Unique homeobox codes delineate all the neuron classes of *C. elegans*

https://doi.org/10.1038/s41586-020-2618-9

Received: 5 February 2020

Accepted: 4 June 2020

Published online: 19 August 2020

Check for updates

Molly B. Reilly[1], Cyril Cros[1], Erdem Varol[2], Eviatar Yemini[1] & Oliver Hobert[1]✉

It is not known at present whether neuronal cell-type diversity—defined by cell-type-specific anatomical, biophysical, functional and molecular signatures—can be reduced to relatively simple molecular descriptors of neuronal identity[1]. Here we show, through examination of the expression of all of the conserved homeodomain proteins encoded by the *Caenorhabditis elegans* genome[2], that the complete set of 118 neuron classes of *C. elegans* can be described individually by unique combinations of the expression of homeodomain proteins, thereby providing—to our knowledge—the simplest currently known descriptor of neuronal diversity. Computational and genetic loss-of-function analyses corroborate the notion that homeodomain proteins not only provide unique descriptors of neuron type, but also have a critical role in specifying neuronal identity. We speculate that the pervasive use of homeobox genes in defining unique neuronal identities reflects the evolutionary history of neuronal cell-type specification.

The classification of neurons into distinct types is an important step towards understanding the logic of the evolution, development and function of the nervous system[1]. Traditionally, the classification of neuron types has relied on anatomical features, and later expanded to include electrophysiological features and eventually molecular markers[1]. The emergence of high-throughput transcriptome profiling, including single-cell sequencing, has deepened our appreciation of the complexity of neuronal cell types among many different animal species, from very simple (for example, cnidarian) to very complex (mammals)[3–6]. Ongoing molecular classifications of neuron type raise a number of questions, including whether there is a minimal descriptor for neuronal identity—that is, whether specific subsets of molecular features exist that are sufficient to capture the full complexity of all neuronal cell types, or whether unique cellular identities can be described only by their combined expression of many different types of gene. Additionally, from a developmental standpoint, many questions remain about how the molecular signatures that characterize individual neuron types are genetically specified during differentiation.

Homeodomain transcription factors, which are encoded by homeobox genes[7], have emerged as possible answers to these questions. Loss-of-function studies in a number of organisms have demonstrated the importance of these transcription factors in the specification of neuronal cell types. For example, in *C. elegans*, the first neuronal-specification genes to be positionally cloned after unbiased mutant screens were homeobox genes (*mec-3*, *unc-4*, *unc-30* and *unc-86*)[8–11]. Subsequent mutant analysis revealed that many additional homeobox genes control neuronal identity in the nematode[12]. Homeobox genes have also surfaced as specifiers of neuronal identity in other organisms[7,13–17], and recent single-cell profiling of many regions of the mouse central nervous system has shown that homeobox genes are the gene family that best distinguishes neuron classes of the central nervous system[4]. A similar discriminatory power for homeobox gene expression—particularly, the combinatorial expression of distinct homeobox genes—has been revealed through the bulk sequencing of 179 distinct, genetically and anatomically identified cell populations in mouse[18]. Transcriptome analysis in the visual system and the ventral nerve cord of *Drosophila* has also revealed that homeobox genes display a more-discriminatory expression profile than that of genes that encode other types of transcription factors[19,20]. However, owing to the complexity of the mouse and fly nervous system, and the resulting incomplete coverage of all neuronal cell types, these previous studies have not been able to test the possibility that the expression of homeobox genes might uniquely identify every cell type in the entire nervous system. We test this possibility here in the context of the nervous system of the *C. elegans* model system. Fine-grained anatomical analysis of the adult hermaphrodite worm has classified its 302 neurons into 118 anatomically distinct types and several additional subtypes[21,22]. We set out to systematically address how much of this diversity in neuronal cell type can be explained by homeobox gene expression and function.

C. elegans homeobox genes

The *C. elegans* genome encodes 102 homeobox genes (Methods), less than half of the number of homeobox genes present in mammalian genomes[2,23,24]. As in other animal genomes, *C. elegans* homeodomain proteins do not constitute the largest family of transcription factors, accounting for only about 10% of all genes that encode transcription factors[25,26]. Of the 102 *C. elegans* homeobox genes, 70 have homologues in other invertebrate and vertebrate genomes, 18 are conserved only in nematodes and 14 are not conserved in any other known species of *Caenorhabditis*[2] (Fig. 1a). *Caenorhabditis elegans* contains representatives of most subclasses of mammalian homeobox genes, characterized by specific sequence features within the homeodomain (for example, paired-type homeodomain) or by the presence of additional domains

[1]Department of Biological Sciences, Howard Hughes Medical Institute, Columbia University, New York, NY, USA. [2]Department of Statistics, Columbia University, New York, NY, USA. ✉e-mail: or38@columbia.edu

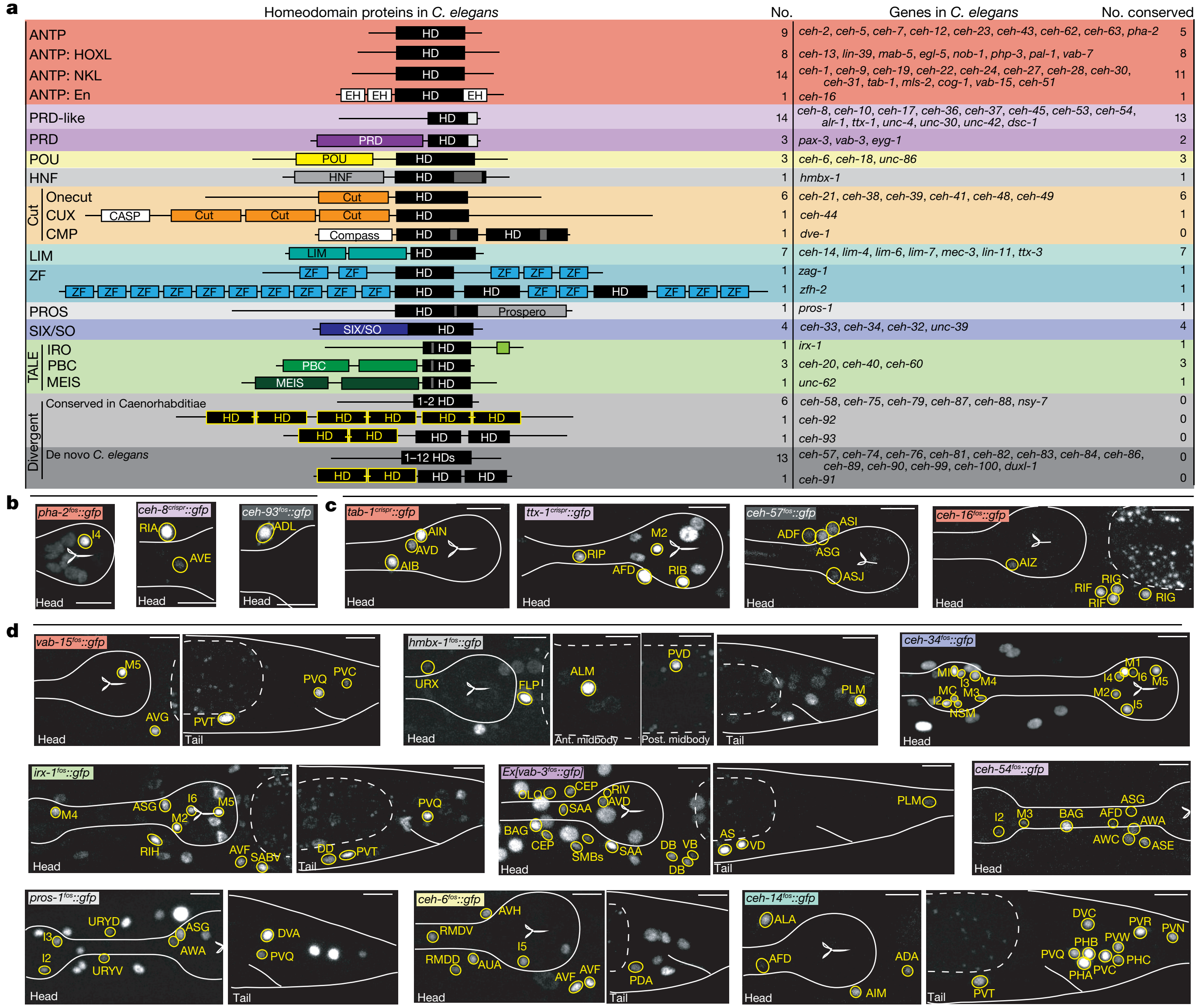
a
Homeodomain proteins in C. elegans
No.
Genes in C. elegans
No. conserved
ANTP | 9 | ceh-2, ceh-5, ceh-7, ceh-12, ceh-23, ceh-43, ceh-62, ceh-63, pha-2 | 5
ANTP: HOXL | 8 | ceh-13, lin-39, mab-5, egl-5, nob-1, php-3, pal-1, vab-7 | 8
ANTP: NKL | 14 | ceh-1, ceh-9, ceh-19, ceh-22, ceh-24, ceh-27, ceh-28, ceh-30, ceh-31, tab-1, mls-2, cog-1, vab-15, ceh-51 | 11
ANTP: En | 1 | ceh-16 | 1
PRD-like | 14 | ceh-8, ceh-10, ceh-17, ceh-36, ceh-37, ceh-45, ceh-53, ceh-54, alr-1, ttx-1, unc-4, unc-30, unc-42, dsc-1 | 13
PRD | 3 | pax-3, vab-3, eyg-1 | 2
POU | 3 | ceh-6, ceh-18, unc-86 | 3
HNF | 1 | hmbx-1 | 1
Cut: Onecut | 6 | ceh-21, ceh-38, ceh-39, ceh-41, ceh-48, ceh-49 | 6
CUX | 1 | ceh-44 | 1
CMP | 1 | dve-1 | 0
LIM | 7 | ceh-14, lim-4, lim-6, lim-7, mec-3, lin-11, ttx-3 | 7
ZF | 1 | zag-1 | 1
1 | zfh-2 | 1
PROS | 1 | pros-1 | 1
SIX/SO | 4 | ceh-33, ceh-34, ceh-32, unc-39 | 4
TALE: IRO | 1 | irx-1 | 1
PBC | 3 | ceh-20, ceh-40, ceh-60 | 3
MEIS | 1 | unc-62 | 1
Divergent: Conserved in Caenorhabditiae | 6 | ceh-58, ceh-75, ceh-79, ceh-87, ceh-88, nsy-7 | 0
1 | ceh-92 | 0
1 | ceh-93 | 0
De novo C. elegans | 13 | ceh-57, ceh-74, ceh-76, ceh-81, ceh-82, ceh-83, ceh-84, ceh-86, ceh-89, ceh-90, ceh-99, ceh-100, duxl-1 | 0
1 | ceh-91 | 0
HD
EH
PRD
POU
HNF
CASP
Cut
Compass
LIM
ZF
Prospero
SIX/SO
PBC
MEIS
1-2 HD
1–12 HDs
b
pha-2fos::gfp
I4
Head
ceh-8crispr::gfp
RIA
AVE
Head
ceh-93fos::gfp
ADL
Head
c
tab-1crispr::gfp
AIN
AVD
AIB
Head
ttx-1crispr::gfp
M2
RIP
AFD
RIB
Head
ceh-57fos::gfp
ADF
ASI
ASG
ASJ
Head
ceh-16fos::gfp
AIZ
RIF
RIG
RIF
RIG
Head
d
vab-15fos::gfp
M5
AVG
Head
PVQ
PVC
PVT
Tail
hmbx-1fos::gfp
URX
FLP
Head
ALM
Ant. midbody
PVD
Post. midbody
PLM
Tail
ceh-34fos::gfp
MI
I3
M4
MC
M3
I2
NSM
M1
I4
I6
M5
M2
I5
Head
irx-1fos::gfp
M4
ASG
I6
M2
M5
RIH
AVF
SABV
Head
PVQ
DD
PVT
Tail
Ex[vab-3fos::gfp]
OLQ
CEP
RIV
AVD
SAA
BAG
CEP
SMBs
SAA
DB
VB
DB
Head
PLM
AS
VD
Tail
ceh-54fos::gfp
I2
M3
BAG
AFD
ASG
AWA
AWC
ASE
Head
pros-1fos::gfp
URYD
ASG
I3
AWA
I2
URYV
Head
DVA
PVQ
Tail
ceh-6fos::gfp
AVH
RMDV
I5
RMDD
AUA
AVF
AVF
Head
PDA
Tail
ceh-14fos::gfp
ALA
AFD
ADA
AIM
Head
DVC
PVR
PVN
PHB
PVW
PVQ
PHC
PVC
PHA
PVT
Tail

Fig. 1 | The homeobox gene family in *C. elegans*, and representative expression patterns. a, Cartoon representations of homeodomain proteins and their associated domains by subfamily. Numbers and names of homeobox genes in *C. elegans*, and the number of conserved homeobox genes in humans, are based on ref. [2]. HD, homeodomain. Yellow 'HD' symbol indicates nematode-specific HOCHOB domain, a derivative of the homeodomain[2]. **b**–**d**, Representative images showing homeobox genes expressed in 1 or 2 neuron classes (**b**), 3 or 4 neuron classes (**c**) or 5 to 18 neuron classes (**d**). Ant, anterior; post, posterior. Neurons were identified by overlap with the NeuroPAL landmark strain, outlined and labelled in yellow. Head structures, including the pharynx, are outlined in white for visualization. Autofluorescence common to gut tissue is outlined with a white dashed line. Ten worms were analysed for each reporter strain. Scale bars, 10 μm. All other expression patterns are shown in Extended Data Figs. 1–8.

(for example, the POU or LIM domain)[2] (Fig. 1a). As in other animal genomes, only a small fraction of all *C. elegans* homeobox genes are of the Antennapedia-like HOX cluster[23,24].

Analysis of homeodomain protein expression

The expression patterns of a number of *C. elegans* homeobox genes have previously been reported, but often without individual neuron resolution and almost entirely with reporter reagents that do not capture the full complement of regulatory sequences[2,12,27] (Supplementary Table 1). To comprehensively analyse the expression pattern of homeodomain proteins throughout the entire nervous system, we used fosmid-based reporter transgenes that contain the full intergenic genomic context of the respective homeobox genes and/or we engineered *gfp* (encoding green fluorescent protein, GFP) into homeobox gene loci using CRISPR–Cas9 genome engineering. As expected, our fosmid and/or endogenous reporter alleles reveal many novel sites of expression of previously reported homeodomain proteins, in addition to providing expression patterns of many dozen previously uncharacterized homeodomain proteins (Supplementary Tables 1, 2). It is important to emphasize that our analysis delineates protein expression, thereby capturing post-transcriptional regulatory events that are not revealed through mRNA-based transcriptomic approaches.

We built an expression atlas of 101 of the 102 homeodomain proteins, including all of the 70 homeodomain proteins that are conserved outside the nematode phylum, plus all of the 18 nematode-specific homeodomain proteins, and 13 of the 14 *C.-elegans*-specific homeodomain proteins (that is, no homologues in the genomes of other *Caenorhabditis* species[2]). This atlas incorporates 97 homeodomain expression patterns that we established ourselves using fosmid reporters and/or CRISPR–Cas9-engineered reporter alleles, complemented with the

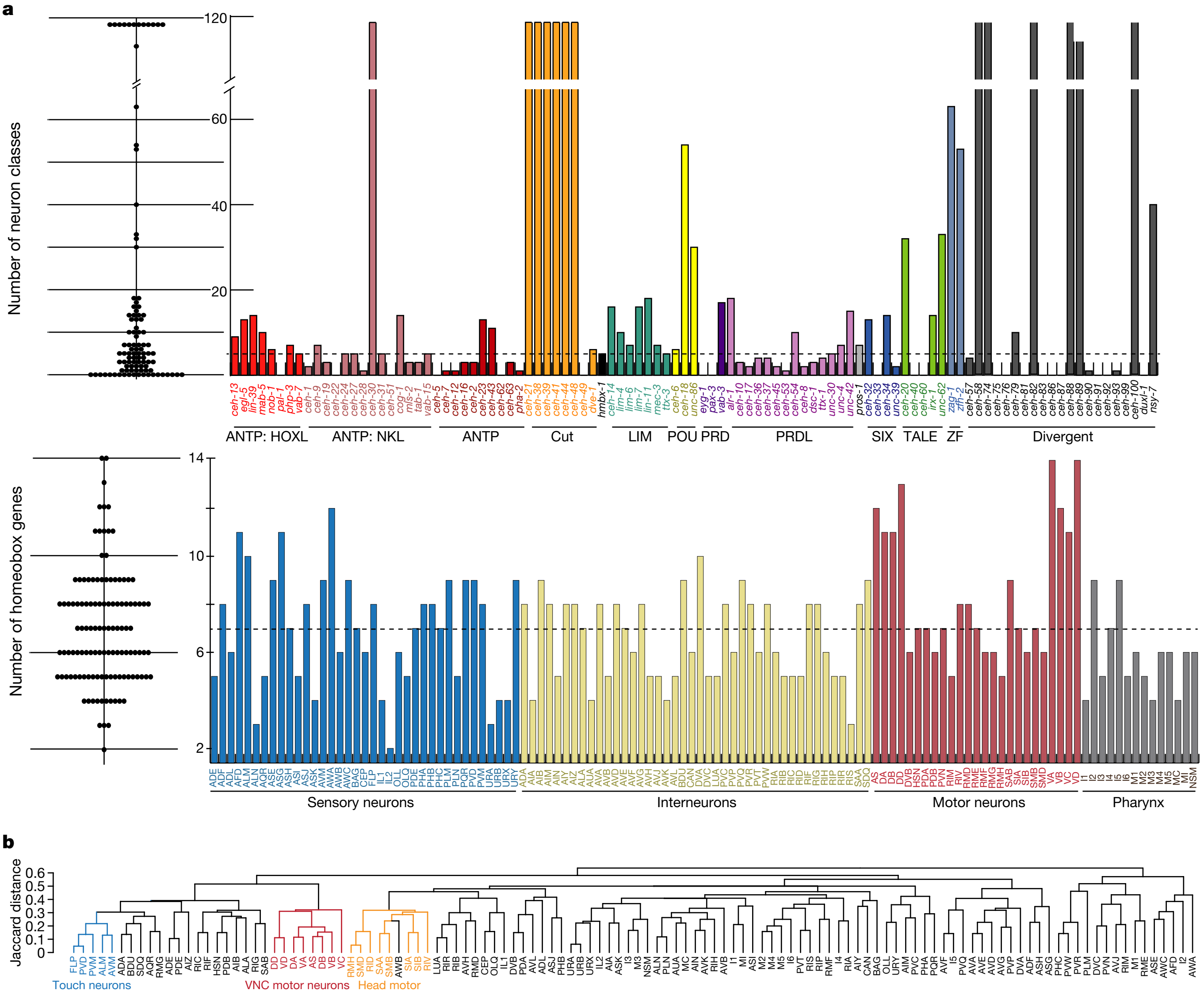

Fig. 2 | Summary of expression patterns of homeodomain proteins across the nervous system, and similarity of neuron classes on the basis of homeodomain codes. a, Top, number of neuron classes in which each homeobox gene is expressed. Left, dot plot distribution. Each dot represents a homeobox gene, and the value associated with the dot represents the number of neuron classes in which this homeobox gene is expressed. Right, histogram showing the number of neuron classes in which each homeobox gene is expressed, organized and coloured by homeobox gene subfamily and shared protein domains. The dashed line at 5 neuron classes denotes the median number of neuron classes in which each homeobox gene is expressed. Bottom, number of homeobox genes expressed in each neuron class. Left, dot plot distribution. Each dot represents a neuron, and the associated value represents the number of homeobox genes expressed in this neuron. Right, histogram displaying the number of homeobox genes expressed in each neuron class (excluding the pan-neuronal homeobox genes), ordered by the neuron type (sensory, motor, interneuron and pharyngeal). The dashed line at 7 homeobox genes denotes the median number of genes per neuron class. **b**, Dendrogram ordering neuron classes on the basis of the similarity of their homeobox gene code. Some examples of functionally related neuron groups are shaded. VNC, ventral nerve cord.

patterns of 4 previously fully characterized homeodomain patterns that were also generated either using fosmid or CRISPR–Cas9-engineered reporter alleles (Supplementary Table 1–3). We comprehensively analysed the expression pattern of all these homeodomain proteins at single-neuron resolution throughout all 302 neurons, using the multicolour-landmark identification transgene NeuroPAL[28]. We focused our expression analysis on mature neurons in the nervous system of late larval stage or young adult-stage worms, because continuous expression throughout the life of postmitotic neurons is usually associated with transcription factors that specify and subsequently maintain terminal neuron identity[12,29].

Notably, we find that 80 of the 101 homeodomain proteins we examined are expressed in the mature nervous system (Fig. 1b–d, Extended Data Figs. 1–7, Supplementary Tables 2, 3). Twelve are expressed in all neurons and many major tissue types; two Cut-type homeobox genes (*ceh-44* and *ceh-48*), as well as the nematode-specific *ceh-58* gene, are exclusively expressed in all neurons, but no other major tissue types (Fig. 1, Extended Data Figs. 3, 7). At the other extreme, seven homeodomain proteins are expressed exclusively in one neuron class (Fig. 1, Extended Data Figs. 1, 2, 5, 7). More than two thirds of the neuron-specific homeodomain proteins are expressed in less than 10% of all neuron classes (Fig. 2a). Neurons that express the same homeodomain protein are not usually related by lineage or by neurotransmitter identity (Extended Data Fig. 8). With the exception of pan-neuronally expressed homeodomain proteins, no two homeodomain proteins are expressed in the exact same combination of neuron classes (Supplementary Table 2). The two homeodomain proteins with the closest similarity in expression are UNC-62 (the orthologue of vertebrate MEIS

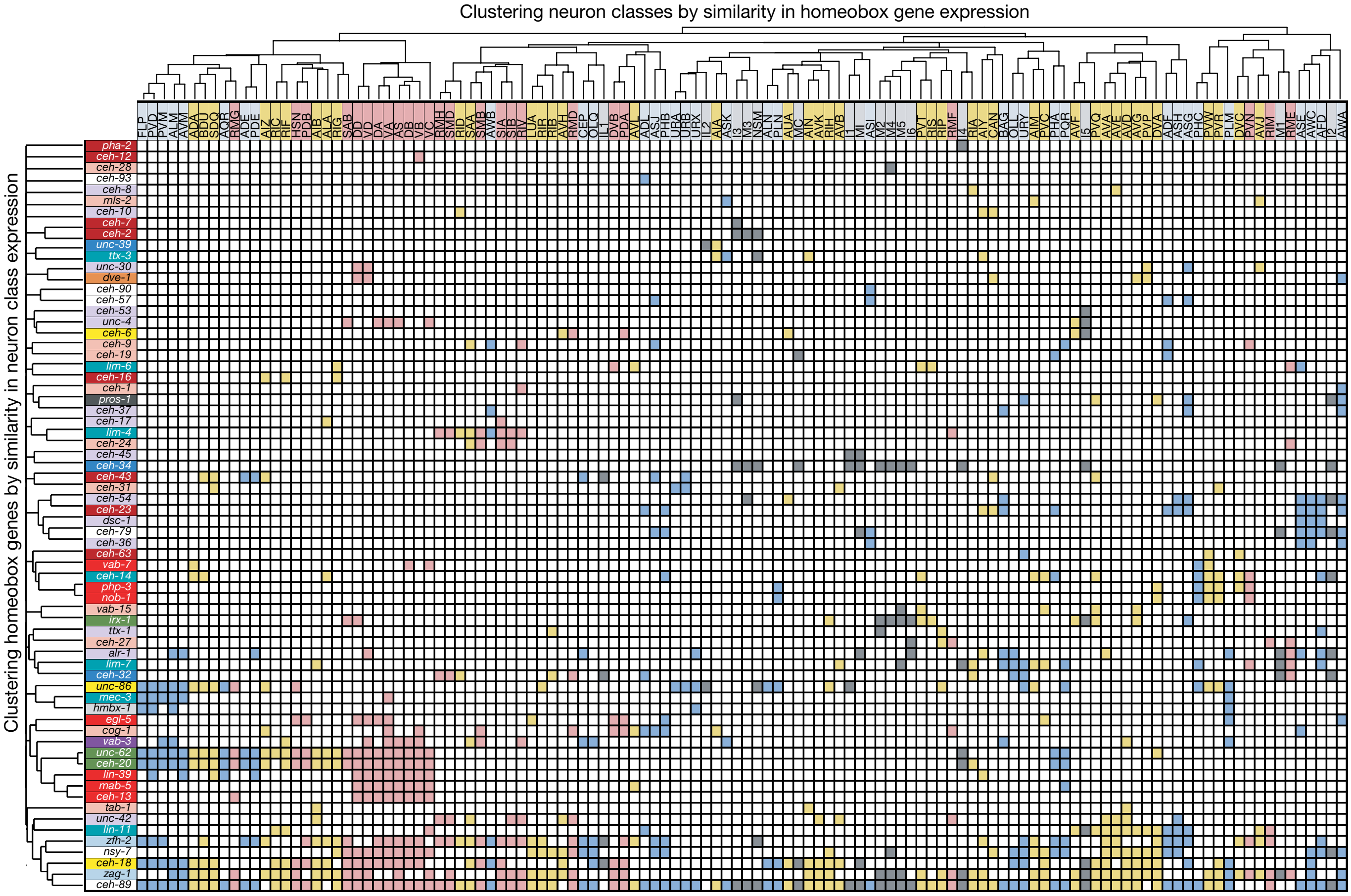

Fig. 3 | Homeodomain expression atlas for entire nervous system of *C. elegans*. Neuron classes are coloured by neuron type (blue, sensory; pink, motor; yellow, interneuron; and grey, pharyngeal) and ordered by similarity between neuron classes defined by the Jaccard index, as in Fig. 2b. Homeodomain proteins are coloured by subfamily, and ordered by similarity of neuron-class expression and sparsity.

proteins), which is expressed in 33 neuron classes, and CEH-20 (orthologous to vertebrate PBX proteins), which is expressed in 32 classes (31 of which are same as the classes that express UNC-62)—consistent with the mutual dependency of function of MEIS and PBX proteins in other organisms[30]. Tandem duplicated homeobox genes retain overlaps in their expression, but in most cases one of the duplicates shows an expression pattern that is much more restricted than the other (Supplementary Table 2).

The expression pattern of members of subclasses of homeodomain proteins (for example, POU, LIM and PRD) do not share obvious features: for example, there is no enrichment of specific homeodomain subclasses in sensory neurons versus interneurons or motor neurons, or in neurons of a specific neurotransmitter identity. The only exceptions are the above-mentioned Cut-type homeodomain proteins, which are either ubiquitously or pan-neuronally expressed. The cellular specificity of homeodomain protein within the nervous system appears to correlate with the extent of conservation. Of the 70 conserved homeodomain proteins, 56 (80%) are expressed in specific subsets of neurons, whereas only 10 out of the 18 (56%) nematode-specific proteins and only 3 of the tested 13 (27%) *C. elegans*-specific homeobox genes are selectively expressed in the nervous system (Extended Data Fig. 7, Supplementary Table 2). Some of the highly unusual *C. elegans*-specific homeodomain proteins[2]—such as CEH-100, which contains an unparalleled number of twelve homeodomains—are expressed in all cells and tissues, whereas the very unusual HOCHOB-type homeodomain protein CEH-91 displays no expression in the mature nervous system (Extended Data Fig. 7). The greater specificity of expression in individual neuronal cell types of conserved homeodomain proteins suggests that neuron-type-specific expression may be an ancestral feature of homeodomain protein expression.

Recently reported single-cell transcriptome sequencing recovered mRNA profiles for 42 of the 118 neuron classes[31,32]. Although these datasets recover homeobox gene transcripts in all of the 42 identified neuron classes, they uncover only little more than half (55%) of the expression profiles that we recovered via our protein expression analysis (Methods), which is probably a testament to the incomplete depth of single-cell RNA sequencing profiles (Supplementary Table 4). Vice versa, there are cases in which a homeobox gene transcript can be detected in cells in which we observe no expression of the corresponding protein (Supplementary Table 4), possibly owing to post-transcriptional regulatory events. Together, the comparison of our protein dataset with single-cell transcriptome data illustrates the limitations of the depth of currently available single-cell transcriptome datasets and the expected discordances between transcript and protein expression.

Homeodomain combinations define neuron types

The most notable feature of the homeodomain protein expression atlas becomes apparent when one considers the patterns of co-expression of homeodomain proteins in distinct neuron classes: each neuron class expresses its own entirely unique combination of homeodomain proteins. Excluding the pan-neuronally expressed homeobox genes, the combinatorial code consists of four homeodomain proteins on average (Fig. 2a). Neuron-type-specific homeodomain codes are generated by the 70 phylogenetically conserved homeobox genes alone

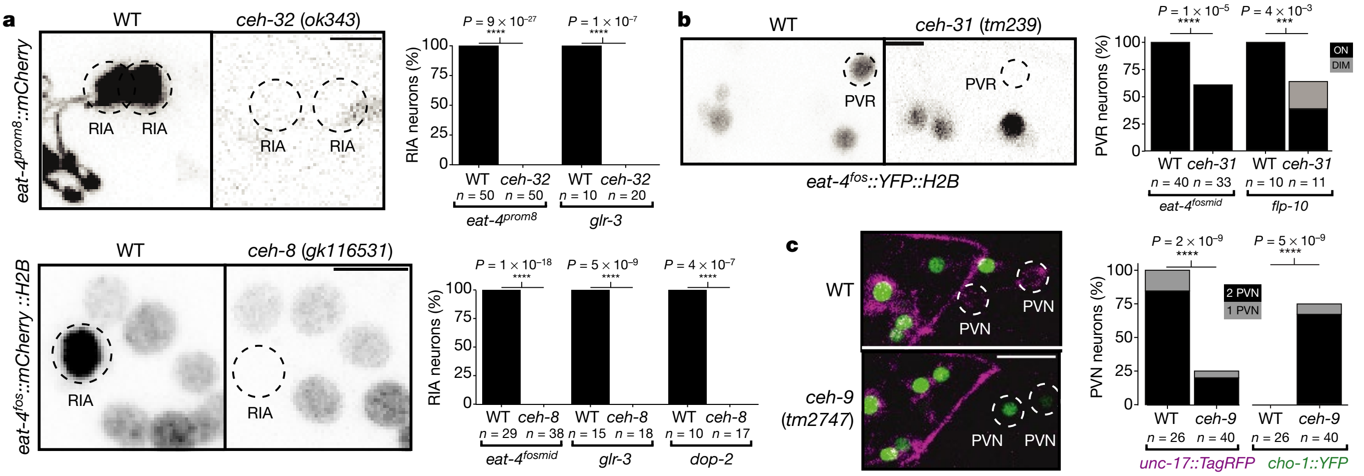

Fig. 4 | Previously uncharacterized homeobox genes act as regulators of neuronal identity. a–**c**, Each panel compares wild-type (WT) worms to null mutants, with about 30 independent worms per condition for neurotransmitter reporters and about 15 worms for other reporters. Graphs show *P* values from Fisher's exact test; ****P* values between 10^{-2} and 10^{-3}, *****P* values below 10^{-3}. Characteristic images were chosen. In **a**, RIA neuron identity is lost in *ceh-8* and *ceh-32* mutant worms, as assessed with multiple markers. Left, *eat-4* expression is lost from RIA neurons in a *ceh-8*- or *ceh-32*-mutant background. Right, quantification of *eat-4* loss in RIA neurons for both *ceh-8* and *ceh-32* mutant worms, as well as *glr-3* and *dop-2* reporters (see Extended Data Fig. 11 for reporter image). **b**, Glutamatergic identity of the PVR neuron is lost in *ceh-31*-mutant worms. Left, *eat-4* expression is lost in *ceh-31*-mutant background. Right, quantification of *eat-4* loss in PVR neurons, as well as a *flp-10* reporter (see Extended Data Fig. 11 for reporter image). **c**, PVN neuron fate change in *ceh-9*-mutant worms. Left, PVN neuron expresses only *unc-17* in a wild-type background. In a *ceh-9*-mutant background, *cho-1* fosmid expression is ectopically activated and *unc-17* expression is lost, indicating a change in cell fate. Right, quantification of *unc-17* and *cho-1* expression in wild-type and mutant worms. Scale bars, 5 μm (**a**, **b**), 10 μm (**c**).

(Extended Data Fig. 9a). Not all of the 70 conserved homeobox genes are required to generate neuron- class-specific codes. We calculated that the expression patterns of a minimal set of 24 conserved homeodomain proteins uniquely identify all 118 neuron class (Extended Data Fig. 9b).

We visualized the complete set of homeodomain codes using their Jaccard distance to construct a dendrogram, grouping neurons on the basis of the similarity of their unique homeodomain protein codes (Fig. 2b, Methods). When comparing this clustering to the relatedness of neuron classes on the basis of other anatomical or functional criteria, a number of expected and unexpected relationships were revealed. Broad classes of functionally related neurons (such as ventral-nerve-cord motor neurons, head motor neurons or touch-receptor neurons) clustered together on the basis of the similarity of their homeodomain protein codes (Fig. 2b). Notably, neurons that share similar codes and fall into related classes are not obviously related by lineage. However, functionally and anatomically related neuron classes can also display very different homeodomain protein codes, as seen—for example—in the case of the two interconnected, anatomically similar and functionally related phasmid sensory-neuron classes PHA and PHB (Fig. 2b, Supplementary Table 3). Conversely, several neuron classes that display no obvious functional or anatomical similarity clustered together on the basis of their homeodomain protein codes. For example, the amphid olfactory neuron AWB displays a code related to that of several head motor neurons.

We also clustered homeodomain proteins on the basis of similarity of their expression patterns. This dendrogram visualizes substantial differences in the expression patterns of individual homeodomain proteins (with a few notable exceptions, such as the MEIS and PBX similarities noted above) (Fig. 3). We used both of our dendrograms (that is, clustering homeodomain proteins on the basis of similarity of expression patterns, as well as clustering of neuron classes on the basis of similarity of homeodomain expression) to order the axes of our homeodomain expression matrix (Fig. 3). By grouping the most similar codes in proximity to each other, this illustrates the uniqueness of each homeodomain code per neuron class, providing the most succinct summary of the expression patterns of homeodomain protein throughout the *C. elegans* nervous system and visualizing the sparsity of this matrix (Fig. 3).

There are 118 anatomically defined neuron classes but 155 distinct combinatorial homeobox codes, which demonstrates that the homeobox codes reveal additional neuronal subclass identities (Extended Data Fig. 10a–c). For example, the six radially symmetric RMD neurons (composed of a dorsal and a ventral left–right symmetric neuron pair, and a lateral left–right symmetric pair) are uniquely defined by the combination of *ceh-89*, *nsy-7*, *unc-42*, *zfh-2* and *zag-1*, but the dorsal and ventral neuron pair is further distinguished by additional expression of *ceh-32* and *ceh-6* and the lateral pair by the additional expression of *cog-1*. The subclassification of the dorsal–ventral and the lateral RMD pair is paralleled by synaptic connectivity differences[21]. Similarly, the inner labial neuron class IL1—composed of six class members (a dorsal, lateral and ventral pair)—can be subdivided into subclasses by differential homeodomain expression patterns (all three neuron pairs co-express *ceh-43*, *ceh-32* and *ceh-18*, but only the dorsal and ventral pairs express *zfh-2*). This subclassification also mirrors the distinct synaptic connectivity patterns of the dorsal and ventral IL1 pairs versus the lateral IL1 pair[21].

Yet another example of homeodomain codes subdividing neuron classes is evident in ventral-nerve-cord motor neurons that are aligned along the anterior–posterior axis (Extended Data Fig. 10c). Distinct homeobox codes uniquely identify all known motor neuron classes (that is, DA, VA, AS and so on), but the expression of HOX cluster proteins further subdivides the identity of individual members of specific motor neuron classes (for example, DA1 versus DA2)—not only towards the tail of the ventral nerve cord (as previously reported[33,34]), but also in mid- and anterior domains of the ventral nerve cord. Moreover, every single post-embryonically generated motor neuron class expresses a diverse set of additional, non-HOX homeodomain proteins in a subclass-specific manner, including VAB-3 (the *C. elegans* orthologue of PAX6), VAB-7 (EVX1 and EVX2) or COG-1 (NKX6) (Extended Data Fig. 10c). Lastly, our homeobox data also revealed left–right asymmetries in the functionally lateralized ASE neuron pair[35], which we find express the homeobox genes *alr-1* and *ceh-23* in the left but not right ASE neuron (Extended Data Figs. 1, 5).

Homeodomain profiles predict neuron identity

We next set out to determine the extent to which the unique homeodomain expression code can account for the known molecular signatures of all *C. elegans* neurons. To this end, we used a Wormbase-curated list of 1,126 published reporter transgenes generated by the *C. elegans* research community over the past few decades[22]. This reporter atlas describes regulatory states for every single neuron type, with a sizable average of 42 reporters expressed per neuron type[22] (Supplementary Table 5). We used a simple multivariate linear regression to ask how well our homeodomain protein expression atlas (the independent variables) fit the remaining genes observed in neurons (the dependent results). We found that we could explain 74% of the reporter atlas expression at single-neuron resolution, using our sparse set of homeobox protein expression. This a significantly better fit than our control dataset ($P < 0.0001$), a randomly shuffled set of homeodomain protein expression data. To further illustrate the fit of our multivariate linear regression, we used this regression to predict reporter expression in each neuron class and correlated this prediction to the known reporter expression in these neuron classes (Extended Data Fig. 11a, Supplementary Table 5). Several classes of neuron have expression that is completely predicted by homeodomain protein expression (exhibiting a correlation coefficient of 1) and all of the remaining neuron classes show moderate-to-strong positive correlations (exhibiting coefficients between 0.5 and 0.95).

Functional relevance of homeobox genes

Experimental validation of the importance of the homeobox code had already been demonstrated by previous genetic loss-of-function analysis, which had shown that 40 of the 80 neuronally expressed *C. elegans* homeodomain proteins indeed have a role in the specification of neuronal identity[8–12] (Supplementary Table 2). We extended this functional analysis by examining homeobox genes that were not previously implicated in the specification of neuronal identity, and examining neurons for which no identity-promoting factor had previously been reported. We found that *ceh-8*, the *C. elegans* orthologue of the vertebrate *RAX* homeobox gene, and the SIX/SO-type homeobox gene *ceh-32*—both of which were uncharacterized in the context of the specification of neuronal identity—define a unique homeodomain expression code for the RIA interneurons (Extended Data Figs. 2, 5). In worms that carry a nonsense allele of either *ceh-8* or *ceh-32*, the RIA interneurons do not acquire a number of distinct features of RIA identity (Fig. 4a, Extended Data Fig. 11b–e).

We further examined whether any of our newly identified expression patterns of homeobox genes can distinguish previously defined, but non-discriminatory homeobox codes. The *unc-86* (an orthologue of *Brn3*) POU and *ceh-14* (an orthologue of *LHX3*) LIM homeobox genes were previously found to specify the identity of distinct classes of neuron—among them, the AIM and PVR neurons[36,37]. We discovered that the BarH homologue *ceh-31* is expressed in PVR—but not AIM—neurons, and that in *ceh-31*-mutant worms, the glutamatergic as well as peptidergic identity of PVR neurons is affected (Fig. 4b, Extended Data Fig. 11c). Similarly, we discovered that the NK-like homeobox gene *ceh-9* is required for the specification of the neurotransmitter mechanistic identity of the PVN neuron (Fig. 4c), a neuron that was previously found to be specified by a combination of *ceh-14* and *unc-3*, both of which also specify the PVC neuron[38]. The *ceh-9* homeobox gene therefore distinguishes *ceh-14*- and *unc-3*-dependent PVN from *ceh-14*- and *unc-3*-dependent PVC identity. Taken together, 74 of the 118 neuron classes of *C. elegans* have so far been found to require at least one (if not multiple) homeobox transcription factors for the correct specification of their identity (Extended Data Fig. 11f).

Conclusions

We have shown here that the expression patterns of a single transcription factor family fully describe the diversity of all neuronal cell types throughout an entire nervous system. Several transcriptome datasets from *Drosophila* and vertebrate nervous systems have also explicitly noted that homeobox genes are the gene family that distinguishes neuron types most effectively[4,18–20]. For example, bulk sequencing of large collections of distinct, labelled cell types throughout the mouse central nervous system also revealed that individual homeobox gene combinations distinguish almost all unique populations of neuronal cells[18]. However, to our knowledge, our analysis is the first to assign unique homeodomain protein codes to a whole nervous system in its entirety and with single-cell resolution. Transcriptome efforts from more-complex nervous systems will need to be scaled up substantially to assess the depth and breadth of combinatorial homeobox codes. As transcriptome datasets do not capture post-transcriptional regulatory events, ideally such transcriptome data need to be complemented by protein expression data, as we have shown here.

Future analysis will reveal whether other families of transcription factors display unique combinatorial expression patterns throughout the nervous system. It is already clear that non-homeodomain types of transcription factors also have critical roles in neuronal identity specification (for example, ref. [12]) but such non-homeodomain transcription factors often cooperate with homeodomain transcription factors in the control of neuron identity in *C. elegans*[33,38–40]. Inspired by Dobzhansky's dictum that 'nothing in biology makes sense except in the light of evolution'[41], we speculate that the potential preponderance of homeobox genes in the specification of neuronal identity may hint at the possibility that homeodomain proteins were recruited into the specification of neuronal identity very early in the evolution of the nervous system. It is possible that a homeodomain transcription factor was used to specify the signal properties of an ancestral 'ur-neuron' (the evolutionarily earliest, most-primitive form of a neuron). Different neuronal cell types could have come into existence through homeobox gene duplication, and an ensuing diversification of expression and target specificity of the duplicated homeodomain proteins. Homeobox expression codes may therefore provide a window in the evolutionary history of neuronal cell types.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2618-9.

1. Zeng, H. & Sanes, J. R. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat. Rev. Neurosci.* **18**, 530–546 (2017).
2. Hench, J. et al. The homeobox genes of *Caenorhabditis elegans* and insights into their spatio-temporal expression dynamics during embryogenesis. *PLoS ONE* **10**, e0126947 (2015).
3. Sebe-Pedros, A. et al. Cnidarian cell type diversity and regulation revealed by whole-organism single-cell RNA-seq. *Cell* **173**, 1520–1534 (2018).
4. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014 (2018).
5. Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
6. Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
7. Gehring, W. J. *Master Control Genes in Development and Evolution: The Homeobox Story* (Yale Univ. Press, 1998).
8. Way, J. C. & Chalfie, M. *mec-3*, a homeobox-containing gene that specifies differentiation of the touch receptor neurons in *C. elegans*. *Cell* **54**, 5–16 (1988).
9. Finney, M., Ruvkun, G. & Horvitz, H. R. The *C. elegans* cell lineage and differentiation gene *unc-86* encodes a protein with a homeodomain and extended similarity to transcription factors. *Cell* **55**, 757–769 (1988).
10. White, J. G., Southgate, E. & Thomson, J. N. Mutations in the *Caenorhabditis elegans unc-4* gene alter the synaptic input to ventral cord motor neurons. *Nature* **355**, 838–841 (1992).
11. Jin, Y., Hoskins, R. & Horvitz, H. R. Control of type-D GABAergic neuron differentiation by *C. elegans* UNC-30 homeodomain protein. *Nature* **372**, 780–783 (1994).
12. Hobert, O. A map of terminal regulators of neuronal identity in *Caenorhabditis elegans*. *Wiley Interdiscip. Rev. Dev. Biol.* **5**, 474–498 (2016).

13. Tsuchida, T. et al. Topographic organization of embryonic motor neurons defined by expression of LIM homeobox genes. *Cell* **79**, 957–970 (1994).
14. Lindtner, S. et al. Genomic resolution of DLX-orchestrated transcriptional circuits driving development of forebrain GABAergic neurons. *Cell Rep*. **28**, 2048–2063 (2019).
15. Stettler, O. & Moya, K. L. Distinct roles of homeoproteins in brain topographic mapping and in neural circuit formation. *Semin. Cell Dev. Biol*. **35**, 165–172 (2014).
16. Tahayato, A. et al. Otd/Crx, a dual regulator for the specification of ommatidia subtypes in the *Drosophila* retina. *Dev. Cell* **5**, 391–402 (2003).
17. Blochlinger, K., Bodmer, R., Jack, J., Jan, L. Y. & Jan, Y. N. Primary structure and expression of a product from cut, a locus involved in specifying sensory organ identity in *Drosophila*. *Nature* **333**, 629–635 (1988).
18. Sugino, K. et al. Mapping the transcriptional diversity of genetically and anatomically defined cell populations in the mouse brain. *eLife* **8**, e38619 (2019).
19. Davis, F. P. et al. A genetic, genomic, and computational resource for exploring neural circuit function. *eLife* **9**, e50901 (2020).
20. Allen, A. M. et al. A single-cell transcriptomic atlas of the adult *Drosophila* ventral nerve cord. *eLife* **9**, e54074 (2020).
21. White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Phil. Trans. R. Soc. Lond. B* **314**, 1–340 (1986).
22. Hobert, O., Glenwinkel, L. & White, J. Revisiting neuronal cell type classification in *Caenorhabditis elegans*. *Curr. Biol*. **26**, R1197–R1203 (2016).
23. Bürglin, T. R. & Affolter, M. Homeodomain proteins: an update. *Chromosoma* **125**, 497–521 (2016).
24. Bürglin, T. R., Finney, M., Coulson, A. & Ruvkun, G. *Caenorhabditis elegans* has scores of homoeobox-containing genes. *Nature* **341**, 239–243 (1989).
25. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
26. Fuxman Bass, J. I. et al. A gene-centered *C. elegans* protein–DNA interaction network provides a framework for functional predictions. *Mol. Syst. Biol*. **12**, 884 (2016).
27. Murray, J. I. et al. Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. *Nat. Methods* **5**, 703–709 (2008).
28. Yemini, E. et al. NeuroPAL: a neuronal polychromatic atlas of landmarks for whole-brain imaging in *C. elegans*. Preprint at https://www.biorxiv.org/content/10.1101/676312v1 (2019).
29. Hobert, O. Terminal selectors of neuronal identity. *Curr. Top. Dev. Biol*. **116**, 455–475 (2016).
30. Merabet, S. & Mann, R. S. To be specific or not: the critical relationship between Hox and TALE proteins. *Trends Genet*. **32**, 334–347 (2016).
31. Packer, J. S. et al. A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* **365**, eaax1971 (2019).
32. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
33. Kratsios, P. et al. An intersectional gene regulatory strategy defines subclass diversity of *C. elegans* motor neurons. *eLife* **6**, e25751 (2017).
34. Schneider, J. et al. UNC-4 antagonizes Wnt signaling to regulate synaptic choice in the *C. elegans* motor circuit. *Development* **139**, 2234–2245 (2012).
35. Hobert, O. Development of left/right asymmetry in the *Caenorhabditis elegans* nervous system: from zygote to postmitotic neuron. *Genesis* **52**, 528–543 (2014).
36. Serrano-Saiz, E. et al. Modular control of glutamatergic neuronal identity in *C. elegans* by distinct homeodomain proteins. *Cell* **155**, 659–673 (2013).
37. Serrano-Saiz, E., Oren-Suissa, M., Bayer, E. A. & Hobert, O. Sexually dimorphic differentiation of a *C. elegans* hub neuron is cell autonomously controlled by a conserved transcription factor. *Curr. Biol*. **27**, 199–209 (2017).
38. Pereira, L. et al. A cellular and regulatory map of the cholinergic nervous system of *C. elegans*. *eLife* **4**, e12432 (2015).
39. Lloret-Fernández, C. et al. A transcription factor collective defines the HSN serotonergic neuron regulatory landscape. *eLife* **7**, e32785 (2018).
40. Doitsidou, M. et al. A combinatorial regulatory signature controls terminal differentiation of the dopaminergic nervous system in C. elegans. *Genes Dev*. **27**, 1391–1405 (2013).
41. Dobzhansky, T. Biology, molecular and organismic. *Am. Zool*. **4**, 443–452 (1964).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Homeobox gene list

Previous sequence analysis identified 103 *C. elegans* homeobox genes[2]. A more recent evaluation of sequences revealed that one gene (*ceh-85*) is a pseudogene (www.wormbase.org), which brings the total number of homeobox genes considered here down to 102.

Generation of expression reagents

Previously reported expression patterns of homeodomain proteins relied, in a few cases, on antibody staining, but the patterns of expression of these proteins in the nervous system were either incompletely or not completely correctly identified (for example, VAB-7 and UNC-30, which are revised in this Article), owing to a lack of molecular landmarks for proper cellular identification. With only three exceptions (*ttx-3*, *unc-86* and *unc-42*, all of which used both fosmid and/or endogenous reporter alleles generated by CRISPR–Cas9), all other previously reported expression patterns of homeobox genes were determined using reporter transgenes that did not contain the entire gene locus, which—as we show here—results in substantial underestimations of expression patterns (summarized in Supplementary Tables 1, 2).

Here we examined the expression patterns of 20 homeodomain proteins by tagging the respective endogenous locus with *gfp* via CRISPR–Cas9 genome engineering. To this end, *gfp* was inserted at the 3′ end of the gene, immediately before the stop codon. For *vab-7*, *lin-11*, *ceh-37* and *zfh-2*, these reporter alleles were generated using the self-excising cassette method for CRISPR–Cas9 genome engineering[42]. *ceh-44* and *ceh-49* reporter alleles were provided by E. Leyva Díaz, and were generated as previously described[43]. CRISPR–Cas9-engineered strains with the strain name PHX were created by Sunybiotech. Sixty homeodomain proteins were examined using available chromosomally integrated fosmid reporters lines generated by ModEncode (not previously examined for neuron-type-specific expression in the nervous system)[44], and an additional six homeodomain proteins were examined using fosmid reporters (made by the ModEncode project[44]) that we injected ourselves. All fosmid reporters included 3′ tagged protein fusions as well. Injections were done into OH15430 (*otis669;pha-1*(*e2123*)) worms at 10 ng/µl with 3 ng/µl *pha-1*(+) and 100 ng/µl OP50 genomic DNA to create independent lines. A list of all reporter strains is provided below.

As expected from the usual compactness of *C. elegans* gene loci and the size of fosmid reporters (about 40 kb of genomic sequence, usually containing several genes up- and/or downstream of the gene of interest), so far we have not found a single instance in which the fosmid reporters do not fully recapitulate expression patterns observed with a reporter allele generated by CRISPR–Cas9 genome engineering. Such comparisons have been explicitly made with the transcription factors *unc-42* (E. Berghoff, pers. comm.), *ttx-3* (V. Bertrand, pers. comm.), *lin-39* (ref. [45]), *unc-3*(ref. [46]) and *che-1* (ref. [47]).

Strain list for expression analysis

All newly generated strains used in this study are publicly available from the *Caenorhabditis* Genetics Center. The strains for the respective homeobox genes are as follows: *alr-1:* OP200; *wgIs200 [alr-1::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-1:* OP571; *wgIs571 [ceh-1::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-12:* OH16368; *otEx7486[ceh-12::TY1::EGFP::3xFLAG + unc-119(+) + pha-1(+)]; ceh-13:* OH16366; *otEx7484[ceh-13::TY1::EGFP::3xFLAG + unc-119(+) + pha-1(+)]; ceh-14:* OP73; *wgIs73 [ceh-14::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-16:* OP82; *wgIs82 [ceh-16::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-17:* OH16369; *otEx7487[ceh-17::TY1::EGFP::3xFLAG + unc-119(+) + pha-1(+)]; ceh-18:* OP533; *wgIs533 [ceh-18::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-19:* OP739; *wgIs739 [ceh-19::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-2:* OP323; *wgIs323 [ceh-2::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-20:* RW12211; *ceh-20(st12211[ceh-20::TY1::EGFP::3xFLAG]); ceh-21, ceh-39, ceh-41:* OP759; *wgIs759 [ceh-41::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-22:* OP389; *wgIs389 [ceh-22::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-23:* PHX1849; *ceh-23(syb1849[ceh-23::GFP)]; ceh-24:* PHX1608; *ceh-24(syb1608[ceh-24::GFP)]; ceh-27:* OP135; *wgIs135 [ceh-27::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-28:* OH16367; *otEx7485[ceh-28::TY1::EGFP::3xFLAG + unc-119(+) + pha-1(+)]; ceh-30:* OP120; *wgIs120 [ceh-30::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-31:* OP370; *wgIs379 [ceh-31::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-32:* OH16477; *ceh-32(ot1040[ceh-32::GFP::LoxP::3x FLAG]) V; ceh-33:* OP575; *wgIs575 [ceh-33::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-34:* OP524; *wgIs524 [ceh-34::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-36:* OP620; *wgIs620 [ceh-36::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-37:* OH16345; *ceh-37(ot1023[ceh-37::GFP::FLAG]); ceh-38:* OP241; *wgIs241 [ceh-38::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-40:* OP232; *wgIs232 [ceh-40::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-43:* OH10447; *otIs339 [ceh-43::gfp; ttx-3::dsred; rol-6]; ceh-44:* OH16219; *ceh-44(ot1015[ceh-44::gfp]); ceh-45:* OH16370; *otEx7488[ceh-45::TY1::EGFP::3xFLAG + unc-119(+) + pha-1(+)]; ceh-48:* OP631; *wgIs631 [ceh-48::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-49:* OH16224; *ceh-49(ot1016[ceh-49::gfp]); ceh-5:* PHX1592; *ceh-5(syb1592[ceh-5::GFP)]; ceh-51:* PHX1551; *ceh-51(syb1551[ceh-51::GFP)]; ceh-53:* OP444; *wgIs444 [ceh-53::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-54:* OP456; *wgIs456 [ceh-54::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-57:* OP706; *wgIs706 [ceh-57::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-58:* PHX2015; *ceh-58(syb2015[ceh-58::GFP)]; ceh-6:* RW10871; *wgIs87[ceh-6::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-60:* DLS395; *ceh-60(rhd395 [HA-mCherry::ceh-60]); ceh-62:* OP416; *wgIs416 [ceh-62::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-63:* OP742; *wgIs741 [ceh-63::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-7:* OP168; *wgIs681[ceh-7::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-74:* OP680; *wgIs680 [ceh-74::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-75:* PHX1884; *ceh-75(syb1884[ceh-75::GFP)]; ceh-76:* OH16487; *ceh-76(ot1042[ceh-76::GFP]) ceh-79:* OP553; *wgIs553 [ceh-79::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-8:* PHX1656; *ceh-8(syb1656 [ceh-8::GFP)]; ceh-81:* OH16479; *otEx7569 [ceh-81:TY1::EGFP::3xFLAG + unc-119(+)]; ceh-82:* OP212; *wgIs212 [ceh-82::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-83:* OP727; *wgIs727 [ceh-83::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-86:* PHX2517; *ceh-86(syb2517[ceh-86::GFP)]; ceh-87:* PHX1955; *ceh-87(syb1995[ceh-87::GFP)]; ceh-88:* OP593; *wgIs593 [ceh-88::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-89:* OH16505; *ceh-89(ot1050[ceh-89::GFP]); ceh-9:* OP690; *wgIs690 [ceh-9::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-90:* OP210; *wgIs210 [ceh-90::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-91:* OH16480; *otEx7570 [ceh-91:TY1::EGFP::3xFLAG + unc-119(+)]; ceh-92:* PHX1610; *ceh-92(syb1610 [ceh-92::GFP)]; ceh-93:* OP554; *wgIs554 [ceh-93::TY1::EGFP::3xFLAG + unc-119(+)]; ceh-99:* OH16481; *otEx7571 [ceh-99:TY1::EGFP::3xFLAG + unc-119(+)]; ceh-100:* OH16488; *ceh-100(ot1043[ceh-100::GFP]); cog-1:* OP541; *wgIs541 [cog-1::TY1::EGFP::3xFLAG + unc-119(+)]; dsc-1:* OP522; *wgIs522[dsc-1::TY1::EGFP::3xFLAG + unc-119(+)]; duxl-1:* OP470; *wgIs470 [duxl-1::TY1::EGFP::3xFLAG + unc-119(+)]; dve-1:* OP398; *wgIs398 [dve-1::TY1::EGFP::3xFLAG + unc-119(+)]; egl-5:* OP54; *wgIs54 [egl-5::TY1::EGFP::3xFLAG + unc-119(+)]; eyg-1:* OP441; *wgIs441 [eyg-1::TY1::EGFP::3xFLAG + unc-119(+)]; hmbx-1:* OP655; *wgIs655 [hmbx-1::TY1::EGFP::3xFLAG + unc-119(+)]; irx-1:* OP536; *wgIs536 [irx-1::TY1::EGFP::3xFLAG + unc-119(+)]; lim-4:* OP681; *wgIs681 [lim-4::TY1::EGFP::3xFLAG + unc-119(+)]; lim-6:* OP387; *wgIs387 [lim-6::TY1::EGFP::3xFLAG + unc-119(+)]; lim-7:* OP15; *wgIs15[lim-7::TY1::EGFP::3xFLAG + unc-119(+)]; lin-11:* OH15910; *lin-11(ot958[lin-11::GFP::FLAG]); lin-39:* OP18; *wgIs18 [lin-39::TY1::EGFP::3xFLAG + unc-119(+)]; mab-5:* OP27; *wgIs27 [mab-5::TY1::EGFP::3xFLAG + unc-119(+)]; mec-3:* OP55; *wgIs55 [mec-3::TY1::EGFP::3xFLAG + unc-119(+)]; mls-2:* OP645; *wgIs654 [mls-2::TY1::EGFP::3xFLAG + unc-119(+)]; nob-1*: JIM271; *stIs10286 [nob-1::GFP::unc-54 3′UTR + rol-6(su1006)]; nsy-7:* OH16371; *otEx7489[nsy-7:TY1::EGFP::3xFLAG + unc-119(+) + pha-1(+)]; pal-1:* OP380; *wgIs380 [pal-1::TY1::EGFP::3xFLAG + unc-119(+)]; pax-3:* OP190; *wgIs190 [pax-3::TY1::EGFP::3xFLAG + unc-119(+)]; pha-2:* OP687; *wgIs687 [pha-2::TY1::EGFP::3xFLAG + unc-119(+)]; php-3:* PHX1549;

php-3(syb1548[php-3::GFP]); pros-1: OP500; *wgIs500 [ceh-26::TY1::EGFP::3xFLAG + unc-119(+)]; tab-1:* PHX1587; *tab-1(syb1587[tab-1::GFP]); ttx-1:* PHX1679; *ttx-1(syb1679[ttx-1::GFP)]; unc-30:* OP395; *wgIs395[unc-30::TY1::EGFP::3xFLAG + unc-119(+)]; unc-39:* OP186; *wgIs186 [unc-39::TY1::EGFP::3xFLAG + unc-119(+)]; unc-4:* PHX1658; *unc-4(syb1658[unc-4::GFP)]; unc-62:* SD1871; *wgIs600 [unc-62::TY1::EGFP::3xFLAG + unc-119(+)]; vab-15:* OP730; *wgIs730 [vab-15::TY1::EGFP::3xFLAG + unc-119(+)]; vab-3:* FQ1092; *wzEx302[vab-3::GFP + Pflp-17::DsRed]; vab-7:* OH15912; *vab-7(ot959[vab-7::GFP::FLAG]); zag-1:* OP83; *wgIs83 [zag-1::TY1::EGFP::3xFLAG + unc-119(+)];* and *zfh-2:* OH16346; *zfh-2(ot1024[zfh-2::GFP::FLAG])*. The *unc-42* reporter lines will be described elsewhere (E. Berghoff and O.H., manuscript in preparation).

Microscopy

Worms were anaesthetized using 100 mM of sodium azide (NaN_3) and mounted on 5% agarose pads on glass slides. Images were acquired using confocal laser scanning microscopes (Zeiss LSM800 and LSM880) and processed using ImageJ software[48]. For expression of reporters, representative maximum intensity projections are shown for the GFP channel as greyscale, and gamma and histogram were adjusted for visibility. For mutant functional analysis, representative maximum intensity projections are shown as an inverted greyscale. NeuroPAL images provided in the Supplementary Information are pseudocoloured in accordance with ref. [28]. All reporter reagents and mutants were imaged at 40× using fosmid or CRISPR reagents, unless otherwise noted.

Examination of expression reagents and neuron identification

Some obviously panneuronal or ubiquitous genes were determined to be expressed in all neurons by crossing the reporter strain with *otIs314*, a *rab-3* fosmid driving TagRFP. For all of the remaining genes, colocalization with the NeuroPAL landmark strain (*otIs669* or *otIs696*) was used to determine the identity of all neuronal expression[28]. For CRISPR–Cas9-generated strains and integrated fosmid strains, the reporter strain was crossed with the NeuroPAL landmark strain. To analyse fosmid expression with available DNA but no integrated strain, fosmid DNA was injected into the NeuroPAL landmark strain OH15430 (*otIs669;pha-1*(*e2123*)) as a rescuing array with *pha-1*(+) DNA. Three extrachromosomal lines were created and analysed for each extrachromosomal fosmid strain to determine the expression of that gene. Generally, the expression of a given reporter gene was stable over all worms scored. In the few cases in which we observed variable expression of fosmid reporter genes (for example, *ceh-8* and *ceh-24*), we generated reporter alleles by CRISPR–Cas9, resulting in more stable expression. In terms of expression level, for every gene expressed in multiple neurons, we noticed different levels of expression in different neuron classes (Extended Data Figs. 1–8). Expression (even if very dim) was counted as present if seen across multiple worms. This is because even the dim expression of a homeodomain transcription factor has been shown have functional phenotypes. For example, *ceh-14* is bright in all neuron types in which it is expressed except AFD and I2, but has previously been shown to control the specification of the AFD neurons[36,49].

We also noticed many cases of additional expression of well-characterized homeobox genes, the expression of which had previously been studied with suboptimal reporter reagents. In some cases, the new sites of expression are relatively dim, whereas in other cases they are strong. Two such examples are a fosmid reporter of the LIM homeobox *mec-3*, which is brightly expressed in previously identified touch neurons[50] and is less bright in posterior VA neurons (which we describe here). By contrast, a CRISPR–Cas9-engineered reporter allele of the *unc-4* locus is—within the context of the ventral nerve cord—equally bright in the previously identified VA and DA motor neuron classes[51], as it is in the AS motor neurons which we identify here as *unc-4*-expressing.

Although we did not notice obvious differences in expression patterns between late larval stages and adult, a number of genes are clearly expressed in additional cells in the embryo.

Clustering using the Jaccard index

To assess the similarities among neuron classes by homeobox genes, we used the Jaccard index. This index is used to measure similarity between finite sample sets by calculating the intersection of those sets divided by their union. For our data, we calculated the number of shared homeobox genes between each neuron class in a pairwise manner, and then divided them by the number of shared and unshared homeobox genes in those pairs. To cluster this data, we created a distance matrix for the degree of dissimilarity between each neuron class based on their homeobox gene codes, calculated as 1 − Jaccard similarity index. With this distance matrix, we clustered our data using the hierarchical clustering tool hclust (available in R), an open source software environment for statistical computing.

We did this same analysis for the degree of similarity among homeobox genes by their expression in shared neuron classes. In this calculation, the number of shared neuron classes between each homeobox gene was counted in a pairwise manner, and then divided by the number of shared and unshared neuron classes in which those genes expressed. We again created a distance matrix (1 − Jaccard index), clustered the data using hclust.

Minimal code of homeobox genes

Given a set of redundant codes of homeodomain coexpression for each neuron, we aim to reduce this codebook to one in which there are no redundancies and each cell is represented by a unique barcode. The problem of codebook reduction is cast as a multidimensional knapsack problem[52] with binary weight constraints. The goal of this problem is to find the minimum set of homeodomain codes such that no two neurons share the same barcode. The global optimum solution is then found through a branch-and-bound scheme[53] that yields the minimum subset of bits that can be conserved from the genetic codebook and that ensures uniqueness of cell barcodes.

Correlation of homeobox and reporter expression

We used a Wormbase-curated list of 1,126 published reporter transgenes available, with new homeobox gene expression data added in Supplementary Table 5. To test for correlations between reporter transgene expression in specific neurons and homeobox gene expression, we removed all homeobox gene expression profiles from the Wormbase-curated list. We then performed a simple linear regression using the lm function in R: we fitted lm(G ~ TF), in which G is the reporter expression by neuron class matrix and TF is the homeobox expression by neuron class 1. To assess the goodness of our fit, we also shuffled the homeobox expression matrix 1,000 times. This gave us an R^2 value of 0.74 for our actual homeobox expression dataset, which compared favourably to the 0.41 achieved with the control shuffled homeobox expression dataset. We then set out to verify how good this correlation was across individual neuron classes, as the number of available reporters they express is variable. The fitted values from the above regression predict an expected reporter expression for each neuron class, on the basis of their homeobox gene expression. For each neuron class, we extracted these fitted values and compared them to the actual transgene expression profiles reported using the correlation function in R (cor) using the standard Pearson method. These correlation values are shown in Extended Data Fig.11a.

Mutant analysis scoring and statistics

Reporter expression was scored as an all-or-nothing phenotype per neuron, with expression in 0, 1 or 2 neurons. Scoring data was processed in R and converted as contingency tables of the number of expressing neurons by genotype. Statistical analysis was then done using Fisher's exact test (under a two-sided null hypothesis), using Holm's method to correct for multiple comparisons. The resulting adjusted *P* values are all below 0.001. No statistical methods were used to determine sample size before the experiment. On basis of the common standard in the field, we aimed for about 30 worms per genotype for neurotransmitter reporters and about 15 worms for other markers.

ceh-32(*ok343*) mutant worms arrested at L1 were maintained with an *otEx7146 ceh-32* fosmid rescue construct. Worms were only counted as *ceh-32* mutants when the *myo-2::mCherry* coinjection marker of this array was not visible at all. The *ceh-32* mutants arrested at L1 were scored against their wild-type counterpart strain at L1, rather than with the rescued worm of the same strain. Owing to the disorganization of their head ganglions, glutamatergic identity in RIA neurons was instead scored using a short integrated *eat-4* promoter fragment (*otIs521*) with a restricted expression pattern in only a subset of glutamatergic neurons[36]. Scoring was done under a Zeiss stereo dissecting scope at high magnification, and representative images from confocal microscopy are shown at 63×. One or two very dim cells were seen in less than 15% of the *ceh-32*-mutant worms under confocal microscopy, but these cells made no axonal projection and their cell body did not match the shape of RIA neurons. Reported *P* values would still be significant if they were conservatively counted as *eat-4*(+) RIA neurons.

For the mutant analysis, the following strains were used: OH13094 *otIs354[cho-1fos::YFP]; otIs518 [eat-4fos::mCherry]*; OH15958 *otIs354[cho-1fos::YFP]; otIs518 [eat-4fos::mCherry]; ceh-8(gk116531)*; *IK705(njIs10[glr-3p::GFP]*; OH15970 *njIs10[glr-3p::GFP]; ceh-8(gk116531)*; OH4793 *otIs173 [F25B3.3::DsRed2 + ttx-3pB::GFP]; otEx980 [dop-2::GFP + pha-1(+)]*; OH16478 *otIs173 [F25B3.3::DsRed2 + ttx-3pB::GFP]; otEx980 [dop-2::GFP + pha-1(+)]; ceh-8(gk116531)*; OH16253 *otIs354[cho-1fos::YFP]; ot907(unc-17::mKate2 CRISPR)*; OH16251 *otIs354[cho-1fos::YFP]; ot907(unc-17::mKate2 CRISPR), ceh-9(tm2747)*; OH16256 *otIs580 [cho-1 fos::mCherry + eat-4fos::YFP]*; OH16201 *otIs580 [cho-1fos::mCherry + eat-4fos::YFP] ceh-31(tm239)*; OH16204 *otIs92[flp-10p::GFP]*; OH16203 *otIs92[flp-10p::GFP]; ceh-31(tm239)*; OH12525 *otIs521[eat-4prom8::tagRFP; ttx-3::gfp]*; OH16314 *otIs521[eat-4prom8::tagRFP; ttx-3::gfp], otIs388[eat-4fos::YFP], ceh-32(ok343) otEx7146[ceh-32 fosmid rescue WRM0637dA10 + myo-2 RFP])*; *IK705 njIs10[glr-3p::GFP]*; and OH16476 *ceh-32(ok343) V; njIs10[glr-3p::GFP]; otEx7146[ceh-32 fosmid rescue WRM0637dA10 + myo-2 RFP]*.

Comparison of homeobox expression with single-cell RNA-sequencing data

To analyse the congruence between available single-cell RNA sequencing (scRNA-seq) data[31,32] and our reported homeodomain expression, we used the provided bootstrap median data (averaging resampled RNA levels 1,000 times) from refs. [31,32], and applied no cut off (that is, any transcripts per million value >0 counted as real expression). We then directly compared the binary expression profiles of the homeobox gene mRNA in isolated neuron classes with our reported homeodomain protein expression (coloured in keys in the figures). We found that the scRNA-seq expression data from the 42 identified L2 neuron classes recapitulated only 38% of our homeodomain protein expression. We calculated this percentage by taking the agreed expression (blue) and dividing it by the agreed expression plus the expression seen only in the homeodomain protein analysis (blue + red). We then asked whether scRNA-seq was able to detect mRNA of our homeodomain proteins at earlier embryonic time points. To this end, we added the scRNA-seq embryo data available for these 42 neuron classes, and found that this increased the coverage to 55%. This percentage was calculated as above with the agreed expression divided by the agreed plus the expression seen only in the homeodomain protein analysis.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All newly generated data, including the expression pattern of every homeobox gene, are available in Supplementary Tables 1, 2. Additionally, whole-worm confocal images of all homeobox genes analysed are available in Extended Data Figs. 1–8. Newly generated reporter strains made during this study are available from the *Caenorhabditis* Genetics Center. The most up-to-date version of the community-curated transgene expression resource used is available in Supplementary Table 5.

Code availability

The R code used to generate the Jaccard distance matrix for the clustering of homeobox genes and neuron classes is available on the GitHub of the O.H. laboratory, at https://github.com/hobertlab/Reilly_2020/tree/master/Jaccard_Distance. The MATLAB code used to create the minimal codebook of homeobox genes is available at https://github.com/hobertlab/Reilly_2020/tree/master/Minimal_Codebook.

42. Dickinson, D. J., Pani, A. M., Heppert, J. K., Higgins, C. D. & Goldstein, B. Streamlined genome engineering with a self-excising drug selection cassette. *Genetics* **200**, 1035–1049 (2015).
43. Dokshin, G. A., Ghanta, K. S., Piscopo, K. M. & Mello, C. C. Robust genome editing with short single-stranded and long, partially single-stranded DNA donors in *Caenorhabditis elegans*. *Genetics* **210**, 781–787 (2018).
44. Sarov, M. et al. A genome-scale resource for in vivo tag-based protein function exploration in *C. elegans*. *Cell* **150**, 855–866 (2012).
45. Feng, W. et al. A terminal selector prevents a Hox transcriptional switch to safeguard motor neuron identity throughout life. *eLife* **9**, e50065 (2020).
46. Patel, T. & Hobert, O. Coordinated control of terminal differentiation and restriction of cellular plasticity. *eLife* **6**, e24100 (2017).
47. Leyva-Díaz, E. & Hobert, O. Transcription factor autoregulation is required for acquisition and maintenance of neuronal identity. *Development* **146**, dev177378 (2019).
48. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
49. Cassata, G. et al. The LIM homeobox gene *ceh-14* confers thermosensory function to the AFD neurons in *Caenorhabditis elegans*. *Neuron* **25**, 587–597 (2000).
50. Way, J. C. & Chalfie, M. The *mec-3* gene of *Caenorhabditis elegans* requires its own product for maintained expression and is expressed in three neuronal cell types. *Genes Dev.* **3** (12A), 1823–1833 (1989).
51. Miller, D. M., III & Niemeyer, C. J. Expression of the unc-4 homeoprotein in *Caenorhabditis elegans* motor neurons specifies presynaptic input. *Development* **121**, 2877–2886 (1995).
52. Kellerer, H., Pferschy, U. & Pisinger, D. *Knapsack Problems* (Springer, 2004).
53. Schrijver, A. *Theory of Linear and Integer Programming* (John Wiley & Sons, 1998).
54. Mukaka, M. M. Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* **24**, 69–71 (2012).

Acknowledgements We thank Q. Chen for expert technical assistance in generating transgenic lines; E. Berghoff for communicating the unpublished expression of *unc-42;* E. Leyva-Díaz for CRISPR-tagging *ceh-44* and *ceh-48;* R. Dowen for sending a *ceh-60* reporter allele; V. Bertrand for communicating unpublished results on *ceh-10* and *ttx-3* reporter alleles; L. Glenwinkel for updating and sharing the community-curated gene expression resource; J. Booth for creation of select extrachromosomal fosmid reporter strains; Y. Ramadan for help with mutant analysis; and T. Bürglin for comments on the manuscript. This work was funded by a predoctoral fellowship to M.B.R. (F31 NS105398), by NIH R21 NS106843, and by the Howard Hughes Medical Institute. Some strains were provided by the CGC, which is funded by NIH Office of Research Infrastructure Programs (P40 OD010440).

Author contributions M.B.R. and O.H. designed the experiments and wrote the manuscript, M.B.R. generated constructs and conducted the expression pattern analysis, C.C. conducted genetic loss-of-function experiments and contributed to writing the paper, E.V. and E.Y. conducted the bioinformatic analysis and contributed to writing the paper.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2618-9.
Correspondence and requests for materials should be addressed to O.H.
Peer review information Peer reviewer reports are available.
Reprints and permissions information is available at http://www.nature.com/reprints.