

Extremely high molecular diversity within the East Asian nematode *Caenorhabditis* sp. 5

GUO-XIU WANG,* SHUANG REN,* YUAN REN,* HUI AI* and ASHER D. CUTTER†‡

*Hubei Key Laboratory of Genetic Regulation and Integrative Biology, HuaZhong Normal University, Wuhan, Hubei, China,

†Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, ON M5S 3B2, Canada, ‡Centre for the

Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON M5S 3B2, Canada

Abstract

Most relatives of the self-fertilizing hermaphroditic nematode model organism *Caenorhabditis elegans* reproduce via obligate outbreeding between males and females, which also represents the ancestral mode of reproduction within the genus. However, little is known about the scope of genetic diversity and differentiation within such gonochoristic species, especially those found outside of temperate Europe and North America. It is critical to understand the evolutionary processes operating in these species to provide a framework for deciphering the evolution of hermaphroditism and a baseline for the application of outcrossing *Caenorhabditis* to problems in evolutionary genetics. Here, we investigate for the first time molecular sequence variation for *Caenorhabditis* sp. 5, a species found commonly in eastern Asia. We identify enormous levels of standing genetic variation that approach the levels observed in the marine broadcast-spawning sea squirt, *Ciona savignyi*. Although we document significant isolation by distance, we demonstrate that the high polymorphism within *C.* sp. 5 is not because of strong differentiation among populations or to the presence of cryptic species. These findings illustrate that molecular population genetic approaches to studying obligately outbreeding species of *Caenorhabditis* will prove powerful in identifying and characterizing functionally and evolutionarily important features of the genome.

Keywords: *Caenorhabditis*, nucleotide variation, phylogeography

Received 30 June 2010; revision received 30 August 2010; accepted 3 September 2010

Introduction

The *Caenorhabditis* genus of nematodes is the focus of a recently burgeoning literature on comparative developmental biology, evolution and genetics (Kammenga *et al.* 2008; Haag 2009). Much of this work is inspired by the experimental tractability of the classic nematode animal model, *C. elegans*—but a protracted question surrounds the generality of inference to be made from molecular and phenotypic variation found in *C. elegans*. Specifically, the derived breeding system of *C. elegans*, in which populations primarily comprise self-fertilizing hermaphrodites, exerts profound consequences for much of the organism's biology (Cutter *et al.* 2009), despite the conveniences of laboratory manipulation

imparted by selfing hermaphroditism. A major unsolved issue needed to fully characterize the evolution and ecology of *Caenorhabditis* nematodes—and *C. elegans*' place within this group—is to better understand species with the ancestral, gonochoristic (=dioecious, male-female) breeding system. At present, analysis of phylogeographical, population genetic and phenotypic variation for gonochoristic *Caenorhabditis* is known reasonably well only for *C. remanei* in this group that contains over 30 species (Kiontke & Fitch 2005; Kammenga *et al.* 2008; Cutter *et al.* 2009; K. Kiontke personal communication). In this study, we characterize for the first time molecular genetic variation in a broad collection of the gonochoristic species *C.* sp. 5—which appears to be restricted in distribution to eastern Asia—to help fill this gap in our understanding of outbreeding taxa that represent the ancestral mode of reproduction for *Caenorhabditis*.

Correspondence: Asher D. Cutter, Fax: 416 978 5878; E-mail: asher.cutter@utoronto.ca

Despite the exceptional insights into biological processes permitted by studying highly selfing species of *Caenorhabditis*, several important problems are impossible to investigate in such species. For example, sexual selection is thought to be largely absent in natural populations of *C. elegans* and *C. briggsae*, yet is probably an important evolutionary force in obligately outbreeding species (LaMunyon & Ward 1999; Cutter 2008c). Investigation of the functional and molecular evolutionary consequences of male–male sperm competition and the potential for intersexual antagonistic coevolution will be best dissected in outbreeders that currently experience such selection in nature. Moreover, data from *C. remanei* (and from *C. sp. 5* presented in this study) indicate that segregating genetic variation and effective population size (N_e) are dramatically greater than in selfing species (Cutter *et al.* 2009). This has important implications for detecting and quantifying adaptation, the evolutionary response to even very weak natural selection and other population genetic processes in the absence of complex demographic population structure and genome-wide linkage disequilibrium (LD) induced by a selfing life-style. A variety of molecular population genetic methods have been developed for such inference, and their statistical power is determined in-part by the extent of genetic variation and linkage, which is problematic for their application to *C. elegans* (Kreitman 2000; Boffelli *et al.* 2004; Nielsen 2005).

C. sp. 5 is an obligately outbreeding species discovered in China independently by Walter Sudhaus and Marie-Anne Felix (Kiontke & Sudhaus 2006; Sudhaus & Kiontke 2007). Like other *Caenorhabditis*, *C. sp. 5* is a small (~1 mm long), transparent bacterivorous nematode amenable to laboratory culture. It is closely related to *C. briggsae* and *C. sp. 9* (Kiontke *et al.* 2007; Cutter 2008a), and by virtue of providing the nearest outgroup for these two sister species, *C. sp. 5* holds an important position for analyses of molecular evolution and comparative phenotypic analysis based on *C. briggsae*. The shape of part of the male reproductive anatomy is distinct in *C. sp. 5* compared to other *Caenorhabditis* (K. Kiontke, personal communication), as are some aspects of cell division in the developing female vulva (Dolgin *et al.* 2008), making it an interesting subject for morphological analysis of the evolution of development. Until recently, only one main isolate was available (JU727), for which an expressed sequence tag library was constructed and used in analyses of molecular divergence and the evolution of biased codon usage (Cutter 2008a; Cutter *et al.* 2008). The evidence of selection-mediated codon bias suggests that *C. sp. 5* likely has a large effective population size, which is necessary for weak translational selection for codon bias to manifest (Bulmer 1991). However, an understanding of the extent of its

species range, the level genetic variation and the diversity of environments inhabited by *C. sp. 5* has been unavailable until now.

In this study, we document the extent of nucleotide polymorphism across the X-chromosome of *C. sp. 5* and quantify the degree to which molecular genetic variation is partitioned among sampling localities and evinces deviation from neutral population genetic expectations for a single panmictic population. Our extensive sampling confirms that this species has a widespread distribution within China, but so far has been isolated elsewhere in the world only in northern Vietnam near its border with China. *C. sp. 5* harbours an exceptionally high level of polymorphism, contributing to an emerging picture of the gonochoristic species of *Caenorhabditis* being hyperdiverse in molecular population genetic variation.

Materials and methods

Nematode strains

We collected 18 isofemale lines of *C. sp. 5* from throughout eastern and central China during 2008 and 2009 according to established procedures (Barrière & Félix 2006) (Table 1). We supplemented these isolates with an additional 4 strains, also from China or from Vietnam near the Chinese border, that were gifts of Marie-Anne Felix (Table 1). Strains are available from the authors or the Felix lab (<http://www.justbio.com/worms>). Initial species determination was made through crosses with other known species and sequencing a portion of the small subunit rDNA. The geographical sources of collection are viewable through an online mapping resource <http://www.google.com/maps/ms?msa=0&msid=117700919974655793194.00046c90c5fa3a4c5837b>. We computed the distance between collection sites using the haversine formula, assuming the radius of the Earth to be 6367.5 km. The most distant pair of sites are separated by 1786 km, and on average the sites are 644 km apart.

Molecular methods

We used PCR primers previously developed for *C. remanei* to amplify nuclear gene fragments from the *C. sp. 5* genome (Cutter 2008b). After initial testing of a large number of primer pairs, we successfully obtained sequence from *C. sp. 5* for 7 primer sets. These primers include the p46, p47, p50, p51, p74, p87 and p94 primer sets described previously (Cutter 2008b). These loci are putatively X-linked and contain coding sequence from a single exon, representing about 4.4 kb of total sequence from each individual. X-linked loci were selected to

Table 1 Collection locality information of strains used in this study

Strain	Latitude	Longitude	Locality	Date	Source
VX0035	25.028N	113.393E	Lechang (Changlai town), Guangdong	9/1/2008	compost soil in cabbage field
VX0036	25.028N	113.393E	Lechang (Changlai town), Guangdong	9/1/2008	snail (<i>Oxychilus</i> sp.)
VX0037	25.028N	113.393E	Lechang (Changlai town), Guangdong	9/1/2008	rotting bamboo leaves (<i>Bambusa textilis</i>)
VX0065	30.514N	114.495E	Wuhan (Jiufeng village), Hubei	9/20/2008	soil near pond
VX0066	29.350N	102.683E	Yaan, Sichuan	9/5/2008	soil under poplar tree
VX0067	30.550N	114.250E	Wuhan (Hanyang town), Hubei	9/30/2008	soil
VX0068	23.617N	110.150E	Yulin, Guangxi	10/23/2008	soil in banana plantation
VX0069	30.519N	114.358E	Wuhan (CCNU), Hubei	10/24/2008	isopod
VX0070	30.514N	114.495E	Wuhan (Jiufeng village), Hubei	9/20/2008	soil in cabbage field
VX0071	23.425N	113.236E	Guangzhou (Huadu district), Guangdong	12/4/2009	compost soil from vegetable field
VX0072	23.425N	113.236E	Guangzhou (Huadu district), Guangdong	12/4/2009	compost soil from vegetable field
VX0073	24.776N	113.271E	Shaoguan (Ruyuan county), Guangdong	12/3/2009	isopod under bamboo
VX0074	24.776N	113.271E	Shaoguan (Ruyuan county), Guangdong	12/3/2009	isopod under bamboo
VX0075	23.139N	113.268E	Guangzhou (Renhe town), Guangdong	11/30/2009	moist soil
VX0076	25.028N	113.393E	Lechang (Changlai town), Guangdong	12/3/2009	compost soil from vegetable field
VX0077	25.028N	113.393E	Lechang (Changlai town), Guangdong	12/3/2009	compost soil under bamboo
VX0078	24.776N	113.271E	Shaoguan (Ruyuan county), Guangdong	12/3/2009	compost soil under bamboo
VX0079	30.519N	114.358E	Wuhan (CCNU), Hubei	8/4/2009	isopod
JU727	25.903N	109.651E	Chengyang, Guangxi	5/6/2005	under a tree
JU1201	31.317N	120.617E	Suzhou, Jiangsu	8/17/2007	small fruit in garden
JU1202	30.300N	120.183E	Hangzhou, Zhejiang	8/25/2007	small fruit in park
JU1423	22.417N	105.583E	near Ba Be Lake, Bac Kan	4/29/2008	rotting fruit

ease analysis, because DNA isolated from a single male (karyotype: 5 diploid autosomes and a haploid X chromosome) eliminates the potential for heterozygous base calls in sequence traces. We refer to the genes of these partial coding sequences using an extension of the nomenclature suggested by Jonathan Hodgkin for other *Caenorhabditis*, namely, an abbreviation representing the species identity followed by the *C. elegans* ortholog gene name (e.g. *Csp5-spc-1*). DNA used as template for PCR was isolated from single male worms of each strain and amplified with the Qiagen Repli-g mini kit as described previously (Cutter 2008b). The lack of heterozygous single nucleotide polymorphism (SNP) calls in sequence traces supports the expectation of sex-chromosome linkage for these loci. Sequencing was unsuccessful for VX0070 for primer set p94, giving a sample size of 21 for this locus in analysis. Sequencing was performed at the University of Arizona UAGC sequencing facility and by the University of Toronto CAGEF sequencing facility. New sequences have been deposited in Genbank under accessions HQ189144-H189296.

Sequence analysis

The programs SEQUENCHER 4.10 and BIOEDIT 7.0.9 were used to manually edit and confirm sequence quality. We then calculated population genetic summary statistics for synonymous and replacement sites separately with DnaSP 5.10 (Librado & Rozas 2009), including

polymorphism from per-site pairwise differences (π) and segregating sites (θ) with Jukes–Cantor multiple-hits correction, the number of segregating sites (S) and the site frequency spectrum (Tajima's D). We also used DnaSP to calculate the number of haplotypes (H), the minimum number of recombination events (R_{\min}) and LD (r^2) for each locus. Pairwise genetic distances used in a test of isolation by distance are based on Nei & Gojobori's (1986) model for synonymous sites (K_s), also calculated in DnaSP. Measures of codon bias (F_{op} , frequency of preferred codons using *C. elegans* codon table; N_c , effective number of codons) were measured for the gene fragments with CodonW (J. Peden, <http://codonw.sourceforge.net>). Divergence with *C. briggsae* of homologous regions (K_a , K_s , K_a/K_s) were calculated using the Goldman and Yang method implemented in Kaks_calculator (Goldman & Yang 1994; Zhang *et al.* 2006). The population recombination rate (ρ) was estimated with LDhat's finite-sites model, and nonparametric tests for recombination were performed in LDhat with the 'Lkmax' method (McVean *et al.* 2002). We used nonlinear fitting in JMP 7 to fit the expected decay in r^2 with distance between polymorphic sites (Weir & Hill 1986) as carried out previously (Cutter *et al.* 2006), as well as a spline fit. A Mantel test on genetic and geographical distance matrices was implemented with custom Perl scripts, assuming independence of loci (1000 matrix permutations). We repeated this Mantel test approach using just a single,

arbitrarily selected strain from those localities for which we had sampled two or more strains; this approximates the 'scattered sample' paradigm of the coalescent process for many demes (Wakeley & Aliacar 2001). We made a rough calculation of effective population size (N_e) assuming demographic equilibrium (i.e., $\pi_{\text{syn}} = \theta = 4 N_e \mu$) and neutral mutation rate (μ) from *C. elegans* (2.7×10^{-9} mutations/site/generation) or *C. briggsae* (2-fold higher than *C. elegans*) (Denver *et al.* 2009; Phillips *et al.* 2009).

Results

Patterns of polymorphism

Nucleotide polymorphism at synonymous sites is very high in our sample of 22 individuals of *C. sp. 5*, averaging 6.84% across seven nuclear loci (Jukes–Cantor corrected; uncorrected average $\pi_{\text{syn}} = 6.40\%$). Polymorphism varied roughly 2.5-fold across loci, with the locus with lowest diversity still being highly variable in the population: $\pi_{\text{syn}} = 4.36\%$ (*Csp5-asp-3*; Table 2). However, polymorphism at replacement sites is very low (average $\pi_{\text{rep}} = 0.05\%$; Table 2), indicating strong constraint on changes to the amino acid sequence encoded by these genes. Indeed, the overall $\pi_{\text{rep}}/\pi_{\text{syn}}$ ratio of 0.0093 implies that more than 99% of mutations to replacement sites are effectively eliminated by purifying selection. Across the 4.4 kb of coding sequence sampled from each individual, we identified 222 SNPs at synonymous sites and 9 at replacement sites.

We calculated the Tajima's (1989) *D* summary of the variant frequency spectrum for synonymous sites and found that none of the loci deviated significantly from the neutral expectation. We note, however, that the average value of *D* across loci is positive (mean

$D_{\text{syn}} = 0.24$), which could be indicative of a modest level cryptic population structure within our sample or a relatively recent population contraction. Given the fairly broad geographical distribution of sampling sites across China, we consider cryptic population structure to be very plausible. As expected with purifying selection targeting replacement sites, Tajima's *D* indicates a trend for skew towards low-frequency variants and a much lower mean value (mean $D_{\text{rep}} = -0.27$) relative to synonymous sites.

We also calculated codon bias for these gene regions as well as their divergence with *C. briggsae* (Table 3), because selection on codon usage has been demonstrated to affect synonymous-site divergence in *C. sp. 5* (Cutter 2008a; Cutter *et al.* 2008). However, our sample of 7 loci is not sufficient to allow strong tests for correlations with polymorphism, and we did not adjust K_s for selection on synonymous sites.

Linkage disequilibrium

Linkage disequilibrium decays detectably across only hundreds of nucleotides between SNP pairs, dropping by half within ~250 bp on average (Fig. 1). When each locus was considered separately (mean 630 bp long), the decay of LD was significant for five of the seven loci (Table 3). Moreover, recombination was evident in all loci, with the minimum number of recombination events (R_{min}) ranging from 2 to 13 within each locus (Table 3). Estimating the population recombination parameter ($\rho = 4 N_e c$, effective population size N_e and recombination rate per generation c) with the finite-sites method implemented in LDhat, we again confirm extensive recombination with mean $\rho = 0.0125$ per site. Consequently, the ratio ρ/π_{syn} averages 0.199, indicating that there is abundant nucleotide variation relative to the level of recombination.

Table 2 Summary of molecular genetic variation statistics†

gene (primer)	π_{syn}^a	π_{rep}^a	θ_{syn}	θ_{rep}	S_{syn}	S_{rep}	D_{syn}	D_{rep}	<i>H</i>
<i>Csp5-asp-3</i> (p46)	0.0436	0.00112	0.0390	0.00074	16	1	0.280	0.895	11
<i>Csp5-spc-1</i> (p47)	0.1131	0	0.0846	0	42	0	0.839	n.a.	10
<i>Csp5-col-41</i> (p50)	0.0677	0	0.0697	0	42	0	-0.320	n.a.	12
<i>Csp5-myo-2</i> (p51)	0.0599	0.00020	0.0480	0.00059	20	1	0.662	-1.162	12
<i>Csp5-pgp-14</i> (p74)	0.0891	0.00089	0.0710	0.00050	42	1	0.638	1.334	13
<i>Csp5-W07E11.1</i> (p87)	0.0480	0.00035	0.0430	0.00056	24	1	0.241	-0.641	11
<i>Csp5-E01G6.1</i> (p94)	0.0573	0.00107	0.0650	0.00265	36	5	-0.629	-1.795	15
average	0.0684	0.00052	0.0600	0.00072	31.7	1.3	0.244	-0.274	12.0

†average pairwise nucleotide polymorphism at synonymous (π_{syn}) and replacement-sites (π_{rep}), number of segregating sites at synonymous (total observed S_{syn} , scaled per site θ_{syn}) and replacement sites (S_{rep} , θ_{rep}), Tajima's (1989) metric of the site frequency spectrum at synonymous (D_{syn}) and replacement sites (D_{rep}), observed number of haplotypes (*H*).

^aJukes–Cantor corrected.

Table 3 Summary of molecular evolution statistics†

Gene (primer)	Length ^a	R_{\min}	ρ^b	F_{op}^c	N_c	K_a^d	K_s^d	K_a/K_s^d	π_{rep}/π_{syn}^e
<i>Csp5-asp-3</i> (p46)	485	2	0.0153**	0.383	52.0	0.0238	0.369	0.0644	0.0257
<i>Csp5-spc-1</i> (p47)	670	7	0.0046 ^{n.s.}	0.721	33.1	0.0020	0.523	0.0039	0
<i>Csp5-col-41</i> (p50)	657	6	0.0088 ^{n.s.}	0.322	51.0	0.0355	0.417	0.0850	0
<i>Csp5-myo-2</i> (p51)	579	3	0.0079**	0.347	44.7	0.0130	0.503	0.0259	0.0033
<i>Csp5-pgp-14</i> (p74)	709	13	0.0295**	0.284	51.2	0.0261	0.885	0.0295	0.0100
<i>Csp5-W07E11.1</i> (p87)	645	3	0.0104**	0.366	50.2	0.0244	1.430	0.0171	0.0073
<i>Csp5-E01G6.1</i> (p94)	680	4	0.0110**	0.333	48.9	0.0230	0.927	0.0248	0.0187
average	632.1	5.4	0.0125	0.394	47.3	0.0211	0.722	0.0358	0.0093

†minimum number of recombination events (R_{\min}), population recombination parameter per site (ρ), frequency of optimal codons (F_{op}), effective number of codons (N_c), divergence with *C. briggsae* at replacement sites (K_a) and synonymous sites (K_s), ratio of pairwise nucleotide polymorphism at replacement and synonymous sites (π_{rep}/π_{syn}); ^agaps excluded; ^bsignificance of recombination ** $P \leq 0.001$ via Lkmax method; ^c*C. elegans* codon preference table; ^dpairwise Goldman–Yang estimates with *C. briggsae*; ^eJukes–Cantor corrected.

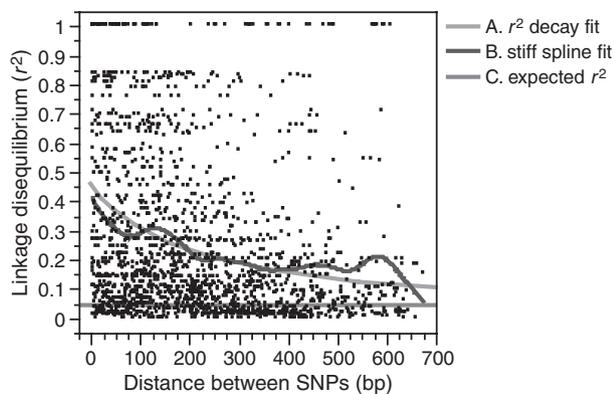


Fig. 1 Decay of linkage disequilibrium (LD) with distance between pairs of polymorphic sites. Data for all loci combined reveal the overall rapid decay of LD and that it approaches expected background levels within hundreds of nucleotides. Monotonically decreasing curve A is the nonlinear fit to the expected decay in LD (Weir & Hill 1986), and curve B is a stiff spline fit. Horizontal line C indicates the expected r^2 , given the sample size (Weir & Hill 1980). Statistical tests of recombination within each locus are given in Table 3.

Tests for structure

We tested for evidence of isolation by distance by performing a Mantel test on the pairwise geographical distances among sampled strains with the pairwise genetic distances among sampled strains (K_s), assuming independence of loci. This analysis revealed a significant, albeit quantitatively weak, positive correlation between genetic and geographical distance (Pearson $r = 0.122$, $P < 0.001$). We repeated the Mantel test with exactly one strain from each locality to better represent a ‘scattered sample’ for the collecting phase of the multi-deme coalescent (Wakeley & Aliacar 2001) and found consistent results to the original analysis (Pearson $r = 0.151$,

$P < 0.001$). This finding of isolation by distance further supports the notion that there is some level of population structure among the sampled individuals, as suggested by the slightly positive values of Tajima’s D . However, the fact that we have just a single iso-female isolate (strain) from most sampling sites precludes further tests of substructure among sampling localities within China (e.g. F_{ST}).

Discussion

Our sample of individuals for *C. sp. 5* identifies substantial nucleotide polymorphism, averaging $\pi_{syn} = 0.068$, which implies that there is about 1 SNP of every 15 silent sites between a random pair of sequences. We find that *C. sp. 5* exhibits weak, but significant, genetic isolation by distance, with individuals from more geographically distant localities tending to have more divergent haplotypes. However, there is pervasive recombination within loci, causing LD to decay rapidly over spans of just a few hundred nucleotides. Thus, the extremely high diversity observed within *C. sp. 5* does not appear to be because of divergence between differentiated lineages or subpopulations that might comprise cryptic species. Instead, we propose that the effective population size of *C. sp. 5* is very large. We expect that large effective population size could be maintained by virtue of the abundant food resource for *C. sp. 5* (bacteria in rotting vegetable matter), an apparently high dispersal ability and its obligately outbreeding mode of reproduction. A rough calculation of long-term N_e for *C. sp. 5* is $3\text{--}6 \times 10^6$, assuming demographic equilibrium and that the estimates of mutation rate from *C. elegans* and *C. briggsae* based on mutation accumulation experiments can extend to *C. sp. 5* (Denver *et al.* 2009; Phillips *et al.* 2009).

The signature of genetic isolation by distance within *C. sp. 5* suggests that this species harbours some degree of population structure across its range in east Asia. With available sampling, which is currently limited to one or a few isofemale strains from any given locality, it is not yet possible to detail the extent of subpopulation differentiation. By considering the total sample in the face of some degree of structure, our analyses likely give slightly biased estimates of some population genetic parameters. For example, within any single locality, we should expect polymorphism, LD and the deficit of low-frequency variants to be somewhat lower than the values reported here. However, the high number of pairwise differences between sequences expected even at a geographical distance of 0 km (Fig. 2) as well as the rapid decay of LD, even in the total sample (Fig. 1), indicate that detailed investigation of individual localities should corroborate the main findings we report here.

C. remanei populations also have high levels of molecular variation at synonymous sites, albeit substantially lower than *C. sp. 5* (Graustein *et al.* 2002; Jovelin *et al.* 2003, 2009; Cutter *et al.* 2006; Cutter 2008b; Jovelin 2009). Homologous loci in *C. remanei* from Ohio, USA show $\pi_{\text{syn}} = 0.022$, although much larger samples of loci in that species give an average of $\pi_{\text{syn}} \sim 0.04$ (Cutter 2008b; Jovelin 2009). Thus, high nucleotide polymorphism appears to be a common feature of obligately outbreeding species of *Caenorhabditis*—and should exemplify the state of those ancestral populations that gave rise to the derived, extant species that are comprised of self-fertile hermaphrodites (*C. elegans*, *C. briggsae*, *C. sp. 11*). *C. sp. 5*, in particular, harbours enormous levels of polymorphism in absolute terms,

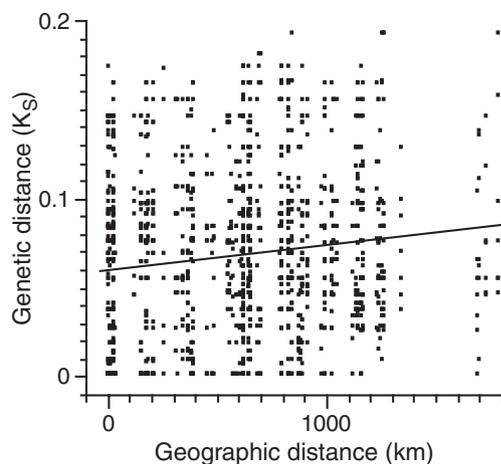


Fig. 2 The average genetic distance between individuals increases with geographical distance. A Mantel test identifies significant isolation by distance ($P < 0.001$). Data for all loci is superimposed in the figure.

rivalling that of the sea squirt *Ciona savignyi* (average $\pi = 0.08$ at 4-fold synonymous sites), which is the metazoan with highest documented molecular diversity across its genome to date (Small *et al.* 2007).

The level of polymorphism seen in *C. sp. 5* and *C. remanei* greatly exceeds the levels observed in self-fertilizing taxa like *C. elegans* and *C. briggsae*, differing by a factor of at least 20 (reviewed in Cutter *et al.* 2009). Moreover, this disparity between the obligate outbreeding species and the self-fertile species is much greater than the roughly 2-fold difference expected from selfing alone. Presuming that polymorphism in these outbreeding species is representative of the values present in the outbreeding ancestors of *C. elegans* and *C. briggsae*, this indicates that other effects beyond selfing alone must have reduced disproportionately the amount of variation in self-fertile lineages (Graustein *et al.* 2002; Sivasundar & Hey 2003; Barrière & Félix 2007; Cutter *et al.* 2009). Such diversity-reducing factors potentially include potent extinction–recolonization metapopulation dynamics, a recent origin of high selfing rate in each species and the consequences of selection at linked sites. These factors are not mutually exclusive, and there is support in the literature from molecular evolution and population genetic data for each of them individually in *C. elegans* and/or *C. briggsae* (reviewed in Cutter *et al.* 2009).

C. sp. 5 is found only in East Asia but in both tropical and temperate latitudes, whereas *C. remanei* appears to be found around the globe but geographically restricted to northerly latitudes (Sudhaus & Kiontke 2007). Other gonochoristic species have more tropical known distributions (e.g. *C. brenneri*) (Sudhaus & Kiontke 2007). We isolated *C. sp. 5* from a variety of sample types that are typical of habitats in which *Caenorhabditis* have been found previously (Kiontke & Sudhaus 2006; Sudhaus & Kiontke 2007), including in phoretic association with isopods and snails, in rotting fruit and vegetation, in compost and in agriculture-associated soil (Table 1). We also identified *C. remanei* and *C. briggsae* at some localities where we also found *C. sp. 5*, indicating that the geographical ranges of these three species overlap at a fairly fine scale within China. Further sampling of *Caenorhabditis* species is necessary to fully document the extent of their species ranges, habitat types, potential niche partitioning and phylogenetic diversity.

One locus, *Csp5-spc-1*, revealed unusual patterns of genetic variation in several respects (Tables 2 and 3). *Csp5-spc-1* had the fewest haplotypes, and yet exhibited the highest synonymous-site diversity among the loci we analysed, despite being the only locus with strongly skewed codon usage that is indicative of selection acting on synonymous sites. Fifteen of 22 samples fall into three differentiated haplotypes for the 670-bp portion of

Csp5-spc-1 we have analysed, one of which includes individuals derived only from localities in southeastern China. This locus also had the most skewed variant frequency spectrum (Tajima's $D = +0.84$) and was one of the two loci for which we could not detect significant decay in LD along its length. These patterns suggest that *Csp5-spc-1*, and its genomic neighbourhood, might experience stronger subpopulation differentiation than other loci. The protein sequence for *Csp5-spc-1* is by far the most highly constrained of the loci in our sample, based on low interspecific divergence, and indeed we observed no replacement-site polymorphisms within *C. sp. 5*. Notably, the ortholog of this locus in *C. remanei* also stood out in population genetic analysis (Cutter 2008b), albeit in a qualitatively different manner (significant negative skew of the variant frequency spectrum with $D = -1.9$, with below-average synonymous-site polymorphism). In *C. elegans*, the alpha spectrin SPC-1 is a long 2427 amino acid protein involved in body morphogenesis, including body-wall muscle development, and is expressed at high levels throughout embryogenesis (Norman & Moerman 2002). Further investigation of the genomic region surrounding *Csp5-spc-1* in a more extensive collection of population samples from different localities would help determine the causes of the unusual patterns of polymorphism at this locus.

Despite the streamlined morphology that is largely devoid of obvious differences among *Caenorhabditis* species, it is becoming clear that individual species harbour enormous genetic variation at the molecular level—in addition to the well-known high molecular divergence between species (Thomas & Wilson 1991; Stein *et al.* 2003; Cutter 2008a). This molecular population genetic biodiversity provides unprecedented opportunity to combine the laboratory experimental tractability of these organisms to characterize organismal function and evolution at the genetic level. In particular, 'SNP shadowing'—an intraspecific analog to 'phylogenetic footprinting'—should prove especially valuable in *C. sp. 5* and related outbreeding taxa for characterizing the evolution and function of regulatory and other sequence features as a precursor to further experimental work (Boffelli *et al.* 2004).

Acknowledgements

We are grateful to Marie-Anne Felix and Josselin Milloz for collecting and sharing strains with us. We thank Cecilia Chan, Alivia Dey, Richard Jovelin and Joanna Kozłowska for discussion and assistance. This research was supported by the National Natural Science Foundation of China (NSFC #30870350, 31071998) and HuaZhong Normal University (G.-X.W.) and by the Natural Sciences and Engineering Research Council of Canada (A.D.C.).

References

- Barrière A, Félix M-A (2006) Isolation of *C. elegans* and related nematodes. In: *WormBook* (ed. The *C. elegans* Research Community), <http://www.wormbook.org>.
- Barrière A, Félix M-A (2007) Temporal dynamics and linkage disequilibrium in natural *Caenorhabditis elegans* populations. *Genetics*, **176**, 999–1011.
- Boffelli D, Weer CV, Weng L *et al.* (2004) Intraspecies sequence comparisons for annotating genomes. *Genome Research*, **14**, 2406–2411.
- Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics*, **129**, 897–907.
- Cutter AD (2008a) Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. *Molecular Biology and Evolution*, **25**, 778–786.
- Cutter AD (2008b) Multilocus patterns of polymorphism and selection across the X-chromosome of *Caenorhabditis remanei*. *Genetics*, **178**, 1661–1672.
- Cutter AD (2008c) Reproductive evolution: symptom of a selfing syndrome. *Current Biology*, **18**, R1056–R1058.
- Cutter AD, Baird SE, Charlesworth D (2006) High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. *Genetics*, **174**, 901–913.
- Cutter AD, Wasmuth JD, Washington NL (2008) Patterns of molecular evolution in *Caenorhabditis* preclude ancient origins of selfing. *Genetics*, **178**, 2093–2104.
- Cutter AD, Dey A, Murray RL (2009) Evolution of the *Caenorhabditis elegans* genome. *Molecular Biology and Evolution*, **26**, 1199–1234.
- Denver DR, Dolan PC, Wilhelm LJ *et al.* (2009) A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proceedings of the National Academy of Sciences, USA*, **106**, 16310–16324.
- Dolgin ES, Félix M-A, Cutter AD (2008) Hakuna nematoda: genetic and phenotypic diversity in African isolates of *Caenorhabditis elegans* and *C. briggsae*. *Heredity*, **100**, 304–315.
- Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, **11**, 725–736.
- Graustein A, Gaspar JM, Walters JR, Palopoli MF (2002) Levels of DNA polymorphism vary with mating system in the nematode genus *Caenorhabditis*. *Genetics*, **161**, 99–107.
- Haag ES (2009) *Caenorhabditis* nematodes as a model for the adaptive evolution of germ cells. *Current Topics in Developmental Biology*, **86**, 43–66.
- Jovelin R (2009) Rapid sequence evolution of transcription factors controlling neuron differentiation in *Caenorhabditis*. *Molecular Biology and Evolution*, **26**, 2373–2386.
- Jovelin R, Ajie BC, Phillips PC (2003) Molecular evolution and quantitative variation for chemosensory behaviour in the nematode genus *Caenorhabditis*. *Molecular Ecology*, **12**, 1325–1337.
- Jovelin R, Dunham JP, Sung FS, Phillips PC (2009) High nucleotide divergence in developmental regulatory genes contrasts with the structural elements of olfactory pathways in *Caenorhabditis*. *Genetics*, **181**, 1387–1397.
- Kammenga JE, Phillips PC, De Bono M, Doroszuk A (2008) Beyond induced mutants: using worms to study natural

- variation in genetic pathways. *Trends in Genetics*, **24**, 178–185.
- Kiontke K, Fitch DHA (2005) The phylogenetic relationships of *Caenorhabditis* and other rhabditids. In: *WormBook* (ed. The C. elegans Research Community), <http://www.wormbook.org>.
- Kiontke K, Sudhaus W (2006) Ecology of *Caenorhabditis* species. In: *WormBook* (ed. The C. elegans Research Community), <http://www.wormbook.org>.
- Kiontke K, Barriere A, Kolotuev I *et al.* (2007) Trends, stasis, and drift in the evolution of nematode vulva development. *Current Biology*, **17**, 1925–1937.
- Kreitman M (2000) Methods to detect selection in populations with applications to the human. *Annual Review of Genomics and Human Genetics*, **1**, 539–559.
- LaMunyon CW, Ward S (1999) Evolution of sperm size in nematodes: Sperm competition favours larger sperm. *Proceedings of the Royal Society of London Series B—Biological Sciences*, **266**, 263–267.
- Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.
- McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, **160**, 1231–1241.
- Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, **3**, 418–426.
- Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.
- Norman KR, Moerman DG (2002) Alpha spectrin is essential for morphogenesis and body wall muscle formation in *Caenorhabditis elegans*. *Journal of Cell Biology*, **157**, 665–677.
- Phillips N, Salomon M, Custer A, Ostrow D, Baer CF (2009) Spontaneous mutational and standing genetic (co)variation at dinucleotide microsatellites in *Caenorhabditis briggsae* and *Caenorhabditis elegans*. *Molecular Biology and Evolution*, **26**, 659–669.
- Sivasundar A, Hey J (2003) Population genetics of *Caenorhabditis elegans*: the paradox of low polymorphism in a widespread species. *Genetics*, **163**, 147–157.
- Small KS, Brudno M, Hill MM, Sidow A (2007) Extreme genomic variation in a natural population. *Proceedings of the National Academy of Sciences, USA*, **104**, 5698–5703.
- Stein LD, Bao Z, Blasiar D *et al.* (2003) The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biology*, **1**, 166–192.
- Sudhaus W, Kiontke K (2007) Comparison of the cryptic nematode species *Caenorhabditis brenneri* sp. n. and *C. remanei* (Nematoda: Rhabditidae) with the stem species pattern of the *Caenorhabditis Elegans* group. *Zootaxa*, **1456**, 45–62.
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
- Thomas WK, Wilson AC (1991) Mode and tempo of molecular evolution in the nematode *Caenorhabditis*: cytochrome oxidase-II and calmodulin sequences. *Genetics*, **128**, 269–279.
- Wakeley J, Aliacar N (2001) Gene genealogies in a metapopulation. *Genetics*, **159**, 893–905.
- Weir BS, Hill WG (1980) Effect of mating structure on variation in linkage disequilibrium. *Genetics*, **95**, 477–488.
- Weir BS, Hill WG (1986) Nonuniform recombination within the human beta-globin gene cluster. *American Journal of Human Genetics*, **38**, 776–778.
- Zhang Z, Li J, Zhao XQ *et al.* (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics*, **4**, 259–263.

Research in Asher Cutter's laboratory addresses genome evolution and the evolution of breeding systems, reproductive isolation, and adaptation in nematodes, focusing on population genetic and molecular evolutionary approaches. Guo-xiu Wang's lab studies population genetics and molecular evolution in *Caenorhabditis* and entomopathogenic Mermithid nematodes. Shuang Ren and Yuan Ren are graduate researchers working with Guo-xiu Wang. Hui Ai's research interests include studying the biodiversity and phylogenetics of nematodes in China.
