A short conserved sequence in the HOT region of the *C. elegans dlg-1* gene largely recapitulates its transcriptional behavior during embryogenesis

Cristina Tocchini[1,*], Palmer Bassett[1], and Susan E. Mango[1,*]

1 Biozentrum, University of Basel, 4056 Basel, Switzerland.

* Co-corresponding authors:

c.tocchini@unibas.ch

susan.mango@unibas.ch

Abstract

High occupancy target (HOT) regions, genomic sites bound by many transcription factors, are found across eukaryotes, but their regulatory roles remain unclear. In *Caenorhabditis elegans*, transcription primarily relies on promoter-proximal elements, and HOT regions frequently coincide with constitutively active promoters. Examining the *dlg-1* locus, we identify a putative HOT region essential for embryonic expression. A small sequence within this region, far upstream of the transcription start site, is essential for *dlg-1* transcriptional levels. In addition, the first intron modestly enhances transcription. These findings reveal that HOT regions in non-canonical positions can play critical roles in gene regulation in *C. elegans*.

Key words

HOT regions, transcriptional regulation, *C. elegans*, first intron, *dlg-1*

Background

Precise regulation of gene expression is crucial for orchestrating complex developmental processes. Transcriptional control relies on a series of regulatory elements located within promoters or in distant genomic regions containing enhancers or repressors (1). High occupancy target (HOT) regions are genomic sites bound by unusually large numbers of transcription factors (2,3), but their contribution to transcriptional regulation remains poorly understood. In *Caenorhabditis elegans*, transcriptional regulation mainly relies on short-distance regulatory sequences within promoters and can carry HOT sites that activate transcription ubiquitously (2,4). *C. elegans* enhancers have been identified by their chromatin modifications, but they typically lack strong HOT sites which are generally not essential for transcription (1,2). In *C. elegans*, genome-wide studies have identified HOT regions that overlap with both promoters and enhancer sequences (2). Whether HOT regions act redundantly or additively as a whole unit, or have distinct sub-HOT regulatory regions to fine-tune gene expression, remains unknown. Here, we dissect the regulatory landscape of the *dlg-1/Discs large* locus during embryonic morphogenesis (5,6). We identify a conserved region that encompasses a previously unreported HOT site that is necessary and sufficient for *dlg-1* expression. This region carries histone marks typical of enhancers and is located far upstream of the transcription start site, suggesting it to be an enhancer. A subdomain within a conserved region possessing enhancer features is sufficient for transcription in all *dlg-1*-positive cells. These sequences act with a second minor HOT site within the first long intron of *dlg-1*, that also has features of enhancers and is sufficient to activate transcription in embryonic epidermal cells. Together, our findings reveal minor HOT sites in enhancers and control general and selective gene expression in *C. elegans*. This expands the repertoire of regulatory strategies used in *C. elegans* transcriptional regulation.

Results and discussion

We focused our analysis of transcriptional regulation on the *dlg-1* locus, which is expressed in diverse tissues (5,6). The *dlg-1* promoter has been previously defined as a seven-kilobase (Kb) sequence that stretches from the transcription start site (TSS) of the gene upstream on the minus DNA strand, *dpyd-1*, to the TSS of the *dlg-1* gene on the plus DNA strand (7) (Figure 1A,B). In analyzing this seven-Kb sequence, we identified an upstream genomic region highly conserved among the *Caenorhabditis* species (Figure 1A,B). This region is bound by several transcription factors (TFs) suggestive of a high occupancy target (HOT) region (2,3) (Figure 1A and S1A). However, large-scale studies aiming to identify HOT regions, did not picked up this genomic region (2), suggesting it might only be a minor HOT region (8). In *C. elegans*, sequences that dictate correct expression levels and patterning are usually located within a few hundred base-pairs (bp) from a gene's TSS (3,9–11). The conserved region is positioned approximately four Kb from the TSS of *dlg-1* and is overlapping and bordered by chromatin marks typically found in enhancers. Specifically, it is enriched in acetylated lysine 27 (H3K27ac) and mono- and di-methylated lysine 4 of histone H3 (H3K4me1/2), and depleted from tri-methylated lysine 4 of histone H3 (H3K4me3) (Figure S1B) (12,13). Indeed, this genomic region has been predicted an enhancer site (13). Cohesins have been shown to organize the DNA in a way that active enhancers make contacts in the 3D space forming "fountains" detectable in Hi-C data (14,15). Although the presence of a fountain at this locus remains to be determined during embryogenesis, ARC-C data showed the putative HOT region interacts with other DNA sequences, including the first intron of *dlg-1* (1).

We hypothesized the minor HOT site could be an enhancer for *dlg-1.* To test this idea, we generated three single-copy, integrated transcriptional reporters carrying a histone H2B sequence fused to an mNeon-Green (mNG) fluorescent protein under the transcriptional control of the following *dlg-1* upstream regions: a full-length sequence (FL) and two deletion sequences spanning from the beginning (HOT) or the end (Short) of the HOT region until the *dlg-1* TSS (Figure 1A,C). Live imaging and quantitation analyses showed minor and insignificant differences between the FL and the HOT reporters regarding cell-type specificity (Figure 1C) and levels of expression (Mean FL= 1.00, STDEV FL = 0.26; mean HOT = 0.93, STDEV HOT = 0.22; Figure 1D), and recapitulated that of the endogenous *dlg-1* gene (5,6). Surprisingly, the Short reporter, containing nearly four Kb upstream the TSS, showed a dramatic reduction in expression levels compared to FL and HOT (Mean Short = 0.29, STDEV Short = 0.08; Figure 1C,D). This reduction was observed in most of the cells where endogenous *dlg-1* is normally expressed (Figure 1C-E). Only developing epidermal cells were still transcriptionally active, as shown previously with an extrachromosomal *dlg-1* transcriptional reporter with truncated upstream region of the TSS (16). smFISH experiments showed that the transcriptional control mediated by the HOT region was specific to *dlg-1*, as

the nearby *dpyd-1* locus was expressed sporadically (Figure 1E). Together, these analyses demonstrate that the HOT region is necessary to (i) enhance *dlg-1* expression in developing epidermal cells, and (ii) allow *dlg-1* expression in a broader range of cell types, such as neuronal and pharyngeal precursors. We conclude that the HOT region constitutes a key regulatory sequence for correct expression levels and patterning of *dlg-1*.

To understand the sufficiency and determine which portion of the HOT region was important for its function in gene expression strength and patterning, we selected four small regions within and surrounding the upstream conserved region (HOT1 = 406 bp; HOT2 = 185 bp; HOT3 = 332 bp; HOT4 = 374 bp) (Figure 1B and 2A,B). Each region was paired with the minimal promoter *Δpes-10* (Figure 2A), commonly used to assess the functionality of candidate regulatory sequences in the context of transcription (17) and generated four transcriptional reporters. As previously described, we performed live imaging and fluorescence quantitation on these reporters to assess the levels and pattern of expression dictated by each sequence. Among the four synthetic promoters designed, only HOT2 and HOT3 were permissive for transcription (Mean FL = 1.00, STDEV FL = 0.19; Mean HOT1 = 0.02, STDEV HOT1 = 0.05; Mean HOT2 = 0.10, STDEV HOT2 = 0.08; Mean HOT3 = 0.94, STDEV HOT3 = 0.35; Mean HOT4 = -0.01, STDEV HOT4 = 0.02; Figure 2B,C). The HOT2 reporter showed scattered and variable expression that differed significantly in expression levels compared to the FL. The HOT3 reporter showed patterning (Figure 2B) and expression levels (Figure 2C) comparable to the FL reporter. These data suggest that a small sequence of 330 bp, which lies more than four Kb from the TSS, is a strong driver of *dlg-1* transcription during embryonic morphogenesis.

We found a second region within the first intron of *dlg-1* was also enriched for TF binding sites (3), making it another candidate HOT region (Figure S1A). Long first introns have been shown to frequently contain regulatory regions used for transcription, such as enhancers (18–20). The enhancers located in introns can drive gene expression in patterns that are distinct from, yet complementary to, those found in the gene's promoter (9). *dlg-1* possesses a long first intron whose sequence is enriched with H3K4me1/2 chromatin modifications, suggesting that it is an enhancer (Figure S1B) (13). To test if the first intron contributed to *dlg-1* transcriptional control, we generated two single-copy integrated lines where the first intron was either paired with a minimal promoter (INT1) or inserted in the FL transgene within the H2B sequence which acted as an actual intron (FL-INT1). The reporter INT1 was created to test the transcriptional abilities of intron 1 alone, whereas the reporter FL-INT1 was engineered to determine if the first intron was changing the transcriptional capabilities of the upstream region (Figure 2A,D). We found that the fluorescence expression driven by the INT1 reporter was confined to developing epidermal cells (Figure 2D) and minimal (13% of the FL; Mean FL = 1.00, STDEV FL = 0.19; Mean INT1 = 0.13, STDEV INT1 = 0.02; Figure 2E), suggesting that this sequence plays a marginal role in *dlg-1* transcription. Interestingly, quantitation analyses

on the FL-INT1 reporter showed a slight, but significant increase (+15%) in fluorescence intensity compared to the reference FL reporter in the tested cells (Mean INT1 = 1.15, STDEV INT1 = 0.22; Figure 2E). This result supports the idea of enhancers providing additive effect to gene transcription. We conclude that *dlg-1* intron 1 complements the promoter, albeit weakly, in a subset of cells.

In summary, we identified two key regulatory regions in the *dlg-1* locus that are critical for modulating gene expression levels and spatial patterning during embryonic morphogenesis: a small portion of a distal HOT region located over four Kb upstream of the TSS (HOT3), and the first intron. HOT regions in *C. elegans* have been predominantly associated with promoters and were often considered non-functional or simply reflective of accessible chromatin. Our data provide clear functional evidence that a distal HOT region can instead act as a *bona fide* enhancer. Specifically, we found that a 330 bp fragment within HOT3 is sufficient to drive strong and spatially-patterned expression in embryonic cells, closely resembling the endogenous *dlg-1* expression profile. In addition, we showed that the first intron acts additively with the upstream region (*i.e.*, HOT3) to boost gene expression specifically in transcriptionally active epidermal cells. The presence of shared TF binding both HOT/enhancer regions, HOT3 and intron 1, suggests that regulatory elements far upstream the TSS and intronic regions can function cooperatively. This discovery challenges previous reports that HOT regions are functionally isolated modules, and instead suggest a possible cooperative enhancer logic.

Conclusions

Together, our findings illustrate how non-canonical regulatory elements (distal HOT regions and intronic enhancers) can be integrated to fine-tune gene expression in a cell-type-specific manner during development. These results not only clarify the functional potential of HOT regions in *C. elegans* but also expand our understanding of how enhancers contribute to precise spatiotemporal control of gene expression in compact genomes like the one of *C. elegans*.

Methods

Nematode culture

C. elegans strains were maintained as previously described (21) at 20°C. Embryos were imaged at the comma stage. For a detailed list of transgenic lines used in this study, see Table S1.

Generation of transgenic lines

Transgenes carrying the different promoters analyzed in this study (amplified from N2 lysate) were fused to an H2B sequence (amplified from N2 lysate) paired to an mNG sequence through a flexible linker (amplified from LP598 lysate), and a *tbb-2* 3'UTR (amplified from N2 lysate). These sequences were assembled (NEBuilder® HiFi DNA Assembly Cloning Kit, New England BioLabs, cat#E5520) in the pCFJ150 vector and inserted into the *ttTi5605* locus on chromosome II of the mosSCI strain EG6699 (22). NEB® 10-beta Competent *E. coli* (High Efficiency) (NEB, cat#C3019) competent cells were used for transformation. Full plasmid sequencing from Microsynth was used to confirm the correctness of the sequences. Plasmids were purified with MIDI preps (NucleoBond Xtra Midi Plus (Macherey - Nagel, cat#740412.10) and single-copy integrated lines were generated following the mosSCI protocol (22). For a detailed list of plasmids used in this study, see Table S1.

smFISH

smFISH experiments were performed as previously described (6,23,24). For a full list of smFISH probes, see Table S1.

A widefield microscope FEI "MORE" with total internal reflection fluorescence (TIRF), equipped with a Hamamatsu ORCA flash 4.0 cooled sCMOS camera, and a Live Acquisition 2.5 software were used for capturing smFISH images. Images were deconvolved with the Huygens software and processed in OMERO (https://www.openmicroscopy.org/omero/). All figures were assembled in Adobe Illustrator (https://www.adobe.com/).

Live imaging: microscopy, quantitation, and analysis

Live imaging was performed on embryos at the comma stage. Embryos were washed from plates with 1 ml of water, collected into Eppendorf tubes, and spun-down ("short") for 5 seconds in a table centrifuge. After removing the supernatant, embryos were resuspended in M9 buffer

and transferred onto poly-L-lysine-coated slides (Thermo Scientific, cat#ER-308B-CE24). A coverslip was placed on the sample and embryos were imaged immediately.

A widefield ZEISS Axio Imager M2 equipped with a ZEISS Colibri LED light source, a Hamamatsu ORCA flash 4.0 camera, and ZEN 2.6 software (blue edition) was used for capturing images for live imaging of *C. elegans* embryos.

Fluorescence signal quantitation of nuclear signal (from H2B::mNG) in the different transgenic lines was performed using Fiji, where the maximum intensity from the same three seam cell nuclei per embryo was quantified.

Statistical analyses were performed using R software (R Core Team, 2021; https://www.R-project.org/). The ggplot2 package (Wickham, 2016; https://ggplot2.tidyverse.org) was used to generate bar plots with error bars (standard deviation). Each dot represents the normalized value to the corresponding FL of each seam cell nucleus. Statistical differences were determined using a *t*-test.

Figure legends

Figure 1. The HOT region in the distal part of the *dlg-1* upstream region is the main driver of transcription. A. Schematic representation of the *dlg-1* locus: black, exons (thick lines) and introns/UTR (thin lines); cyan, the first long intron; light grey, distal-most sequence of the upstream region; dark purple, sequence in and around the HOT region; dark grey, proximal-most sequence of the upstream region. **B.** Snapshot of the upstream region of the *dlg-1* TSS from JBrowse showing curated genes, conservation, and schematics used to generate the transgenic lines in this study. **C.** Upper-right side: schematic representation of the transgenic constructs used in this figure: H2B::mNG (mNG) coding sequence containing three synthetic introns (green) and a *tbb-2* 3'UTR (white) were under the control of upstream regions with different lengths (white). The fluorescent images are from live *C. elegans* embryos (outlined by dashed lines) at the comma stage and corresponding zoom-ins (squared boxes). The three seam cell nuclei used for quantitation are highlighted in the zoom-ins with dashed lines. A reference nucleus to identify the same measured nuclei in all the conditions is highlighted with an asterisk. The transgene expressed in the exemplifying embryos are stated above each image and accompanied with a small schematic. **D.** Bar plot with error bars (standard deviation) normalized to the FL. Each dot represents the maximum fluorescent intensity of a seam cell nucleus. ns: not significant; ***: $p < 0.01$. For raw data, see Table S1. **E.** Fluorescence images of a *C. elegans* embryo at the comma stage. Upper panels, section of the surface of the embryo; lower panels, section of the middle part of the same embryo. Panels from left to right: smFISH signal from *dlg-1* mRNA using probes complementary to exon region (magenta); smFISH signal from *dpyd-1* pre-mRNA using probes complementary to intron regions (cyan); antibody stain signal from GFP fused to DLG-1 (yellow); and merges. Scale bar: 10 µm.

Figure 2. A small conserved sequence within the HOT region and the first intron of *dlg-1* contribute to transcriptional regulation. A. Upper part: schematic representation of the *dlg-1* locus as in Fig. 1A. Lower part: schematic representations of the types of transgenic constructs used in this figure: H2B::mNG (mNG) coding sequence containing three synthetic introns (green) and a *tbb-2* 3'UTR (white) were under the control of (left schematics) upstream regions with different lengths (white), or (right schematics) portions of the different tested regions (white) followed by a minimal *Δpes-10* promoter (MINp, orange). **B,D.** Upper-right side: schematic representations of the types of transgenic constructs used in these panels: H2B::mNG (mNG) coding sequence containing three synthetic introns (green) and a *tbb-2* 3'UTR (white) were under the control of portions of the HOT region (purple, B) or the first intron (cyan, D) followed by a minimal *Δpes-10* promoter (MINp, orange). The fluorescent images

are from live *C. elegans* embryos (outlined by dashed lines) at the comma stage and corresponding zoom-ins (squared boxes) of the three seam cell nuclei used for quantitation (smaller dashed lines depict nuclei boundaries). A reference cell to identify the same three seam cells measured in all the conditions in highlighted with an asterisk. The transgenes expressed in the embryos are stated above each image and accompanied with a small schematic. **C,E.** Bar plot with error bars (standard deviation) normalized to the FL. Each dot represents the maximum fluorescent intensity of a seam cell nucleus (three nuclei per embryo). ns: not significant; ***: $p < 0.01$. For raw data, see Table S1.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data are available in the described sources in the Methods.

Competing interests

The authors declare that they have no competing interests.

Funding

The study has been partially funded by the Forschungsfonds (Excellent Junior Researcher) from the University of Basel (U.570.1040) from CT and the Swiss National Science Foundation (310030_197713) from SEM.

Authors' contribution

Conception of the work (CT and SEM); design of the work (CT); acquisition and analysis (CT and PB); interpretation of data (CT); drafted the work (CT); revised the work (SEM and PB). All authors have read and approved the current version of this manuscript and have agreed to its submission for publication.

Acknowledgments

We thank current and previous lab members of the Mango group for scientific discussions, and WormBase. Some strains were provided by the CGC, which is funded by the NIH Office of Research Infrastructure Programs (P40 OD010440).

References

1. Huang N, Seow WQ, Appert A, Dong Y, Stempor P, Ahringer J. Accessible Region Conformation Capture (ARC-C) gives high-resolution insights into genome architecture and regulation. Genome Res. 2022 Feb;32(2):357–66.

2. Chen RAJ, Stempor P, Down TA, Zeiser E, Feuer SK, Ahringer J. Extreme HOT regions are CpG-dense promoters in *C. elegans* and humans. Genome Res. 2014 Jul;24(7):1138–46.

3. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al. Integrative Analysis of the *Caenorhabditis elegans* Genome by the modENCODE Project. Science. 2010 Dec 24;330(6012):1775–87.

4. Van Nostrand EL, Kim SK. Integrative analysis of *C. elegans* modENCODE ChIP-seq data sets to infer gene regulatory interactions. Genome Res. 2013 Jun;23(6):941–53.

5. Bossinger O, Klebes A, Segbert C, Theres C, Knust E. Zonula Adherens Formation in Caenorhabditis elegans Requires dlg-1, the Homologue of the Drosophila Gene discs large. Developmental Biology. 2001 Feb;230(1):29–42.

6. Tocchini C, Rohner M, Guerard L, Ray P, Von Stetina SE, Mango SE. Translation-dependent mRNA localization to *Caenorhabditis elegans* adherens junctions. Development. 2021 Dec 15;148(24):dev200027.

7. McMahon L, Legouis R, Vonesch JL, Labouesse M. Assembly of *C. elegans* apical junctions involves positioning and compaction by LET-413 and protein aggregation by the MAGUK protein DLG-1. Journal of Cell Science. 2001 Jun 15;114(12):2265–77.

8. Kvon EZ, Stampfel G, Yáñez-Cuna JO, Dickson BJ, Stark A. HOT regions function as patterned developmental enhancers and have a distinct *cis* -regulatory signature. Genes Dev. 2012 May 1;26(9):908–13.

9. Fuxman Bass JI, Tamburino AM, Mori A, Beittel N, Weirauch MT, Reece-Hoyes JS, et al. Transcription factor binding to Caenorhabditis elegans first introns reveals lack of redundancy with gene promoters. Nucleic Acids Research. 2014 Jan 1;42(1):153–62.

10. Gaudet J, Muttumu S, Horner M, Mango SE. Whole-Genome Analysis of Temporal Gene Expression during Foregut Development. Ronald H. A. Plasterk, editor. PLoS Biol. 2004 Oct 19;2(11):e352.

11. Zhong M, Niu W, Lu ZJ, Sarov M, Murray JI, Janette J, et al. Genome-Wide Identification of Binding Sites Defines Distinct Functions for Caenorhabditis elegans PHA-

4/FOXA in Development and Environmental Response. Copenhaver GP, editor. PLoS Genet. 2010 Feb 19;6(2):e1000848.

12. Spicuglia S, Vanhille L. Chromatin signatures of active enhancers. Nucleus. 2012 Mar;3(2):126–31.

13. Jänes J, Dong Y, Schoof M, Serizay J, Appert A, Cerrato C, et al. Chromatin accessibility dynamics across C. elegans development and ageing. eLife. 2018 Oct 26;7:e37344.

14. Isiaka BN, Semple JI, Haemmerli A, Thapliyal S, Stojanovski K, Das M, et al. Cohesin forms fountains at active enhancers in *C. elegans* [Internet]. Genomics; 2023 [cited 2025 Jun 12]. Available from: http://biorxiv.org/lookup/doi/10.1101/2023.07.14.549011

15. Kim J, Wang H, Ercan S. Cohesin organizes 3D DNA contacts surrounding active enhancers in *C. elegans* [Internet]. Genomics; 2023 [cited 2025 Jun 12]. Available from: http://biorxiv.org/lookup/doi/10.1101/2023.09.18.558239

16. Sheffield M, Loveless T, Hardin J, Pettitt J. C. elegans Enabled Exhibits Novel Interactions with N-WASP, Abl, and Cell-Cell Junctions. Current Biology. 2007 Oct;17(20):1791–6.

17. Fire A, Harrison SW, Dixon D. A modular set of lacZ fusion vectors for studying gene expression in Caenorhabditis elegans. Gene. 1990 Jan;93(2):189–98.

18. Bradnam KR, Korf I. Longer First Introns Are a General Property of Eukaryotic Gene Structure. Christoffels A, editor. PLoS ONE. 2008 Aug 29;3(8):e3093.

19. Brabin C, Appleford PJ, Woollard A. The Caenorhabditis elegans GATA Factor ELT-1 Works through the Cell Proliferation Regulator BRO-1 and the Fusogen EFF-1 to Maintain the Seam Stem-Like Fate. Hardin JD, editor. PLoS Genet. 2011 Aug 4;7(8):e1002200.

20. Jorgensen EM, Mango SE. The art and design of genetic screens: Caenorhabditis elegans. Nat Rev Genet. 2002 May 1;3(5):356–69.

21. Brenner S. The genetics of Caenorhabditis elegans. Genetics. 1974 May;77(1):71–94.

22. Frøkjær-Jensen C, Wayne Davis M, Hopkins CE, Newman BJ, Thummel JM, Olesen SP, et al. Single-copy insertion of transgenes in Caenorhabditis elegans. Nat Genet. 2008 Nov;40(11):1375–83.

23. Tocchini C, Mango SE. An adapted MS2-MCP system to visualize endogenous cytoplasmic mRNA with live imaging in Caenorhabditis elegans. Ketting RF, editor. PLoS Biol. 2024 Mar 1;22(3):e3002526.

24. Tsanov N, Samacoits A, Chouaib R, Traboulsi AM, Gostan T, Weber C, et al. smiFISH and FISH-quant – a flexible single RNA detection approach with super-resolution capability. Nucleic Acids Res. 2016 Dec 15;44(22):e165–e165.

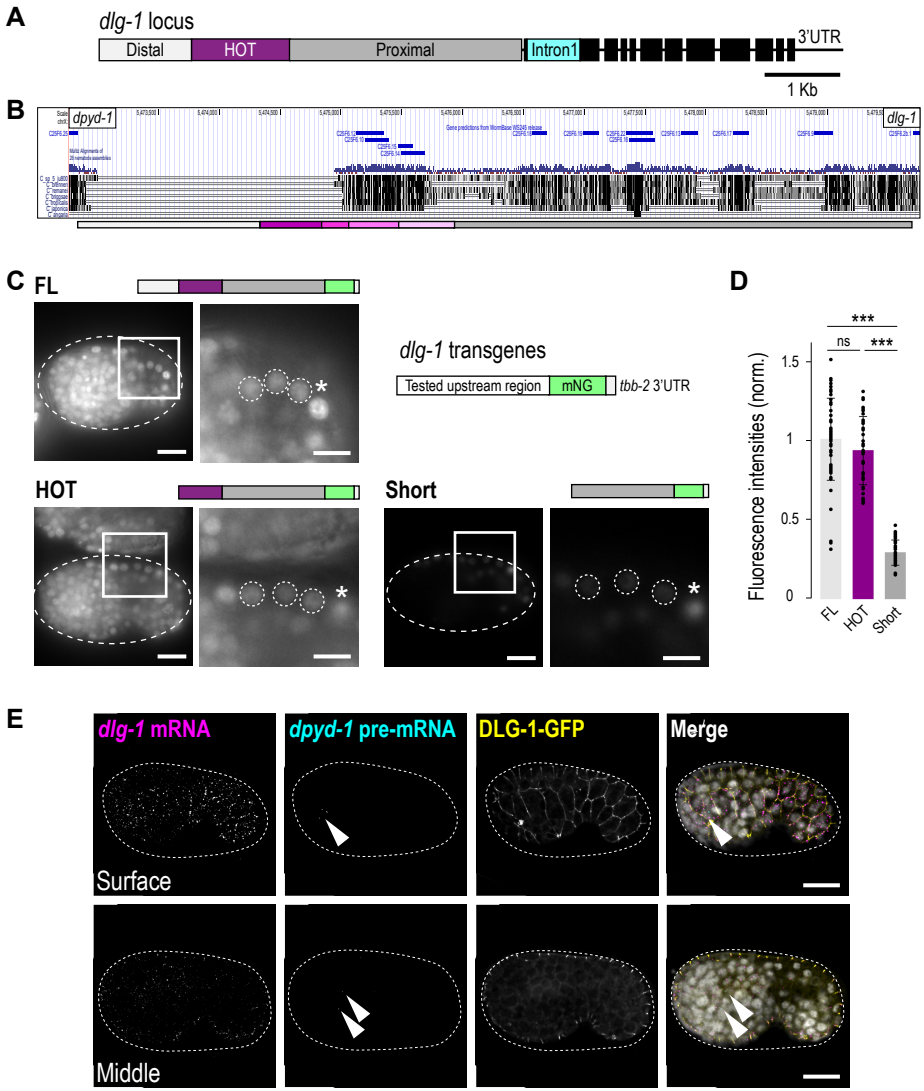

Fig.1, Tocchini *et al.*

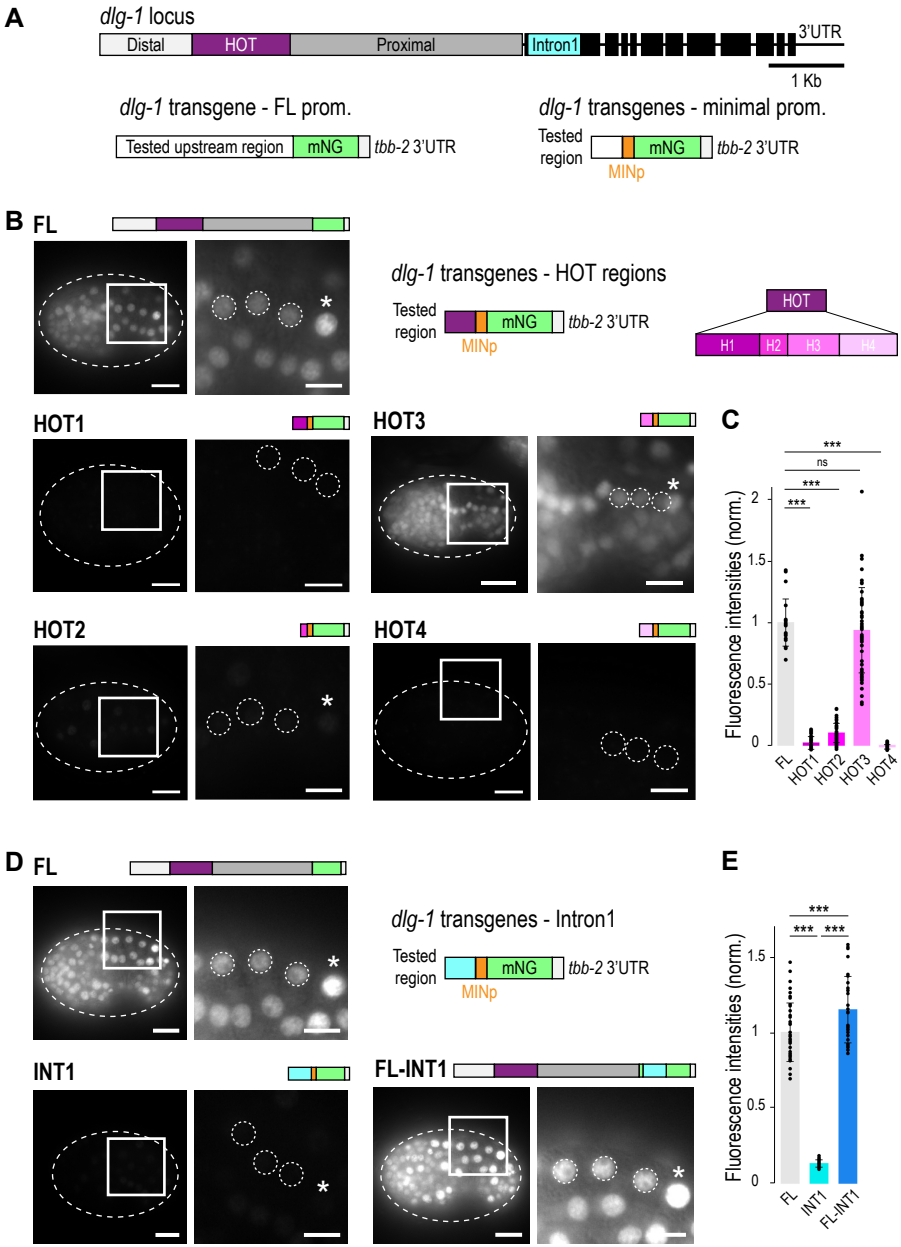

Fig.2, Tocchini *et al.*