A Proposal for Comparative Genomics in Support the modENCODE Project

Organizers:
Fabio Piano (NYU: fp1@nyu.edu) and Peter Cherbas (Indiana University: cherbas@indiana.edu)

Contributors:
Karin Kiontke (NYU), Paul Sternberg (CalTech), Robert Waterston (UW), David Fitch (NYU), Asher Cutter (U. Toronto), Marie-Anne Felix (IJM, Paris), Manolis Kellis (MIT), Michael Eisen (LBNL, UC Berkeley), Artyom Kopp (UC, Davis).

Summary

This paper proposes the sequencing of 8 additional fruitfly and 7 nematode species selected specifically for their potential to enhance modENCODE's goal of complete functional annotation of DNA elements in the genomes of *Drosophila melanogaster* and *Caenorhabditis elegans*. While the modENCODE Project is pursuing that goal primarily by systematic, large-scale experimental analyses, those experiments are already being guided and enhanced by available comparative genomics data. It is our thesis that additional sequencing will greatly improve our ability to identify functional elements in the *D. melanogaster* and *C. elegans* genomes especially in conjunction with experimental data being generated by modENCODE. These species were chosen to fill "phylogenetic discovery gaps" – the absence of sequence data at critical evolutionary distances identified by recent theoretical and empirical studies - as ideal for functional annotation. We present evidence for the need for additional species, the rationale for our particular choice of evolutionary distances and species and an analysis of the likely benefit of these sequences to modENCODE.

Background and Justification

Every type of DNA element leaves a record of its existence in the history of natural selection that can be recovered from the genomes of closely related species. These evolutionary fingerprints have been exploited to annotate genomes in many contexts, including recent work in both of the modENCODE target species. Both the benefits of comparative annotation and the need for additional species are exemplified by analyses of the now 12 fully-sequenced *Drosophila* genomes [1, 2].

One of the principal motivations for the sequencing projects that led to the 12 *Drosophila* genomes was to develop general methodologies to discover and refine functional elements from comparative data that would be applicable to humans, and to empirically investigate how discovery power scaled with evolutionary distance for different classes of functional elements. The *Drosophila* 12 Genomes Consortium used the new sequences and the distinct evolutionary signatures of different functional elements to confirm, refine and augment annotated collections of protein-coding exons and transcripts, RNA genes and structures, miRNAs, pre- and post-transcriptional regulatory motifs and regulatory targets. Many of the new annotations have been validated by cDNA sequencing, human curation, small RNA sequencing, and other methods.

Specifically, the analysis identified thousands of novel functional elements in *D. melanogaster*, including: (a) 1,193 new protein-coding exons, of which 948 were assessed by manual curation and directed cDNA sequencing, validating 83% of those and leading to the creation of many new genes; (b) 394 new RNA structures, with candidate roles in post-transcriptional and translational control; (c) 41 novel microRNAs, of which 24 were validated experimentally; (d) 145 regulatory motifs in promoters, enhancers, introns, 3'UTRs and intergenic regions, most of which show tissue-specific expression and strong functional enrichments; (e) 46,525 binding sites for transcription factors and microRNAs, whose high confidence was validated by correlation with ChIP binding and experimentally confirmed microRNA targets.

In addition, the analysis revealed many new insights on the biology and regulation of animal genomes, including: (a) 149 genes with apparent stop-codon readthrough, 123 novel polycistronic transcripts, and several candidate programmed frameshifts with potential roles in regulation, localization and function of the corresponding protein products; (b) evidence of post-transcriptional regulation of many regulators, suggesting abundant feedback loops, and many new RNA structures involved in A-to-I RNA editing; (c) evidence that some miRNA loci yield multiple functional products, from both hairpin arms or from both DNA strands, increasing the versatility and complexity of miRNA-mediated regulation, and with important implications for Hox regulation; (d) evidence of redundancy between pre-transcriptional and post-transcriptional regulation, with particularly heavy targeting of transcription factors, by both miRNAs and transcription factor motifs.

From these analyses, it is clear that comparing genome data at different evolutionary distances has provided a significant amount of functional information for the reference species. It worth noting that the distances from *D. melangaster* of the 11 other genomes is from about 0.1 to over >2 subs/ss [4]. The *Caenorhabditis* genomes currently being completed will result in 5 *Caenorhabditis* species that will span a similar evolutionary distance [5]. Despite this tremendous success, it is clear that we have not yet fully exploited the power of comparative genomics to aid in the annotation of *D. melanogaster* or *C. elegans*. Stark et al. [2] investigated the likely benefit of additional sequencing by analyzing the discovery power of subsets of available data. As expected from earlier theoretical studies [6], recovery consistently increased with the total evolutionary distance spanned by the comparison; multi-species comparisons outperformed pairwise comparisons at the same total evolutionary distance; and the total evolutionary distance needed to identify specific classes of functional elements scaled inversely with the size of the element. Long proteins coding exons (greater than 300 nucleotides) were recovered at high rates with even small numbers of closely related species, suggesting that there is little room for improvement in their identification. In contrast, the recovery of most other classes of functional elements has not yet saturated even when all available species are included. Several types of very small elements – such as specific instances of transcription factor binding sites – were not as well-recovered as longer elements even when all available sequences were included, and their reliable discovery would particularly benefit from the additional species. While these studies have been done most extensively for *Drosophila*, similar patterns have been seen within the currently available genomes surrounding *C. elegans* (P. Sternberg, unpublished).

These analyses demonstrate that additional sequence data will contribute to the annotation of functional elements using multiple closely related species. Particularly in the context of modENCODE, these types of analyses can help shed light on every aspect of the biology of *D. melanogaster* and *C. elegans*, and provide a powerful complement to the various modENCODE projects, including for the identification of novel genes and transcripts, the discovery and characterization of small regulatory RNAs, the annotation of large and small chromatin domains of developmental importance, and the genome-wide binding of sequence-specific transcription factors. In each case, evaluating the level of conservation of each candidate region, and the particular intensity and type of selection it is under, will be invaluable in recognizing functionally important sequence elements associated with each type of element, increasing the resolution with which we can identify the DNA sequence elements responsible for each biochemical event, and distinguishing important functional elements under selection from biochemically-active but selectively-neutral regions [7].

But which species will provide the most useful data? Because recovery power scales with evolutionary distance, the naïve answer is simply to choose species at maximal evolutionary distance. However, two factors argue against selecting species that are too distant from the targeted species: 1) Alignment error: Comparative methods – especially the evolutionary signature-based methods shown by Stark et al. to be so effective – depend upon accurate alignments. However, alignment accuracy, and ultimately the ability to make alignments at all, decays with evolutionary distance, and Pollard et al. have demonstrated that there is a critical evolutionary distance beyond which alignment error precludes the effective use of comparative methods to recover functional elements [8]. 2) Functional divergence: Comparative annotation requires that the targeted element be present in the species being compared, and the probability that a functional element is present and has retained its function decays with evolutionary distance.

Several factors argue that the best strategy is to obtain data from a range of evolutionary distances with respect to the reference species. First, scaling analyses of the 12 *Drosophila* genomes data and of mammals suggest that the optimal distance for comparative identification of exons is approximately 0.5 subs/ss. Second, simulations suggest that the accuracy of alignments of non-coding DNA, and the ability to recover transcription factor binding sites and other short functional elements, decays after a distance of 1.0 [8]. Third, empirical studies show that optimal pair-wise distance for protein-coding gene identification is in the range of 0.5 to 1.0 subs/ss [4]. Fourth, preliminary results from a nGASP competition (to annotate gene models in *C. elegans*), show that algorithms that combine transcriptional data and comparative genomic data from the currently available *Caenorhabditis* species significantly outperform in sensitivity and specificity algorithms that only use one or the other types of data (going from a maximum sensitivity/specificity of 63.6/40.6 to 84.3/59.1. L. Stein, unpublished). Fifth, as seen in the *Drosophila* case, the additional sequences continued to provide useful comparative data and the discovery power continued to scale with each additional species without apparent saturation [4]. Finally, the bulk of the eutherian mammals are at an evolutionary distance of 0.2 to 0.5 subs/ss from humans, and thus analyses of the proposed species of *Drosophila* and *Caenorhabditis* species will provide valuable experience for the comparative annotation of the human genome.

Given the above considerations, our strategy here is to span a range of distances rather than focusing on a single distance. For *Drosophila* species, there is a paucity of data ranging between 0.2 and 0.8 and for *C. elegans* the gap is on either side of ~1 subs/ss. Spanning a range of distance also helps alleviate a number of challenges. 1) There is a fair amount of ambiguity in estimates of the optimal evolutionary distances. 2) Different types of functional elements have different optimal distances for a combination of reasons involving the size of the element and the precise nature of purifying selection on the element. 3) Functional elements are gained at lost at varying rates, and spanning a range of evolutionary distances maximizes the number of functional elements from the reference species that have orthologs in at least one comparative species.

Unfortunately, for both *D. melanogaster* and *C. elegans* “comparative sweet spots” remain largely uncovered by sequenced species. The species targeted by the 12 genomes project left a large gap around the 0.5 subs/ss distance we now believe to be optimal for many comparative analyses. The species of the melanogaster subgroup (*D. yakuba* and *D. erecta*) are approximately 0.2 subs/ss from *D. melanogaster*) while the next closest species, *D. ananassae* (is approximately 1.0 subs/ss from *D. melanogaster*). Fortunately, the remarkable diversity of *Drosophila* allows us to select additional species at the appropriate evolutionary distances from *D. melanogaster*.

For example, inclusion of additional species at this 'evolutionary sweet spot' significantly increased discovery power for short protein-exons (50-150 nucleotides). A comparison of all *melanogaster* subgroup species (total neutral branch length 0.4 substitutions per site) recovered 75% of exons at 99% specificity, while inclusion of *D. ananassae* in the comparison (increasing the total branch length to 1.3) recovered 90% of exons at the same specificity (a pairwise comparison of *D. melanogaster* with *D. ananassae* alone, at branch length 1.0, led to 85% recovery) [4].

These results are even more pronounced for the discovery of individual motif instances (typically 6-8 nucleotides), which we evaluated by comparing the conservation of known motifs to that of randomly shuffled control motifs. We found that for both transcription factor and microRNA motifs, the average signal-to-noise ratio increased from 2:1 to 3:1 for transcription factor motifs, and from 2.5:1 to 8:1 for microRNA motifs with the inclusion of additional species (from 6 to 12 species) [9]. Moreover, as distant species may typically lose individual motif instances due to evolutionary divergence, we expect these results to further improve with the addition of multiple species at this 'evolutionary sweet spot' of conservation.

For *C. elegans* the challenge has been in having access to species in the evolutionary vicinity of *C. elegans*. Recently, four related species have been selected to add a comparative dimension to the functional and evolutionary analysis of the *C. elegans* genome. Ideally, more than four species would have been selected. However, at that time, no other known species within a useful genetic distance from *C. elegans* were available. Over the last three years, due to a surge in collecting activity, a set of several new culturable *Caenorhabditis* species have been discovered (M-A Felix et al., unpublished). Molecular analysis place four of these species close to each other and near *C. elegan*s within the so-called *Elegans* group and the remainder outside this group (K. Kiontke et al., unpublished, see Figure 2). Notably, the two most distant of these *Elegans*-group species are as divergent as *D. melanogaster* and *D. anannasae*, well within the range of genetic distances among the 12 sequenced *Drosophila* genomes and are thus expected to be as useful in analyzing the *C. elegans* genome as has already been amply demonstrated in *Drosophila*. The additional four species selected outside the Elegans group will add additional points currently missing in the phylogenetic space surrounding *C. elegans.* Together, these species are chosen as the most valuable among all available species to help annotate the *C. elegans* genome.

So far, our justifications for including additional genomes for comparative sequencing relative to *C. elegans* have derived from the additional differences that accumulated in the evolution of additional species lineages. Additionally, a finer phylogenetic resolution increases the accuracy of assigning orthology and paralogy and determining which elements have been lost or gained. For example, many genes in the *C. elegans* genome exist in multigene families making orthology assignment challenging (see e.g. [10]). Thus, to allow accurate functional predictions, it is important to elucidate when gene duplications occurred relative to species divergences to identify orthologous genes/elements (thus likely to share function) and paralogous genes/elements (thus likely to diverge in function).

Species choice for Drosophila:

We have selected 8 additional species representing the major unsequenced lineages of the melanogaster group. These species are listed – in two priority groups -- in Table 1. Their relationships to the already sequenced species and the "phylogenetic discovery gaps" are shown in Fig. 1.

We also request, in Priority group 1, sequencing of *D. melanogaster*, strain Oregon R – a standard strain that is being used by experimentalists in several modENCODE projects. Oregon R is being used in those projects because it is healthier and faster-growing than the sequenced strain. Initial transcriptional profiling indicates the two strains are very similar, but clearly identification of elements will be enhanced by sequencing of Oregon R.

	Dist*	**subgroup**	**Artyom Kopp (UC-Davis)**			**Tucson Drosophila Stock Center**	
Priority 1.			**available?**	**isofemale?**	**inbreed? (full-sib)**	**available?**	**isofemale?**
D. ficusphila	0.80	ficusphila	X	X		X	
D. biarmipes	0.70	suzukii	X			X	X
D. elegans	0.72	elegans	X			X	
D. kikkawai	0.89	montium		X	X	X	
D. melanogaster, OregonR	*See explanation below*		X				
Priority 2.							
D. eugracilis	0.76	eugracilis	X	X		X	X
D. takahashii	0.65	takahashii	X				
D. rhopaloa	0.66	rhopaloa	X	X			
D. bipectinata	0.99	ananassae	X	X	X	X	X

Table 1.

* substitutions per neutral site with respect to *D. melanogaster* [3]. Note: *D. ananassae* = 0.151.

Notes on the species:

- ***D. ficusphila.*** An old, probably inbred line is available from the TDSC, and additional isofemale lines are available. The species is Southeast Asian; fruit-feeding, ecology not well studied. It is Moderately easy to maintain, not sure if it will tolerate much inbreding. There is some history of evo-devo work.

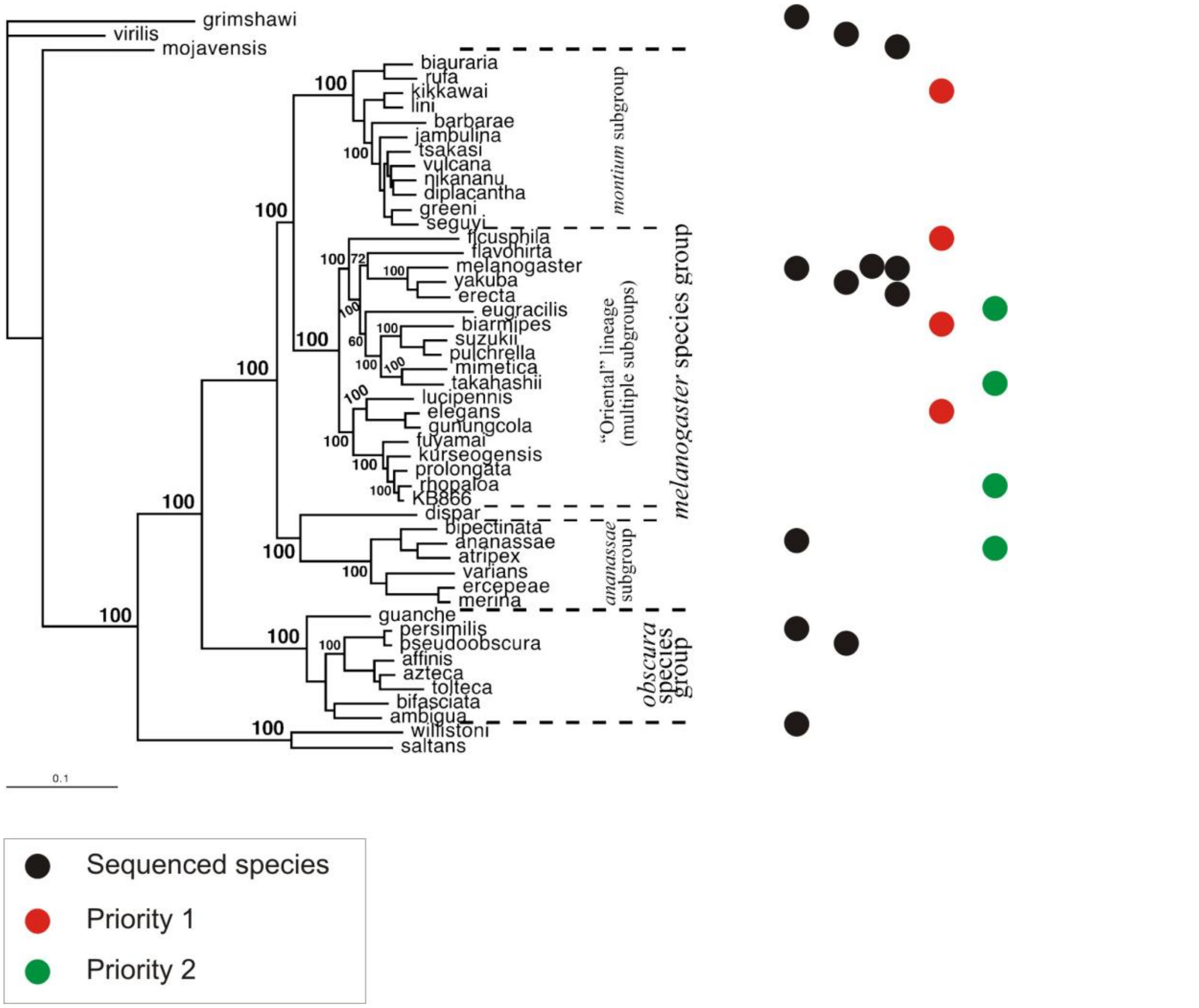

Figure 1. The figure shows the distribution of sequenced species and new requests among the *Sophophora.* Note the clusters of black dots (sequenced species) in the immediate vicinity of *D. melanogaster* and at the extremes of the tree. Note also how the requested species (red and green dots) fills the gaps at the most informative distances from *D. melanogaster*.

- ***D. biarmipes.*** An isofemale line is available from Tucscon. *D. biarmipes* is Part of a morphologically diverse lineage, mainly South Asian; fruit-feeding, ecology not well studied. It is very prolific and easy to maintain, should tolerate inbreeding. There is some some genetic and evo-devo literature.
- ***D. elegans.*** A 40 year-old line is available from Tucson. It is a southeast Asian; flower-feeding in the wild, some ecological information is available. It is Moderately easy to maintain, morphologically variable and has close relatives with which it can hybridize. A genetic map exists and there is a history of some genetic and evo-devo work. Additional lines are available .
- ***D. kikkawai.*** An isofemale line is available from Tucson This is one of the best studied species in the montium subgroup. It is a widespread tropical species, originally from SE Asia, invasive in Africa and S. America. It is prolific and easy to maintain, tolerates full-sib inbreeding, and is morphologically variable. There is some history of genetic and evo-devo work. Inbred lines are available.

- ***D. eugracilis.*** An isofemale line is available from Tucson and additional lines are available. It is Southeast Asian; fruit-feeding, ecology not well studied. It is Prolific and easy to maintain, should tolerate inbreeding.
- ***D. takahashii.*** A southeast Asian species, fruit-feeding, ecology not well studied. It is Prolific and easy to maintain, and should tolerate additional inbreeding. Additional lines are available.
- ***D. rhopaloa.*** One line is available from Artyom Kopp (Davis). The species is Southeast Asian; fruit-feeding, ecology not well studied. It is moderately easy to maintain.
- ***D. bipectinata.*** Several 50yo lines as well as recent isofemale lines available from Tucson. Many highly inbred, inversion-free strains are available. It is a distant relative of *D. ananassae*, very prolific and easy to maintain. Genetic maps, polytene chromosome maps, and morphological markers are available. It is an emerging model of phenotypic evolution, speciation, and population genetics. It is part of a morphologically diverse lineage and is itself morphologically variable and has close relatives with which it can hybridize. It is a widespread tropical species, originally from SE Asia, invasive in Africa and S. America.

Species choice for Caenorhabditis:

We have selected 7 species in two priority groups.

<u>Priority Group 1</u>

• *Caenorhabditis* sp. 9 (JU1325)
• *Caenorhabditis* sp. 11 (JU1373)
• *Caenorhabditis* sp. 7 strain JU1199

<u>Priority Group 2</u>

• *Caenorhabditis* sp. 5 strain JU727 (20x inbred line JU800)
• *Caenorhabditis* sp. 10 strain JU1333
• Two species from: *Caenorhabditis* sp. 6, *Caenorhabditis* sp. 3, *C. drosophilae*, *Caenorhabditis* sp. 8, *Caenorhabditis* sp. 2.

These species represent the closest species to *C. elegans* currently in culture. Four of the selected species are within the so-called *Elegans* group and four are outside. In addition, to help with the assembly and the annotation of the gene structures for each species, we propose to incorporate deep sequencing from a mixed stage cDNA library for each *Caenorhabditis* species excepting *C. elegans* and *C. briggsae* for which this is currently completed or in progress (R. Waterston et al., unpublished).

In addition to guiding the functional analysis of the *C. elegans* genome, these data will jump-start comparative genome analysis in this group of organisms, in which every cell can be observed *in vivo* and where tools like RNAi and transgenesis are available for functional analysis in several species (see species descriptions, below). The diverse natural histories among *Drosophila* and *Caenorhabditis* will provide unique opportunities to study the evolutionary forces underlying fundamental problems in molecular evolution. Broad comparative analysis in the genus will inform the genetic basis for distinct life histories, such as species with male-female versus hermaphroditic reproduction, and life cycles that include stages of quiescence (i.e., the dauer larva). Thus, in addition to the clear impact on modENCODE, these sequence projects are expected to spur a wave of evolutionary studies within and among these groups. The molecular phylogeny of the group selected is shown in Figure 1. This phylogeny includes all *Caenorhabditis* species currently in culture and available for molecular analyses with the exception of *C.* sp. 12. This species discovered in the spring of 2008 hybridizes with *C.* sp. 3.

<u>Priority Group 1:</u> three species.

All species of these groups have been isogenized or are self-fertilizing, ready for large-scale sequencing.

• *Caenorhabditis* sp. 9 strain JU1325 (20x inbred line JU1419)

This species was discovered recently in rotting flowers and leaves that were sampled in the Zoo/Botanical Garden of Trivandrum, Kerala, India. It is gonochoristic (male-female). Morphologically, it shows several salient differences from the other species of the *Elegans* group, which are otherwise noted for morphological uniformity. Intriguingly, in the laboratory, *C.* sp. 9 produces fertile offspring with *C. briggsae*. Since *C. briggsae* was sampled only 500m away from the sample site of *C.* sp. 9, it is likely that these two species occur sympatrically, and provides the potential for a model of speciation genetics in *Caenorhabditis*. A preliminary estimate shows that the genetic distance between *C. briggsae* and *C.* sp. 9 is similar to the distance between *Drosophila erecta* and *Drosophila yakuba*, making this the closest species pair in the *Elegans* group.

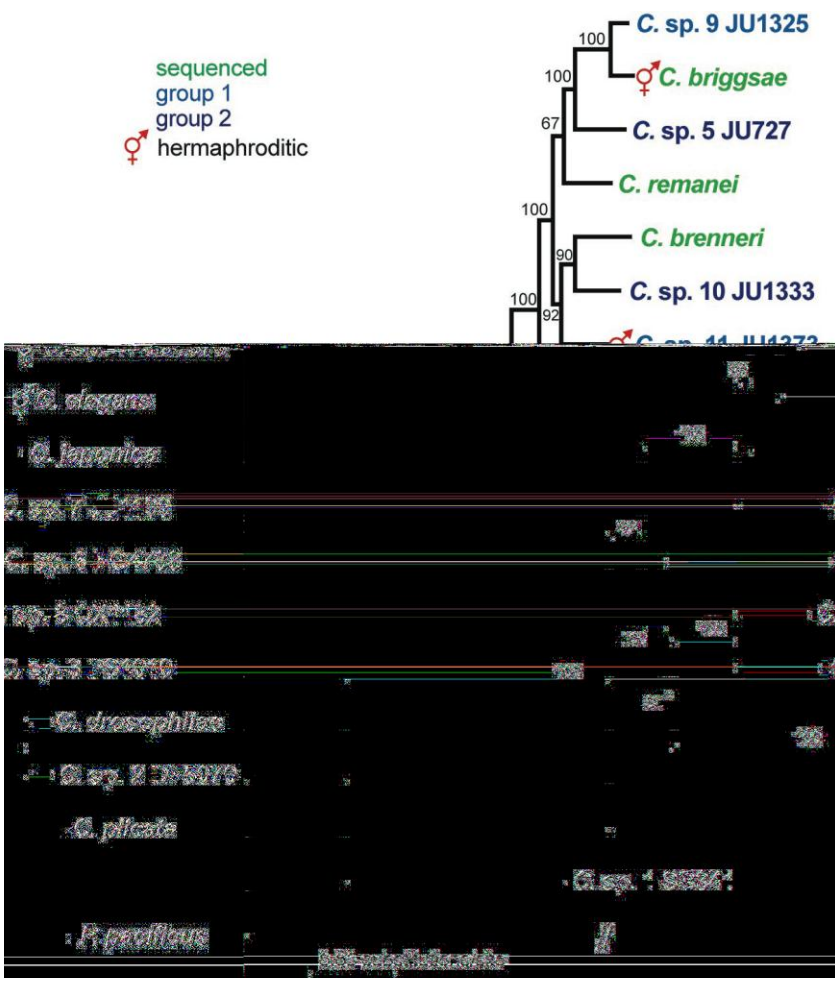

Figure 2: Molecular phylogeny of the *Caenorhabditis* species currently in culture, reconstructed using weighted maximum parsimony (transversions are weighted double transitions). The numbers at the branches denote jackknife values for 1000 replications with 2 addition sequence replications. The genes used were: near full length SSU and LSU rRNA genes, part of the gene for the largest subunit of RNA polymerase II, part of *lin-44*, *par-6*, *pkc-3* and orthologs

•***Caenorhabditis*** **sp. 11, strain JU1373**
This species was isolated from decaying flowers of the torch ginger (*Etlingeria elatior* a commercially grown plant) sampled on La Reunion, an island near Madagascar and again from a rotting *Duguetia surinamensis* fruit in French Guyana. Like *C. elegans* and *C. briggsae*, this species is composed of self-fertilizing hermaphrodites and facultative males, and it is likely that hermaphroditism evolved independently in *C.* sp. 11. Because of this reproductive mode, we expect a very low level of polymorphism in *C.* sp. 11.

• ***Caenorhabditis*** **sp. 7 strain JU1199 (20x inbred line JU1286)**
This species was isolated from a rotting citrus fruit sampled in Begoro, Ghana. *C.* sp. 7 is gonochoristic. The molecular phylogeny, in agreement with morphological characters, shows that this species branches off before the radiation of the *Elegans* group like *C. japonica* which was previously chosen as an outgroup for the other genome-sequenced species. However, unlike *C. japonica*, *C.* sp. 7 is easier to culture in the laboratory, does not suffer pronounced inbreeding depression, and is susceptible to RNAi by feeding. These features make this species an excellent candidate for experimental research.

Priority Group 2: *Four species*

• ***Caenorhabditis*** **sp. 5 strain JU727 (20x inbred line JU800)**
This gonochoristic species was first isolated from a soil sample collected in a rural area in Chengyang, Guanxi, China. It has since been isolated 4 additional times in China and Vietnam. Before the discovery of *C.* sp. 9, it was the closest known relative of *C. briggsae*, prompting several recent studies to include this species in analysis. *C.* sp. 5 breaks the branch between the pair *C. briggsae/C.* sp. 9 and the other species of the *Elegans* group. It is inbred and isogenized and ready for genome sequencing.

• ***Caenorhabditis*** **sp. 10 strain JU1333**
This species was isolated repeatedly from different rotting fruits (e.g. cocoa) collected in a garden near Periyar and a plantation in Kanjirapalli, Kerala, India. This is the fourth new species inside of the *Elegans* group. It is gonochoristic and the closest known relative of *C. brenneri,* the genome of which is already sequenced.

• **A final two species will be selected from the *Drosophilae* group** --- which would be morphologically and genetically quite different from the *Elegans* group plus *C. japonica* and *C.* sp. 7. Within this clade, a 300Kb sequence is available for PS1010 which has shown that it is useful as an outgroup. However, within this group the other species may offer more long term potential for being good satellite model systems in which to perform functional analysis. The two species in this group will provide an intermediate distance between *C. elegans* and *C.* sp 1, which is a distantly related *Caenorhabditis* species completing the range of genetic distances to be sampled for analysis of the *C. elegans* genome. We will select the best candidate species from this group based on ease of isogenizing the strain and largest impact based on any data that will be accumulated during the next few months (see below).

Caenorhabditis sp. 6 (EG4788) was recently isolated from a rotten apple in Amares, Portugal. This species is gonochoristic. It is one of the few *Caenorhabditis* species that are sensitive to RNAi by feeding.

Caenorhabditis sp. 3 (PS1010, RGD1, RGD2) is regularly found in palm trees infested with the sugar cane weevil *Metamasius hemipterus,* a major crop pest, in Florida and Trinidad.

An intimate association of the nematodes with these beetles is likely. A small part of the genome (~300Kb) of strain PS1010 was sequenced and regulatory elements were analyzed for some genes (S. Kuntz et al, in preparation). This strain suffers from inbreeding depression, but there are other strains available for generation of healthier isogenic lines. Polymorphism levels appear to be high. The discovery of *C.* sp. 12 which produces fertile hybrids with *C.* sp. 3 opened up possibilities for evolutionary studies within *Caenorhabditis.*

C. drosophilae is a regular colonizer of rotting columnar cacti in Arizona. It is tightly associated with the cactophilic fly *Drosophila nigrospiracula. C. drosophilae* is one of only 5 out of the 17 cultivable *Caenorhabditis* species which are not found predominantly in anthropogenic habitats (the others are *C. plicata*, *C. japonica*, *C.* sp. 1 and *C.* sp. 2). *C. drosophilae* is very closely related to *C.* sp. 2, another cactophilic species from Europe. Although both species are reproductively isolated, they appear genomically very similar. Because of their narrowly defined ecological niche, their specific island habitat and their specificity for a phoretic host, *C.* sp. 3 and *C. drosophilae* are most promising candidates for future ecological and population biological investigations.

Sequence Quality and Strategy

The goals of this proposal are to obtain comparative genomic data to advance the goals of modENCODE. Therefore, each genome should be sequenced at a level depth and quality that will facilitate assembly and comparisons to the reference species. Repeated sequences, including gene families, transposons, and heterochromatin in *Drosophila* confound genome assembly, and directed finishing work is required to produce contiguous sequences spanning large segments of chromosomes. Ideally, many genomic regions would be spanned by Mb-scale assemblies. However, we have devised a *minimally useful set of parameters* keeping in mind the current evolution of the sequencing technologies. Our minimal parameters are based on the following considerations. The average gene in the *Elegans* group and *Drosophila* is ~5-10 kb (*C. elegans*: 100,000 kb / 20,000 genes = 5 kb; *D. melnaogaster:* 120,000 Kb/ 14,000 = 8.7Kb), and since there are few data to suggest long range-acting sequences, contigs that are at least the size of a single gene are important. However, many genes with complex transcriptional regulation are larger than 5 kb (up to 40 kb for C. elegans and larger than 100Kb for Drosophila); elements within one gene’s introns are known to affect an adjacent gene; and operons can have 2-10 genes. Thus, contigs that are at least 20 kilobases will provide most of the key information. We therefore need 97% of the contigs to be 20 kb or greater. We suggest that a mixed strategy using Illumina (7X paired-end; ~2 kb separation), 454 (1X paired-end; ~10 kb separation) and 0.1X fosmid-end reads is likely to achieve the desired assemblies.

RNASeq.

In addition to sequencing these species we also propose to conduct deep sequencing runs of cDNA libraries for each of the sequenced genomes. A minimum of 1 full flow cell run on an Illumina/Solexa machine per species is required to provide deep enough coverage to assemble many transcripts, and thus help with the annotation. Deep sequencing of cDNA for the 7 *Caenorhabditis* species proposed here, plus *C. remanei* and *C. japonica*, requires at least 9 full flow cell runs. In addition, 8 full flow cell runs for the *Drosophila* species are proposed.

References

1. Clark, A.G., et al. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature, 2007. 450:203-18.
2. Stark, A., et al. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. Nature, 2007. 450:219-32.
3. Barmina, O. and A. Kopp. Sex-specific expression of a HOX gene associated with rapid morphological evolution. Dev Biol, 2007. 311:277-86.
4. Lin, M.F., et al. Performance and scalability of discriminative metrics for comparative gene identification in 12 *Drosophila* genomes. PLoS Comput Biol, 2008. 4:e1000067.
5. Cutter, A.D. Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. Mol Biol Evol, 2008. 25:778-86.
6. Eddy, S.R. A model of the statistical power of comparative genome sequence analysis. PLoS Biol, 2005. 3:e10.
7. Birney, E., et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 2007. 447:799-816.
8. Pollard, D.A., et al. Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. BMC Bioinformatics, 2006. 7:376.
9. Kheradpour, P., et al. Reliable prediction of regulator targets using 12 *Drosophila* genomes. Genome Res, 2007. 17:1919-31.
10. Stein, L.D., et al. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. PLoS Biol, 2003. 1:E45.