A first version of the Caenorhabditis elegans Promoterome

Denis Dupuy  
*Dana-Farber Cancer Institute*

Qian-Ru Li  
*Dana-Farber Cancer Institute*

Bart Deplancke  
*University of Massachusetts Medical School*

*See next page for additional authors*

Follow this and additional works at: [http://escholarship.umassmed.edu/pgfe_pp](http://escholarship.umassmed.edu/pgfe_pp)

Part of the [Genetics and Genomics Commons](http://escholarship.umassmed.edu/pgfe_pp)

Repository Citation

Dupuy, Denis; Li, Qian-Ru; Deplancke, Bart; Boxem, Mike; Hao, Tong; Lamesch, Philippe; Sequerra, Reynaldo; Bosak, Stephanie; Doucette-Stamm, Lynn; Hope, Ian A.; Hill, David E.; Walhout, Albertha J. M.; and Vidal, Marc, "A first version of the Caenorhabditis elegans Promoterome" (2004). *Program in Gene Function and Expression Publications and Presentations*. 20.  
[http://escholarship.umassmed.edu/pgfe_pp/20](http://escholarship.umassmed.edu/pgfe_pp/20)

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in Program in Gene Function and Expression Publications and Presentations by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
A first version of the Caenorhabditis elegans Promoterome

Authors
Denis Dupuy, Qian-Ru Li, Bart Deplancke, Mike Boxem, Tong Hao, Philippe Lamesch, Reynaldo Sequerra, Stephanie Bosak, Lynn Doucette-Stamm, Ian A. Hope, David E. Hill, Albertha J. M. Walhout, and Marc Vidal

Rights and Permissions
Citation: Genome Res. 2004 Oct;14(10B):2169-75. Link to article on publisher’s site

This article is available at eScholarship@UMMS: http://escholarship.umassmed.edu/pgfe_pp/20
A First Version of the *Caenorhabditis elegans* Promoterome

Denis Dupuy, Qian-Ru Li, Bart Deplancke, et al.

*Genome Res.* 2004 14: 2169-2175
Access the most recent version at doi:10.1101/gr.2497604

**References**

This article cites 43 articles, 18 of which can be accessed free at:
http://genome.cshlp.org/content/14/10b/2169.full.html#ref-list-1

Article cited in:
http://genome.cshlp.org/content/14/10b/2169.full.html#related-urls

**Email alerting service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions
A First Version of the *Caenorhabditis elegans* Promoterome

Denis Dupuy, 1 Qian-Ru Li, 1 Bart Deplancke, 2 Mike Boxem, 1 Tong Hao, 1 Philippe Lamesch, 1 Reynaldo Sequerra, 4 Stephanie Bosak, 4 Lynn Doucette-Stamm, 4 Ian A. Hope, 3 David E. Hill, 1 Albertha J.M. Walhout, 2 and Marc Vidal 1, 5

1 Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; 2 Program in Gene Function and Expression and Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA; 3 School of Biology, University of Leeds, Leeds, LS2 9JT United Kingdom; 4 Agencourt Biosciences Corporation, Beverly, Massachusetts 01915, USA

An important aspect of the development of systems biology approaches in metazoa is the characterization of expression patterns of nearly all genes predicted from genome sequences. Such "localizome" maps should provide information on where (in what cells or tissues) and when (at what stage of development or under what conditions) genes are expressed. They should also indicate in what cellular compartments the corresponding proteins are localized. *Caenorhabditis elegans* is particularly suited for the development of a localizome map since all its 959 adult somatic cells can be visualized by microscopy, and its cell lineage has been completely described. Here we address one of the challenges of *C. elegans* localizome mapping projects: that of obtaining a genome-wide resource of *C. elegans* promoters needed to generate transgenic animals expressing localization markers such as the green fluorescent protein (GFP). To ensure high flexibility for future uses, we utilized the newly developed MultiSite Gateway system.

Biological processes result from complex networks of interactions between proteins, nucleic acids, and metabolites (Barabasi and Oltvai 2004). Although conventional approaches aimed at discovering the function of genes on-the-fly at one-at-a-time have been successful at elucidating many biological principles, they are not well suited to unravel the structure, function, and dynamics of complex molecular networks. With the availability of (nearly) complete genome sequences, approaches have been developed recently to investigate the systems properties of biological networks (Ideker et al. 2001; Vidal 2001; Kitano 2002). The *C. elegans* genome annotation provided the first draft of a comprehensive set of metazoan genes (*The C. elegans Sequencing Consortium* 1998). Subsequently, genome-wide expression profiling experiments (transcriptome mapping; Reinke et al. 2000; Kim et al. 2001), high-throughput (HT) RNAi analyses (phenome mapping; Fraser et al. 2000; Gonczy et al. 2000; Pfau et al. 2000; Kamath et al. 2003; Simmer et al. 2003) and HT yeast two-hybrid projects (interactome mapping; Walhout et al. 2000a, 2002; Davy et al. 2001; Boulton et al. 2002; Reboul et al. 2003; Li et al. 2004) were used to initiate studies of various biological processes in the worm at a systems level. However, the approaches mentioned above mainly probe the structure of the proteome network and the function of its components, without providing precise information on dynamic aspects. Thus additional information is needed about gene expression patterns at the cellular and/or subcellular levels ("localizome" mapping). Ultimately, localizome maps should provide information on where (in what cells or tissues) and when (at what stage(s) of development or under what conditions) genes are expressed. They should also indicate in what cellular compartments the corresponding proteins are localized.

Gene expression pattern information is being collected for multiple species, including human (http://biomap.ims.u-tokyo.ac.jp/), mouse (http://genex.hgu.mrc.ac.uk/), *Caenorhabditis elegans* (http://www.informatics.jax.org/menus/expression_menu.shtml), Xenopus (http://www.dkfz-heidelberg.de/abt0135/axeldb.htm), and *C. elegans* (http://129.11.204.86:591/, http://nematode.lab.nig.ac.jp/db/index.html/). These projects employ different techniques, such as antibody staining, mRNA collection from specific tissues, reporter genes strategies and in situ mRNA localization to generate gene expression pattern data.

Each approach has specific advantages and disadvantages, and provides information on different aspects of gene expression. For example, antibody staining provides direct information about protein localization in vivo, but the time and cost involved in generating antibodies for each predicted protein currently precludes application of this technique on a proteome-wide scale. Expression patterns can also be determined by in situ hybridization against specific mRNAs, as currently employed for *C. elegans* (Tabara et al. 1996). The drawbacks of this technique include potentially high background levels and potential cross-hybridization to homologous mRNAs. Finally, expression patterns can be determined by generating transgenic animals carrying promoters fused in vitro to *LacZ* (Hope 1991) or GFP (Chalfie et al. 1994) reporter genes.

[All clones in promoterome v1.1 are available to the scientific community through Open Biosystems (http://www.openbiosystems.com) and MRC geneservice (http://www.hgmp.mrc.ac.uk/geneservice).]

*Corresponding author.
E-MAIL marc_vidal@dfci.harvard.edu; FAX (617) 632-5739.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2497604.
The GFP reporter strategy has important advantages for localizome studies. First, GFP expression can be examined in living animals, thus changes in GFP expression levels and patterns can be followed over time during development. Second, GFP can also be expressed as a fusion with an endogenous ORF to investigate subcellular localization in vivo. Such GFP-based localizome data have already proven extremely valuable in yeast (Kumar et al. 2002; Huh et al. 2003). Finally, once a GFP reporter strain has been created, it can be used repeatedly to examine changes in expression in response to external stimuli, pharmacological agents or alterations of gene function. For these reasons, we chose to utilize the GFP reporter strategy to generate expression pattern data for C. elegans. This approach requires the production of C. elegans transgenic strains expressing GFP under the control of a variety of worm promoters (Hope et al. 1998). Here we describe the cloning of the initial version of a C. elegans promoter resource or “promoterome,” and how this resource can be used to generate gene expression pattern and protein localization data.

Ideally, a promoterome resource should have many other downstream applications, such as the study of physical interactions between promoters and the transcription complexes that regulate them. A global map of protein–DNA interactions would greatly benefit our understanding of the regulatory mechanisms that give rise to the expression patterns observed in vivo. Approaches used to investigate protein–DNA interactions include the yeast one-hybrid system (Deplancke et al. 2004) and chromatin immunoprecipitation followed by DNA microarray analysis (ChIP–CHIP or chip²). The latter has been successfully used to generate a map of the yeast transcriptional regulatory network (Lee et al. 2002). The availability of a cloned C. elegans promoterome will facilitate the production of both yeast one-hybrid reporter constructs and promoters microarray.

To generate a promoterome resource that can be easily transferred to a variety of vectors we decided to use a novel implementation of the Gateway technology (Hartley et al. 2000; Wallhout et al. 2000b). In addition to the versatility of the original Gateway recombinational cloning system, MultiSite Gateway (Cheo et al. 2004) allows the transfer of two or more DNA fragments from different Entry clones into a single Destination vector. Promoters are thus flanked by attL4 and attR1 sites, so that promoterome Entry clones can be cloned in fusion with ORFeome Entry clones, which are flanked by attL1 and attL2 recombination sites (Reboul et al. 2003). In summary, thanks to MultiSite Gateway, both resources can be used either independently or together (Fig. 1).

RESULTS

Use of MultiSite Gateway to Generate promoter::GFP and promoter::ORF::GFP

Transgenic Worms

The flexibility of the MultiSite Gateway recombinational cloning system described in this issue (Cheo et al. 2004) allows HT generation of various types of gene fusion constructs needed for different genome-wide localizome projects. Since our goals are to generate and analyze both promoter::GFP and promoter::ORF::GFP fusions, we confirmed that promoter sequences flanked by the Gateway sites attB4 and attB1 (B4-promoter-B1-GFP-R2) can faithfully recapitulate known expression patterns (see also, Hope et al. 2004). Furthermore, we demonstrated that promoter fusions containing attB sites can appropriately drive the expression of proteins fused to GFP (B4-promoter-B1-attB-GFP).

We selected three genes with well-described expression patterns: myo-2 is expressed in pharyngeal muscular cells (Okkema et al. 1993); mec-7 is in a subset of sensory neurons (touch cells; Hamelin et al. 1992); and hlh-8 in the M lineage and in the adult vulval muscle cells (Eimer et al. 2002). For each gene, we cloned into a MultiSite Entry vector a 2.5 kb DNA fragment located directly upstream of the predicted start codon. PCR amplicons of these “promoter” sequences were obtained using worm genomic DNA as template with primers flanked by attB sites (attB4 for the forward primer and attB1R for the reverse primer), and then cloned into the pDONR P4-P1R vector to generate attL4-promoter-attR1 Entry clones. The promoters were subsequently transferred into pDEST-D03 Destination vector to generate promoter::GFP fusions (see Methods). Transgenic lines were generated with each construct using micro particle bombardment (Praitis et al. 2001). The three resulting expression patterns matched previous observations obtained with clones not containing any Gateway sites (Fig. 2). We therefore concluded that the presence of attB sites in the final construction does not interfere noticeably with proper expression.

In addition, we established our ability to fuse promoter Entry clones with ORF Entry clones to analyze the subcellular localization patterns of protein::GFP fusion constructs. We produced six distinct promoter::ORF::GFP constructs by combining...
the mec-7 or myo-2 promoters with each of three ORFs (Y71H2R.4, T26A5.9, and B0035.8) from ORFeome v1.1. Out of 18 clones examined by restriction analysis (three for each construct), 15 displayed the correct fragment pattern on an agarose gel (not shown). One of each construct was sequenced and found to encode the expected ORF fused in frame to the GFP encoding sequence, downstream of the expected promoter.

To examine whether MultiSite-generated fusions recapitulate subcellular localization of the ORF, the two constructs harboring B0035.8, the histone H2B HIS-48 encoding ORF, were used to generate transgenic worms (Pmyo-2::his-48::GFP and Pmec-7::his-48::GFP). In both cases expression was observed in the expected cells, with fluorescence signal located in the nuclei of these cells as expected for a histone-GFP fusion protein (Fig. 2B, D). This experiment demonstrates the potential of combining the Gateway cloned promoters with ORFeome clones to produce promoter::ORF::reporter constructs that accurately reflect both gene expression pattern and subcellular localization of a protein.

Defining the C. elegans Promoterome
Promoters are commonly defined as DNA regions located upstream of transcription start sites and containing the necessary binding elements for proper transcriptional regulation. However, gene expression can be regulated by additional cis-acting elements, or enhancers, that can be located distantly upstream and/or in the introns, and/or downstream of genes (e.g., Conradt and Horvitz 1999). In the context of a genome-wide project it would not be possible to tackle at once all the complexity of gene regulation. We thus decided, as a first pass, to focus on the role of the proximal upstream regions in gene expression.

We decided to apply the definition that was used for S. cerevisiae promoters (Lee et al. 2002). For each gene, the “promoter” is defined here as the upstream intergenic region (IGR), specifically the region of DNA that spans from the ATG of the ORF to the extremity of the closest 5′ ORF (see Fig. 3A).

During the cloning of the ORFeome, we observed that cloning efficiency decreases as PCR fragment size increases above 2 kb (Reboul et al. 2001). Therefore, we set an upper size limit of 2 kb for the PCR fragments to ensure high cloning efficiency. This size restriction implies that regulatory elements located more than 2 kb upstream of the ATG will not be included.

We inspected ~200 articles describing expression patterns obtained by a promoter::reporter-approach in C. elegans. The size of the promoter regions employed varied from a few hundred base pairs to several kilobasepairs and was often dictated by the location of convenient restriction sites used in cloning (Fig. 3B). We noted that many of the published expression patterns derived from small “promoters” (red bars in Fig. 3B) encompassing only a fraction of the actual IGR (blue bars in Fig. 3B), suggesting that the inclusion of the entire IGR is not always necessary to obtain specific expression. Moreover, sequence comparison between C. elegans and C. briggsae shows a dramatic decrease in the level of homology 1500 bp upstream of the predicted ATG for most genes with a long IGR (G. Achaz, pers. comm.). Thus our strategy should result in the inclusion of most upstream cis-regulatory elements.

Several predicted intergenic regions are extremely small (the smallest IGR is only 3 bp), prompting us to set a lower size limit of 300 bp. Overall, 60% of predicted IGRs are fully included within these limits (Fig. 3C). For each predicted gene in the C. elegans genome, we determined the IGR and designed PCR primers using the OSP primer prediction software (Hillier and Green 1991).

For the sake of simplicity we refer to the individual clones constituting the promoterome as “promoter Entry clones”, although they also include the 5′ untranslated region and they might not contain all relevant regulatory sites (Fig. 3A). As ORFeome clones do not contain the original start codon (Reboul et al. 2001), our promoter Entry clones also add a translation start codon to allow proper translation of promoter::ORF::reporter constructs.

For this project’s design we used WormBase release WS93, which predicts 21,583 ORFs in the genome. This number includes internal splice variants, so the actual number of predicted start codons is 19,540. The number of actual promoters is further decreased by the presence of operons in the C. elegans genome, where a unique promoter drives the expression of several ORFs (Blumenthal et al. 2002; Lercher et al. 2003). The final count of unique promoters is 18,151. As a first pass we focused on cloning
the size of the initial PCR product (Table 1). In our analysis, there was no size bias in the genomic PCR success rate or in the efficiency of transferring fragments from Entry clones to Destination vectors (not shown). We therefore concluded that an enhanced competitive advantage towards cloning primer–dimers was the cause of our lower success rate for fragments bigger than 1.5 kb (see Methods).

The identity of each cloned promoter was established by sequencing. Of 6018 good-quality traces (Ewing et al. 1998), 5526 corresponded to the expected DNA fragment (92%). Most of the erroneous sequences corresponded to the Donor vector, which carries an ∼2-kb sized ccdB cassette that resembles a large promoter clone by electrophoretic analysis.

We developed a C. elegans promoterome database (PromoteromeDB; http://vital.dfc.ih..harvard.edu/promoteromdb) to integrate and disseminate the data collected in the promoterome-cloning project. The interface to this database displays IGR size, primer sequences, cloning and sequencing information, as well as links to corresponding entries in WormDB (http://wormdb.dfc.ih.ard. uva; Vaglio et al. 2003) and WormBase (http://www.wormbase.org; Stein et al. 2001) for each promoter.

**DISCUSSION**

The first version of the promoterome generated here contains a third of the predicted C. elegans “promoters” in a cloning format that allows their transfer into various destination vectors in HT settings.

Through the production of promoter:reporter and promoter:ORF–reporter constructs the promoterome can be used to generate gene expression pattern and protein localization data on a genome-wide scale. However, to date, the generation of transgenic strains carrying low-copy integrated constructs is a very tedious and time-consuming process that constitutes the bottleneck of such a project. It will thus take a collaborative effort from the C. elegans community to generate gradually evolving versions of the worm localizome.

As mentioned earlier, the scale of the project described herein led us to disregard regulatory elements contained in introns and the 3′ UTR, or located far upstream of the transcription start. As a result, a fraction of the expression patterns derived from the promoterome may not accurately reflect the native expression pattern of the corresponding proteins as could be observed with antibodies. However, despite these limitations, the construction of a localizome data set should provide insights into the function of numerous otherwise uncharacterized genes (Werner 2003). Moreover, the generation of a standardized localizome map should allow integration of this map with other functional genomics data sets. For example, integrating the spatiotemporal component derived from localization studies with the interactome map (Li et al. 2004) should allow the generation of tissue-specific interaction maps that would be more relevant to biological processes. Indeed, for an interaction to occur in vivo...
both partners have to be present in the same cell/tissue and at the same time during development and, so far, the C. elegans protein–protein interaction map fails to take into account the complex manner in which gene expression is regulated.

Loss-of-function phenotypes obtained by RNAi are commonly utilized to infer the role of a given gene. Although this approach is very powerful, only 10.3% of the ORFs assayed gave rise to an observable phenotype in N2 animals; the use of the RNAi sensitive mutant strain rrf-3(pk1426) only increased this number to 12.8% (Simmer et al. 2003). The promoterome could be used to deliver hairpin RNAi constructs to specific tissues or cells (Wesley et al. 2001). This may facilitate the identification of phenotypes for genes or tissues that are refractory to traditional RNAi delivery.

The availability of the promoterome should also allow for the production of a wide variety of GFP reporter strains. These can be utilized to characterize phenotypes otherwise too subtle to be observed, as recently demonstrated in a screen for genes involved in C. elegans fat accumulation (Ashrafi et al. 2003).

Finally, the promoterome will constitute a tool to study transcriptional regulatory networks through protein-DNA interactions studies. Its potential to be applied in a large-scale yeast one-hybrid setting is demonstrated elsewhere in this issue (Deplancke et al. 2004). Overlapping the localizome map and the data from protein-DNA interaction maps should allow better understanding of the molecular mechanisms that drive development. The use of a common resource of promoters to generate both data sets should greatly facilitate their integration and specifically the identification of transcription-factor binding elements (TFBEs). The information gathered about TFBEs in the IGRs should allow the identification of TFBEs located in other regions such as the introns, 3' untranslated regions or even within the ORFs themselves.

The promoterome cloning strategy described herein can be easily adapted to other organisms of interest provided that its annotated genome sequence is available. We believe that the combined use of ORFeomes and promoteromes resources will prove very powerful to unravel the complexity of gene regulatory networks in various model organisms.

### METHODS

#### Large-Scale Cloning of Promoters

Individual PCR reactions were performed in a 50-µl reaction volume using 50 ng of wild-type (N2) C. elegans genomic DNA as template and one unit of High Fidelity Platinum Taq polymerase (Invitrogen). To facilitate the analysis of the size of PCR products, primers (Illumina) were organized in 96-well plates by increasing size of the IGR. PCR products were analyzed on 1% agarose Gels. Ninety-five percent (95%) of the reactions gave visible PCR products of the expected size (primer pairs were non-purified PCR product was transferred to 8 µl of BP reaction mix containing 100 ng of Entry vector, 0.5 µl of BP clonase and 2 µl of 5X BP reaction buffer. After overnight incubation at 25°C, 2 µl of BP reaction was transformed into 20 µl of E. coli DH5α thermocompetent cells (Inoue et al. 1990), using ABgene Thermo-Fast 96-well plates. Transformed cells were spotted on LB agar+kanamycin plates using a Genmate workstation (Tecan). In pilot experiments, we isolated eight colonies for each of 94 promoters, to evaluate our background level. We observed that 86.6% of promoters under 1500 bp display the expected size in at least six out of eight colonies. In contrast, bigger promoters could be split into two categories: 32% showed at least six out of eight
correctly sized clones, while 68% showed less than two out of eight correct clones. In most cases the clones of the wrong size corresponded to primer-dimers. We therefore decided to isolate two colonies per clone, and select one that displayed the expected insert size by PCR. All clones displaying the correct insert size were reamplified and archived as both bacterial glycerol stocks and plasmid DNA mini-preps (a schematic of the protocol is shown in Fig. 4B).

Using MultiSite Gateway to Express GFP Constructs In Vivo

To generate promoter::GFP and promoter::ORF::GFP clones, we generated two MultiSite Gateway compatible destination vectors: pDest-DD03 and pDest-MB14. To generate pDest-DD03, a HindIII-Kpn fragment containing the attR4-ccb-attR2 cassette was moved from pDEST-6 (Invitrogen) into the unc-119 insert vector pDP#MM106b (Prairits et al. 2001). To generate pDEST-MB14, a Kpn-Apal fragment containing the GFP encoding sequence and the unc-119 5′ UTR was moved from pKD95.79 (Fire et al. 1999) into pDP#MM106b. Next, an EcolV fragment containing the attR4-ccb-attR2 cassette was moved from pDEST-6 (Invitrogen) into this construct digested with Acc65I and made blunt by fill-in with Klenow. The unc-119 rescuing fragment present in these vectors can be used as a marker to select transgenic worms. Multi-Site LR reactions were performed in 10-µl volumes, but otherwise according to manufacturers specifications (Invitrogen). Transgenic worm-strains were generated using a ballistic transformation protocol as previously described (Prairits et al. 2001).

ACKNOWLEDGMENTS

This work was funded by the National Cancer Institute (NCI 4 R33 CA97516-02 awarded to AJMW, IAH, and MV), the Leukemia Research Foundation (MB), the King Baudouin Foundation and the Belgian American Educational Foundation (Henri Benedictus–BAEF Fellowship in Biomedical Engineering awarded to BD), and the Fu Fellowship (QL). Thanks to Tracey Clingelinsmith for her indispensable administrative assistance, John Nahory and Judith Hernandez from Illumina for their technical help.

REFERENCES


WEB SITE REFERENCES
http://bodymap.ims.u-tokyo.ac.jp/; Bodymap Human and Mouse Gene Expression Database.
http://genex.hgu.mrc.ac.uk; EMAP, the Edinburgh Mouse Atlas Project.
http://www.informatics.jax.org/menus/expression_menu.shtml; Mouse Genome Informatics.
http://www.dkfz-heidelberg.de/abt0135/axeldb.htm; A Xenopus laevis database.
http://129.11.204.86:591; The Hope Laboratory Expression Pattern Database.
http://nematode.lab.nig.ac.jp/db/index.html; NEXTDB, the Nematode Expression Pattern Database.
http://worfdb.dfci.harvard.edu/; WorfDB, the C. elegans ORFeome Cloning Project Database.
http://www.hgmp.mrc.ac.uk/geneservice; MRC geneservice.
http://www.openbiosystems.com; Open Biosystems.
http://vidal.dfci.harvard.edu/promoteromedb; PromoteromeDB the C. elegans Promoterome Database.

Received February 23, 2004; accepted in revised form May 4, 2004.