© 2020. Published by The Company of Biologists Ltd | Development (2020) 147, dev190330. doi:10.1242/dev.190330

RESEARCH ARTICLE

How affinity of the ELT-2 GATA factor binding to *cis*-acting regulatory sites controls *Caenorhabditis elegans* intestinal gene transcription

Brett R. Lancaster* and James D. McGhee‡

ABSTRACT

We define a quantitative relationship between the affinity with which the intestine-specific GATA factor ELT-2 binds to *cis*-acting regulatory motifs and the resulting transcription of *asp-1*, a target gene representative of genes involved in *Caenorhabditis elegans* intestine differentiation. By establishing an experimental system that allows unknown parameters (e.g. the influence of chromatin) to effectively cancel out, we show that levels of *asp-1* transcripts increase monotonically with increasing binding affinity of ELT-2 to variant promoter TGATAA sites. The shape of the response curve reveals that the product of the unbound ELT-2 concentration *in vivo* [i.e. ($ELT\text{-}2_{free}$) or ELT-2 'activity'] and the largest ELT-XXTGATAAXX association constant (K_{max}) lies between five and ten. We suggest that this (unitless) product [$K_{max}\times(ELT\text{-}2_{free})$ or the equivalent product for any other transcription factor] provides an important quantitative descriptor of transcription-factor/regulatory-motif interaction in development, evolution and genetic disease. A more complicated model than simple binding affinity is necessary to explain the fact that ELT-2 appears to discriminate *in vivo* against equal-affinity binding sites that contain AGATAA instead of TGATAA.

KEY WORDS: Transcription, GATA Factor, ELT-2, Binding affinity, *cis*-acting regulatory motif, *C. elegans*, Intestine, *asp-1*, Protease

INTRODUCTION

During animal development, the transcription of a single gene by RNA Polymerase II is regulated by scores, perhaps hundreds, of different proteins (Carey et al., 2009; Workman and Abmayr, 2014; Peter and Davidson, 2015, 2016; Furlong and Levine, 2018). In this study, we focus on arguably the earliest and most instructive step in this overall process: the binding of a specific activating transcription factor to a *cis*-acting regulatory motif in the control region of a developmentally regulated tissue-specific target gene. We wish to understand how the affinity of interaction between this transcription factor and its binding site influences the level or rate of target gene transcription. This general problem has been approached multiple times in the past, most often in yeast or cultured cells; however, it has been surprisingly difficult to settle on an unambiguous, let alone universal, answer. Many previous studies have concluded that increased binding affinity of a transcriptional activator does indeed have a positive influence on target gene transcription (Bain et al., 2012); however, other studies have found target gene transcription to be insensitive to transcription factor affinity or even anti-correlated (Meijsing et al., 2009). Experimental limitations have included unknown levels of free (unbound) transcription factors *in vivo* following induction or transfection, i.e. incompletely defined binding isotherms (Bain et al., 2012). In many studies, transcription factor affinity has been only one parameter among many that determines target gene transcription levels. Other parameters include: (1) chromatin accessibility (Grossman et al., 2017), which, in cases in which this has been looked at more closely, can reveal a detailed interplay between transcription factor affinity and nucleosome positioning (Lam et al., 2008; Rajkumar et al., 2013); (2) nearby binding of auxiliary transcription factors (Sasse et al., 2015; Grossman et al., 2017); (3) the form of the embedding regulatory network (e.g. feed-forward loops), especially in time-dependent systems (Sasse et al., 2015); and (4) more elaborate mechanisms, such as proposed allosteric changes in transcription factor conformation dictated by a particular DNA-binding sequence (Meijsing et al., 2009; Weikum et al., 2017).

In this study, we define a quantitative relationship between transcription factor binding affinity and target gene transcript levels for a gene associated with the differentiation of a specific cell lineage within a developing multicellular animal. The *Caenorhabditis elegans* intestine is a clonally derived and relatively homogeneous set of cells (Sulston et al., 1983), the differentiation of which is largely controlled by a single transcriptional activator, the zinc-finger GATA factor ELT-2 (Hawkins and McGhee, 1995; McGhee et al., 2007, 2009; Dineen et al., 2018). The gene selected as an ELT-2 target is *asp-1*, which encodes the *C. elegans* intestinal-specific aspartic acid protease ASP-1 (Tcherepanova et al., 2000), the transcription of which is almost entirely dependent on ELT-2 (McGhee et al., 2009; Dineen et al., 2018). We establish an experimental system that allows unknown parameters (e.g. the influence of 'chromatin') to effectively cancel out, thereby allowing us to isolate the transcriptional consequences of normal physiological levels of ELT-2 binding to variable-affinity XXTGATAAXX sites in the *asp-1* promoter. We show that: (1) levels of *asp-1* transcripts increase monotonically with increasing binding affinity of ELT-2 to variant promoter XXTGATAAXX sites; (2) the shape of the response curve determines an important relationship between the unbound ELT-2 concentration *in vivo* [i.e. ($ELT\text{-}2)_{free}$ or ELT-2 'activity'] and the tightest association constant (K_{max}) to a TGATAA site; and (3) ELT-2 is able to functionally discriminate *in vivo* against binding sites that contain AGATAA rather than TGATAA, even though the binding

Department of Biochemistry and Molecular Biology, University of Calgary, Cumming School of Medicine, Alberta Children's Hospital Research Institute, Calgary, Alberta T2N 4N1, Canada.
*Present Address: AstraZeneca Canada, 1004 Middlegate Road, Mississauga, Ontario L4Y 1M4, Canada.

‡Author for correspondence (jmcghee@ucalgary.ca)

J.D.M-G., 0000-0001-7266-7275

Handling Editor: Susan Strome
Received 7 March 2020; Accepted 6 June 2020

Development (2020) 147, dev190330. doi:10.1242/dev.190330

affinity to these two different sequences can be closely comparable, i.e. for non-TGATAA target sites, a more complicated model than simple ELT-2 binding affinity must be invoked.

RESULTS

Genes expressed in the differentiated *C. elegans* intestine are controlled by extended TGATAA sites

We found 44 examples in which experimental mutation of *cis*-acting sequence motifs significantly diminished the expression of particular genes in the differentiated *C. elegans* intestine. (details and references are collected in Table S1). Fig. 1 shows the summarizing sequence logo; the predominant site was clearly a TGATAA sequence but with significant information content in the flanking two base pairs, both upstream and downstream. It is well established that TGATAA-like sites are enriched [and (A/C/G)GATAA-like sites are correspondingly depleted] in the regulatory regions of all genes transcribed in the *C. elegans* intestine, from embryos to adults [Pauli et al., 2006; McGhee et al., 2007, 2009; Dineen et al., 2018; Table S2 reproduces the position frequency matrix from McGhee et al. (2009)]. We have argued that these sites are primarily the direct targets of the intestine-specific GATA-type transcription factor ELT-2: ELT-2 protein binds to similar sites both *in vitro* (Hawkins and McGhee, 1995; Goszczynski et al., 2016; Wiesenfahrt et al., 2016) and *in vivo* (Mann et al., 2016; Wiesenfahrt et al., 2016). A subset of these sites is also a direct target of a second intestinal GATA-factor, ELT-7 (Dineen et al., 2018); however, ELT-7 can be removed without overt consequences (McGhee et al., 2007; Sommermann et al., 2010; Dineen et al., 2018) and our *in vivo* experiments were conducted in its absence.

Additionally, in the 12 base pairs upstream and 12 base pairs downstream of the XXTGATAAXX motif, the information content of these collected sites was essentially at background levels (Fig. 1), consistent with the absence of a co-factor binding in a constant and close relationship to ELT-2. Although there are certainly genes expressed in the *C. elegans* intestine that are co-regulated by ELT-2 and some other transcription factor (Neves et al., 2007; Sinclair and Hamza, 2010; Romney et al., 2011; Roh et al., 2015; Goszczynski et al., 2016), the relative disposition of the factors varies between different co-regulated promoters. We interpret the sequence logo data (Fig. 1) to be consistent with a model in which the isolated binding of ELT-2 by itself provides the dominant contribution to target gene activation, an important simplification for our analysis.

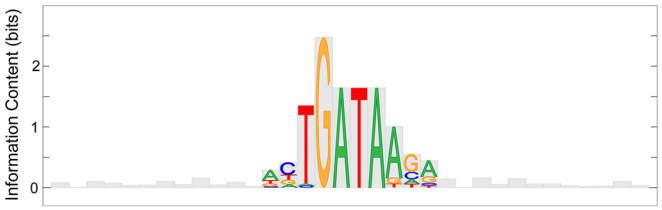


Fig. 1. Genes expressed in the differentiated *C. elegans* intestine are controlled by extended TGATAA sites. Sequence logo displaying the information content of 44 …GATA… sequences that activate transcription of a variety of *C. elegans* intestinal genes (details and references are provided in Table S1). Information content is calculated using the base composition of the full promoters (2000 repeat-masked base pairs upstream of the initiation codon) as background (64% AT). Information content is shown for 12 bp upstream and downstream of the extended XXTGATAAXX motif.

In vitro binding affinity of ELT-2 to the TGATAA motif is strongly influenced by flanking dinucleotides

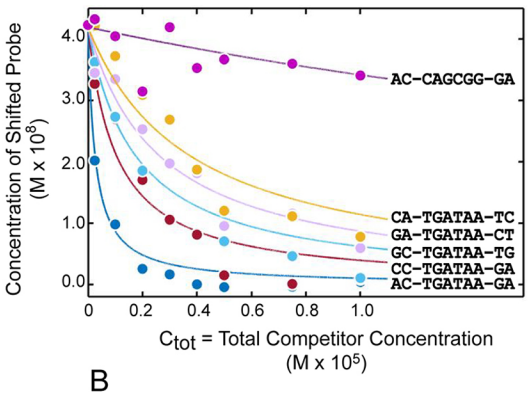
An *in vitro* competitive band shift assay [electrophoretic mobility shift assay (EMSA)] was used to measure the affinity of (full-length) ELT-2 protein binding to a series of XXTGATAAXX sequences, relative to its binding affinity to the preferred sequence ACTGATAAGA (Fig. 1); this preferred sequence will turn out to have the highest affinity but such agreement is not necessary (see below). Experimental details of the binding competition are provided in the Materials and Methods section. The supplementary Materials and Methods describes how the competition data were analysed in order to produce estimates of $K_{rel}=K_C/K_A$, i.e. the ratio of the ELT-2 binding affinity to competitor oligodeoxynucleotide *C* (association constant K_C M^{-1}) to the ELT-2 binding affinity to the labelled (and most tightly binding) oligodeoxynucleotide *A* (association constant K_A M^{-1}). Representative gel images are shown in Fig. S1; representative competition isotherms are shown in Fig. 2A; and numerical estimates of K_{rel} for the series of XXTGATAAXX motifs used in this study are presented in Fig. 2B. The primary conclusion from this section is that alterations in two base pairs upstream and downstream of the core TGATAA motif can modulate ELT-2-binding affinity by ∼tenfold.

ELT-2 binds to TGATAA and AGATAA motifs with comparable affinity

GATA factors in vertebrates bind to a *cis*-regulatory motif of the general form (A/T)GATA(A/G) (Patient and McGhee, 2002). Indeed, *in vitro* measurements show that the residue preceding the core GATA-binding sequence of vertebrate GATA factors is an A or T with approximately equal frequency (Khan et al., 2017). In contrast, the functional motifs that regulate intestinal genes in *C. elegans* show much lower degeneracy (Fig. 1, Table S1) and TGATA appears to be favoured over AGATA by ∼30-fold (see also Table S2). We wished to test whether this increased specificity of the functional GATA motifs in *C. elegans* is imposed by the intrinsic sequence preferences of ELT-2 binding or by some other feature of the transcriptional process. Fig. 2C shows the results of a competitive EMSA experiment in which the ELT-2-binding affinity to an …AC**A**GATAAGA… containing double-stranded oligodeoxynucleotide is compared with that of the otherwise identical …AC**T**GATAAGA… containing oligodeoxynucleotide. Contrary to the implications of the data featured in Fig. 1 and Tables S1,S2, ELT-2 binds *in vitro* to the oligodeoxynucleotide containing the AGATAA motif with ∼45% of the affinity with which it binds to the otherwise identical TGATAA control motif. This conclusion is validated and extended by an experimental approach in which multiple degenerate double-stranded oligodeoxynucleotides are incubated with ELT-2, and the bound and unbound fractions electrophoretically separated, followed by sequencing [a low resolution implementation of the Spec-Seq procedure (Zuo and Stormo, 2014; Stormo et al., 2015)]. As explained in more detail in the supplementary Materials and Methods, we estimate that ELT-2 binds to an XAGATA sequence with 78±16% or 94±33% of the affinity that it binds to an XTGATA sequence, depending upon whether the identity of X is considered or ignored, respectively. Thus, the intrinsic *in vitro* sequence preference of ELT-2 appears to be similar to that of vertebrate GATA factors. However, we demonstrate below that the *in vivo* transcriptional potency of an AGATA motif is much lower than that of a TGATA motif in spite of comparable binding affinity to ELT-2.

Development (2020) 147, dev190330. doi:10.1242/dev.190330

A



B

Sequence	K_{rel} = Affinity Relative to ACTGATAAGA		
	#1	#2	Mean ± SD
ACTGATAAGA	1	1	1
AATGATAAGA	0.48	0.43	0.45 ± 0.03
ACTGATAATA	0.39	0.37	0.38 ± 0.01
CCTGATAAGA	0.25	0.21	0.23 ± 0.02
GCTGATAATG	0.15	0.13	0.14 ± 0.01
GATGATAACT	0.10	0.08	0.09 ± 0.01
CATGATAATC	0.07	0.11	0.09 ± 0.02
ACCAGCGGGA	0.01	0.00	0.00 ± 0.00
ACAGATAAGA	0.45	0.43	0.44 ± 0.01

C

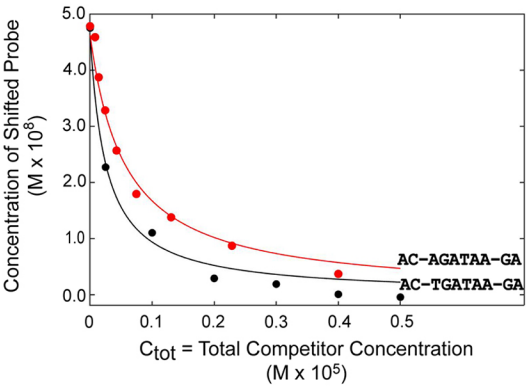


Fig. 2. Analysis of competitive band shift assays in order to determine K_{rel}. (A) Full-length ELT-2 protein was incubated with a mixture of two self-complementary hairpin oligodeoxynucleotides. Oligodeoxynucleotide A contained the highest affinity sequence, …ACTGATAAGA…, and was labelled with fluorescein. Competitor oligodeoxynucleotide C contained a variant, …XXTGATAAXX…, and was unlabelled. Bound and free species were separated by electrophoresis and the amount of bound *A* was measured by fluorescence, as a function of increasing amounts of total competitor *C*. Data were analyzed as described in the supplementary Materials and Methods in order to obtain numerical estimates of the relative binding constant $K_{rel}=K_C/K_A$. These numerical estimates were used to calculate the competition curves; two sets of representative data are shown. (B) Numerical estimates of K_{rel} derived from competition curves, such as those shown in A, and as explained in more detail in the text and in the supplementary Materials and Methods. Estimates of K_{rel} in column 1 were derived by direct competition with the highest affinity sequence, …ACTGATAAGA…; estimates of K_{rel} in column 2 were obtained independently by direct competition with the more weakly binding sequence, …GCTGATAATG…; K_{rel} was then calculated by simple ratio. (C) Competitive band shift assay to show that ELT-2 binds to an ACAGATAAGA motif with approximately half of the affinity that it binds to the most tightly bound motif, ACTGATAAGA.

Quantifying the influence of XXTGATAAXX motif affinity on *in vivo* transcription rates of a *C. elegans* intestine-specific gene

In order to measure the transcriptional consequences *in vivo* of ELT-2 binding to a particular XXTGATAAXX site (or sites) in the promoter of an intestinal gene, we developed an experimental system that we refer to, for shorthand, as SQRIPT (simultaneous quantitation of reporter transcripts). *C. elegans* is routinely transformed by injecting plasmids into the syncytial gonad of the adult hermaphrodite; the injected plasmids assemble into an extrachromosomal multicopy array that might contain a hundred copies (or more) of the transforming plasmids (Mello et al., 1991; Stringham et al., 1992; Meister et al., 2010), which are passed on to ∼50% of next-generation animals. Most experimental analyses of transcriptional regulation in *C. elegans* have been performed using these arrays; the general consensus is that genes expressed from these transgenic arrays are correctly regulated, at least to a good first approximation, and reports of misregulation are rare (Hope, 1991; Boulin et al., 2006). The properties of these multicopy transgenic arrays provide the key rationale for the SQRIPT assay: that control and test constructs can be made to differ at only a small number of base pairs (typically fewer than ten). These constructs can then be incorporated in equal stoichiometry into the arrays, such that each experimentally manipulated test construct can be compared with an unperturbed control construct in the same (ideally identical) environment. Additional features of SQRIPT will be noted once more specific properties of the assay are described.

Our current version of SQRIPT is based on the *C. elegans asp-1* gene, which encodes a major intestine-specific aspartic acid protease [a homologue of cathepsin D (Tcherepanova et al., 2000)]. *asp-1* transcripts are first detected in late embryogenesis, reach peak levels in mid-larval stages and then decline modestly (∼twofold) in adulthood (data from modENCODE assembled in www.wormbase.org). The *asp-1* gene has no introns, is highly expressed and transcript levels are reduced 40- to 50-fold in an *elt-2* null mutant (measured at the arrested L1 stage) (McGhee et al., 2009; Dineen et al., 2018). There are eight TGATAA sites distributed over 6.5 kb of upstream flanking region but for our experiments, we confined our analysis to the ∼1.4 kb immediately upstream of the ATG start codon, which has been shown previously to drive intestine-specific reporter expression (Tcherepanova et al., 2000). As shown in Fig. 3A, this region contains two TGATAA sites lying just upstream of the *asp-1* transcription initiation site; ChIP-Seq experiments detect ELT-2 binding to this region *in vivo*, with the only significant ELT-2 peak aligning with the two TGATAA sites (Wiesenfahrt et al., 2016).

We produced two variants of the *asp-1* coding region by introducing a KpnI site at different positions so that transcripts produced by the two reporters (R1 and R2) *in vivo* can be distinguished (Fig. 3A). Each reporter differed by one base pair from the wild-type sequence and by two base pairs from each other; the encoded proteins remained unchanged. The basic assay is shown schematically in Fig. 3B. In a typical experiment, a variant of the *asp-1* 1.4 kb promoter fragment (e.g. with a mutated TGATAA site) is used to drive the expression of *asp-1* reporter R2; the wild-type version of the promoter is used to drive the expression of *asp-1* reporter R1. Equal amounts of these two constructs, test and control, are mixed with an *unc-119*-rescuing plasmid (Maduro, 2015) and injected into host strain JM189 [*unc-119(ed3)* III*; elt-7(tm840) asp-1(tm666)* V*; elt-4(ca16)* X]. (Although the ELT-7 and ELT-4 endodermal GATA factors make little or no contribution to *asp-1* transcription, respectively, incorporating the null mutations into the

Development (2020) 147, dev190330. doi:10.1242/dev.190330

host strain removes the possibility that either could act through experimentally introduced variant TGATAA sites.) For each construct being tested, several independent transgenic strains are produced, propagated and harvested. Both RNA and DNA are isolated. RNA is reverse transcribed and amplified by PCR using the *asp-1* primers shown in Fig. 3A; the resulting cDNA is digested with KpnI and digestion products are separated by electrophoresis; the full sequence of reactions is performed in triplicate. In order to correct for any inequality in reporter stoichiometry, R1 and R2 copies in the genomic DNA are amplified using the same primers and the relative amounts of KpnI digestion products quantified. As will be shown in the following sections, the SQRIPT assay has a dynamic range of 10- to 20-fold and a precision of ∼10% in measuring the relative transcriptional activity of any particular promoter-modified construct.

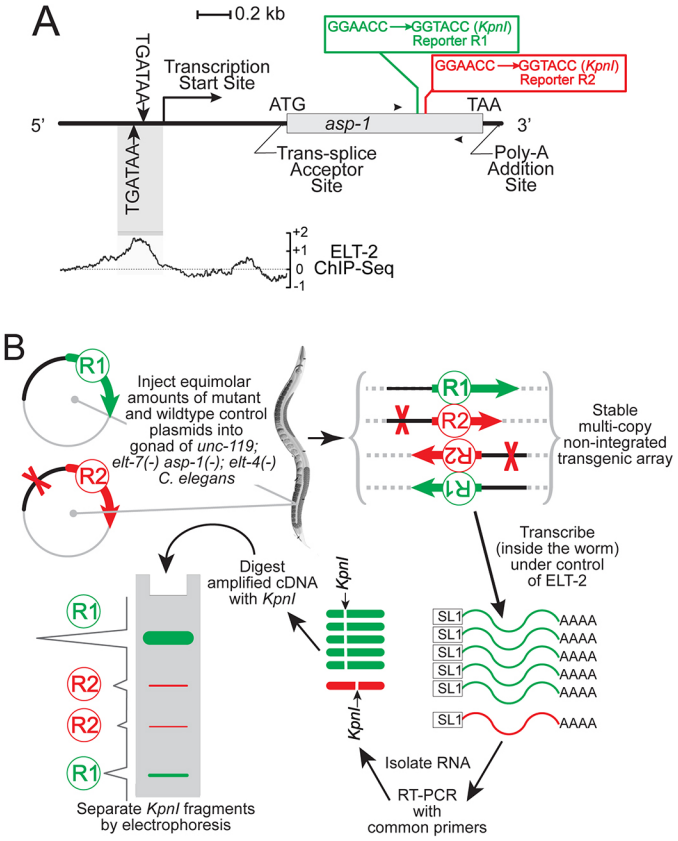


Fig. 3. The SQRIPT assay for simultaneous quantitation of reporter transcripts. (A) The *C. elegans asp-1* gene encodes a highly expressed intestine-specific aspartic protease. *asp-1* transcription is controlled by two TGATAA sites that lie immediately upstream of the transcription start site and that align with the only significant peak (black bar) of ELT-2 bound *in vivo* [ChIP-seq data from Wiesenfahrt et al. (2016)]. Two distinguishable reporter versions of the *asp-1* gene, R1 (green) and R2 (red), were constructed by the insertion of KpnI sites, as indicated. (B) Key steps in the SQRIPT assay to compare the transcriptional influence of two versions of the *asp-1* promoter. Reporter R1 is controlled by the wild-type *asp-1* promoter; reporter R2 is controlled by an *asp-1* promoter variant (indicated by the red 'X'). Equal concentrations of R1 and R2 plasmid DNA are injected into strain JM189 (*unc-119 III; elt-7 asp-1V; elt-4 X)*. Transgenic animals are identified by UNC-119 rescue and propagated as a stable multicopy transgenic strain. The arrangement of reporters R1 and R2 in the array is not known and could occur in both orientations. The basis of the SQRIPT assay is that, overall, the environments of R1 and R2 are expected to be highly similar. RNA is isolated, reporter cDNA is synthesized by RT-PCR and the different levels of reporter R1 and R2 are measured quantitatively by KpnI digestion and subsequent electrophoresis to separate the distinguishable digestion products.

TGATAA sites act synergistically to activate *asp-1* transcription *in vivo*

Fig. 4A shows the relative transcript levels measured when both reporters (R1 and R2) are activated by the same wild-type promoter; the relative transcript levels were measured as 1.06±0.15 (mean±s.d.), i.e. there was no significant bias *in vivo* between the two reporters (unpaired, two-tailed Student's *t*-test $P>0.2$). Fig. 4A also shows that the destruction of either of the two TGATAA sequences reduced reporter transcript levels to 10 to 20% of the level measured with the wild-type reporter. In other words, these two motifs are acting neither redundantly (in which case, reporter transcript levels would have remained unchanged in the single mutants) nor additively (single mutant reporter transcript levels would have been approximately half of wild-type levels), but rather the two sites appear to be acting synergistically or cooperatively. This synergy is not complete because reporter transcripts were reduced by a further 50-60% if both TGATAA sites were destroyed simultaneously (unpaired, two-tailed Student's *t*-test $P<0.001$). The synergistic/

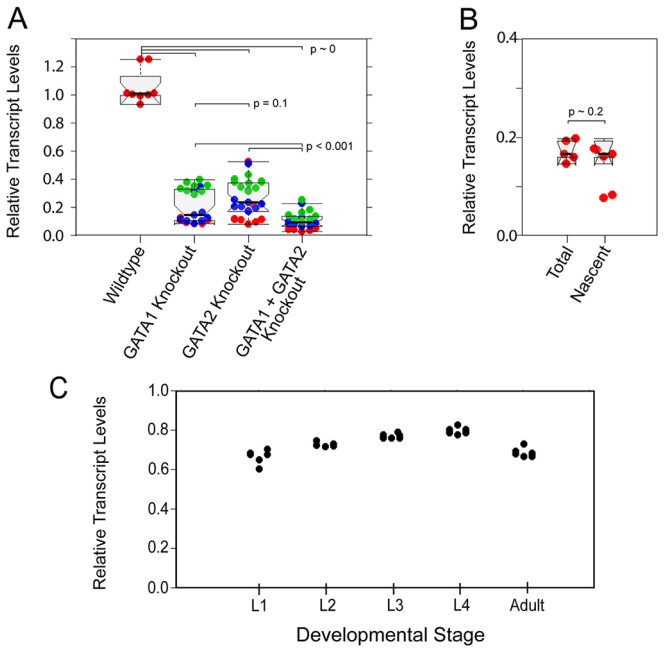


Fig. 4. SQRIPT assay characterization of *asp-1* transcription. (A) The two TGATAA sites in the *asp-1* promoter act synergistically to drive reporter expression. 'Wildtype' data represent the relative transcript levels measured when the expression of both reporter R1 and R2 are driven by the wild-type *asp-1* promoter. 'GATA1 Knockout' and 'GATA2 Knockout' data measure the effect on relative reporter transcript levels of ablating the upstream or the downstream *asp-1* promoter TGATAA site, respectively. 'GATA1+GATA2 Knockout' data measure relative reporter transcript levels produced when both upstream and downstream TGATAA sites are ablated. Different colour points correspond to data obtained from independent transgenic strains; different points of the same colour correspond to data obtained from replicate assays with a single transgenic strain. Unpaired, two-tailed Student's *t*-test probabilities are as indicated. (B) Relative reporter transcript levels are the same when measured from either total RNA or from nascent RNA. Individual data points represent replicate assays. Unpaired, two-tailed Student's *t*-test probability indicated. (C) Relative reporter transcript levels do not strongly depend on the developmental stage. Relative transcript levels were measured for transgenic strains in which reporter R2 was driven by a promoter containing two copies of a CCTGATAAGA motif replacing the wild-type TGATAA versions. Plots were assembled using RStudio; whiskers encompass all data points not judged to be outliers; boxes represent the interquartile range (i.e. 25–75% of the data).

Development (2020) 147, dev190330. doi:10.1242/dev.190330

cooperative behaviour of the two *asp-1* TGATAA sites was qualitatively validated using GFP as a reporter (Fig. S2). We draw the following conclusions from Fig. 4A: (1) the two *asp-1* promoter TGATAA sites act largely but not completely synergistically; (2) the five GATA sites in the 1.4 kb *asp-1* promoter that are not TGATAA make only minor contributions to promoter activity; and (3) independent transgenic strains produced with the same injection mixture give similar results.

Fig. 4B,C describes two further important features of the SQRIPT assay. To test whether reverse transcription of nascent RNA produced the same estimate of relative transcript levels as did reverse transcription of total RNA, nuclear run-ons were performed according to Kruesi et al. (2013) with nascent mRNA being affinity isolated based on incorporation of bromouridine. As shown in Fig. 4B, relative reporter transcript levels were not significantly different (*t*-test, P=0.17) whether they were measured using total or nascent RNA, suggesting that the SQRIPT assay measures differences in the rates of transcript initiation. Although an effect on transcript elongation or degradation cannot be ruled out, such an explanation would seem unlikely considering the high degree of similarity between the two transcript sequences and the equivalent results produced when reporters are interchanged. Fig. 4C shows that the relative reporter transcript levels produced by a modified *asp-1* promoter changed only modestly from embryo to adult. Supporting this observation, Fig. S3 shows similar data obtained with two different *asp-1* variant promoters. A practical consequence of these results is that conclusions will not be strongly influenced by imperfect age-matching of different samples from different strains.

ELT-2 affinity to the XXTGATAAXX promoter motifs controls *asp-1* transcription *in vivo*

XXTGATAAXX sequences with known K_{rel} (Fig. 2) were inserted into the SQRIPT reporters, such that each variant reporter had two copies of the same variant replacing the two TGATAA copies in the wild-type *asp-1* promoter. Three independent transgenic strains were produced for each construct and the transcript levels of the variant reporters were measured (at the L4/young adult stage) relative to transcript levels of wild-type control reporters incorporated into the same transgenic array. Fig. 5 plots the relative transcript levels measured for a particular test promoter versus the relative ELT-2 affinity constant (K_{rel}) measured *in vitro* for the TGATAA variant present (as pairs) in each promoter. The important conclusions are that: (1) transcriptional activity of a variant *asp-1* promoter is highest when both XXTGATAAXX sites correspond to the strongest ELT-2 binding sequence, ACTGATAAGA; and (2) transcript levels decrease monotonically as ELT-2 affinity decreases. The shape of the 'relative transcript levels versus K_{rel}' response curve has important implications for ELT-2/target gene behaviour *in vivo* and we therefore explored this more quantitatively.

We showed above in Fig. 4A that reporter transcription was greatly reduced when either of the two TGATAA sequences in the *asp-1* promoter were ablated. We now explore an initial model in which a variant *asp-1* promoter activates reporter transcription if and only if both of the two TGATAA sites are occupied by bound ELT-2. As described in more detail in the supplementary Materials and Methods, the relationship between (y=measured relative transcript level) and (x=K_{rel}) is:

$$y/y_{max} = [(x \times K_{max} \times [\text{ELT-2}_{free}])/(1 + x \times K_{max} \times [\text{ELT-2}_{free}])]^2.$$

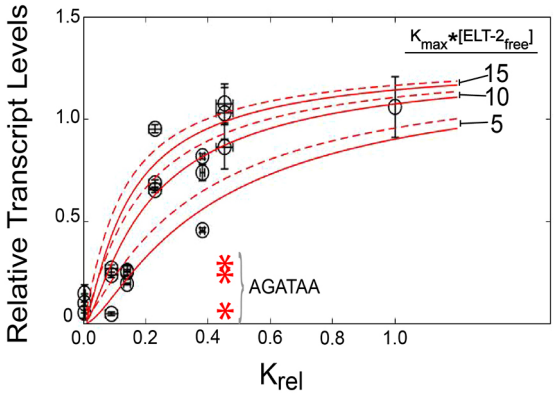


Fig. 5. Relationship between reporter transcript levels and the ELT-2-binding affinity to extended XXTGATAAXX sites in the *asp-1* promoter. Test reporters were constructed in which both of the wild-type TGATAA sites were replaced by variant XXTGATAAXX, in which the two base pairs flanking the core TGATAA sites were varied to produce values of K_{rel} ranging from ~0 to 1. Transgenic strains were produced and relative reporter transcript levels were measured using the SQRIPT assay. Individual data points (circles) for the same K_{rel} represent independent transgenic strains produced by the same reporter constructs; error bars derive from replicate assays within one transgenic strain. Solid red lines are calculated as described in the text, assuming that the two TGATAA sites are completely synergistic and with trial K_{max}×[ELT-2$_{free}$] values of 5, 10 or 15, and for a single trial value of maximum relative transcription activity of 1.3 (at infinite ELT-2 levels). The dashed lines are calculated using the same parameters but allowing for partial synergy between the two TGATAA sites. The red asterisks correspond to the relative transcript levels measured for a reporter in which the TGATAA sites in the *asp-1* promoter were replaced with variant AGATAA sites with K_{rel} corresponding to 0.45 (see Fig. 2B,C).

The two parameters to be derived from the curve shapes of Fig. 5 are: (1) y_{max}=the maximum relative transcript level that would be obtained at 'infinite' ELT-2 concentrations; and (2) the unitless product K_{max}×[ELT-2$_{free}$], where K_{max} is the absolute affinity (association constant) of ELT-2 to the most preferred sequence ACTGATAAGA and [ELT-2$_{free}$] is the normal effective free concentration of ELT-2 protein *in vivo* (i.e. ELT-2 activity). (K_{max} refers to ELT-2 affinity *in vivo* and may or may not be equivalent to K_A used in analyzing the *in vitro* binding competitions described above. All we are proposing is that relative affinities of different motifs are the same *in vivo* and *in vitro*.) Minimizing the sums of the squares of the deviations of the measured data points from the trial-parameterized relationship defined above (see supplementary Materials and Methods) shows that y_{max} must be in the range of 1.1 to 1.5 but the fit is generally insensitive to choice. The more important conclusion is that the product K_{max}×[ELT-2$_{free}$] must be in the range of 10±5, i.e. clearly greater than 1. Examples of fits to the data are shown in Fig. 5 for the case of y_{max}=1.3 and the product K_{max}×[ELT-2$_{free}$] chosen as 5, 10 or 15 (solid lines). The implications of these particular parameter values will be discussed below. The dashed lines in Fig. 5 show a similar analysis for a model in which either single or double occupancy of the XXTGATAAXX sites can activate reporter transcription to the extent measured in Fig. 4A (see supplementary Materials and Methods for more details); as expected, this extension produces only a modest change in the calculated binding curve.

AGATAA sequences do not obey the relative transcript level versus K_{rel} relation defined for TGATAA sequences

We are now in a position to resolve a potential paradox that emerges from the above analyses. As just discussed, Fig. 5 shows that the

Development (2020) 147, dev190330. doi:10.1242/dev.190330

transcriptional activity of reporters controlled by XXTGATAAXX sequences is dominated by the affinity of ELT-2 for these variant motifs. Fig. 2C (and supplementary Materials and Methods) shows that ELT-2 binds to AGATA sequences with close to the same affinity that it binds to the matched TGATA sequence. Yet, the sequence logo shown in Fig. 1 indicates that functional *cis*-acting regulatory motifs in *C. elegans* intestinal promoters are strongly favoured to be TGATAA rather than AGATAA. We investigated whether an AGATAA sequence replacing a TGATAA in our transcriptional reporters would show the same transcriptional behaviour if these two sequences had the same affinity for ELT-2. We thus compared the behaviour of two different core motifs, AATGATAAGA and ACAGATAAGA, that were chosen as they have the same relative affinity for ELT-2 measured *in vitro* (Fig. 2). Two copies of the selected ACAGATAAGA motif were inserted into the appropriate SQRIPT reporters replacing the wild-type TGATAA motifs, transgenic animals were produced and relative transcript levels were measured. The AGATAA-dependent relative transcript levels are plotted on Fig. 5 as the red asterisks. We conclude that *asp-1* promoters in which a TGATAA sequence is replaced by an AGATAA sequence with the same ELT-2 affinity do not obey the relative transcript level versus relative ELT-2-binding affinity relationship dynamic defined for TGATAA sites. In fact, the AGATAA-containing promoter approaches inactivity, whereas the promoter containing the TGATAA equal-affinity counterpart approaches maximum transcriptional activity. We conclude that there must be at least one additional layer of specificity beyond simple ELT-2 affinity that controls the transcription of intestinal genes.

DISCUSSION

Properties of the core TGATAA motifs that drive intestinal gene expression in *C. elegans*

Among the collection of *cis*-acting GATA motifs shown experimentally to influence *in vivo* expression of *C.elegans* intestinal genes (Fig. 1, Table S1), the most frequent core motif is TGATAA but with significant additional information present in the two base pairs upstream and the two base pairs downstream. Variations in these flanking dinucleotides can lead to a ∼tenfold variation in the binding affinity to ELT-2 (Fig. 2). Variations in the nucleotides or dinucleotides immediately flanking the core binding motifs of other transcription factors have also been shown to modulate interaction affinities (Levo et al., 2015; Schöne et al., 2016; Rudnizky et al., 2018).

As judged by *in vivo* functional assays (Fig. 1), as well as by computational identification of over-represented promoter motifs (Pauli et al., 2006; McGhee et al., 2007, 2009; Dineen et al., 2018), the most frequent decameric sequence controlling intestinal genes in *C. elegans* is ACTGATAAGA. This same sequence turns out to have the highest binding affinity to ELT-2 (Fig. 2) but such correspondence is not necessary; there are both bacterial and eukaryotic examples in which evolution appears to have selected lower affinity 'sub-optimal' transcription factor binding sites in gene promoters (Sadler et al., 1983; Crocker et al., 2015; Farley et al., 2015). We thus wished to investigate whether the degree to which a particular extended TGATAA motif (not just ACTGATAAGA) is over-represented in intestinal promoters of *C. elegans* reflects its binding affinity to ELT-2. Table S2 contains a position frequency matrix [PFM; reproduced from McGhee et al. (2009)] collecting over-represented sequence motifs computationally identified in the promoters of intestine-specific genes expressed in embryos, larvae and adults. Each PFM entry was converted to a log-odds 'statistical weight' (see Eqn 7-3 by Stormo, 2013), summed over the ten entries corresponding to the 10 bp binding sequences that had their relative ELT-2-binding affinities measured in Fig. 2A,B. This overall statistical weight was then plotted versus the logarithm of the corresponding K_{rel} (i.e. both variables should then be proportional to a free energy). As seen in Fig. S4, the relationship is satisfyingly linear. We interpret this linearity to suggest that *cis*-acting TGATAA motifs regulating *C. elegans* intestinal transcription are selected, at least in part, on the basis of their binding affinity to ELT-2: the higher the binding affinity to ELT-2, the more likely it is that the motif is present in intestinal promoters.

We note a potentially interesting feature of the endodermal promoter TGATAA sites in *C. elegans*: at least a subset of these sites are functional targets of ELT-7 in addition to ELT-2, and possibly, at least in the early embryo, of END-1/END-3 as well (Dineen et al., 2018). Although the current experiments were performed in the absence of ELT-7 and after END-1/3 have decayed, one could imagine that the information-rich gene-controlling sequences shown in Fig. 1 (and Table S1) represent some evolutionary or physiological compromise between different sequence preferences for the four individual endodermal GATA factors. However, we also note that *C. elegans* endodermal GATA factors appear to possess a remarkable degree of interchangeability; in particular, if placed under the appropriate promoters, both ELT-2 and ELT-7 can individually replace all three of the other endodermal GATA factors (Wiesenfahrt et al., 2016; Dineen et al., 2018). Plausible scenarios have been proposed to explain how this interchangeability could have arisen during evolution (Wiesenfahrt et al., 2016; Maduro, 2020).

Features of SQRIPT

The experimental system that forms the basis of the current study has several advantages over previous methods of defining the relationship between transcription factor binding affinity and target gene activity. These advantages are as follows: (1) outputs of the two reporters are measured directly as transcripts rather than reporter proteins, turnover rates for the two reporter transcripts are likely to be more similar than for two different protein reporters, and overall, the assay measures relative transcription initiation rates, not elongation rates nor transcript stabilities (Fig. 4B); (2) chromatin arrangements over the two highly similar reporter gene sequences can reasonably be expected to be similar, which might not be the case for genes expressing two different protein reporters; (3) the transcription of reporters is regulated at the normal *in vivo* physiological levels of free ELT-2 protein, an important feature that will be considered below; and (4) the similar treatments and environments of test and control constructs allow reliable normalization. Expanding on this last feature, we suggest that the many unknown parameters associated with the *in vivo* regulation of transcription, e.g. nucleosome arrangements, histone modifications, biases between *in vitro* and *in vivo* affinity measurements, etc., are likely to be the same (or highly similar) for the test and control constructs. The major rationale of the SQRIPT assay is that the effects of these unknown parameters can be assumed to 'cancel out', thereby allowing the role of binding affinity in gene transcription to be measured in isolation.

TGATAA motif synergy and activity throughout development

Using the quantitative SQRIPT assay, we showed that the two TGATAA motifs in the *asp-1* promoter were neither redundant nor additive but acted in an almost completely synergistic or cooperative manner, i.e. ablation of either one of the two TGATAA motifs

Development (2020) 147, dev190330. doi:10.1242/dev.190330

lowered reporter expression to a level similar to that observed when both motifs were ablated (Fig. 4A). One molecular mechanism that could explain such synergy is that two (simultaneous) ELT-2/TGATAA-binding events are required to displace a resident inhibitory nucleosome [e.g. Morgunova and Taipale (2017); Zhu et al. (2018)]. Consistent with such a model, the two TGATAA sites in the *C. elegans asp-1* promoter are spaced 60 bp apart, well within the span of a single nucleosome core particle. Furthermore, apparently homologous pairs of TGATAA sites, spaced between 54 and 106 bp apart, can be found in *asp-1* promoters from related caenorhabditid nematodes (Fig. S5). [We note that, in each of these homologous promoters, one of the TGATAA motifs (ACTGATAAGA) is the sequence that binds most tightly.] Table S1 lists several further examples of *C. elegans* intestinal promoters with TGATAA sites that have been reported to act synergistically, at least to some degree; the distance between these paired sites ranges from 9 to 65 bp, all well below the size of a nucleosome core. In contrast, we have described the behaviour of the major *elt-2* enhancer in which four conserved TGATAA sites are spaced 186, 210 and 235 bp apart and act as if they are largely redundant (Wiesenfahrt et al., 2016). Further experiments will be required to test this cooperative nucleosome displacement model in which TGATAA sites that act synergistically are spaced less than 145 bp apart but TGATAA sites that act redundantly are spaced more than 145 bps apart.

We also used the SQRIPT assay to show that the relationship between promoter TGATAA affinity and reporter transcript levels remains approximately the same between newly hatched larvae and adults (Fig. 4C, Fig. S3). Such invariance suggests that the basic molecular mechanisms relating *asp-1* transcription to ELT-2 binding are qualitatively the same in different developmental stages, arguing against a model in which gene transcription later in life adopts a 'locked-in' configuration in which individual transcription factors such as ELT-2 have been supplanted by, for example, a stably propagating chromatin structure.

The gene response function relating ELT-2 affinity to *asp-1* transcript levels

The most important result in this study is shown in Fig. 5, namely the quantitative relationship between the relative transcript levels produced by a test promoter and the relative ELT-2 association constants (K_{rel}) for the pair of TGATAA sites in this particular promoter. The shape of this curve has important implications for understanding the molecular mechanisms by which ELT-2 interacts with *cis*-acting promoter motifs to drive intestinal gene transcription. Qualitatively, the free ELT-2 levels *in vivo* (i.e. [ELT-2_{free}]) cannot be so high that low-affinity sites are saturated (i.e. the curve of Fig. 5 is steep at low K_{rel}). Likewise, the free ELT-2 levels *in vivo* cannot be so low that high-affinity sites are far from saturation (i.e. the curve of Fig. 5 plateaus at higher K_{rel}). Using a simple thermodynamic model incorporating either complete or partial synergy between the paired TGATAA sites, we estimate that the product of K_{max}×[ELT-2_{free}] is ∼10, in which K_{max} is the *in vivo* ELT-2 association constant to the highest affinity XXTGATAAXX DNA sequence. Both parameters, K_{max} (1/M) and [ELT-2_{free}] (M), are difficult to measure individually but we suggest that the dimensionless product of K_{max}×[ELT-2_{free}] is the more useful parameter to know: it provides a quantitative measure of system responsiveness and the degree to which variants in *cis*-acting binding sites can be expected to influence associated transcription.

The gene response function shown in Fig. 5 summarizes the manner in which the *asp-1* promoter responds *in vivo* to ELT-2 interaction with the pair of TGATAA sites: binding affinity is paramount. However, ELT-2-binding affinity to a *cis*-acting motif is not sufficient to determine promoter response because this same relationship does not apply to a core binding site containing AGATAA. Rather, there must be at least one additional criterion applied to ELT-2-binding sites in order to explain promoter behaviour. One candidate for this additional criterion could be an allosteric change induced in ELT-2 conformation by binding to certain sequences (e.g. TGATAA) but not to other sequences (e.g. AGATAA), as suggested for glucocorticoid receptor binding (Meijsing et al., 2009; Schöne et al., 2016). Any free-energy change necessary to drive this postulated conformational change in ELT-2 would be expected to be incorporated into the overall free energy of binding to this particular sequence, and the TGATAA and AGATAA sites being compared were chosen to have equivalent overall affinities. We thus suggest that any additional criterion is more likely to be applied downstream of the initial ELT-2 binding: e.g. a complex of ELT-2 with a TGATAA site might be able to accommodate binding of a particular co-factor but a complex with an equal-affinity AGATAA site cannot. Overall, these considerations reveal the complexities of a sequence logo like that shown in Fig. 1. Within the TGATAA series of core motifs, 'information' reflects evolutionary selection for tightness of binding (Fig. S4). However, the sequence logo also incorporates additional information, such as the disfavouring, or essentially vetoing, of AGATAA sites. We note that binding motifs for several additional *C. elegans* GATA factors, both endodermal (ELT-7 but not END-1/3) and hypodermal (ELT-3, ELT-6 and EGL-18) also appear to be enriched in TGATAA sequences (Shao et al., 2013; Narasimhan et al., 2015); it will be interesting to determine whether these other GATA factors can, like ELT-2, discriminate *in vivo* against AGATAA sequences independently of binding affinity.

It will be important to define a similar quantitative response curve as shown in Fig. 5 for other transcription factors, both in the *C. elegans* intestine and elsewhere. Are all transcription factors present at *in vivo* free concentrations that are 'above the dissociation constant' for interacting with their preferred site? Or do different transcription factors operate at different effective *in vivo* free concentrations that result, in turn, in a different range of *in vivo* occupancy levels? The Fig. 5 response curve also has implications for attempts to interpret mutations in binding motifs in terms of *in vivo* consequences, either in the context of evolution or genetic disease. For example, genome-wide association studies regularly identify alterations in candidate transcription factor binding sites associated with human disease (Deplancke et al., 2016; Vockley et al., 2017). Even if, as is customarily assumed, the consequences of such alterations are due to changes in transcription factor binding affinity, and even if, as is often the case, changes in transcription factor binding affinity can be predicted from available position weight matrices, the practical implications for the individual or for the evolving organism are not clear. Whether there are effective changes in the degree of transcription factor occupancy of this mutated site *in vivo*, with concomitant changes in target gene transcription, will depend upon the free effective concentration (activity) of the particular transcription factor within the affected cells.

MATERIALS AND METHODS

Competitive EMSAs

Each XXTGATAAXX variant was embedded within the same 26 bp sequence that contained the distal TGATAA site of the *asp-1* promoter, followed by four unpaired C nucleotides, followed by the reverse complement of the initial 26 bp sequence, thereby enabling the formation

Development (2020) 147, dev190330. doi:10.1242/dev.190330

of a double-stranded hairpin. Oligodeoxynucleotides to be used as probes were 5′-labelled with fluorescein amidite (FAM); probe sequences are presented in the supplementary Materials and Methods. Competitive binding reactions were prepared by mixing full-length ELT-2 protein (purified from baculovirus-infected insect cells) into a solution of 250 µM FAM-labelled hairpin oligonucleotide, variable amounts of unlabelled 'competitor' hairpin oligodeoxynucleotide (0 to 4 mM), 10 ng/µl poly(dI-dC).poly(dI-dC) and 1× binding buffer [25 mM Tris-HCl (pH 7.5), 50 mM KOAc, 20 µM MgOAC, 1 mM DTT, 1% NP40 and 100 ng/µl bovine serum albumin]. The quantity of ELT-2 protein per reaction was adjusted in order to shift ∼15% of the probe in the absence of a competitor. Binding reactions were incubated in the dark at room temperature for 20 min, then 10× loading buffer (0.25% Orange G and 20% Ficoll) was added (2 µl per 20 µl reaction) and reactions were loaded onto a 6% polyacrylamide gel prepared with 0.5× Tris Borate EDTA (TBE). Electrophoresis was performed in the dark at 100 V in 0.5× ice-cold TBE for 1 h or until the dye reached the end of the gel. Gels were imaged using a SYBR Green filter and the images were exported as unscaled 8-bit TIF files. Band intensities (shifted=bound; unshifted=free) were quantitated using ImageJ. The relative affinities of ELT-2 binding to …AGATA… and …TGATA… sequences were measured using the Spec-Seq method, closely following the protocol provided by Stormo et al. (2015). The production and sequencing of the degenerate libraries, as well as procedures used to extract relative affinities, are described more fully in the supplementary Materials and Methods.

Production and growth of transgenic *C. elegans* strains

Site-directed mutagenesis was performed using overlap extension PCR (Ho et al., 1989); two successive rounds were used to mutate the two TGATAA sites in the variant *asp-1* promoters. For reporters R1 and R2, respectively, base pair 795 and base pair 840 of the *asp-1* coding region (with A of the ATG=1) were changed from A to T; both changes introduced a unique KpnI site without changing protein sequence. Full sequences are provided in the supplementary Materials and Methods.

Transgenic worm strains used in SQRIPT experiments were created by standard gonadal injection of strain JM189 [*unc-119(ed4) III; asp-1(tm666) elt-7(tm840) V; elt-4(ca16) X*] to produce extrachromosomal multicopy arrays (Mello et al., 1991). Reporter plasmids, as well as the *unc-119* rescuing plasmid pDP#MM016B (Maduro, 2015), were each injected at a concentration of 50 µg/ml.

Simultaneous quantification of reporter transcripts (SQRIPT)

Transgenic populations were expanded at room temperature on nematode growth media plates (either 150 mm or 35 mm diameter) covered with a lawn of *E. coli* OP50. The high transmission rate of *unc-119*-containing transgenic arrays, combined with the increased health and fecundity of rescued animals, resulted in ∼75% of the harvested animals containing the transgenic array. Worms were washed from unstarved plates and excess bacteria removed either by filtering through Nytex screens or by repeated centrifugations. In a typical experiment, the mass of collected worms, suspended in nuclease-free water, was 100-200 mg. Nuclear run-on transcription (Fig. 4B) was performed as described by Kruesi et al. (2013) but omitting α-^{32}P-CTP from the reaction. Compared with the standard growth procedure just described, worms collected for nuclear run-ons were expanded to ∼threefold greater population sizes and received a final wash with, and were resuspended in, ice-cold nuclear isolation buffer [250 mM sucrose, 10 mM Tris-HCl (pH 7.9), 10 mM $MgCl_2$, 1 mM EGTA, 0.25% NP-40, 1 mM dithiothreitol, 4 U/ml RNAse inhibitor cocktail and protease inhibitors (Roche, used at 1× concentration, as specified by the manufacturer)]. One half of each sample, either worms or nuclei, was used for RNA extraction and the other half was used for DNA extraction. RNA extraction was performed using TRIzol Reagent (Thermo Fisher Scientific, 1556018) following the manufacturer's protocol with minor modifications. DNA extraction was performed by prolonged protease digestion, organic extractions and ethanol precipitation (McGhee et al., 1981).

Reverse transcriptase (RT) PCR was performed using the QuantiTect Reverse Transcription Kit (Qiagen, 205313) following the manufacturer's protocol but using a primer (oBL22, sequence in the supplementary Materials and Methods) specific to the *asp-1*-coding sequence. Duplicate or triplicate RT reactions were performed for each sample and each reaction contained up to 1 µg of RNA. A 6 µl aliquot of each finished RT reaction was added directly to a PCR reaction (final volume, 60 µl), together with primers oBL21 and oBL22 (sequences in the supplementary Materials and Methods) to amplify the segment from R1 and R2 reporters that contains the inserted KpnI site. The same fragment was amplified in parallel from DNA samples using the same primers. PCR products were purified by spin column and digested for 1 h at 37°C with 4 U of KpnI (up to 500 ng DNA per 10 µl digestion reaction). Digestion products were separated by electrophoresis on an Agilent D1000 ScreenTape in an Agilent 2200 TapeStation (performed by the University of Calgary Core DNA Services). Control experiments showed that a 'promoterless' reporter produces only a low level of background transcripts (5-8% of wild-type levels), indicating that introduced plasmids are not massively rearranged upon assembly into the transforming array and that there is minimal 'readthrough' transcription from adjacent plasmids incorporated into the array. This 'no-promoter' background rate was used to correct all subsequent measurements. Further control experiments (Fig. S6) showed that: (1) there is little PCR amplification bias between the two reporter sequences; and (2) heteroduplexes can form during PCR amplification of the two highly similar reporter sequences but their effect can be easily corrected. Figs S7-S12 illustrate the calculations used to define K_{rel} and the product K_{max} × [ELT-2free], as well as the Spec-Seq method used to define ELT-2 binding preferences.

Acknowledgements

The authors gratefully acknowledge the expert contribution of Barbara Goszczynski, who performed the Spec-Seq analysis described in the supplementary Materials and Methods. We also thank Dr Erin Osborne Nishimura (Colorado State University) for many helpful discussions.

Competing interests

The authors declare no competing or financial interests.

Author contributions

Conceptualization: B.R.L., J.D.M.; Methodology: B.R.L., J.D.M.; Software: J.D.M.; Investigation: B.R.L., J.D.M.; Writing - original draft: B.R.L., J.D.M.; Writing - review & editing: B.R.L., J.D.M.; Supervision: J.D.M.; Project administration: J.D.M.; Funding acquisition: J.D.M.

Funding

This work was supported by operating grants from the Canadian Institutes of Health Research and from the Natural Sciences and Engineering Research Council of Canada (67135 and RGPIN/04133-2017, respectively, to J.D.M.). B.R.L. received salary support from the Alberta Children's Hospital Foundation.

Supplementary information

Supplementary information available online at
https://dev.biologists.org/lookup/doi/10.1242/dev.190330.supplemental

Peer review history

The peer review history is available online at
https://dev.biologists.org/lookup/doi/10.1242/dev.190330.reviewer-comments.pdf

References

Bain, D. L., Yang, Q., Connaghan, K. D., Robblee, J. P., Miura, M. T., Degala, G. D., Lambert, J. R. and Maluf, N. K. (2012). Glucocorticoid receptor-DNA interactions: binding energetics are the primary determinant of sequence-specific transcriptional activity. *J. Mol. Biol.* **422**, 18-32. doi:10.1016/j.jmb.2012.06.005

Boulin, T., Etchberger, J. F. and Hobert, O. (2006). Reporter gene fusions. WormBook. 1-23. (ed. The C. elegans Research Community): http://www.wormbook.org . doi:10.1895/wormbook.1.106.1

Carey, M. F., Peterson, C. L. and Smale, S. T. (2009). *Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniques*, 2nd edn. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

Crocker, J., Abe, N., Rinaldi, L., McGregor, A. P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F. et al. (2015). Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell* **160**, 191-203. doi:10.1016/j.cell.2014.11.041

Deplancke, B., Alpern, D. and Gardeux, V. (2016). The genetics of transcription factor DNA binding variation. *Cell* **166**, 538-554. doi:10.1016/j.cell.2016.07.012

Dineen, A., Osborne Nishimura, E., Goszczynski, B., Rothman, J. H. and McGhee, J. D. (2018). Quantitating transcription factor redundancy: The relative roles of the ELT-2 and ELT-7 GATA factors in the C. elegans endoderm. *Dev. Biol.* **435**, 150-161. doi:10.1016/j.ydbio.2017.12.023

Farley, E. K., Olson, K. M., Zhang, W., Brandt, A. J., Rokhsar, D. S. and Levine, M. S. (2015). Suboptimization of developmental enhancers. *Science* **350**, 325-328. doi:10.1126/science.aac6948

Furlong, E. E. M. and Levine, M. (2018). Developmental enhancers and chromosome topology. *Science* **361**, 1341-1345. doi:10.1126/science.aau0320

Goszczynski, B., Captan, V. V., Danielson, A. M., Lancaster, B. R. and McGhee, J. D. (2016). A 44 bp intestine-specific hermaphrodite-specific enhancer from the C. elegans vit-2 vitellogenin gene is directly regulated by ELT-2, MAB-3, FKH-9 and DAF-16 and indirectly regulated by the germline, by daf-2/insulin signaling and by the TGF-beta/Sma/Mab pathway. *Dev. Biol.* **413**, 112-127. doi:10.1016/j.ydbio.2016.02.031

Grossman, S. R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., Tewhey, R., Isakova, A., Deplancke, B., Bernstein, B. E. et al. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc. Natl. Acad. Sci. USA* **114**, E1291-E1300. doi:10.1073/pnas.1621150114

Hawkins, M. G. and McGhee, J. D. (1995). elt-2, a second GATA factor from the nematode Caenorhabditis elegans. *J. Biol. Chem.* **270**, 14666-14671. doi:10.1074/jbc.270.24.14666

Ho, S. N., Hunt, H. D., Horton, R. M., Pullen, J. K. and Pease, L. R. (1989). Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene* **77**, 51-59. doi:10.1016/0378-1119(89)90358-2

Hope, I. A. (1991). 'Promoter trapping' in Caenorhabditis elegans. *Development* **113**, 399-408.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S. R., Tan, G. et al. (2017). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260-D266. doi:10.1093/nar/gkx1126

Kruesi, W. S., Core, L. J., Waters, C. T., Lis, J. T. and Meyer, B. J. (2013). Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. *eLife* **2**, e00808. doi:10.7554/eLife.00808

Lam, F. H., Steger, D. J. and O'Shea, E. K. (2008). Chromatin decouples promoter threshold from dynamic range. *Nature* **453**, 246-250. doi:10.1038/nature06867

Levo, M., Zalckvar, E., Sharon, E., Dantas Machado, A. C., Kalma, Y., Lotam-Pompan, M., Weinberger, A., Yakhini, Z., Rohs, R. and Segal, E. (2015). Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res.* **25**, 1018-1029. doi:10.1101/gr.185033.114

Maduro, M. F. (2015). 20 Years of unc-119 as a transgene marker. *Worm* **4**, e1046031. doi:10.1080/21624054.2015.1046031

Maduro, M. F. (2020). Evolutionary dynamics of the SKN-1 → MED → END-1,3 regulatory gene cascade in *Caenorhabditis* endoderm specification. *G3 (Bethesda)* **10**, 333-356. doi:10.1534/g3.119.400724

Mann, F. G., Van Nostrand, E. L., Friedland, A. E., Liu, X. and Kim, S. K. (2016). Deactivation of the GATA transcription factor ELT-2 is a major driver of normal aging in C. elegans. *PLoS Genet.* **12**, e1005956. doi:10.1371/journal.pgen.1005956

McGhee, J. D., Wood, W. I., Dolan, M., Engel, J. D. and Felsenfeld, G. (1981). A 200 base pair region at the 5′ end of the chicken adult beta-globin gene is accessible to nuclease digestion. *Cell* **27**, 45-55. doi:10.1016/0092-8674(81)90359-7

McGhee, J. D., Sleumer, M. C., Bilenky, M., Wong, K., McKay, S. J., Goszczynski, B., Tian, H., Krich, N. D., Khattra, J., Holt, R. A. et al. (2007). The ELT-2 GATA-factor and the global regulation of transcription in the C. elegans intestine. *Dev. Biol.* **302**, 627-645. doi:10.1016/j.ydbio.2006.10.024

McGhee, J. D., Fukushige, T., Krause, M. W., Minnema, S. E., Goszczynski, B., Gaudet, J., Kohara, Y., Bossinger, O., Zhao, Y., Khattra, J. et al. (2009). ELT-2 is the predominant transcription factor controlling differentiation and function of the C. elegans intestine, from embryo to adult. *Dev. Biol.* **327**, 551-565. doi:10.1016/j.ydbio.2008.11.034

Meijsing, S. H., Pufall, M. A., So, A. Y., Bates, D. L., Chen, L. and Yamamoto, K. R. (2009). DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science* **324**, 407-410. doi:10.1126/science.1164265

Meister, P., Towbin, B. D., Pike, B. L., Ponti, A. and Gasser, S. M. (2010). The spatial dynamics of tissue-specific promoters during C. elegans development. *Genes Dev.* **24**, 766-782. doi:10.1101/gad.559610

Mello, C. C., Kramer, J. M., Stinchcomb, D. and Ambros, V. (1991). Efficient gene transfer in C.elegans: extrachromosomal maintenance and integration of transforming sequences. *EMBO J.* **10**, 3959-3970. doi:10.1002/j.1460-2075.1991.tb04966.x

Morgunova, E. and Taipale, J. (2017). Structural perspective of cooperative transcription factor binding. *Curr. Opin. Struct. Biol.* **47**, 1-8. doi:10.1016/j.sbi.2017.03.006

Narasimhan, K., Lambert. S. A., Yang, A. W. H., Riddell, J., Mnaimneh, S., Zheng, H., Albu, M., Najafabadi, H. S., Reece-Hoyes, J. S., Bass, J. I. F., Walhout, A. J. M., Weirauch, M.T. and Hughes, T. R. (2015). Mapping and analysis of *Caenorhabditis elegans* transcription factor sequence specificities. *eLife*. **4**, e06967. doi:10.7554/eLife.06967

Neves, A., English, K. and Priess, J. R. (2007). Notch-GATA synergy promotes endoderm-specific expression of ref-1 in C. elegans. *Development* **134**, 4459-4468. doi:10.1242/dev.008680

Patient, R. K. and McGhee, J. D. (2002). The GATA family (vertebrates and invertebrates). *Curr. Opin. Genet. Dev.* **12**, 416-422. doi:10.1016/S0959-437X(02)00319-2

Pauli, F., Liu, Y., Kim, Y. A., Chen, P. J. and Kim, S. K. (2006). Chromosomal clustering and GATA transcriptional regulation of intestine-expressed genes in C. elegans. *Development* **133**, 287-295. doi:10.1242/dev.02185

Peter, I. S. and Davidson, E. H. (2015). *Genomic Control Process: Development and Evolution*. Academic Press.

Peter, I. S. and Davidson, E. H. (2016). Implications of developmental gene regulatory networks inside and outside developmental biology. *Curr. Top. Dev. Biol.* **117**, 237-251. doi:10.1016/bs.ctdb.2015.12.014

Rajkumar, A. S., Denervaud, N. and Maerkl, S. J. (2013). Mapping the fine structure of a eukaryotic promoter input-output function. *Nat. Genet.* **45**, 1207-1215. doi:10.1038/ng.2729

Roh, H. C., Dimitrov, I., Deshmukh, K., Zhao, G., Warnhoff, K., Cabrera, D., Tsai, W. and Kornfeld, K. (2015). A modular system of DNA enhancer elements mediates tissue-specific activation of transcription by high dietary zinc in C. elegans. *Nucleic Acids Res.* **43**, 803-816. doi:10.1093/nar/gku1360

Romney, S. J., Newman, B. S., Thacker, C. and Leibold, E. A. (2011). HIF-1 regulates iron homeostasis in Caenorhabditis elegans by activation and inhibition of genes involved in iron uptake and storage. *PLoS Genet.* **7**, e1002394. doi:10.1371/journal.pgen.1002394

Rudnizky, S., Khamis, H., Malik, O., Squires, A. H., Meller, A., Melamed, P. and Kaplan, A. (2018). Single-molecule DNA unzipping reveals asymmetric modulation of a transcription factor by its binding site sequence and context. *Nucleic Acids Res.* **46**, 1513-1524. doi:10.1093/nar/gkx1252

Sadler, J. R., Sasmor, H. and Betz, J. L. (1983). A perfectly symmetric lac operator binds the lac repressor very tightly. *Proc. Natl. Acad. Sci. USA* **80**, 6785-6789. doi:10.1073/pnas.80.22.6785

Sasse, S. K., Zuo, Z., Kadiyala, V., Zhang, L., Pufall, M. A., Jain, M. K., Phang, T. L., Stormo, G. D. and Gerber, A. N. (2015). Response element composition governs correlations between binding site affinity and transcription in glucocorticoid receptor feed-forward loops. *J. Biol. Chem.* **290**, 19756-19769. doi:10.1074/jbc.M115.668558

Schöne, S., Jurk, M., Helabad, M. B., Dror, I., Lebars, I., Kieffer, B., Imhof, P., Rohs, R., Vingron, M., Thomas-Chollier, M. et al. (2016). Sequences flanking the core-binding site modulate glucocorticoid receptor structure and activity. *Nat. Commun.* **7**, 12621. doi:10.1038/ncomms12621

Shao, J., He, K., Wang, H., Ho, W. S., Ren, X., An, X., Wong, M. K., Yan, B., Xie, D., Stamatoyannopoulos, J. and Zhao, Z. (2013). Collaborative regulation of development but independent control of metabolism by two epidermis-specific transcription factors in *Caenorhabditis elegans*. *J. Biol. Chem.* **288**, 33411-33426. doi:10.1074/jbc.M113.487975

Sinclair, J. and Hamza, I. (2010). A novel heme-responsive element mediates transcriptional regulation in Caenorhabditis elegans. *J. Biol. Chem.* **285**, 39536-39543. doi:10.1074/jbc.M110.167619

Sommermann, E. M., Strohmaier, K. R., Maduro, M. F. and Rothman, J. H. (2010). Endoderm development in Caenorhabditis elegans: the synergistic action of ELT-2 and -7 mediates the specification–>differentiation transition. *Dev. Biol.* **347**, 154-166. doi:10.1016/j.ydbio.2010.08.020

Stormo, G. D. (2013). *Introduction to Protein-DNA Interactions: Structure, Thermodynamics and Bioinformatics*. Cold Spring Harbor Laboratory Press.

Stormo, G. D., Zuo, Z. and Chang, Y. K. (2015). Spec-seq: determining protein-DNA-binding specificity by sequencing. *Brief Funct Genomics* **14**, 30-38. doi:10.1093/bfgp/elu043

Stringham, E. G., Dixon, D. K., Jones, D. and Candido, E. P. (1992). Temporal and spatial expression patterns of the small heat shock (hsp16) genes in transgenic Caenorhabditis elegans. *Mol. Biol. Cell* **3**, 221-233. doi:10.1091/mbc.3.2.221

Sulston, J. E., Schierenberg, E., White, J. G. and Thomson, J. N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.* **100**, 64-119. doi:10.1016/0012-1606(83)90201-4

Tcherepanova, I., Bhattacharyya, L., Rubin, C. S. and Freedman, J. H. (2000). Aspartic proteases from the nematode Caenorhabditis elegans. Structural organization and developmental and cell-specific expression of asp-1. *J. Biol. Chem.* **275**, 26359-26369. doi:10.1074/jbc.M000956200

Vockley, C. M., Barrera, A. and Reddy, T. E. (2017). Decoding the role of regulatory element polymorphisms in complex disease. *Curr. Opin. Genet. Dev.* **43**, 38-45. doi:10.1016/j.gde.2016.10.007

Weikum, E. R., Knuesel, M. T., Ortlund, E. A. and Yamamoto, K. R. (2017). Glucocorticoid receptor control of transcription: precision and plasticity via allostery. *Nat. Rev. Mol. Cell Biol.* **18**, 159-174. doi:10.1038/nrm.2016.152

Wiesenfahrt, T., Berg, J. Y., Osborne Nishimura, E., Robinson, A. G., Goszczynski, B., Lieb, J. D. and McGhee, J. D. (2016). The function and

regulation of the GATA factor ELT-2 in the C. elegans endoderm. *Development* **143**, 483-491. doi:10.1242/dev.130914

Workman, J. L. and Abmayr, S. M. (2014). *Fundamentals of Chromatin*. Springer.

Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S. O., Nitta, K. R., Morgunova, E., Taipale, M. et al. (2018). The interaction landscape between transcription factors and the nucleosome. *Nature* **562**, 76-81. doi:10.1038/s41586-018-0549-5

Zuo, Z. and Stormo, G. D. (2014). High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding. *Genetics* **198**, 1329-1343. doi:10.1534/genetics.114.170100

Table S1. Collected examples of functional "GATA" sites in the promoters of *C. elegans* intestinal genes[a]

Click here to Download Table S1

Table S2											
	Position Frequency Matrix from McGhee et al 2009 Figure 2A										
		Position									
		1	2	3	4	5	6	7	8	9	10
Base	A	0.64	0.12	0.03	0	1	0	1	0.96	0.1	0.66
	C	0.07	0.5	0.01	0	0	0	0	0	0.26	0.08
	G	0.15	0.17	0	1	0	0	0	0.03	0.58	0.17
	T	0.15	0.21	0.96	0	0	1	0	0.01	0.06	0.09

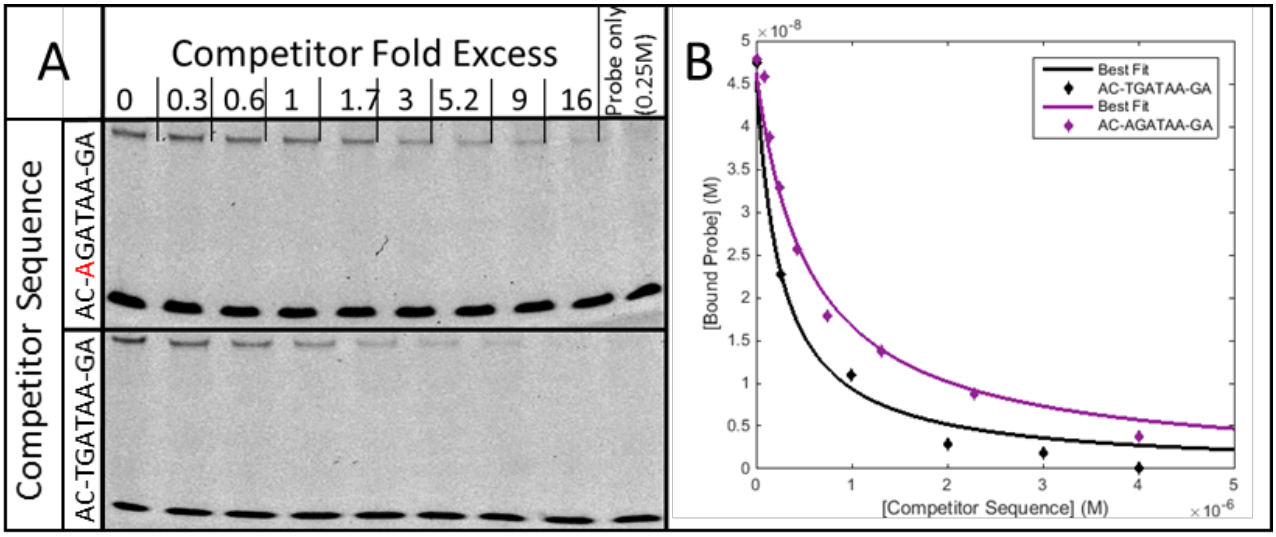


Figure S1. Representative gel images of competitive EMSA assays used to produce estimates of relative binding constants.

A. Free (unshifted) and ELT-2 Bound (shifted) bands are detected by the fluorescence (inverted contrast) of FAM-Labelled double-stranded oligodeoxynucleotide hairpins containing the highest affinity site ACTGATAAGA. Bottom panel shows self-competition of the labelled ACTGATAAGA-containing double stranded oligodeoxynucleotide with an unlabelled double stranded oligodeoxynucleotide containing the same high affinity site. The upper panel shows competition of the highest affinity labelled probe with an unlabelled double-stranded oligodeoxynucleotide containing the same sequence but with AGATAA replacing TGATAA.

B. Analysis of competitive EMSA gels shown in (A) in order to yield estimates of K_{rel}, the relative binding constant to ELT-2. The measured amount of bound (shifted) fluorescent probe is plotted against the total concentration of added competitor. Curves are generated as described in Supplementary Methods and used to yield estimates of K_{rel}.

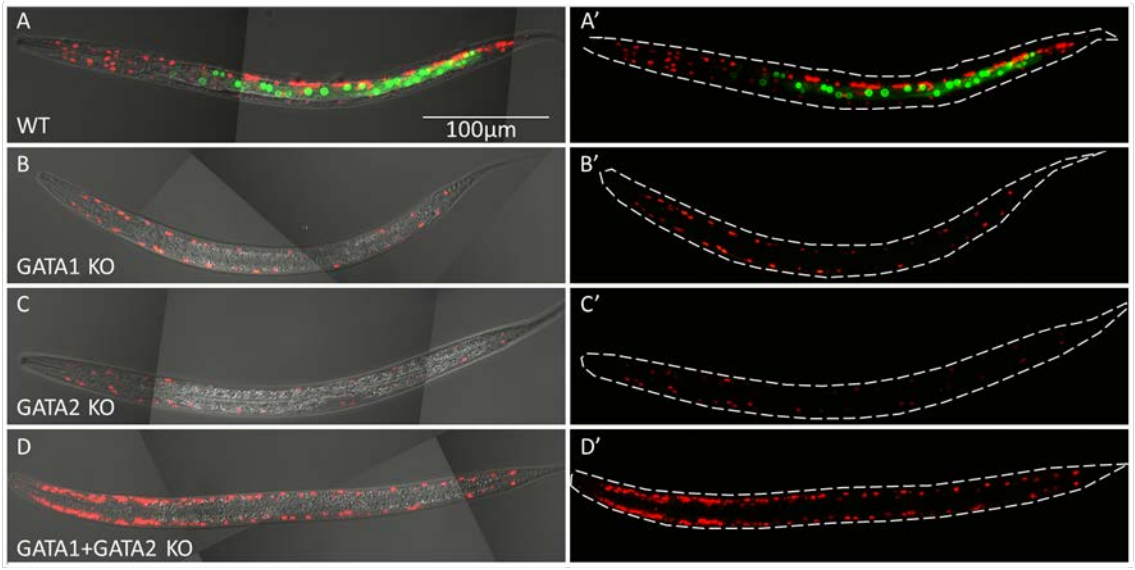


Figure S2: Analysis of WT and TGATAA KO *asp-1* Promoter Activity by Transgenic GFP Reporter Fluorescence

Green fluorescence is observed in the intestinal cells when GFP reporters are placed under transcriptional control of the WT *asp-1* promoter (A and A') but little fluorescence is observed when the *asp-1* promoter TGATAA sites are ablated, either singly (B, B', C, C') or together (D, D'). Red fluorescence reflects body wall muscle expression of a *myo-3promoter::rfp* reporter to indicate successful transgenesis. Left panels (A, B, C, D) = superimposed fluorescence (merged green + red) + differential interference contrast images; Right panels (A', B', C', D') = fluorescence (merged green + red). These images are representative of the majority of worms observed. Images were captured on a Zeiss Axioplan2i microscope with a Hamamatsu Orca digital camera and AxioVision software (version 4.8.1). All images were captured with the same settings and exposure time and displayed with the same brightness and contrast.

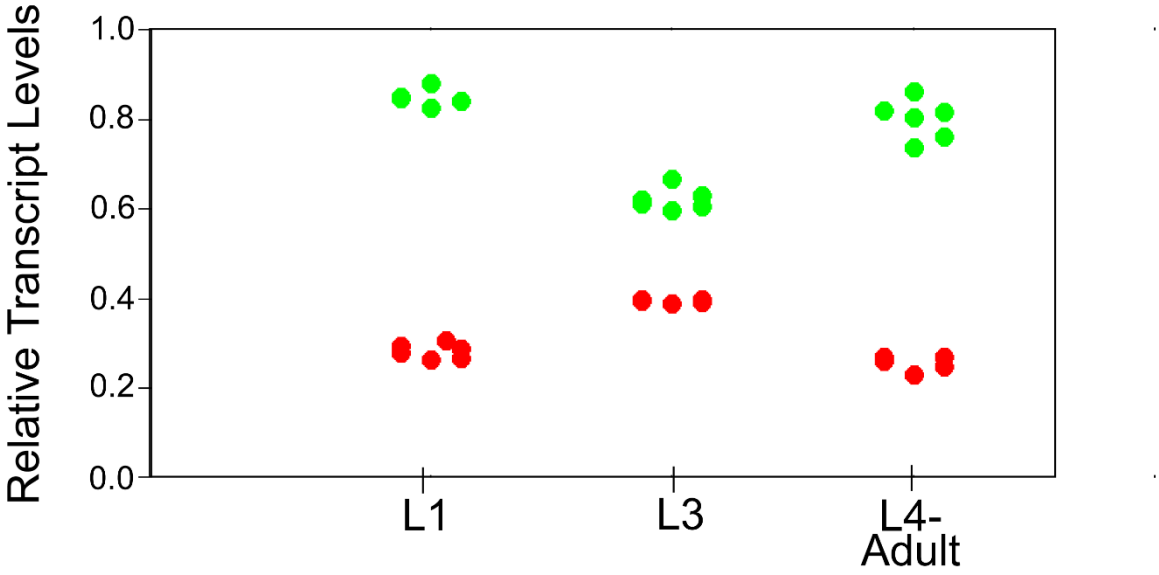


Figure S3. Relative Promoter Activity Measured at Different Developmental Stages Relative transcript levels produced by an *asp-1* variant promoter containing two copies of …AATGATAAGA… (green circles) or two copies of …GCTGATAATG… (red circles) replacing the two wildtype copies of the core TGATAA motifs. The data extends the analysis of Figure 4C of the main text to a weaker promoter and to a slightly stronger promoter.

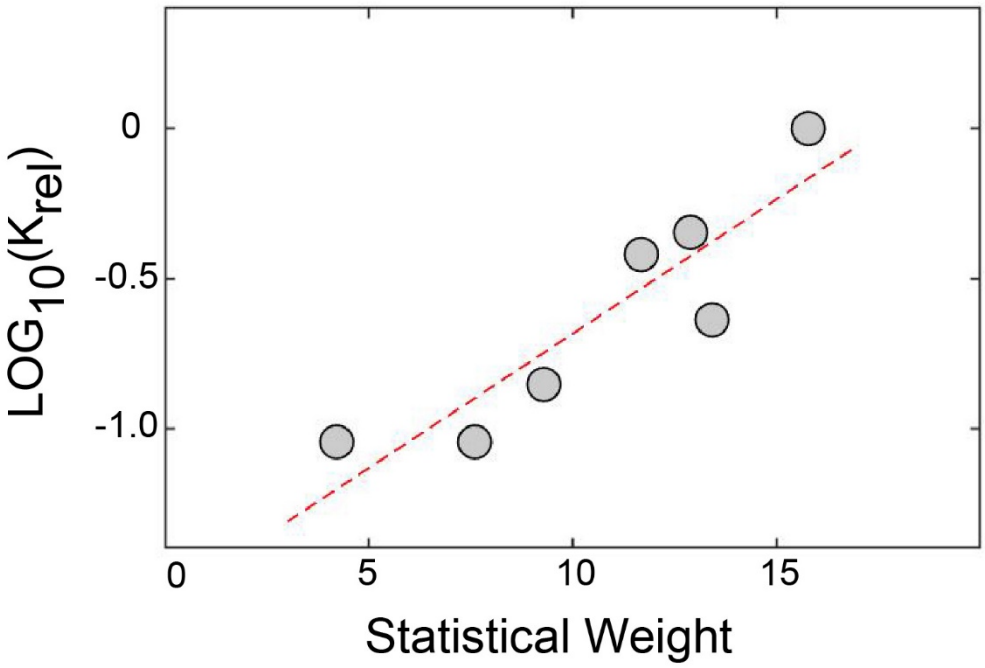


Figure S4

As described in more detail in the text, the position frequency data reproduced in Supplementary Table S2 were used to calculate a Statistical Weight for each of the decameric sequences for which the ELT-2 relative binding affinity had been measured. These statistical weights were then plotted against the logarithm of the relative binding constant. The dashed line represents the least squares regression fit to the data. The fact that this relation is linear is consistent with the conclusion that the frequency with which a *cis*-acting TGATAA regulatory site appears in the promoters of *C. elegans* intestinal genes is determined, in part, by the strength of binding of this sequence to ELT-2.

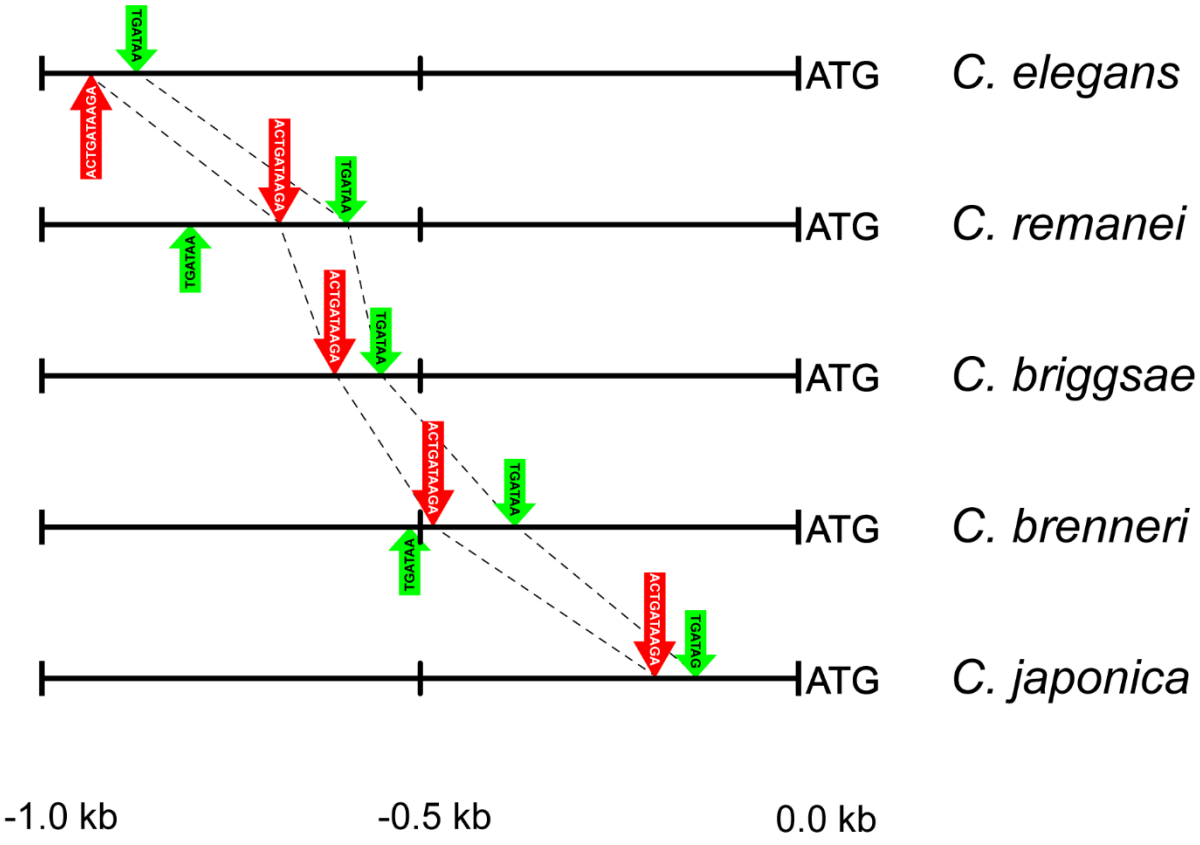


Figure S5. Conservation of Paired TGATAA sites in *Caenorhabditis asp-1* Promoter Homologs.
All TGATAA sequences are shown for the 1 kb regions upstream of *asp-1* homologs in related *Caenorhabditis* species (counting from the ATG initiation codon). The decameric sequence, ACTGATAAGA, with the highest binding affinity to ELT-2 is shown as the red arrow. All other sequences containing TGATAA are shown as the green arrows. In the Discussion section of the main text, we note that each homologous promoter contains two TGATAA sites (connected by the dashed lines) spaced <145 bps apart; one of these sites is the ELT-2 binding sequence with the highest affinity.

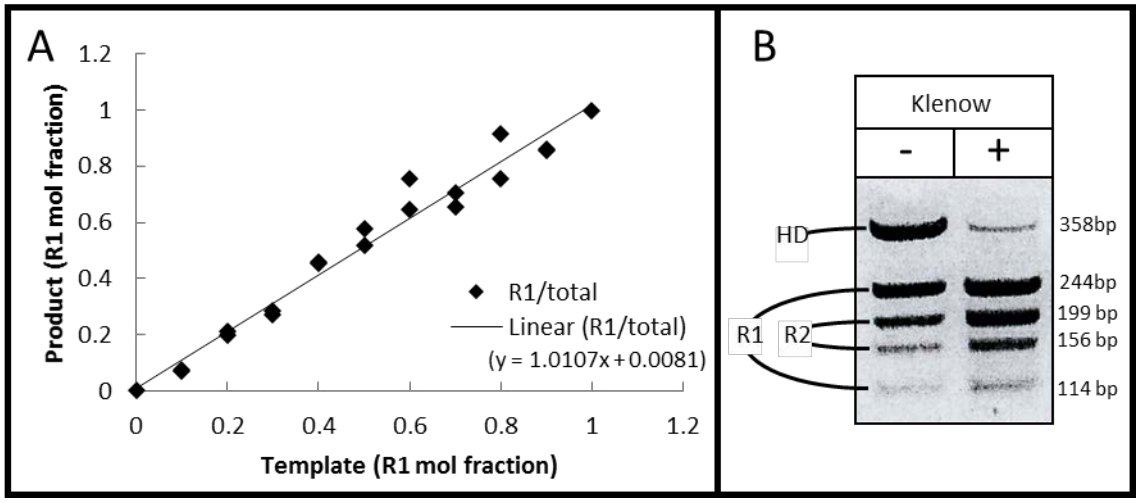


Figure S6: Validation of SQRIPT Procedures

(A) Plasmids containing reporter sequences R1 or R2 were mixed in defined mol fractions as PCR templates. Following amplification, *KpnI* digestion and electrophoresis, pixel intensities of the *KpnI* restriction fragments were quantitated. The slope of the best fit line is close to 1, indicating minimal amplification bias.

(B) In some R1 + R2 amplifications, a fraction of the amplification product was resistant to exhaustive *KpnI* digestion. We deduced that this product was a heteroduplex formed between the two highly similar reporter fragments. First, re-amplification of this resistant product regenerated fragments that were *KpnI* sensitive. Secondly, incorporation of a final extension period using Klenow polymerase with additional added primers removed the majority of the resistant product, as shown in the "+" lane. In practice, the intensity measured for any *KpnI* resistant product was apportioned equally between intensities assigned to reporter R1 and R2.

Supplementary Methods

Analysis of Competitive Band Shifts (See Figure 2 of main text)

Let the labelled oligodeoxynucleotide (present as a double stranded self-complementary hairpin "oligo") be represented as A.

Let the unlabelled competitor oligo be represented as C. C is also a self-complementary oligodeoxyonucleotide hairpin, identical to A except in the residues within 3 base pairs of the core GATA sequence.

Let the protein ligand (in the present case ELT-2) be represented as L.

Let the total concentration (activity) of A in the binding experiment be represented as A_{tot} (M).
Then $A_{tot} = A_b + A_f$ (M) where the subscripts b and f represent bound and free oligo, respectively.
Likewise for the competitor oligo: $C_{tot} = C_b + C_f$ (M).
Similarly for the protein ligand: $L_{tot} = L_f + A_b + C_b$ (M) where ligand can be either free in solution (L_f) or bound in a one-to-one complex with oligo A or oligo C ---(1)

Let the association constant of the ligand L binding to oligo A be K_A (1/M) and to competitor oligo C be K_C (1/M), where $K_A = A_b/(A_f \cdot L_f)$ ---(2) and $K_C = C_b/(C_f \cdot L_f)$ (1/M). ---(3)

And (rearranging the ratio of (3) to (2)): $C_b = A_b \cdot (K_C/K_A) \cdot (C_f/A_f)$ ---(4)

A typical competition experiment begins by mixing oligo A with ligand L under conditions such that most of oligo A remains free, aiming for $A_b \sim 0.15\ A_{tot}$. The precise fraction is measured on sample lane 1 (no added competitor C). Competitor oligo C is then mixed in increasing concentrations maintaining the same L_{tot} and A_{tot} (and with constant sample volumes) and the amount of A_b is measured (on sample lanes 2,3,4,....) for each total concentration of C. The objective is to measure A_b as a function of added C (i.e. C_{tot}) and then to estimate the ratio of the affinity constants, $K_{rel} = K_C/K_A$. In these experiments, A_{tot} and C_{tot} are known (within the accuracy of the assumption that their extinction coefficients are equal). The total protein concentration L_{tot} is constant in each sample but is known only approximately for reasons explained below. In practice, L_{tot} is estimated from a self-competition experiment in which unlabelled A competes with labelled A.

Equations (1) to (4) above are now combined and rearranged in order to obtain A_b as a function of C_{tot} and in a form that will allow $K_{rel} = K_C/K_A$ to be estimated. Other parameters are constants, either measured independently of the experiment (A_{tot}, C_{tot}) or, as just noted, estimated from a separate self-competition experiment (L_{tot}).

$$A_b = K_A \cdot A_f \cdot L_f$$

$$= K_A \cdot (A_{tot} - A_b) \cdot (L_{tot} - A_b - C_b) \quad \text{---(5)}$$

$$= K_A \cdot (A_{tot} - A_b) \cdot (L_{tot} - A_b - A_b \cdot (K_C/K_A) \cdot (C_{tot}/A_{tot})) \quad \text{---(6)}$$

where the C_b term in equation (5) is replaced by C_b from equation (4) above, at the same time replacing C_f/A_f with C_{tot}/A_{tot}. The basis of this approximation is that, under the experimental conditions, most of either oligo will not be bound, i.e. $A_f \lesssim A_{tot}$, $C_f \lesssim C_{tot}$, and the ratio C_{tot}/A_{tot} will be a better approximation than either of the individual quantities.

Equation (6) is expanded and rearranged as a quadratic in A_b, substituting $K_{rel} = K_C/K_A$.

$$A_b^2 \cdot [1 + (K_{rel}) \cdot (C_{tot}/A_{tot})] - A_b \cdot [1/K_A + A_{tot} + L_{tot} + A_{tot} \cdot (K_{rel}) \cdot (C_{tot}/A_{tot})] + A_{tot} \cdot L_{tot} = 0 \quad ---(7)$$

Solving the quadratic (and choosing the negative root in order that $A_b \leq A_{tot}$)

$$A_b = \{(1/K_A + A_{tot} + L_{tot} + K_{rel} \cdot C_{tot}) - [(1/K_A + A_{tot} + L_{tot} + K_{rel} \cdot C_{tot})^2 - 4 \cdot (1 + K_{rel} \cdot (C_{tot}/A_{tot})) \cdot A_{tot} \cdot L_{tot}]^{1/2}\}/\{2 \cdot (1 + (K_{rel}) \cdot (C_{tot}/A_{tot}))\} \quad ---(8)$$

To obtain numerical estimates of K_{rel} for any particular competitor C, the first step is to perform a self-competition of labelled A with unlabelled A in order to obtain estimates of K_A and L_{tot}. An initial estimate of the total ELT-2 concentration present in each binding reaction (obtained from conventional protein assays) is likely to be an overestimate because a fraction of the ELT-2 protein could be inactive in the binding assay; in addition, the effective ELT-2 concentration (i.e. its activity) will be lowered because of binding to components of the reaction other than the specific sequence probes, in particular poly(dI-dC) added to suppress non-specific binding. Using the simple vector commands available in MATLAB, trial values of L_{tot} and K_A are used to calculate A_b (for each of the known values of A_{tot} in the titration experiment) using Equation (8) above; subsequently the sum of the squares of the deviations between these trial values of A_b and the observed set of values for A_b are calculated. The following figure (Supplementary Figure S7) uses the data obtained from a typical self-competition experiment to plot this sum of the squares of the deviations as a function of K_A for a range of values of L_{tot} as shown on each curve. It can be seen that minimizing the sum of the squares of the deviations provides well defined estimates for L_{tot} of ~20 nanomolar and for K_A of ~2x10^7 (1/M).

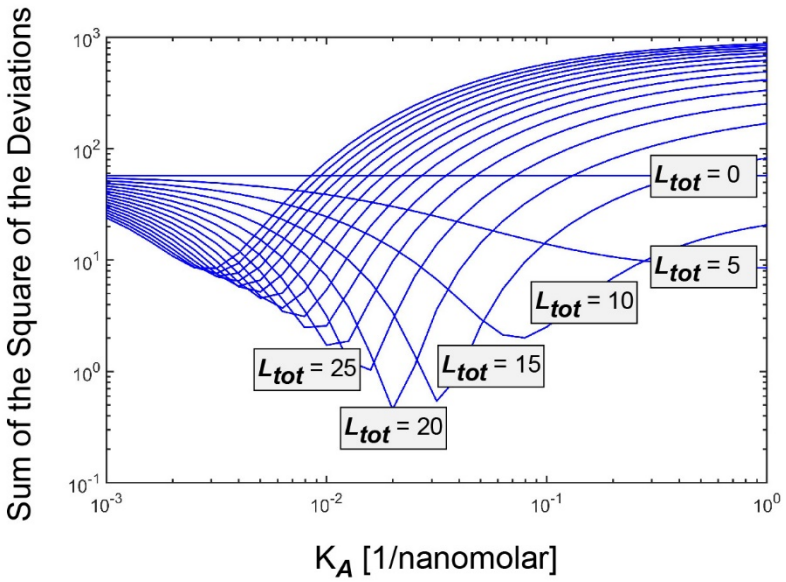


Figure S7

Using the values of L_{tot} and K_A determined by the self-competition experiment, the known values of A_{tot} and C_{tot} and the measured values of A_b, a similar strategy was used to obtain an estimate of $K_{rel} = K_C/K_A$. The sum of the squares of the deviations between calculated A_b (from Equation (8)) and measured A_b is then plotted as a function of trial values of K_{rel} , as shown in Supplementary Figure S8. For this particular set of binding data, the best estimate of K_{rel} is 0.12.

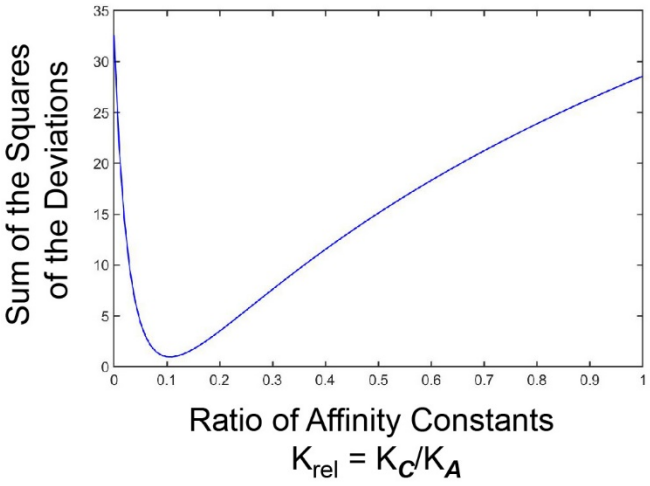


Figure S8

Relative Binding Affinity of ELT-2 to …AGATA… compared to …TGATA… motifs.

We measured the relative binding affinity of ELT-2 to AGATA… vs. TGATA… sequences using the method of "SpecSeq", as developed by Stormo and coworkers [1, 2]. We began with a degenerate "library" of oligodeoxynucleotides, synthesized according to the formula:

`tcctactcttctgtatgtcgNNNNGATANNNNcctaaccgactccgttaatt`

where the lower case sequences bind to appropriate primers, first to render the entire library double-stranded and labelled with fluorescein at the 5'-end of one strand and secondly to amplify ELT-2 bound and unbound fractions by PCR prior to sequencing.

We first determined the quantity of poly(dI-dC). poly(dI-dC) to be added to each EMSA reaction such that non-specific binding of ELT-2 is suppressed but specific binding remains. Increasing amounts of poly(dI-dC). poly(dI-dC) were added to a mixture of ELT-2 protein and a double stranded oligodeoxynucleotide (5'-labelled with FAM) that contains a tightly binding ACTGATAAGA motif (left series of lanes) or a mutated non-binding GATA motif (ACGTCGCCGA; right series of lanes), followed by electrophoresis. Images of typical EMSA gels are shown in Supplementary Figure S9, where it can be seen that 10-20 nanograms of poly (dI-dC). poly(dI-dC) per reaction is able to suppress non-specific binding, while at the same time leaving specific binding apparently unchanged.

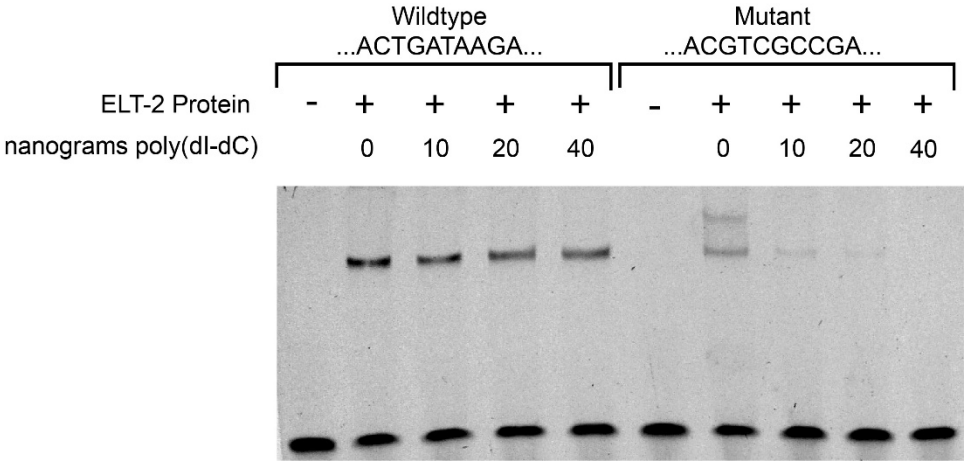
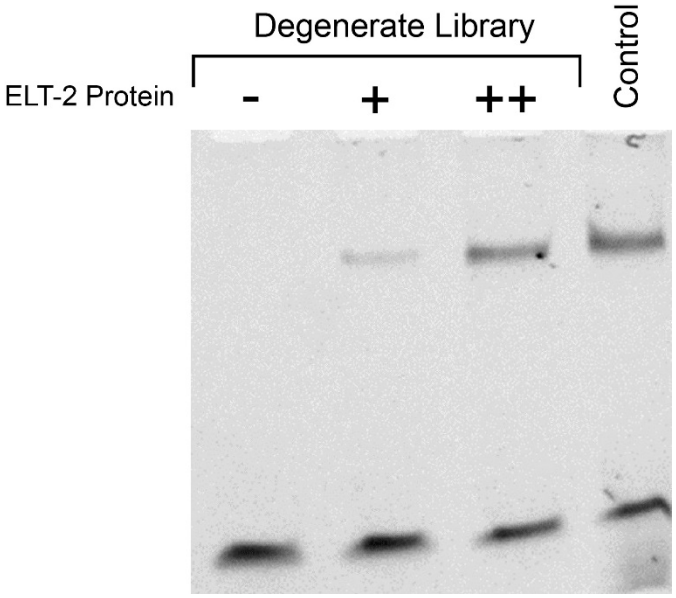


Figure S9

Figure S10 shows a fluorescent image of a typical band shift gel using the degenerate libraries, with no ELT-2 protein (left) and then two different levels of ELT-2 protein (synthesized in baculovirus infected insect cells) mixed in with the double stranded degenerate fluorescently-labelled library. In the lane on the right, ELT-2 is binding to a single-sequence (containing the tight binding motif ACGATAAGA) double stranded oligonucleotide that has the same length and that serves as a positive control. Bound and free sequences were excised from the gel, PCR-amplified, ligated to sequencing primers and processed for next generation sequencing (Illumina MiSeq). For each sample, sequencing returned between 600,000 and 700,000 reads, of which >95% contained an expected core sequence that allowed it to be identified (i.e. ctgtatgtcgNNNNGATANNNNcctaaccgac) and then extracted using a Perl script. Individual 12-mer motifs within the extracted sequences were counted using a MatLab script.

Supplementary Figure S10



Following Zuo and Stormo [2], the relative affinity of AGATA… motifs relative to TGATA… motifs, assuming independent contributions of individual base pairs to the binding affinity, can be assessed by the following ratio of ratios:

(Number of AGATA…Reads in Bound Fraction)/(Number of AGATA…Reads in Unbound Fraction)
(Number of TGATA…Reads in Bound Fraction)/(Number of TGATA…Reads in Unbound Fraction)

The resulting estimate of relative affinity was calculated to be 0.67 and 0.90 for the two different loadings of ELT-2 protein.

The next more complicated model assumes nearest neighbour interactions between the A (or the T) residue and the residue lying immediately upstream (i.e., both A and T have the constant G on their 3'-side). In this case the above formula was re-applied but first to compare the relative affinities of AAGATA… and ATGATA… sequences, then of CAGATA… and CTGATAA… sequences and so on. The resulting estimate of relative affinity, averaged over the four possible nearest neighbour residues, was 0.91 +/- 0.43 and 0.97 +/- 0.27 for the two different loadings of ELT-2 protein.

In principle, next nearest neighbour effects could also be investigated but we judge that the results are sufficient to demonstrate that ELT-2 does not greatly favour binding of TGATA… over AGATA…

Thermodynamic Modelling of the Gene-Response Curve Relating Relative Levels of Reporter Transcripts to Relative Affinity Constants of the Promoter xxTGATAAxx Sites

Let K_{max} (1/M) be the affinity constant for ELT-2 binding to the TGATAA site with the highest naturally occurring affinity (ACTGATAAGA).

Let K (1/M) be the affinity constant for ELT-2 binding to any other individual TGATAA site.

Let L (M) be the concentration (activity) of free unbound ELT-2 in the nucleus *in vivo*.

Thus the (dimensionless) product K·L is an important descriptor of a simple hyperbolic binding curve. That is, if K·L = 1, the site is half-occupied; if K·L <<1, the site is largely free, and; if K·L >> 1, the site is approaching saturation.

Based on the results of Figure 4 (main text), we initially assume that, in our current SQRIPT reporter system, **both** of the two TGATAA sites in each test promoter must be occupied in order for reporter transcription to occur. We further assume that the two sites are occupied independently and that the probability of a site being occupied can be calculated from simple equilibrium considerations.

Thus, the probability that any particular TGATAA site is occupied is given by:

$$\Theta = K{\cdot}L/(1+K{\cdot}L)$$

Since, in our experiments, both TGATAA sites in the same reporter promoter are identical, the transcript levels from this reporter are proportional to Θ^2, i.e. transcript levels from an individual promoter $= \propto \cdot \Theta^2$ where $\propto$ is a constant (ignoring units because $\propto$ will cancel out).

Thus, the transcript levels produced from any reporter if both TGATAA sites were to be completely saturated $= \propto$ (i.e. $\Theta = 1$ as would occur at "infinite" *in vivo* ELT-2 levels).

And now express all transcript levels as y = Relative Transcript Levels, i.e. the transcript levels produced by an individual reporter relative to the transcript levels produced by the reporter driven by the wildtype promoter and incorporated into the same transforming array. y is the parameter measured in our *in vivo* SQRIPT experiments.

Therefore, the maximum possible relative transcript levels (produced when both TGATAA sites are fully saturated) is given by

$$y_{\Theta=1} = \propto/(\text{ transcript levels produced by wildtype control promoter})$$

and y = the Relative Transcript Level produced by any other reporter is given by

$$y = \propto \cdot \Theta^2/(\text{ transcript levels produced by wildtype control promoter})$$

Thus: $y = y_{\Theta=1} \cdot \Theta^2$

And now replace $K = K_{rel} \cdot K_{max}$ in the expression for Θ;

$$y = y_{\Theta=1} \cdot \Theta^2$$

$$= y_{\Theta=1} \cdot \{K{\cdot}L/(1+K{\cdot}L)\}^2$$

$$= y_{\Theta=1} \cdot \{K_{rel} \cdot K_{max} \cdot L/(1+K_{rel} \cdot K_{max} \cdot L)\}^2$$

This rearrangement allows us to plot y (measured in the *in vivo* SQRIPT experiments) as a function of K_{rel}, which was measured by the competitive band shift experiments. We treat $y_{\Theta=1}$ and the product $K_{max} \cdot L$ as adjustable parameters that can be determined by fitting the data. Optimal values for the two parameters can easily be determined to appropriate precision by numerical trials using the vector commands provided by Matlab. The figure below (Supplementary Figure S11) shows the sums of the squares of the differences between the observed y and a trial value of y calculated assuming particular numerical values of $y_{\Theta=1}$ and $K_{max} \cdot L$. As can be seen, the values

that minimize the sums of the squares of the deviations are: $y_{\Theta=1}$ is roughly 1.4 +/- 0.3; $K_{max}\cdot L$ lies in the range from 5 to 10. Calculated Relative Transcript Levels (y) as a function of K_{rel} (ranging from 0 to 1) are superimposed on the actual SQRIPT data in Figure 5 of the main text, using $y_{\Theta=1} = 1.3$ and $K_{max}\cdot L = 10$. The data are fit somewhat less well using $K_{max}\cdot L = 5$ or 15.

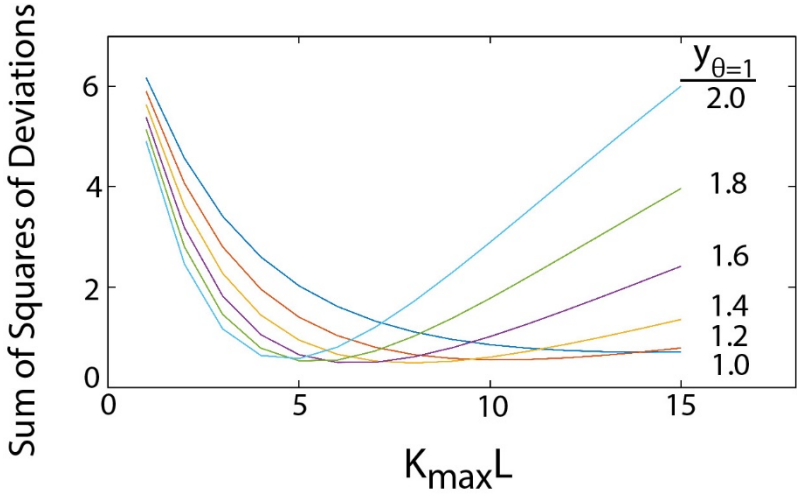


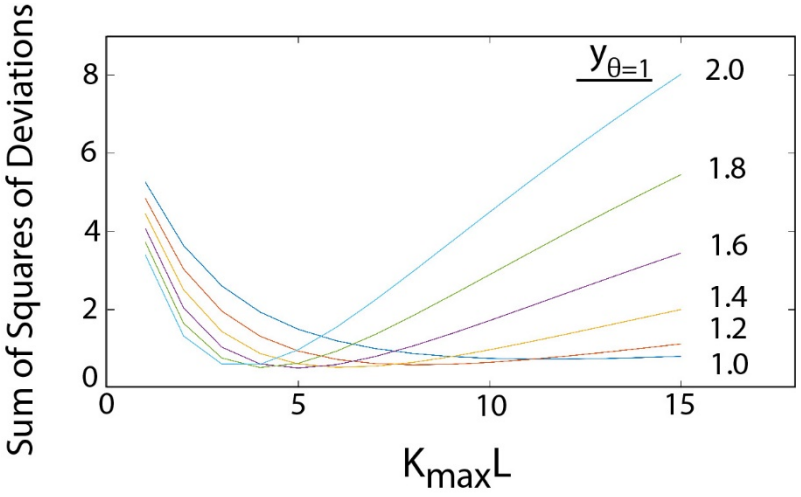
Figure S11

The above model used to predict the y vs. K_{rel} data has assumed that single occupancy of either of the promoter's two TGATAA sites does not activate reporter transcription; only double occupancy is productive. How much do promoter states with single TGATAA occupancy contribute to the observed transcript levels? Consider the strain in which GATA2 has been knocked out (Figure 4 of main text), leaving the reporter promoter under control of the single ACTGATAAGA site, i.e. the optimal site. The mean relative transcript levels produced by this reporter is 0.27 +/- 0.13 (SD), compared to the mean relative transcript levels for the double TGATAA knockout of 0.10 +/- 0.06 (SD), i.e. roughly 15% higher. Thus a revised version of the double-occupancy-only model would stipulate that transcript levels produced if a single TGATAA site were completely occupied $\cong 0.15 \cdot \propto$, where $\propto$ represents the transcript levels produced when both TGATAA sites are completely occupied. Thus the transcript levels produced by a reporter would now be calculated as: $\propto \cdot \Theta^2 + 2 \cdot 0.15 \cdot \propto \cdot \Theta \cdot (1 - \Theta)$, i.e. for single occupancy, either of the two TGATAA sites can be occupied but the other site must be unoccupied. The maximum transcript levels would remain the same, since both sites are then fully occupied. Hence, the expression for relative transcript levels (normalized to the relative transcript levels produced when $\Theta = 1$) is now given by :

$$y = y_{\Theta=1} \cdot \{0.7 \cdot \Theta^2 + 0.3 \cdot \Theta\}$$

$$= y_{\Theta=1} \cdot \{0.7 \cdot [K_{rel}\cdot K_{max}\cdot L/(1+K_{rel}\cdot K_{max}\cdot L)]^2 + 0.3 \cdot K_{rel} \cdot K_{max}\cdot L/(1+K_{rel}\cdot K_{max}\cdot L)\}$$

The same numerical trials are then conducted with Matlab, varying $y_{\Theta=1}$ and $K_{max}\cdot L$ and , as shown in the figure (Supplementary Figure S12), plotting the sums of the squares of the differences between the observed and calculated y values. Compared to the previous figure for the double occupancy only model, the goodness of fit (i.e. the minimum achieved value of the sums of the squares of the deviations) remains largely unchanged but the best values of the two parameters change slightly; $y_{\Theta=1}$ is estimated to be slightly higher at 1.6 and $K_{max}\cdot L$ is now estimated to be slightly lower, closer to 5. The dashed lines in Figure 5 of the main text show how well this revised model fits the experimental data.



Supplementary Figure S12

Sequences of Reporters and Collected Primers

Sequence of the *asp-1* gene used in SQRIPT reporters R1 and R2. Genomic sequence begins 1368 bps upstream of the ATG codon (highlighted in green). The two TGATAA motifs are highlighted in yellow. Protein coding sequence is shown in uppercase. The two single base pair mutations that convert the wildtype *asp-1* sequence into reporters R1 and R2, thereby inserting novel *KpnI* sites, are indicated in cyan.

```
    1 gagacatccc gcccccattt taagtgcata ataagtgtat ttagacaaaa
   51 tccccactgg cgctactcca ccaatcattg agaagaaatt cagccttctt
  101 gtatgaaaaa tgctgaaaaa actgcaaaac ttggccaaaa aactctaaat
  151 cagaacgaaa attcaagaaa ccaacgttaa aatctcccac acaataccca
  201 aaattttcaa aaatcttaca ctaaaataat aataataata ccttctgtac
  251 ttttctacag agttcgtcct aagtcatgtg ctaaactggt cacaataaac
  301 tattcttatt ctctaaaaat ctctaaaaat agctcgtact gtactttccc
  351 ccacctaacc acaattagta caatagtaca attacctttc ggctcctttg
  401 ctatactctt tctcttctaa aaccctcttc TTATCAgtgc aaaagcagta
  451 aaaagtgaaa ctaagaagaa atcggaaaac gagaccaaga accTGATAAg
  501 atttctgaaa ccaattgctg catgaggcta attaaacacg aatgacgtaa
  551 aaggaggagg ggggttggag ccggagtttg ggggtactat aaaagatgag
  601 cggagaggtg gagaagcata ttatcttttt ttgttaggtt tctggttttg
  651 cagaactttc tagaaggttt ttcttttagt aataaaaaaa taattactat
  701 gtttgactat gaagttttaa gggtttttat aaaaacttat gtattaatgt
  751 tcatattaaa atgcttttgg gctgtttttt aggcctaacc ctacttttta
  801 cagtgtattt catgtattac accaggaatt agctttcaag atctctaagg
  851 tttaacaact tttaaagttc aaaactaggc tctagctcgg ttttccggtt
  901 ttttaaagcc tagctgaacc tgagactgtc tgcgaccggt tcaagctgac
  951 ctatactgag ttagactatt ggaaactgcc tcaaatcgtg ttaagcttca
 1001 gacccacata aacgggctaa aatctaccca gaatgtttta gagcgtctta
 1051 aagtccctca gacagtcttt aaacgactta tattgcctga ggatagcagc
 1101 ctgaaaaccg ccctagattt ggcatagctt catttcaaac gtttagatcg
 1151 acctggaggg cttaaaagtt ccgtcctata acagcctaga tatcctgaga
 1201 ataggcaaac cctagaccgt actagaatgt ctcagaaggc cttagacccc
 1251 catccaaaag gttcagatct cctccagact gcttcagata gttttaaatt
 1301 gtcctcagac tcgatgtttc ttcttcagaa ccctccaatc cttccctaaa
 1351 ctcttccttc ttccaggtaA TGCAGACCTT CGTTTTGCTC GCCCTTGTGG
 1401 CGGCATGCTC CGCAGCAGTC ATCCAGGTGC CAACGCACAA GACCGAGTCA
 1451 CTCCGTGCCA AGCTCATCAA GGAGGGCAAG TACACCGCCT TCTTGGCTTC
 1501 ACAGCACGCC GCTCGTGCTC AACAGCTCAA CACCGGATTC CAGCCATTCG
 1551 TCGACTACTT CGATGACTTC TACCTCGGAA ACATCACCCT CGGAACTCCA
 1601 CCACAGCCAG CCACCGTCGT TCTTGACACC GGATCATCCA ACCTTTGGGT
 1651 TATCGATGCC GCATGCAAGA CCCAGGCTTG CAACGGATAC CCAGACTCTG
 1701 GATACACCAA GCAGAAGTTC GACACCACCA AGTCGACCAC CTTCGTGAAG
 1751 GAGACCCGCA AGTTCTCGAT CCAATACGGA TCCGGATCCT GCAACGGATA
 1801 CCTCGGAAAG GATGTTCTTA ACTTCGGAGG ACTCACCGTC CAGTCTCAAG
 1851 AGTTCGGAGT TTCCACCCAC CTCGCCGACG TCTTCGGATA CCAACCAGTT
 1901 GACGGAATCC TCGGACTCGG ATGGCCAGCA CTCGCCGTCG ACCAGGTCGT
 1951 CCCACCAATG CAGAACCTCA TCGCCCAAAA GCAATTGGAC GCTCCACTCT
 2001 TCACTGTCTG GCTTGACCGC AACCTCCAGA TCGCCCAAGG AACCCCAGGA
 2051 GGTCTCATCA CCTACGGAGC CATCGACACC GTCAACTGCG CCAAGCAAGT
 2101 CACCTACGTT CCATTGAGCG CCAAGACCTA CTGGCAATTC CCACTCGACG

                 GGTACC = reporter R1
 2151 CGTTCGCAGT CGGAACCTAC TCTGAGACCA AGAAGGATCA AGTCATCTCC

           GGTA CC = reporter R2
 2201 GACACCGGAA CCTCATGGCT CGGAGCACCA AACACCATCG TCTCCGCCAT
 2251 CGTCAAGCAG ACCAAGGCCG TCTTCGACTG GTCCACCGAG CTTTACACCG
 2301 TCGACTGCTC CACCATGAAG ACCCAGCCAG ACCTCATCTT CACCATCGGA
 2351 GGAGCCCAAT TCCCAGTCAA GTCTGTCGAG TACGTCCTTG ACCTTCAACT
 2401 CGGAGGTGGA AAGTGCGCTC TCGCTGTCTT CTCTATGGGA TCCGGAGGAT
 2451 TCGGACCATC ATGGATTCTT GGAGACACCT TCATCCGTCA ATACTGTAAC
 2501 GTCTACGATA TCGGAAACGG CCAAATCGGA TTCGCCACCG CCGTCCACAA
 2551 GGGATTGTAA gaatggtgtt tttcctgtat ggttatgtat tgctttagtg
 2601 tacaatttgg acacaattct ttgcctcaat tctttgtctc gaataaaatc
 2651 tttaatttct ga
```

FAM labelled self-complementary oligodeoxynucleotides used as probes in competitive band shift assays. TGATAA sites (and reverse complements) are highlighted in yellow and cyan, respectively.

(FAM-labelled AC-TGATAA-GA probe)
oBL76 FAM-aaccctcttcttatcagtgcaaaagcccccgctttgcactgataagaagagggtt

(FAM-labelled GC-TGATAA-TG probe
oBL77 FAM-aaccctcttcattatcagcgcaaaagcccccgctttgcgctgataatgagagggtt

Primers used in SQRIPT cDNA synthesis and PCR amplification of reporter fragments

oBL21 ggaggtctcatcacctacgg (forward primer for amplifying reporter fragment)

oBL22 ctccgagttgaaggtcaagg (reverse primer for amplifying reporter fragment AND primer for reverse transcription reaction)

References for Supplementary Methods

1. Stormo, G.D., Z. Zuo, and Y.K. Chang, *Spec-seq: determining protein-DNA-binding specificity by sequencing.* Brief Funct Genomics, 2015. **14**(1): p. 30-8.
2. Zuo, Z. and G.D. Stormo, *High-resolution specificity from DNA sequencing highlights alternative modes of Lac repressor binding.* Genetics, 2014. **198**(3): p. 1329-43.