natureresearch

Peer Review File

Manuscript Title: Unique homeobox codes delineate all C. elegans neuron classes

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

This paper presents a massive set of data and a striking conclusion about specification of neurons and the generation of neural diversity.

The authors took advantage of the diverse yet limited and extremely well-characterized set of C. elegans neurons in an attempt to define the transcription factor (TF) 'code' that might specify them.

They focused their attention of the 100+ homeodomain proteins in the C elegans genome, many of which are know to play significant roles in the transcriptional regulation of a large number of neurons, in worms and in many species.

In a huge effort, they tagged these proteins in fosmids or by CRISPR generating fusion proteins that presumably represent the true expression pattern of these genes in neurons. It should be noted that while transgenes might not be the best approach to address precise expression patterns, the specific features of the C. elegans genome and past experience validate this approach. In any case, the large set of genes tested represents a statistically significant survey to compensate for any potential small differences in the expression pattern of transgenes.

The authors therefore tagged all conserved homeoproteins and those conserved in worms and analyzed their expression in all C elegans neurons.

The results are striking:
- Few homeoproteins are ubiquitous and only on class is pan-neuronal (Cut-like)
- No two homeproteins have the exact same expression pattern
- All worm neurons can be defined by a unique combination of homeodomain proteins
- On average four homeodomain proteins are sufficient to define the identity of each neuron
- The various subclasses of homeoproteins do not correspond to any specific 'features' of a neuron (more later about this).
- This code allows them to define sub-identities for neurons that superficially appear identical but clearly differ in their code based on R-L, D-V or A-P position, in the latter case through Hox genes. They describe numerous such cases that are all strong confirmations of their model.
- No rules can be defined as to the relationship between related combinations of homeoproteins and lineage relationship or functional proximity like neurotransmitters
- single-cell transcriptomics (at least that published in C elegans) falls quite short from identifying all expression patterns, while mRNA expression sometimes does not correlate with protein expression.
- Functional studies show that these homeodomains play important roles in the specification of these neurons
- C elegans specific homeoproteins do not appear to be involved in this code, which leaves open their function: where are they expressed? This is not addressed but it is a side question

All of this is striking and of high significance.

However, the paper would greatly benefit from addressing the following points:

- The authors performed these experiments with homeodomain proteins and most readers will conclude that homeoproteins are 'special' for the generation of neural diversity as compared to any other class of TFs. However, the data do not say that (and the authors do not explicit say that either!). It is possible (and I would say highly likely) that one could have performed the very same experiments with any other large class of TFs, or even with a random subset of 100 TFs, and obtained similar data. The fact that the various subclasses of homeoproteins that diverged from each other many hundreds millions years ago do not define specific subroutine of neurons argues that homeoproteins in general are not 'special', The authors should clearly explain this.

- To test this point, the authors should be able to respecify neurons by changing the code. Surprisingly, they do not attempt to do these obvious experiments. The authors should show examples of fate-switching by altering the homeobox-code (which should be an easy experiment in worms). It is likely that the authors have already the data. Considering the current situation and the likely impossibility to perform these experiments for quite some time, I suggest that the authors significantly tone down the rhetoric about homeoproteins among all TFs being particularly important for identity.

- One thing that the authors could easily do in the next few weeks without having access to the lab, and is mentioned in the introduction, is to use existing published data to test whether their model is valid in other species. The Davis paper (referred to as a bioRxiv paper although it is now published in eLife) should contain enough information (using bulk sequencing, which is deeper than scRNAseq) to test whether the 50+ classes of neurons in the fly brain similarly express combinations of homeobox genes. It is a simple bioinformatic analysis that should have almost as much power as the current work in C elegans.

In conclusion, this paper will make people think again about combinatorial control by TFs; it might bias the thinking towards homeoproteins. This might provide a way to repurpose neurons by reprogramming them, maybe with homeoproteins only, or with other combinations of TFs obtained from similar types of analysis, including scRNAseq.

Referee #2 (Remarks to the Author):

This is a hugely ambitious and impressively comprehensive study of every one of the almost 100 homeobox genes of C. elegans and how they conspire to specify neuronal identity. Such a study could ONLY be done in C. elegans where the 302 neurons are so well characterized and the expression patterns of 100 GFP fusion genes can be described with such cellular precision. The authors have found, using human intelligence rather than artificial intelligence, that the co expression of four homeobox genes can rather perfectly predict the neural type, a classification of about 1 in 100 based on four homeobox genes. I am convinced that it works and that there is deep deep biology that explains why it works: their discussion hypothesizes that homeobox genes perhaps specified neurons from the beginning of metazoans so that the current set of about 100 homeobox genes is the highly ramified duplicated and diverged homeoboxes that subserve the evolved complexity of behavior. This is a quite reasonable idea. But there is also good old fashioned genetics that supports the project: when geneticists have screened for a defect in this or that neural type, homeobox transcription factors (plus other transcription factors as well) have emerged from many of the genetic screens. So clearly homeoboxes are not redundant for neural specification and they can vary in our genetic screens and in evolution to generate diverse neural phenotypes. And homeobox genes constitute about 20% of the cell type specific transcription factors in any animal. So it makes sense that combinations of homeobox signatures may be sufficient to explain neuronal types.

The paper is very succinctly written, given how huge the datasets and the statistical analyses are.

I thought that every figure was important and would not change anything. Except the color coding in Figure 4A. This red green color blind male could distinguish blue and a lot of reddish or greenish neuron types that seemed to be of 3 classes but I could not begin to try to read it. You lose the best 4% of your readers with a color coding like that.

I also tried to look in a phylogenetic analysis for whether there are any animals that might have lost particular homeobox genes during evolution and lost a class of neuron types at the same time. The finding that 4 homeobox genes are enough to specify a neural type predicts that organisms that lost particular classes of neuron types (or are more primitive and never duplicated and diverged their homeobox genes) should be missing neural classes that are lost when the homeobox is lost. The web site below looks at PBC-20 to show that some acetylcholine receptors seem to be lost when PBC is lost. It is probably too simplistic an analysis but I suggest that the authors spend an hour or two sniffing around loss of homeobox genes in animal evolution. In a more systematic approach, they could grab the homeobox protein sequences from ALL the homeoboxes from C. elegans, two insects, two fish, two mammals, two cnidarians, two ascomycota and make a tree. See the homeoboxes that have been lost in one animal lineage and see if it correlates with loss of the neural types that depend on that homeobox in the 4 homeobox signatures of neural cell types. In Nematostella for example they may see what homeobox classes it retains and see if there are neuronal signatures that their C elegans analysis would suggest that these cnidarians have lost. Perhaps GABA ergic neurons disappear in that lineage or GABA and cholinergic that share a Hox code.

This suggestion is NOT something that I demand from the authors in any revision. It is just a suggestion for a look see. And it may give some surrogate genetic proof of the homeobox signature of neural classes. And a simple "one week of comparative genomics" solution to the "prove it" demands that another reviewer may, in the usual bombast of reviewers, demand.

I dont want to add my name to the review because I happen to like anonymity in reviews, but of course there are hints in the review of who I am.

http://genetics.mgh.harvard.edu/phylodomain/index.php?user_org=Caenorhabditis_elegans&user_gene=ceh-20__PBC&user_num=50

Author Rebuttals to Initial Comments:

Author rebuttals written in red

Referee #1 (Remarks to the Author):

We thank the reviewer for the comments. Before we respond to the specific comments, we want to alert the reviewer to a few minor additions to the manuscript, which are the result of analysis done while the manuscript was under review:

1) we included the expression patterns of 3 additional CRISPR/gfp tagged, C.elegans-specific, i.e. non-conserved homeodomain proteins (ceh-76, ceh-89, ceh-100). This now brings up the analysis above the “magic” 100. None of those added genes change the homeodomain “codes”. Those three patterns follow the somewhat curious pattern that we have observed for other C.elegans-specific homeodomain proteins (one is ubiquitous, the other not in neurons and the third very broadly throughout the nervous system); perhaps that’s what “novel genes” do.

2) the total number of C.elegans homeobox genes was corrected down from 103 and 102 since one of the genes was recently re-assigned by Wormbase curators as a pseudogene.

3) We played around with the homeobox code a little more and found two things, one of which we regret not having noted before because it is very exciting: In the previous version of the manuscript, we had stated that the subsets of homeobox genes that are C.elegans-specific (i.e. only found in C.elegans, but not in other Caenorhabditis species) do not contribute to the code (i.e. if we don't consider their expression, the remaining homeobox genes still codify all distinct neuron types). We now noted that we can even strip away the homeobox genes that are conserved, but only conserved in nematodes, and still find that the remaining homeobox genes generate enough codes. In other words, the 71 conserved-to-human homeobox genes alone build all the neuron-type specific codes. We find this quite exciting from an evolutionary point of view.

We also took this one step further – and this is of course pure theory now – and ask how few of those 71 expression patterns of the conserved homeodomain proteins are SUFFICIENT to "codify" 118 neuron classes. We find that 24 patterns alone do the job. It's not clear what this means other than demonstrating something that everyone intuitively presumes: These codes are "overbuilt", i.e. you could do with much less.

This paper presents a massive set of data and a striking conclusion about specification of neurons and the generation of neural diversity.
The authors took advantage of the diverse yet limited and extremely well-characterized set of C. elegans neurons in an attempt to define the transcription factor (TF) 'code' that might specify them.
They focused their attention of the 100+ homeodomain proteins in the C elegans genome, many of which are know to play significant roles in the transcriptional regulation of a large number of neurons, in worms and in many species.
In a huge effort, they tagged these proteins in fosmids or by CRISPR generating fusion proteins that presumably represent the true expression pattern of these genes in neurons. It should be noted that while transgenes might not be the best approach to address precise expression patterns, the specific features of the C. elegans genome and past experience validate this approach. In any case, the large set of genes tested represents a statistically significant survey to compensate for any potential small differences in the expression pattern of transgenes.
The authors therefore tagged all conserved homeoproteins and those conserved in worms and analyzed their expression in all C elegans neurons.
The results are striking:
- Few homeoproteins are ubiquitous and only on class is pan-neuronal (Cut-like)
- No two homeproteins have the exact same expression pattern
- All worm neurons can be defined by a unique combination of homeodomain proteins
- On average four homeodomain proteins are sufficient to define the identity of each neuron
- The various subclasses of homeoproteins do not correspond to any specific 'features' of a neuron (more later about this).
- This code allows them to define sub-identities for neurons that superficially appear identical but clearly differ in their code based on R-L, D-V or A-P position, in the latter case through Hox genes. They describe numerous such cases that are all strong confirmations of their model.
- No rules can be defined as to the relationship between related combinations of homeoproteins and lineage relationship or functional proximity like neurotransmitters
- single-cell transcriptomics (at least that published in C elegans) falls quite short from identifying all expression patterns, while mRNA expression sometimes does not correlate with protein expression.
- Functional studies show that these homeodomains play important roles in the specification of

these neurons
- C elegans specific homeoproteins do not appear to be involved in this code, which leaves open their function: where are they expressed? This is not addressed but it is a side question

All of this is striking and of high significance.
However, the paper would greatly benefit from addressing the following points:

- The authors performed these experiments with homeodomain proteins and most readers will conclude that homeoproteins are 'special' for the generation of neural diversity as compared to any other class of TFs. However, the data do not say that (and the authors do not explicit say that either!). It is possible (and I would say highly likely) that one could have performed the very same experiments with any other large class of TFs, or even with a random subset of 100 TFs, and obtained similar data. The fact that the various subclasses of homeoproteins that diverged from each other many hundreds millions years ago do not define specific subroutine of neurons argues that homeoproteins in general are not 'special', The authors should clearly explain this.

Yes, absolutely, fair enough. As the reviewer points out, we do not claim that homeodomain proteins are special; we took pains to avoid this in order to leave room for the possibility that other TF families may also generate neuron type-specific codes. We now include a sentence in the final paragraph of the Discussion that explicitly states that other TF families will need to be analyzed to see whether they display such combinatorial properties as well.

- To test this point, the authors should be able to respecify neurons by changing the code. Surprisingly, they do not attempt to do these obvious experiments. The authors should show examples of fate-switching by altering the homeobox-code (which should be an easy experiment in worms). It is likely that the authors have already the data. Considering the current situation and the likely impossibility to perform these experiments for quite some time, I suggest that the authors significantly tone down the rhetoric about homeoproteins among all TFs being particularly important for identity.

Yes, understood. In the concluding paragraph we now clearly acknowledge that other non-homeodomain TF do play a role in neuron identity specification (and that other families may or may not reveal such combinational codes as well). We do maintain, however, that it is notable that at least in C. elegans, in the vast majority of cases in which a non-homeodomain protein has been found to be involved in neuronal identity specification, it IS collaborating also with at least one homeodomain transcription factor. Also, as reviewer #2 points out (and as we allude to in the introductory paragraphs), it remains notable that the unbiased “classic” screens for neuronal function/identity in C. elegans appear to have predominantly retrieved homeobox genes. Not clear whether this is statistically significant (hence we didn’t delve on this in the manuscript), but it is notable.

In regard to changing the code – we do appreciate that the reviewer does not demand code changing experiments (there are single studies by us and the Chalfie lab in the past that have changed neuronal identity via homeobox misexpression, but that’s anecdotal). But we wish to emphasize that our computational analysis does suggest that homeodomain protein combinations are able to account for the specificity of the molecular signatures of individual neuron types. Yes, it's not as good as experimental evidence, but it's more than just having loss of function studies alone.

- One thing that the authors could easily do in the next few weeks without having access to the lab, and is mentioned in the introduction, is to use existing published data to test whether their model is valid in other species. The Davis paper (referred to as a bioRxiv paper although it is now published in eLife) should contain enough information (using bulk sequencing, which is deeper than scRNAseq) to test whether the 50+ classes of neurons in the fly brain similarly express combinations of homeobox genes. It is a simple bioinformatic analysis that should have almost as much power as the current work in C elegans.

We thank the reviewer for this suggestion – it revealed a really neat point that is now added to the manuscript: We took all fly homeobox genes and queried the Davis et al bulk sequencing datasets (the reviewer is right to point out that it's important that this is bulk, but not scRNA because of depth). This query was easy because Davis et al present the data in a very accessible database. The result is very nice: Each of the 60+ fly neuron types also display a unique homeobox code! This finding is now included in a new Suppl. Figure and mentioned in the text. Moreover, the Sugino et al. bulk sequencing data from different regions of the mouse brain also shows that set of as little as 8 homeobox genes can distinguish between more than 99% of the different cell types, something we now also mention in the text. We emphasize, though, that in both cases we're dealing with very very low sampling of cell types (plus: the Sugino datasets are not exactly homogenous populations of cells) and this theme may or may not hold if one were to look at other neuron types. Here is where the strength of our analysis comes in, i.e. we look at an entire nervous system.

In conclusion, this paper will make people think again about combinatorial control by TFs; it might bias the thinking towards homeoproteins. This might provide a way to repurpose neurons by reprogramming them, maybe with homeoproteins only, or with other combinations of TFs obtained from similar types of analysis, including scRNAseq.

Referee #2 (Remarks to the Author):

We thank the reviewer for these comments. Before we respond to the specific comments, we want to alert the review to a few changes/additions to the manuscript, which are the results of analysis done while the manuscript was under review:

1) we included the expression patterns of 3 additional CRISPR/gfp tagged homeobox genes (C.elegans-specific genes), which now brings up the analysis above the "magic" 100.

2) the total number of C.elegans homeobox genes was corrected down from 103 and 102 since one of the genes was recently re-assigned as a pseudogene.

We also alert this reviewer to a nice data addition prompted by Reviewer #1: We examined bulk sequencing datasets from ~60 neuron classes in the fly and found that each neuron class displayed a unique set of homeobox combinations.

This is a hugely ambitious and impressively comprehensive study of every one of the almost 100 homeobox genes of C. elegans and how they conspire to specify neuronal identity. Such a study could ONLY be done in C. elegans where the 302 neurons are so well characterized and the expression patterns of 100 GFP fusion genes can be described with such cellular precision. The authors have found, using human intelligence rather than artificial intelligence, that the co expression of four homeobox genes can rather perfectly predict the neural type, a classification of about 1 in 100 based on four homeobox genes. I am convinced that it works and that there is deep deep biology that explains why it works: their discussion hypothesizes that homeobox genes perhaps specified neurons from the beginning of metazoans so that the current set of about 100 homeobox genes is the highly ramified duplicated and diverged homeoboxes that subserve the evolved complexity of behavior. This is a quite reasonable idea. But there is also good old fashioned genetics that supports the project: when geneticists have screened for a defect in this or that neural type, homeobox transcription factors (plus other transcription factors as well) have emerged from many of the genetic screens. So clearly homeoboxes are not redundant for neural specification and they can vary in our genetic screens and in evolution to generate diverse neural phenotypes. And homeobox genes constitute about 20% of the cell type specific transcription factors in any animal. So it makes sense that combinations of homeobox signatures may be sufficient to explain neuronal types.

The paper is very succinctly written, given how huge the datasets and the statistical analyses are. I thought that every figure was important and would not change anything. Except the color coding in Figure 4A. This red green color blind male could distinguish blue and a lot of reddish or greenish neuron types that seemed to be of 3 classes but I could not begin to try to read it. You lose the best 4% of your readers with a color coding like that.

Color change done and verified by color-blind lab member.

I also tried to look in a phylogenetic analysis for whether there are any animals that might have lost particular homeobox genes during evolution and lost a class of neuron types at the same time. The finding that 4 homeobox genes are enough to specify a neural type predicts that organisms that lost particular classes of neuron types (or are more primitive and never duplicated and diverged their homeobox genes) should be missing neural classes that are lost when the homeobox is lost. The web site below looks at PBC-20 to show that some acetylcholine receptors seem to be lost when PBC is lost. It is probably too simplistic an analysis but I suggest that the authors spend an hour or two sniffing around loss of homeobox genes in animal evolution. In a more systematic approach, they could grab the homeobox protein sequences from ALL the homeoboxes from C. elegans, two insects, two fish, two mammals, two cnidarians, two ascomycota and make a tree. See the homeoboxes that have
been lost in one animal lineage and see if it correlates with loss of the neural types that depend on that homeobox in the 4 homeobox signatures of neural cell types. In Nematostella for example they may see what homeobox classes it retains and see if there are neuronal signatures that their C elegans analysis would suggest that these cnidarians have lost. Perhaps GABA ergic neurons disappear in that lineage or GABA and cholinergic that share a Hox code.

This suggestion is NOT something that I demand from the authors in any revision. It is just a suggestion for a look see. And it may give some surrogate genetic proof of the homeobox signature of neural classes. And a simple "one week of comparative genomics" solution to the "prove it" demands that another reviewer may, in the usual bombast of reviewers, demand.

I dont want to add my name to the review because I happen to like anonymity in reviews, but of course there are hints in the review of who I am.

http://genetics.mgh.harvard.edu/phylodomain/index.php?user_org=Caenorhabditis_elegans&user_gene=ceh-20__PBC&user_num=50

We are glad that the reviewer has been prodding us to look at homeobox genes of other species. A few interesting observations emerged from this: As we were considering the issue of conservation, we took another look at our own data and noted something we only-half appreciated before: We had already stated that the C. elegans-specific homeobox genes (i.e. those ONLY present in C. elegans and NOT in other Caenorhabditis species) do NOT contribute to the all-nervous-system code. We simply took this now a step further and considered ONLY those homeobox genes that are conserved from C. elegans to human (71 genes). And those are totally sufficient to distinguish all neuron classes! So, going “parallel” (flies) or “up” (vertebrates), there are no obvious losses of homeobox genes to be considered. What about going simpler (e.g. cnidarians) – here again we were quite surprised to appreciate what has already been discussed quite a bit in the literature: (1) the last common bilaterian ancestor already has almost 60 homeobox genes; (2) if one goes outside bilateria to the arguably simplest nervous systems of cnidarians there are very few losses, if that. Hydra has ~50 homeobox genes and in Nematostella, the complement of homeobox genes has even expanded to ~130 (more than C. elegans)!

I would argue that this means that gain and losses of homeobox genes themselves may not matter a great deal - it’s rather a conserved set of homeobox genes that carry the load of specifying neuronal identity. That’s totally consisent with the patterns that we observed with the C. elegans-specific homeobox genes; most of them are ubiquitous or not at all in the nervous system. Of course, expression changes of these homeobox codes do hugely matter over the course of evolution (because new combos are expected to generate new neuron types), but the gain and loss of genes themselves may not.

It’s perhaps also worth pointing out that even if there were clear gains or losses of homeobox genes, it would presently be hard to correlate this with the gain or loss of specific neuronal cell types. “Gain” and “loss” of a specific trait implies that there IS a truly homologous traits shared by many species – which is then gained or lost in individual species. The problem is that we’re not quite there yet in regard to neuronal cell types – but we soon will: With now emerging scRNA datasets in C. elegans (the published ones are not yet deep enough, but some unpubl. ones are), we will soon be in a position to compare neuron types across phylogeny and identify truly homologous cell types. I think that this is a necessary prerequisite to then argue that the gain or loss of specific cell type MAY (or may not) correlate with the gain or loss of specific homeobox genes. But again, I would argue that novel cell types (via new homeobox codes) may not arise by gain/loss of a genetic locus, but rather by gain/loss of expression of homeobox genes.

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The revised paper along with the rebuttal does not change significantly the manuscript but it clarifies the points that were raised by the reviewers.
- The new point (not requested by reviewers!) about homeodomain proteins only conserved in nematodes not being involved in the code is an important addition that supports a more general role for this code in evolution.
- The addition of the search for a similar code in the recently published Drosophila (and mouse) data also serves an important generalization of the concept put forward by this paper. The authors rightfully mention that the mouse data are not homogeneous. However they claim that the Drosophila datas represents a "very very low sampling of cell types". 60+ neurons is more than half of the 118 classes in C elegans! So, this result is very significant!
- Finally about code-changing, I am still surprised that the authors have not performed at least a few such experiments....and I do not understand their argument that "our computational analysis does suggest that homeodomain protein combinations are able to account for the specificity of the molecular signatures of individual neuron types"!!

Therefore, this remains a very interesting concept that is presented and it will be utterly interesting to test whether the homeodomain protein code is sufficient to specify cell types, or if other transcription factors are also necessary. It is true that the many mutations that lead to loss of neuronal identity are in homeobox genes, although they belong to families that diverged many hundreds millions of years ago.

Referee #2 (Remarks to the Author):

The authors have addressed my very minor comments and the same weight of comments from Reviewer 1. I vote for publish AS IS. Congratulations to the authors. A very fine paper.

Author Rebuttals to First Revision:

N/A