Review

Modeling genomic regulatory networks with big data

Hamid Bolouri

Division of Human Biology, Fred Hutchinson Cancer Research Center (FHCRC), 1100 Fairview Avenue North, PO Box 19024, Seattle, WA 98109, USA

High-throughput sequencing, large-scale data generation projects, and web-based cloud computing are changing how computational biology is performed, who performs it, and what biological insights it can deliver. I review here the latest developments in available data, methods, and software, focusing on the modeling and analysis of the gene regulatory interactions in cells. Three key findings are: (i) although sophisticated computational resources are increasingly available to bench biologists, tailored ongoing education is necessary to avoid the erroneous use of these resources. (ii) Current models of the regulation of gene expression are far too simplistic and need updating. (iii) Integrative computational analysis of large-scale datasets is becoming a fundamental component of molecular biology. I discuss current and near-term opportunities and challenges related to these three points.

Gene regulatory networks (GRNs)

The past few years have witnessed dramatic milestones in high-throughput sequencing, large-scale data generation, cloud computing, and computational biology. Supra-exponential improvements in the throughput and cost of DNA sequencing (http://www.genome.gov/sequencingcosts/) have been accompanied by improvements in accuracy and reductions in the required sample size. These improvements have in turn led to the widespread adoption of a broad range of sequencing-based technologies (reviewed in [1]) to characterize not only genomes but also the regulatory interactions that allow genomes to specify cellular structure, function, and behavior.

GRNs are defined as the set of interactions among genes and their products (RNAs and proteins) that determine the isoforms, location (cell type), timing, and rate of RNA expression [2] (see Figure 1 for examples). With the possible exception of some metabolic and physiological processes, GRNs are the primary drivers of cellular behavior and function.

Because GRNs are ultimately specified by the digital code of DNA, they are uniquely accessible to both high-throughput sequencing-based technologies and to computational modeling and analysis. At the same time, GRNs are both complex (i.e., can exhibit hard-to-predict/nonlinear behaviors) and complicated (i.e., they are composed of large numbers of component parts and interactions). For this reason, mathematical and computational approaches are essential in GRN research.

Cellular behaviors have traditionally been characterized as being mediated through highly distinct processes (e.g., DNA replication) and pathways (e.g., the canonical WNT signaling pathway). However, because of widespread interactions among cellular processes and pathways, the use of unbiased, genome-wide technologies is essential to the discovery and characterization of GRNs.

In addition to the bedrock of 'classical' *cis*-regulatory analysis, GRN modeling today is buttressed by four cornerstones: (i) high-throughput technologies, (ii) integrative analysis of complementary data types, (iii) leveraging large-scale public datasets, (iv) computational modeling and analysis. This article reviews recent developments and discusses their implications for future research.

To maintain coherence and brevity, this review will focus on developments in human GRN modeling and analysis. Diverse new GRN modeling opportunities are also opening up in both well-studied and less-studied organisms. These and the complex GRNs underlying interactions between hosts and commensal or pathogenic organisms are beyond the scope of the present review.

Types and uses of human GRN modeling

A model is any representation of a system that can facilitate its analysis, communication, or documentation [3]. Modeling is at the heart of GRN research at multiple levels. At the most basic level, statistical models are at the heart of all high-throughput data analysis. For example, statistical models are commonly used to characterize DNA fragment length distribution as a first step towards the identification of transcription factor (TF) binding peaks in ChIP-seq (chromatin immunoprecipitation followed by high-throughput DNA sequencing) data.

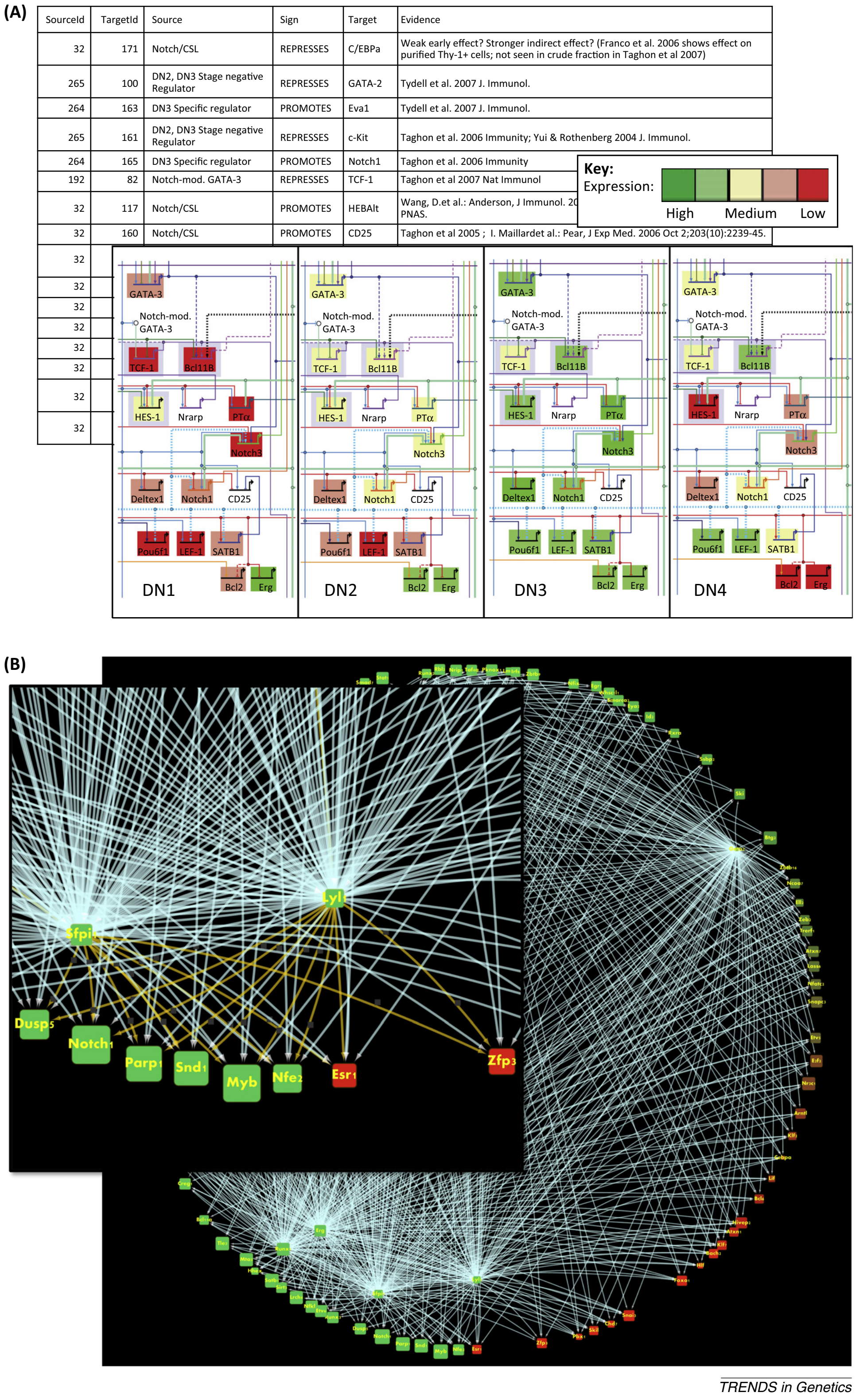
Given filtered data, methods such as network inference [4], guilt-by-association (e.g., through network or expression clustering; see Figures 2 and 3), and enrichment/overrepresentation analysis (e.g., to identify the impacted pathways or processes [5]) are used to organize genes and their products into broad-brush conceptual models. These models can then be refined and extended by integrating multiple data types each highlighting a different

Corresponding author: Bolouri, H. (HBolouri@fhcrc.org).

Keywords: gene regulatory networks; modeling; network biology; big data; computational biology; bioinformatics; systems biology.

0168-9525/$ – see front matter

http://dx.doi.org/10.1016/j.tig.2014.02.005



(A)

SourceId	TargetId	Source	Sign	Target	Evidence
32	171	Notch/CSL	REPRESSES	C/EBPa	Weak early effect? Stronger indirect effect? (Franco et al. 2006 shows effect on purified Thy-1+ cells; not seen in crude fraction in Taghon et al 2007)
265	100	DN2, DN3 Stage negative Regulator	REPRESSES	GATA-2	Tydell et al. 2007 J. Immunol.
264	163	DN3 Specific regulator	PROMOTES	Eva1	Tydell et al. 2007 J. Immunol.
265	161	DN2, DN3 Stage negative Regulator	REPRESSES	c-Kit	Taghon et al. 2006 Immunity; Yui & Rothenberg 2004 J. Immunol.
264	165	DN3 Specific regulator	PROMOTES	Notch1	Taghon et al. 2006 Immunity
192	82	Notch-mod. GATA-3	REPRESSES	TCF-1	Taghon et al 2007 Nat Immunol
32	117	Notch/CSL	PROMOTES	HEBAlt	Wang, D.et al.: Anderson, J Immunol. 20 PNAS.
32	160	Notch/CSL	PROMOTES	CD25	Taghon et al 2005 ; I. Maillardet al.: Pear, J Exp Med. 2006 Oct 2;203(10):2239-45.
32					
32					
32					
32					
32					
32					
32					
32					

Key:
Expression: 🟩 🟩 🟨 🟫 🟥
High Medium Low

TRENDS in Genetics

Figure 1. Examples of gene regulatory network analysis, documentation, and visualization. **(A)** Part of BioTapestry visualization of a proposed early T cell specification gene regulatory network (GRN) (adapted from: http://www.its.caltech.edu/~tcellgrn/Oldnetwork.html). Each gene (symbol with a bent arrow) is represented as having a regulatory region (horizontal line) and a transcriptional output (arrow). A transcription factor (TF)–DNA binding interaction is depicted as an arrow incident on the regulatory region of a gene. Protein–protein interactions are depicted by circles with incident and output arrows. The background color of each gene indicates the fold-change in expression of the gene at a particular developmental stage. Snapshots of the network over four developmental stages are shown [double negative (DN) 1 to 4]. In the interactive viewer, clicking on a gene brings up a table showing the experimental data supporting the indicated regulatory interactions. **(B)** Cytoscape visualization of potential T cell specification gene regulatory interactions derived from ChIP-seq and gene expression data. Arrows represent regulatory interactions. Node colors and sizes represent gene expression levels at early and late developmental stages. The inset shows a zoomed-in view of the lower portion of the network. Using Cytoscape utilities, the user can quickly and easily identify a set of genes coregulated by Sfpi1 and Lyl1 (edge arrows highlighted in gold). This example network was derived during a 1.5 h introductory laboratory session by novice computational biology students (see http://www.bu.edu/computationalimmunology/summer-school/ for details).

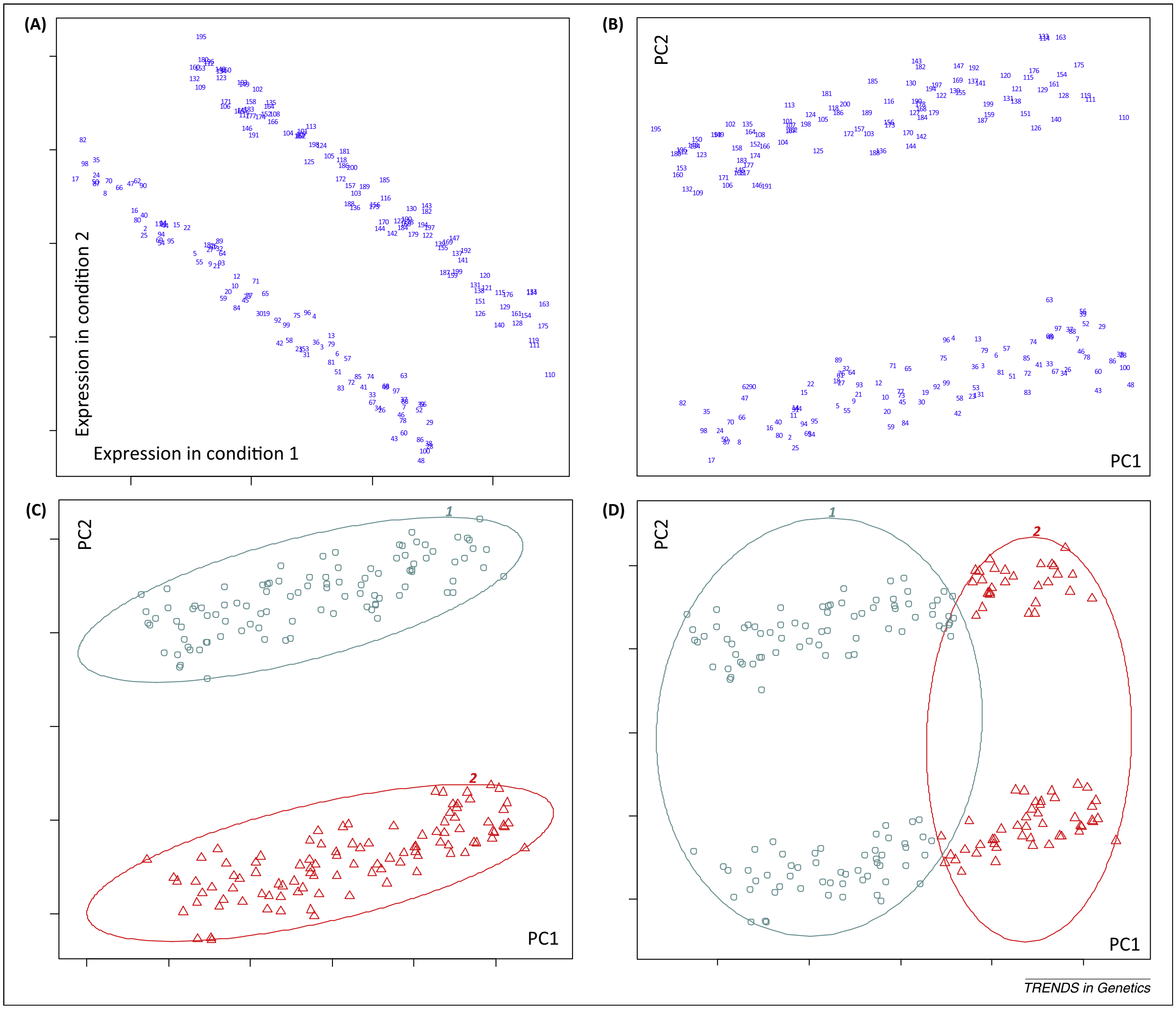


Figure 2. An example of how choices in data processing impact the findings. **(A)** Scatter plot of example simulated data showing the expression of 200 genes in two conditions. This simulated dataset was intentionally generated to have two distinct characteristics. First, genes near the lower-right (upper-left) corner of the plot are high (low) in condition 1 and low (high) in condition 2. In addition, there are two well-separated groups of genes distinguished by having generally higher (lower) expression in both conditions. **(B)** The same data plotted in the space of the principal components (PC) (in this case, because there were only two conditions, there are only two PCs). Note that the two PCs optimally distinguish the two characteristics of the data. **(C,D)** Unsupervised hierarchical clustering identifies different features of the data depending on how distances between clusters are measured (clusters marked by oblongs and labeled 1 and 2). In 'single-linkage' clustering (C), the distance between two clusters is defined as the distance between their nearest elements. In 'complete-linkage' clustering (D), the distance between two clusters is defined as the distance between their farthest elements.

aspect of the system {e.g., mRNA and microRNA (miRNA) expression data from RNA-seq (whole transcriptome shotgun RNA sequencing), TF DNA-binding data from ChIP-seq, and protein interaction data from mass spectrometry [6] or yeast two-hybrid assays [7]}.

Depending on the available data, GRN models may represent snapshots of the state of the GRN in different cells (for examples, see [8] and Figure 1), they may capture the dynamic trajectory of GRN state changes in a particular group of cells over time (e.g., [9]), or they may quantitatively predict the kinetics of expression of individual genes (e.g., [10]).

Importantly, all of the above GRN model types can be used to predict the impact of genetic, epigenetic, and environmental perturbations on cells (see below), or to predict potential drug targets (e.g., [11]). Drug target discovery based on GRN models was recently reviewed comprehensively in [12] and therefore will not be discussed further here.

Using GRN models to address disease

In the application of GRN modeling to diseases it is helpful to consider disease-causing perturbations as impacting a GRN in two distinct ways: (1) changes in the interactions of the component genes (e.g., [13]) impact the 'view from the genome' (i.e., the set of all regulatory interactions encoded in the genome [14]). (2) Changes in the cell type(s) in which a gene is expressed, the time of onset and duration of

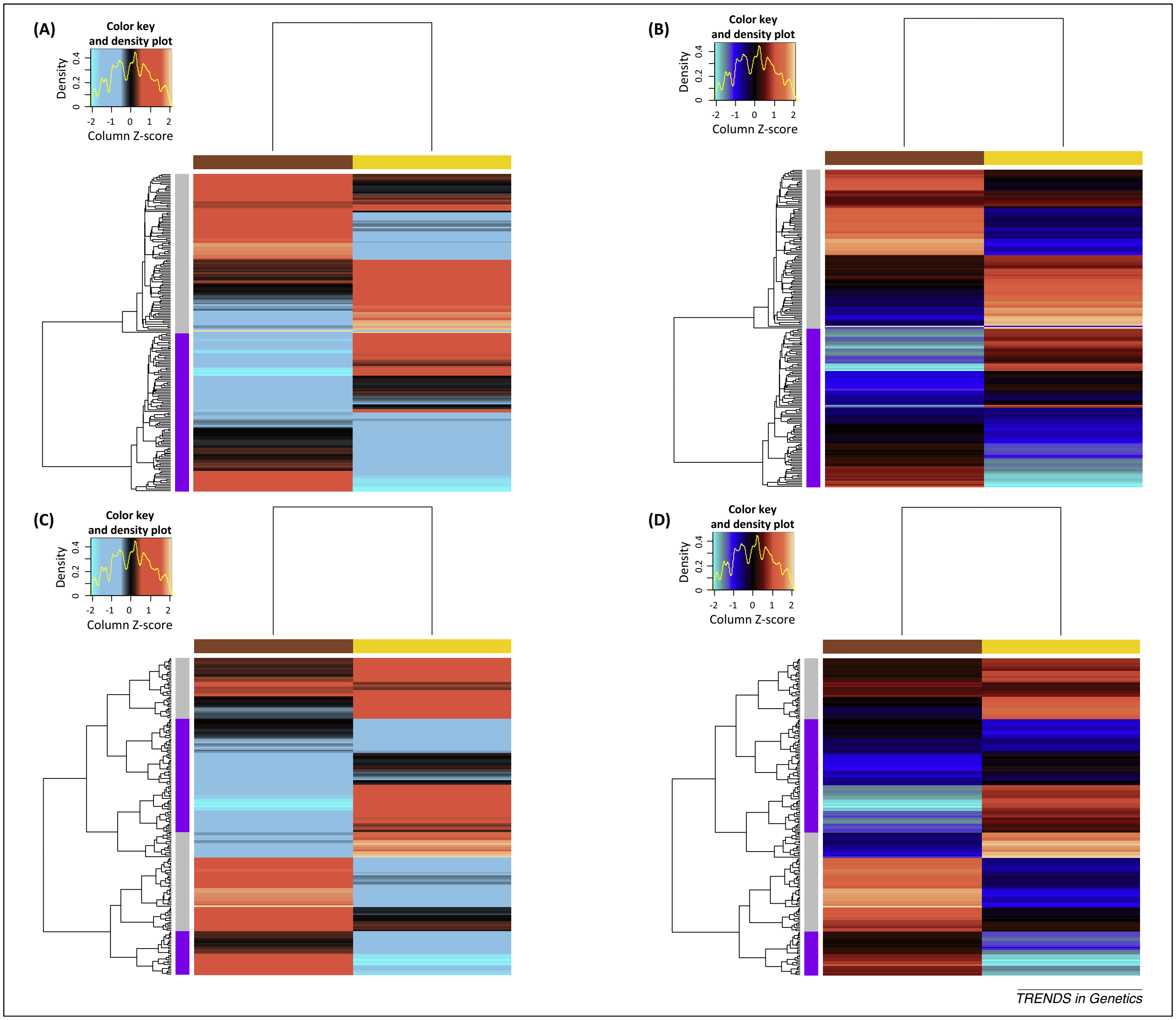


Figure 3. Example of the impact of choices in data visualization. **(A,B)** Heatmaps of the same data as in Figure 2 clustered as in Figure 2C. **(C,D)** Heatmaps of the same data as in Figure 2 clustered as in Figure 2D. Each vertical column (marked by the gold and brown bar) represents data from one condition. Each row represents the expression levels of one gene across the two conditions (note identical row dendrograms in A,B and C,D). The purple and gray side-bar colors mark the two well-separated groups of genes visible in Figure 2A. The only difference between (A) and (C) – or between (B) and (D) – is the color scale (as indicated at the top-left of each plot). Note how (A) and (C) give the impression of more homogeneous gene expression clusters. This example highlights the need to include dendrogram and color-scale information in heatmap plots, something that is often overlooked by inexpert users.

transcriptional activity (e.g., brief versus long-lasting), or the magnitude of gene expression (e.g., [10]).

The impact of the latter group of perturbations (e.g., activating or loss of function mutations) can often be predicted by appropriately modifying GRN models for healthy cells. To predict the effects of group 1 perturbations (i.e., those affecting interactions) we need to know all changes in interactions. For example, a DNA-binding fusion oncoprotein such as TMPRSS2:ERG (transmembrane protease, serine 2, fused to the ETS-related gene ERG) or RUNX1:ETO [Runt-related transcription factor 1, fused to the eight-twenty-one (ETO) nuclear corepressor] may bind to new DNA loci and activate or repress genes that neither of the fusion partners regulate individually [15]. Alternatively, a gene product may lose a subset of its interactions, for example when a particular DNA binding site is mutated [16]. Because functional protein–DNA interactions are often the result of complex multiprotein interplay, computational prediction of lost and gained interactions is currently challenging, but promising new experimental approaches have recently been demonstrated [17,18].

Another way in which GRN models of healthy cells can be used to address human disease is through the identification of candidate disease-causing genes. Candidate disease-causing gene lists can be produced by a wide variety of approaches, including genome-wide association studies (GWAS), DNA sequencing, expression profiling, and RNA interference (RNAi) and synthetic-lethality screens. GRN structure analysis can help rank candidate genes. For example, a recent study [19] suggests that, in both protein interaction and gene regulatory networks, essential genes tend to be more central and highly connected. Thus, a simple approach to ranking candidate genes would be to prioritize genes that are more highly connected and more central in GRN models. For a review of candidate gene prioritization tools using interaction networks and guilt-by-association see [20] and the associated Gene Prioritization Portal (http://homes.esat.kuleuven.be/~bioiuser/gpp/).

An additional approach to discovering candidate disease-causing genes is 'genetical genomics', which typically generates multiple candidate quantitative trait loci (QTLs) identifying associations between gene expression and aberrant splicing, chromatin state, or TF binding in a given cell type (reviewed in [21]). In this context, GRN models of disease-related pathways and processes can be used to rank variants and target genes by disease relevance and potential to be causal (see e.g., [22,23]).

The cumulative efforts of many disease-focused projects have also led to the development of generally useful GRN modeling resources. For example, many cancer cell lines are also used to study cellular pathways and processes not specific to cancer. The Cancer Cell Line Encyclopedia (CCLE, http://www.broadinstitute.org/ccle) includes data on mutations in ~1600 genes, and genome-wide data on gene expression and copy-number variations in 947 human cell lines. Integration of these data with existing interaction and pathway databases will aid the development of cell line-specific GRN models and further our understanding of the impact of DNA sequence variants on GRN structure, function, and behavior.

Several dedicated online resources (e.g., https://genome-cancer.ucsc.edu and http://www.cbioportal.org) allow interactive mining and exploration of large-scale datasets from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) project. Of special interest here is the platform-independent, menu-driven geWorkbench (genomics workbench) of TCGA (http://wiki.c2b2.columbia.edu/workbench), which offers a large collection of general-purpose 'plug-in' modules including the popular ARACNE (algorithm for the reconstruction of accurate cellular networks) GRN inference algorithm [24], and the MARINa master regulator TF detection algorithm [25].

The impact of high-throughput technologies

ChIP-seq, RNA-seq, and miRNA-seq are by now well-established (for a review of ChIP-seq and related emerging technologies, see [26]). The October 2012 data release from the ENCODE project (Encyclopedia of DNA Elements; http://encodeproject.org, see http://genome.ucsc.edu/encode/pubs.html for publications) spans 4060 experiments using 33 experimental approaches. Likewise, as of October 2013, the National Institutes of Health (NIH) Roadmap Epigenomics project (http://www.roadmapepigenomics.org/) has performed 3176 experiments in 61 tissue types.

Together, the ENCODE and Epigenomics projects provide a unique resource of multiple cell lines in which gene expression, chromatin regulatory state (via DNA methylation assays and histone-modification ChIP-seq), chromatin accessibility (via DNase I-seq and FAIRE-seq, see below), RNA polymerase II (Pol2) activity state, and the binding patterns for multiple TFs have been characterized. As discussed below, these datasets enable the prediction of cell type-specific gene regulatory interactions and greatly facilitate GRN modeling in other cell types by revealing the 'view from the genome' GRN in humans.

In addition to antibodies for Pol2, phosphorylated (transcriptionally active) Pol2, Pol3, and 17 histone modification states, the ENCODE project currently lists antibodies for 221 protein isoforms (https://genome.ucsc.edu/encode/antibodies.html, accessed October 2013), suggesting that ChIP-seq is currently feasible for only about 10% of the estimated 1500–3000 [27,28] human TFs (the 2012 ENCODE release of uniformly processed TF ChIP-seq peaks, encompassing 690 experiments, 161 distinct factors, and 91 cell types, can be accessed at http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeAwgTfbsUniform).

In contrast to ChIP-seq, DNase I-seq and FAIRE–seq (formaldehyde-assisted identification of regulatory elements followed by tag sequencing) experiments offer unbiased genome-wide identification of (nucleosome-depleted, open-chromatin) regulatory regions [29]. Moreover, very high coverage DNase I/FAIRE-seq can be used to delineate at nucleotide resolution the binding footprints of individual TFs [30].

Because DNase I hypersensitivity (HS) regions are short and well-defined, searching them for occurrences of known TF binding motifs can yield high-confidence binding-site predictions. As part of the ENCODE effort, researchers [31] searched for *in vitro* determined TF binding motifs in DNase I HS regions within 5 kb of the transcription start-sites of 475 TFs in 41 cell types. The GRNs inferred in this way recapitulate many known cell-specific interactions, confirming the potential of this approach.

Searching within DNase I HS and footprint regions has allowed high-confidence predictions of TF binding within these regions. However, motif searches typically cannot distinguish between members of TF families. This uncertainty is alleviated in ChIP-seq data where the most enriched motifs should correspond to the immunoprecipitated factor and its co-factors. Accordingly, all ENCODE ChIP-seq binding peaks have been searched for matches to known TF binding motifs and also for statistically overrepresented novel motifs [32], and these data are publicly available for mining (http://compbio.mit.edu/encode-motifs/).

In parallel efforts, TF ChIP-seq peaks have been demonstrated to cluster spatially in the genome [33,34] and these cluster regions were shown to correspond closely with regulatory regions marked by DNase I and histone modification assays [35]. Clustered TF binding regions may distinguish functionally active regulatory DNA binding from inactive, random DNA binding.

Correlations between histone modification patterns and active/poised enhancers and promoters are well established (for a review of epigenetic regulation, see [36]). A recent study combined the latest large-scale data releases and two previously demonstrated machine-learning methods to identify candidate regulatory regions genome-wide [37]. These and other annotations resulting from integrative analysis of the ENCODE data are now available as browser tracks from the University of California, Santa Cruz – UCSC (http://genome.ucsc.edu/encode/analysis.

html) and ENSEMBL (http://ensembl.org/info/website/tutorials/encode.html) genome browsers.

To date, GRN modeling and analysis efforts have tended to focus on RNA and gene-level data. Recent technological improvements are driving the development of more nuanced GRN models. In particular, high-depth RNA sequencing is increasingly providing isoform abundance information [38], thus allowing GRN models to include alternative splicing events. At the same time, phosphoproteomics is maturing to allow the integration of post-translational modifications into GRN models [39].

Associating distal enhancers with specific genes has been challenging so far. Many previous studies have used the binding sites of the boundary element/insulator protein CTCF (CCCTC-binding factor) to impose bounds on the possible targets of a candidate enhancer region. However, CTCF is a multifunctional protein with complex regulatory roles (reviewed in [40]). ChIA-PET (chromatin interaction analysis using paired-end tag sequencing) and genome-wide high-resolution chromatin conformation capture techniques such as HiC now provide improved methods for mapping DNA–DNA interactions (reviewed in [41]).

The studies reviewed above deliver snapshots of GRN states under specified conditions. Recent developments enabling large-scale measurements of the production and decay rates of mRNA and protein levels enable modeling of GRN dynamics. Specifically, dynamic transcriptome analysis (DTA) uses metabolic RNA labeling to measure the production and decay rate of RNAs [42], whereas 'ribosome profiling' measures translational efficiency by deep sequencing of ribosome-protected mRNA fragments [43,44].

The implications of big data

As discussed above, high-throughput data are driving the development of more detailed, more mechanistic, and more predictive GRN models. In addition to the CCLE, ENCODE, TCGA, and Roadmap Epigenomics projects discussed above, many national and international projects are now generating large volumes of sequencing-based data. Examples include the National Cancer Institute (NCI) TARGET project (Therapeutically Applicable Research to Generate Effective Treatments; http://ocg.cancer.gov/programs/target), the 1000 Genomes project (http://www.1000genomes.org/), the NIH Gene-Tissue Expression program (http://commonfund.nih.gov/GTEx/), the International Cancer Genome (http://dcc.icgc.org/) and Epigenome (http://ihec-epigenomes.org/) Consortia, and many more.

There is a pressing need for careful filtering and interpretation of these data (see examples in Figure 4). Efforts such as the NIH Big Data to Knowledge (BD2K, http://bd2k.nih.gov/) initiative are needed to ensure that large datasets can be easily accessed, understood, searched, cross-referenced, combined, and interpreted. Such efforts will ultimately make big data directly available to bench biologists for mining. In the meantime, there are exciting opportunities for computational biologists to develop cell type- and disease-specific GRN models by filtering, comparing, and integrating data from multiple sources and experiments. For example, the Bioconductor AnnotationHub (http://www.bioconductor.org/packages/2.13/bioc/html/AnnotationHub.html) currently makes 5324 human datasets readily accessible to users of the R programming language. Genome-wide GRN models built from such data can be used to inform the construction of more focused small-scale GRN models of specific cellular processes and pathways.

To enable reproducible research, data, script, and workflow sharing facilities offered by dedicated tools such as Synapse (https://www.synapse.org) and by multifunctional software suites such as GenePattern and Galaxy (discussed below) can be used to make the models derived from big data publicly available along with their associated data, scripts and workflows. Such sharing will aid rapid and widespread uptake, ensure reproducibility, and catalyze the development of future GRN models.

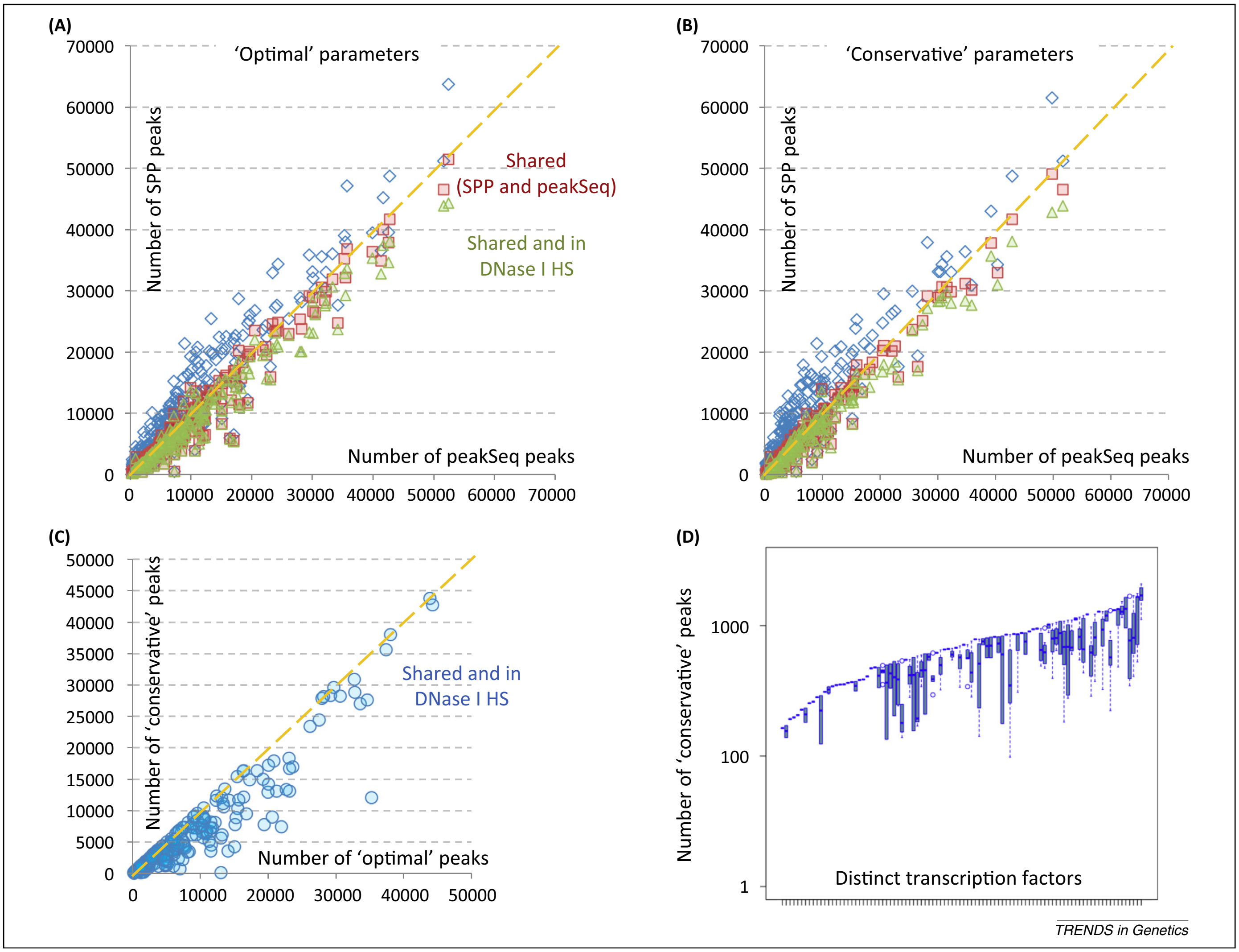
Developments and trends in computational tools

GRN models can span from genetic interaction maps to physical interaction graphs to models of network dynamics and gene expression kinetics. A very broad range of software tools have been developed to address GRN modeling and analysis at all these levels. For brevity, here I summarize only recent trends and cite some examples.

In integrative model-building, multiple types of experimental data and computational predictions are combined to infer gene regulatory interactions from large-scale data [45]. A large number of databases provide protein–protein and protein–DNA interaction information extracted from the literature and curated to varying degrees. GeneMania (http://genemania.org/) integrates pathway and interaction data with coexpression, colocalization, and gene ontology information from a multitude of sources. A confidence weight is associated with each reported interaction, making GeneMania a convenient and useful starting point for GRN modeling (for two other useful starting points, see http://www.ihop-net.org/UniPub/iHOP and http://string-db.org/).

Many new and established tools provide specialized data analysis resources – typically via platform-independent downloadable software or through web browsers. Such tools include over-representation/enrichment analysis servers such as GeneTrail (http://genetrail.bioinf.uni-sb.de/), TargetMine (http://targetmine.nibio.go.jp/), and David (http://david.abcc.ncifcrf.gov/); Gene-E for expression clustering (http://www.broadinstitute.org/cancer/software/GENE-E); the MATISSE (module analysis via topology of interactions and similarity sets)–DEGAS (dysregulated gene set analysis via subnetworks) suite for the identification of GRN modules and pathways from expression data (http://acgt.cs.tau.ac.il/matisse/); and Genomica (http://genomica.weizmann.ac.il/), which identifies functionally related gene groups and enrichment patterns in expression data.

GRN models have several distinctive features not shared by signaling and other network models. Changes in gene expression patterns typically occur on a much slower timescale than protein–protein interactions and post-translational modifications. Viewed from the perspective of GRNs, many signaling and other processes can be viewed as simple switching 'events'. This realization greatly



TRENDS in Genetics

Figure 4. Challenges in interpreting large-scale datasets. Shown are data relating to 254 transcription factor ChIP-seq datasets uniformly processed by two peak-calling methods (SPP: http://compbio.med.harvard.edu/Supplements/ChIP-seq/ and peakSeq: http://compbio.med.harvard.edu/Supplements/ChIP-seq/) by the ENCODE Project (see http://tinyurl.com/ENCODE-SPP-Peaks and http://tinyurl.com/ENCODE-PeakSeqPeaks). **(A,B)** Scatter plots of the numbers of peaks called by PeakSeq (*x* axis) and SPP (*y* axis) using optimized (A) and conservative (B) peak-calling parameters. Counts for all called peaks are shown as blue diamonds. Counts of overlapping SPP and PeakSeq peaks are shown as red squares, and counts of shared peaks that also fall in ENCODE DNase I HS regions are shown as green triangles. Note that in all cases, the calls from the two methods are highly concordant. **(C)** Comparison of conservatively and optimally called peaks after both have been filtered by the requirement to be shared across the two peak-calling methods and to overlap DNase I HS regions. The existence of a few outliers suggests that a fraction of optimally called peaks in a subset of experiments may be false-positives, a feature not apparent from the plots in (A) and (B). **(D)** The number of conservatively called peaks per TF varies by two orders of magnitude. For each TF, a boxplot shows the range in the number of peaks across different cell types and experimental conditions. Analysis of such characteristics would require highly reliable peak-calling. Abbreviations: ChIP-seq, chromatin immunoprecipitation followed by high-throughput DNA sequencing; ENCODE, Encyclopedia of DNA Elements; HS, hypersensitivity; SPP, sequence processing pipeline; TF, transcription factor.

simplifies representations of GRNs. By contrast, many features specific to GRNs need to be captured and documented in GRN models [46]. Examples include chromatin state regulation, the spatial organization of TF binding sites on DNA, interactions among regulatory factors and among regulatory regions, alternative transcription start-sites, alternative splicing, and the regulatory logic that determines whether a gene is expressed given a set of factors bound at specific sites.

BioTapestry ([47], http://www.biotapestry.org/, see Figure 1A for an example BioTapestry model) is a platform-independent menu-driven drawing, visualization, and documentation tool designed specifically for GRN modeling. It offers a range of resources that directly address many of the above needs.

The past few years have seen a maturing of software suites that aim to provide a comprehensive analysis toolset. Prominent examples of such efforts are Bioconductor (http://bioconductor.org), Cytoscape (http://cytoscape.org/), Galaxy (http://galaxyproject.org/, [48]), GenePattern (http://www.broadinstitute.org/cancer/software/genepattern), and GenomeSpace (http://www.genomespace.org/).

Each of these resource hubs has its own distinctive style and focus. For example, Bioconductor 2.13 provides 749 open-source packages written in the R language. Bioconductor resources are primarily focused on data interpretation and aimed at bioinformaticians with script-writing skills (see http://www.rstudio.com/shiny/ for a very promising web-based menu-driven interface to R scripts).

In contrast to Bioconductor, Cytoscape provides a menu-driven user interface through the platform-independent Java language, and is primarily focused on integrative modeling and analysis of interaction networks. Like the Bioconductor project, which accepts and curates third-party R packages, Cytoscape enables third-party researchers to contribute to its resources through 'Apps' (Java plugins that perform specific functions within Cytoscape).

A notable feature of Galaxy is that it combines a web-based, menu-driven user interface with the ability for users and providers to save and share work-flows. These features, combined with Galaxy's ability to integrate third-party tools (including R/Bioconductor packages) empower bench biologists to run complex data-processing pipelines without the need for script writing. GenePattern has capabilities and resources very similar to Galaxy. It currently has a larger toolset, additional data visualization capabilities, and a more visually-rich user interface.

GenomeSpace integrates other resource integrators, including Cytoscape, Galaxy, Genepattern, and Genomica (see below). Notably, users of GenomeSpace can easily transfer/re-use data and analysis results among these software suites.

Finally, as noted earlier, large-scale measurements of mRNA and protein production and decay rates are becoming feasible, allowing quantitative modeling of transcriptional dynamics (reviewed in [49]). A large number of tools are available for simulation modeling and analysis of dynamical system properties – such as intrinsic noisiness, controllability (e.g., for cellular re-programming), and robustness to environmental and genetic perturbations (see http://sbml.org for a list of over 250 such tools).

Opportunities and challenges ahead

The tools reviewed above all require the user either to install the software locally or to upload data to a remote server. Raw and aligned read data from high-throughput sequencing experiments are typically gigabytes in size, making data transfer to and from remote servers cumbersome. At the same time, multiple central processing units (CPUs) and large amounts of memory are necessary for data analysis. Installation, configuration, and maintenance of software on such platforms often require system-administration skills not available in most individual research laboratories.

At present, initial data processing (e.g., read alignment, ChIP-seq peak calling, RNA-seq differential gene expression analysis) is often performed by the facility generating the data, and researchers often focus on downstream analysis of the summary tables generated by the facility. Although this approach may be adequate for routine experiments, many – perhaps most – discovery-research experiments are better served by iterative reprocessing of the raw data based on initial findings. For example, if peaks from two different ChIP-seq experiments are found to overlap to a surprising degree, one may wish to search the neighborhoods of apparently non-overlapping peaks to see if they were in fact flanked by peaks that were below the calling threshold in the first round of analysis.

These considerations increasingly point to a need to colocate high-throughput data with the computational tools to analyze the data [50]. In support of this trend, web-based cloud-computing bioinformatics platforms such as GenePattern, GenomeSpace, and Galaxy in the public domain, as well as DNAnexus (https://www.dnanexus.com/) in the commercial domain, already provide the necessary infrastructure.

Low-cost high-throughput technologies and cloud-based integrative databases and tools are making GRN modeling and analyses more accessible to bench biologists. The enormous and growing range of methods provided by the Bioconductor project can now be tapped through Cytoscape, Galaxy, and GenePattern, bringing intuitive, menu-driven workflows and sophisticated computational resources to bench biologists. This is an inevitable and very welcome trend, but it also brings challenges and pitfalls.

With only a few mouse clicks, a user can now perform highly complex statistical, mathematical, and algorithmic operations and create sophisticated data visualizations. But users who fail to understand the theoretical principles underlying a point-and-click computational tool risk using inappropriate methods, incorrect algorithmic parameters, and improperly pre-processed input data. To illustrate this point, some example scenarios are presented and discussed in Figures 2–4. These considerations highlight a pressing need for widely accessible ongoing training in computational biology for bench biologists at all levels.

Our understanding of how TFs regulate gene expression (reviewed in [51]) remains far from complete. With the explosion of new data in the past few years there is a need and an opportunity to develop new models of transcriptional regulation that better explain the effects of engineered as well as disease-causing perturbations. For example, an assumption underlying virtually all qualitative and quantitative GRN models is that the concentrations of regulatory TFs determine the rate of transcription (e.g., [3,52]). Remarkably, recent findings suggest that the duration of TF dwell time on DNA – rather than its average occupancy – can determine target gene expression [53]. Similar arguments can be made regarding recent discoveries of widespread polymerase pausing [54], the regulation of enhancer–promoter interactions [40,55,56], and so on.

To date, most GRN modeling efforts have focused on TF–TF and TF–gene interactions. The availability of large-scale data and high-throughput technologies is now enabling far more nuanced hierarchical and modular GRN models incorporating cross-regulation of gene expression with histone modifications, DNA methylation, miRNA expression, RNA splicing, and so on. These added dimensions will in turn permit GRN models better to predict cellular differentiation, re-programming, aging, and disease.

More nuanced models will necessitate a shift away from sharply defined GRN modules and pathways (which are primarily aids to human comprehension) towards hierarchical networks composed of recurring 'functional building blocks' [57] each consisting of no more than a handful of components and performing a distinct regulatory function.

Characterizations of protein localization, abundance, post-translational modifications, and *in vivo* interaction dynamics have so far been challenging. Recent breakthroughs (reviewed in [58]) and a better understanding of the many modes of translational regulation (reviewed in [59]) offer the exciting possibility of 'completing the circle' in GRN models from DNA to RNA to protein and back to DNA.

Acknowledgments

I thank Bill Longabaugh (Institute for Systems Biology) and Paul Shannon (Bioconductor team, FHCRC) for many helpful comments on an early draft of this review.

References

1 Soon, W.W. *et al.* (2013) High-throughput sequencing for biology and medicine. *Mol. Syst. Biol.* 9, 640

2 Davidson, E.H. (2006) *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*, Academic

3 Bolouri, H. (2008) *Computational Modeling of Gene Regulatory Networks: A Primer*, Imperial College Press

4 Maetschke, S.R. *et al.* (2013) Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Brief. Bioinform.* http://dx.doi.org/10.1093/bib/bbt034

5 Geistlinger, L. *et al.* (2011) From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics* 27, i366–i373

6 Mazloom, A.R. *et al.* (2011) Recovering protein–protein and domain–domain interactions from aggregation of IP-MS proteomics of coregulator complexes. *PLoS Comput. Biol.* 7, e1002319

7 Baker, M. (2012) Proteomics: the interaction map. *Nature* 484, 271–275

8 Yosef, N. *et al.* (2013) Dynamic regulatory network controlling TH17 cell differentiation. *Nature* 496, 461–468

9 Calzone, L. *et al.* (2010) Mathematical modelling of cell-fate decision in response to death receptor engagement. *PLoS Comput. Biol.* 6, e1000702

10 Purvis, J.E. *et al.* (2012) p53 dynamics control cell fate. *Science* 336, 1440–1444

11 Chen, E.Y. *et al.* (2012) Expression2Kinases: mRNA profiling linked to multiple upstream regulatory layers. *Bioinformatics* 28, 105–111

12 Csermely, P. *et al.* (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol. Ther.* 138, 333–408

13 Bandyopadhyay, S. *et al.* (2010) Rewiring of genetic networks in response to DNA damage. *Science* 330, 1385–1389

14 Bolouri, H. and Davidson, E.H. (2002) Modeling transcriptional regulatory networks. *Bioessays* 24, 1118–1129

15 Okumura, A.J. *et al.* (2008) t(8;21)(q22;q22) Fusion proteins preferentially bind to duplicated AML1/RUNX1 DNA-binding sequences to differentially regulate gene expression. *Blood* 112, 1392–1401

16 Wang, S. *et al.* (2013) An enhancer element harboring variants associated with systemic lupus erythematosus engages the TNFAIP3 promoter to influence A20 expression. *PLoS Genet.* 9, e1003750

17 Patwardhan, R.P. *et al.* (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* 30, 265–270

18 Akhtar-Zaidi, B. *et al.* (2012) Epigenomic enhancer profiling defines a signature of colon cancer. *Science* 336, 736–739

19 Khurana, E. *et al.* (2013) Interpretation of genomic variants using a unified biological network approach. *PLoS Comput. Biol.* 9, e1002886

20 Moreau, Y. and Tranchevent, L.C. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat. Rev. Genet.* 13, 523–536

21 Gaffney, D.J. (2013) Global properties and functional complexity of human gene regulatory variation. *PLoS Genet.* 9, e1003501

22 Huang, J. *et al.* (2011) eResponseNet: a package prioritizing candidate disease genes through cellular pathways. *Bioinformatics* 27, 2319–2320

23 Erten, S. *et al.* (2011) DADA: Degree-aware algorithms for network-based disease gene prioritization. *BioData Min.* 4, 19

24 Margolin, A.A. *et al.* (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7 (Suppl. 1), S7

25 Lefebvre, C. *et al.* (2010) A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol. Syst. Biol.* 6, 377

26 Furey, T.S. (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nat. Rev. Genet.* 13, 840–852

27 Vaquerizas, J.M. *et al.* (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* 10, 252–263

28 Fulton, D.L. *et al.* (2009) TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.* 10, R29

29 Song, L. *et al.* (2011) Open chromatin defined by DNase I and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* 21, 1757–1767

30 Neph, S. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489, 83–90

31 Neph, S. *et al.* (2012) Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150, 1274–1286

32 Kheradpour, P. and Kellis, M. (2013) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* http://dx.doi.org/10.1093/nar/gkt1249

33 Yan, J. *et al.* (2013) Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* 154, 801–813

34 Chen, X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 133, 1106–1117

35 Bolouri, H. and Ruzzo, W.L. (2012) Integration of 198 ChIP-seq datasets reveals human cis-regulatory regions. *J. Comput. Biol.* 19, 989–997

36 Rivera, C.M. and Ren, B. (2013) Mapping human epigenomes. *Cell* 155, 39–55

37 Hoffman, M.M. *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* 41, 827–841

38 Zhao, K. *et al.* (2013) GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome Biol.* 14, R74

39 Huang, S.S. *et al.* (2013) Linking proteomic and transcriptional data through the interactome and epigenome reveals a map of oncogene-induced signaling. *PLoS Comput. Biol.* 9, e1002887

40 Merkenschlager, M. and Odom, D.T. (2013) CTCF and cohesin: linking gene regulatory elements with their targets. *Cell* 152, 1285–1297

41 Smallwood, A. and Ren, B. (2013) Genome organization and long-range regulation of gene expression by enhancers. *Curr. Opin. Cell Biol.* 25, 387–394

42 Rabani, M. *et al.* (2011) Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* 29, 436–442

43 Brar, G.A. *et al.* (2012) High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* 335, 552–557

44 Ingolia, N.T. *et al.* (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* 7, 1534–1550

45 Bebek, G. *et al.* (2012) Network biology methods integrating biological data for translational science. *Brief. Bioinform.* 13, 446–459

46 Longabaugh, W.J. *et al.* (2009) Visualization, documentation, analysis, and communication of large-scale gene regulatory networks. *Biochim. Biophys. Acta* 1789, 363–374

47 Longabaugh, W.J. (2012) BioTapestry: a tool to visualize the dynamic properties of gene regulatory networks. *Methods Mol. Biol.* 786, 359–394

48 Goecks, J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 11, R86

49 Coulon, A. *et al.* (2013) Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nat. Rev. Genet.* 14, 572–584

50 Stein, L.D. (2010) The case for cloud computing in genome informatics. *Genome Biol.* 11, 207

51 Spitz, F. and Furlong, E.E. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* 13, 613–626

52 Bolouri, H. and Davidson, E.H. (2003) Transcriptional regulatory cascades in development: initial rates, not steady state, determine network kinetics. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9371–9376

53 Lickwar, C.R. *et al.* (2012) Genome-wide protein–DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature* 484, 251–255

54 Adelman, K. and Lis, J.T. (2012) Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* 13, 720–731

55 Li, W. *et al.* (2013) Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* 498, 516–520

56 Chu, C. *et al.* (2011) Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell* 44, 667–678

57 Longabaugh, W. and Bolouri, H. (2006) Understanding the dynamic behavior of genetic regulatory networks by functional decomposition. *Curr. Genomics* 7, 333–341

58 Altelaar, A.F. *et al.* (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.* 14, 35–48

59 Kong, J. and Lasko, P. (2012) Translational control in cellular and developmental processes. *Nat. Rev. Genet.* 13, 383–394