Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Systems Biology

Network-based approaches that exploit inferred transcription factor activity to analyze the impact of genetic variation on gene expression

Harmen J. Bussemaker[1,2], Helen C. Causton[3], Mina Fazlollahi[4], Eunjee Lee[4] and Ivor Muroff[1]

Abstract

Over the past decade, a number of methods have emerged for inferring protein-level transcription factor activities in individual samples based on prior information about the structure of the gene regulatory network. We discuss how this has enabled new approaches for dissecting trans-acting mechanisms that underpin genetic variation in gene expression.

Addresses

[1] Department of Biological Sciences, Columbia University, New York, NY 10027, USA
[2] Department of Systems Biology, Columbia University, New York, NY 10032, USA
[3] Department of Pathology and Cell Biology, Columbia University Medical Center, New York, NY 10032, USA
[4] Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029, USA

Corresponding author: Bussemaker, Harmen J (hjb2004@columbia.edu)

Current Opinion in Systems Biology 2017, **2**:98–102

This review comes from a themed issue on **Regulatory and metabolic networks (2017)**

Edited by **Sang Yup Lee** and **Christina Leslie**

For a complete overview see the Issue and the Editorial

Available online 17 April 2017

http://dx.doi.org/10.1016/j.coisb.2017.04.002

2452-3100/© 2017 Elsevier Ltd. All rights reserved.

Keywords

Gene regulatory networks, Transcription factor activity, Systems genetics, QTL mapping, aQTL, cQTL.

Introduction

The structural and functional unit of life is the cell. Thanks to remarkable recent advances in genomics, cells have been probed at many levels (genome sequence, mRNA transcript abundance, epigenetic modification, *etc.*) and in a variety of genetic and physiological contexts. The Roadmap Epigenomics Consortium (roadmapepigenomics.org) has generated a large collection of functional genomics data for primary cells and tissues, including histone modification patterns, DNA accessibility, DNA methylation, and RNA expression [1]. In a separate initiative, the Genome-Tissue Expression (GTEx) project (gtexportal.org) has collected parallel genotype and RNA-seq data for dozens of tissues across hundreds of human individuals [2]. The Cancer Genome Atlas (TCGA) project (cancergenome.nih.gov) provides a similar resource for a variety of human tumors [3]. These rich compendia provide new opportunities for dissecting the genetic and molecular mechanisms that underlie cellular behavior. To derive actionable knowledge from them, however, new integrative computational methods are needed.

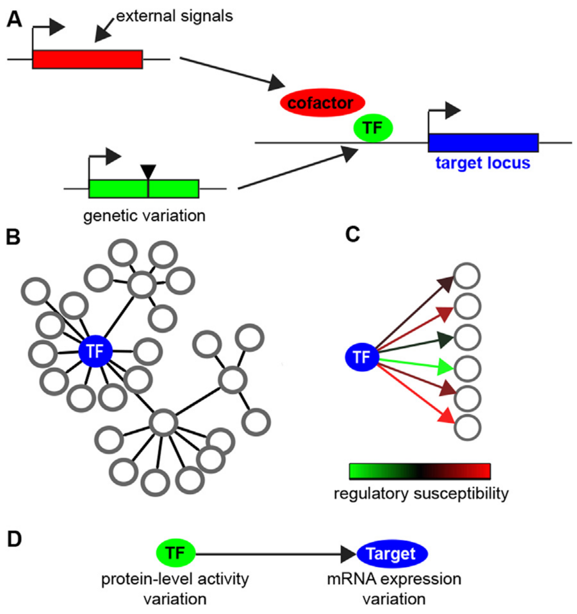
The genome sequence of individual organisms from the same species varies naturally within a population, as new combinations of alleles are created in each generation by genetic recombination between parental genomes during meiosis. As a consequence of this variation, myriad measurable traits differ between individuals; these can include classic variables such as the rate of cell division, but also intermediate phenotypic variables such as transcript abundances for all genes. A more subtle kind of quantitative trait is the cell's sensitivity to changes in external parameters, such as responsiveness to a particular drug. In general, the internal state of the cell may vary along any of a large number of regulatory dimensions or "pathways." It is important to establish which natural variables are the most useful for quantifying cellular state.

Protein-level regulatory activity of transcription factors as a hidden variable

Transcription factors (TF) are proteins that bind to DNA with an affinity that depends on the base sequence at the protein-DNA interface. Defining the binding specificity of TFs in the form of "motifs" or "weight matrices" [4] has received much attention in the literature. A number of high-throughput assays for probing *in vitro* protein-DNA interactions have been developed, based on DNA microarray [5,6] or deep sequencing [7–10] technology. The development of computational methods for building accurate sequence-to-affinity models from such data [11] continues to be an active area of research.

However, DNA binding specificity is only one aspect of TF function. It is equally important to know to what extent the nuclear presence of a TF contributes to the mRNA expression level of its target genes in any given cellular state. This state can be influenced by genetic variation among individuals, by the epigenetic state of the cell (*i.e.*, it is different between cell types, tissues or energetic states), and by external agents (*e.g.*, hormones or drugs). Over the past decade and a half, a number of computational methods have been developed that are capable of inferring the protein-level regulatory activity of a transcription factor from the mRNA expression levels of its known or predicted target genes [12]. This activity can be thought of as the net global effect on gene expression of active TF protein in the nucleus. It is usually quantified as a differential activity relative to a reference sample. Regulation of the subcellular distribution of the TF (nuclear vs. cytoplasmic) through post-translational modification is a major mode of TF activity modulation: the genome resides in the nucleus, and therefore cytoplasmic TF protein cannot contribute to transcriptional activation or repression. Cell-type-specific co-factors can also modulate the ability of a TF to influence gene expression (Figure 1A).

Figure 1


A
external signals
cofactor
TF
target locus
genetic variation
B
TF
C
TF
regulatory susceptibility
D
TF
protein-level activity variation
Target
mRNA expression variation

A network-level perspective on genetic variation in gene expression. (**A**) Genetic polymorphisms can affect the protein-level activity of a transcription factor (TF) or the modulating effect of a co-factor. This in turn affects the mRNA expression level of target genes. (**B**) Prior information about target gene sets (regulons) can be used to infer TF activity. (**C**) Alternatively, prior information about TF-target connectivity can be used for the same purpose. (**D**) By considering the relationship between the activity of a TF and the mRNA expression level of a putative target gene, the strength of their functional connection (*i.e.*, the “susceptibility” or “responsiveness” of the gene to changes in the activity of the TF) can be estimated.

The simplest methods for inferring TF activity are based on prior information on the set of regulatory targets, sometimes referred to as the “regulon”, of a given TF (Figure 1B). The “T-profiler” method [13], introduced over a decade ago, uses a standard statistical test (Student’s t-test) to analyze the extent to which the mean expression of the genes that belong to the regulon, and those that do not, differs; in cases where a normal distribution of the expression values cannot be assumed, a non-parametric variant of the t-test (Wilcoxon–Mann–Whitney test) can be used. “Gene Set Enrichment Analysis” (GSEA) [14], which relies on an *ad hoc* statistic inspired by the Kolmogorov–Smirnov test and requires sampling from an empirical null distribution to assess statistical significance, has also been widely used to analyze differential expression of the level of gene sets. The regulons used in these analyses can be defined based on a variety of criteria: such as the occurrence of TF binding motifs near the gene, or evidence of *in vivo* occupancy from chromatin-immunoprecipitation (ChIP) assays [15]. In a related approach, the correlation structure of transcript abundance across a large number of conditions is analyzed in order to define regulons [16], which can subsequently be used to predict TF activity using the GSEA statistic [17]. Approaches that analyze the mutual information between motif-based and expression-based gene sets have also been used to infer TF activity [18].

In an alternative approach to inferring TF activity, the degree to which a gene is expected to respond to a change in the activity of the TF is treated as a continuous variable, ranging from strong targets that have high-affinity binding sites in their upstream promoter region to weak targets with low-affinity sites. This motivates an approach based on linear regression (Figure 1C), in which the independent variable (the “predictor” or x-variable) is constructed by using a consensus motif or scoring matrix to scan the cis-regulatory sequence of each gene, while the differential expression value for the gene plays the role of dependent variable (the “response” or y-variable). The regression coefficient or “slope” can be interpreted as the differential TF activity. The earliest implementation of this idea was the REDUCE algorithm, which used the number of matches to a simple DNA binding consensus motif as a predictor [19]. Subsequent studies showed how the motif-based predictor could be improved by considering the expression response of each gene to deletion of the TF [20], by using weight matrices to quantify binding affinity [21], or by considering the pattern of conservation between orthologous promoter regions [22]. Other extensions of this regression-based approach have also been described [23,24].

Transcription factor activity as a genetic trait

The ability to infer TF activity in individual samples provides a unique opportunity for understanding how genetic variation can drive phenotypic variation via trans-acting mechanisms (Figure 1A). Lee et al. [25] introduced the concept of "aQTLs" in which the inferred TF activity is treated as a quantitative trait. In their approach, linear regression, across all genes, of genotype-specific expression changes on total promoter affinities is first used to infer the activity of a large number TFs for which the DNA binding specificity is known in the form of a scoring matrix [25]. Genetic polymorphisms that modulate TF activity are subsequently mapped using similar methods as for mapping quantitative trait loci (eQTLs) that affect the mRNA expression level of an individual gene [26].

Locus Expression Signature Analysis (LESA; [27]) is a variant of the aQTL approach. In this study, genome-wide signatures that quantify the gene expression response to genetic perturbation at a particular locus were constructed for a panel of mouse tumors created through viral insertional mutagenesis, as well as from TCGA data. These signatures were then further characterized in order to link genetic loci to the TFs whose activity they perturb, or to identify therapeutic locus-drug combinations [27]. Empirically defined regulons have also been used to infer TF activity across TCGA samples and analyze the regulatory connectivity between TFs and their upstream signaling pathways [28].

DNA-binding proteins are not the only trans-acting factors involved in regulating gene expression. Control of transcript stability by RNA binding regulatory factors is also important [18,29–32]. An early study used eQTL data to analyze post-transcriptional regulation by exploiting modules of co-expressed genes [33]. More recently, the methodology for mapping aQTLs was adapted to identify genetic polymorphisms that modulate the activity of RNA binding proteins acting as post-transcriptional regulators of gene expression [34].

Regulatory network connectivity as a genetic trait

Thus far, we have touched upon two distinct views on how genetic variation can impact gene expression. The first and most widely explored considers *cis*-acting polymorphisms. For instance, a single-nucleotide polymorphism (SNP) in a TF binding site can change the affinity for the TF, and thereby influence TF occupancy [35–37]. A number of methods aim to explain the impact of SNPs on local chromatin accessibility in terms of their effect on TF binding [11,38–41]. Variation in the amino-acid sequence of the TFs has recently been shown to be more common than previously assumed, and can have a significant impact on DNA binding specificity [42]. More subtle regulatory traits such as the degree to which transcriptional initiation is spread over multiple positions within a particular promoter have also been shown to have a genetic component [43]. In the second view, *trans*-acting loci exert their effect along a causal path that begins at the polymorphism, continues via signal transduction pathways upstream of the TF, affects the protein-level activity of the TF, and finally proceeds downstream to affect the transcript abundance of target genes.

The techniques outlined above may be exploited further to analyze transcriptional network structure, by analyzing the relationship between the inferred activity of a TF and the mRNA expression level of its putative targets across a large number of cell states. Inspired by techniques for refining TF-target networks [44,45], Gao et al. [46] analyzed the extent to which TF binding to a gene's promoter implies that the gene's expression level is regulated by the TF. By performing regression of the mRNA level of each putative target gene on inferred TF activity, the sensitivity of the gene to changes in TF activity (its regulatory "susceptibility") was estimated from the data [31,46]. Alternative approaches have also been used to distinguish functional from non-functional TF binding sites [47].

The degree to which a gene responds to a given TF is context-specific; for instance, it can depend on co-factors that modulate the subcellular distribution of a TF or its interactions with its DNA binding sites. An early attempt to map modulators of regulatory connectivity was based on expression data alone [48]; however, the need to avoid confounding between TF-target and TF-modulator correlation posed a limitation. The emergence of parallel genotype and expression data across populations provided a new opportunity to pursue this line of analysis, and allowed network connectivity QTLs (or "cQTLs") to be mapped [49]. This type of analysis led to the prediction that activation of the gene expression response to mating pheromone by the transcription factor Ste12p in the yeast *Saccharomyces cerevisiae* was modulated by a non-synonymous single-nucleotide polymorphism (SNP) in the co-factor Dig2, which was validated by reporter gene experiments on allele replacements strains [49]. A similar approach has been used in a human cancer context [28]. These successful case studies illustrate how genetic factors that modulate the response of a cell to external perturbation (mating pheromone in yeast, or a drug in human) may be systematically uncovered.

Conclusion

A conceptual framework now exists for analyzing variation in genome expression from the perspective of trans-acting pathways. The different classes of approaches that have been proposed so far each have their own limitations. Those based on empirical prior definition of sets of TF targets or "regulons" rely on observation of the behavior of cells to define the connectivity of the gene regulatory networks. They require dedicated compendia of functional genomics and are therefore relatively expensive. Alternative approaches that explicitly rely on interpretation of the non-coding genome sequence in terms of TF binding affinities have so far mostly been limited to model organisms. They are intrinsically more difficult to implement, as they essentially require one to predict expression from sequence, a notoriously hard problem. However, with the increasing availability of information on the *in vitro* DNA binding specificity of human transcription factors, as well as cell-type specific information about chromatin context, they are now poised to become mainstream tools for elucidating the molecular mechanisms through which sequence variation drives changes in gene expression in normal and diseased human tissue.

Funding

This research was funded by National Institutes of Health grant R01HG003008.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest

1. Roadmap Epigenomics Consortium: **Integrative analysis of 111 reference human epigenomes**. *Nature* 2015, **518**:317–330.

2. GTEx Consortium: **The genotype-tissue expression (GTEx) project**. *Nat Genet* 2013, **45**:580–585.

3. Tomczak K, Czerwinska P, Wiznerowicz M: **The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge**. *Contemp Oncol (Pozn)* 2015, **19**:A68–A77.

4. Stormo GD: **DNA binding sites: representation and discovery**. *Bioinformatics* 2000, **16**:16–23.

5. Berger MF, *et al.*: **Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities**. *Nat Biotechnol* 2006, **24**:1429–1435.

6. Warren CL, *et al.*: **Defining the sequence-recognition profile of DNA-binding molecules**. *Proc Natl Acad Sci USA* 2006, **103**: 867–872.

7. Stormo GD, Zhao Y: **Determining the specificity of protein-DNA interactions**. *Nat Rev Genet* 2010, **11**:751–760.

8. Jolma A, *et al.*: **Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities**. *Genome Res* 2010, **20**:861–873.

9. Slattery M, *et al.*: **Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins**. *Cell* 2011, **147**: 1270–1282.

10. * Isakova A, *et al.*: **SMiLE-seq identifies binding motifs of single and dimeric transcription factors**. *Nat Methods* 2017, **351**: 1450–1454.

Using a versatile combination of microfluidics technology and massively parallel sequencing, the authors generate *in vitro* transcription factor binding specificity data of exceptional quality.

11. Weirauch MT, *et al.*: **Evaluation of methods for modeling transcription factor sequence specificity**. *Nat Biotechnol* 2013, **31**:126–134.

12. Bussemaker HJ, Foat BC, Ward LD: **Predictive modeling of genome-wide mRNA expression: from modules to molecules**. *Annu Rev Biophys Biomol Struct* 2007, **36**:329–347.

13. Boorsma A, *et al.*: **T-profiler: scoring the activity of predefined groups of genes using gene expression data**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W592–W595.

14. Subramanian A, *et al.*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci USA* 2005, **102**: 15545–15550.

15. Boorsma A, *et al.*: **Inferring condition-specific modulation of transcription factor activity in yeast through regulon-based analysis of genomewide expression**. *PLoS One* 2008, **3**:e3112.

16. Basso K, *et al.*: **Reverse engineering of regulatory networks in human B cells**. *Nat Genet* 2005, **37**:382–390.

17. Alvarez MJ, *et al.*: **Functional characterization of somatic mutations in cancer using network-based inference of protein activity**. *Nat Genet* 2016, **48**:838–847.

18. Elemento O, Slonim N, Tavazoie S: **A universal framework for regulatory element discovery across all genomes and data types**. *Mol Cell* 2007, **28**:337–350.

19. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression**. *Nat Genet* 2001, **27**: 167–171.

20. Wang W, *et al.*: **A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae***. *Proc Natl Acad Sci USA* 2002, **99**:16893–16898.

21. Foat BC, Morozov AV, Bussemaker HJ: **Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE**. *Bioinformatics* 2006, **22**:e141–e149.

22. Ward LD, Bussemaker HJ: **Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences**. *Bioinformatics* 2008, **24**:i165–i171.

23. Balwierz PJ, *et al.*: **ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs**. *Genome Res* 2014, **24**:869–884.

24. Osmanbeyoglu HU, *et al.*: **Linking signaling pathways to transcriptional programs in breast cancer**. *Genome Res* 2014, **24**: 1869–1880.

25. Lee E, Bussemaker HJ: **Identifying the genetic determinants of transcription factor activity**. *Mol Syst Biol* 2010, **6**:412.

26. Brem RB, *et al.*: **Genetic dissection of transcriptional regulation in budding yeast**. *Science* 2002, **296**:752–755.

27. Lee E, *et al.*: **Identifying regulatory mechanisms underlying tumorigenesis using locus expression signature analysis**. *Proc Natl Acad Sci USA* 2014, **111**:5747–5752.

28. Chen JC, *et al.*: **Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks**. *Cell* 2014, **159**:402–414.

29. Keene JD: **RNA regulons: coordination of post-transcriptional events**. *Nat Rev Genet* 2007, **8**:533–543.

30. Gerber AP, Herschlag D, Brown PO: **Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast**. *PLoS Biol* 2004, **2**:E79.

31. Foat BC, *et al.*: **Profiling condition-specific, genome-wide regulation of mRNA stability in yeast**. *Proc Natl Acad Sci USA* 2005, **102**:17675–17680.

32. Lee E, *et al.*: **Inferred miRNA activity identifies miRNA-mediated regulatory networks underlying multiple cancers**. *Bioinformatics* 2016, **32**:96–105.

33. Lee SI, *et al.*: **Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification**. *Proc Natl Acad Sci USA* 2006, **103**:14062–14067.

34. Fazlollahi M, *et al.*: **Harnessing natural sequence variation to dissect posttranscriptional regulatory networks in yeast**. *G3 (Bethesda)* 2014, **4**:1539–1553.

35. Borneman AR, *et al.*: **Divergence of transcription factor binding sites across related yeast species**. *Science* 2007, **317**: 815–819.

36. Deplancke B, Alpern D, Gardeux V: **The genetics of transcription factor DNA binding variation**. *Cell* 2016, **166**:538–554.

37. Reddy TE, *et al.*: **Effects of sequence variation on differential allelic transcription factor occupancy and gene expression**. *Genome Res* 2012, **22**:860–869.

38. Kelley DR, Snoek J, Rinn JL: **Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks**. *Genome Res* 2016, **26**:990–999.

39. Zhou J, Troyanskaya OG: **Predicting effects of noncoding variants with deep learning-based sequence model**. *Nat Methods* 2015, **12**:931–934.

40. Lee D, *et al.*: **A method to predict the impact of regulatory variants from DNA sequence**. *Nat Genet* 2015, **47**:955–961.

41. Varshney A, *et al.*: **Genetic regulatory signatures underlying islet gene expression and type 2 diabetes**. *Proc Natl Acad Sci USA* 2017, **114**:2301–2306.

42. * Barrera LA, *et al.*: **Survey of variation in human transcription factors reveals prevalent DNA binding changes**. *Science* 2016, **351**:1450–1454.

Using computational structural analysis and protein binding microarray (PBM) technology, the authors show that most human individuals carry transcription factor variants with altered DNA binding specificity.

43. Schor IE, *et al.*: **Promoter shape varies across populations and affects promoter evolution and expression noise**. *Nat Genet* 2017, **49**:550–558.

44. Ihmels J, *et al.*: **Revealing modular organization in the yeast transcriptional network**. *Nat Genet* 2002, **31**:370–377.

45. Liao JC, *et al.*: **Network component analysis: reconstruction of regulatory signals in biological systems**. *Proc Natl Acad Sci USA* 2003, **100**:15522–15527.

46. Gao F, Foat BC, Bussemaker HJ: **Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data**. *BMC Bioinformatics* 2004, **5**:31.

47. Cusanovich DA, *et al.*: **The functional consequences of variation in transcription factor binding**. *PLoS Genet* 2014, **10**:e1004226.

48. Wang K, *et al.*: **Genome-wide identification of post-translational modulators of transcription factor activity in human B cells**. *Nat Biotechnol* 2009, **27**:829–839.

49. * Fazlollahi M, *et al.*: **Identifying genetic modulators of the connectivity between transcription factors and their transcriptional targets**. *Proc Natl Acad Sci USA* 2016, **113**:E1835–E1843.

Using a systems genetics approach based on inferred transcription factor activity, the authors identify and validate polymorphisms in co-factors that modulate how sensitive the cell is to external signals.