CHAPTER 10

Gene-Centered Regulatory Network Mapping

Albertha J.M. Walhout

Program in Gene Function and Expression and Program in Molecular Medicine, University of Massachusetts Medical School, USA

Abstract

The *Caenorhabditis elegans* hermaphrodite is a complex multicellular animal that is composed of 959 somatic cells. The *C. elegans* genome contains ~20,000 protein-coding genes, 940 of which encode regulatory transcription factors (TFs). In addition, the worm genome encodes more than 100 microRNAs and many other regulatory RNA and protein molecules. Most *C. elegans* genes are subject to regulatory control, most likely by multiple regulators, and combined, this dictates the activation or repression of the gene and corresponding protein in the relevant cells and under the appropriate conditions. A major goal in *C. elegans* research is to determine the spatiotemporal expression pattern of each gene throughout development and in response to different signals, and to determine how this expression pattern is accomplished. Gene regulatory networks describe physical and/or functional interactions between genes and their regulators that result in specific spatiotemporal gene expression. Such regulators can act at transcriptional or post-transcriptional levels. Here, I

METHODS IN CELL BIOLOGY, VOL 106

0091-679X/10 $35.00
DOI 10.1016/B978-0-12-544172-8.00010-4

will discuss the methods that can be used to delineate gene regulatory networks in *C. elegans.* I will mostly focus on gene-centered yeast one-hybrid (Y1H) assays that are used to map interactions between non-coding genic regions, such as promoters, and regulatory TFs. The approaches discussed here are not only relevant to *C. elegans* biology, but can also be applied to other model organisms and humans.

I. Introduction

Complex multicellular model organisms such as *C. elegans* need to faithfully develop from a fertilized oocyte into a complete and fully functioning animal that is composed of different cell and tissue types. After development is completed, metazoan organisms also need mechanisms for homeostasis and to adequately respond to physiological and environmental cues, in order to find mating partners, to detect food, and to avoid pathogens. For correct functionality, cells and tissues need to compute an appropriate biological output based on the input they receive. Such an output can, for instance, be to differentiate, to move, or to enter the dauer stage. Biological outputs result from interactions between the different biomolecules cells and tissues contain, including the genome, proteins and RNA molecules as well as small molecules such as metabolites.

Developmental and post-developmental processes are controlled, at least in part, by the specific spatiotemporal expression of each of the ~20,000 protein-coding genes in the *C. elegans* genome. Each gene/protein is likely controlled by multiple regulators and at multiple levels (Figs. 1 and 2). First, genes are transcribed into

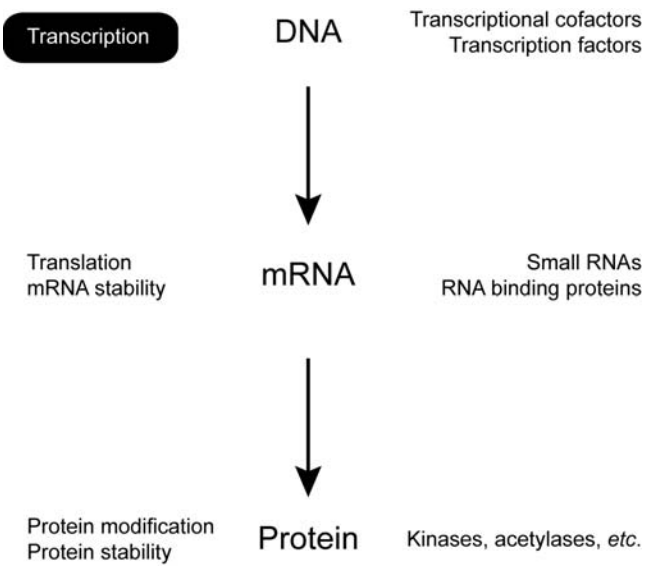

Fig. 1 The different levels of differential gene expression. This review focuses mainly on the transcription, and to a lesser extent, on microRNAs.

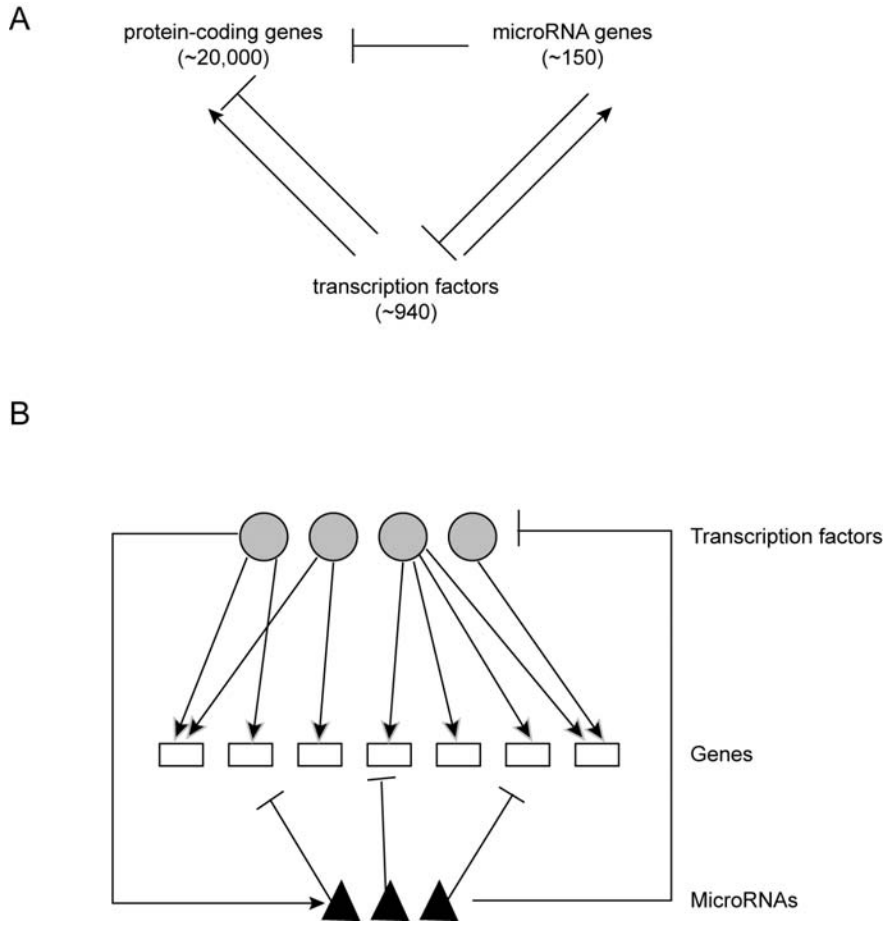

Fig. 2 A) The first levels of gene regulation of the ~20,000 protein-coding genes are controlled transcriptionally by ~940 TFs that can either activate or repress transcription and post-transcriptionally by ~150 microRNAs that repress mRNA translation and/or stability. B) Cartoon of a gene regulatory network involving genes, TFs, and microRNAs.

mRNAs, and this is controlled by the action of regulatory transcription factors (TFs) that can repress or activate gene expression by directly interacting with the genome. Second, mRNA stability and translation are controlled by small RNAs such as microRNAs, and by RNA binding proteins that frequently interact with the 3′UTR of their target mRNAs. Third, after translation, proteins can be stabilized or destabilized due to post-translational modifications by, for example, kinases or acetylases. Finally, sub-cellular mRNA and protein localization can be subject to control mechanisms as well.

The delineation of the complex networks that comprehensively describe the physical and regulatory interactions at each of these levels and between all

biomolecules is a daunting task. Here, I will focus specifically on *C. elegans* gene regulatory networks that control gene expression at the transcriptional and post-transcriptional levels. I will briefly discuss the methods that can be used to identify the players in gene regulatory networks, as well as approaches to identify interactions between them, with a primary focus on gene-centered yeast one-hybrid (Y1H) assays that are used to identify interactions between non-coding regulatory DNA regions and TFs.

II. Gene Regulatory Networks

Gene regulatory networks are composed of two main components: nodes and edges. The network nodes are the players involved, that is, the genes and their regulators. The edges are the physical and/or regulatory relationships between the nodes (Fig. 2B). Gene regulatory networks are different from better-known protein–protein interaction networks, because gene regulatory networks are both bipartite and directional. They are bipartite because there are two types of nodes: genes and regulators, although of course some genes are themselves regulators of other genes or proteins. Gene regulatory networks are directional because regulators control genes and usually not the other way around. In order to map and characterize gene regulatory networks, one needs to first identify the nodes. For the genes this means to identify the non-coding genomic DNA sequences that participate in the control of gene expression, and for the regulators this means to identify which protein-coding genes encode TFs, RNA binding proteins, and other regulators, as well as to determine the complete collection of regulatory RNA molecules. Here, I will mostly focus on TFs and microRNAs, and the types of genic regions they interact with.

III. Identifying Gene Regulatory Network Nodes

A. Regulatory Regions

Different parts of a gene can contribute to its regulation. The more complex an organism, the more complex its gene regulation is. In *C. elegans* there are two main regulatory regions: gene promoters in the genome and 3′UTRs in mRNAs.

1. Promoters

A gene promoter is the genomic DNA sequence immediately upstream of the transcription start site. Generally, promoters are composed of a basal element where the general transcriptional machinery binds (e.g., RNA polymerase II and general TFs), and the proximal gene promoter that serves as a landing site for regulatory TFs. Since the majority of *C. elegans* genes are subject to trans-splicing, precise transcription start sites have not been determined for most genes. However, 5′UTRs are

short compared to more complex organisms such as humans, and for practical purposes, promoters can therefore be defined as the region immediately upstream of the translational start site. It is difficult to determine the 5′ start point of gene promoters. However, since most intergenic regions are shorter than 2 kb, most studies have limited their analyses to this length (Deplancke *et al.*, 2004; Dupuy *et al.*, 2004; Hunt-Newbury *et al.*, 2007). Importantly, it has been shown that this region, when fused to a reporter gene such as that encoding the green fluorescent protein (GFP) often drives gene expression in a manner that recapitulates the expression of the endogenous gene (Dupuy *et al.*, 2004; Grove *et al.*, 2009; Hunt-Newbury *et al.*, 2007; Martinez *et al.*, 2008b; Reece-Hoyes *et al.*, 2007).

To facilitate the system-level analysis of gene expression, a clone resource comprised of ∼6000 *C. elegans* promoters, referred to as the Promoterome, has been generated (Dupuy *et al.*, 2004). This resource is based on the Gateway cloning system and consists of promoter Entry clones that can be easily transferred to various Destination vectors by a simple recombination reaction (Hartley *et al.*, 2000; Walhout *et al.*, 2000b). Destination vectors that can be used to analyze gene regulatory networks include a GFP vector for the creation of transgenic animals to study promoter activity *in vivo*, and Y1H vectors for the identification of TFs that can interact with the promoter (see below). So far, systematic efforts have determined the *in vivo* activity of ∼350 TF-encoding gene promoters (Grove *et al.*, 2009; Reece-Hoyes *et al.*, 2007), ∼1800 additional gene promoters (Hunt-Newbury *et al.*, 2007), and 73 microRNA gene promoters (Martinez *et al.*, 2008b). Many of the corresponding transgenic lines are available to the community through the *C. elegans* genetics center (CGC).

2. 3′ UTRs

The 3′UTR is the untranslated region in the mRNA, immediately downstream of the stop codon. This region is subject to post-transcriptional control by microRNAs and RNA binding proteins. Recently, a comprehensive collection of 3′UTRs has been delineated for most *C. elegans* genes (Mangone *et al.*, 2010). Cloning these 3′UTRs into Gateway-compatible vectors will provide a resource for experimental gene regulatory network mapping that is similar to the ORFeome (see below) and Promoterome resources.

3. Other Genic Regulatory Regions

It is not clear to what extent other regulatory regions function in gene regulatory networks in *C. elegans*. So far, transcriptional studies have mostly focused on promoters. However, it is clear that other regions, such as introns and sequences downstream of the gene, can also play a role. Similarly, microRNAs and RNA binding proteins could target regions outside 3′UTRs within their mRNA targets. Systematic studies are required to elucidate the relative role different genic regions play in complex gene regulatory networks.

B. Regulators

1. Transcription Factors

TFs provide the first level of gene control. They bind directly to DNA through their sequence-specific DNA binding domain and can be grouped into families based on the type of DNA binding domain they possess (Reece-Hoyes *et al.*, 2005). Well-known DNA binding domains include the homeodomain, the basic helix-loop-helix (bHLH) domain, C2H2 zinc fingers, the ETS domain, the bZIP domain, and C4-type zinc fingers found in nuclear hormone receptors (NHRs). TFs can be predicted in a genome of interest by searching the complete collection of proteins for the presence of a known DNA binding domain. This is usually done by computational methods, for instance using Interpro (Mulder *et al.*, 2003) or SMART (Letunic *et al.*, 2004) databases. However, we have found that visual inspection of predicted DNA binding domains using knowledge of their sequence and structure is highly useful as well. Indeed, by doing so we increased the predicted set of *C. elegans* TFs from ~600 (Ruvkun and Hobert, 1998) to 940, or ~5% of all protein-coding genes (Reece-Hoyes *et al.*, 2005; Vermeirssen *et al.*, 2007b). Most *C. elegans* TF-encoding genes encode a single splice variant; however in some cases multiple variants are present, and some of these may encode proteins with different DNA binding domains (Reece-Hoyes *et al.*, 2005). Interestingly, different TF variants can have different biological functions. For instance, different variants of the forkhead protein DAF-16 were recently found to be expressed in distinct patterns and to confer different functions related to metabolism and aging (Kwon *et al.*, 2010). Several proteins have been identified that can bind *C. elegans* gene promoters but that do not possess a known DNA binding domain (Deplancke *et al.*, 2006a; Vermeirssen *et al.*, 2007a). Thus, the total collection of *C. elegans* TFs may be slightly larger, but is likely not to exceed 1000 (unpublished data).

More than 12,000 *C. elegans* full-length open reading frames (ORFs) have been cloned into a Gateway-compatible resource called the ORFeome (Lamesch *et al.*, 2004; Reboul *et al.*, 2003). We obtained the TF-encoding ORFs from this resource and supplemented that with TF-encoding ORFs that we cloned *ab initio* (Deplancke *et al.*, 2004; Vermeirssen *et al.*, 2007b). The resulting clone collection currently contains ~90% of all full-length TFs and can be directly used in assays for the delineation of gene regulatory networks such as Y1H assays (unpublished data, see below).

2. MicroRNAs

MicroRNAs regulate gene expression post-transcriptionally by sequence-specific but imperfect basepairing with the 3′UTR of their target mRNAs. It has been estimated that the *C. elegans* genome encodes more than 110 microRNAs (Lehrbach and Miska, 2008). Some of these have been identified genetically (e.g., *lin-4, let-7*), some have been predicted computationally (Lim *et al.*, 2003), and others were more recently found by deep sequencing small RNA populations purified from

worms (Friedlander *et al.*, 2008; Kato *et al.*, 2009). As with TFs, microRNAs can also be grouped into families, based on their seed sequence, the part with which they basepair with their target genes. It is not yet clear whether all *C. elegans* microRNAs have been identified. Indeed, it may be that additional microRNAs will be uncovered when the animal is exposed to particular conditions, in males or dauers, or when sequencing techniques further improve to detect microRNAs of very low abundance.

3. Other Regulators

In addition to TFs and microRNAs, other RNA and protein molecules contribute to differential gene regulation. These include RNA binding proteins, transcriptional co-factors, and signaling molecules such as kinases and phosphatases, as well as endogenous siRNAs and, perhaps, long non-coding RNAs. Systematic computational and experimental analyses will shed light on the number of molecules in each class of regulators.

C. Delineating Gene Regulatory Network Edges

1. TF-Target Gene Interactions

Interactions between TFs and their target genes can be identified using two conceptually different and highly complementary strategies. The first are TF-centered (protein-to-DNA); they start with a TF of interest and identify the genes with which this factor interacts. The second are gene-centered (DNA-to-protein); they start with a gene of interest and identify the TFs with which it interacts (Fig. 3).

2. Transcription Factor-Centered Methods: ChIP

The most widely used TF-centered method is chromatin immunoprecipitation (ChIP). In ChIP assays, an anti-TF antibody is used to precipitate TFs *in vivo*. Briefly, worm extracts are first treated with formaldehyde to crosslink proteins to proteins and proteins to DNA. After precipitation of the TF, associated DNA molecules can be identified 1) by PCR using primer sets of interest (Deplancke *et al.*, 2006a); 2) by cloning and sequencing (Oh *et al.*, 2006); 3) using microarrays that tile the entire *C. elegans* genome (Tabuchi *et al.*, 2011; Whittle *et al.*, 2009); or more recently 4) by deep sequencing (e.g., 454 or Solexa). Controls include a non-relevant antibody and, if possible, mutant animals that do not express the TF of interest (Walhout, 2011).

ChIP is a powerful method to identify TF-target gene interactions that occur *in vivo*. However, it is mostly limited to TFs that are highly and/or broadly expressed throughout the lifetime of the animal, and to TFs for which ChIP-grade antibodies are available. It is, however, also feasible to use ChIP in transgenic animals that overexpress an epitope-tagged TF. Although ChIP is usually the method of choice when one is interested in one or a few TFs, it is less suitable when one is interested in

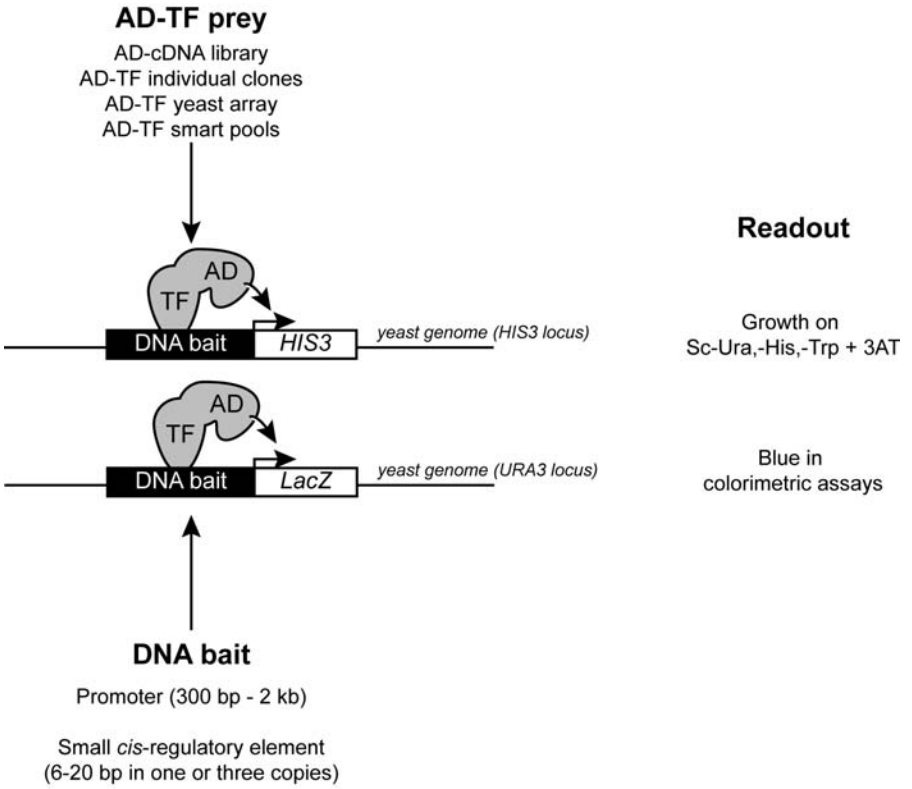

Fig. 3 Both TF-centered and gene-centered methods can be used for the identification of TF-target gene interactions.

a single gene (or a set of genes) and wants to identify the TFs that contribute to its (their) regulation. This is because all 940 TFs would have to be tested and under all relevant developmental and physiological conditions. Detailed discussion and protocols for ChIP in worms are provided elsewhere (Mukhopadhyay *et al.*, 2008).

3. Gene-Centered Methods: Y1H Assays

Y1H assays provide a genetic method for the gene-centered identification of TF-target gene interactions. The Y1H system is conceptually similar to yeast two-hybrid (Y2H) assays that have been used extensively to map *C. elegans* protein–protein interaction networks (Li *et al.*, 2004; Walhout *et al.*, 2000a, 2002). Here, I will discuss the principles of the Y1H system. Detailed Y1H protocols are available elsewhere (Deplancke *et al.*, 2006b).

The Y1H system uses a reporter gene readout in yeast to detect interactions between a "DNA bait" and a "protein prey" (e.g., TF) (Fig. 4). The first step in Y1H assays involves the selection of the DNA bait. In most of the cases, this will be a gene promoter or a small *cis*-regulatory element. Next, the DNA bait is cloned upstream of two reporter genes, *HIS3* and *LacZ* (Fig. 4). Traditionally, this was done by restriction enzyme/ligation-based methods (Li and Herskowitz, 1993). However, this is difficult to standardize and thus not amenable to the high-throughput settings that are required for regulatory network studies. To enable high-throughput cloning of DNA baits, we have combined the Y1H system with Gateway cloning, a recombination-based method that is compatible with the Promoterome resource (Deplancke *et al.*, 2004). With this method, multiple DNA baits can be transferred to the Y1H reporter Destination vectors simultaneously (e.g., in 96-well plates).

After cloning, the two DNA bait::reporter constructs are linearized and integrated into the genome of a suitable yeast strain. DNA bait::*HIS3* constructs are integrated into a mutant *HIS3* locus and plated on media lacking histidine. There is enough background His3 expression conferred by the basal yeast promoter present in the DNA bait::reporter constructs to enable growth on media lacking histidine. When the same construct is used in a protein–DNA interaction assay, however, the media are supplemented with 3-aminotriazole (3AT), a competitive inhibitor of the His3 enzyme. That way, the growth of the yeast depends on an increase in expression of His3, conferred by an interacting AD-TF hybrid protein (see below and Fig. 4). DNA bait::*LacZ* constructs contain a wild-type *URA3* gene and are integrated into a mutant *URA3* locus, thereby rescuing the Ura3 deficiency when plated on media lacking uracil. The DNA bait::reporter constructs do not carry a yeast origin of replication and, therefore, the formation of colonies is strictly dependent on their

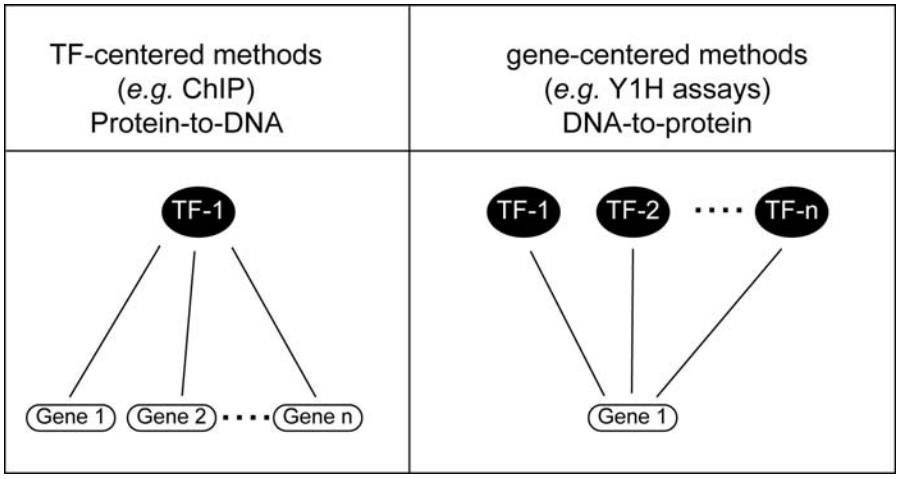

Fig. 4 Cartoon of yeast one-hybrid assays. AD – Gal4 transcription activation domain; TF – transcription factor.

integration into the yeast genome. Integrations are generally done sequentially, either by first integrating the DNA bait::*HIS3* or the DNA bait::*LacZ* construct, and following with the other. However, it is possible to integrate both constructs simultaneously, but the efficiency will be much lower and only a handful of colonies is usually obtained (unpublished data).

After picking integrant colonies, they need to be tested for background reporter gene expression (auto-activation). Levels of auto-activation can differ between integrants from the same DNA bait::reporter construct, most likely because of differences in copy number (Deplancke *et al.*, 2004). The degree of auto-activation of DNA bait::*HIS3* strains is determined by plating the colonies on media lacking histidine, and with increasing concentrations of 3AT (5, 10, 20, 40, 60, and 80 mM). Preferably colonies are selected that do not confer growth on low concentrations (5–40 mM) of 3AT. The degree of auto-activation of DNA bait::*LacZ* strains is determined by a colorimetric assay where white indicates no expression and darker shades of blue indicate increasing induction of βGalactosidase. Colonies with little or no blue should be selected where possible. In our hands 10–20% of all DNA baits exhibit high levels of auto-activation. These baits are difficult to use in Y1H assays although interacting TFs can sometimes be detected, particularly in directed Y1H assays (Vermeirssen *et al.*, 2007b).

After obtaining double integrant DNA bait strains that exhibit the lowest possible levels of auto-activation, the actual Y1H experiment can be performed to detect interacting TFs. In Y1H assays, TFs are fused to the transcription activation domain (AD) of the yeast Gal4 protein. This ensures that both activators and repressors of transcription can be detected. In other words, only physical protein–DNA interactions are examined in Y1H assays. AD-TF clones can be obtained from different sources and can be introduced into the DNA bait strain in different ways (Fig. 4). In our Y1H system, AD-TF clones carry wild-type yeast *TRP1* gene and, therefore, colonies containing the plasmid are selected on media lacking tryptophan.

The most commonly used method is by transforming an AD-cDNA library into haploid DNA bait strains (Arda *et al.*, 2010; Deplancke *et al.*, 2004, 2006a, 2006b; Martinez *et al.*, 2008a; Vermeirssen *et al.*, 2007a). Another source for such haploid transformations was created by cherry-picking relevant clones from the ORFeome, transferring them to the AD Y1H Destination vector by Gateway cloning, and combining them into a single AD-TF mini library (Deplancke *et al.*, 2004). This library consists of ~650 full-length TFs. Screening such a mini library enables the detection of TFs that are underrepresented in non-normalized cDNA libraries. Since TFs are often of low abundance, this can be very useful. In library screens, interacting TFs are identified by yeast colony PCR and sequencing. We have also developed mini pools of individual AD-TF clones that can be introduced into DNA bait strains by transformation (Vermeirssen *et al.*, 2007b). These pools are designed using a "Smart pool" strategy, based on a Steiner Triple System that is used in combinatorial mathematics. We have generated these pools as well as the scripts to deconvolute the resulting interactions. This method is useful for higher throughput, cost-effective Y1H experiments because it does not rely on extensive prey sequencing. Single AD-TF

clones can of course also be transformed individually when particular pre-defined interactions are to be examined (Reece-Hoyes *et al.*, 2009).

In addition to transformation into haploid DNA bait strains, AD-TF clones can also be introduced by mating. For this, we have transformed the AD-TF clones (∼755 in the first iteration) into yeast of mating type α, which is compatible for mating with the DNA bait strains that have the "a" mating type (Vermeirssen *et al.*, 2007b). DNA bait strains are mated with the AD-TF clone array and positives are examined in diploids. Each of these different methods for introducing AD-TFs into DNA bait strains has advantages and disadvantages (Vermeirssen *et al.*, 2007b). Generally, transformation detects more interacting TFs than mating. However, mating is fast, less labor-intensive, and much less costly. Further, interactions detected by mating are highly reproducible. When comparing library screens to more directed experiments with smart pools or individual clones, it is clear that many more protein–DNA interactions are found by the latter methods. However, with directed experiments only cloned TFs can by definition be found, which in our current collection is about 850 (∼90%) (Vermeirssen *et al.*, 2007b) (unpublished data). Proteins that do not have a recognizable DNA binding domain can only be retrieved in unbiased cDNA library screens (Deplancke *et al.*, 2006a; Vermeirssen *et al.*, 2007a). However, we do include these in TF resources after confirming their capability of interacting with *C. elegans* promoters and obtaining a suitable clone.

4. MicroRNA–mRNA Interactions

Putative interactions between the 3′UTRs of mRNAs and microRNAs are mostly identified genetically or computationally predicted using one or more algorithms that are publicly available. These include PicTar (Lall *et al.*, 2006), MiRanda (Griffiths-Jones *et al.*, 2006), TargetScan (Lewis *et al.*, 2005), RNA hybrid (Rehmsmeier *et al.*, 2004), and mirWIP (Hammell *et al.*, 2008). These algorithms are challenging to use because they are often too greedy (high rate of false positive predictions), or too stringent (high rate of false negative predictions). In order to alleviate this, at least to some extent, we have previously used predictions that were found by at least two of the four algorithms used (Martinez *et al.*, 2008a). Future experimental approaches will shed light onto physical and functional microRNA–mRNA interactions that occur *in vivo* (Lall *et al.*, 2006; Zisoulis *et al.*, 2010).

5. Other Regulatory Interactions

In addition to protein–DNA and microRNA–mRNA interactions, other relationships are involved in gene control. An important class involves sequence-specific RNA binding proteins that interact with the 3′UTR of mRNAs. It is not yet clear how many sequence-specific RNA binding proteins are encoded by the *C. elegans* genome, and only few have been studied genetically or biochemically. For instance, detailed binding sites have been determined *in vitro* for MEX-3, MEX-5, and a handful of other RNA binding proteins (Farley *et al.*, 2008; Pagano *et al.*, 2007,

2009). However, the functionality of most RNA binding proteins and their mRNA targets remains largely unexplored.

IV. Gene Regulatory Network Visualization and Analysis

The identification of physical and functional relationships between genes and their regulators is only the first step in the characterization of gene regulatory networks. Lists of interactions are usually difficult to navigate through. Network models, however, provide a visually attractive method for gene regulatory network analysis. We usually use the publicly available Cytoscape tool (Shannon *et al.*, 2003) for network visualization and analysis (Arda *et al.*, 2010; Deplancke *et al.*, 2006a; Grove *et al.*, 2009; Martinez *et al.*, 2008a; Vermeirssen *et al.*, 2007a). Subsequently, we use a variety of tools for network analysis. Most notably we use topological overlap coefficient analysis to compare gene expression patterns and to identify TF or gene network modules. These methods are discussed elsewhere (Arda *et al.*, 2010; Arda and Walhout, 2009; Ravasz *et al.*, 2002; Vermeirssen *et al.*, 2007a).

A. Gene Regulatory Network Validation

As with any method, the identification of physical and functional interactions between genes and their regulators is subject to issues related to both assay sensitivity and assay specificity (Walhout, 2011). Sensitivity refers to the proportion of real interaction that can be identified by the assay; interactions that cannot be detected are referred to as false negatives. Specificity refers to the proportion of interactions detected that are real, that is, that do occur *in vivo* and/or that have a biological consequence. Interactions that are detected but that are not "biologically meaningful" are referred to as false positives.

1. False Negatives

Previously, we estimated the coverage of our Y1H screens to be ~35% (Deplancke *et al.*, 2006a). This number is based on a very small number of available published interactions, but is very similar to the coverage obtained with Y2H (Braun *et al.*, 2009). There are several reasons that not all possible TF-promoter interactions can be detected by Y1H assays: 1) Several TFs bind DNA as obligatory dimers. Although homodimers can be detected, the Y1H assay currently is not configured to detect heterodimers. In the future, we hope to develop approaches that enable the detection of heterodimeric TF-DNA interactions in directed Y1H assays. 2) We will not find TFs that depend on specific post-translational modification or co-factor interactions with *C. elegans* proteins for DNA binding. 3) We can obviously only find TFs that are available in the TF resource used. However, it is highly encouraging to note that we have already detected interactions for about 25% of all predicted *C. elegans* TFs with only ~1% of all gene promoters.

2. False Positives

As with Y2H assays, there are two types of false positives with Y1H assays: *technical false positives* that cannot be reproduced in the same assay and *biological false positives* that represent genuine Y1H/Y2H interactions that nonetheless do not occur *in vivo*. To keep the rate of technical false positives low several issues need to be taken into consideration. First, it is best to only consider interactions that score positively for both Y1H reporters, that is, that induce growth on media lacking histidine and containing 3AT and that are bluer than an "AD only" control. Second, it is important to make sure that the TF retrieved is in frame (only relevant to cDNA library screens). Third, all Y1H interactions need to be retested in fresh DNA bait cells (i.e., from a frozen stock that has not been used in the screen itself), either by gap-repair (Walhout and Vidal, 2001) or by directly transforming an AD-TF clone. This is necessary because baits can mutate in yeast and give rise to a colony with an apparent interaction phenotype that is not reproducible (Walhout and Vidal, 1999). Fourth, it is absolutely critical to integrate DNA bait::reporter constructs into the yeast genome. We have tried to perform the assay with replicating plasmids, but the background expression was highly variable, probably due to different plasmid copy numbers. Finally, it is important to note that interactions obtained with highly auto-active DNA baits are more difficult to assess and may be less specific. We have developed an interaction scoring scheme to assess the results obtained from Y1H library screens (Vermeirssen *et al.*, 2007a).

Biological false positives are more challenging to assess. First, the genome itself is the same in every cell and thus, when a TF is expressed in any given cell one may expect the interaction to occur. However, the nucleosome occupancy likely varies in different cell types and this may prevent interactions from occurring *in vivo*. The integration of the DNA baits into the yeast genome ensures that they are incorporated into chromatin and, thus, Y1H assays are not based on interactions with naked DNA. However, it could be that the integration of the DNA baits in yeast only partially recapitulates the chromatin state in any *C. elegans* cell *in vivo*. In Y1H assays, we can find multiple members of a TF family binding to a particular DNA bait. This could be because these members have very similar DNA binding specificities and that this does not reflect *in vivo* functionality. However, we, and others, have found that multiple members of a TF family can bind the same DNA targets *in vivo* and can function redundantly (Hollenhorst *et al.*, 2007; Ow *et al.*, 2008). For instance, multiple TFs with a FLYWCH DNA binding domain were found to interact with microRNA promoters in Y1H assays and to redundantly repress microRNA expression in the early *C. elegans* embryo (Ow *et al.*, 2008). It is also important to note that not all TF-DNA interactions lead to a regulatory consequence. For instance, ChIP has identified numerous interactions that do not have an apparent biological function (Li *et al.*, 2008). This should be taken into account when physical interactions are being assessed by regulatory assays such as target gene expression in TF mutants or by TF knockdown with RNAi. Finally, different validation assays each have their own rate of false negatives, that is, they cannot detect every single genuine

interaction. For instance, assays that are performed with mixed populations of animals can easily miss interactions that occur only in a few cells or only during a short developmental time. Indeed, in our study of the B0507.1 promoter, we found a reduction in expression upon loss of the TF CES-1 only in the spermatheca, rectal gland, and pharyngeal-intestinal valve and, since these are not large tissues, this would be extremely difficult to detect in mixed population whole animal assays such as qPCR (Reece-Hoyes *et al.*, 2009).

V. Future Challenges

The comprehensive mapping of gene regulatory networks in *C. elegans* has only just started. Future studies are needed to complete transcriptional networks by high-throughput Y1H assays, and by other complementary assays such as ChIP. In addition, it will be highly useful to systematically generate promoter::GFP constructs and corresponding transgenic *C. elegans* lines for all worm genes. Such lines can then be used to examine promoter activity under different experimental or physiological conditions and to validate transcriptional networks, for instance using TF mutants or TF knockdown. Further, the continued experimental analysis of microRNAs and other small RNAs will be of extremely high value. Experimental methods also need to be developed and applied to assess other regulatory networks, such as those involving RNA binding proteins, signaling molecules, and metabolites. Finally, it will be exciting to go beyond static network models that represent a compilation of the interactions that can occur in the animal and to incorporate the dynamics and levels of gene and regulator expression and activation throughout the lifetime of the nematode.

Acknowledgments

I thank members of my lab for their hard work and especially John Reece-Hoyes and Lesley MacNeil for critical reading of the manuscript. Work in my lab is supported by the National Institutes of Health (DK068429 and GM082971) and by the Ellison Medical Research Foundation.

References

Arda, H. E., Taubert, S., Conine, C., Tsuda, B., Van Gilst, M. R., Sequerra, R., Doucette-Stam, L., Yamamoto, K. R., and Walhout, A. J. M. (2010). Functional modularity of nuclear hormone receptors in a *C. elegans* gene regulatory network. *Mol. Syst. Biol.* **6**, 367.

Arda, H. E., and Walhout, A. J. M. (2009). Gene-centered regulatory networks. *Briefings Funct. Genomic. Proteomic* doi:10.1093/elp049 **9**, 4–12.

Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J. M., Murray, R. R., Roncari, L., de Smet, A. S., Venkatesan, K., Rual, J. F., Vandenhaute, J., Cusick, M. E., Pawson, T., Hill,

D. E., Tavernier, J., Wrana, J. L., Roth, F. P., and Vidal, M. (2009). An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* **6**, 91–97.

Deplancke, B., Dupuy, D., Vidal, M., and Walhout, A. J. M. (2004). A gateway-compatible yeast one-hybrid system. *Genome Res.* **14**, 2093–2101.

Deplancke, B., Mukhopadhyay, A., Ao, W., Elewa, A. M., Grove, C. A., Martinez, N. J., Sequerra, R., Doucette-Stam, L., Reece-Hoyes, J. S., Hope, I. A., Tissenbaum, H. A., Mango, S. E., and Walhout, A. J. M. (2006a). A gene-centered *C. elegans* protein-DNA interaction network. *Cell* **125**, 1193–1205.

Deplancke, B., Vermeirssen, V., Arda, H. E., Martinez, N. J., and Walhout, A. J. M. (2006b). Gateway-compatible yeast one-hybrid screens. *CSH Protocols* doi:10.1101/pdb.prot4590 .

Dupuy, D., Li, Q., Deplancke, B., Boxem, M., Hao, T., Lamesch, P., Sequerra, R., Bosak, S., Doucette-Stam, L., Hope, I. A., Hill, D., Walhout, A. J. M., and Vidal, M. (2004). A first version of the *Caenorhabditis elegans* promoterome. *Genome Res.* **14**, 2169–2175.

Farley, B. M., Pagano, J. M., and Ryder, S. P. (2008). RNA target specificity of the embryonic cell fate determinant POS-1. *RNA* **14**, 2685–2697.

Friedlander, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewski, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* **26**, 407–415.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34**, D140–D144.

Grove, C. A., deMasi, F., Barrasa, M. I., Newburger, D., Alkema, M. J., Bulyk, M. L., and Walhout, A. J. (2009). A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell* **138**, 314–327.

Hammell, M., Long, D., Zhang, L., Lee, A., Carmack, C. S., Han, M., Ding, Y., and Ambros, V. (2008). mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *Nat. Methods* **5**, 813–819.

Hartley, J. L., Temple, G. F., and Brasch, M. A. (2000). DNA cloning using *in vitro* site-specific recombination. *Genome Res.* **10**, 1788–1795.

Hollenhorst, P. C., Shah, A. A., Hopkins, C., and Graves, B. J. (2007). Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the *ETS* gene family. *Genes Dev.* **21**, 1882–1894.

Hunt-Newbury, R., Viveiros, R., Johnsen, R., Mah, A., Anastas, D., Fang, L., Halfnight, E., Lee, D., Lin, J., Lorch, A., McKay, S., Okada, H. M., Pan, J., Schulz, A. K., Tu, D., Wong, K., Zhao, Z., Alexeyenko, A., Burglin, T., Sonnhammer, E., Schnabel, R., Jones, S. J., Marra, M. A., Baillie, D. L., and Moerman, D. G. (2007). High-throughput *in vivo* analysis of gene expression in *Caenorhabditis elegans*. *PLoS Biol.* **5**, e237.

Kato, M., de Lencastre, A., Pincus, Z., and Slack, F. J. (2009). Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during *Caenorhabditis elegans* development. *Genome Biol.* **10**, R54.

Kwon, E. S., Narasimhan, S. D., Yen, K., and Tissenbaum, H. A. (2010). A new DAF-16 isoform regulates longevity. *Nature* **466**, 498–502.

Lall, S., Grun, D., Krek, A., Chen, K., Wang, Y. -L., Dewey, C. N., Sood, P., Colombo, T., Bray, N., MacMenamin, P., Kao, H. -L., Gunsalus, K. C., Pachter, L., Piano, F., and Rajewski, N. (2006). A genome-wide map of conserved microRNA targets in *C. elegans*. *Curr. Biol.* **16**, 460–471.

Lamesch, P., Milstein, S., Hao, T., Rosenberg, J., Li, N., Sequerra, R., Bosak, S., Doucette-Stam, L., Vandenhaute, J., Hill, D. E., and Vidal, M. (2004). *C. elegans* ORFeome version 3.1: increasing the coverage of ORFeome resources with improved gene predictions. *Genome Res.* **14**, 2064–2069.

Lehrbach, N. J., and Miska, E. A. (2008). Functional genomic, computational and proteomic analysis of *C. elegans* microRNAs. *Brief. Funct. Genomic. Proteomic.* **7**, 228–235.

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P., and Bork, P. (2004). SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* **32**, D142–D144.

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20.

Li, J. J., and Herskowitz, I. (1993). Isolation of the ORC6, a component of the yeast origin recognition complex by a one-hybrid system. *Science* **262**, 1870–1874.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. -O., Han, J. -D. J., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J. -F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Bauemgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., van den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stam, L., Gunsalus, K., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D., and Vidal, M. (2004). A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543.

Li, X. Y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendriks, C. L., Chu, H. C., Ogawa, N., Inwood, W., Sementchenko, V., Beaton, A., Weiszmann, R., Celniker, S. E., Knowles, D. W., Gingeras, T. R., Speed, T. P., Eisen, M. B., and Biggin, M. D. (2008). Transcription factors bind thousands of active an inactive regions in the Drosophila blastoderm. *PLoS Biol.* **6**, e27.

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., Burge, C. B., and Bartel, D. P. (2003). The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* **17**, 991–1008.

Mangone, M., Macmenamin, P., Zegar, C., Piano, F., and Gunsalus, K. C. (2008). UTRome.org: a platform for 3′UTR biology in C. elegans. *Nucleic Acids Res.* **36**, D57–D62.

Mangone, M., Manoharan, A. P., Thierry-Mieg, D., Thierry-Mieg, J., Han, T., Mackowiak, S., Mis, E., Zegar, C., Gutwein, M. R., Khivansara, V., Attie, O., Chen, K., Salehi-Ashtiani, K., Vidal, M., Harkins, T. T., Bouffard, P., Suzuki, Y., Sugano, S., Kohara, Y., Rajewsky, N., Piano, F., Gunsalus, K. C., and Kim, J. K. (2010). The landscape of C. elegans 3′UTRs. *Science* **329**, 432–435.

Martinez, N. J., Ow, M. C., Barrasa, M. I., Hammell, M., Sequerra, R., Doucette-Stamm, L., Roth, F. P., Ambros, V., and Walhout, A. J. M. (2008a). A *C. elegans* genome-scale microRNA network contains composite feedback motifs with high flux capacity. *Genes Dev.* **22**, 2535–2549.

Martinez, N. J., Ow, M. C., Reece-Hoyes, J., Ambros, V., and Walhout, A. J. (2008b). Genome-scale spatiotemporal analysis of *Caenorhabditis elegans* microRNA promoter activity. *Genome Res.* **18**, 2005–2015.

Mukhopadhyay, A., Deplancke, B., Walhout, A. J., and Tissenbaum, H. A. (2008). Chromatin immunoprecipitation (ChIP) coupled to detection by quantitative real-time PCR to study transcription factor binding to DNA in Caenorhabditis elegans. *Nat. Protoc.* **3**, 698–709.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R. R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S. E., Pagni, M., Peyruc, D., Ponting, C. P., Selengut, J. D., Servant, F., Sigrist, C. J., Vaughan, R., and Zdobnov, E. M. (2003). The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31**, 315–318.

Oh, S. W., Mukhopadhyay, A., Dixit, B. L., Raha, T., Green, M. R., and Tissenbaum, H. A. (2006). Identification of direct targets of DAF-16 controlling longevity, metabolism and diapause by chromatin immunoprecipitation. *Nat. Genet.* **38**, 251–257.

Ow, M. C., Martinez, N. J., Olsen, P., Silverman, S., Barrasa, M. I., Conradt, B., Walhout, A. J. M., and Ambros, V. R. (2008). The FLYWCH transcription factors FLH-1, FLH-2 and FLH-3 repress embryonic expression of microRNA genes in C. elegans. *Genes Dev.* **22**, 2520–2534.

Pagano, J. M., Farley, B. M., Essien, K. I., and Ryder, S. P. (2009). RNA recognition by the embryonic cell fate determinant and germline totipotency factor MEX-3. *Proc. Natl Acad. Sci. U. S. A.* **106**, 20252–20257.

Pagano, J. M., Farley, B. M., McCoig, L. M., and Ryder, S. P. (2007). Molecular basis of RNA recognition by the embryonic polarity determinant MEX-5. *J. Biol. Chem.* **282**, 8883–8894.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555.

Reboul, J., Vaglio, P., Rual, J. F., Lamesch, P., Martinez, M., Armstrong, C. M., Li, S., Jacotot, L., Bertin, N., Janky, R., Moore, T., Hudson Jr., J. R., Hartley, J. L., Brasch, M. A., Vandenhaute, J., Boulton, S., Endress, G. A., Jenna, S., Chevet, E., Papasotiropoulos, V., Tolias, P. P., Ptacek, J., Snyder, M., Huang, R., Chance, M. R., Lee, H., Doucette-Stamm, L., Hill, D. E., and Vidal, M. (2003). *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* **34**, 35–41.

Reece-Hoyes, J. S., Deplancke, B., Barrasa, M. I., Hatzold, J., Smit, R. B., Arda, H. E., Pope, P. A., Gaudet, J., Conradt, B., and Walhout, A. J. (2009). The *C. elegans* Snail homolog CES-1 can activate gene expression *in vivo* and share targets with bHLH transcription factors. *Nucleic Acids Res.* **37**, 3689–3698.

Reece-Hoyes, J. S., Deplancke, B., Shingles, J., Grove, C. A., Hope, I. A., and Walhout, A. J. M. (2005). A compendium of *C. elegans* regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biol.* **6**, R110.

Reece-Hoyes, J. S., Shingles, J., Dupuy, D., Grove, C. A., Walhout, A. J., Vidal, M., and Hope, I. A. (2007). Insight into transcription factor gene duplication from *Caenorhabditis elegans* Promoterome-driven expression patterns. *BMC Genomics* **8**, 27.

Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. *RNA* **10**, 1507–1517.

Ruvkun, G., and Hobert, O. (1998). The taxonomy of developmental control in *Caenorhabditis elegans*. *Science* **282**, 2033–2041.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504.

Tabuchi, T., Deplancke, B., Osato, N., Zhu, L. J., Barrasa, M. I., Harrison, M. M., Horvitz, H. R., Walhout, A. J. M., and Hagstrom, K. (2011). Chromosome-biased binding and gene regulation by the *Caenorhabditis elegans* DRM complex. *PLoS Genet.* **7**, e1002074.

Vermeirssen, V., Barrasa, M. I., Hidalgo, C., Babon, J. A. B., Sequerra, R., Doucette-Stam, L., Barabasi, A. L., and Walhout, A. J. M. (2007a). Transcription factor modularity in a gene-centered *C. elegans* core neuronal protein-DNA interaction network. *Genome Res.* **17**, 1061–1071.

Vermeirssen, V., Deplancke, B., Barrasa, M. I., Reece-Hoyes, J. S., Arda, H. E., Grove, C. A., Martinez, N. J., Sequerra, R., Doucette-Stamm, L., Brent, M., and Walhout, A. J. M. (2007b). Matrix and Steiner-triple-system smart pooling assays for high-performance transcription regulatory network mapping. *Nat. Methods* **4**, 659–664.

Walhout, A. J. M. (2011). What does biologically meaningful mean?. A perspective on gene regulatory network validation. *Genome Biol.* **12**, 109.

Walhout, A. J. M., Reboul, J., Shtanko, O., Bertin, N., Vaglio, P., Ge, H., Lee, H., Doucette-Stam, L., Gunsalus, K. C., Schetter, A. J., Morton, D. G., Kemphues, K. J., Reinke, V., Kim, S. K., Piano, F., and Vidal, M. (2002). Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline. *Curr. Biol.* **12**, 1952–1958.

Walhout, A. J. M., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., and Vidal, M. (2000a). Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122.

Walhout, A. J. M., Temple, G. F., Brasch, M. A., Hartley, J. L., Lorson, M. A., van den Heuvel, S., and Vidal, M. (2000b). GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. *Methods Enzymol.* **328**, 575–592.

Walhout, A. J. M., and Vidal, M. (1999). A genetic strategy to eliminate self-activator baits prior to high-throughput yeast two-hybrid screens. *Genome Res.* **9**, 1128–1134.

Walhout, A. J. M., and Vidal, M. (2001). High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods* **24**, 297–306.

Whittle, C. M., Lazakovitch, E., Gronostajski, R. M., and Lieb, J. D. (2009). DNA-binding specificity and *in vivo* targets of *Caenorhabditis elegans* nuclear factor I. *Proc. Natl Acad. Sci. U. S. A.* **106**, 12049–12054.

Zisoulis, D. G., Lovci, M. T., Wilbert, M. L., Hutt, K. R., Liang, T. Y., Pasquinelli, A. E., and Yeo, G. W. (2010). Comprehensive discovery of endogenous Argonaute binding sites in *Caenorhabditis elegans*. *Nat. Struct. Mol. Biol.* **17**, 173–179.