leads to a strong emphasis to identify genes that are associated with a particular biological process (e.g., differentiation) or cell identity (e.g., progenitor cells). Once a candidate gene is identified, it is commonly characterized and perturbed to understand its association with the phenotype of interest. Downstream experiments may be undertaken to identify its binding partners and associated pathways. However, it is increasingly clear that there is a need to fully unravel the complex

mechanism underlying many developmental processes. Therefore, we must take a different approach—a *systems* approach.

Systems biology is the study of *complex* biological organization and processes, including most processes in developmental biology, in terms of the molecular constituents and the interactions among them (Kirschner 2005). A systems biology study is typically characterized by three hallmarks: (1) generation and analysis of genome-scale high throughput ‘omic’ data (i.e., systems-level data) to reveal the key components in the system, (2) emphasis on identifying emergent properties of a biomolecular system (i.e., systems-level properties) that are usually not apparent when only an individual molecular component is being studied, and (3) application of computational and mathematical techniques for hypothesis generation or data interpretation. It can be argued that the primary goal of systems biology is to investigate the collective behavior of many biomolecules simultaneously to reveal a variety of systems level properties, such as robustness and self-sustaining feedback (Kitano 2002). The recent surge of interest in systems biology can be attributed to advances in high-throughput experimental technologies, bioinformatics, growing recognition of interdisciplinary research, and the widespread availability of Internet resources for biomolecular data. With the increasing volume of high throughput data and open source bioinformatic tools, the role of systems biology is likely to become more prominent. Through the merger of advanced experimental biochemistry, genetics, bioinformatics, and biophysics, the scientific community is now capable of taking a systems approach to tackle some of ‘big picture’ questions that were intractable in the past, such as those in animal development (Davidson et al. 2003; Peter and Davidson 2011).

High throughput genomic technology

One enabling factor of modern systems biology is the availability of high-throughput genomic technologies; such as microarray and high-throughput ‘next-generation’ sequencing. The principles behind many genomic technologies are very similar: The targeted DNA (or reverse transcribed RNA) fragments are selected to enrich for a particular subset of DNA fragments of interest; these enriched DNA fragments are then identified by direct sequencing or microarray hybridization. For sequencing, the enriched DNA fragments are partially sequenced at one or both ends (usually up to 100 bp), and the resulting short sequence reads are mapped onto a reference genome to identify where the original fragments of interest are located in the genome. High-throughput sequencing can be coupled with other experimental preparations to generate genome-wide profiles of

different genomic features, such as RNA-seq for gene expression, ChIP-seq for transcription factor binding to DNA and histone modification, and DNase-seq for chromatin accessibility. Here, I review some of the technologies that are commonly used in the human ENCyclopedia of DNA Elements (ENCODE) consortium (The ENCODE Project Consortium 2011). For RNA-seq, expressed mRNA sequences are reverse transcribed, amplified, and sequenced (Marioni et al. 2008). The resulting sequence reads are then mapped onto the genome (and transcriptome) to identify genes that are expressed. Therefore, RNA-seq is a technology for gene expression measurement. ChIP-seq is a general method for detecting protein–DNA interactions in vivo (Barski et al. 2007). In a chromatin immunoprecipitation (ChIP) experiment, the DNA–protein complex is either sonicated or digested by MNase to obtain short (100–600 bp) DNA fragments. The target DNA fragments are immunoprecipitated by an antibody specific to the target protein of interest, such as a transcription factor, a modified histone protein, or a histone variant. As such, the immunoprecipitated DNA fragments are enriched for genomic DNA that are originally bound by that protein in vivo. Therefore, ChIP-seq can be used to obtain genome-wide profile of TF binding (e.g., binding of Oct4 in embryonic stem cells), histone modification (e.g., H3K27me3), and histone variant (e.g., H2A.Z) (Ho et al. 2012). DNase-seq is a method to identify DNase hypersensitivity sites (DHS) at a genome-wide scale (Boyle et al. 2008; Song and Crawford 2010). Open chromatin is much more likely to be sensitive to action of DNase I digestion, therefore DHS are typical of open chromatin. A similar technology of DNase-seq is called FAIRE-seq (Gaulton et al. 2010).

Advances in regulatory network modeling and analysis

A common task in systems biology is to uncover the gene regulatory network that controls various normal and abnormal cellular processes. A gene regulatory network is a useful tool for hypothesis generation, critical evaluation of alternative hypotheses, and summarization of knowledge of a system (Ho and Charleston 2011). In the context of studying developmental gene regulation, a properly generated network of transcription factors, signaling pathways, and important downstream target genes can be very helpful in elucidating the properties of complex developmental processes, such as robustness, self-sustaining feedback, and stochastic fate changes. In addition, a useful gene regulatory network should lead to new hypotheses and an experimental plan for validation. A network model itself is of little value if it does not have any predictive or explanatory power.

It is important to note that a gene regulatory network is an abstraction of a number of physical (direct) and functional

(indirect causal) regulatory interactions between genes and the proteins they code for. ChIP-seq is a powerful technology to discover an important class of physical interactions at a genome-wide scale: protein–DNA interactions between regulatory factors and DNA. These regulatory proteins can include master regulators of embryonic stem cells (Chen et al. 2008), lineage-specific factors (Nielsen et al. 2008), and general enhancer-associated proteins such as p300 (Visel et al. 2009). The genome-wide binding profiles of these transcription factors are usually highly specific to a tissue or a developmental stage. Therefore, comparative analysis of these profiles allows us to gain insight into dynamic changes in regulatory network during development.

Besides these genomic data, there are other datasets that can enrich a systems biology study by grouping related genes into modules and connecting genes and proteins with their interaction partners. Grouping genes according to their pathway or gene-set involvement is a good way to understand the coordinated regulation of these genes. For example, annotated gene sets can be obtained from KEGG (Kanehisa and Goto 2000) and gene ontology (Ashburner et al. 2000). A protein often exerts its action through interaction with one or more protein partners, so analysis of protein–protein interactions (PPIs) can help identify links between multiple proteins. High quality PPI networks can be used to connect proteins that may interact in vitro (Ravasi et al. 2010; Stelzl et al. 2005). Many validated or predicted PPI networks are stored in public databases, such

as STRING (von Mering et al. 2003) and BioGRID (Stark et al. 2011).

In systems developmental biology, we often overlook a very informative class of data: gene expression results from perturbation experiments that are reported in the literature. From my own experience, we can often extract a wealth of data related to causal gene regulation for a specific organ in the literature. For example, in a recent effort to recover a gene regulatory network in early tooth development, our group curated over 1,000 pieces of gene expression data by manually reading over 100 peer-reviewed research articles published since 1993 (O'Connell et al. 2012). These gene expression data (measured by in situ hybridization and qRT-PCR) were based mainly on two types of perturbation experiments: (1) treatment of explanted tissues with signaling molecules, or (2) knockout and knockdown of a gene in mouse models. Using a statistical model, we were able to integrate this large collection of low-throughput perturbation-based regulatory data with our own high-throughput experimental data to infer an intertissue gene regulatory network that controls early tooth development (Fig. 2). One may hope to use automatic text mining techniques to aid the task of literature text mining (Rodriguez-Esteban 2009), but their practical use in reliably extracting perturbation experimental results is currently limited (Rodriguez-Esteban et al. 2009). The task of manually extracting and curating gene expression results from the literature may seem daunting at first, but the potential use of the compiled dataset is immense.

Fig. 2 A systems approach to study developmental gene regulation. **a** Experimental data, literature-derived regulatory evidence, and prior background knowledge can be integrated computationally to construct a gene regulatory network model of a developmental system. **b** Using this approach, we have identified an intertissue Wnt-Bmp feedback

circuit that accounts for the epithelial-mesenchymal signaling dynamics during early tooth development. **c** The temporal expression patterns of Bmp4 and Wnt signaling ligands can be simulated by the feedback circuit model. This figure is adapted with permission from Fig. 2 of O'Connell et al. (2012)

In terms of regulatory network inference, an important conceptual advancement is the incorporation of physical interactions and causal regulatory data into network inference. Physical interaction data can be obtained from transcription-factor–DNA binding data and PPI. Causal gene regulatory relationships can be obtained from genetic or signaling perturbation experiments (Tegnér and Björkegren 2007). Many algorithms have been adapted to make use of these physical and causal interaction information jointly for inferring gene regulatory network. One proposal was to construct a *physical network model* to connect functional perturbation experimental results with a physical network of protein–protein and protein–DNA interactions (Yeang et al. 2004). In the initial study, the physical network was used to define all possible edges between nodes, and the functional perturbation data was used to select the active edges and assign edge direction using a probabilistic graphical model. This basic approach has been extended methodologically by using different optimization approaches. They are implemented into different algorithms, such as SPINE (Ourfali et al. 2007), ResponseNet (Lan et al. 2011; Yeager-Lotem et al. 2009), and physical module network (Novershtern et al. 2011).

Systems biology of developmental gene regulation

Understanding the gene regulatory network is essential to understanding the control of various complex processes in a developing cell. To translate the network structure into gene expression (or signaling pathway activity) dynamics, one can use ordinary differential equations or other network modeling techniques to simulate gene expression behavior under different parameter settings (Ho and Charleston 2011). When analyzing the structure of a network, one should pay attention to specific topological features, such as local clustering and feedback circuits. In particular, feedback circuits often play a central role in many developmental processes because they provide a mechanism to allow information feedback. One example that demonstrates the importance of feedback mechanism comes from our recent study of early tooth development (O’Connell et al. 2012). Our work was designed to study the mechanism of reciprocal signaling interactions between juxtaposed dental epithelium and dental mesenchyme during early tooth development. We reconstructed a network for the dental epithelium and a network for the dental mesenchyme separately, and then combined them through diffusible extracellular signaling molecules. Using this approach, we discovered an intertissue Wnt-Bmp signaling feedback circuit (Fig. 2b). Simulation of an ordinary differential equation model of this Wnt-Bmp circuit revealed that the observed reciprocal signaling dynamics between the two tissues can be explained by the circuit structure

itself (Fig. 2c). This is an example in which a large collection of individual ‘reductionist’ experimental data is used to gain insight into systems-level properties. Also, this work highlights the importance of studying signaling-mediated interactions between two cell types. Since such intertissue signaling interactions are commonplace, a similar signaling feedback mechanism may be a general feature of organogenesis. Therefore, the study of interaction between GRNs from neighboring cell types may potentially lead to many new discoveries in cell development. Similar systems biology approaches have been applied to study other developing tissues or organs, such as the vertebrate neural tube (Balaskas et al. 2012), limb bud (Bénazet et al. 2009), and hematopoietic cells (Kirouac et al. 2009, 2010).

An important premise of systems biology is that different biological molecules and cellular processes interact to exert a multitude of biological functions. In the context of developmental biology, signaling pathway connect to the core transcriptional regulatory network to affect the epigenome of a developing cell through interactions with the epigenetic machinery (Mohammad and Baylin 2010). Therefore, beyond transcription factor-based and signaling pathway-based regulatory network analysis, a current research direction is the study of epigenomic regulation of the chromatin landscape during cellular development (Reik 2007). Maintaining the proper chromatin structure is critical to control gene expression. The chromatin is classically categorized into euchromatin, for open accessible chromatin, and heterochromatin, for close inaccessible chromatin. In addition, the chromatin is marked by different histone modifications (Gaspar-Maia et al. 2011; Zhou et al. 2011). Different domains of chromatin mark changes dynamically during development. An example is that the domains of heterochromatin, which are often marked by histone modifications H3K27me3 or H3K9me3, are expanded in differentiated cells compared to pluripotent stem cells (Hawkins et al. 2010). In addition to studying individual marks, it is possible to define combinatorial ‘chromatin states’ based on multiple histone marks (Ernst et al. 2011; Kharchenko et al. 2011). For example, the promoters of developmental genes in embryonic stem cells are marked by both active mark H3K4me3 and repressive mark H3K27me3. These promoters are termed bivalent promoters (Bernstein et al. 2006). Genes that contain bivalent promoters are commonly developmental genes that are not expressed in embryonic stem cells, but are expressed in differentiated cells. A future research direction is to link cellular signaling pathways with the epigenetic machinery, such as histone acetyltransferase and histone deacetylases (Mohammad and Baylin 2010). The next generation of gene regulatory network analysis should incorporate information about transcription factor binding, cell–cell signaling interactions, and interactions with the epigenetic regulatory machinery.

Public data repositories and bioinformatic tools

An increasingly large amount of data are made publicly available through a variety of web-based databases (Ng et al. 2006; Armit 2007). Some commonly used databases are listed in Table 1. Most of these databases are maintained regularly to ensure high data quality. In particular, I would like to draw your attention to some very useful genomic resources from two large consortia: ENCODE and the model organism ENCODE (modENCODE) projects (Contrino et al. 2011; The ENCODE Project Consortium 2011). These consortia generate and disseminate a variety of genome-scale data for multiple cell lines, primary tissues, and embryonic stages in human (ENCODE), fly, and worm (modENCODE). In most cases, multiple data types are generated for each cell line or tissue type, such as gene expression, transcription factor binding, histone modification, DNase hypersensitivity, replication origin, and nucleosome positioning. Most of these data are generated either by high-throughput sequencing or microarrays. Besides these general genomic resources, there are many databases or web resources for individual organs or diseases. For example, we have recently published a database for mouse tooth development, ToothCODE (O'Connell et al. 2012), and a database for mouse ocular lens gene expression, iSyTE (Lachke et al. 2012).

One often needs to write custom computer scripts or programs to analyze genome-scale data. Before embarking on a software development project, it is important to keep in mind that many useful bioinformatic tools already exist for common analysis tasks. For example, there are numerous useful software packages in the Bioconductor tool and annotation collection (<http://www.bioconductor.org/>). Common gene-set

enrichment analysis tasks can be performed through DAVID (Dennis et al. 2003). Visualization and analysis of biological networks can be performed in Cytoscape (Shannon et al. 2003). Visualization of genomic datasets can be performed in various public genome browsers, including UCSC Genome Browser (Rhead et al. 2010) and IGV (Thorvaldsdóttir et al. 2012). For alignment of short sequence reads, one can use Bowtie (Langmead et al. 2009). There are also software packages for many common ChIP-seq and RNA-seq analysis tasks (Kharchenko et al. 2008; Trapnell et al. 2009; Zhang et al. 2008). For those who prefer a web-based graphical interface, many common functionalities are implemented in Galaxy (Taylor et al. 2007).

Before fully delving into these public data, here are a few cautionary notes. First, data must be properly normalized with appropriate controls. For example, when performing meta-analysis on multiple microarray gene expression data downloaded from GEO, it is important to check that the dynamic range of the expression values is comparable across multiple datasets. The profiles of the same technical replicates generated by two laboratories may be different due to batch effect. If necessary, these datasets need to be jointly normalized before pursuing any further analysis (Leek et al. 2010). Besides microarray data analysis, the use of appropriate background genomic control is important in ChIP-seq analysis. Failure to use an appropriate input DNA to normalize a ChIP-seq profile may reduce the accuracy of protein binding sites identification and reconstruction of enrichment profile around important regulatory sites (Ho et al. 2011). Another important point to bear in mind is that the results are often dependent on the cell type, developmental stage, treatment conditions, and genetic background of the samples. For example, a gene

Table 1 A list of public databases for systems developmental biology

Database	URL	Organisms
ENCODE	http://genome.ucsc.edu/ENCODE/	<i>H. sapiens</i>
modENCODE	http://www.modencode.org/	<i>D. melanogaster</i> , <i>C. elegans</i>
FANTOM	http://fantom.gsc.riken.jp/	<i>H. sapiens</i> , <i>M. musculus</i>
Human Epigenome Atlas	http://www.genboree.org/epigenomeatlas/	<i>H. sapiens</i>
TCGA	http://cancergenome.nih.gov/	<i>H. sapiens</i>
GEO	http://www.ncbi.nlm.nih.gov/geo/	Various
UCSC Genome Browser	http://genome.ucsc.edu/	Various
EmouseAtlas	http://www.emouseatlas.org/emap/home.html	<i>M. musculus</i>
GWAS Catalog	http://www.genome.gov/gwastudies/	<i>H. sapiens</i>
MGI	http://www.informatics.jax.org/	<i>M. musculus</i>
WormBase	http://www.wormbase.org	<i>C. elegans</i> and related species
FlyBase	http://flybase.org/	<i>D. melanogaster</i> and related species
miRBase	http://www.mirbase.org/	Various
Yeast Genome Database	http://www.yeastgenome.org/	<i>S. cerevisiae</i>

expression profile obtained from the whole pancreas can be quite different from the profile obtained from FAC-sorted islet beta cells. Recognition of cell-type specificity is a very important consideration when comparing different samples downloaded from a public database. In short, compilation and processing of raw datasets are often the most time-consuming steps in a computational biology study. It is therefore imperative to spend enough time to ensure that the data are properly pre-processed before proceeding to downstream analysis. Remember: garbage in, garbage out!

The roles of a computational biologist in the age of systems biology

Prior to the availability of large-scale genomic databases, a biologist had to directly obtain relevant data by experimentation. In particular, developmental biology is a field that has been primarily dominated by ‘wet lab’ experimental biologists. Nowadays, with access to these vast amounts of computational tools and genome-scale experimental data, many interesting biological questions can, at least in part, be investigated through careful integrative analysis of public datasets. Computational biologists now have an unprecedented opportunity to address many long-standing biological questions in developmental biology. How can a ‘dry lab’ computational biologist contribute to this emerging field of systems developmental biology? One essential attribute is our ability to build useful biological models using diverse experimental data. There is ample room for a computational biologist to develop new research projects in the field of systems developmental biology—either as an individual investigator or as a member of an interdisciplinary team. Here, I name three categories of opportunity: (1) integrative and comparative analysis of multiple published datasets, (2) collaboration with experimental biologists to generate and analyze new experimental data, and (3) development of computational tools and databases.

Key insights can often be obtained by integrative analysis of multiple related public datasets. This kind of integrative analysis may involve comparison of the gene expression of the same cell types in multiple mutants, or comparison of gene regulatory features in multiple species. As exemplified in earlier sections, this type of integrative meta-analysis of multiple datasets can generate new insights that otherwise cannot be obtained when the datasets are analyzed separately. From a systems perspective, it is possible to provide a more robust and multi-gene based definition of important cell types, such as stem cells (Müller et al. 2011). Furthermore, it should be recognized that fundamental

understanding of cellular developmental program is essential in tackling broader problems in human diseases and regenerative medicine. Many human congenital diseases are the result of one or more embryonic developmental defects. In other words, genes that are important in developmental gene regulation are also likely candidate disease-associated genes, such as human cataract genes (Lachke et al. 2012). Thus, a solid understanding of developmental gene expression and gene regulation also plays an important role in the study of human diseases.

A computational biologist can also be a key partner in an interdisciplinary collaboration team of experimental biologists. Besides performing data analysis, a computational biologist should play an active role in defining the research problem, designing the experiments, and providing critical insight into the interpretation of the data. Often, we can make use of existing public datasets to estimate whether a specific hypothesis would likely be supported or refuted before undertaking an experiment. This has implication in terms of designing experiments.

Lastly, a computational biologist can contribute new computational methods and tools for public use. With advances of genomic technologies, there is a constant need for fast and reliable computational tools to analyze genome-scale data. In particular, there is a need for analysis pipeline for high-throughput sequencing data, such as RNA-seq, ChIP-seq, and other emerging applications of high-throughput sequencing. Another opportunity is the development of new databases that store, categorize, and allow integrative analysis of multiple public datasets for the same cell types (e.g., stem cells), organ (e.g., tooth), and diseases (e.g., diabetes). When developing these tools and databases, one should also strive to verify and validate the correctness of the resulting software or database. Although this may be difficult given the resource limitation of a local laboratory, there are software testing strategies to assist this task (Chen et al. 2009; Xie et al. 2011).

It is an exciting time to be a computational biologist who studies developmental biology. The widespread application of genomic technologies is apt to give rise to many high quality datasets. Instead of directly interrogating the nature, we have the privilege to mine the data from public web-based repositories. As long as we are aware of the limitation of the biological data at hand, we can solve many interesting developmental biology questions with these data and tools from physics, mathematics, statistics, and computer science. This is the time to revisit some long-standing problems in developmental biology, such as the origin of self-organization, robustness, developmental autonomy, stochasticity in cell fate choices, and their relationships with human diseases.

Acknowledgment This work is partially supported by the National Institutes of Health (NIH) Common Fund through National Institute of Biomedical Imaging and Bioengineering Grant RL9EB008539.

Conflict of interest The author does not have any conflict of interest.

References

- Armit C (2007) Developmental biology and databases. *Organogenesis* 3:70–73
- Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:25–29. doi:10.1038/75556
- Balaskas N, Ribeiro A, Panovska J et al (2012) Gene regulatory logic for reading the sonic hedgehog signaling gradient in the vertebrate neural tube. *Cell* 148:273–284. doi:10.1016/j.cell.2011.10.047
- Barski A, Cuddapah S, Cui K et al (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129:823–837. doi:10.1016/j.cell.2007.05.009
- Bénazet J-D, Bischofberger M, Tiecke E et al (2009) A self-regulatory system of interlinked signaling feedback loops controls mouse limb patterning. *Science* 323:1050–1053. doi:10.1126/science.1168755
- Bernstein BE, Mikkelsen TS, Xie X et al (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* 125:315–326. doi:10.1016/j.cell.2006.02.041
- Boyle A, Davis S, Shulha H et al (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132:311–322. doi:10.1016/j.cell.2007.12.014
- Chen X, Xu H, Yuan P et al (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133:1106–1117. doi:10.1016/j.cell.2008.04.043
- Chen T, Ho JW, Liu H, Xie X (2009) An innovative approach for testing bioinformatics programs using metamorphic testing. *BMC Bioinforma* 10:24. doi:10.1186/1471-2105-10-24
- Contrino S, Smith RN, Butano D et al (2011) modMine: flexible access to modENCODE data. *Nucleic Acids Res.* doi:10.1093/nar/gkr921
- Davidson EH (2006) The regulatory genome: gene regulatory networks in development and evolution, 1st ed. Academic, New York
- Davidson EH, McClay DR, Hood L (2003) Regulatory gene networks and the properties of the developmental process. *Proc Natl Acad Sci USA* 100:1475–1480
- Dennis G Jr, Sherman BT, Hosack DA et al (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 4:P3
- Ernst J, Kheradpour P, Mikkelsen TS et al (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473:43–49. doi:10.1038/nature09906
- Gaspar-Maia A, Alajem A, Meshorer E, Ramalho-Santos M (2011) Open chromatin in pluripotency and reprogramming. *Nat Rev Mol Cell Biol* 12:36–47. doi:10.1038/nrm3036
- Gaulton KJ, Nammo T, Pasquali L et al (2010) A map of open chromatin in human pancreatic islets. *Nat Genet* 42:255–259. doi:10.1038/ng.530
- Grobstein C (1967) Mechanisms of organogenetic tissue interaction. *Natl Cancer Inst Monogr* 26:279–299
- Hawkins RD, Hon GC, Lee LK et al (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* 6:479–491. doi:10.1016/j.stem.2010.03.018
- Ho JWK, Charleston MA (2011) Network modelling of gene regulation. *Biophys Rev* 3:1–13. doi:10.1007/s12551-010-0041-4
- Ho JWK, Bishop E, Karchenko P et al (2011) ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics* 12:134. doi:10.1186/1471-2164-12-134
- Ho JWK, Alekseyenko AA, Kuroda MI, Park PJ (2012) Genome-wide mapping of protein-DNA interactions by ChIP-seq. Tag-based Approaches for Next-Generation Sequencing: Merging two High-throughput Technologies. Wiley, Weinheim, pp 139–152
- Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 28:27–30. doi:10.1093/nar/28.1.27
- Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 26:1351–1359. doi:10.1038/nbt.1508
- Kharchenko PV, Alekseyenko AA, Schwartz YB et al (2011) Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* 471:480–485. doi:10.1038/nature09725
- Kirouac DC, Madlambayan GJ, Yu M et al (2009) Cell-cell interaction networks regulate blood stem and progenitor cell fate. *Mol Syst Biol* 5:293. doi:10.1038/msb.2009.49
- Kirouac DC, Ito C, Csaszar E et al (2010) Dynamic interaction networks in a hierarchically organized tissue. *Mol Syst Biol.* doi:10.1038/msb.2010.71
- Kirschner MW (2005) The meaning of systems biology. *Cell* 121:503–504. doi:10.1016/j.cell.2005.05.005
- Kitano H (2002) Systems biology: a brief overview. *Science* 295:1662–1664. doi:10.1126/science.1069492
- Lachke SA, Ho JWK, Kryukov GV et al (2012) iSyTE: Integrated Systems Tool for Eye Gene Discovery. *Investig Ophthalmol Vis Sci* 53:1617–1627. doi:10.1167/iov.11-8839
- Lan A, Smoly IY, Rapaport G et al (2011) ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Res* 39:W424–W429. doi:10.1093/nar/gkr359
- Langmead B, Trapnell C, Pop M, Salzberg S (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25. doi:10.1186/gb-2009-10-3-r25
- Leek JT, Scharpf RB, Bravo HC et al (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 11:733–739. doi:10.1038/nrg2825
- Marioni JC, Mason CE, Mane SM et al (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18:1509–1517. doi:10.1101/gr.079558.108
- Mohammad HP, Baylin SB (2010) Linking cell signaling and the epigenetic machinery. *Nat Biotechnol* 28:1033–1038. doi:10.1038/nbt1010-1033
- Müller F-J, Schuldt BM, Williams R et al (2011) A bioinformatic assay for pluripotency in human cells. *Nat Methods* 8:315–317. doi:10.1038/nmeth.1580
- Ng A, Bursteinas B, Gao Q et al (2006) Resources for integrative systems biology: from data through databases to networks and dynamic system models. *Brief Bioinform* 7:318–330. doi:10.1093/bib/bbl036
- Nielsen R, Pedersen TÅ, Hagenbeek D et al (2008) Genome-wide profiling of PPAR γ :RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. *Genes Dev* 22:2953–2967. doi:10.1101/gad.501108
- Novershtern N, Regev A, Friedman N (2011) Physical module networks: an integrative approach for reconstructing transcription regulation. *Bioinformatics* 27:i177–i185. doi:10.1093/bioinformatics/btr222
- O’Connell DJ, Ho JWK, Mammoto T et al (2012) A Wnt-Bmp feedback circuit controls intertissue signaling dynamics in tooth organogenesis. *Sci Signal* 5:ra4. doi:10.1126/scisignal.2002414
- Ourfali O, Shlomi T, Ideker T et al (2007) SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics* 23:i359–i366. doi:10.1093/bioinformatics/btm170

- Peter IS, Davidson EH (2011) A gene regulatory network controlling the embryonic specification of endoderm. *Nature*. doi:10.1038/nature10100
- Ravasi T, Suzuki H, Cannistraci CV et al (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* 140:744–752. doi:10.1016/j.cell.2010.01.044
- Reik W (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* 447:425–432. doi:10.1038/nature05918
- Rhead B, Karolchik D, Kuhn RM et al (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res* 38:D613–D619. doi:10.1093/nar/gkp939
- Rodriguez-Esteban R (2009) Biomedical text mining and its applications. *PLoS Comput Biol* 5:e1000597. doi:10.1371/journal.pcbi.1000597
- Rodriguez-Esteban R, Roberts PM, Crawford ME (2009) Identifying and classifying biomedical perturbations in text. *Nucleic Acids Res* 37:771–777. doi:10.1093/nar/gkn986
- Saxén L, Thesleff I (1992) Epithelial-mesenchymal interactions in murine organogenesis. *CIBA Found Symp* 165:183–193, discussion 193–198
- Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504. doi:10.1101/gr.1239303
- Song L, Crawford GE (2010) DNase-Seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* 2010: pdb.prot5384. doi: 10.1101/pdb.prot5384
- Spemann H, Mangold H (2001) Induction of embryonic primordia by implantation of organizers from a different species. 1923. *Int J Dev Biol* 45:13–38
- Stark C, Breitkreutz B-J, Chatr-Aryamontri A et al (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* 39: D698–D704. doi:10.1093/nar/gkq1116
- Stelzl U, Worm U, Lalowski M et al (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122:957–968. doi:10.1016/j.cell.2005.08.029
- Taylor J, Schenck I, Blankenberg D, Nekrutenko A (2007) Using galaxy to perform large-scale interactive data analyses. *Curr Protoc Bioinforma Chapter 10:Unit 10.5*. doi: 10.1002/0471250953.bi1005s19
- Tegnér J, Björkegren J (2007) Perturbations to uncover gene networks. *Trends Genet* 23:34–41. doi:10.1016/j.tig.2006.11.003
- The ENCODE Consortium Project (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 9: e1001046. doi:10.1371/journal.pbio.1001046
- Thorvaldsdóttir H, Robinson JT, Mesirov JP (2012) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. doi:10.1093/bib/bbs017
- Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105–1111. doi:10.1093/bioinformatics/btp120
- Visel A, Blow MJ, Li Z et al (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457:854–858. doi:10.1038/nature07730
- von Mering C, Huynen M, Jaeggi D et al (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 31:258–261. doi:10.1093/nar/gkg034
- Xie X, Ho JWK, Murphy C et al (2011) Testing and validating machine learning classifiers by metamorphic testing. *J Syst Softw* 84:544–558. doi:10.1016/j.jss.2010.11.920
- Yeang C-H, Ideker T, Jaakkola T (2004) Physical network models. *J Comput Biol* 11:243–262. doi:10.1089/1066527041410382
- Yeger-Lotem E, Riva L, Su LJ et al (2009) Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. *Nat Genet* 41:316–323. doi:10.1038/ng.337
- Zhang Y, Liu T, Meyer C et al (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 9:R137. doi:10.1186/gb-2008-9-9-r137
- Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* 12:7–18. doi:10.1038/nrg2905