Review in Advance first posted online on September 3, 2015. (Changes may still occur before final publication online and in print.)

# Comparative Analysis of Gene Regulatory Networks: From Network Reconstruction to Evolution

Dawn Thompson,[1] Aviv Regev,[1,2] and Sushmita Roy[3,4]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142

[2]Howard Hughes Medical Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02140

[3]Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, Wisconsin 53715; email: sroy@biostat.wisc.edu

[4]Wisconsin Institute for Discovery, University of Wisconsin–Madison, Madison, Wisconsin 53715

Annu. Rev. Cell Dev. Biol. 2015. 31:6.1–6.30

The *Annual Review of Cell and Developmental Biology* is online at cellbio.annualreviews.org

This article's doi:
10.1146/annurev-cellbio-100913-012908



## Keywords

gene regulatory networks, comparative functional genomics, evolution, *cis*-regulatory elements, coexpression modules, comparative sequencing, comparative expression, network analysis

## Abstract

Regulation of gene expression is central to many biological processes. Although reconstruction of regulatory circuits from genomic data alone is therefore desirable, this remains a major computational challenge. Comparative approaches that examine the conservation and divergence of circuits and their components across strains and species can help reconstruct circuits as well as provide insights into the evolution of gene regulatory processes and their adaptive contribution. In recent years, advances in genomic and computational tools have led to a wealth of methods for such analysis at the sequence, expression, pathway, module, and entire network level. Here, we review computational methods developed to study transcriptional regulatory networks using comparative genomics, from sequences to functional data. We highlight how these methods use evolutionary conservation and divergence to reliably detect regulatory components as well as estimate the extent and rate of divergence. Finally, we discuss the promise and open challenges in linking regulatory divergence to phenotypic divergence and adaptation.

## Contents



## INTRODUCTION

Gene expression is a key factor shaping organism phenotype (Jacob & Monod 1961) and is controlled by intricate, integrated regulatory circuits spanning multiple molecular levels. These levels include the regulation of transcription through transcription factor (TF) binding, the interaction of chromatin remodelers or noncoding RNAs with regulatory DNA, higher order chromosome organization, post-transcriptional control of RNA levels through RNA transport, processing, modification, sequestration, and degradation, and control of protein translation, stability, and activity, especially through signal transduction networks and post-translational modifications such as phosphorylation and acetylation. At each layer, we distinguish between regulators—e.g., TFs and chromatin remodelers at the transcriptional layer—and targets—e.g., the gene promoters to which a TF binds—such that gene expression levels are determined through a regulatory network connecting regulators to targets at each level. Dissecting such regulatory networks is one of the fundamental challenges of systems biology.

Time and again, changes in gene regulation and expression have been postulated to be major forces in adaptation and evolution (Emerson & Li 2010, Gasch et al. 2004, Jordan et al. 2005, King & Wilson 1975, Romero et al. 2012, Thompson & Regev 2009, Tirosh & Barkai 2011, Weirauch & Hughes 2010, Wohlbach et al. 2009, Wray 2007, Zheng et al. 2011). Numerous studies have implicated regulatory variation in generating the phenotypic diversity of both unicellular and multicellular organisms, both within populations and between species. Regulatory divergence between species has been studied across many kingdoms of life, including bacteria (McAdams et al. 2004), fungi (Gasch et al. 2004, Thompson et al. 2013, Tirosh & Barkai 2011, Tirosh et al. 2006, Wohlbach et al. 2011), flies (Bradley et al. 2010, Jiménez-Guri et al. 2013, Kalinka et al. 2010, Paris et al. 2013, Prud'homme et al. 2007, Wittkopp et al. 2008), worms (Silver et al. 2012, Yanai & Hunter 2009), plants (Ichihashi et al. 2014), fish (Brawand et al. 2014, Chan et al. 2010, Jones et al. 2012), and mammals (Barbosa-Morais et al. 2012, Brawand et al. 2011, Khaitovich et al. 2006, Necsulea et al. 2014, Merkin et al. 2012). Similarly, regulatory mechanisms and gene expression

**Network:** collection of edges specifying the regulators associated with all genes

divergence has been studied between individuals of the same species, including in yeast (Brem 2005, Kvitek et al. 2008), fish (Oleksiak et al. 2002), flies (Denver et al. 2005), worms (Grishkevich et al. 2012), plants (Lasky et al. 2014, West et al. 2007), and humans (Battle et al. 2013, Bryois et al. 2014, Li et al. 2014, Stranger et al. 2007). Across species, several phylogenetic studies have demonstrated the role of regulatory divergence in phenotypic variation, although many regulatory changes may be neutral (Jordan et al. 2005, Khaitovich et al. 2004, Lynch 2007, Yanai et al. 2004). Genetic variants associated with many complex human diseases and phenotypes often map to noncoding regulatory regions (Maurano et al. 2012), and studies are beginning to show how this variation is mechanistically associated with gene expression changes and physiological effects related to the etiology of human disease (Emilsson et al. 2008, Farh et al. 2015, Lee et al. 2014, Ye et al. 2014).

Comparative studies of gene regulatory networks within and between species can thus serve two important goals. First, leveraging the genetics implicit in the differences and similarities between species or individuals can help dissect the mechanisms that govern changes in expression (Wittkopp 2007), thus aiding in the reconstruction of a regulatory network. Second, comparing gene regulation mechanisms and expression levels between organisms or individuals can help shed light on the driving forces and mechanisms underlying adaptation and evolution (Necsulea & Kaessmann 2014, Romero et al. 2012, Wohlbach et al. 2009, Wray 2007).

Achieving these goals requires two interrelated efforts. One is to collect relevant data on gene expression and regulation mechanisms across species or individuals; the other is to develop computational methods that use these data to infer gene regulatory networks, compare networks across species, and infer evolutionary histories. In recent years, tremendous technological advances have been made in the accurate measurement of gene expression and regulation at all levels. In parallel, advances in computational methodologies specifically geared to examining regulatory profiles and networks across multiple genomes have been instrumental in gleaning insights into regulatory variation within and between species. Such research is setting the stage to understand the impact of regulatory divergence on the overall phenotype of organisms and to provide an integrated view of how different functional levels contribute to downstream gene expression. Notably, although gene regulation acts at all levels, from transcription to protein synthesis, nearly all studies have focused on messenger RNA (mRNA) levels and transcriptional mechanisms, as data for these two aspects remain the most readily available for systematic genomic experiments. As a result, most computational methods have been developed with these data in mind. Both regulatory networks (De Smet & Marchal 2010) and comparative analysis of gene expression (Necsulea & Kaessmann 2014, Romero et al. 2012) have been the topics of recent reviews. In addition, several reviews by us and others describe experimental approaches to studying the evolution of gene regulation and regulatory networks (Li & Johnson 2010, Thompson & Regev 2009, Tirosh & Barkai 2011, Wohlbach et al. 2009).

In this review, we focus on computational approaches that aim to systematically infer transcriptional gene regulatory networks from comparative data and to study the evolution of these networks. We briefly introduce transcriptional regulatory networks and then describe computational approaches for detecting changes in network components and assessing the rate of change. We examine these approaches from three perspectives, which cover nearly all studies in this field: (*a*) transcriptional regulation through TF binding to specific *cis*-regulatory sequences, (*b*) the transcriptional state as measured by mRNA levels and chromatin state, and (*c*) a systems-level view of entirely inferred networks. We conclude with an outlook on challenges to understanding the contribution of regulatory divergence to phenotypic divergence, which requires integrating multiple levels of regulation, including at the translational, pre-transcriptional, and post-translational levels.

## MAJOR STRATEGIES FOR MODELING TRANSCRIPTIONAL REGULATORY NETWORKS

**Node:** regulatory proteins such as transcription factors and target genes

**Edge:** a connection between a regulatory protein and a target gene, specifying that the protein regulates the gene

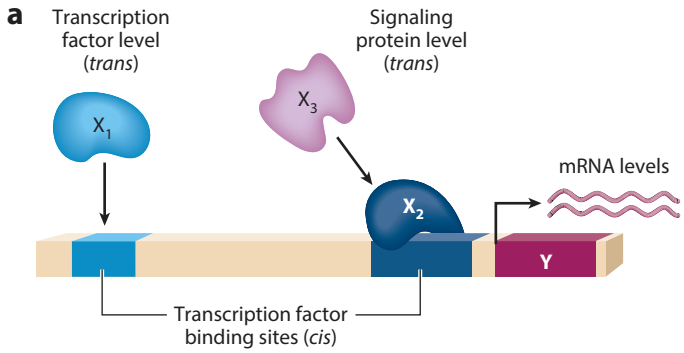
A transcriptional regulatory network specifies which *trans* regulators (e.g., proteins and noncoding RNAs) and *cis* sequence elements control the expression of target genes, in which contexts, and in what way. In many studies, this information is abstracted as a graph in which nodes represent regulators (e.g., TFs, signaling proteins, and chromatin remodelers) and their target genes (**Figure 1*a***) and edges represent directed regulation from a regulator to a target (**Figure 1*b***). Edges are determined by physical interactions between a regulator and a target, as when a TF physically interacts with binding sites at the promoters of its target genes. The expression level of a target gene is a function of the regulator's state as well as *cis* sequence features, such as a TF's binding affinity for a gene's promoter (**Figure 1*a***). A *trans* regulator's state may reflect a protein's level, location, and modifications. A regulator may directly regulate a target's expression, as when a TF binds to a promoter or enhancer, or may do so indirectly, as when a signaling protein modifies a TF and affects its binding to a promoter (**Figure 1*a***).

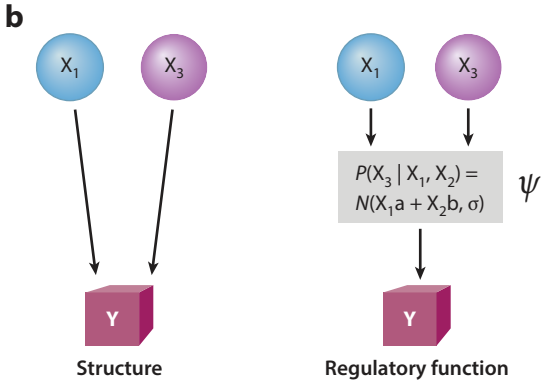
A regulatory network model has two components: (*a*) the structure, which specifies the regulators of a target gene (**Figure 1*b***), and (*b*) the regulatory function, encoded as a mathematical function (**Figure 1*b***), which describes how individual and combined regulatory inputs specify a target gene's expression. Several different mathematical functions at different levels of resolution are commonly used to relate regulatory inputs to expression output. Among them are Boolean functions (Dunn et al. 2014), ordinary differential equations (Greenfield et al. 2013, Molinelli et al. 2013), and probabilistic functions (Bonneau 2008, de Jong 2002, Friedman 2004, Kim et al. 2009, Markowetz & Spang 2007, Pe'er & Hacohen 2011, Segal et al. 2005). Although differential equation–based models and stochastic models based on a master equation (Thattai & van Oudenaarden 2001) can readily represent biochemical events and nonlinear interactions, they cannot always be fit with available data. Probabilistic graphical models, such as Bayesian networks, Markov networks, and factor graphs, are powerful paradigms for modeling regulatory networks (Friedman 2004, Markowetz & Spang 2007, Segal et al. 2005) and can handle noise and uncertainty in measurements. Although genome-wide gene expression data are commonly used as the sole or primary source to construct these models, some approaches integrate multiple types of data (Cheng et al. 2011, Ciofani et al. 2012, Marbach et al. 2012, Novershtern et al. 2011a, Segal et al. 2003).

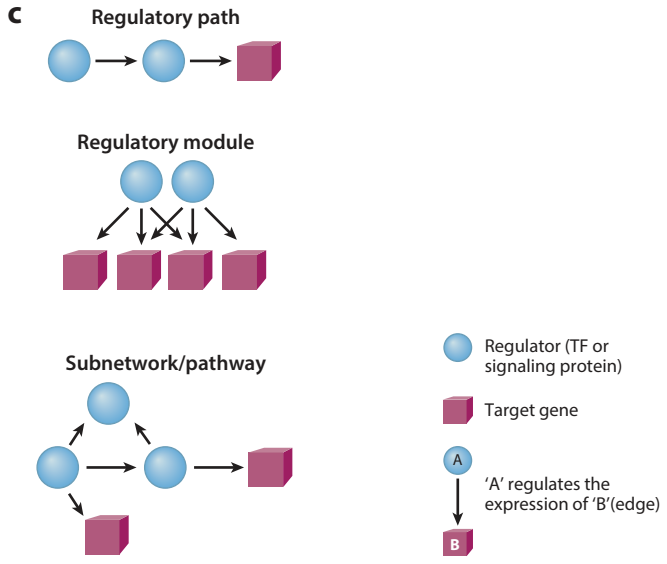
Five key experimental genomic strategies can help decipher the components and interactions in a regulatory network (**Figure 2**): (*a*) RNA profiling, today normally with RNA-seq (Hoheisel 2006,

**Figure 1**

Key features in regulatory network biology. (*a*) Regulators interact directly or indirectly with a gene's promoter. *Trans* regulators such as transcription factors (TFs; $X_1$ and $X_2$, *blue blobs*) bind to DNA, whereas signaling proteins ($X_3$, *purple blob*) modify TFs, affecting their regulatory activity and indirectly impact a target gene's downstream expression. (*b*) Many network models are specified by their structure (topology) and regulatory functions ($\psi$). The structure specifies the connections, or regulatory function edges (*arrows*) between regulators ($X_1$ and $X_3$, *blue and purple spheres*) and their target (Y, *red cubes*). The regulatory function ($\psi$) is associated with specific parameters that capture the nature of regulation (e.g., linear or logistic) and its strength. (*c*) Regulatory subnetworks. Regulators (*nodes*), target genes (*nodes*), and regulatory connections (edges) are assembled into network substructures that represent aspects of biological pathways. Examples shown (*top to bottom*) are a linear regulatory path with a series of regulators leading from an upstream regulator to a target gene, a regulatory module of coexpressed genes associated with a shared regulatory program, and a subnetwork of interconnected regulators and target genes. Regulatory paths and modules are specific cases of subnetworks.


a
Transcription factor level (trans)
Signaling protein level (trans)
X1
X3
X2
mRNA levels
Y
Transcription factor binding sites (cis)


b
X1
X3
Y
Structure
X1
X3
P(X3 | X1, X2) = N(X1a + X2b, σ)
ψ
Y
Regulatory function


c
Regulatory path
Regulatory module
Subnetwork/pathway
Regulator (TF or signaling protein)
Target gene
A
'A' regulates the expression of 'B'(edge)
B

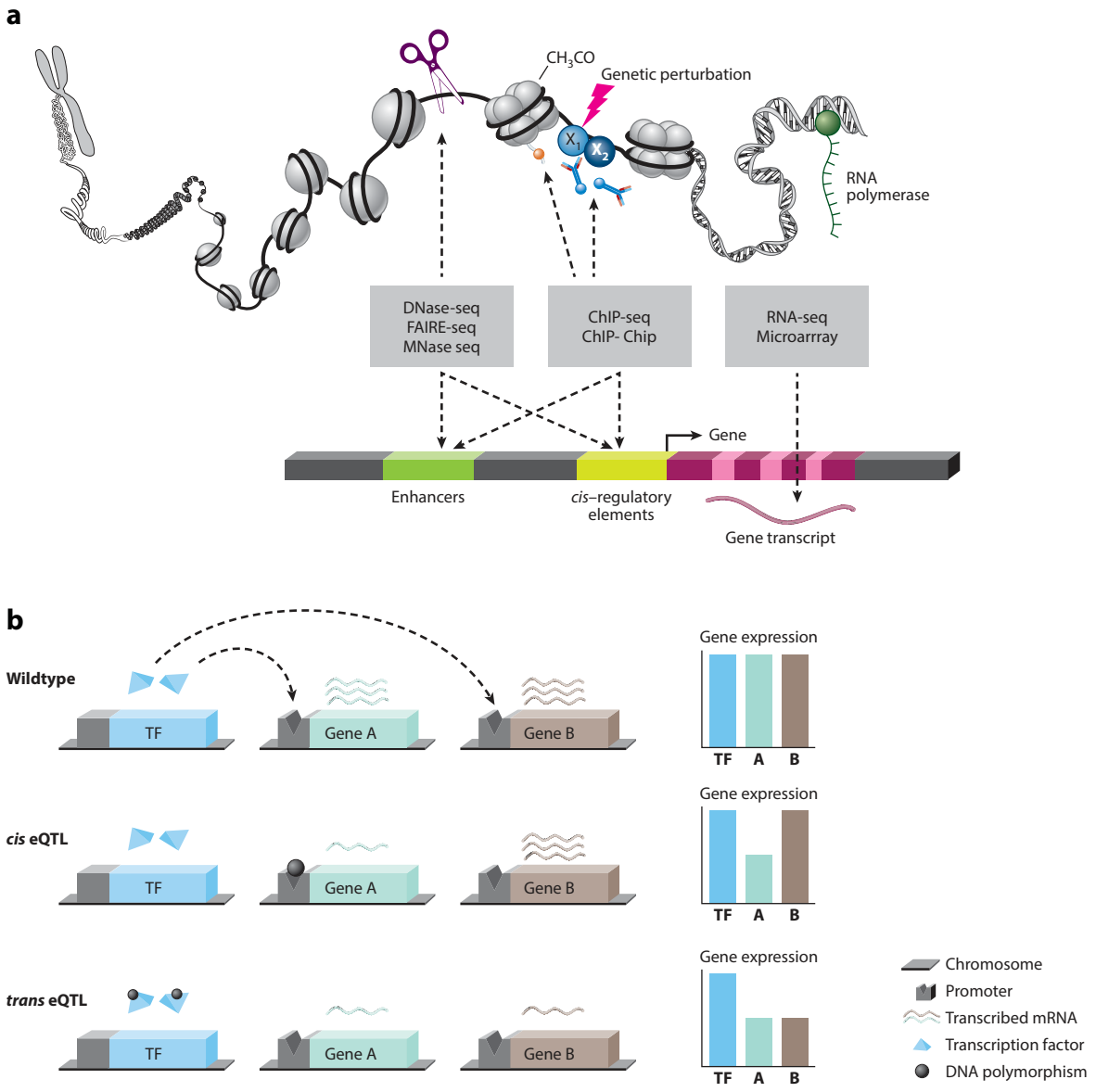



**Figure 2**

Key experimental genomic strategies to decipher the components and interactions of regulatory networks. (*a*) Chromatin accessibility profiling with DNase I hypersensitivity (DNase-seq), FAIRE-seq, ATAC-seq or MNase-seq measures open chromatin structure and nucleosome positions by employing enzymes (*scissors*) that cleave DNA that is not bound by nucleosomes (*shaded gray spheres*). Protein-DNA interactions for transcription factors (TFs; $X_1$ and $X_2$, *shaded blue spheres*) or histone modification states (*orange sphere*, $CH_3CO$) are measured by ChIP-seq or ChIP-chip assays. These technologies are used to identify *cis*-regulatory elements where TFs bind on gene promoters or distally on enhancers. RNA profiling is performed by RNA-seq or microarray analysis. Genetic perturbation (*pink lightning bolt*) of regulators by knockout, overexpression, genome editing, or knockdown followed by RNA profiling is used to validate predictions of regulatory networks. (*b*) Expression quantitative trait loci (eQTL) analyses: a genomic strategy to link genetic changes to their effects on gene expression in *cis* or in *trans*. A DNA polymorphism (*black sphere*) in a promoter (*notched gray box*) in Gene A alters binding of the TF (*blue pyramids*), reducing expression in *cis*. A *trans* change is illustrated by a DNA polymorphism in a gene encoding a TF that alters binding to the promoters of both Gene A and Gene B, resulting in reduced expression of both.

Nagalakshmi et al. 2008, Wang et al. 2009); (*b*) measuring protein-DNA interactions for TF or histone modification states with chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Furey 2012, Park 2009) or microarray (ChIP-chip assays) (Harbison et al. 2004, Johnson et al. 2007); (*c*) chromatin accessibility profiling by DNase I hypersensitivity (John et al. 2013, Song & Crawford 2010) and ATAC-seq (Buerostro et al. 2013**)** and nucleosome positioning with MNase-seq (Barski et al. 2007) and FAIRE-seq (Giresi et al. 2007); (*d*) genetic perturbation of regulators by knockout (Kemmeren et al. 2014), overexpression, knockdown (Amit et al. 2009, Chua et al. 2006, Nishiyama et al. 2009, Novershtern et al. 2011b, Walker et al. 2007), or finer editing (Hsu et al. 2014, Sander & Joung 2014), followed by profiling of RNA levels; and (*e*) expression quantitative trait loci (eQTL) analyses, which can link genetic changes to their effects on gene expression in *cis* or in *trans* (Brem 2002, Fraser et al. 2011, Göring et al. 2007, Majewski & Pastinen 2011, Rockman & Kruglyak 2006, Stranger & Raj 2013). Each strategy addresses a distinct aspect of regulatory networks: RNA profiling addresses the targets and output of the network; protein-DNA interaction and chromatin accessibility profiles help determine the molecular structure of the network; and genetic strategies, either engineered or based on natural variation, determine function and causality.

Most of these strategies have been deployed, at least to some extent, in a cross-species, comparative genomics setting, providing both excellent input for and substantial need for computational method development and application. First and foremost, many studies have collected comparative transcriptional profiles across phylogenies, in either a condition- or tissue-specific manner, including in yeasts (Thompson et al. 2013, Tirosh & Barkai 2011, Tirosh et al. 2006), flies (Jiménez-Guri et al. 2013, Kalinka et al. 2010, Paris et al. 2013, Prud'homme et al. 2007, Wittkopp et al. 2008), worms (Grishkevich et al. 2012, Yanai & Hunter 2009), plants (Ichihashi et al. 2014), fish (Brawand et al. 2014), frogs (Yanai et al. 2011), and mammals (Barbosa-Morais et al. 2012, Brawand et al. 2011, Merkin et al. 2012). Second, factor-specific ChIP-seq or ChIP-chip experiments have measured the binding of orthologous proteins to genomic loci across multiple species in a phylogeny, including in yeast (Baker et al. 2012, Borneman et al. 2007, Lavoie et al. 2010, Tuch et al. 2008), flies (He et al. 2011, Paris et al. 2013), vertebrates (Schmidt et al. 2010), and mammals (Ballester et al. 2014, Kutter et al. 2011, Perry et al. 2012, Schmidt et al. 2012, Stefflova et al. 2013, Wong et al. 2014). Similarly, epigenomic profiles of chromatin states, such as histone modifications and DNA methylation, have been collected across multiple species in matched tissues and developmental stages (Cotney et al. 2013, Ho et al. 2014, Villar et al. 2015, Xiao et al. 2012). Third, when collected across a set of species, global maps of open chromatin, created with DNase I hypersensitivity assays (Stergachis et al. 2014, Vierstra et al. 2014) or nucleosome position profiles compiled with MNase-seq (Tsankov et al. 2010, 2011; Tsui et al. 2011), allow a comprehensive comparison of entire regulatory landscapes, focusing on those elements that are accessible. Finally, an increasing number of approaches and studies in the emerging field of genetical genomics link genetic variants to gene expression traits in *cis* or in *trans* (**Figure 2*e***). Importantly, two new studies have examined transcriptomes across multiple species after knockout of a TF (Maguire et al. 2014, Wong et al. 2014), together with binding profiles of that TF (Wong et al. 2014). Others have leveraged natural genetic variation across individuals for eQTL analysis in humans (Dixon et al. 2007, Emilsson et al. 2008, Farh et al. 2015, Göring et al. 2007, Lee et al. 2014, Schadt et al. 2008, Stranger et al. 2007, Ye et al. 2014); worms (Grishkevich et al. 2012, Li et al. 2006); fish (Oleksiak et al. 2002); flies (Denver et al. 2005, Genissel et al. 2008), including recombinant inbred lines (Anholt & Mackay 2004); plants (Lasky et al. 2014, West et al. 2007); and mice (Peirce et al. 2004). Meiotic segregants or species hybrids (Thompson & Regev 2009, Wittkopp et al. 2008) in yeast (Brem 2002, Brem et al. 2005, Tirosh et al. 2006) and fly (Wittkopp et al. 2008) have also been investigated in this way.

## COMPARATIVE ANALYSIS OF REGULATORY GENOMICS DATA IN INFERRING NETWORKS AND STUDYING EVOLUTION

**Module:** set of genes coexpressed under different experimental conditions; often representative of a specific biological pathway

**TFBS:** transcription factor binding site

***cis*-regulatory module (CRM):** a collection of clustered binding sites

**Subnetwork/ Pathway:** set of genes and the regulatory connections among them; represents a smaller component of a regulatory network; a subnetwork captures a more fine-grained dependency structure than a gene expression module

As noted above, comparative analysis of gene regulation data presents two important possibilities. First, comparing across species allows us to better infer functional regulatory relationships, either because of the conservation of functional elements through purifying selection (Haerty & Ponting 2014, Lindblad-Toh et al. 2011, Siepel et al. 2005, Ward & Kellis 2012a, Xie et al. 2005) or because of the ability to link genetic changes across strains or species to their functional impact (e.g., on TF binding or mRNA levels) (Emerson & Li 2010, Zheng et al. 2011). Such inference should result in more accurate or comprehensive models of regulatory networks. Second, given inferred networks across species, analyzing reconstructed models through an evolutionary and phylogenetic lens can shed light on the evolutionary processes and associated adaptations. Such analyses can help address questions about the constraints that operate on quantitative traits such as expression levels and how these constraints act on individual genes, modules, and entire pathways. Thus, computational analysis methods for comparative data on gene regulation must tackle two challenges: the inference of gene regulation, which receives additional power from evolutionary considerations, and rigorous phylogenetic analysis of network evolution.

In practice, these challenges are tackled in the context of the different components and entities in gene regulatory network models, three of which are the focus of most studies today (**Figure 3**; **Table 1**). One line of research focuses on regulatory interactions with DNA, most prominently the identification and analysis across species of TF binding sites (TFBSs), either from sequence data alone or from TF binding information; sites that are clustered in a genomic region and form a *cis*-regulatory module (CRM) are of special interest. A second focus is on understanding the evolution of transcriptional or epigenomic states, at either the single-locus or gene-module level. Finally, emerging approaches consider an entire pathway or subnetwork, including regulators, their targets, and the regulatory interactions involved.

At each level, computational methods must (*a*) identify the relevant entities (e.g., regulatory DNA, gene module, or subnetwork) in the studied species, thus essentially inferring a regulatory model; (*b*) compare the entities across species or strains to estimate conservation and divergence; and (*c*) assess the rate of change, in particular testing whether changes deviate from a neutral model of evolution, and if so, how selection is acting on an entity. An entity is said to evolve neutrally if the change across species is linear with respect to the divergence time, i.e., when it is under neither purifying nor diversifying selection (Khaitovich et al. 2004).

### Comparative Analysis of Regulatory Sequence Elements

Divergence in regulatory sequence elements has been implicated in the evolution of complex traits (Villar et al. 2014, Wray 2007, Wray et al. 2003). Mutations in these elements can alter the affinity of a TF for a specific genomic region, either by partially affecting binding or resulting in a complete gain or loss of a binding site.

Detection of *cis*-regulatory elements from sequence data alone is notoriously challenging (Hardison & Taylor 2012), given the short and partially degenerate nature of the sites and the many perfect-match sequence sites that either are not bound by a factor or are bound without any apparent impact on gene expression. These characteristics both provide important opportunities and pose challenges for comparative analysis of *cis*-regulatory elements. First, phylogenetic conservation (or finer constrained patterns) can provide an important signal for detecting functional elements. Second, assessing how sequences change can highlight the sequences essential for binding specificity. Finally, monitoring the gain and loss of sites across a phylogeny can shed light

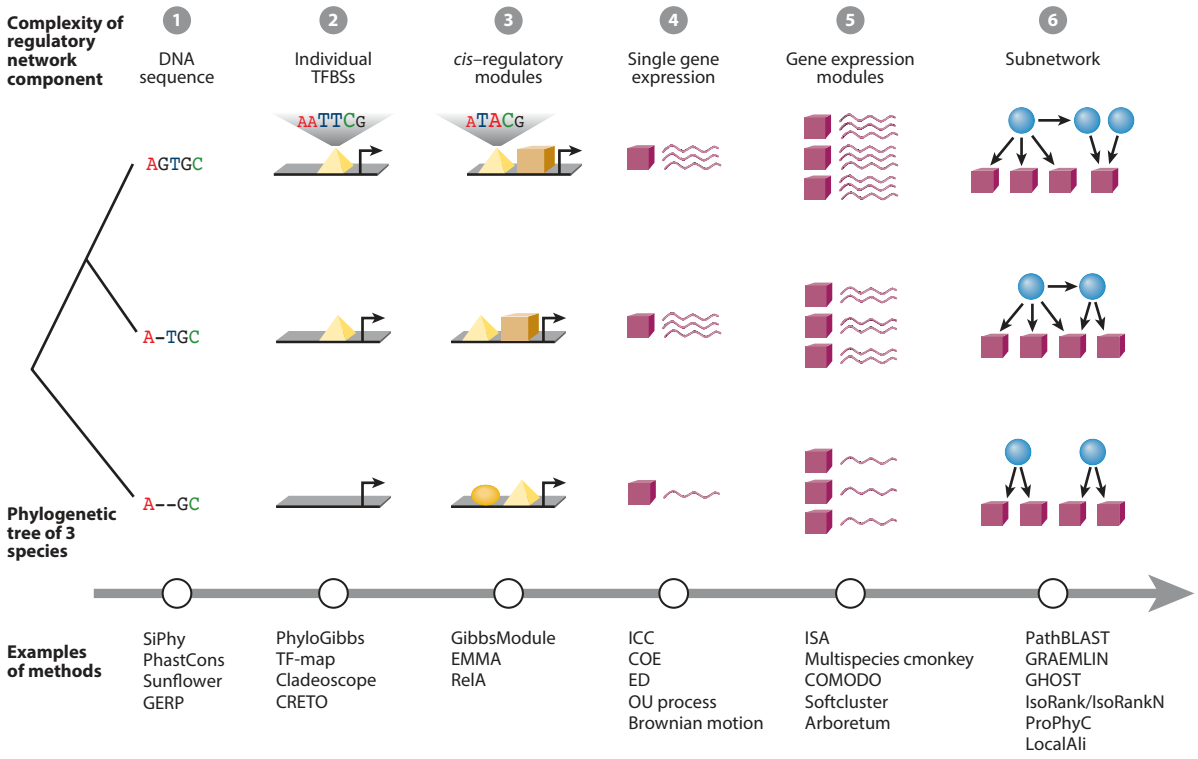



**Figure 3**

Comparative analysis of gene regulatory networks is performed at multiple levels: (*a*) DNA sequences, (*b*) individual TFBSs, (*c*) *cis*-regulatory modules, (*d*) expression levels of individual genes, (*e*) sets of coexpressed genes or modules, and (*f*) subnetworks. Shown on the left is a phylogenetic tree of three species. Representative methods, tools, and analytical measures used to study each level are shown at the bottom of the figure. Abbreviations: COE, conservation of expression; COMODO, conserved modules across organisms; CRETO, Cis-Regulatory Element Turn-Over; ED, expression divergence; EMMA, Evolutionary Model-based *cis*-regulatory Module Analysis; GERP, Genomic Evolutionary Rate Profiling; ICC, iterative comparison of coexpression; ISA, iterative signature algorithm; OU, Ornstein-Uhlenbeck; ReLA, REgulatory region Local Alignment tool; SiPhy, SIte-specific PHYlogenetic analysis.

on evolutionary constraints on gene regulation. Mirroring these opportunities, computational methods for comparative analysis of regulatory elements fall into two broad classes: those that detect reliable *cis*-regulatory elements across multiple species and those that assess *cis*-regulatory divergence across species. We highlight key approaches in each category.

## Leveraging Conservation to Detect *cis*-Regulatory Elements and Modules Across Multiple Species

Detecting mechanistically functional *cis*-regulatory elements, defined here as elements bound by regulatory factors that impact transcription of their cognate target genes, is notoriously difficult (Bailey 2008, Hardison & Taylor 2012, Whitfield et al. 2012). Extensive methods based on sequence and other experimental data from a single species are available for inferring the presence of *cis*-regulatory elements (Hardison & Taylor 2012, Stormo & Zhao 2010, Yáñez-Cuna et al. 2013). Almost invariably, the DNA-sequence specificity of a TF is represented as a sequence motif

**Table 1 Computational tools to address various problems in the comparative analytics of regulatory networks**

| Method or measure | Level of regulatory network | Type of analysis | Description |
|---|---|---|---|
| GERP | DNA sequence | Detection of sequence constraint | Tool to study constraint in DNA sequences |
| SiPhy | DNA sequence | Detection of site-specific sequence constraint | Probabilistic model of sequence evolution that has different rate parameters for individual site locations |
| PhastCons | DNA sequence | Detection of site-specific sequence constraint | Model of sequence evolution to find sequences evolving neutrally or under constraint; based on PhyloHMM |
| Sunflower | TFBS | Assessment of selection on binding sites | HMM-based approach to measuring selection on binding sites; adapts the dN/dS ratio to the context of regulatory regions |
| CRETO | TFBS | Assessment of selection on binding sites | Probabilistic phylogenetic model to study evolution of individual binding sites |
| Cladeoscope | TFBS | Sequence-specific motif learning and binding-site detection | Approach to identify motifs in multiple, distant species |
| EMMA | CRM | Detection of CRMs | Probabilistic model of background sequence and CRM evolution to detect CRMs |
| ReLA | Regulatory sequence element | Detection of general regulatory regions | Model based on local alignments of binding sites of different TFs |
| TF-map | *cis*-element | Alignment of promoters | Tool that maps binding site occurrences to the names of TFs and performs sequence alignment based on these data |
| ICC | Individual gene | Conservation of gene expression | Method based on Pearson's correlation coefficient; applicable when conditions measured across species are the not the same |
| COE | Individual gene | Conservation of gene expression | Approach based on Pearson's correlation coefficient; applicable for cross-species comparison when matched conditions are available |
| ED | Individual gene | Divergence of gene expression | Procedure based on a Euclidean distance metric of expression divergence; requires matching across conditions |
| Brownian motion | Individual gene/ gene module | Selection on expression | Model that assumes expression evolves at the same rate on all lineages |
| Ornstein-Uhlenbeck process | Individual gene | Selection on expression | Model that can be used to detect lineage-specific selection |
| COMODO | Expression module | Detection of coexpressed gene modules across two species | Approach that exploits homology and coexpression to simultaneously find coexpressed modules in two species; a module includes homologous genes as well as additional nonhomologous, species-specific genes that serve as linkers |
| Softcluster | Expression module | Detection of coexpressed gene modules across multiple species | *k*-means algorithm with a heuristic to favor coclustering of orthologous genes |

(*Continued*)

**Table 1** (*Continued*)

| Method or measure | Level of regulatory network | Type of analysis | Description |
|---|---|---|---|
| Arboretum | Expression module | Detection of coexpressed gene modules across multiple species | Probabilistic model of module evolution and detection; based on a Gaussian mixture model of module expression |
| ISA | Expression module | Detection of coexpressed gene modules in one species | Algorithm that identifies biclusters in one species at a time, where each bicluster represents a set of genes that are coexpressed in a subset of conditions |
| Multispecies cMonkey | Expression module | Detection of coexpressed gene modules across multiple species | Method that identifies biclusters in multiple species; starts with seeds of orthologous genes that are coexpressed in each species and iteratively adds new genes to each module |
| PathBLAST | Network | Network alignment | Local network alignment for two species |
| NetworkBLAST-M | Network | Network alignment | Local network alignment that is applicable to multiple species |
| GRAEMLIN | Network | Network alignment | Global network alignment that is applicable to multiple species |
| GHOST | Network | Network alignment | Global network alignment |
| IsoRank/IsoRankN | Network | Network alignment | Global network alignment based on spectral methods; IsoRank works with two species, whereas IsoRankN works with more than two species |
| LocalAli | Network | Network alignment | Local network alignment program that makes use of a phylogeny to find conserved subnetworks |
| ProPhyC | Network | Network refinement | Approach that refines a set of extant regulatory network reconstructions; makes use of a probabilistic model of network evolution |

Abbreviations: COE, conservation of expression; COMODO, conserved modules across organisms; CRETO, Cis-Regulatory Element Turn-Over; CRM, *cis*-regulatory module; dN/dS ratio, ratio of nonsynonymous to synonymous substitutions; ED, expression divergence; EMMA, Evolutionary Model-Based *cis*-Regulatory Module Analysis; GERP, Genomic Evolutionary Rate Profiling; HMM, hidden Markov model; ICC, iterative comparison of coexpression; ISA, iterative signature algorithm; ReLA, REgulatory region Local Alignment tool; SiPhy, SIte-specific PHYlogenetic analysis; TF, transcription factor.

(**Figure 3**) and formally encoded as a position weight matrix (PWM), which specifies the DNA base pair composition at each position of the motif.

Phylogenetic signals, especially conservation, have been used for more than a decade to determine the position of sites acted upon by purifying selection in a group of species (Boffelli et al. 2003, Bulyk 2003, Frazer et al. 2003, Margulies et al. 2003, Skipper 2004, Stark et al. 2007), based on the intuitive notion of a phylogenetic footprint (Blanchette & Tompa 2002). Initially, such approaches were site (i.e., instance) specific, relying on an alignment of orthologous sequences to identify a conserved element using programs such as PhyloGibbs (Siddharthan et al. 2005) and Phylogenetic Motif Elicitation (PhyME; Sinha et al. 2004). Given the inability to align regulatory DNA beyond certain evolutionary distances, methods have also emerged to compare the set of sites associated with orthologous genes across more distant species (Gordân et al. 2010, Habib et al. 2012). Some approaches use comparative data to both discover the motifs and score them. Others rely on an existing library of motifs—for example, those determined by experimental approaches in one species (Daniel & Newburger 2009, Kheradpour & Kellis 2014, Mathelier et al. 2013, Neph et al. 2012)—scoring and refining motif instances based on evolutionary conservation.

**Position weight matrix (PWM):** commonly used to represent the DNA sequence binding affinity of a transcription factor

**Hidden Markov model (HMM):** a probabilistic model commonly used to model sequential data; provides the basis of many sequence analysis tools

The input for the alignment-based method PhyloGibbs (Siddharthan et al. 2005) is several multiple sequence alignments of orthologous promoters and one alignment for every gene. The program uses a Bayesian framework that combines a sampling-based, motif-finding approach with phylogenetic information on orthologous promoters. Because alignment of regulatory regions is challenging, methods that need an input alignment are applicable only to closely related species. Several hybrid approaches exist, in which an initial, possibly noisy alignment is gradually refined while the presence of a sequence motif is simultaneously detected (Bais et al. 2007).

Conversely, Cladeoscope (Habib et al. 2012) is an alignment-free approach that can be applied over great phylogenetic distances. It requires an independent initial set of motifs from which to learn species-specific motifs. Being alignment free, Cladeoscope can be used to compare regulatory sequences between very distant species; the available initial library impacts its sensitivity, however, such that it is more sensitive in species closer to its source. Indeed, when applied to 23 species of Ascomycete yeasts, Cladeoscope performed better in species more closely related to *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, from which most motif data were derived. The absence of motifs detected in distant species may reflect divergence in TF-binding specificity rather than true loss (Habib et al. 2012).

In many multicellular organisms, multiple binding sites are typically organized in close proximity to one another on individual gene promoters within CRMs, and it is reasonable to hypothesize that selection may act not only on individual elements but also on entire CRMs (Villar et al. 2014). This selection may occur at two levels. First, within a CRM, a site may only move when paired with another, dependent site; for example, cooperative binding of certain factors may require them both (Villar et al. 2014). Second, the entire CRM may evolve as a single unit (Duque et al. 2014) moving, being duplicated, or being lost as a whole.

Identification and comparative analysis of CRMs and the enhancers that harbor them have been the subject of intense research. For example, GibbsModule (Xie et al. 2008) extends PhyloGibbs to find CRMs, relying on an input sequence alignment, whereas Evolutionary Model-based *cis*-regulatory Module Analysis (EMMA; He et al. 2009) is an alignment-free method that uses a statistical model of the evolution of CRMs and an input library of motifs to simultaneously align regulatory regions and detect clusters of TFBSs. Briefly, EMMA represents a CRM as a hidden Markov model (HMM), with a background state to generate nonbinding sites and $k$ states to represent each of the $k$ motifs. EMMA has distinct evolutionary models for the background sequence and binding sites, the latter based on a population genetics model of different evolutionary constraints for sites in a motif with different extents of degeneracy. Applied to simulated and real data from two *Drosophila* species, EMMA improved alignments of regulatory regions relative to existing methods that assumed a fixed input sequence alignment, as measured by the ability to recover conserved binding sites in aligned regions and accurately discriminate between true and false CRMs. EMMA identified CRMs in regions that were previously classified nonfunctional sites and further showed that distal CRMs were more conserved than promoter-proximal CRMs. Another approach, REgulatory region Local Alignment tool (ReLA; González et al. 2012), uses a local alignment strategy to identify regulatory regions that are characterized by the presence of conserved clusters of binding sites. Such regions can represent enhancers as well as promoters. In a nutshell, ReLA converts occurrences of binding sites into symbols of a different alphabet and aligns the sequences in this new alphabet to identify contiguous, aligned regions of binding sites. ReLA identifies experimentally validated enhancers and promoters with as good or higher precision and recall as other approaches for identifying transcription start sites or conserved TFBSs, and it can serve as a useful annotation tool for less well-characterized genomes.

Finally, related strategies are used to identify general noncoding or functional regions, not necessarily those characterized by TFBSs. Most notably, PhastCons (Siepel et al. 2005) finds

general conserved coding and noncoding regions based on a two-state phylogenetic hidden Markov model (PhyloHMM; Siepel & Haussler 2004), in which one state is for conserved elements and one for nonconserved elements. PhastCons has become the gold standard in large comparative genomics projects seeking to identify phylogenetic footprints of conserved elements (Brawand et al. 2014, Lindblad-Toh et al. 2011, Pique-Regi et al. 2011), and it has proven useful for identifying noncoding functional regions, including enhancers (Brawand et al. 2014).

## Assessing Conservation and Divergence of *cis*-Regulatory Elements and Modules

A key goal is to reconstruct and monitor the process of evolution by comparing *cis*-regulatory elements across species. It is critical not to conduct such analyses in a circular fashion. In particular, any evolutionary constraint introduced by the method used to develop the model cannot then be purported to be a conclusion of the resulting model. An obvious example is conservation: If conservation is aggressively used to call sites, it is not surprising that conservation is subsequently observed across the resulting model. Overall, computational approaches to compare *cis*-regulatory models across species can be partitioned into those that only quantify the extent of divergence and those that additionally inform us about the rate of divergence in specific lineages, distinguishing neutrality from positive or negative selection.

*Cis*-regulatory site evolution has been explored from two perspectives: the evolution of individual bases within a site and an overall site's gain and loss, known as turnover. Although the two are interrelated when site turnover happens through a series of point mutations, the former emphasizes effects on binding specificity, whereas the latter highlights overall regulatory impact.

Two popular tools for assessing the conservation and divergence of single base pairs are applicable to *cis*-regulatory elements: Genomic Evolutionary Rate Profiling (GERP; Cooper et al. 2005) and SIte-specific PHYlogenetic analysis (SiPhy; Garber et al. 2009). GERP explicitly measures evolutionary constraint on sequence quantitatively (Cooper et al. 2005). It computes the mutation rate for each column of a multiple sequence alignment for a set of species. It then compares these rates with a neutral rate of mutation for the species under consideration and returns candidate elements that have fewer mutations than would be expected from the neutral model (i.e., the Hasegawa, Kishino, and Yano model of DNA sequence evolution; the neutral rate is estimated using only those species that do not have a gap in that column). SiPhy considers more general patterns of conservation, including constraint that is reflected in the biased substitution of specific bases. Using a rate matrix, SiPhy models the evolution of DNA sequences as a continuous-time Markov process for each branch on the species tree. Rate matrix entries reflect site-specific biases, the neutral substitution rates of bases, and the overall rate of substitution. Like GERP, SiPhy estimates rate parameters for each column, but it also uses an HMM to identify contiguous, variable-length regions exhibiting similar levels of constraint. SiPhy identifies a greater number of constrained sequences compared to rate-based methods such as GERP or PhastCons, many of which are degenerate. It has been used to compute constraint in the human genome using whole genome alignments from 29 mammals (Lindblad-Toh et al. 2011), as well as to identify regulatory variants in the human genome (Ward & Kellis 2012a,b).

Overall constraint can be leveraged or interpreted in a more nuanced way, tailored specifically to the molecular reality of a DNA sequence bound by a TF, by considering how a change in a site's sequence would impact binding affinity. For example, the site-level selection (SS) method, first used to study the evolution of regulatory sequences in 12 *Drosophila* species (Kim et al. 2009), not only quantifies the rate of substitution from one site to another but determines each site's fitness level based on its binding affinity (Kim et al. 2009). The Predicted Expression-Based

Site Evolution Simulator (PEBSES) and Predicted Expression-Based CRM Evolution Simulator (PEBCRES) methods have since extended the SS method to model selection on a site within a CRM and on an entire CRM, respectively (Duque et al. 2014). A complementary approach (Tanay et al. 2004a) used promoter sequence alignment across four yeast species to compute a normalized substitution rate between pairs of DNA octamers that differed by one base pair. Next, a selection network was generated, in which nodes represented octamers and edges were weighted by the rate of substitution from one octamer to another. Clustering the selection network identified families of motifs that shared a high substitution rate and were isolated from other octamers with lower substitution rates. The organization of the selection network provided several insights into the affinity of multimodal TFs and their functional diversity.

In another example, the method Sunflower (Hoffman & Birney 2010) studies selection on binding sites by capturing the impact of a mutation in a TFBS on a TF's binding affinity. Sunflower uses an analog of the ratio of nonsynonymous to synonymous (dN/dS) substitutions between two protein-coding sequences, where dT (analogous to dN) is the change in binding affinity resulting from a mutation. dT is obtained from the relative entropy between the posterior distribution of states before and after a simulated disruption in binding sites, using an HMM with one state for the unbound state and multiple additional states for each column of the TF's PWM. Sunflower has revealed that dT/dS values are typically uncorrelated with dN/dS values for the protein coding portions of both individual genes and functional gene classes, suggesting that positive selection acts on processes either at the transcriptional or the protein-coding level, but not both.

Finally, some approaches study divergence at the binding-site level, relying on either computational determination of binding sites without alignments (as for Cladeoscope, discussed above) or on comparative in vivo binding data sets created using ChIP-chip or ChIP-seq experiments. In some cases, an underlying assumption is that although the entire promoter sequence may be evolving neutrally, and the positions of the binding site may change, binding site occurrence may be under constraint; these methods are set to either leverage or capture this constraint. An early example of this strategy, applicable only to pairs of species, was TF-map (Blanco et al. 2006). Here, the difficulty of aligning fast-evolving promoter sequences was addressed by representing them in a different alphabet, in which each character represented a motif instance.

A second generation of approaches relaxes this assumption even further, quantifying turnover even in the absence of an alignment by considering the binding sites associated with a gene, locus, or even the genome as a whole. For example, Cis-Regulatory Element Turn-Over (CRETO; Otto et al. 2009) is a maximum likelihood–based approach that models the number of binding sites for a specific TF in a phylogeny, without requiring binding site positions to be conserved between species. Using an explicit model of evolution parameterized by a rate of gain and loss of binding sites in different clades of the phylogeny, CRETO can also provide insight into potential selective forces that act on binding sites (described below). As CRETO only analyzes a genome-wide aggregate count of binding sites, it does not capture gene-level binding information. It is most useful when applied to factors associated with tightly regulated gene modules, all sharing the same regulatory program (e.g., methionine genes in yeast; Gasch et al. 2004). Conversely, Cladeoscope (Habib et al. 2012) analyzes turnover at either a per-gene or per-module level. When applied to 23 species of Ascomycete yeasts, Cladeoscope showed rapid turnover of binding sites at the gene level. Despite this rapid turnover, association of biological processes with the gene targets of a TF remained conserved, demonstrating that constraint operates at the level of biological processes and pathways associated with TF targets.

The challenge of reliably detecting sites while not overly relying on sequence conservation is substantially relaxed when comparative TF binding data are available. Such data reflect direct evidence of regulatory interactions between TFs and DNA across multiple species (Borneman

et al. 2007, He et al. 2011, Schmidt et al. 2010, Tuch et al. 2008) or regions with binding-accessible chromatin based on DNase I hypersensitivity assays (Stergachis et al. 2014, Thurman et al. 2012, Vierstra et al. 2014). Several studies have compared binding profiles across species, either focusing on per-site comparisons (Bardet et al. 2012; Schmidt et al. 2010, 2012) or per-ortholog comparisons (Borneman et al. 2007, Tuch et al. 2008). The former strategy requires alignment, to relate bound sites between species, but does not require knowing the association of a site with a regulated target gene, making it a compelling option for mammalian genomes, in which much binding occurs at distal enhancers. The latter strategy does not require alignment, as it is the site-target association that is compared across orthologs, but it does require knowing which target gene is associated with each site. It is thus useful for either lower eukaryotes, such as yeast, or instances when binding and regulation occur at the proximal promoter.

Both strategies have been successfully applied in relevant settings. For example, a continuous-time Markov chain analysis of individual binding instances, derived from ChIP-chip data, captured insertion-deletion (INDEL) events genome wide (Zheng & Zhao 2013). This approach projects each bound peak onto the corresponding species entry in a multiple sequence alignment and converts the positions of the multiple sequence alignment into an alphabet of GAP (G), BINDING (B), or NO-BINDING (N), each represented by a distinct state in the continuous-time Markov chain. The Markov chain's rate matrix contains a rate parameter that reflects loss or gain of a binding site for each branch of the tree; this parameter is estimated using Expectation Maximization (EM). Application of this approach to published binding data for the TF CEBPA (Schmidt et al. 2010) has shown that sites that are conserved at the nucleotide level are under strong selective constraint, as measured by GERP scores, and are unlikely to be disrupted by INDELs. However, neutral regions display accelerated transition rates in TF binding regions (B) compared with nonbinding regions (NB), with accelerated binding site turnover. An example of the per-ortholog comparison was the study of the evolution of the target set of the yeast TF Mcm1 (Tuch et al. 2008). In particular, the authors developed a maximum likelihood–based evolutionary model to estimate the rate of gain and loss of the Mcm1 targets, from Mcm1 binding profiles across three yeast species (Tuch et al. 2008). Cladeoscope later used a similar model to estimate turnover from computational predictions of TF gene targets based on the aggregated affinity of a TF for a gene promoter (Habib et al. 2012).

## Comparative Gene Expression Across Species: From Genes to Modules

Whereas *cis*-regulatory elements and the factors that bind them represent key regulatory components of a network and allow us to infer and compare regulatory mechanisms across species, genome-wide expression profiles, especially mRNA profiles, are a major output of transcriptional regulatory networks (**Figure 3**). In particular, they provide a comprehensive molecular phenotype from which physiological functions and organismal phenotypes can be inferred, thus connecting network organization to function and regulatory evolution to adaptation (Necsulea & Kaessmann 2014, Romero et al. 2012, Zheng et al. 2011).

Changes in expression across species are studied at either the level of individual gene orthologs or the level of sets of genes coexpressed in a module or pathway. As for *cis*-regulatory sequences, all methods aim to quantify the extent of expression or regulation divergence, and some also assess the extent to which the observed divergence is explained by a neutral model versus selection (**Figure 3**). Performing any comparison requires matching orthologs across species (a computational challenge addressed in other reviews, typically based on sequence data alone; see Kuzniar et al. 2008, Koonin 2005), and, in the case of module-level analysis, inferring modules across species, which can be performed independently for each species or in a phylogenetically informed way, as discussed below.

Two major strategies are used to estimate the extent of divergence or similarity in orthologous gene expression between two species: comparing gene coexpression even when profiles are not matched between species and comparing expression when matched measurements are available. A compelling early method to compare unmatched samples in a pair of species was iterative comparison of coexpression (ICC; Tirosh & Barkai 2007). ICC measures the similarity in coexpression profiles between two orthologs, with similarity being iteratively computed for each gene with an ortholog in the other species. In the first iteration, the ICC value is a Pearson's correlation coefficient computed between two vectors, one for each species. Each vector reflects the correlation of a gene with all other genes in that species, limited to genes with matched orthologs. In subsequent iterations, the ICC value is a weighted correlation, where the weight is the ICC value from the previous iteration. This iterative procedure enables ICC to favor genes whose expression correlations are more conserved across species. A related measure that requires matched conditions is expression divergence (ED) (Tirosh et al. 2006). ED more directly considers the correlation in expression profiles between orthologs while controlling for technical variation, differences in the magnitude and kinetics of expression, and scaling differences. A related measure of ED that also requires matched conditions is conservation of expression (COE) (Shay et al. 2013). A more principled approach to comparing expression between matched conditions relies on a likelihood ratio test (Khan et al. 2013). For every pair of orthologs, a likelihood ratio is computed for a gene's expression under (*a*) the null model of no differential expression between two orthologs and (*b*) a second model that captures a species-specific component of expression. A likelihood ratio can be converted into a *P*-value using the $\chi^2$ distribution, assigning statistical significance to divergence in expression while controlling for the false discovery rate.

Although gene-pair comparisons are helpful, they may be prone to errors in measurement and errors in correctly resolving orthologs, as well as phenotypic changes due to expression divergence of individual genes and are less directly interpretable than most researchers would prefer. Conversely, gene modules—sets of coexpressed and/or co-regulated genes—are a hallmark of gene regulatory networks (Tanay et al. 2004b), and analysis at the module level has repeatedly been shown to increase both statistical robustness and biological interpretability in single-species studies (Mitra et al. 2013). Furthermore, as noted above, analysis of comparative *cis*-regulatory elements suggests that selection may act to conserve the regulation of a pathway or a process rather than the regulation of any particular gene in a module (Habib et al. 2012). Thus, there has been substantial effort to develop methods to infer and compare regulatory modules between species, either pairwise or simultaneously across multiple species in a phylogeny.

Some strategies identify modules of orthologous genes whose coexpression is conserved in pairs of species, and these methods do not typically require data from matched conditions (Bergmann et al. 2003, Tanay et al. 2005, Zarrineh et al. 2011). One such approach first identifies modules of coexpressed genes in each species independently and then matches the modules across pairs of species based on significant overlap in the orthologous genes within the two modules (Tanay et al. 2005). A related approach, conserved modules across organisms (COMODO), detects sets of coexpressed genes by extending the ICC concept of conserved correlation of expression from a single gene to multiple genes (Zarrineh et al. 2011). COMODO seeds modules of highly coexpressed genes identified independently for each species, and finds initial pairs with significant orthology overlap, and then expands these modules to include additional genes. In both cases, the matched pair of orthologous modules includes orthologs coexpressed in both species and additional, species-specific genes.

Other strategies rely on matched conditions to infer modules across larger numbers of species. These strategies are distinct from pairwise methods because of their ability to consider expression levels of genes (rather than only coexpression), as well as the phylogenetic relationships between

multiple species. As in the pairwise case, modules can include both conserved and species-specific members, and phylogenetic structure can enhance the power to reliably identify clade-specific and lineage-specific patterns. One early strategy, Softcluster (Kuo et al. 2010), concatenates expression data from each species by matching corresponding experimental conditions and then uses *k*-means clustering with an added heuristic to favor orthologous genes being in the same cluster. We recently developed a more flexible strategy, Arboretum (Roy et al. 2013), that directly models gene orthology relationships across multiple species in a complex phylogeny. Arboretum relies on a generative probabilistic model of module evolution and a Gaussian mixture model to represent the expression modules in each species, and assigns genes to modules in each extant and ancestral species. It also models gene duplication events and infers ancestral modules. As a result, Arboretum can handle genes with complex orthology relationships, resulting from both duplication and loss. Notably, however, Arboretum does not infer the expression pattern in ancestral nodes.

Both Softcluster and Arboretum assume that a given gene may belong to only one module capturing its expression profile across all investigated conditions in a given species. This is a reasonable assumption when the number of conditions investigated is small, as was the case in both studies used to introduce the programs, but it is less reasonable when many diverse conditions are investigated in each species. In that case, a gene may belong to different modules under different conditions, a situation captured by biclustering methods. For example, the Iterative Signature Algorithm (ISA) is applied in a cross-species setting by first identifying modules in one species and then finding the set of orthologs of that module's members in another species, followed by a two-stage application of ISA to identify conserved and coexpressed modules. More generally, cMonkey identifies biclusters in multiple species by identifying seed biclusters for every pair of species, followed by refinement of biclusters based on species-specific information (Waltman et al. 2010). Notably, cMonkey can incorporate diverse data, such as information on binding sites and metabolic pathways.

## Estimating Selection Pressure on Gene Expression

Fundamental questions about assessing the selective pressures acting on gene expression and coexpression (i.e., regulation) are at the crux of reasoning about regulatory evolution and its adaptive importance (Romero et al. 2012, Necsulea & Kaessmann 2014). However, unlike models of sequence evolution, which are employed as the basis for analysis of *cis*-regulatory evolution, models of the evolution of expression and coexpression are still in their early stages. Arguably, a unified model has not yet been fully developed. In the past decade, both model-based and empirical measures of selection on gene expression have been developed. Model-based approaches assume an explicit model of the evolution of quantitative traits in a phylogeny (Bergmann et al. 2003), most commonly the Ornstein-Uhlenbeck (OU) model or the Brownian Motion (BM) model (see below).

Empirical strategies to assess selection on expression rely on ranking or nonparametric statistical tests that aim to reflect some prior, informal assumption about how selected expression patterns would behave compared to nonselected ones. For example, one approach discussed by Romero et al. (2012) relies on the intuition that positive selection manifests as low interindividual variation within a species and differential expression across species. Genes can thus be ranked based on the difference between expression variation within and between species, with genes that are candidates for positive selection expected to rank highly. Such rank-based approaches are empirical and straightforward: They make few assumptions about the model of selection and its contributing factors and are well-suited for studies in nonmodel systems (Romero et al. 2012). An earlier study, the first to examine whether divergence in gene expression between species results

from positive selection or neutral processes (Khaitovich et al. 2004), relied on the assumption that mutations cause changes in expression independent of the absolute level of expression. Hence, the squared difference in the log of expression of orthologous genes should scale linearly with phylogenetic distance under a neutral model. To determine whether gene expression between humans and chimpanzees evolved neutrally, the authors used both the Kolmogorov-Smirnov (KS) and Wilcoxon rank-sum tests to compare two distributions: one distribution captured the squared difference of the mean expression of each gene between the two species, and the other distribution the squared difference of the mean expression of a set of pseudogenes, considered to be evolving neutrally. A significant *P*-value for any test suggested that a neutral model could not explain the expression differences between the species. The authors also devised a test for positive selection that compared the distributions of diversity with the divergence ratio in intact genes versus pseudogenes. Again, KS and Wilcoxon rank-sum tests were used to assess statistical differences between the two distributions. However, when applied to these data, no significant change was detected using either this approach or related strategies, supporting a neutral model of expression evolution among the species studied.

Conversely, model-driven approaches rely on statistical models of expression evolution as a continuous trait to quantify the extent of selection in a lineage-specific manner in a complex phylogeny. The two popular models for expression evolution are the BM and OU models. The BM model is the simplest, assuming that a quantitative trait evolves according to a neutral model or by random drift (i.e., there is no selection). The only parameter in the model is the intensity of random fluctuations in a trait of interest, which is the same for all branches of the tree (Butler & King 2004). The OU approach explicitly models selection for each branch, such that selection acts to push the current value of the trait toward its optimal value. Although all lineages evolve neutrally in the BM model, the OU model enables different selective forces to act on lineages, each represented by branch-specific arguments. Both models are represented as multivariate Gaussian models whose parameters are estimated using maximum likelihood. However, in the BM model, means are lineage-specific estimates of a trait, whereas in the OU model, means are linear combinations of other parameters defining the optimal values of the trait in ancestral paths. The OU process was recently used to study the evolution of tissue-specific expression in eight mammals and chicken (Brawand et al. 2011). It found a large number of genes under positive selection, many in the testis, and depletion in positive selection on gene expression in the brain. More sophisticated models are still emerging, and others are likely to come. For example, mixture models can help capture selection pressure on expression (Eng et al. 2009) by attempting to decompose the total variance in expression of a gene across species into its correlated phylogenetic signal, assumed to be given by a neutral model; the uncorrelated phylogenetic signal, to capture selection; and residual experimental variance.

Finally, such approaches can be applied not only to individual genes, but also to entire predefined gene modules (**Figure 3**). For example, a recent study (Schraiber et al. 2013) developed a maximum likelihood framework based on the BM model to assess the evolution of expression of genes in a module. In this model, the rate of expression evolution of a module's genes follows an inverse gamma distribution whose parameters are fit to maximize the observed expression in extant species. Applying this approach to mRNA profiles from four sensu stricto *Saccharomyces* yeast species and interspecies hybrids identified several modules that display non-neutral regulatory variation. Another study (Fraser et al. 2011) analyzed predefined modules for bias in the direction of changes in expression that can be explained by a particular variant in *cis*. It revealed modules associated with locomotion, growth, and memory that were under lineage-specific selection between two mouse species. Module-level analyses are thus particularly valuable in relating sequence regulatory variants with a measurable non-neutral impact to higher-level phenotypes.

## Comparative Analysis of Subnetworks and Pathways

A comprehensive understanding of regulatory networks and their evolution would, by necessity, combine both regulatory inputs and their impact on target gene expression within a single integrative network model (**Figure 3**). Comparative network analysis can assist in the inference of such models (network reconstruction), their comparison between species (network alignment), and quantification of any selective pressures at the network level (network selection).

Although computational approaches for network reconstruction in a single species have been an area of intensive research for more than a decade (Bonneau 2008, Kim et al. 2009, De Smet & Marchal 2010), approaches that leverage comparative data or a phylogenetic approach are only emerging: Some methods use comparative data to refine regulatory network reconstructions first obtained independently in individual species, whereas others construct a regulatory network for multiple species simultaneously. The refinement process takes a network inferred in a single species and uses comparative data to improve it. For example, ProPhyC (Zhang & Moret 2012) is a probabilistic approach that iteratively refines a network based on the assumption that networks whose evolution follows that observed in the species genome are more likely to be true. This assumption is also made by the related methods RefineML and RefineFast (Zhang & Moret 2008). ProPhyC allows diverse network evolution models, including edge gain and loss and gene (i.e., node) duplication and loss. Other strategies aim to jointly infer networks across multiple species. For example, one approach uses a probabilistic framework of regulatory network evolution to estimate regulatory edges in each extant and ancestral species (Xie et al. 2011). In this model, the connection between a TF and a target gene is modeled using a hidden variable (edge state, **Figure 1*c***) that determines the generative process of a sequence or the expression level of a given gene. If a TF regulates a gene, its sequence is governed by an HMM that captures the TF's binding site; if it does not regulate the gene, its sequence and expression state are governed by a background sequence and expression model. The expression cluster to which the gene is preassigned determines the target gene's expression. Thus, the regulatory connections themselves evolve on the phylogenetic tree, modeled by a continuous-time Markov process with parameters for the gain and loss of regulatory edges learned using EM (Xie et al. 2011). Although these approaches show initial promise, the limited number of multispecies data sets has curtailed method development. As more functional genomics profiles are collected, it will be possible to develop and test approaches that more systematically integrate data across multiple species, allowing changes in genes and interactions.

More extensive work has been done to develop methods that can compare network models—independently defined in each species and typically represented as graphs—between species. Network alignment methods, analogous to sequence alignment methods, aim to find conserved subnetworks between a pair of networks (**Figure 3**). Although most network alignment methods were initially developed in the context of protein-protein interaction graphs, which tend to not have directional information on edges, they can be applied to regulatory network models as well provided that directionality in the regulatory network can be ignored. By analogy to sequence alignment methods, network alignment methods can be distinguished by whether they perform pairwise (reviewed in Clark & Kalita 2014) or multiple (Flannick et al. 2006, Liao et al. 2009) network alignment and by whether they perform local alignments of subgraphs or global alignments of entire graphs, the latter primarily useful for closely related species. Among local alignment methods are the pairwise PathBLAST (Kelley et al. 2004), its multiple network extension NetworkBLAST-M (Kalaev et al. 2009), and LocalAli (Hu & Reinert 2015), whereas IsoRank (Singh et al. 2008), IsoRankN (Liao et al. 2009), GRAEMLIN (Flannick et al. 2006), and GHOST (Patro & Kingsford 2012) are all global network alignment methods.

All alignment methods aim to create a new graph, in which nodes represent an orthologous group of proteins and edges represent conserved interactions among the proteins. Conserved subnetworks are then identified by matching nodes in one species to nodes in another species, such that both sequence similarity and network similarity are maximized in the alignment. Methods vary in their definition of network similarity. For example, in PathBLAST (Kelley et al. 2004), similarity is defined by conserved edges in the local neighborhood of a gene. IsoRank, IsoRankN, and GHOST are based on spectral methods (von Luxburg 2007) that exploit global topological properties to define the relevance or connectivity of a specific graph and to define node similarity. Although most multispecies methods are agnostic of phylogeny and do not handle gene duplication, LocalAli (Hu & Reinert 2015) and GRAEMLIN (Flannick et al. 2006) do use phylogeny both to refine conserved subnetworks in extant species and to infer ancestral subnetworks, which are important for reconstructing the evolutionary process (Baker et al. 2012, Li & Johnson 2010).

Finally, although studying the impact of selective and neutral forces on network evolution addresses fundamental questions, it is significantly more complicated than studying the impact of these forces on sequence or even expression evolution. This is because very few network reconstructions exist, and those that do are incomplete. Furthermore, organizational properties of networks, such as network modularity or scale-free distributions, although intriguing, could arise through nonadaptive forces such as mutation, recombination, and random drift. An explicit model that accounts for such nonadaptive forces is important to understand the interplay of selection and nonadaptive forces in shaping such topological features (Lynch 2007). In practice, most studies on real systems, rather than simulated data, have examined how events such as gene duplication, often hypothesized to play a major role in evolution (Ohno 1970, Taylor & Raes 2004), give rise to the network observed in a single, extant species (Presser et al. 2008, Teichmann & Babu 2004). An alternate strategy would take network models predefined independently in multiple extant species (Gibson & Goldberg 2009, Patro et al. 2012, Pinney et al. 2007), infer an ancestral network, and then estimate rates of divergence by parameterizing a probabilistic model of network evolution and estimating the model within a maximum likelihood or a Bayesian framework. This model would evolve through edge gain and loss and gene duplication, divergence, and loss. However, as discussed above, the limited availability of reliable network reconstructions in extant species limits the feasibility of this approach at the moment. Good theoretical models of network evolution—ideally with empirical support—could eventually help improve multispecies reconstruction methods, yielding more accurate species-specific networks.

## FUTURE OUTLOOK: HOW DO REGULATORY NETWORK CHANGES ADAPTIVELY IMPACT PHENOTYPE?

Studying the evolution of regulatory networks promises insight into the evolution of complex traits. In this review, we highlighted the major opportunities offered by a comparative perspective to examining regulatory networks, discussed computational challenges, and outlined available methods to systematically study regulatory network evolution. Computational approaches to detect changes at the sequence level, including in both coding and noncoding DNA, have significantly advanced. Going forward, three major challenges remain. They include developing methods that (*a*) integrate comparative regulatory information from multiple molecular levels (e.g., *cis*-regulatory elements, TF binding, chromatin organization, mRNA levels, and protein levels) into network models across species; (*b*) assess selection pressure on network elements and topological properties, and on networks as a whole; and (*c*) predict phenotype from genotype by incorporating network knowledge.

With regard to the first challenge, genomic studies are increasingly collecting integrative data across species, highlighting the opportunity and need for novel computational methods. For example, a recent study (Khan et al. 2013) collected RNA and protein profiles from three mammalian species and found that protein levels are under greater evolutionary constraint than mRNA levels. Another recent study (Wong et al. 2014) tackled the burning need to relate network structure (e.g., TF binding to *cis*-regulatory elements) to function (i.e., impact on expression). Genome-wide binding of three tissue-specific TFs was measured by ChIP-seq, and genome-wide mRNA levels following TF knockout in four related mouse species were measured using RNA-seq. Most cross-species differences in TF binding were not reflected by functional changes in expression, highlighting the critical importance of functional strategies. Moreover, comparative epigenomic data are now increasingly collected in mammals (Xiao et al. 2012, Villar et al. 2015) and other animals (Ho et al. 2014), and yeast (Tsankov et al. 2010, 2011; Tsui et al. 2011), although comparative methods to analyze these data are just emerging (Ho et al. 2014). Finally, as comparative functional genomics studies expand to measure post-translational modifications (Boekhorst et al. 2008, Freschi et al. 2014, Nakagami et al. 2010, Still et al. 2013), integrated approaches that combine multiple types of data, including transcriptomic, proteomic, and epigenomic data, will be important for examining the relative contribution of different modes of regulation to phenotypic divergence.

With regard to the second challenge, computational methods that incorporate specific connectivity and organizational properties of networks are needed to study selection on networks and how these properties relate to adaptive evolution. As in studying selection pressure on sequences and expression, one needs parameterized models to capture network evolution under a neutral model versus under lineage-specific evolution. Furthermore, one must be able to disentangle the various sources of network differences between species, including differences owing to errors in network reconstruction. Such a model could be formulated within a probabilistic framework with parameters to account for each network operation, such as edge insertion and deletion and gene duplication and divergence, and parameters could be estimated for different networks in extant species. An important barrier is obtaining better reconstructions for extant species, which requires collection of comprehensive gene expression data across multiple conditions for multiple species. Network reconstruction algorithms that leverage phylogenies with realistic models of network evolution are also important. Further incorporating a population genetics framework to study the role of different selection forces on the topological and structural properties of a network could be useful. Initial theoretical models have shown that some properties of networks may be governed by phenotypic selection, but intrinsic properties of an organism such as mutational bias, expression cost, and expression dynamics could also influence network structure (Tsuda & Kawata 2010).

With regard to the final challenge, genome-wide association and eQTL studies examine how changes in genotype influence changes in expression and other complex traits within a population. A regulatory network-based predictive model that can predict complex phenotypes, especially those associated with species fitness, can provide insights into how regulatory networks shape the adaptability of species to changing environmental conditions. However, almost all of these studies have examined sequence variation independent of regulatory network structure. As much of the genetic variation associated with complex traits in genome-wide association studies resides in regulatory regions (Maurano et al. 2012, Thurman et al. 2012), these variants likely impact genes through a regulatory network; indeed entire pathways may be targets of variants that act in *trans*. Thus, approaches that link regulatory variants to downstream phenotypes through regulatory network mechanisms will become increasingly important (Califano et al. 2012). One approach toward addressing this problem is based on network information flow. Here, the goal is to link regulatory variants in some genes to downstream effector genes through an underlying network

(Cho et al. 2012). A second strategy is to extend approaches developed to predict phenotypes correlated with fitness from protein sequence variation (Jelier et al. 2011). Although the study cited here focuses on protein coding regions, this approach provides one possible way to predict phenotypes from genomic sequences, especially when considering a small number of strains or species.

In summary, a large number of computational tools have been developed and used to study the evolution of regulatory sequences. These tools help us to study evolutionary dynamics at the level of individual base pairs, single regulatory sites, or larger regulatory elements by finding sequence-specific motifs in poorly annotated species and identifying parts of the genome under accelerated or constrained evolution. Recently, our repertoire of functional genomics data has grown, as such data are more readily and extensively collected. Computational approaches to compare the transcriptome and epigenome are emerging and have been important for systematic annotation and comparison of the noncoding portions of DNA, which constitute a significant fraction of mammalian genomes. We are in the early stages of methods for regulatory network evolution and to link changes in the network to changes in phenotype. As more data sets measuring genome-wide states of the regulation machinery become available, advances in approaches to map networks and model their evolutionary dynamics will become increasingly important for studying selection on regulatory networks and its impact on complex phenotypes.

## DISCLOSURE STATEMENT

A.R. is an SAB member of Thermo Fisher Scientific, a consultant with the Driver Group, and an SAB member of Syros Pharmaceuticals.

## ACKNOWLEDGMENTS

We thank Leslie Gaffney for help with preparation of figures. D.T. is supported by the NIH (R01CA119176-01). A.R. is supported by HHMI, the NIH (R01CA119176-01, P50HG006193), and the Sloan Foundation. S.R. is supported by an NSF CAREER award (NSF DBI: 1350677) and a Sloan Foundation research fellowship.

## LITERATURE CITED

Amit I, Garber M, Chevrier N, Leite AP, Donner Y, et al. 2009. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science* 326(5950):257–63

Anholt RRH, Mackay TFC. 2004. Quantitative genetic analyses of complex behaviours in *Drosophila*. *Nat. Rev. Genet.* 5(11):838–49

Bailey TL. 2008. Discovering sequence motifs. *Methods Mol. Biol.* 452:231–51

Bais AS, Grossmann S, Vingron M. 2007. Simultaneous alignment and annotation of cis-regulatory regions. *Bioinformatics* 23(2):e44–49

Baker CR, Booth LN, Sorrells TR, Johnson AD. 2012. Protein modularity, cooperative binding, and hybrid regulatory states underlie transcriptional network diversification. *Cell* 151(1):80–95

Ballester B, Medina-Rivera A, Schmidt D, Gonzàles-Porta M, Carlucci M, et al. 2014. Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *eLife* 3:e02626

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338(6114):1587–93

Bardet AF, He Q, Zeitlinger J, Stark A. 2012. A computational pipeline for comparative ChIP-seq analyses. *Nat. Protoc.* 7(1):45–61

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* 129(4):823–37
Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, et al. 2013. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* 24(1):14–24
Bergmann S, Ihmels J, Barkai N. 2003. Similarities and differences in genome-wide expression data of six organisms. *PLOS Biol.* 2(1):e9
Blanchette M, Tompa M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* 12(5):739–48
Blanco E, Messeguer X, Smith TF, Guigó R. 2006. Transcription factor map alignment of promoter regions. *PLOS Comput. Biol.* 2(5):e49
Boekhorst J, van Breukelen B, Heck A Jr, Snel B. 2008. Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. *Genome Biol.* 9(10):R144
Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, et al. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299(5611):1391–94
Bonneau R. 2008. Learning biological networks: from modules to dynamics. *Nat. Chem. Biol.* 4(11):658–64
Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, et al. 2007. Divergence of transcription factor binding sites across related yeast species. *Science* 317(5839):815–19
Bradley RK, Li XY, Trapnell C, Davidson S, Pachter L, et al. 2010. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLOS Biol.* 8(3):e1000343
Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478(7369):343–48
Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, et al. 2014. The genomic substrate for adaptive radiation in African cichlid fish. *Nature* 513(7518):375–81
Brem RB. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* 296(5568):752–55
Brem RB. 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *PNAS* 102(5):1572–77
Brem RB, Storey JD, Whittle J, Kruglyak L. 2005. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* 436(7051):701–3
Bryois J, Buil A, Evans DM, Kemp JP, Montgomery SB, et al. 2014. Cis and trans effects of human genomic variants on gene expression. *PLOS Genet.* 10(7):e1004461
Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10:1213–18
Bulyk ML. 2003. Computational prediction of transcription-factor binding site locations. *Genome Biol.* 5(1):201
Butler MA, King AA. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Am. Nat.* 164(6):683–95
Califano A, Butte AJ, Friend S, Ideker T, Schadt E. 2012. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* 44(8):841–47
Chan YF, Marks ME, Jones FC, Villarreal G Jr, Shapiro MS, et al. 2010. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* 327(5963):302–5
Cheng C, Yan KK, Hwang W, Qian J, Bhardwaj N, et al. 2011. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLOS Comput. Biol.* 7(11):e1002190
Cho D-Y, Kim Y-A, Przytycka TM. 2012. Chapter 5: Network biology approach to complex diseases. *PLOS Comput. Biol.* 8(12):e1002820
Chua G, Morris QD, Sopko R, Robinson MD, Ryan O, et al. 2006. Identifying transcription factor functions and targets by phenotypic activation. *PNAS* 103(32):12045–50
Ciofani M, Madar A, Galan C, Sellars M, Mace K, Pauli F, et al. 2012. A validated regulatory network for Th17 cell specification. *Cell* 151(2):289–303
Clark C, Kalita J. 2014. A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics* 30(16):2351–59
Cooper GM, Stone EA, Asimenos G, NISC Comp. Seq. Progr., Green ED, et al. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* 15(7):901–13

Cotney J, Leng J, Yin J, Reilly SK, DeMare LE, et al. 2013. The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell* 154(1):185–96
Daniel E, Newburger MLB. 2009. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* 37(Database issue):D77–82
de Jong H. 2002. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* 9(1):67–103
Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, Thomas WK. 2005. The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nat. Genet.* 37(5):544–48
De Smet R, Marchal K. 2010. Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* 8(10):717–29
Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. 2007. A genome-wide association study of global gene expression. *Nat. Genet.* 39(10):1202–7
Dunn SJ, Martello G, Yordanov B, Emmott S, Smith AG. 2014. Defining an essential transcription factor program for naïve pluripotency. *Science* 344(6188):1156–60
Duque T, Samee MA, Kazemian M, Pham HN, Brodsky MH, Sinha S. 2014. Simulations of enhancer evolution provide mechanistic insights into gene regulation. *Mol. Biol. Evol.* 31(1):184–200
Emerson JJ, Li W-H. 2010. The genetic basis of evolutionary change in gene expression levels. *Philos. Trans. R. Soc. B* 365(1552):2581–90
Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, et al. 2008. Genetics of gene expression and its effect on disease. *Nature* 452(7186):423–28
Eng KH, Bravo HC, Keles S. 2009. A phylogenetic mixture model for the evolution of gene expression. *Mol. Biol. Evol.* 26(10):2363–72
Farh KK-H, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, et al. 2015. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518:337–43
Flannick J, Novak A, Srinivasan BS, McAdams HH, Batzoglou S. 2006. Graemlin: general and robust alignment of multiple large interaction networks. *Genome Res.* 16(9):1169–81
Fraser HB, Babak T, Tsang J, Zhou Y, Zhang B, et al. 2011. Systematic detection of polygenic *cis*-regulatory evolution. *PLOS Genet.* 7(3):e1002023
Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC. 2003. Cross-species sequence comparisons: a review of methods and available resources. *Genome Res.* 13(1):1–12
Freschi L, Osseni M, Landry CR. 2014. Functional divergence and evolutionary turnover in mammalian phosphoproteomes. *PLOS Genet.* 10(1):e1004062
Friedman N. 2004. Inferring cellular networks using probabilistic graphical models. *Science* 303:799–805
Furey TS. 2012. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Sci. Transl. Med.* 13(12):840–52
Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* 25(12):i54–62
Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB. 2004. Conservation and evolution of *cis*-regulatory systems in ascomycete fungi. *PLOS Biol.* 2(12):e398
Genissel A, McIntyre LM, Wayne ML, Nuzhdin SV. 2008. *Cis* and *trans* regulatory effects contribute to natural variation in transcriptome of *Drosophila melanogaster*. *Mol. Biol. Evol.* 25(1):101–10
Gibson TA, Goldberg DS. 2009. Reverse engineering the evolution of protein interaction networks. *Pac. Symp. Biocomput.* 2009:190–202
Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* 17(6):877–85
González S, Montserrat-Sentís B, Sánchez F, Puiggròs M, Blanco E, et al. 2012. ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites. *Bioinformatics* 28(6):763–70
Gordân R, Narlikar L, Hartemink AJ. 2010. Finding regulatory DNA motifs using alignment-free evolutionary conservation information. *Nucleic Acids Res.* 38(6):e90
Göring HHH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, et al. 2007. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Cell Biol.* 39(10):1208–16

Greenfield A, Hafemeister C, Bonneau R. 2013. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics* 29(8):1060–67
Grishkevich V, Ben-Elazar S, Hashimshony T, Schott DH, Hunter CP, Yanai I. 2012. A genomic bias for genotype-environment interactions in *C. elegans*. *Mol. Syst. Biol.* 8(1):587
Habib N, Wapinski I, Margalit H, Regev A, Friedman N. 2012. A functional selection model explains evolutionary robustness despite plasticity in regulatory networks. *Mol. Syst. Biol.* 8(1):619
Haerty W, Ponting CP. 2014. No gene in the genome makes sense except in the light of evolution. *Annu. Rev. Genomics Hum. Genet.* 15:71–92
Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431(7004):99–104
Hardison RC, Taylor J. 2012. Genomic approaches towards finding *cis*-regulatory modules in animals. *Nat. Rev. Genet.* 13(7):469–83
He Q, Bardet AF, Patton B, Purvis J, Johnston J, et al. 2011. High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nat. Genet.* 43(5):414–20
He X, Ling X, Sinha S. 2009. Alignment and prediction of *cis*-regulatory modules based on a probabilistic model of evolution. *PLOS Comput. Biol.* 5(3):e1000299
Ho JWK, Jung YL, Liu T, Alver BH, Lee S, et al. 2014. Comparative analysis of metazoan chromatin organization. *Nature* 512(7515):449–52
Hoffman MM, Birney E. 2010. An effective model for natural selection in promoters. *Genome Res.* 20(5):685–92
Hoheisel JD. 2006. Microarray technology: beyond transcript profiling and genotype analysis. *Nat. Rev. Genet.* 7(3):200–10
Hsu PD, Lander ES, Zhang F. 2014. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* 157(6):1262–78
Hu J, Reinert K. 2015. LocalAli: An evolutionary-based local alignment approach to identify functionally conserved modules in multiple networks. *Bioinformatics* 31(3):363–72
Ichihashi Y, Aguilar-Martínez JA, Farhi M, Chitwood DH, Kumar R, et al. 2014. Evolutionary developmental transcriptomics reveals a gene network module regulating interspecific diversity in plant leaf shape. *PNAS* 111(25):E2616–21
Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3:318–56
Jelier R, Semple JI, Garcia-Verdugo R, Lehner B. 2011. Predicting phenotypic variation in yeast from individual genome sequences. *Nat. Genet.* 43(12):1270–74
Jiménez-Guri E, Huerta-Cepas J, Cozzuto L, Wotton KR, Kang H, et al. 2013. Comparative transcriptomics of early dipteran development. *BMC Genomics* 14(1):123
John S, Sabo PJ, Canfield TK, Lee K, Vong S, et al. 2013. Genome-scale mapping of DNase I hypersensitivity. *Curr. Protocols Mol. Biol.* Chapter 27:Unit 21.27
Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316(5830):1497–502
Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484(7392):55–61
Jordan IK, Mariño-Ramírez L, Koonin EV. 2005. Evolutionary significance of gene expression divergence. *Gene* 345(1):119–26
Kalaev M, Bafna V, Sharan R. 2009. Fast and accurate alignment of multiple protein networks. *J. Comput. Biol.* 16(8):989–99
Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, et al. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* 468(7325):811–14
Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, Ideker T. 2004. PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res.* 32(Web Server issue):W83–88
Kemmeren P, Sameith K, van de Pasch LA, Benschop JJ, Lenstra TL, et al. 2014. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* 157(3):740–52
Khaitovich P, Enard W, Lachmann M, Pääbo S. 2006. Evolution of primate gene expression. *Nat. Rev. Genet.* 7(9):693–702
Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, et al. 2004. A neutral model of transcriptome evolution. *PLOS Biol.* 2(5):e132

Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y. 2013. Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* 342(6162):1100–4
Kheradpour P, Kellis M. 2014. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* 42(5):2976–87
Kim HD, Shay T, O'Shea EK, Regev A. 2009. Transcriptional regulatory circuits: predicting numbers from alphabets. *Science* 325(5939):429–32
Kim J, He X, Sinha S. 2009. Evolution of regulatory sequences in 12 *Drosophila* species. *PLOS Genet.* 5(1):e1000330
King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188(4184):107–16
Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* 39:309–38
Kuo D, Tan K, Zinman G, Ravasi T, Bar-Joseph Z, Ideker T. 2010. Evolutionary divergence in the fungal response to fluconazole revealed by soft clustering. *Genome Biol.* 11(7):R77
Kutter C, Brown GD, Gonçalves A, Wilson MD, Watt S, et al. 2011. Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. *Nat. Genet.* 43(10):948–55
Kuzniar A, van Ham RC, Pongor S, Leunissen JA. 2008. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.* 24(11):539–51
Kvitek DJ, Will JL, Gasch AP. 2008. Variations in stress sensitivity and genomic expression in diverse *S. cerevisiae* isolates. *PLOS Genet.* 4(10):e1000223
Lasky JR, Des Marais DL, Lowry DB, Povolotskaya I, McKay JK, et al. 2014. Natural variation in abiotic stress responsive gene expression and local adaptation to climate in *Arabidopsis thaliana*. *Mol. Biol. Evol.* 31(9):2283–96
Lavoie H, Hogues H, Mallick J, Sellam A, Nantel A, Whiteway M. 2010. Evolutionary tinkering with conserved components of a transcriptional regulatory network. *PLOS Biol.* 8(3):e1000329
Lee MN, Ye C, Villani AC, Raj T, Li W, et al. 2014. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* 343(6175):1246980
Li H, Johnson AD. 2010. Evolution of transcription networks—lessons from yeasts. *Curr. Biol.* 20(17):R746–53
Li X, Battle A, Karczewski KJ, Zappala Z, Knowles DA, et al. 2014. Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. *Am. J. Hum. Genet.* 95(3):245–56
Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, et al. 2006. Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLOS Genet.* 2(12):e222
Liao C-S, Lu K, Baym M, Singh R, Berger B. 2009. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics* 25(12):i253–58
Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478(7370):476–82
Lynch M. 2007. The evolution of genetic networks by non-adaptive processes. *Nat. Rev. Genet.* 8(10):803–13
Maguire SL, Wang C, Holland LM, Brunel F, Neuvéglise C, et al. 2014. Zinc finger transcription factors displaced SREBP proteins as the major sterol regulators during Saccharomycotina evolution. *PLOS Genet.* 10(1):e1004076
Majewski J, Pastinen T. 2011. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.* 27(2):72–79
Marbach D, Roy S, Ay F, Meyer PE, Candeias R, et al. 2012. Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks. *Genome Res.* 22(7):1334–49
Margulies EH, Blanchette M, NISC Comp. Seq. Progr., Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res.* 13(12):2507–18
Markowetz F, Spang R. 2007. Inferring cellular networks—a review. *BMC Bioinformat.* 8(Suppl. 6):S5
Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, et al. 2013. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 42(Database issue):D142–47
Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337(6099):1190–95
McAdams HH, Srinivasan B, Arkin AP. 2004. The evolution of genetic regulatory systems in bacteria. *Nat. Rev. Genet.* 5(3):169–78

Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* 338(6114):1593–99
Mitra K, Carvunis AR, Ramesh SK, Ideker T. 2013. Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* 14(10):719–32
Molinelli EJ, Korkut A, Wang W, Miller ML, Gauthier NP, et al. 2013. Perturbation biology: inferring signaling networks in cellular systems *PLOS Comput. Biol.* 9(12):e1003290
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320(5881):1344–49
Nakagami H, Sugiyama N, Mochida K, Daudi A, Yoshida Y, et al. 2010. Large-scale comparative phosphoproteomics identifies conserved phosphorylation sites in plants. *Plant Physiol.* 153(3):1161–74
Necsulea A, Kaessmann H. 2014. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat. Rev. Genet.* 15(11):734–48
Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, et al. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505(7485):635–40
Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, et al. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489(7414):83–90
Nishiyama A, Xin L, Sharov AA, Thomas M, Mowrer G, et al. 2009. Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors. *Cell Stem Cell* 5(4):420–33
Novershtern N, Regev A, Friedman N. 2011a. Physical module networks: an integrative approach for reconstructing transcription regulation. *Bioinformatics* 27(13):i177–85
Novershtern N, Subramanian A, Lawton LN, Mak RH, Haining WN, et al. 2011b. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* 144(2):296–309
Ohno S. 1970. *Evolution by Gene Duplication*. Berlin: Springer-Verlag
Oleksiak MF, Churchill GA, Crawford DL. 2002. Variation in gene expression within and among natural populations. *Nat. Genet.* 32(2):261–66
Otto W, Stadler PF, López-Giraldéz F, Townsend JP, Lynch VJ, Wagner GP. 2009. Measuring transcription factor–binding site turnover: a maximum likelihood approach using phylogenies. *Genome Biol. Evol.* 1:85–98
Paris M, Kaplan T, Li XY, Villalta JE, Lott SE, Eisen MB. 2013. Extensive divergence of transcription factor binding in *Drosophila* embryos with highly conserved gene expression. *PLOS Genet.* 9(9):e1003748
Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10(10):669–80
Patro R, Kingsford C. 2012. Global network alignment using multiscale spectral signatures. *Bioinformatics* 28(23):3105–14
Patro R, Sefer E, Malin J, Marçais G, Navlakha S, Kingsford C. 2012. Parsimonious reconstruction of network evolution. *Algorithms Mol. Biol.* 7(1):25
Pe'er D, Hacohen N. 2011. Principles and strategies for developing network models in cancer. *Cell* 144(6):864–73
Peirce JL, Lu L, Gu J, Silver LM, Williams RW. 2004. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC Genet.* 5:7
Perry GH, Melsted P, Marioni JC, Wang Y, Bainer R, et al. 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res.* 22(4):602–10
Pinney JW, Amoutzias GD, Rattray M, Robertson DL. 2007. Reconstruction of ancestral protein interaction networks for the bZIP transcription factors. *PNAS* 104(51):20449–53
Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* 21(3):447–55
Presser A, Elowitz MB, Kellis M, Kishony R. 2008. The evolutionary dynamics of the *Saccharomyces cerevisiae* protein interaction network after duplication. *PNAS* 105(3):950–54
Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. *PNAS* 104(Suppl. 1):8605–12
Rockman MV, Kruglyak L. 2006. Genetics of global gene expression. *Nat. Rev. Genet.* 7(11):862–72
Romero IG, Ruvinsky I, Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. *Nat. Rev. Genet.* 13(7):505–16

Roy S, Wapinski I, Pfiffner J, French C, Socha A, et al. 2013. Arboretum: reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Res.* 23(6):1039–50

Sander JD, Joung JK. 2014. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* 32(4):347–55

Schadt EE, Molony C, Chudin E, Hao K, Yang X, et al. 2008. Mapping the genetic architecture of gene expression in human liver. *PLOS Biol.* 6(5):e107

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves A, et al. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* 148(1–2):335–48

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 328(5981):1036–40

Schraiber JG, Mostovoy Y, Hsu TY, Brem RB. 2013. Inferring evolutionary histories of pathway regulation from transcriptional profiling data. *PLOS Comput. Biol.* 9(10):e1003255

Segal E, Pe'er D, Regev A, Koller D, Friedman N. 2005. Learning module networks. *J. Mach. Learn. Res.* 6:557–88

Segal E, Yelensky R, Koller D. 2003. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* 19(Suppl. 1):i273–82

Shay T, Jojic V, Zuk O, Rothamel K, Puyraimond-Zemmour D, et al. 2013. Conservation and divergence in the transcriptional programs of the human and mouse immune systems. *PNAS* 110(8):2946–51

Siddharthan R, Siggia ED, van Nimwegen E. 2005. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLOS Comput. Biol.* 1(7):e67

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15(8):1034–50

Siepel A, Haussler D. 2004. Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.* 11(2–3):413–28

Silver DH, Levin M, Yanai I. 2012. Identifying functional links between genes by evolutionary transcriptomics. *Mol. Biosyst.* 8(10):2585–92

Singh R, Xu J, Berger B. 2008. Global alignment of multiple protein interaction networks with application to functional orthology detection. *PNAS* 105(35):12763–68

Sinha S, Blanchette M, Tompa M. 2004. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinform.* 5(1):170

Skipper M. 2004. The puzzling side of the human genome. *Nat. Rev. Genet.* 5(7):482

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(2):pdb.prot5384

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* 450(7167):219–32

Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, et al. 2013. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* 154(3):530–40

Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, et al. 2014. Conservation of *trans*-acting circuitry during mammalian regulatory evolution. *Nature* 515(7527):365–70

Still AJ, Floyd BJ, Hebert AS, Bingman CA, Carson JJ, et al. 2013. Quantification of mitochondrial acetylation dynamics highlights prominent sites of metabolic regulation. *J. Biol. Chem.* 288(36):26209–19

Stormo GD, Zhao Y. 2010. Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.* 11(11):751–60

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. 2007. Population genomics of human gene expression. *Nat. Genet.* 39(10):1217–24

Stranger BE, Raj T. 2013. Genetics of human gene expression. *Curr. Opin. Genet. Dev.* 23(6):627–34

Tanay A, Gat-Viks I, Shamir R. 2004a. A global view of the selection forces in the evolution of yeast *cis*-regulation. *Genome Res.* 14(5):829–34

Tanay A, Regev A, Shamir R. 2005. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *PNAS* 102(20):7203–8

Tanay A, Sharan R, Kupiec M, Shamir R. 2004b. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *PNAS* 101(9):2981–86

Taylor JS, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.* 38(1):615–43
Teichmann SA, Babu MM. 2004. Gene regulatory network growth by duplication. *Nat. Genet.* 36(5):492–96
Thattai M, van Oudenaarden A. 2001. Intrinsic noise in gene regulatory networks. *PNAS* 98(15):8614–19
Thompson DA, Roy S, Chan M, Styczynsky MP, Pfiffner J, et al. 2013. Evolutionary principles of modular gene regulation in yeasts. *eLife* 2:e00603
Thompson DAA, Regev A. 2009. Fungal regulatory evolution: *cis* and *trans* in the balance. *FEBS Lett.* 583(24):3959–65
Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* 489(7414):75–82
Tirosh I, Barkai N. 2007. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol.* 8(4):R50
Tirosh I, Barkai N. 2011. Inferring regulatory mechanisms from patterns of evolutionary divergence. *Mol. Syst. Biol.* 7:530
Tirosh I, Weinberger A, Carmi M, Barkai N. 2006. A genetic signature of interspecies variations in gene expression. *Nat. Genet.* 38(7):830–34
Tsankov A, Yanagisawa Y, Rhind N, Regev A, Rando OJ. 2011. Evolutionary divergence of intrinsic and *trans*-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization. *Genome Res.* 21(11):1851–62
Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ. 2010. The role of nucleosome positioning in the evolution of gene regulation. *PLOS Biol.* 8(7):e1000414
Tsuda ME, Kawata M. 2010. Evolution of gene regulatory networks by fluctuating selection and intrinsic constraints. *PLOS Comput. Biol.* 6(8):e1000873
Tsui K, Dubuis S, Gebbia M, Morse RH, Barkai N, et al. 2011. Evolution of nucleosome occupancy: conservation of global properties and divergence of gene-specific patterns. *Mol. Cell. Biol.* 31(21):4348–55
Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD. 2008. The evolution of combinatorial gene regulation in fungi. *PLOS Biol.* 6(2):e38
Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, et al. 2014. Mouse regulatory DNA landscapes reveal global principles of *cis*-regulatory evolution. *Science* 346(6212):1007–12
Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, et al. 2015. Enhancer evolution across 20 mammalian species. *Cell* 160(3):554–66
Villar D, Flicek P, Odom DT. 2014. Evolution of transcription factor binding in metazoans—mechanisms and functional implications. *Nat. Rev. Genet.* 15(4):221–33
von Luxburg U. 2007. A tutorial on spectral clustering. *Stat. Comput.* 17(4):395–416
Walker E, Ohishi M, Davey RE, Zhang W, Cassar PA, et al. 2007. Prediction and testing of novel transcriptional networks regulating embryonic stem cell self-renewal and commitment. *Cell Stem Cell* 1(1):71–86
Waltman P, Kacmarczyk T, Bate AR, Kearns DB, Reiss DJ, et al. 2010. Multi-species integrative biclustering. *Genome Biol.* 11(9):R96
Wang Z, Gerstein M, Snyder M. 2009. RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10(1):57–63
Ward LD, Kellis M. 2012a. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337(6102):1675–78
Ward LD, Kellis M. 2012b. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* 40(Database issue):D930–34
Weirauch MT, Hughes TR. 2010. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet.* 26(2):66–74
West MAL, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, et al. 2007. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. *Genetics* 175(3):1441–50
Whitfield TW, Wang J, Collins PJ, Partridge EC, Aldred SF, et al. 2012. Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* 13(9):R50
Wittkopp PJ. 2007. Variable gene expression in eukaryotes: a network perspective. *J. Exp. Biol.* 210(Pt. 9):1567–75

Wittkopp PJ, Haerum BK, Clark AG. 2008. Regulatory changes underlying expression differences within and between *Drosophila* species. *Nat. Genet.* 40(3):346–50

Wohlbach DJ, Kuo A, Sato TK, Potts KM, Salamov AA, et al. 2011. Comparative genomics of xylose-fermenting fungi for enhanced biofuel production. *PNAS* 108(32):13212–17

Wohlbach DJ, Thompson DA, Gasch AP, Regev A. 2009. From elements to modules: regulatory evolution in Ascomycota fungi. *Curr. Opin. Genet. Dev.* 19(6):571–78

Wong ES, Thybert D, Schmitt BM, Stefflova K, Odom DT, Flicek P. 2014. Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. *Genome Res.* 25(2):167–78

Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nat. Rev. Genet.* 8(3):206–16

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, et al. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* 20(9):1377–419

Xiao S, Xie D, Cao X, Yu P, Xing X, et al. 2012. Comparative epigenomic annotation of regulatory DNA. *Cell* 149(6):1381–92

Xie D, Cai J, Chia N-Y, Ng HH, Zhong S. 2008. Cross-species de novo identification of cis-regulatory modules with GibbsModule: application to gene regulation in embryonic stem cells. *Genome Res.* 18:1325–35

Xie D, Chen CC, He X, Cao X, Zhong S. 2011. Towards an evolutionary model of transcription networks. *PLOS Comput. Biol.* 7(6):e1002064

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* 434(7031):338–45

Yanai I, Graur D, Ophir R. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS* 8(1):15–24

Yanai I, Hunter CP. 2009. Comparison of diverse developmental transcriptomes reveals that coexpression of gene neighbors is not evolutionarily conserved. *Genome Res.* 19(12):2214–20

Yanai I, Peshkin L, Jorgensen P, Kirschner MW. 2011. Mapping gene expression in two *Xenopus* species: evolutionary constraints and developmental flexibility. *Dev. Cell* 20(4):483–96

Yáñez-Cuna JO, Kvon EZ, Stark A. 2013. Deciphering the transcriptional *cis*-regulatory code. *Trends Genet.* 29(1):11–22

Ye CJ, Feng T, Kwon HK, Raj T, Wilson MT, et al. 2014. Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* 345(6202):1254665

Zarrineh P, Fierro AC, Sánchez-Rodríguez A, De Moor B, Engelen K, Marchal K. 2011. COMODO: an adaptive coclustering strategy to identify conserved coexpression modules between organisms. *Nucleic Acids Res.* 39(7):e41

Zhang X, Moret BME. 2008. Boosting the performance of inference algorithms for transcriptional regulatory networks using a phylogenetic approach. In *Algorithms in Bioinformatics*, ed. KA Crandall, J Lagergren, pp. 245–58. Berlin: Springer

Zhang X, Moret BME. 2012. Refining regulatory networks through phylogenetic transfer of information. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9(4):1032–45

Zheng W, Gianoulis TA, Karczewski KJ, Zhao H, Snyder M. 2011. Regulatory variation within and between species. *Annu. Rev. Genomics Hum. Genet.* 12(1):327–46

Zheng W, Zhao H. 2013. Studying the evolution of transcription factor binding events using multi-species ChIP-Seq data. *Stat. Appl. Genet. Mol. Biol.* 12(1):1–15