Changes in Postural Syntax Characterize Sensory Modulation and Natural Variation of *C. elegans* Locomotion

Roland F. Schwarz[1], Robyn Branicky[2], Laura J. Grundy[2], William R. Schafer[2], André E. X. Brown[3]*

1 European Bioinformatics Institute, Hinxton, United Kingdom, **2** MRC Laboratory of Molecular Biology, Cambridge, United Kingdom, **3** MRC Clinical Sciences Centre, Imperial College London, London, United Kingdom

* andre.brown@imperial.ac.uk

Abstract

Locomotion is driven by shape changes coordinated by the nervous system through time; thus, enumerating an animal's complete repertoire of shape transitions would provide a basis for a comprehensive understanding of locomotor behaviour. Here we introduce a discrete representation of behaviour in the nematode *C. elegans*. At each point in time, the worm's posture is approximated by its closest matching template from a set of 90 postures and locomotion is represented as sequences of postures. The frequency distribution of postural sequences is heavy-tailed with a core of frequent behaviours and a much larger set of rarely used behaviours. Responses to optogenetic and environmental stimuli can be quantified as changes in postural syntax: worms show different preferences for different sequences of postures drawn from the same set of templates. A discrete representation of behaviour will enable the use of methods developed for other kinds of discrete data in bioinformatics and language processing to be harnessed for the study of behaviour.

Author Summary

Technology for recording neural activity is advancing rapidly and whole-brain imaging with single neuron resolution has already been demonstrated for smaller animals. To interpret such complex neural recordings, we need comprehensive characterizations of behaviour, which is the principal output of the brain. Animal tracking can increasingly be performed automatically but an outstanding challenge is finding ways to represent these behavioural data. We have focused on the movement of the nematode worm *C. elegans* to develop a quantitative representation of behaviour as a series of distinct postures. Each posture is analogous to a word in a language and so we can directly count the number of phrases that makes up the *C. elegans* behavioural repertoire. *C. elegans* has a very small nervous system but we find that its behavioural repertoire is still complex. As with human languages, there is a large number of possible phrases, but most are rarely used. When comparing different populations of worms or worms in different environments, we find

that the difference between their behaviour is due to a subset of their entire repertoire. In the language analogy, these would correspond to idiomatic phrases that distinguish groups of speakers. A quantitative understanding of the nature of behavioural variation will inform research on the function and evolution of neural circuits.

Introduction

Animal behaviour can be difficult to study because it involves complex movements that vary continuously in time. This is especially true for limbed animals such as insects and mammals, because they have many degrees of freedom and can adopt an even larger number of different postures. Detailed characterization without explicit posture quantification is possible [1], but researchers commonly use alternative methods for controlling behavioural complexity. In ethology, complexity is often controlled through careful observation to identify a set of behavioural categories that can then be scored to assess differences between individuals, between species, and over time [2]. This has the advantage that observations can be carried out in an animal's natural habitat where it is able to behave freely. However, the choice of categories may introduce unwanted bias and inter-observer variability can reduce reproducibility [3]. In laboratory-based studies it is possible to impose constraints on the animal, ranging from physically restraining some parts of the body to training an animal to perform a predefined task such as pressing a lever or solving a maze. Constraints reduce measurement dimensionality and can aid interpretation but may also mask variation which may be important for understanding the genetic or neural control of behaviour [4].

Here we adopt an increasingly common [5,6] intermediate approach in which we record animals in a laboratory environment, but with minimal constraints. *Caenorhabditis elegans* is a useful animal for addressing questions of behavioural representation because it has a simple morphology and its movement can be confined to the two-dimensional surface of an agar plate. It has also been characterized exhaustively in several relevant respects: in addition to a complete genome sequence [7], we know the developmental lineage of all of its cells from the single-celled embryo to the adult [8], and it has the most complete connectome—the wiring diagram of its neurons—of any animal [9]. These reference data are powerful in part because of their completeness, making it possible to characterize differences due to genetic, evolutionary, or environmental perturbations systematically. An analogous systematic characterization of *C. elegans* locomotion could also be a useful tool for studying behaviour.

To move on the surface of an agar plate, *C. elegans* propagates bends along its body. It is not clear whether the locomotor wave is generated by a central pattern generator [10], but there is evidence that once generated, waves are driven by proprioceptive coupling between body segments [11]. During navigation, *C. elegans* switches between bouts of forward and backwards locomotion [12] and uses both abrupt and gradual turns during directed behaviours [13,14] and when searching in a relatively featureless environment [15,16]. While performing these manoeuvres, *C. elegans* adopts postures in a low dimensional shape space [17]. However, knowing the dimensionality of the space still leaves open the question of which regions of shape space worms visit and which trajectories they tend to follow. The method we describe here is meant to address this question directly.

We previously described a method of constructing a dictionary of behavioural motifs [18], which are short patterns of repeated motion, and showed that such a dictionary could be useful for identifying and comparing mutant phenotypes in *C. elegans*. However, because motifs can overlap there will be significant redundancy in the representation before completeness is

achieved. To solve this problem, we have developed an alternative method of characterising locomotion in terms of a sequence of discrete postures, which serve as building blocks for more complex motifs. We found that 90 distinct postures are sufficient to account for 82% of postural variance. A discrete representation enables efficient enumeration of postural sequences and makes it possible to address questions about behavioural plasticity in response to stimulation and natural variation in spontaneous locomotion. We found that freely moving worms have a core set of frequently used sequences and a much larger set of rarely used sequences. When worms adapt their behaviour in response to environmental conditions such as a lack of food or the presence of an attractive chemical, we find that they adjust the syntax of their locomotion (i.e. their preference for different postural sequences), but that their shapes are still accounted for using the same set of template postures. Similarly, between populations of *C. elegans* isolated from different parts of the world, we find that they have a similar rank-frequency distribution of behaviours, but with significant variations in sequence preference.

Results

Ninety template postures capture wild type worm shapes on food

We tracked single worms using a Dino-Lite USB microscope mounted on a motorized stage, as previously described [19]. Worms were segmented from the background using the Otsu threshold and the skeleton was calculated by tracing the midline along the worm body connecting the head and the tail. We then measured the tangent angle at 49 points equally spaced along the worm and subtracted the average angle to achieve a position and orientation independent measure posture [17]. These angles change as the worm bends its body during crawling. Although a worm's posture varies continuously through time, at any given moment the current posture can be approximated by the closest matching template from a pre-defined set of postures. This is analogous to constructing a histogram, in which discrete bins are used to represent data drawn from a continuous distribution.

The first step was to select a set of template postures that was as small as possible while still having enough elements to capture most of the observed variability. Instead of using uniformly spaced bins to define the templates we used k-means clustering, which attempts to minimize the distance between k templates (the cluster centres) and the data (the set of input postures measured from wild-type worms crawling on an agar plate seeded with bacterial food). As a distance metric, we used the squared difference between the tangent angle vectors of two worm skeletons. As the number of template postures is increased (k in k-means), the R^2 between the original postures and their nearest neighbours improves (**Fig 1A**). After 90 postures, the rate of improvement plateaus at a small value (**Fig 1A**, inset). We therefore used 90 postures for subsequent analyses. The most common postures in this set of templates are sinuous and correspond to different phases of locomotion at different curvatures (**Fig 1B**). The rare postures include more highly curved shapes that worms adopt during turns as well as straighter shapes that are occasionally observed during dwelling. The full set of 90 postures ordered by frequency is plotted in **S1 Fig**. Using this set of template postures, locomotion can be represented compactly. At any point in time, a worm's posture is matched to its nearest-neighbour in the set of template postures and so a bout of locomotion can be simply recorded as a series of labels indicating which of the representative postures is closest (**Fig 1C**).

Posture sequences are complex but countable

Repeated patterns of behaviour can now be found simply by finding repeating sequences of postures, including subsequences sampled from longer segments. It is also possible to match similar behaviours performed at different speeds by removing repeated elements from a
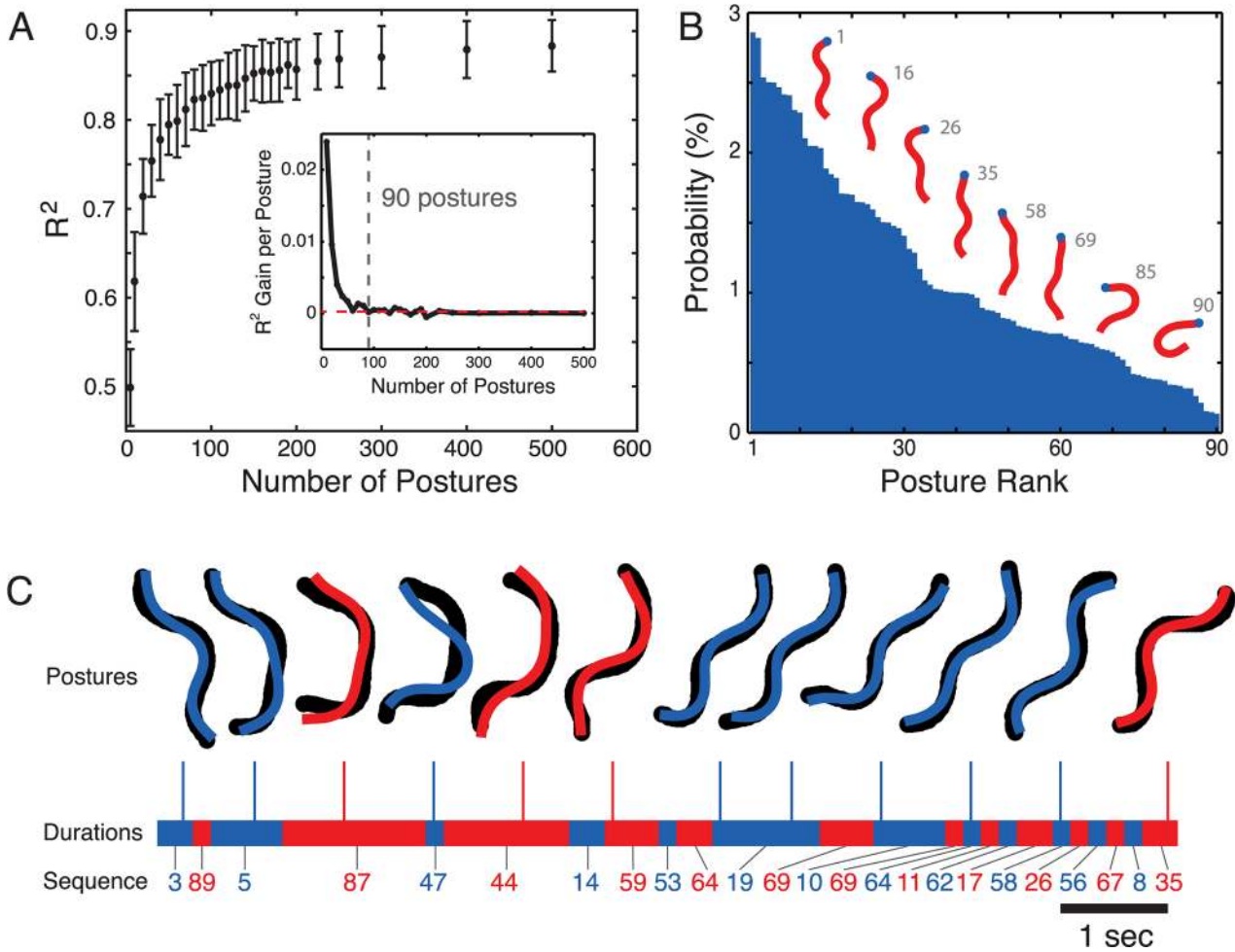
Fig 1. Discrete representation of worm locomotion by clustering. (A) As the number of template postures increases, the R^2 of the fit between original body angles and nearest neighbour matches increases (points show mean ± standard deviation for 50 worms, 15 000 frames each). The inset shows the rate of increase in R^2 for increasing k. The red dashed line is the average increase in R^2 per posture for $k = 100-500$ (2.4×10^{-4} per posture). **(B)** Probability of observing each of the 90 template postures. Selected postures are shown as skeletons with the head indicated by the blue dot. The grey label indicates their rank by frequency with 1 being the most common posture. **(C)** Locomotion is represented by a series of discrete template postures. In each frame, the original skeleton (black) is represented by its nearest neighbour (coloured overlay) in the set of 90 template postures. The duration spent in each posture is also recorded and shown as the alternating coloured bands. Each segment of repeated postures is collapsed to a single label and it is this sequence of states that represents locomotion (i.e. {3, 3, 89, 5, 5, 5, 5, 87, 87, 87, . . .} becomes {3, 89, 5, 87}).

doi:10.1371/journal.pcbi.1004322.g001

sequence. For example, if a worm is moving slowly, its nearest-neighbour template posture might not change over several frames leading to a sequence with repeats such as {3, 3, 89, 5, 5, 5, 5, 87, 87, 87, . . .} (see durations bar in Fig 1C). We reduce this sequence to {3, 89, 5, 87} so that it will be recognized regardless of how slowly the transitions are made. We used this simple kind of non-uniform time warping for all of the sequence results presented in this paper.

In a discrete representation, the behavioural repertoire can be enumerated directly by counting the variety of sequences that are observed. We do this by counting the number of observed *n*-grams. An *n*-gram is any sequence of *n* states, including overlaps. In the sequence {3, 89, 5, 87} the bigrams are {3, 89}, {89, 5}, and {5, 87}, the trigrams are {3, 89, 5} and {89, 5, 87}, and the sequence is itself a single 4-gram. For longer sequences, there are more possible *n*-grams but because of repeats, the number of *unique* *n*-grams will be less than the total number observed. There is therefore a sublinear growth in the number of unique *n*-grams as more

sequences are observed and the degree of this difference is a measure of the structure in the sequence. **Fig 2A** shows how the number of unique *n*-grams, for *n* = 2–5, grows with total sequence length for the laboratory strain N2. Individual blue lines show data from 100 worms, each recorded for 15 minutes, randomly selected from the total data set of 1262 individuals. Orange lines show accumulation curves for sequences that have been shuffled to destroy temporal order but maintain the relative frequency of postures.

The accumulation curves grow less than linearly, reflecting the fact that worm locomotion is repetitive. It is notable though, that even after over 100 000 postures have been observed, new unique trigrams are still being observed at a rate of approximately 1 in 25 (or about 1 previously unobserved trigram every 12 seconds after a total observation time of 14 hours). This combination of repetitiveness with a large total repertoire can be seen more clearly in a Zipf plot of *n*-gram frequencies (**Fig 2B**), which shows the frequency of sequences plotted against their rank on a log-log plot. The distribution is heavy-tailed in the sense that there are more high-frequency behaviours than would be expected for an exponential distribution with the same mean (**S2 Fig**). This corresponds to a highly skewed usage of the behavioural repertoire with the top 1% of trigrams being used 32.2 ± 0.6% of the time during crawling on food (mean ± standard deviation for the blue curves in Fig 2B).

We used the OpenGRM NGram library [20] to model posture sequences as Markov chains, in which the probability of the next posture in the sequence is a function only of the previous *n*– 1 states. In the simplest case of a bigram model, the probability of the next state in the sequence depends only on the current state, but for worm locomotion, a bigram model does not capture the distribution of longer *n*-grams (**S3 Fig**). In contrast, when we fit a trigram model using observed sequences and used it to generate new sequences, the simulated sequences had a similar distribution not only of trigrams, but also 4- and 5-grams (Fig 2A and 2B, black lines). Increasing *n* to 4 or 5 improves the fit quality, but only modestly (S2 Fig). We therefore chose to focus on trigrams for most of the subsequent analysis.

Optogenetic activation changes posture *n*-gram frequencies

To determine how worms adapt their behavioural repertoire in response to stimulation, we first used optogenetics to induce behavioural transitions at predictable times in freely crawling worms. We chose two strains for these experiments that produce robust behavioural responses to optogenetic stimulation. ZX46 worms were a useful starting point because they express channelrhodopsin under the control of the *unc-17* promoter in the worm's cholinergic neurons, and thus activation by blue light results in the simultaneous contraction of body wall muscles and an obvious change in locomotion. Previous experiments on this strain showed that activation leads to a deep dorsal bend [21]. Although dorsal bends are rare under normal circumstances, they are represented in the set of 90 N2 on-food postures and the fit quality during activation actually increases somewhat (**Fig 3A**). This is because curved postures have a high total variance leading to an R^2, the fraction of explained variance, near 1. Plots of relative posture probability over time (**Fig 3B**) show some postures that are over-represented during the stimulus and a large number of under-represented postures. A rank sum test comparing posture probabilities during and outside of stimulation periods shows that dorsal bends are over-represented (**Fig 3C**) during stimulation, as previously reported for this strain [21]. The same effect is also seen in the bigram probabilities, although there was insufficient power to detect over-represented trigrams. The under-represented postures are sinuous postures common during normal locomotion. This is due to a decrease in forward locomotion during photo-activation of the cholinergic neurons as revealed by the under-represented bigrams and trigrams.
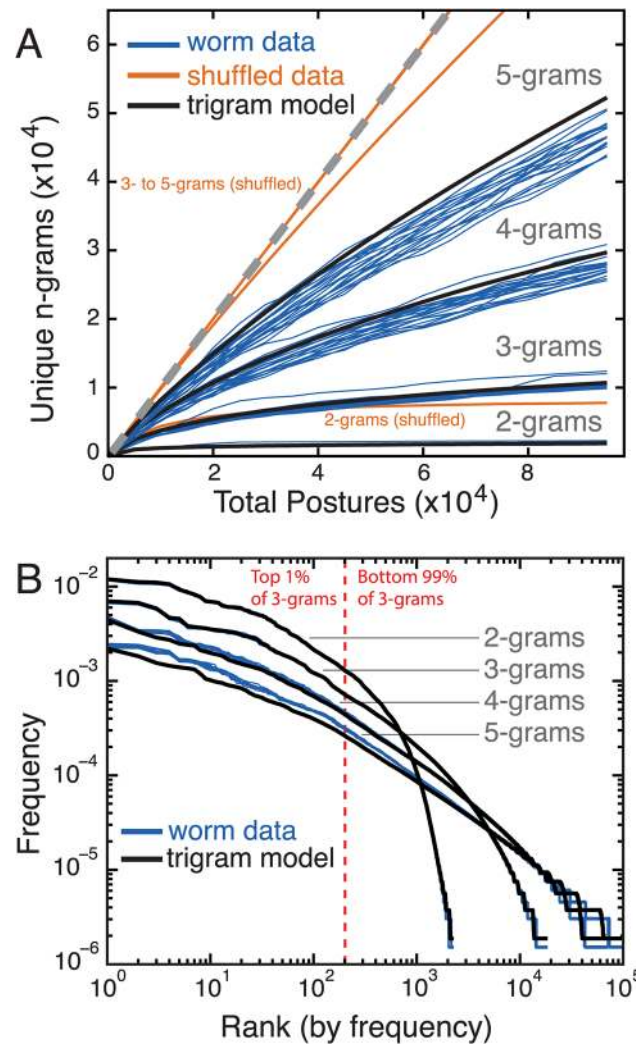
Fig 2. Worm locomotion sequences are complex but can be fit by a trigram model. (A) *n*-gram accumulation curves for bigrams (sequences of two postures) to 5-grams (sequences of 5 postures). Over time, as more postures are observed (Total Postures), the number of unique *n*-grams that is observed (analogous to vocabulary size) grows sub-linearly. The grey dashed line has slope 1. Each blue line is calculated from data from 100 randomly chosen worms from the entire data set of 1262 individuals. Orange lines are averages calculated from the same data but with the sequences randomly shuffled. Black lines are averages calculated for data generated from a trigram model of the posture sequences. **(B)** Zipf plot of *n*-grams. The frequency distribution of *n*-grams ranked from most to least frequent. Each blue line is calculated from 400 randomly chosen worms. Black lines are averages calculated for data generated from a trigram model of the posture sequences. The red dashed line indicates the rank that divides the top 1% most frequent trigrams from the remaining 99%.

Photoactivation of AQ2026 worms, which express channelrhodopsin under the control of the *sra-6* promoter, reliably induces an escape response through the activation of the ASH sensory neurons [22,23]. The behaviour is readily observed, but is more naturalistic than the deep dorsal bends seen in ZX46 worms since it occurs through activation of a sensory neuron that normally detects aversive stimuli and elicits an escape response. Again, the discrete representation fits the induced postures well (**Fig 3D**). The plot of posture probability over time shows both over- and under-represented postures during and after stimulation (**Fig 3E**). Escape responses in *C. elegans* typically consist of a reversal followed by a turn, which we can detect by
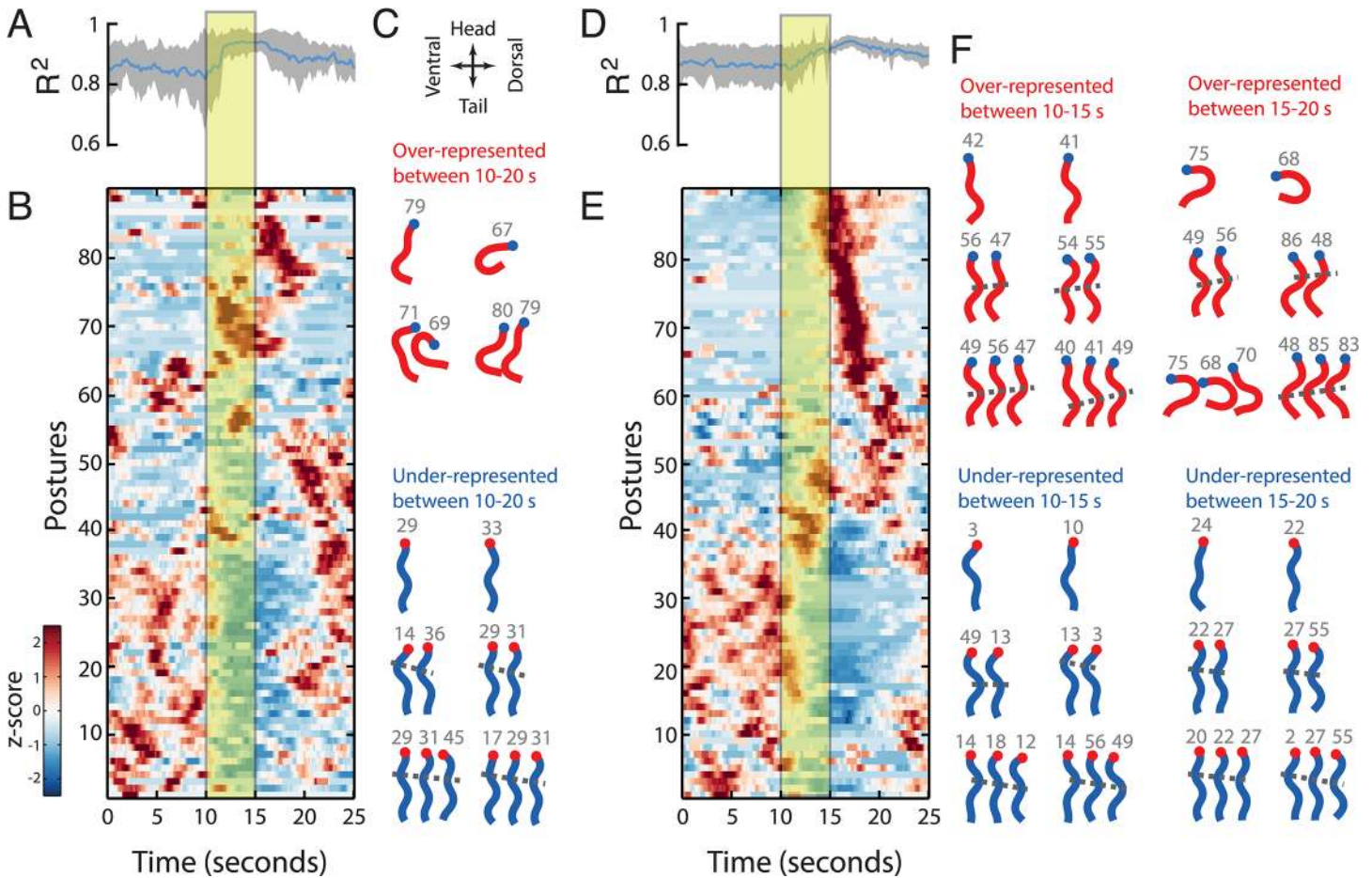
Fig 3. Optogenetic activation of subsets of neurons drives behavioural transitions. The left half of the figure (A-C) shows data from ZX46 worms expressing channelrhodopsin in the cholinergic motor neurons and the right half (D-F) from AQ2026 worms expressing channelrhodopsin in ASH, a neuron that detects aversive stimuli, as well as PVQ and ASI. **(A, D)** R^2 between original body angles and nearest-neighbour template over time. The blue line is the mean and the gray shading indicates the standard deviation ($n = 23$ in A; $n = 34$ in D). Yellow box indicates light stimulation period. **(B, E)** The relative probability of observing each posture over time. Colour indicates whether a posture is over-represented (red) or underrepresented (blue) at each time. Postures with similar temporal profiles are clustered independently in B and E for visualization. **(C, F)** The top 2 n-grams that are most significantly overrepresented (red) or underrepresented (blue) during the indicated time periods for $n = 1$–3 for each strain. Grey numbers show the row number for each posture in the corresponding heatmap. Dots indicate worm head, dorsal and ventral sides are as indicated by the arrows. The gray dashed lines approximately connect body bends from one posture to the next and can be used to distinguish forward from reverse locomotion. Lines with a negative slope indicate body bends propagating backwards and therefore forward locomotion. Lines with a positive slope show reversals.

doi:10.1371/journal.pcbi.1004322.g003

splitting the analysis into three parts: outside stimulus, during stimulus, and 0–5 seconds after stimulus. During stimulation the bigrams and trigrams representing reversals are significantly more likely while forward locomotion is less likely (**Fig 3F**) and in the 5 seconds after stimulus, ventral turns and reversals are more likely while forward locomotion is still supressed.

Worms change their postural syntax in different environments

We next recorded worm locomotion in different environments to see how they adapt their behavioural repertoire in response to variation in sensory stimulation. To add to the data for worms on bacterial food, we also recorded N2 worms crawling on an agar surface with no food, and with no food but in the presence of the chemoattractant benzaldehyde [24]. N2 animals rarely leave a patch of food [25] and spend some time dwelling leading to a confined path (**Fig 4A**). In the absence of food, worms show a clear switch in behaviour, spend more time
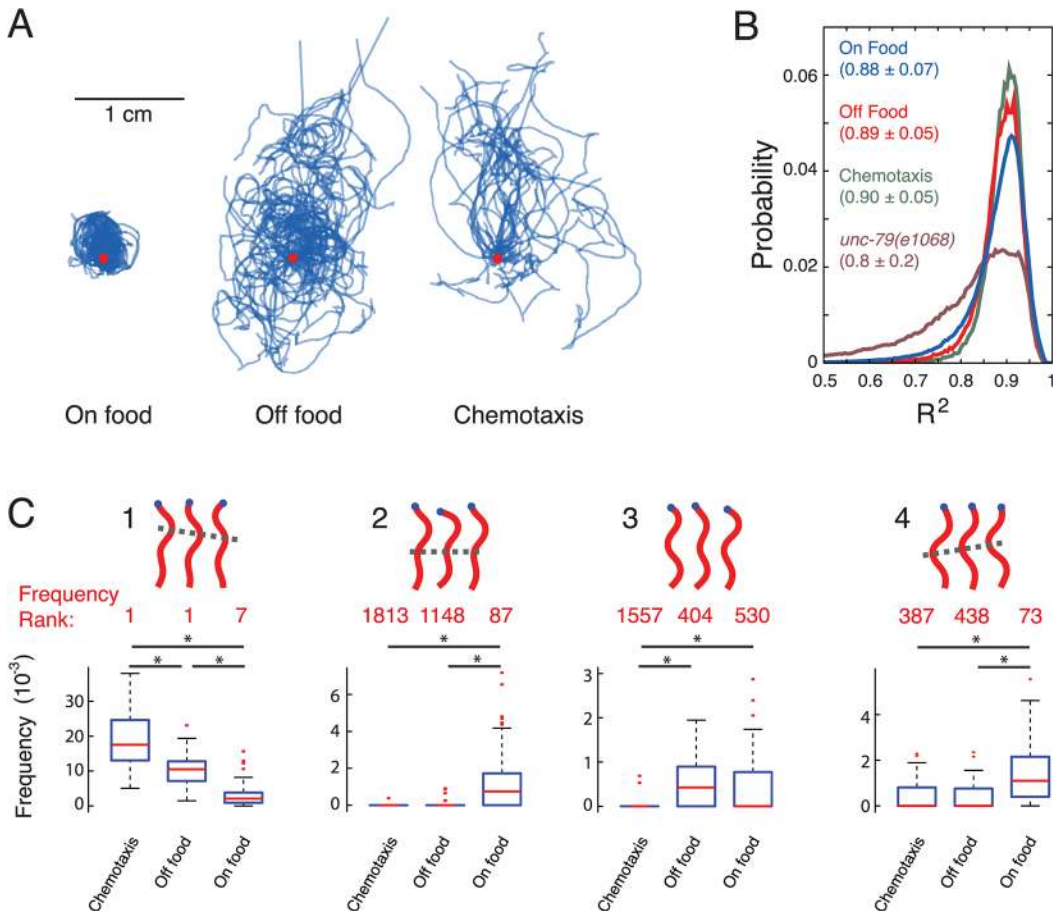
Fig 4. Worm behaviour is compositional across different conditions. (A) Worm tracks in different conditions: on food ($n = 34$), off food ($n = 23$), and off food with an attractant (benzaldehyde) ($n = 25$). Tracks for different worms start at the red dot and are aligned with their long-axis arranged vertically. Worms on food explore the region of the food but rarely leave, while off food they explore a larger area. In the presence of an attractant, worms explore a large area, but in a more directed way. **(B)** A plot shows the distribution of R^2 values for worms in the different conditions. The fit quality is no worse for worms off food or performing chemotaxis even though the postures are derived only for worms on food. For comparison, the distribution of fits is also shown for an uncoordinated mutant *unc-79(e1068)* that is poorly fit using the wild type on food postures. Numbers in the legend are mean ± standard deviation of the R^2 distributions. **(C)** Comparisons between the frequencies of four selected trigrams in the three conditions. The rank of each sequence in the repertoire of worms in each condition is shown in red. * indicates $p < 4.1 \times 10^{-4}$, chosen to control the false discovery rate at 5% across multiple comparisons.

doi:10.1371/journal.pcbi.1004322.g004

roaming and initiate a local search behaviour consisting of reversals and turns. After longer periods of starvation, they show increasingly directed locomotion [15,16]. During chemotaxis off food, their change in behaviour compared to worms off food with no chemo attractant is more subtle; they still spend most of their time moving, but their paths are less random as they will, on average, move up the gradient of attractant [13,14].

We used the same set of 90 templates derived for N2 animals on food to fit the shapes of worms off food and during chemotaxis. The fit quality is no worse in either of the off food conditions, indicating that there is not an increase in outlying postures that are not captured by the template set ([Fig 4B](Fig 4B)). In fact, the poorly fit tail of the distribution is reduced off food compared to on food. The better fit may be due too noisier skeletonization of worms imaged on food and partly because worms off food are less likely to adopt some postures characteristic of pauses that are among the worst fitting postures ([S4 Fig](S4 Fig)). As a positive control, we also measured the distribution of fit qualities for a visibly uncoordinated mutant *unc-79(e1068)* and found a clear increase in outlying postures not captured by the wild type on food template postures.

We also tested for differences in *n*-gram frequencies that worms performed in the different conditions by pooling their sequences and getting the complete set of unique *n*-grams observed across the three conditions, for *n* = 1–3. We counted the number of occurrences of each *n*-gram in each condition and used a rank sum test to find sequences that were used more or less frequently in one of the conditions. We excluded *n*-grams that were observed less than 5 times in total in the two conditions being compared. Note that we did not exclude behaviours that were rare in only one of the conditions. For example, an *n*-gram with a count of 10 in one condition and 0 in another would still be included in the comparison. We corrected for multiple comparisons using an adjusted *p*-value threshold calculated to control the false discovery rate at 5% [26]. Using this procedure, we found that for the on food vs. off food comparison 377 trigrams were significantly more common in one condition, whereas for the subtler off food vs. chemotaxis comparison we found 11. These differences in trigram frequencies are not entirely explained by differences in posture preference, since in shuffled sequences that preserve posture frequency differences, only 2 trigrams were found to be significantly different between the on and off food conditions and none were significantly different between the off food and chemotaxis conditions. Four examples of distinguishing trigrams are shown in **Fig 4C**. The first one is a bout of forward locomotion that is common in all conditions, but progressively less common in worms performing chemotaxis, crawling off food, and crawling on food. This is consistent with expectations for the persistence of locomotion in these conditions. The second trigram starts with the same posture as the first, but it appears in a different context. Instead of appearing during forward locomotion, the posture appears during a pause with a head swing. This change in context makes trigram 2 much less common in the off food conditions even though trigrams 1 and 2 share a posture in common. This implies that behavioural adaptation is at least partly characterized by shifts in postural syntax. Trigram 3 shows that behaviours that are rare in all conditions can still show significant differences in frequency and trigram 4 shows an example of a reversal which is more common on food than in either of the off food conditions, consistent with worms showing more persistent forward locomotion when searching off food.

Natural variation in behaviour is captured by a subset of posture *n*-grams

In addition to environmental variation, we also quantified the behavioural variation that arises over short evolutionary times between different populations of the same species. We tracked 17 strains of *C. elegans* isolated from different regions of the world and quantified their behaviour using the discrete representation. We found that the template postures derived from N2 worms on food could also fit the wild isolate data and that the heavy-tailed frequency distribution of trigrams is observed in all strains (**Fig 5A**). The distribution is skewed with the top 1% of trigrams being used 38 ± 4% of the time during crawling (mean ± standard deviation, *n* = 17 strains).

We performed rank sum tests between all strain pairs to find trigrams with significant frequency differences using a *p*-value threshold adjusted to control the false discovery rate at 5% across all comparisons. There are trigrams with significant differences across the frequency spectrum; in fact most behaviours with a significant frequency difference are rare in at least one of the strains where the behaviours were found as a hit (**Fig 5B**). Not all of the hits counted in the histogram in Fig 5B correspond to unique trigrams because some trigrams had significantly different frequencies in more than one strain (**Fig 5C**). The number of *unique* trigrams that are different in at least one strain is 1258, which is approximately 5% of the 2.3×10^4 distinct trigrams observed in the full wild isolate data set. The relative frequencies for these 1258 trigrams are shown in a heatmap in **Fig 5D**. Trigrams 25 and 26 show dorsal turns. 27 and 28
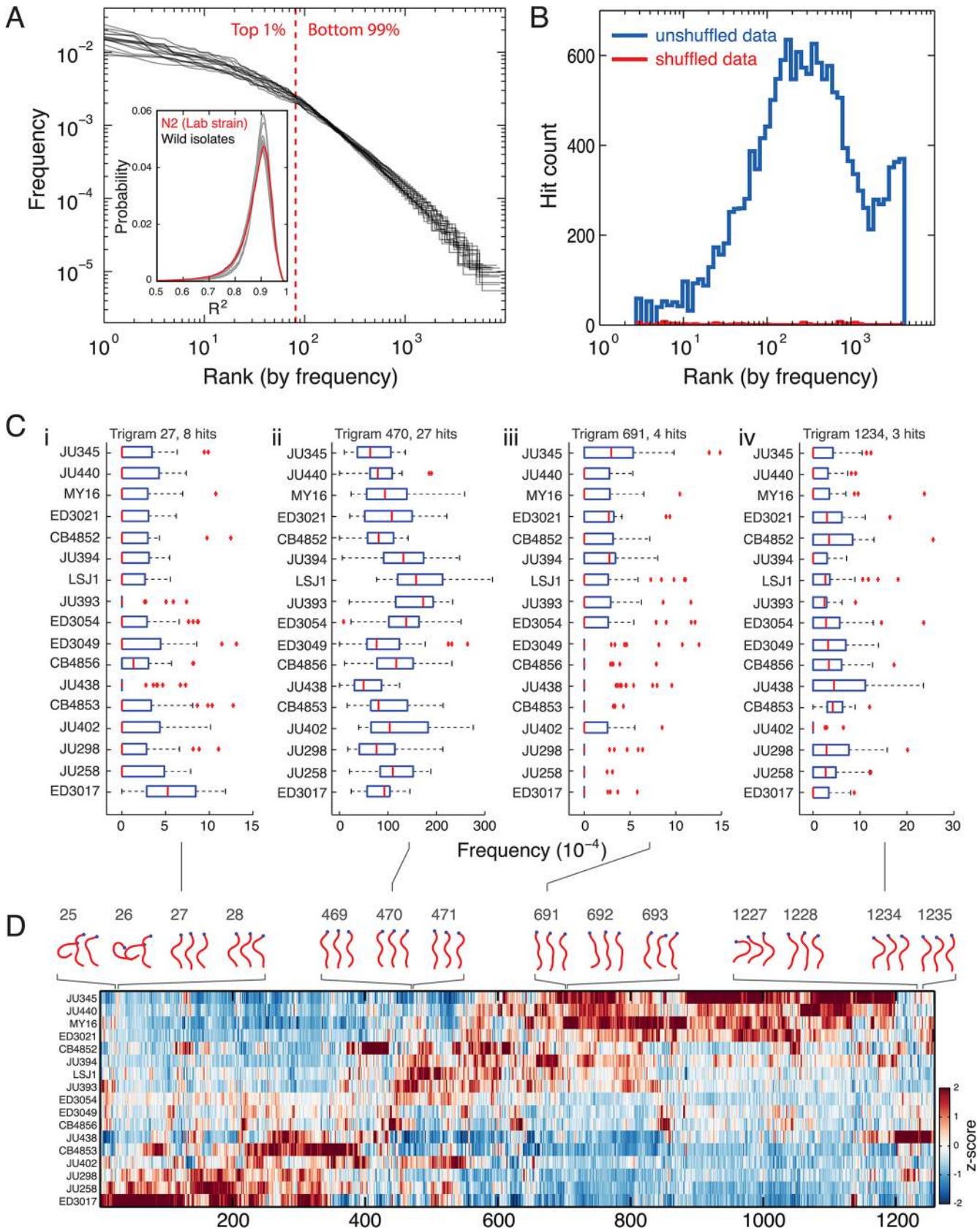
Fig 5. Form of trigram frequency distribution is conserved across wild isolates despite differences in preference for particular sequences. (A) Zipf plots for 17 wild isolates of *C. elegans* all show a heavy tailed distribution similar to N2 worms. The red dashed line shows the rank dividing the top 1% of trigrams in the repertoire from the remaining 99%. The inset shows the distribution of R^2 values for fits between 90 N2-derived templates and the original postures for the 17 wild isolates as well as for N2 itself for comparison. The overall mean of the distributions is 0.86 with standard deviation 0.08. **(B)** Histogram of hit counts (not hit density) versus frequency on a log scale. Hits are trigrams with significantly different frequencies in at least one pairwise rank sum test comparing strains. Note that the bin widths increase to be constant on a log-scale. Blue line is the original data and the red line is for trigrams derived from the same data after random shuffling of the sequences. **(C)** Box plots of trigram frequencies for 4 trigrams showing hits in several strains. The trigram number indicates its column in the heatmap in D. The *p*-value threshold was set at 1.1×10^{-4} to control the false discovery rate at 5% across comparisons. **(D)** A heatmap showing the relative frequency of all of the trigrams with significantly different frequencies in at least one strain comparison. Strains and trigrams have been hierarchically clustered. Sample trigrams are plotted above the heatmap with their columns indicated by the gray numbers.

are also bent more dorsally than ventrally, but 27 represents forward locomotion while 28 represents a reversal. 469–471 are all bouts of forward locomotion, 691–693 are pauses with relatively straight postures. 1227 is a ventral turn and 1234 and 1235 are reversals.

Discussion

We have developed a simple representation of behaviour as a series of discrete postures to systematically enumerate the behavioural repertoire of *C. elegans* during locomotion. The concept of a complete description of behaviour was motivated in part by similar efforts devoted to other aspects of worm biology, most notably the genome, cell lineage, and connectome. However, the repertoire of behavioural sequences cannot be said to be complete in a similar way. Although we have a complete record of the observed sequences, accumulation curves suggest that more recordings would continue to reveal novel sequences, even when considering behaviours with durations of about a second in an animal with a simple morphology in a relatively uniform environment. A notable feature of the observed repertoire is the simultaneous presence of repetition (the heavily-used top 1%) and variety (the large set of relatively rarely used sequences). This feature may result from two demands on an animal's behavioural repertoire: locomotion should be efficient in a given environment but must also be flexible to deal with unexpected conditions or to respond to new sensory information. The repertoires of 17 wild isolates of *C. elegans* have a similar rank-frequency distribution and we hypothesize that such heavy-tailed distributions will generalize beyond nematodes and will in fact be a widespread characteristic of animal behavioural repertoires.

Some of the rarely observed sequences in the repertoire are surely due to measurement noise, but since the frequency of some rare behaviours is modulated in response to sensory stimuli (for example behaviour 3 in Fig 4C), measurement noise is not the only factor. It should also be emphasized that although the repertoire is large in this representation, that does not preclude the existence of an alternative representation that will reveal an underlying simplicity. Indeed, seemingly simple nonlinear systems are well known to be able to produce complex dynamics [27] and the search for such models in the context of animal behaviour remains an exciting challenge [28–30]. The observation that only 5% of the trigram repertoire is modulated by natural variation between wild isolates (Fig 5D) may be a reflection of a small number of behavioural modules that can be adjusted genetically within the constraints of natural selection [31,32].

New imaging technology has made it possible to record neural activity from the majority of *C. elegans* neurons simultaneously in restrained animals [33] and during locomotion [34]. The usefulness of this data will depend to some extent on how the corresponding behavioural data is represented and quantified. This is also the case for optogenetically induced behavioural changes as reported here and many times previously in *C. elegans* (e.g. [35–39]). Recent work in *Drosophila* larvae suggests that an unsupervised detection of locomotion patterns can be informative [40]. The discrete approach that we report here is also unsupervised in the sense

that repeated patterns of behaviour are detected based on locomotion patterns observed within strains without initially using information about different classes (environmental conditions or strains). A conceptually similar approach has proven to be useful for quantifying drug-induced behaviour in mice [41]. It is likely that no behavioural representation will be optimal for all studies, but that different representations will provide complementary insight. The advantages of trigram counting are that it is relatively unbiased, can reliably connect behaviours performed at different speeds, provides a complete record of all of the observed behaviours (up to the resolution of the template postures), and that it can be summarized intuitively as frequencies of short behaviours that can be directly visualized. For some applications the variance that is missed by the discrete templates may be critical. One option would be to increase the number of templates, but it may also be that a different representation would be called for. Another limitation of the discrete representation as presented here is that we have not used the information contained in the duration that worms spend in each posture. The time that worms spend performing each behavioural sequence is relatively straightforward to incorporate and would increase the power of the representation to resolve differences in dynamics between strains.

In the wild, animals must adapt to a variety of environments that can change unpredictably. One strategy for dealing with this complexity is to use a finite set of behavioural primitives that can be composed in different sequences to generate adaptive responses [42]. This is precisely the kind of variation that we observed for worms changing their behaviour in response to new environments (Fig 4). Using the same set of postures, they change their behaviour at the level of postural syntax. Compositionality in behaviour also suggests a mechanism for generating complex behaviours by combining primitive actions. Hierarchical composition has long been recognized as a principle of animal behaviour [43,44] and has also been proposed as an efficient representation for learning in robotics [42]. The possible connection between motor learning and compositionality raises the interesting question of how the repertoire observed in adults is acquired and modified during development.

Materials and Methods

Worm tracking

Locomotion data was recorded for single worms as previously described [19]. For on food tracking, single worms from each strain were picked to the centre of a 25 mm agar plate with a spot of *E. coli* OP50 at the centre, allowed to habituate to the new plate for 30 minutes, and then recorded for 15 minutes. The worm side (whether it was on its left or right side) was manually annotated using a stereomicroscope before transferring plates to the tracking microscope. In addition to the previously recorded N2 and DR96 *unc-76(e911)* data, we used the following wild isolate strains: CB4852, CB4853, CB4856, ED3017, ED3021, ED3049, ED3054, JU258, JU298, JU345, JU393, JU394, JU402, JU438, JU440, LSJ1, and MY16. For experiments off food, worms were briefly transferred to plates without food and then picked to the centre of a 55 mm agar plate and recorded immediately. Recording was stopped once worms reached the edge of the plate. To maximize the time of recording, worms in the chemotaxis assays were picked to a location near the edge of a 55 mm agar plate opposite the location of the attractant spot. 1 μl of the attractant, benzaldehyde (diluted 1:100 in EtOH), was added to the plate 10 mm from the edge.

Optogenetics

The strains used for optogenetics were ZX46 *zxIs6 [punc-17::ChR2(H134R)::yfp]* V [21] and AQ2026 *ljIs105 [sra-6::ChR2::yfp, unc-122::gfp]* [22]. Worms were transferred to retinal plates at the L4 stage, left at 20°C, and assayed either 24 hours later as adults (AQ2026) or in the next

generation (ZX46). Retinal plates were made by seeding NGM plates with OP50 mixed with all-trans retinal (100mM in ethanol) in a ratio of 1000:4. For channel rhodopsin activation, worms were illuminated for 5 seconds with 1 mW/mm^2 blue light (440–460 nm) using LEDs (Luxeon III LXHL-PR09 from Lumileds, CA) controlled by a Mindstorms LEGO Intelligent NXT Brick. Each worm was stimulated three times, with an interstimulus interval of 30 seconds.

Posture discretization

The skeleton of worms in each frame was found after segmentation by tracing the midline between the two points of highest curvature (the head and tail) of the largest connected component in the image (the worm body) and down sampling to 49 equally spaced points [19]. The skeletons were then converted to a position and orientation independent representation by measuring the angle between each of the 49 skeleton points and subtracting the mean angle [17]. Missing skeleton angles were linearly interpolated. Missing data resulted from camera dropped frames, stage motion, or failure to skeletonize because of poor segmentation or coiled worms. On average 13 ± 9% (mean ± standard deviation) of frames cannot be segmented with each of the three causes contributing approximately equally to missing data. 5000 skeletons were randomly selected from each of 20 randomly selected N2 worms recorded crawling on food. Worms were flipped so that all worms appear to be on their right side when viewed on the page. This means that dorsal and ventral turns are consistent across individuals during analysis. This set of skeletons was then clustered using k-means clustering to identify a set of representative template postures. k was varied from 5 to 500 to determine the value for k after which increasing k only marginally improves the quality of the fit. k was chosen as 90 for all subsequent analysis.

For each worm video, we down sampled the data from 30 frames per second to 6 frames per second and then found the best matching template in each frame of the down sampled data. The best match was determined in each frame by finding the template posture with the minimum squared difference to the skeleton in that frame. Locomotion was then represented as a series of state labels between 1 and 90. Repeated states were collapsed to a single state (e.g. {1, 1, 2, 2, 2, 5} would become {1, 2, 5}).

Posture probability analysis

For the optogenetics data, the frequency of each posture over time is averaged over three trials per worm and further averaged over the individuals in each experiment (n = 26 for ZX46 and n = 34 for AQ2026). Each row was smoothed with a moving average filter with a width of 1.3 seconds. The smoothed frequencies of each posture over time were then normalized by subtracting the mean and dividing by the standard deviation of each row to show relative up- and down-regulation over time. To aid visualization, the rows were then clustered vertically so that rows with similar temporal profiles were near each other. The clustering was agglomerative hierarchical clustering using Euclidean distance and average linkage. The clustering was performed independently for the data in Fig 3B and 3E so the rows in each plot do not correspond to the same postures.

n-gram model

The *n*-gram language model analyses were implemented in C and Python using the OpenFST [45] and OpenGRM *n*-gram [20] libraries. Postural sequences from N2 worms on food were compiled into finite-state archives (FAR) and standard *n*-gram count models were built for n = 1–5 and normalized into probabilistic models without additional smoothing (see e.g. [46]).

For the simulation studies and for each *n*, 1000 random sequences were generated from the normalized finite-state models. Beginning at the start state an outgoing arc is chosen at random proportionally to the log-probability associated with the arc, the postural state is emitted and the process continues until an end state transition is reached.

n-gram statistics

The total set of unique *n*-grams was determined for each strain or condition (for the on food, off food, and chemotaxis set, the total set contained 14 209 trigrams and for the 17 wild isolates it contained 22 638 trigrams). Counts of unique *n*-grams in each individual were then used to compare strains or conditions to each other. For each comparison, any *n*-gram that was observed a total of 5 times or less in both groups being compared was ignored. For the remaining *n*-grams, counts were compared using a rank sum test and *p*-value thresholds were chosen using the Benjamini-Yekutieli procedure to control the false discovery rate at 5% [26].

Supporting Information

S1 Fig. The complete set of 90 posture templates derived for N2 animals crawling on food using *k*-means clustering. The postures are ordered from most frequent (top left) counting down each column in turn to the least frequent (bottom right).
(EPS)

S2 Fig. Frequency distribution of *n*-grams compared to exponential distributions and Weibull distributions with parameters set by maximum likelihood estimation. Blue curves are each calculated for 400 worms resampled with replacement from the total N2 data set. Each of the 10 curves is fit separately to an exponential (red curves) and a Weibull (black curves) distribution. A-B show results for *n* = 2–5.
(EPS)

S3 Fig. Accumulation curves (left column) and Zipf rank-frequency plots (right column). Blue lines show data from different randomly selected sets of N2 worms crawling on food (100 worms selected for each curve). Black lines show the corresponding results for data generated from *n*-gram models with different values of *n* shown in A-C.
(EPS)

S4 Fig. (A) R^2 plotted for each posture individually (mean ± standard deviation) for data collected on food, off food, and during chemotaxis. The mean fit quality is consistently slightly lower on food than in the two off food conditions, consistent with slightly noisier segmentation when worms are on a patch of food. Furthermore, the two postures with the lowest R^2 values (34 and 71) are used less frequently in the off food conditions than when worms are on food (B).
(EPS)

Acknowledgments

We thank Mario de Bono and Kate Weber for providing us with the wild isolates used in this study and Alexander Gottschalk for providing ZX46 worms.

Author Contributions

Conceived and designed the experiments: RB WRS AEXB. Performed the experiments: RB LJG. Analyzed the data: RFS AEXB. Wrote the paper: RFS RB WRS AEXB.

References

1. Berman GJ, Choi DM, Bialek W, Shaevitz JW. Mapping the stereotyped behaviour of freely moving fruit flies. Journal of The Royal Society Interface. 2014; 11: 20140672–20140672. doi: 10.1098/rsif.2014.0672

2. Lehner PN. Handbook of ethological methods. 2nd ed. Cambridge ; New York: Cambridge University Press; 1996.

3. Burghardt GM, Bartmess-LeVasseur JN, Browning SA, Morrison KE, Stec CL, Zachau CE, et al. Perspectives—Minimizing Observer Bias in Behavioral Studies: A Review and Recommendations: Minimizing Observer Bias in Behavioral Research. Ethology. 2012; 118: 511–517. doi: 10.1111/j.1439-0310.2012.02040.x

4. Gomez-Marin A, Paton JJ, Kampff AR, Costa RM, Mainen ZF. Big behavioral data: psychology, ethology and the foundations of neuroscience. Nature Neuroscience. 2014; 17: 1455–1462. doi: 10.1038/nn.3812 PMID: 25349912

5. Anderson DJ, Perona P. Toward a Science of Computational Ethology. Neuron. 2014; 84: 18–31. doi: 10.1016/j.neuron.2014.09.005 PMID: 25277452

6. Dell AI, Bender JA, Branson K, Couzin ID, de Polavieja GG, Noldus LPJJ, et al. Automated image-based tracking and its application in ecology. Trends in Ecology & Evolution. 2014; 29: 417–428. doi: 10.1016/j.tree.2014.05.004

7. The C. elegans Sequencing Consortium. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. Science. 1998; 282: 2012–2018. doi: 10.1126/science.282.5396.2012 PMID: 9851916

8. Sulston J. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Developmental Biology. 1977; 56: 110–156. doi: 10.1016/0012-1606(77)90158-0 PMID: 838129

9. White JG, Southgate E, Thomson JN, Brenner S. The Structure of the Nervous System of the Nematode Caenorhabditis elegans. Philosophical Transactions of the Royal Society B: Biological Sciences. 1986; 314: 1–340. doi: 10.1098/rstb.1986.0056

10. Gjorgjieva J, Biron D, Haspel G. Neurobiology of Caenorhabditis elegans Locomotion: Where Do We Stand? BioScience. 2014; doi: 10.1093/biosci/biu058

11. Wen Q, Po MD, Hulme E, Chen S, Liu X, Kwok SW, et al. Proprioceptive Coupling within Motor Neurons Drives C. elegans Forward Locomotion. Neuron. 2012; 76: 750–761. doi: 10.1016/j.neuron.2012.08.039 PMID: 23177960

12. Faumont S. The Awake Behaving Worm: Simultaneous Imaging of Neuronal Activity and Behavior in Intact Animals at Millimeter Scale. Journal of Neurophysiology. 2006; 95: 1976–1981. doi: 10.1152/jn.01050.2005 PMID: 16319197

13. Pierce-Shimomura JT, Morse TM, Lockery SR. The fundamental role of pirouettes in Caenorhabditis elegans chemotaxis. J Neurosci. 1999; 19: 9557–9569. PMID: 10531458

14. Iino Y, Yoshida K. Parallel Use of Two Behavioral Mechanisms for Chemotaxis in Caenorhabditis elegans. Journal of Neuroscience. 2009; 29: 5370–5380. doi: 10.1523/JNEUROSCI.3633-08.2009 PMID: 19403805

15. Gray JM, Hill JJ, Bargmann CI. A circuit for navigation in Caenorhabditis elegans. Proc Natl Acad Sci USA. 2005; 102: 3184–3191. doi: 10.1073/pnas.0409009101 PMID: 15689400

16. Salvador LCM, Bartumeus F, Levin SA, Ryu WS. Mechanistic analysis of the search behaviour of Caenorhabditis elegans. Journal of The Royal Society Interface. 2014; 11: 20131092–20131092. doi: 10.1098/rsif.2013.1092

17. Stephens GJ, Johnson-Kerner B, Bialek W, Ryu WS. Dimensionality and Dynamics in the Behavior of C. elegans. PLoS Computational Biology. 2008; 4: e1000028. doi: 10.1371/journal.pcbi.1000028 PMID: 18389066

18. Brown AEX, Yemini EI, Grundy LJ, Jucikas T, Schafer WR. A dictionary of behavioral motifs reveals clusters of genes affecting Caenorhabditis elegans locomotion. Proc Natl Acad Sci USA. 2013; 110: 791–796. doi: 10.1073/pnas.1211447110 PMID: 23267063

19. Yemini E, Jucikas T, Grundy LJ, Brown AEX, Schafer WR. A database of Caenorhabditis elegans behavioral phenotypes. Nature Methods. 2013; 10: 877–879. doi: 10.1038/nmeth.2560 PMID: 23852451

20. Roark B, Sproat R, Allauzen C, Riley M, Sorensen J, Tai T. The OpenGrm open-source finitestate grammar software libraries. In Proceedings of the Association for Computational Linguistics (System Demonstrations). 2012. pp. 61–66.

21. Liewald JF, Brauner M, Stephens GJ, Bouhours M, Schultheis C, Zhen M, et al. Optogenetic analysis of synaptic function. Nature Methods. 2008; 5: 895–902. doi: 10.1038/nmeth.1252 PMID: 18794862

22. Ezcurra M, Tanizawa Y, Swoboda P, Schafer WR. Food sensitizes *C. elegans* avoidance behaviours through acute dopamine signalling: Food sensitizes *C. elegans* avoidance behaviours. The EMBO Journal. 2011; 30: 1110–1122. doi: 10.1038/emboj.2011.22 PMID: 21304491

23. Guo ZV, Hart AC, Ramanathan S. Optical interrogation of neural circuits in Caenorhabditis elegans. Nature Methods. 2009; 6: 891–896. doi: 10.1038/nmeth.1397 PMID: 19898486

24. Bargmann C. Chemosensation in C. elegans. WormBook. 2006; doi: 10.1895/wormbook.1.123.1

25. Avery L, You Y-J. C. elegans feeding. WormBook. 2012; doi: 10.1895/wormbook.1.150.1

26. Benjamini Y, Yekutieli D. The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics. 2001; 29: 1165–1188. doi: 10.2307/2674075

27. Strogatz SH. Nonlinear dynamics and chaos with applications to physics, biology, chemistry, and engineering. Cambridge, MA: Westview Press; 2000.

28. Cole B. Fractal time in animal behaviour: the movement activity of Drosophila. Animal Behaviour. 1995; 50: 1317–1324. doi: 10.1016/0003-3472(95)80047-6

29. Li J-S, Huston JP. Non-linear dynamics of operant behavior: a new approach via the extended return map. Rev Neurosci. 2002; 13: 31–57. PMID: 12013025

30. Maye A, Hsieh C, Sugihara G, Brembs B. Order in Spontaneous Behavior. PLoS ONE. 2007; 2: e443. doi: 10.1371/journal.pone.0000443 PMID: 17505542

31. Avery L. Caenorhabditis elegans behavioral genetics: where are the knobs? BMC Biology. 2010; 8: 69. doi: 10.1186/1741-7007-8-69 PMID: 20504291

32. Weber JN, Peterson BK, Hoekstra HE. Discrete genetic modules are responsible for complex burrow evolution in Peromyscus mice. Nature. 2013; 493: 402–405. doi: 10.1038/nature11816 PMID: 23325221

33. Schrödel T, Prevedel R, Aumayr K, Zimmer M, Vaziri A. Brain-wide 3D imaging of neuronal activity in Caenorhabditis elegans with sculpted light. Nature Methods. 2013; 10: 1013–1020. doi: 10.1038/nmeth.2637 PMID: 24013820

34. Nguyen J, Shipley F, Linder A, Plummer G, Shaevitz J, Leifer A. Whole-brain calcium imaging with cellular resolution in freely behaving C. elegans. Preprint. Available: arXiv:150103463. Accessed 6 April 2015.

35. Nagel G, Brauner M, Liewald JF, Adeishvili N, Bamberg E, Gottschalk A. Light Activation of Channelrhodopsin-2 in Excitable Cells of Caenorhabditis elegans Triggers Rapid Behavioral Responses. Current Biology. 2005; 15: 2279–2284. doi: 10.1016/j.cub.2005.11.032 PMID: 16360690

36. Stirman JN, Crane MM, Husson SJ, Wabnig S, Schultheis C, Gottschalk A, et al. Real-time multimodal optical control of neurons and muscles in freely behaving Caenorhabditis elegans. Nature Methods. 2011; 8: 153–158. doi: 10.1038/nmeth.1555 PMID: 21240278

37. Leifer AM, Fang-Yen C, Gershow M, Alkema MJ, Samuel ADT. Optogenetic manipulation of neural activity in freely moving Caenorhabditis elegans. Nature Methods. 2011; 8: 147–152. doi: 10.1038/nmeth.1554 PMID: 21240279

38. Shipley FB, Clark CM, Alkema MJ, Leifer AM. Simultaneous optogenetic manipulation and calcium imaging in freely moving C. elegans. 2013;000943. Preprint. Available: http://biorxiv.org/lookup/doi/10.1101/000943. Accessed 6 April 2015.

39. Kocabas A, Shen C-H, Guo ZV, Ramanathan S. Controlling interneuron activity in Caenorhabditis elegans to evoke chemotactic behaviour. Nature. 2012; 490: 273–277. doi: 10.1038/nature11431 PMID: 23000898

40. Vogelstein JT, Park Y, Ohyama T, Kerr RA, Truman JW, Priebe CE, et al. Discovery of brainwide neural-behavioral maps via multiscale unsupervised structure learning. Science. 2014; 344: 386–392. doi: 10.1126/science.1250298 PMID: 24674869

41. Kafkafi N, Mayo CL, Elmer GI. Mining mouse behavior for patterns predicting psychiatric drug classification. Psychopharmacology. 2014; 231: 231–242. doi: 10.1007/s00213-013-3230-6 PMID: 23958942

42. Flash T, Hochner B. Motor primitives in vertebrates and invertebrates. Current Opinion in Neurobiology. 2005; 15: 660–666. doi: 10.1016/j.conb.2005.10.011 PMID: 16275056

43. Dawkins R. Hierarchical organisation: a candidate principle for ethology. Growing Points in Ethology. Cambridge: Cambridge University Press; 1976. pp. 7–54.

44. Lashley KS. The problem of serial order in behavior. In: Jeffress LA, editor. Cerebral Mechanisms in Behavior. New York: John Wiley Press; 1951.

45. Allauzen C, Riley M, Schalkwyk J, Skut W, Mohri M. OpenFst: a general and efficient weighted finite-state transducer library. Proceedings of the 12th international conference on Implementation and application of automata. Prague, Czech Republic: Springer-Verlag; 2007. pp. 11–23.

46. Manning CD. Foundations of statistical natural language processing. Cambridge, Mass: MIT Press; 1999.