molecular systems biology

# A genomic bias for genotype-environment interactions in *C. elegans*

Vladislav Grishkevich, Shay Ben-Elazar, Tamar Hashimshony, Daniel H. Schott, Craig P. Hunter, and Itai Yanai

*Corresponding author: Itai Yanai, Technion - Israel Institute of Technology and Craig P Hunter - Harvard University*

**Review timeline:**

| | |
|---|---|
| Submission date: | 15 January 2012 |
| Editorial Decision: | 17 February 2012 |
| Revision received: | 19 March 2012 |
| Editorial Decision: | 17 April 2012 |
| Revision received: | 26 May 2012 |
| Editorial Decision: | 30 April 2012 |
| Revision received: | 02 May 2012 |
| Accepted: | 02 May 2012 |

## Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 17 February 2012

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the three referees who agreed to evaluate your manuscript. As you will see from the reports below, the referees find the topic of your study of potential interest. They raise, however, substantial concerns on your work, which, I am afraid to say, preclude its publication in its present form.

The reviewers agreed that this work presents a potentially interesting gene expression dataset, but they had important concerns regarding the conclusiveness of the main claims derived from the analysis of these data, raising issues regarding both the rigor and depth of the analysis. All three reviewers indicated that they felt that additional evidence was needed to confirm and validate these findings. In particular, the first reviewer indicates that dissection of "at least genotype-environment to a molecular level" would be needed, and, similarly, the second reviewer felt this work remained rather descriptive and required additional confirmation (suggesting comparisons to trans-eQTL data). All reviewers felt that more rigorous statistical analysis would be needed to support these findings, and they make several specific suggestions along these lines (e.g. multivariate analysis).

Given that these reviewers were cautiously positive about the observations made in this work, and since they make a series of specific recommendations regarding additional analyses and sources of evidence that may help support your findings, we would like offer you the opportunity to submit a revised manuscript. The editor would like to emphasize, though, that the reviewers clearly felt that substantial additional evidence was needed to support these findings and to raise this work beyond a

 

descriptive level. Addressing these concerns convincingly will likely require additional experimental work and analyses.

*** PLEASE NOTE *** As part of the EMBO Publications transparent editorial process initiative (see our Editorial at http://www.nature.com/msb/journal/v6/n1/full/msb201072.html), Molecular Systems Biology will publish online a Review Process File to accompany accepted manuscripts. When preparing your letter of response, please be aware that in the event of acceptance, your cover letter/point-by-point document will be included as part of this File, which will be available to the scientific community. More information about this initiative is available in our Instructions to Authors. If you have any questions about this initiative, please contact the editorial office (msb@embo.org).

Sincerely,
Editor - Molecular Systems Biology
msb@embo.org

---

Referee reports:

Reviewer #1 (Remarks to the Author):

This is an interesting study that describes the genomic properties of genes whose expression shows genotype-environment interactions, using the maternal mRNA load in C. elegans natural isolate strains as a model system. The main take home message is that gene expression wit strong genotype-environment interactions is associated with larger regulatory regions, more upstream motifs and low expression levels, but is not enriched for putative cis regulatory mutations comparing between strains. Thus genotype-environment interactions are predicted to be biased towards 'trans' rather than 'cis' changes. Overall the manuscript and figures are extremely clear.

Suggestions:

The properties of genes with strong expression plasticity (responsiveness) across different environmental conditions have been quite extensively described in yeast. For example it has been shown that genes with high plasticity (and high inter-individual variation, 'noise') tend to have a TATA box promoter and also a nucleosome occluded upstream region. This work in yeast is very relevant to the current study and should be highlighted in the introduction. Another approach in yeast has been the dissection of genotype-environment interactions in recombinant inbred lines, for example the work of the Krugylak and Cohen laboratories, in one example dissecting genotype-environment interactions to bp resolution.

One addition that would certainly strengthen this manuscript would be the dissection of at least one genotype-environment interaction to a molecular level. This is of course challenging given the limited number of genotypes. I feel that, at least, more could be understood about the groups of co-regulated genes than is currently presented.

The authors only examine a few gene properties of the many that are available for C. elegans, and I think they should test more. Moreover they should try and tease out the possible co-correlations / independence of the various features. For example:

The authors test intergenic size, but what about other size properties such as gene length, intron size, first intron size etc?

What about the particular types of predicted or experimentally determined transcription factor binding sites?

What about the expression level of a gene? And its expression breadth across tissues? Can these

 

account for the relationship with motif content or intergenic distance (and vice versa)?

What about chromatin properties?

Another question is whether the authors detect a general correlation in the plasticity of a gene's expression (across conditions) and its change across genotypes, as has been reported in yeast?

Where have the 4 genome sequences been deposited, and with which accession numbers? Without these sequences being publicly available it would be impossible to repeat or extend these analyses.

Further, I consider it essential that the authors provide the results of their raw analyses as text tables in the supplement so that others can repeat and extend this work. i.e. a table with each gene, its orthologs and its various expression and genomic/motif properties across the species should be provided. This would be a very useful resource for other researchers and would fit with the aims of MSB to provide data alongside a manuscript.

Suggested changes to the text:

Throughout the text, the authors should indicate effect sizes, not just p-values (as always in genomic studies, p-values can be highly significant simply because of the large numbers of genes being considered).

Further, are the p-values corrected for multiple testing? What is the expected FDR in each analysis?

p5 first paragraph: edit 'indicating' -> 'suggesting'

p4 second to last paragraph: how sensitive are the conclusions to the rather arbitrary definitions of 'long' and 'mid-range' etc?

The use of the term 'housekeeping' gene as one with stable expression across genotypes and environments is unusual. I suggest using a different term.

Reviewer #2 (Remarks to the Author):

Review of "A genomic bias for genotype-environment interactions in C. elegans"

Grishkevich et al.

This paper investigates the genomic properties which underlie genotype-environment interactions in C. elegans. This is an interesting subject of which not much is known. A priori, the authors define that the phenotype in relation to the environment is determined by gene expression. Obviously this is not the case, indeed it was found that the impact of the environment on the phenotype is for a large part not determined by the transcriptome. This seems trivial but the authors take a short-cut (too short) in the abstract and the introduction of their paper (see Feder, M.E., Walser, J.-C. 2005 Journal of Evolutionary Biology 18 (4) , pp. 901-910 on the limitations of gene expression for genotype-environment interactions). Based on SNP analysis they conclude that genes with environmental interactions are highly regulated.

Major comments
The analysis was restricted to a set of predefined 198 genes after rigorous criteria of :

- Within the linear range of the microarray
- A minimum on variation
- Then ANOVA to detect genes with GxE.

I would like to see this analysis with a window of variation. So how does the analysis of gene properties depend on the expression variation of the selected genes. This would illustrate how robust

 

their analysis is across different levels of gene expression variation and whether it is biased to genes with a particular variation level.

Then, the criterium of intergenic length: the authors do not give a reason why this would play a role? At the moment the paper is rather descriptive than analytical.

On page 4, line 6 the authors state that slightly longer intergenic regions suggest an extensive promoter region may be a liability..... but the authors could find this out. They have the sequence so they could test this speculation/new hypothesis.

On page 4, the authors apply a KS (distribution free) test. Applying this test assumes a deviation form normality of the data. But Fig. 2A shows that data and illustrates a normal-distribution. If this is the case, a different test should be used.

The authors would do good to explain briefly why shorter intergenic regions are consistent with a "potential simple" requirement for regulation. This is one of the pillars of the paper, it is one of the foundations of the mss. They refer to their own paper in 2011 but it seems speculative to me as it is written now.

The fact that interaction genes have higher expression levels than environmental and genotypic genes is taken as a "notion" that interaction genes are under distinct regulation. This is a speculative result and can also be interpreted the other way.
The last 5 sentences of the second paragraph on page 4 are not conclusive to me. This is asking the same question but starting at the other side. If you use the same equation but start at the other end, you will get the same results.

On page 5, 1st paragraph, the authors found that genes with GXE were not different in their SNP content suggesting their expression are caused by trans - effects. The authors should make a stronger case if they could validate this by interrogating trans-eQTL from studies in C. elegans in different environments. For instance data from Rockman et al. 2010; or Li et al. 2006. At the moment the paper is interesting but at this point the authors could make their case much more substantial by validating their findings with the results of eqtl studies.

Minor comments
The mss is not written very clearly. For instance, line 5, ...precise genotype...: what does it mean? Line: 6-10: The definition and the example given of GxE are very difficult to understand and complex. The authors should better use a standard definition given in the literature like in: MacKay, T.F.C., Stone, E.A., Ayroles, J.F. 2009 Nature Reviews Genetics 10 (8) , pp. 565-577. Line 10-14: difficult to understand. Please use plain and clear wording.

The added sentences on medical implications throughout the text are obsolete. These are not well worked out but just added to make the paper seem more relevant to humans. Either expand on it or delete them.

In the second paragraph, line 7-10 the authors state that upstream regulatory factors are the same as trans-effects as defined in eQTL systems. Please expand on this to make more clear to the reader why this is.

Then the authors aim to elucidate the principles (vague) and "to organize these into a general method"....but the authors do not develop a method, they have performed a genomic investigation. The wording is unclear.

On page 3 the authors use "genomic signature"..again new jargon..please rephrase it and be consistent.

Line 9, abstract, However, gene with interactions....: the authors mean environment interactions? This can also be understood as epistatic interactions.

 

Reviewer #3 (Remarks to the Author):

Grishkevich and colleagues have generated a beautiful, well-conceived and -executed dataset of gene expression in multiple C. elegans strains in multiple environmental conditions. The data couldn't be nicer and the authors are to be commended for taking such care and devoting such obvious thought to collecting very high quality data. In addition, the figures are gorgeous.
Their ANOVA identified several hundred genes with expression that depends on genotype, environment, and their interaction. After considering the GxE genes in relation to their genomic characteristics, they conclude that the GxE genes tend to have more complicated cis-regulation than average but no higher level of promoter SNP density, and therefore they are likely affected by genetic variation in trans. This is a reasonable inference from the data.
I have one real substantive concern about the analyses: they are entirely univariate, though the variables considered (and others not considered) are all correlated through genome architecture. In C. elegans, SNP density, gene size, and gene function are all related to chromosomal position; smaller, more conserved, less SNP-dense genes are concentrated in the chromosomal cores/clusters/centers. The authors subject the data to a series of univariate KS tests, but I think a multivariate approach would be more informative, even if it just suggests that many of the variables are collinear and cannot be disentangled.
Finally, I wonder if the authors can elaborate on their impression of the pattern of gene-by-environment interaction. The only example discussed in detail is scrm-4, but looking at Fig. 1A, it seems to me that CB4856 is expressing at maximal levels under all conditions (i.e., expression is constitutively saturated), so the GxE effect is due not to differential regulation across environments but instead to the fact that CB4856 cannot increase expression above its basal (maximal) level. This seems like a different pattern from one in which some strains but not others turn a gene on in response to an environmental perturbation.

Minor points:

In analyses using expression level as a gene-level variable, it's not clear in the text that the value comes only from N2, and it's nowhere indicated why this is the relevant value instead of the mean or median across genotypes & environments.

Why are the animals grown on B. subtilis? That's not the conventional lab environment for C. elegans.

On page 5, it's claimed that the higher SNP density in genotypic genes indicates that these genotypic effects are caused by local changes; I think the data "suggest" such an inference, but they do not "indicate" it.

1st Revision - authors' response 19 March 2012

 

Dear Editor,

Thank you for submitting our manuscript to review. We were pleased that the reviewers appreciated the importance of the central claim of our manuscript; that genes with genotype-environment interactions are enriched for highly regulated genes whose differential expression across strains and conditions is most likely due to *trans* effects. The three reviewers made excellent suggestions. The first reviewer rightfully pointed out that we neglected to cite important work, that an additional experiment would provide additional evidence, that more genomic properties ought to be examined, that availability of the data is a must, and that a sensitivity analysis ought to be included. The second reviewer requested that we cite an important limitation of mRNA level analyses, that we examine correlations between the genomic properties, asked for clarifications as well as more precise definitions in the text, and suggested comparing with eQTL data. The last reviewer also asked for a multivariate analysis, an analysis of the patterns in the set of genes with genotype-environment interactions, and suggested a better control in one of the analyses.

We agreed with the points made by the reviewers and have addressed their concerns in the attached revised manuscript. In particular, we present a new experiment that moves beyond the descriptive realm that the first reviewer addressed. We set out to actually cause genotype-environment interactions using a transgenic *C. elegans* strain and comparing with the previous data. Indeed we found interactions and these had the predicted genomic properties: larger intergenic regions and mid-range expression levels. This experiment provides a powerful new piece of evidence for our hypothesis that highly regulated genes are poised to receive genotype-environment interactions. We also include, among other new analyses, multivariate and sensitivity analyses. The changes led to our inclusion of many additional control analyses resulting in new figures and tables in the supplementary. All of the changes are detailed in the response to the reviewers. We hope that you and the reviewers will find the revision fitting for Molecular Systems Biology and look forward to hearing from you.

With best wishes,

Itai Yanai, for the authors

**Point-by-point response to reviewers**

**Reviewer 1.**

**Concern: Accurate citation of previous work.** "The properties of genes with strong expression plasticity (responsiveness) across different environmental conditions have been quite extensively described in yeast. For example it has been shown that genes with high plasticity (and high inter-individual variation, 'noise') tend to have a TATA box promoter and also a nucleosome occluded upstream region. This work in yeast is very relevant to the current study and should be highlighted in the introduction. Another approach in yeast has been the dissection of genotype-environment interactions in recombinant inbred lines, for example the work of the Krugylak and Cohen laboratories, in one example dissecting genotype-environment interactions to bp resolution."

**Response and Action:** We agree with this point and have added references to the notion the genes with plasticity also have TATA-boxes and a depletion of nucleosones in their proximal promoter: "Furthermore, work in yeast has shown that genes with high expression plasticity tend to have a TATA-box in their promoter (Tirosh et al., 2006) and also a nucleosome occluded upstream region (Tirosh and Barkai, 2008)." We also add the references to the Kruglyak and Cohen papers in the second paragraph of the Introducation.

**Concern: Additional experimental work.** "One addition that would certainly strengthen this manuscript would be the dissection of at least one genotype-environment interaction to a molecular level. This is of course challenging given the limited number of genotypes. I feel that, at least, more could be understood about the groups of co-regulated genes than is currently presented.

**Response:** We agree with the reviewer that additional support would strengthen the central notion of the paper. Our main point is that highly regulated genes are poised to receive gene-environment interactions due to trans effects. Thus, following the reviewer's line of thought, we reasoned that it would be interesting to examine a particular molecular change and detect the gene-environment interactions arising from it. We now report on an experiment in which we query for genotype-environment interaction in a transgenic strain with interrupted function of two genes. The results support the claims of the manuscript.

**Action:** We added the following to the text (last paragraph of Results): "Our results suggest that genes with long promoters and a mid-range level of expression have a disproportionately higher likelihood to develop genotype-environment interactions following *trans* changes. We next asked if a transgenic strain with introduced mutations will produce genotype-environment interactions with this same pattern. Therefore, we compared expression levels across the five conditions on the same microarray platform in triplicate for the N2 strain and a nematode strain deficient for *sid-1* and *haf-6* function in the N2 background (HC445). As expected, *sid-1* and *haf-6* transcripts were significantly reduced ($P<10^{-200}$ and $10^{-70}$, respectively). Querying the data

for genotype-environment interactions we detected 12 genes with significant genotype (N2 vs. *sid-1/haf-6*) -environment interactions ($P$<0.005, two-way ANOVA, Table S4). Consistent with the above results, these 12 interaction genes also showed increased intergenic distances and higher expression on average (Fig. 3). Although the $P$-value for the intergenic genes was greater than 0.1, when examining the 100 genes with the best $P$-values, we found a $P$<0.001. This independent analysis provides strong support for our findings from the geographically distributed strains that interaction genes are highly regulated and that the genotype-environment interaction is due to *trans* effects." We also added figure 3: "Figure 3 Genes with genotype-environment interactions following functional disruption of *sid-1*/*haf-6* also show the hallmarks of highly regulated genes. A. Distributions of intergenic distances, shown as boxplots, comparing the 12 genes with a genotype-environment interaction in the *sid-1/haf-6* analysis (mutant and N2 strain across the five environments, $P$<0.005) with the background set and the 198 interaction genes in the geographical isolates analysis (Figs. 1,2). B. The data for expression levels in the same format." We also added a Supplementary Table with the identities of the 12 genes.

**Concern: "**The authors only examine a few gene properties of the many that are available for C. elegans, and I think they should test more. Moreover they should try and tease out the possible co-correlations / independence of the various features. For example: The authors test intergenic size, but what about other size properties such as gene length, intron size, first intron size etc? What about the particular types of predicted or experimentally determined transcription factor binding sites? What about the expression level of a gene? And its expression breadth across tissues? Can these account for the relationship with motif content or intergenic distance (and vice versa)? What about chromatin properties?"

**Response:** We agree with the reviewer that additional measures must be examined. Many of these we already checked and we now present these as well as the reviewer's suggestions that were novel to us. Interestingly, interaction genes have a higher nucleosome occupancy than the genes of the other sets. The point about co-correlations is addressed below in a response to a similar point made by reviewer 3. We could not examine transcription factor binding sites as these are not well identified in C. elegans at present.

**Action:** We revised the text, adding the following to the results: "The properties of intergenic length and motif concentration are significantly correlated ($P$<10-16, correlation coefficient, Table S2) providing evidence for the notion that longer intergenic lengths indeed reflect increased regulation. These results implicate the interaction genes as a class of highly regulated genes in which the promoter sequence is longer and more motif-dense. Examining other genomic properties, we further found that interaction genes are also enriched in their nucleosome occupancy at the promoter region consistent with our observation of their high expression variability (Fig. S7) (Tirosh and Barkai, 2008)." To the Supplementary Figures we added the Figure along with this Caption: "Figure S7. Genes with genotype-environment interaction are not significantly different from the background in their gene length (A), number of exons (B), length of the first exon (C), combined intron length (D) and protein length (E), but have increased

nucleosome occupancy at their promoters (F) (P<10^-30, KS-test relative to all genes). Nucleosome data was obtained from Valouev, A., et al. (2008). Genome Research 18: 1051-1063."

**Concern: Consistent with yeast data:** "Another question is whether the authors detect a general correlation in the plasticity of a gene's expression (across conditions) and its change across genotypes, as has been reported in yeast?"
**Response and action:** We thank the reviewer for reminding us to state this observation. We have now added the following sentence to the results: "Consistent with previous work in yeast (Tirosh et al., 2006), we found that the set of genes that vary across genotypes and the set of genes that vary across environments significantly overlap ($P$ < 10-200, Hypergeomentric distribution)."

**Concern: Availability of the data:** "Where have the 4 genome sequences ben deposited, and with which accession numbers? Without these sequences being publicly available it would be impossible to repeat or extend these analyses. Further, I consider it essential that the authors provide the results of their raw analyses as text tables in the supplement so that others can repeat and extend this work. i.e. a table with each gene, its orthologs and its various expression and genomic/motif properties across the species should be provided. This would be a very useful resource for other researchers and would fit with the aims of MSB to provide data alongside a manuscript."
**Response and Action:** We agree that data must be freely and conveniently available. We have now added a supplementary table as suggested by the reviewer (Table S5). This Table includes all of the properties used throughout the work: expression data, genomic properties, and SNP data. We further deposited the sequencing data to a central database and we note this in the Methods as follows: "The complete sequencing data has been submitted to the NCBI SRA database with accession ID SRP011413.1 for the study. The accessions for the particular strains are SRS299995.1 (CB4857), SRS299996.1 (RC301), SRS299997.1 (CB4856), and SRS299999.1 (AB2)." We also provide a table describing all of the detected SNPs: "The SNPs in mpileup format are included as Table S6."

**Concern:** "Throughout the text, the authors should indicate effect sizes, not just p-values (as always in genomic studies, p-values can be highly significant simply because of the large numbers of genes being considered).
**Response:** We agree with the reviewer, however we point out that for all of our P-values there are the associated figures from which the effect sizes may be directly examined.

**Concern:** "Further, are the p-values corrected for multiple testing? What is the expected FDR in each analysis?"
**Response:** Yes. The FDR is 0.01. As we noted in "Figure S5. Determination of the ANOVA

significance thresholds. The plots show the cumulative fraction of genes found significant for each *P*-value cutoff. The randomized sets refer to a permutation of all replicate data (75 columns) per gene. For a false discovery of 0.01 the FDR corrected *P*-values were 5.1x10-4, 5x10-4, and 4.7x10-4, for the genotypic (g), environmental (e) and interaction (i) gene sets, respectively."

**Concern:** "p5 first paragraph: edit 'indicating' -> 'suggesting'"
**Action:** Done.

**Concern:** "p4 second to last paragraph: how sensitive are the conclusions to the rather arbitrary definitions of 'long' and 'mid-range' etc?"
**Response:** This is a fair point. We now include a sensitivity analysis in the supplementary materials.
**Action:** To the text we added the following text: "These trends are supported by the complete pattern of enrichments for interactions along the dimensions of intergenic distance and expression level as shown in Figure S8.". To the Supplementary we have added Figure S8 with the following caption: "Figure S8. Genes with long intergenic distances and mid-range expression levels are enriched for genotype-environment interactions. We divided genes into five populated bins with borders 0, 472, 928, 1,818, 4,214, and the maximum distance 57kb. We also binned genes according to their expression with steps of 0.5 log10 expression level units. For each set of genes with a particular combination of intergenic distance bin and expression level bin we compute the enrichment with genes with genotype-environment interactions. This is indicated in the graph according to the -log10 of the *P*-value of the enrichment as computed using the cumulative hypergeometric distribution."

**Concern:** The use of the term 'housekeeping' gene as one with stable expression across genotypes and environments is unusual. I suggest using a different term.
**Response and action:** We agree and now used the term "constitutively expressed" instead of "housekeeping" throughout.

**Reviewer #2:**

**Concern**: "A priori, the authors define that the phenotype in relation to the environment is determined by gene expression. Obviously this is not the case, indeed it was found that the impact of the environment on the phenotype is for a large part not determined by the transcriptome. This seems trivial but the authors take a short-cut (too short) in the abstract and the introduction of their paper (see Feder, M.E., Walser, J.-C. 2005 Journal of Evolutionary Biology 18 (4) , pp. 901-910 on the limitations of gene expression for genotype-environment interactions).
**Response:** We thank for the reviewer for this very useful comment. We agree that this point must be made in the manuscript.

**Action:** We have added the following to the discussion section. “We expect that future work will be directed towards generalizing the approach to developmental time-points, cell types, and conditions. These can be expected to provide an understanding of how genotype-environment interactions arise in the transcriptome; a readily assessable and quantifiable phenotype of the genetic material. However, gene expression in the fullest sense must include protein activity and contributions to fitness (Feder and Walser, 2005) and these provide a challenging goal for the greater understanding of the influence of the genotype and the environment on the organism.”

**Concern:** The analysis was restricted to a set of predefined 198 genes after rigorous criteria of : - Within the linear range of the microarray - A minimum on variation - Then ANOVA to detect genes with GxE. … I would like to see this analysis with a window of variation. So how does the analysis of gene properties depend on the expression variation of the selected genes. This would illustrate how robust their analysis is across different levels of gene expression variation and whether it is biased to genes with a particular variation level.
**Response and Action:** We now include a sensitivity analysis indicating that the results are robust to changes in the threshold of our minimum variation filter. In the text we revised the sentence: “We filtered the dataset to score only those genes with a range of expression within the linear dynamic range of the microarray (2 to 5 log10 units, see Fig. S1) and a minimum level of variation (0.5 log10 units, see Table S1 for robustness to this parameter)” and in the Supplemental we added Table S1: “Table S1. Sensitivity analysis to threshold of expression difference used to define filtered gene set. We selected 0.5 as the threshold for required expression changes but the results hold for all other thresholds.”

**Concern:** “the criterium of intergenic length: the authors do not give a reason why this would play a role? At the moment the paper is rather descriptive than analytical. On page 4, line 6 the authors state that slightly longer intergenic regions suggest an extensive promoter region may be a liability..... but the authors could find this out. They have the sequence so they could test this speculation/new hypothesis. The authors would do good to explain briefly why shorter intergenic regions are consistent with a "potential simple" requirement for regulation. This is one of the pillars of the paper, it is one of the foundations of the mss. They refer to their own paper in 2011 but it seems speculative to me as it is written now.”
**Response and action:** To address this point now provide an additional analysis, also requested by Reviewer 3 that shows that genes with longer promoters also have more motifs. This is indicated by the following text: “The properties of intergenic length and motif concentration are significantly correlated (Table S2) providing evidence for the notion that longer intergenic lengths indeed reflect increased regulation.” We further provide support for the notion that intergenic distance is related to the depth of regulation by citing additional literature: “We first asked whether intergenic lengths might vary across sets of genes with particular expression patterns, since the intergenic distance upstream of a gene’s coding region is a proxy for the length of the promoter (Davidson, 2006).”

**Concern:** On page 4, the authors apply a KS (distribution free) test. Applying this test assumes a deviation form normality of the data. But Fig. 2A shows that data and illustrates a normal-distribution. If this is the case, a different test should be used.
**Response:** We maintain that the KS test is the correct method in this case. The distributions do not all follow a normal distribution and thus statistically cannot be compared by parametric methods. In general the analyses in the manuscript make use of the distribution-free KS test for this reason.

**Concern:** The fact that interaction genes have higher expression levels than environmental and genotypic genes is taken as a "notion" that interaction genes are under distinct regulation. This is a speculative result and can also be interpreted the other way.
**Response:** The test of expression levels simply reveals a difference among the examined gene sets and thus we do not agree that there is any speculation here.
**Action:** To clarify this a bit more we added the phrase "relative to the other gene classes" in the end of that sentence.

**Concern:** "The last 5 sentences of the second paragraph on page 4 are not conclusive to me. This is asking the same question but starting at the other side. If you use the same equation but start at the other end, you will get the same results."
**Response:** Not necessarily. The analysis sought to test whether in general the class of genes with long promoters and intermediate expression are enriched for interactions. Interaction genes may be highly-regulated but in general highly regulated genes may not have interactions. This is why the test is important, as supported by a related comment by Reviewer 1.
**Action:** We clarified this point in text. "Since intergenic distance and basal expression levels may be thought of as proxies for highly regulated genes, we asked whether such a class of genes is enriched for genes with genotype-environment interactions."

**Concern: "**On page 5, 1st paragraph, the authors found that genes with GXE were not different in their SNP content suggesting their expression are caused by trans - effects. The authors should make a stronger case if they could validate this by interrogating trans-eQTL from studies in C. elegans in different environments. For instance data from Rockman et al. 2010; or Li et al. 2006. At the moment the paper is interesting but at this point the authors could make their case much more substantial by validating their findings with the results of eqtl studies."
**Response and Action:** This is a good suggestion. We tested whether our set of interaction genes are known to be influenced by distant regulators using the available eqtl data. Unfortunately, the data did not show enrichment; most likely because of the small sample sizes; the Rockman eqtl study only had 140 genes with trans effects that overlap with our filtered gene set. We hope however that the additional sid-1/haf-6 experiment will satisfy the reviewer's point about further substance to the main claims of the paper. The additional experiment examining the interactions

revealed at the perturbation of two genes showed that here also highly regulated genes are enriched for changes. Furthermore, since we know that the genetic changes are only to *sid-1* and *haf-6* there is good evidence that the observed interactions are *trans* nature.

**Concern: "**For instance, line 5, ...precise genotype...: what does it mean? Line: 6-10: The definition and the example given of GxE are very difficult to understand and complex. The authors should better use a standard definition given in the literature like in: MacKay, T.F.C., Stone, E.A., Ayroles, J.F. 2009 Nature Reviews Genetics 10 (8) , pp. 565-577. Line 10-14: difficult to understand. Please use plain and clear wording."
**Action:** We have now revised our introduction of genotype-environment interactions: "A genotype-by-environment interaction occurs when effect of a locus change in magnitude or direction in different environments (Mackay et al., 2009)."

**Concern:** The added sentences on medical implications throughout the text are obsolete. These are not well worked out but just added to make the paper seem more relevant to humans. Either expand on it or delete them.
**Response:** We agree with the reviewer that the statements may be obsolete in the context of the manuscript.
**Action:** We have now removed the instances mentioning this in the abstract, introduction, and discussion. In the abstract it was replaced with the following: "This observation may provide a deeper understanding into the origin of the extraordinary gene expression diversity present across even closely related species."

**Concern:** In the second paragraph, line 7-10 the authors state that upstream regulatory factors are the same as trans-effects as defined in eQTL systems. Please expand on this to make more clear to the reader why this is.
**Response:** We thank the reviewer for catching this ambiguity.
**Action:** We now rephrased the sentence as: "In particular, evidence has been provided for the notion that much of the observed gene expression variation within a species is due to changes at distant genomic positions (trans changes) (Li et al., 2006; Smith and Kruglyak, 2008; Tirosh et al., 2009; Wittkopp et al., 2004; Wittkopp et al., 2008)."

**Concern:** "Then the authors aim to elucidate the principles (vague) and "to organize these into a general method"....but the authors do not develop a method, they have performed a genomic investigation. The wording is unclear."
**Response and action:** We have now attempted to simplify this sentence: "Here, we describe an investigation into the genomic properties of genes exhibiting genotype-environment interactions."

**Concern:** "On page 3 the authors use "genomic signature"..again new jargon..please rephrase it

and be consistent."
**Response and action:** We now removed the jargon. "To study genotype-environment interactions at a genomic level, mRNA was collected from *C. elegans* embryos extracted from animals of five distinct geographical isolates (genotypes) examined in five conditions and subjected to microarray analysis (see Fig. 1A)."

**Concern:** "Line 9, abstract, However, gene with interactions....: the authors mean environment interactions? This can also be understood as epistatic interactions."
**Response and action:** Yes, thank you. It was changed to: "However, genes with genotype-environment interactions do not significantly differ from background in terms of their promoter SNPs."

**Reviewer #3:**

**Concern:** "I have one real substantive concern about the analyses: they are entirely univariate, though the variables considered (and others not considered) are all correlated through genome architecture. In C. elegans, SNP density, gene size, and gene function are all related to chromosomal position; smaller, more conserved, less SNP-dense genes are concentrated in the chromosomal cores/clusters/centers. The authors subject the data to a series of univariate KS tests, but I think a multivariate approach would be more informative, even if it just suggests that many of the variables are collinear and cannot be disentangled."
**Response and action:** We have now added this analysis to the Results section and refer to the full table of correlations in the Supplementary materials: From the Results section: "The properties of intergenic length and motif concentration are significantly correlated ($P<10^{-16}$, correlation coefficient, Table S2) providing evidence for the notion that longer intergenic lengths indeed reflect increased regulation." and "We defined a set of presumably highly regulated genes as those with long intergenic distance (>5kb) and a mid-range of expression (>2.5 and <3.5 log10 units); these two properties are only weakly correlated (Table S2)." Supplementary Materials: "Table S2. Pairwise correlations among the four examined genomic properties."

**Concern:** "Finally, I wonder if the authors can elaborate on their impression of the pattern of gene-by-environment interaction. The only example discussed in detail is scrm-4, but looking at Fig. 1A, it seems to me that CB4856 is expressing at maximal levels under all conditions (i.e., expression is constitutively saturated), so the GxE effect is due not to differential regulation across environments but instead to the fact that CB4856 cannot increase expression above its basal (maximal) level. This seems like a different pattern from one in which some strains but not others turn a gene on in response to an environmental perturbation.
**Response:** We agree with the reviewer that this description was missing from the original submission. We have opted for providing the full dataset as Table S5 as well as a graphic description of all of the interaction genes, along with a description of our observed overall trends.

**Action:** We included a Supplementary Figure 6 with all of the profiles: "**Figure S6.** Expression profiles for all 198 identified genes with genotype-environment interactions in the same format as Figure 1A. We point out the following classes of patterns: Class 1: changes are unique to one (or a few) combinations of the environment and the genotype. Examples of this are F44E5.4, *hex-2, spe-11,* C38D9.2*, dct-10, hsp-16.11,* F44F4.1, *ifta-1,* and W01C9.4. Class 2: For a particular genotype, variation across conditions not found in the other genotypes. Examples of this are: *cpin-1, plk-3, fkb-5,* F54E4.3, *sago-2,* C11E4.6, *fbxa-192,* B0303.7. Class 3: For a particular environmentally induced expression there is particular variation across the genotypes. Examples are *odc-1, gcy-37,* R01E6.7*,* Y77E11A.2*, asna-2*. Class 4: There is variation across both the genotype and environment dimensions and in the intersection there is a non-additive change. Examples of this are *scrm-4* (as in Fig. 1A), F07B7.2, and K10B3.5."

**Concern:** "In analyses using expression level as a gene-level variable, it's not clear in the text that the value comes only from N2, and it's nowhere indicated why this is the relevant value instead of the mean or median across genotypes & environments.
**Response and Action:** We agree with the reviewer that the more relevant metric is one that summarizes the values for each gene. The plot is nearly identical and in the revised version this now stands as Figure 2C. The relevant sentence in the caption now reads: "Expression levels were defined according to the median across genotypes and environments."

**Concern:** "Why are the animals grown on B. subtilis? That's not the conventional lab environment for C. elegans."
**Response:** E. coli is known to be slightly pathogenic to C. elegans hence we used B. subtilis in all but one condition.
**Action:** In the Methods section we added: "B. subtilis was used here as the standard food source in all but one of the conditions since it is preferred by C. elegans relative to the E. coli OP50 strain (Garsin et al., 2003)"

**Concern:** "On page 5, it's claimed that the higher SNP density in genotypic genes indicates that these genotypic effects are caused by local changes; I think the data "suggest" such an inference, but they do not "indicate" it."
**Response:** Changed to "suggesting" (as also pointed out by Reviewer 1),

2nd Editorial Decision 17 April 2012

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the two referees who agreed to evaluate this revised study. As you will see, the referees felt that the revisions made had improved this work, and they are cautiously supportive. The first reviewer does have some important remaining concerns and makes suggestions for changes, which we would ask you to carefully address in a final revision of the present work.

The most essential outstanding issue appears to be a further request for a deeper multivariate analysis of the genomic features studied in this work. Reviewer #1 did not feel that the correlations presented in Supp. Table 2 adequately addressed the original concern raised by Reviewer #3 (who was not available to review this revision). The current analysis does not test for entanglement with chromosome positions effects, which does appear to be a relevant issue, given the enrichment for essential, conserved genes in chromosome centers, and was clearly central to the concern originally raised by Reviewer #3. The reviewers also seem to feel that a more explicit multivariate statistical framework is needed for this analysis (perhaps MANOVA or a non-parametric alternative).

In addition, the first reviewer has a series of further points that may require discussion or clarifications.

*PLEASE NOTE* As part of the EMBO Publications transparent editorial process initiative (see http://www.nature.com/msb/journal/v6/n1/full/msb201072.html), Molecular Systems Biology now publishes online a Review Process File with each accepted manuscript. Please be aware that in the event of acceptance, your cover letter/point-by-point document will be included as part of this file, which will be available to the scientific community. Authors may opt out of the transparent process at any stage prior to publication (contact us at msb@embo.org). More information about this initiative is available in our Instructions to Authors.

Thank you for submitting this paper to Molecular Systems Biology.

Yours sincerely,

Editor - Molecular Systems Biology
msb@embo.org

---

Referee reports:

Reviewer #1 (Remarks to the Author):

(1) Referee 3 raises what could be an important point about chromosome organisation in C. elegans: "In C. elegans, SNP density, gene size, and gene function are all related to chromosomal position; smaller, more conserved, less SNP-dense genes are concentrated in the chromosomal cores/clusters/centers. The authors subject the data to a series of univariate KS tests, but I think a multivariate approach would be more informative, even if it just suggests that many of the variables are collinear and cannot be disentangled." I think the authors should perform the requested analysis.

(2) The authors now cite some of the relevant work form yeast, but they don't really discuss how the data from C. elegans relate to that from yeast (or to that from fly, for example). In yeast there is gene expression data across both strains and species in different environmental conditions (see work from Gasch and Tirosh, amongst others). Are the trends presented here upheld in yeast? In many ways yeast is a better system to test the main hypothesis because gene regulatory elements are much better annotated than in worm. Also, to what extent can the switch from 'stress-induced' to 'constitutively

 

expressed' account for the differences among strains (http://www.ncbi.nlm.nih.gov/pubmed/219309160)?

(3) The authors could speculate more on why (mechanistically) genes with genotype x environment interactions have larger regulatory regions and more intermediate expression levels than genes with just genotype or just environment effects. Are the same features associated with genes that have 'complex' expression patterns across environments or genotypes?

(4) The authors give the GO (and other) enrichments for the g x e interaction genes, but not for the genotype only/environment only/environment + genotype (no interaction) sets. Indeed this last set (genotypic + environmental, but no interaction) should be used as an additional comparison set in figure 2, I think.

Reviewer #2 (Remarks to the Author):

I believe the authors did everything within their reach to improve thwe mss. They have addressed most issues adequately. I still have some doubts about the verification of the trans-effect QTLs but their additional analysis with the mutants are supportive, although not conclusive.

2nd Revision - authors' response 26 May 2012

We thank you for the constructive feedback. We now submit the revised manuscript including all of the changes requested by you and Reviewer 1. It is clear that this process has improved the manuscript since Reviewer 1 caught that we had inadvertently missed a comment by Reviewer 3 to examine chromosomal locations. The lack of a bias for the gene with genotype-environment interaction provides additional support for our conclusions. We also include a multivariate analysis as you have requested. Finally, we also added additional text to the Discussion section as requested by Reviewer 1.

We hope you find this revised manuscript suitable for publication in Molecular Systems Biology.

Point-by-point Response to Reviewers

Reviewer #1:

**Concern 1: "***Referee 3 raises what could be an important point about chromosome organisation in C. elegans: "In C. elegans, SNP density, gene size, and gene function are all related to chromosomal position; smaller, more conserved, less SNP-dense genes are concentrated in the chromosomal cores/clusters/centers. The authors subject the data to a series of univariate KS tests, but I think a multivariate approach would be more informative, even if it just suggests that many of the variables are collinear and cannot be disentangled." I think the authors should perform the requested analysis."*

**Response:** We thank the reviewer for catching this oversight on our part. Indeed we agree with the original comment of Reviewer 3 to check whether our four defined gene classes (genotypic, environmental, interaction, and constitutively expressed) are biased towards particular chromosomal locations. We have now done this analysis and performed a chromosome localization analysis, included in the supplement.
**Action:** To the main text we added the following sentence: "The longer intergenic regions of the interaction genes may be alternatively explained by a bias in their chromosomal location – since C. elegans chromosomal ends are gene poor (Barnes et al., 1995; Consortium, 1998) – however, we did not detect such an enrichment (Table S3). These results implicate the interaction genes as a class of highly regulated genes in which the promoter sequence is longer and includes more motifs." In the Supplementary we added Table S3. We also added chromosomal location as a parameter in Table S2. Finally we added a multivariate analysis requested by the editor: "**Figure S7. Multivariate**

 

**analysis of the four gene sets.** MANOVA was invoked on the matrix with parameters for the intergenic distance, expression levels, number of SNPs and chromosomal location. CISRED motifs were excluded from the analysis as they are only available for a smaller subset of the genes; when included the results are similar. The figure shows a clear separation of the constitutively expressed genes and the others, as well as the interaction genes from the others. The distribution is similar to that found for expression levels and for intergenic regions (Figure 2)."

**Concern 2:** *"The authors now cite some of the relevant work from yeast, but they don't really discuss how the data from C. elegans relate to that from yeast (or to that from fly, for example). In yeast there is gene expression data across both strains and species in different environmental conditions (see work from Gasch and Tirosh, amongst others). Are the trends presented here upheld in yeast? In many ways yeast is a better system to test the main hypothesis because gene regulatory elements are much better annotated than in worm. Also, to what extent can the switch from 'stress-induced' to 'constitutively expressed' account for the differences among strains (http://www.ncbi.nlm.nih.gov/pubmed/219309160)?"*

**Response:** We agree with reviewer that our results should be placed in better context of other work.
**Action:** We add the following paragraph to the discussion section: "In summary, we have provided evidence that genotype-environment interactions are enriched for highly regulated genes whose differential expression across strains and conditions is most likely due to *trans* effects. Recent studies have explored genotype-environment interactions at the level of a handful of genes and the genome level, for single- and multicellular organisms, strains of the same species and distantly related species (Li et al., 2006; Smith and Kruglyak, 2008; Tirosh et al., 2009; Tirosh et al., 2011; Wittkopp et al., 2004), providing insights into the underlying evolutionary role of genotype-environmental interaction. For example, a recent comparison of gene expression across four yeast species and four environmental conditions revealed a high level of divergence in transcriptional responses to different environments across species (Tirosh et al., 2011). Interestingly, many genes specifically induced in a particular environment in some species often showed high and constitutive expression across all conditions in the other species, suggesting that interactions may occur by transitions between alternative expression programs. Other studies addressed the mechanism of genotype-environmental interaction (Smith and Kruglyak, 2008; Tirosh et al., 2009; Wittkopp et al., 2008). Particularly, changes leading to constitutive expression across environments tended to be *cis*, while *trans* changes were typically condition-specific (Tirosh et al., 2009). Moreover, *cis* changes were implicated as the dominant mechanism of expression divergence among species, while trans changes appeared to account for most of the expression diversity within a species (Li et al., 2006; Wittkopp et al., 2008). Most of the detected *trans* effects were the consequence of genetic changes to sensory genes, rather than to transcription factors (Tirosh et al., 2009). Overall, these studies highlight the dominant role of *trans* effects in genotype-environment interactions in a population, agreeing well with our own result that interaction genes tend to be more highly regulated."

**Concern 3:** *"The authors could speculate more on why (mechanistically) genes with genotype x environment interactions have larger regulatory regions and more intermediate expression levels than genes with just genotype or just environment effects. Are the same features associated with genes that have 'complex' expression patterns across environments or genotypes?"*

**Response:** We agree with the reviewer that some discussion of the possible mechanisms that could explain the differences between the gene sets are important.
**Action:** We now add a paragraph to our discussion section. "We suggest that genes with apparently more complex promoters and mid-range levels of expression are more highly regulated. A growing body of research points to the complexity embodied in particular promoters which correlates with promoter length (Davidson, 2006). As a gene is under the regulation of multiple transcription factors, the binding sites are accommodated in a longer promoter. The relationship of highly regulated genes and mid-range expression levels is less clear. However, if a gene is under the regulation of many transcription factors, one might expect, on average, higher expression due to a higher likelihood for expression at any particular instance. In addition, the higher expression may be attributed to higher basal expression levels, since for a highly regulated gene this may be the sum of leaky expression across all of its regulating TFs."

**Concern 4:** *"The authors give the GO (and other) enrichments for the g x e interaction genes, but not for the genotype only/environment only/environment + genotype (no interaction) sets. Indeed*

 

*this last set (genotypic + environmental, but no interaction) should be used as an additional comparison set in figure 2, I think."*

**Response:** The reviewer is right to point out that we should include the same enrichment analysis for the other datasets, which we have now done. For clarity, we chose not to add the genotype and environment but no interaction set as this would render the figures too difficult to read. The results for this set however are as expected indistinguishable from the genotypic and environmental datasets.
**Action:** We added the Table S8: "Table S8. Functional gene sets with enrichment for the gene sets. Same format as Table S4 for the genotypic and environmental groups, as well as the group of genes that are both genotypic and environmental but lacking interaction."

3rd Editorial Decision 30 April 2012

Thank you again for submitting your work to Molecular Systems Biology. We are now satisfied with the changes made to the scientific content of this work, and we feel that this work is, in principle, acceptable for publication. A few minor remaining format and content issues do remain, however, which we would ask you to address to final revision of this work.

The manuscript, at present, is longer than our Report format requires, and the editor feels it could be improved by the use of more concise language, particularly in the Introduction and Discussion sections.

The editor has attached an edited version of the manuscript with some suggestions for changes. In particular, I felt the first paragraph of the discussion tended to overlap content-wise with the introduction. I recognize this text was inserted to address a reviewer's comments, but a shorter more focused paragraph may be clearer. You might also consider merging the results and discussion sections entirely, which may help to further streamline the manuscript.

Thank you for submitting this paper to Molecular Systems Biology.

Sincerely,

Editor - Molecular Systems Biology
msb@embo.org

3rd Revision - authors' response 02 May 2012

We are pleased that you found the revision appropriate for publication in Molecular Systems Biology. We thank you for the edits to the text. In particular, we are fine with the edits made to the first paragraph of the Discussion; however we suggest that it can also be removed without too much loss if the text deviates too much from the character count limit of a Report. Your call.

 