Google 'EarthWorm'

William A Mohler

A three-dimensional digital atlas allows cell-by-cell navigation of *Caenorhabditis elegans*.

'Where in the worm are we?' is a phrase commonly uttered when viewing high-magnification micrographs of *Caenorhabditis elegans*. This tiny roundworm—modern molecular biology's smallest and best understood metazoan animal—is still big and complex enough to become easily lost in. Between individuals of this species, however, the number and fates of cells are essentially invariant, far more consistent than repeated aerial photos of the same square mile of a road system. A paper in this issue of *Nature Methods* makes a first leap toward capturing the worm anatomy in a comprehensive digital atlas of nuclear coordinates[1].

Eugene Myers and colleagues present a new quantitative three-dimensional (3D) compilation of decades-old information on cell positions in the newly hatched *C. elegans* larva, as well as new techniques to resurvey the landscape quickly and accurately[1]. The result is a 'you are here' experience that rivals the carefree confidence of a motorist guided by a global positioning system (**Fig. 1**).

In a progression of papers, these authors have previously described critical methods that made the new work possible. First came a scheme for mathematically relaxing the curls of worms fixed in mid-wiggle: calculating for each uniquely nonconformist larva a realistic straightened geometry[2]. This allows different individual worms to be conceptually superimposed *in silico*, both to aid manual annotation of cells and as a framework for automated cell identification and statistical analysis of cell positions. Next came a computer interface to help a human expert accurately assign cell names, lineage ancestries and fates to nuclei in confocal microscopy image stacks slicing through a worm[3].

The current paper by Long *et al.*[1] is remarkable in that it brings the arrangement of cells in the nematode into the grasp of both experimentalists and informaticians. The result of this work is a set of three distinct resources that can be used by researchers at large to much better understand the makings of tiny but intricate creatures. The digital atlas, itself a visually bland supplementary table, springs upon the mind's eye when appreciated through an interactive 3D visualization interface or via graphs of sequence and adjacency in cell organization, statistically derived for a quorum of worms. This catalog of nuclear positions is likely the resource that will be accessed most often by the several thousand 'worm people' studying *Caenorhabditis*.

A second crucial biological resource—easily overlooked under the glare of high technology—is the authors' codification of rules for manually assigning cell identities throughout the body plan. This is a concise and valuable compilation of information from diverse sources into a verbal guide to pinpointing the right suspect in any visual study of the young larva. It draws on data in some of the most venerated of *C. elegans* papers[4–6] and on modern renderings in the ever-improving online collection of nematode microanatomy[7]. In this new paper[1], repeated and accurate manual annotation of a handful of specimens provides the authors the essential reference data for the atlas and also guides their training of an automated image-to-information pipeline that can replicate the laborious identification and annotation process quite faithfully with little human attention.

The third resource is the logic and code base for this ingenious automated approach to mapping any newly imaged larva to the atlas. Long *et al.*[1] detail a disassembly line of image-processing algorithms that take the raw 3D image of a worm with fluorescently labeled DNA and

Figure 1 | One leg of a fanciful journey through the midsection of the *C. elegans* larva, guided by the digital atlas of Long *et al.*[1]. WPS, worm positioning system. INT5R and D4A are two of the 558 cells in the worm, with their specific lineage ancestries given in parentheses.

William A. Mohler is at the University of Connecticut Health Center, Department of Genetic and Developmental Biology and Center for Cell Analysis and Modeling, Farmington, Connecticut, USA.
e-mail: wmohler@neuron.uchc.edu

carefully break it down to its constituent cells: pulling it straight, filling holes, finding edges and then prying apart stubborn 'clumps' of clustered nuclei (while pasting together single nuclei accidentally shattered in the process). The resulting 'segmented worm' model (digital jargon; it is still a nematode, not an annelid) is then spun and stretched in space to register its cells with those of the reference atlas. Finally, the cells can be annotated in either fully manual (very accurate) or fully automatic (very fast) modes, or by an iterative course of manual and automatic processes that reconciles speed with fidelity. This workflow should serve as an important model in the blossoming science of tidying up messy biological tissues into orderly lists and maps.

Yet the task of identifying all 558 cells of the hatchling worm in confocal images is still beyond the grasp of human and machine. Long *et al.*[1] settle on 357 cells that are consistent in statistical comparisons between individual worms. Most of the 201 'lost' cells are in the nerve ring and other ganglia. So although 'the mind of a worm' had been read ultrastructurally down to its synapses 23 years ago[8], the *C. elegans* brain is still inscrutable at the level of the confocal microscope or of higher-throughput analysis.

Hauntingly, 350 cells mark the current barrier for fidelity of cell identification by automatic lineage tracing in the *C. elegans* embryo[9,10]. Coincidence? Or is this as far as segmentation algorithms working on images at conventional optical resolution can go? Super-resolution light microscopy could improve the resolving power of nuclear position mapping, although sub-diffraction-limit imaging of fluorescence in complex 3D specimens is a substantial challenge. At least one approach, 4Pi microscopy, can give ~100-nm resolution even through the full depth of a nematode larva (W.A.M. and Joerg Bewersdorf, unpublished observations).

Where to now with this new digital atlas atop the dashboard? The automated mapping method is already being used to record expression of fluorescent reporter genes in each of the young worm's cells: the beginnings of a whole new type of anatomic-genomic data resource. Another obvious use will be the discrimination of subtle anatomical disturbances caused by mutations that have been heretofore classified as having no phenotype. Yet another eventual possibility is to use the nuclear-position atlas to coax larvae into revealing their precise biological age. Although it increases tremendously in size during development, the worm adds few (if any) new cells to most of its somatic tissues. Rather, tissues grow through cellular hypertrophy, such that the shifting distances between nuclei could serve as an internal clock of postembryonic development. In this way, mapping of postembryonic gene expression could approach the cell-by-cell, minute-by-minute resolution that is already produced by autolineaging in embryos.

Future itinerants of the worm anatomy and life cycle might never lose their sense of time or direction! So, too, for other species? In anatomy and genomics, where the worm leads, others soon follow.

1. Long, F., Peng, H., Liu, X., Kim, S.K. & Myers, E. *Nat. Methods* **6**, 667–672 (2009).
2. Peng, H., Long, F., Liu, X., Kim, S.K. & Myers, E.W. *Bioinformatics* **24**, 234–242 (2008).
3. Peng, H., Long, F. & Myers, E.W. *Bioinformatics* **25**, 695–697 (2009).
4. Albertson, D.G. & Thomson, J.N. *Phil. Trans. R. Soc. Lond. B* **275**, 299–325 (1976).
5. Sulston, J.E. & Horvitz, H.R. *Dev. Biol.* **56**, 110–156 (1977).
6. Sulston, J.E., Schierenberg, E., White, J.G. & Thomson, J.N. *Dev. Biol.* **100**, 64–119 (1983).
7. Altun, Z.F. & Hall, D.H. http://www.wormatlas.org/ (2002–2009).
8. White, J.G., Southgate, E., Thomson, J.N. & Brenner, S. *Phil. Trans. R. Soc. Lond. B* **314**, 1–340 (1986).
9. Bao, Z. *et al. Proc. Natl. Acad. Sci. USA* **103**, 2707–2712 (2006).
10. Murray, J.I. *et al. Nat. Methods* **5**, 703–709 (2008).

The 'rare biosphere': a reality check

Jens Reeder & Rob Knight

Methods for error correction and classification of metagenomic datasets suggest that the rare biosphere is not as large as previously assumed.

High-throughput sequencing is revolutionizing our view of microbial communities, both in terms of which species are present and which genes are present in a particular community. In this issue, two papers report breakthroughs targeted at the first of these two questions[1,2]. Finding out 'who' is present in a given community is likely to be crucial for understanding what that community does, yet differences in methods can yield very different estimates from the same set of samples[3]. One consistent theme from high-throughput sequencing studies, however, is that the vast majority of sequences appear to come from species that have not previously been characterized, whether the environments sequenced are marine[4], soil[5], the human hand[6] or the human gut[3]. If this 'rare biosphere' hypothesis—that vast amounts of uncharacterized diversity lurk in every environment—is correct, the task of building reference databases becomes much more challenging.

Quince *et al.*[2] describe a new method of reducing noise in high-throughput pyrosequencing reads derived from the small subunit (16S) rRNA gene; this gene is widely used as a phylogenetic marker because it is present in all organisms and has slow- and fast-evolving regions that allow reliable reconstruction of phylogeny and make it particularly suited for community analysis[7]. The problem is that most environments contain many more microorganisms than have actually been observed even with the depth of sequencing that pyrosequencing allows. Techniques are available for extrapolating the total number of species from a finite sample, but they rely heavily on the number of species observed only once and are thus sensitive to any unique reads that are produced as experimental artifacts.

Quince and colleagues' analysis of an artificial community in which all the species are known, suggests that because of sequencing errors these estimates may be at least an order of magnitude too high (for example, 350 species were inferred using a standard pipeline based on the programs MUSCLE and DOTUR, when the true population diversity was only 40 species). The main problem (**Fig. 1**) is that even when very few reads overall contain multiple errors (sequencing errors or chimeras), all of these errors give rise to unique

Jens Reeder and Rob Knight are in the Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado, USA.
e-mail: rob.knight@colorado.edu