2009

Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes

Antoine Barrière
The University of Chicago

Shiaw-Pyng Yang
Washington University in St Louis

Elizabeth Pekarek
The University of Chicago

Cristel G. Thomas
University of Maryland - College Park

Eric S. Haag
University of Maryland - College Park

See next page for additional authors

Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes

Antoine Barrière, Shiaw-Pyng Yang, Elizabeth Pekarek, et al.

Genome Res. 2009 19: 470-480 originally published online February 9, 2009
Access the most recent version at doi:10.1101/gr.081851.108

Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes

Antoine Barrière,[1] Shiaw-Pyng Yang,[2] Elizabeth Pekarek,[1] Cristel G. Thomas,[3] Eric S. Haag,[3,4] and Ilya Ruvinsky[1,4]

[1]*Department of Ecology and Evolution and Institute for Genomics and Systems Biology, The University of Chicago, Chicago, Illinois 60637, USA;* [2]*Genome Sequencing Center, Washington University, St. Louis, Missouri 63108, USA;* [3]*Department of Biology and Molecular and Cell Biology Program, University of Maryland, College Park, Maryland 20742, USA*

The majority of nematodes are gonochoristic (dioecious) with distinct male and female sexes, but the best-studied species, *Caenorhabditis elegans*, is a self-fertile hermaphrodite. The sequencing of the genomes of *C. elegans* and a second hermaphrodite, *C. briggsae*, was facilitated in part by the low amount of natural heterozygosity, which typifies selfing species. Ongoing genome projects for gonochoristic *Caenorhabditis* species seek to approximate this condition by intense inbreeding prior to sequencing. Here we show that despite this inbreeding, the heterozygous fraction of the whole genome shotgun assemblies of three gonochoristic *Caenorhabditis* species, *C. brenneri*, *C. remanei*, and *C. japonica*, is considerable. We first demonstrate experimentally that independently assembled sequence variants in *C. remanei* and *C. brenneri* are allelic. We then present gene-based approaches for recognizing heterozygous regions of WGS assemblies. We also develop a simple method for quantifying heterozygosity that can be applied to assemblies lacking gene annotations. Consistently we find that ~10% and 30% of the *C. remanei* and *C. brenneri* genomes, respectively, are represented by two alleles in the assemblies. Heterozygosity is restricted to autosomes and its retention is accompanied by substantial inbreeding depression, suggesting that it is caused by multiple recessive deleterious alleles and not merely by chance. Both the overall amount and chromosomal distribution of heterozygous DNA is highly variable between assemblies of close relatives produced by identical methodologies, and allele frequencies have continued to change after strains were sequenced. Our results highlight the impact of mating systems on genome sequencing projects.

[Supplemental material is available online at http://www.genome.org.]

Whereas originally large genomes were sequenced using minimal tiling paths of genomic DNA clones (*C. elegans* Sequencing Consortium 1998), more recently the whole-genome shotgun (WGS) method has greatly expedited the sequencing pipeline. An important step in this approach is the post-sequencing assembly of relatively short sequence reads. Because assembly critically relies on finding perfect (or nearly perfect) overlaps between individual reads, this methodology is most efficient for species with haploid or nearly homozygous diploid genomes with relatively few repetitive sequences. Indeed the first genomes sequenced using the WGS approach contained little heterozygosity because they were derived from highly inbred laboratory strains (e.g., mouse [Mouse Genome Sequencing Consortium 2002]) or selfing hermaphroditic species (e.g., *Caenorhabditis briggsae* [Stein et al. 2003]). Even the genomes of non-inbred, outcrossing species (e.g., human) can be assembled, although with greater difficulty, provided they have relatively low genetic diversity (Venter et al. 2001; Istrail et al. 2004; Levy et al. 2007).

Reconstructing genomes from WGS reads proves more challenging for highly polymorphic species. If DNA from a single outbred individual provides sufficient material for sequencing, then each haplotype has the same read coverage and can be assembled independently (e.g., two species of ascidian *Ciona* [Vinson et al. 2005; Kim et al. 2007; Small et al. 2007] and the sea urchin *Strongylocentrotus purpuratus* [Sea Urchin Genome Sequencing Consortium 2006]). But WGS sequencing is now being applied to genomes from small outcrossing organisms with large natural effective population sizes and considerable genetic diversity (Richards et al. 2005; *Drosophila* 12 Genome Consortium 2007; Ghedin et al. 2007), and additional approaches may be required. In particular, genomes that are a mosaic of homozygous and heterozygous regions are particularly challenging, as paralogous genes and alleles can be easily confused in fragmented WGS assemblies, especially when the former are recent duplications and the latter are highly differentiated.

The nematode *Caenorhabditis elegans* was chosen for genetic study in part for its mode of reproduction, which combines self-fertile hermaphrodites and facultative cross-fertile males (Brenner 1974). Vigorous, completely homozygous strains of hermaphroditic nematodes are easily obtained, and this facilitated the sequencing and assembly of the genomes of both *C. elegans* (*C. elegans* Sequencing Consortium 1998) and *C. briggsae* (Stein et al. 2003), a closely related selfing species. However, *C. elegans* and *C. briggsae* independently evolved selfing from gonochoristic ancestors, which had distinct male and female sexes (Kiontke et al. 2004; Nayak et al. 2005; Hill et al. 2006), and the majority of extant species in the family Rhabditidae retain this ancestral mating system (Kiontke et al. 2007). The Washington University Genome Sequencing Center is currently in various stages of finishing the WGS assemblies of three gonochoristic *Caenorhabditis* species: *C. remanei*, *C.*

[4]Corresponding authors.
E-mail ruvinsky@uchicago.edu; fax (773) 702-9740.
E-mail ehaag@umd.edu; fax (301) 314-9358.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.081851.108.

19:470–480 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

brenneri, and *C. japonica*. Although anatomically very similar to their selfing relatives, they are known to have dramatically higher levels of genetic variation (Graustein et al. 2002; Cutter and Payseur 2003; Haag and Ackerman 2005; Cutter et al. 2006). Therefore, we sought to analyze the effects of allogamic sexual reproduction on the sequences and WGS assemblies of otherwise similar genomes of gonochoristic *Caenorhabditis* nematodes.

Results

The expected amount of retained heterozygosity following inbreeding

Because natural isolates of *C. brenneri*, *C. remanei*, and *C. japonica* were known to harbor high levels of genetic variation, strains selected for sequencing were first inbred. As attempts to establish inbred lines by strict sibling (sib) mating invariably failed (S. Baird, pers. comm.), a modified isofemale-line approach was used. Each generation was founded by a single gravid female that mated with one or more males; therefore, all individuals were full or, if multiple matings occurred, half-siblings. Assuming extreme cases of exclusive full or half-sib inbreeding and no selection, we can compute the expected range of retained heterozygosity. For full sibs, the extent of expected heterozygosity (h_t) is given by $h_t \approx 1.17h_0\,(0.809)^t$, when $t \gg 1$, where h_0 is the initial heterozygosity and t is the number of generations of inbreeding (Nagylaki 1992). For half-sibs, $h_t \approx 1.106h_0\,(0.890)^t$. After 20 generations of inbreeding, h_t is ~1.7% of h_0 in the full-sib case and 10.8% in the half-sib case.

Many loci in the sequenced genomes of *C. remanei* and *C. brenneri* are represented by two alleles

From the individual reads of the WGS sequences of the *C. brenneri* genome (not yet assembled at the time), we manually assembled the orthologs of 23 *C. elegans* genes (Table 1; Supplemental Table S1). Among these, nine assembled as two distinct sequence variants. Whereas some of these variants displayed considerable divergence at the nucleotide level, most were over 99% identical at the amino acid level (Supplemental Table S1). Gene models and depths of sequence coverage for *C. brenneri* genes are presented in Supplemental Figure S1. Because nearly 40% of investigated *C. brenneri* loci were found to be dimorphic (i.e., represented by two similar sequence variants), we sought to confirm these observations by applying the same methodology to other gonochoristic nematodes. Indeed, two independent contigs were assembled from the NCBI sequence traces for a number of *C. remanei* and *C. japonica* genes (Table 2; Supplemental Table S1). We also found that three *C. remanei* genes (*fem-1, fem-3,* and *tra-3*) exist as two distinct variants in the assembled sequence.

The pairs of independently assembled variants homologous to single-copy *C. elegans* genes could represent either genuine copy-number differences or residual heterozygosity that was retained despite inbreeding prior to sequencing. Three lines of experimental evidence indicate that most of these variants are indeed alleles. First, all three possible genotypes were observed in individual nematodes for loci tested with variant-specific genotyping assays (Table 3). Second, for two loci in *C. remanei* and one locus in *C. brenneri*, the two forms segregated in genetic crosses (Table 4). Third, we found considerable changes in the relative frequencies of variants between the inbred sequenced strains and their pre-inbreeding ancestors in both *C. brenneri* and *C. remanei* (see below). These results provide experimental confirmation that many sequence variants from nematode WGS assemblies represent alternative alleles rather than paralogs.

Table 1. Manually assembled *C. brenneri* homologs of *C. elegans* queries

		Diversity		Genome position	
Gene name	Variants	Exon	Noncoding	*C. elegans*	*C. briggsae*
lin-17A	2	0.024	0.118	I: 2.7	I: 4.9
lin-17B	2	NA	0.035	NA	NA
fog-1	2	0.017	0.074	I: 3.2	I: 3.2
ric-19	2	0.014	0.086	I: 3.8	I: 1.5
unc-11	2	0.003	0.046	I: 3.8	I: 1.5
ksr-2	1			I: 12.1	I: 10.1
sur-2	2	0.033	0.259	I: 14.8	I: 10.9
F10E7.9	2	0.014	0.031	II: 7.1	II: 0.5
ulp-4	2	0	0.064	II: 8.1	II: 3.1
let-23	1			II: 9.2	II: 0.7
lin-7	2	0.009	0.1	II: 14.3	II: 0.3
acr-20	2	0 (0)	0.011	II: 14.4	II: 0.2
mes-2	2	0.007	0.017	II: 14.4	II: 0.3
par-2	1			III: 1.1	III: 12.2
oig-1	1			III: 6.5	III: 1.4
unc-47	1			III: 10.2	III: 4.4
unc-119	1			III: 10.9	III: 8.1
unc-25	2	0.047	0.377	III: 12.9	III: 0.8
ama-1	1			IV: 4.3	IV: 13.5
flp-13	2	0.027	0.072	IV: 7.7	IV: 5.9
C55F2.2	2	0.014	0.147	IV: 7.9	IV: 10.8
C04G2.1	2	0.028	0.046	IV: 10.1	IV: 7.8
mec-3	1			IV: 10.5	IV: 12.3
lin-3	2	0.006	0.015	IV: 11.1	IV: 5.7
bcc-1	2	0.012	0.03	IV: 11.1	IV: 5.7
let-99	1			IV: 12.6	III: 5.4
lag-2	1			V: 3.2	V: 2.6
unc-46	2	0	0.046	V: 5.1	V: 5.2
snb-1	1			V: 6.7	V: 9.9
unc-42	2	0.020	0.067	V: 9.8	V: 11.3
sel-10	1			V: 13.8	V: 0.2
ssu-1	1			V: 20.1	V: 15.2
sli-1	1			X: 0.7	X: 0.3
gap-1	1			X: 2.2	X: 7.7
lin-18	1			X: 4.0	X: 8.9
cdf-1	1			X: 6.5	X: 11.8
sng-1	1			X: 7.3	X: 12.7
unc-18	1			X: 7.7	X: 13.1
unc-115	1			X: 10.1	X: 7.2
alr-1	1			X: 11.1	X: 5.7
unc-84	1			X: 13.6	X: 19.2
unc-7	1			X: 15.1	X: 18.2
mec-4	1			X: 16.8	X: 1.5

Chromosomal distribution of heterozygous loci in *C. brenneri* and *C. remanei*

Because our initial sample of genes indicated that potentially large amounts of the gonochoristic nematode assemblies were heterozygous, we sought to verify that heterozygosity was in fact general. At the time, only the *C. remanei* assembly had an associated set of gene predictions, and genetic linkage maps are not available for *C. brenneri* and *C. remanei*. However, all *Caenorhabditis* nematodes studied so far have five pairs of autosomes and either one (in males) or two (in females/hermaphrodites) X chromosomes (Baird 2002; Hillier et al. 2007). Further, both microsynteny (i.e., that two genes closely linked in one species are also linked in another) and chromosomal synteny (i.e., that two genes on the same chromosome in one species are also found on the same chromosome in another species) are high between *C. elegans* and *C. briggsae* (Kuwabara and Shah 1994; Haag and Kimble 2000; Haag et al. 2002; Hillier et al. 2007). Because the phylogenetic distance between these two species is the same as that between *C. elegans* and *C. brenneri* (Kiontke et al.

Table 2. Manually assembled *C. remanei* and *C. japonica* homologs of *C. elegans* queries

		Diversity		Genome position	
Gene name	Variants	Exon	Noncoding	*C. elegans*	*C. briggsae*
C. remanei					
sur-2	1			I: 14.8	I: 10.9
tra-2	1			II: 7.0	II: 5.1
acr-14	1			II: 8.2	II: 3.3
par-2	1			III: 1.1	III: 12.2
lin-12	1			III: 9.1	III: 8.7
glp-1	1			III: 9.1	III: 8.7
ama-1	1			IV: 4.3	IV: 13.5
flp-13	2	0.006	0.014	IV: 7.7	IV: 5.9
C55F2.2	2	0.008	0.011	IV: 7.9	IV: 10.8
C04G2.1	1			IV: 10.1	IV: 7.8
lin-3	2	0.002	0.008	IV: 11.1	IV: 5.7
bcc-1	2	0.009	0.019	IV: 11.1	IV: 5.7
odr-3	1			V: 13.2	V: 8.8
C. japonica					
sur-2	2	0.047	0.117	I: 14.8	I: 10.9
acr-14	2			II: 8.2	II: 3.3
par-2	1			III: 1.1	III: 12.2
lin-12	1			III: 9.1	III: 8.7
glp-1	1			III: 9.1	III: 8.7
ama-1	1			IV: 4.3	IV: 13.5
flp-13	1			IV: 7.7	IV: 5.9
C55F2.2	1			IV: 7.9	IV: 10.8
C04G2.1	1			IV: 10.1	IV: 7.8
lin-3	1			IV: 11.1	IV: 5.7
lag-2	1			V: 3.2	V: 2.6
odr-3	1			V: 13.2	V: 8.8

2004), we used the *C. elegans* linkage map as a reasonable proxy for the genomes of both *C. brenneri* and *C. remanei*.

The 23 original *C. brenneri* loci described above reside on all six chromosomes in *C. elegans*, but heterozygous loci were found only on chromosomes I, II, III, and IV, and absent from V and X. We also noticed two pairs of closely linked heterozygous loci—*lin-17*/*ric-19* and *flp-13*/C55F2.2. To confirm these observations, we examined 19 additional genes (SOM) (Fig. 1; Table 1; Supplemental Table S1). Of these, seven that were predicted to be in close proximity to heterozygous loci were all represented by two contigs. We next selected 12 genes from previously unsampled regions of chromosomes IV, V, and X. Of these, *unc-46 IV* and *unc-42 V* were heterozygous, but none of the X-linked loci were. The fraction of autosomal loci (among the genes selected without regard to specific chromosomal position) that show evidence of heterozygosity is ~0.45. Given that none of the 11 X-linked genes showed evidence of two allelic variants, it is highly unlikely ($p \approx 0.55^{11} = 1.4 \times 10^{-3}$) that the X chromosome has the same density of heterozygous genes as do the autosomes.

Because gene predictions of the preliminary assembly of the *C. remanei* genome were available from WormBase (http://ftp.wormbase.org), we developed a gene-based method to systematically scan the assembly for likely heterozygous regions. Our approach began by first defining a set of 14,530 single-copy genes in *C. elegans* (62% of the total). We then used these to query the predicted *C. remanei* proteome via BLASTP searches. Best hits that identified a second, very similar protein prediction in *C. remanei* (see Methods for details) were retained and further filtered to remove pairs with low sequence identity and those that resided on the same contig and thus likely represented tandem duplicates. Despite these filters some bona fide recently duplicated paralogs may not be distinguished from alleles based on their sequences alone. However, separately assembled alleles should differ from paralogs in having approximately half the WGS read coverage (or "depth"), because the total read depth of a heterozygous locus is distributed between two different alleles. Thus, dimorphic gene predictions that represent allelic variants should have lower read depth than those representing homozygous paralogous genes. Finally, true heterozygosity should exist in blocks defined by recombination. The correspondence between regions of high dimorphic gene density and low WGS read depth thus provides a simple computational assay for allelism (Fig. 2).

The above analysis indicates that relatively few (6–29) genes on chromosomes I, II, III, V, and X are heterozygous in the *C. remanei* assembly. In contrast, approximately three-quarters of the total number of dimorphic loci (414) lie in the central part of chromosome IV (Supplemental Table S2). This approach was validated by the recovery of all dimorphic genes discovered in manual analyses. As predicted, a clear depression of read depth was seen in most of the dimorphic gene predictions matching *C. elegans* queries in this region, as well as in smaller regions of LGI and LGV that also showed high density of dimorphic genes (Fig. 2; Supplemental Table S2). This phenomenon is also observed for manually assembled *C. brenneri* genes (Supplemental Figs. S1, S2), further supporting our inference of heterozygosity in that species.

The high concentration of dimorphic genes on the *C. remanei* LGIV suggests that most of it was heterozygous in the inbred strain PB4641 at the time of sequencing. The specific region involved, that homologous to the central 12 Mb of *C. elegans* chromosome IV, comprises roughly 10% of the genome. The extreme apparent heterozygosity of PB4641 LGIV is not due to a general suppression of recombination, however. Seventeen of 382 chromosomes scored for *Cr-fem-1* and *Cr-fem-3* were recombinant, giving a map distance of roughly 4.4 cM. This is similar to the 2.2 cM distance between their *C. elegans* orthologs, which lie near the center of LGIV.

Species differences in degree of allelic divergence

We observed that some allele pairs differed by single nucleotide polymorphisms (SNPs), while others had insertion/deletion (indel) differences of up to 260 bp and nonalignable regions (Supplemental Fig. S3; Supplemental Table S1). We estimated the average extent of nucleotide diversity in the coding regions of *C. brenneri* and *C. remanei* as 1.8×10^{-2} and 0.6×10^{-2}, respectively, the vast majority of it being synonymous (Supplemental Table S1), whereas diversity in the noncoding regions is as much as an order of magnitude higher (Table 1; Supplemental Table S1). These estimates are over 20-fold higher than those for the human genome (Sachidanandam et al. 2001), and are comparable to those of *Drosophila* species (Begun et al. 2007).

We found four loci (*lin-3*, *flp-13*, C55F2.2, and *bcc-1*) whose orthologs are represented by two alleles in both *C. brenneri* and

Table 3. Genotypes of randomly picked individuals from *C. brenneri* PB2801 and *C. remanei* PB4641

Species	Gene	AA	AB	BB
C. brenneri	*fog-1*	38	0	0
C. brenneri	*sur-2*	5	14	6
C. brenneri	*lin-7*	39	6	1
C. remanei	*fem-1*	22	44	26
C. remanei	*fem-3*	30	39	15

Table 4. Segregation of sequence variants in genetic crosses using the sequenced strains of *C. remanei* and *C. brenneri*

	Parental genotypes	AA progeny	AB progeny	BB progeny
Cr-fem-1	AB × AB	7	18	3
(C. remanei)	AB × BB	0	19	13
	AA × BB	0	24	0
Cr-fem-3	AB × AB	10	26	9
(C. remanei)	AA × AB	28	33	0
Cbn-sur-2	AA × BB	0	30	0
(C. brenneri)	AA × AB	18	12	0
	AB × AB	5	19	7

C. remanei. In all four cases allelic divergence was consistently higher in *C. brenneri* (Tables 1, 2; Supplemental Table S1). For example, when only coding sequences were considered, pairs of *C. brenneri* alleles were 1.3- to 4.5-fold more divergent than their *C. remanei* counterparts. Interestingly, we find that the sequenced strain of *C. brenneri* retained more heterozygosity than its *C. remanei* counterpart (see above and below), even though both have been subjected to similar inbreeding schemes prior to sequencing.

Estimating the overall extent of retained heterozygosity without gene predictions

When we performed these analyses, only the *C. remanei* WGS assembly had a set of gene predictions associated with it. However, our manual searches suggested that the *C. brenneri* assembly had an extraordinarily high degree of heterozygosity. To confirm this on a genome-wide basis, we developed an alternative to the gene-based method described above for *C. remanei*. In this method, we estimated the copy number for every position of the gonochoristic WGS assemblies, for the WGS *C. briggsae* assembly, and for the completed, heterozygosity-free genome of *C. elegans*. These values were then converted into assembly-wide proportions of bases present in each copy-number category. For hermaphroditic species, all multicopy DNA should stem from paralogy, whereas for gonochoristic species it will represent a mixture of paralogy and independently assembled alleles. The largest effect of retained alleles is expected to be the elevation of the fraction of the assembly present in two copies relative to a completely homozygous genome. However, the size of all multi-copy classes will be inflated beyond their true (e.g., haploid or homozygous genome) values to some extent.

The fractions of bases in the various assemblies that are present in two through six copies are shown in Figure 3. As expected, the two hermaphroditic species (*C. elegans* and *C. briggsae*) have low multicopy content, almost all of which must reflect paralogy, whereas the three gonochoristic species have elevated fractions of multicopy sequences, reflecting an input of independently assembled alleles. However, the amount of two-copy DNA inferred from this analysis varies more than fivefold in the gonochoristic species, with *C. brenneri* having the largest fraction at 33%, *C. japonica* lowest at 6%, and *C. remanei* lying between the two. These values are consistent with our previous inference that WGS sequences of *C. brenneri* have an especially large number of independently assembled alleles, while the other two species having retained less heterozygosity.

As it would be useful to be able to estimate the amount of heterozygosity in a WGS assembly without relying on gene predictions, we investigated how copy number distributions, like those in Figure 3, could be used for this purpose. Because gonochoristic WGS assemblies contain both paralogous DNA and two copies of independently assembled alleles, the total fraction of the genome assembly composed of sequence found in two copies, d_2, is

$$d_2 = \frac{(2h_2 + p_2)G}{A} \quad (1)$$

In the above equation, h_2 is the fraction of the single-copy portion of the genome that is heterozygous, p_2 is the proportion of the genome that is two-copy due to paralogy, G is the genome size, and A is the WGS assembly size.

We further note that not all bases of heterozygous regions will be scored as being present in two copies when alleles are highly differentiated (in which case apparent copy number, d, will be one) or contain repetitive elements ($d > 2$). In recognition of this possibility, the contribution of h_2 to d_2 can be corrected by multiplying h_2 by f, the fraction of heterozygous bases that are recognized as two copy. Doing this and solving for h_2, we obtain

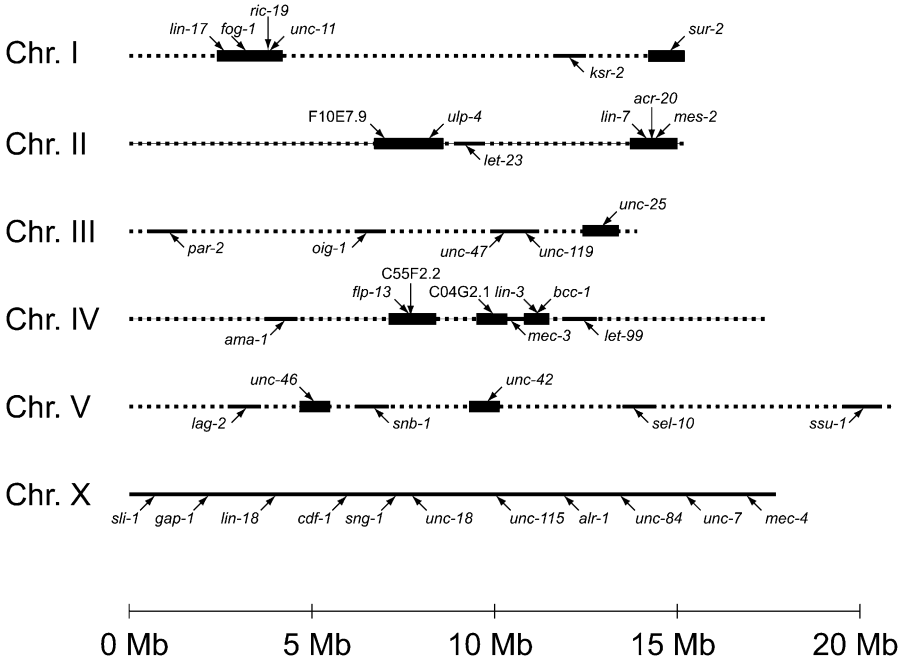

Figure 1. Genomic distribution of inferred heterozygous (thick lines) and homozygous (thin solid lines) regions in the sequenced *C. brenneri* genome. Chromosomal locations are assigned based on positions of *C. elegans* orthologs. Genes represented by two alleles are shown *above* chromosomes; homozygous loci are shown *below* chromosomes. Regions of currently undetermined status are represented by dashed lines.

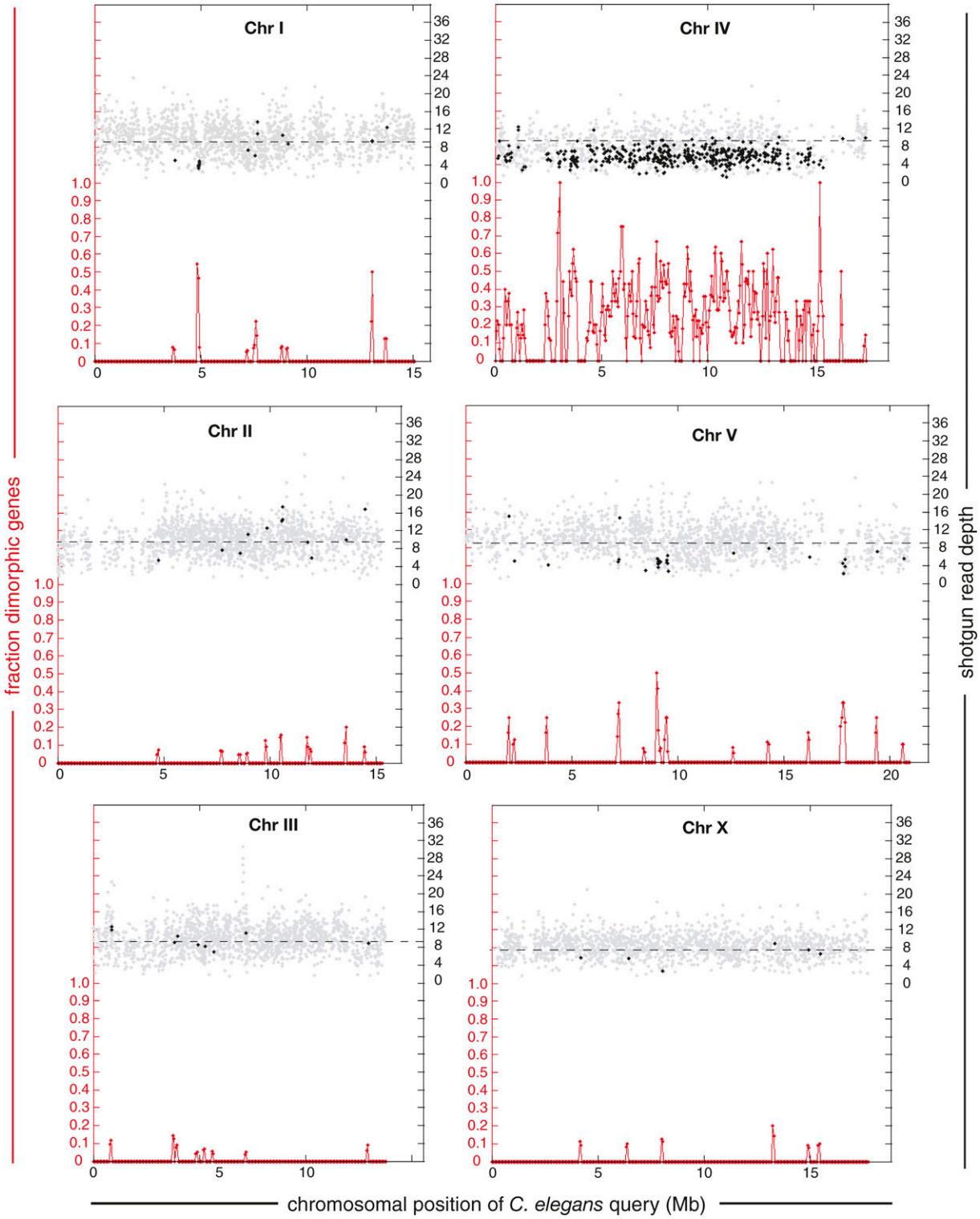

Figure 2. Gene-based, genome-wide survey for heterozygosity in the preliminary *C. remanei* assembly Cr01. 10,322 single-copy *C. elegans* genes were used to query the assembly. The fraction of total queries that identified two distinct yet highly similar gene predictions within a 100-kb sliding window (with 50-kb steps) along the *C. elegans* chromosome is plotted at the *bottom* of each panel. *Left* scales (red) refer only to these values. The *upper* portion of each panel depicts the WGS read depth for queries that have an apparent singleton *C. remanei* homolog (gray diamonds), and the mean depth for doublet homologs (black diamonds). *Right* scales (black) refer only to these values. The small proportion of queries identifying more than two variants are not shown in the depth analysis. Regions in which doublet homologs occur in clusters and have consistently low mean read depth are inferred to be heterozygous. By this criterion, regions of the *C. remanei* genome that are syntenic with *C. elegans* LGI at 5 Mb, LGV at 9 and 18 Mb, and nearly all of LGIV are heterozygous. The mean WGS read depth for the singleton homologs in each query chromosome is plotted with a dashed line. The singleton read depth for chromosome X (8.23×) lies between the genome-wide 9.2× and the 6.9× expected for equal sex ratio, likely due to the substantially smaller size and genome copy number of males relative to females.

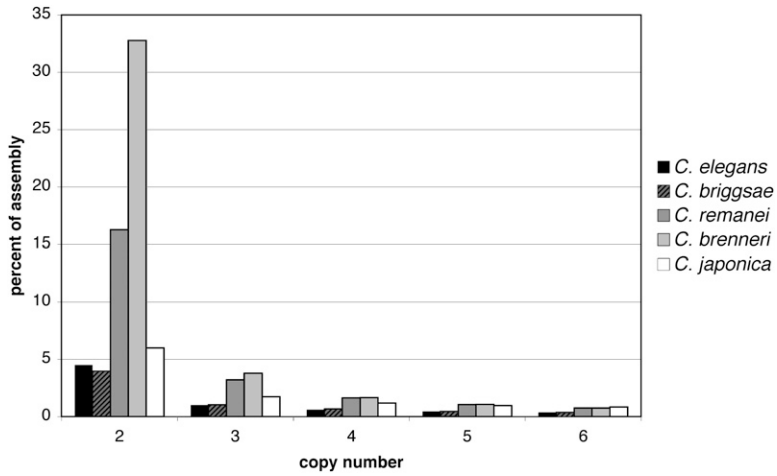

Figure 3. Estimated copy number distributions for five genome assemblies. For each species, a sliding query window of 1000 bp with 500 bp steps was used to identify nonself matches in the assembly. The percentages reported are relative to the size of the total assembly, not to an inferred actual genome size. For the hermaphroditic *C. elegans* (sequenced by a minimum clone tiling path method) and *C. briggsae* (sequenced by WGS), all sequences with copy number of two or more likely represent true copy number variation in the genome. For the three gonochoristic species, however, each bin potentially represents a mix between truly paralogous DNA and retained alleles. As the apparent single-copy sequence is at least 55% in all assemblies, the majority of the unrecognized alleles are expected to lie in the two-copy category. Single-copy DNA is not shown.

the following expression for converting the fraction of the WGS assembly that is computationally recognizable as being present in two copies (d_2) into a fraction of the single-copy portion of the genome that is present in heterozygous form (h_2):

$$h_2 = \frac{d_2 A}{2fG} - \frac{p_2}{2f} \quad (2)$$

If all copy number classes are affected equally by heterozygosity, h_2 is also a reasonable estimate of genome-wide heterozygosity. In principle, estimates of the contribution of heterozygosity to higher-copy classes could also be made, but they are complicated by the possibility of only some members of a gene family being heterozygous and weakened by the small numbers relative to the two-copy case.

In practice, evaluation of Equation 2 relies upon a mixture of known parameter values and estimates. For *C. remanei*, d_2 is 0.163 (Fig. 3), A is 145 Mb, and G is estimated by DNA fluorescence as ~131 Mb (J.S. Johnston, pers. comm.). To estimate f, we first applied the same method presented in Figure 3 to the set of inferred alleles in LGIV (Fig. 2). This analysis found that 80% of the bases in allelic genes were scored as being present in exactly two copies. As expected, this discrepancy is due to the presence of both diverged regions and repetitive elements that are found in at least one other site in the genome (data not shown). The estimate for f, then, is 0.80. The amount of true paralogy, p_2, is unknown for *C. remanei*, but the *C. elegans* and *C. briggsae* d_2 values, which are entirely dependent upon paralogs, are similar to each other. The average of these values, 0.04, is thus a reasonable estimate of the *C. remanei* p_2. With these parameter values, h_2 for the *C. remanei* assembly is 8.8%. This is similar to the gene-based estimate of 10% (see above). We can use a similar procedure to estimate h_2 for the *C. brenneri* assembly. Using fluorescence measurements for G (150 Mb; J.S. Johnston, pers. comm.), assembly size for A (190 Mb), and a value of f calculated from the 42-gene test set described above (0.76), h_2 is estimated to be 24.9%, which is somewhat lower than the estimate of 40%–45% derived from a manually curated set of genes (see above). It must be noted that the values of h_2 derived above estimate the heterozygous fraction of the *genome* (more accurately the fraction of nucleotides that reside in heterozygous regions), whereas manually curated searches estimated the fraction of *genes* that are heterozygous. Therefore, while the two values are expected to be correlated, they will not necessarily be the same.

Retention of heterozygosity is caused by recessive deleterious alleles

Since all strains experienced lengthy laboratory culture (and therefore decreased population size) prior to intentional inbreeding, the observed level of retained heterozygosity in *C. brenneri* is inconsistent with the expectations derived with the assumption of no selection. For. *C. remanei*, the overall value is not substantially different from the all-half-sib expectation, yet it is surprising that most of chromosome IV has retained heterozygosity (Fig. 2). We therefore considered potential selective mechanisms that may have promoted the maintenance of heterozygosity. One would be that deleterious recessive alleles segregating in the natural populations of these species (Cutter et al. 2006; Dolgin et al. 2007) were rendered homozygous by inbreeding. If the fitness effects of some of these were large, the fraction of heterozygous individuals in each subsequent generation would not decline as rapidly as would be expected in the absence of such selective pressure. Two lines of evidence support this interpretation.

First, consistent with the expectation of inbreeding depression in the sequenced strains, we found them to have significantly lower fitness (a combined measure of fecundity and egg to adult viability) than the ancestral strains from which they were derived. We measured the fitness of individual strains by counting the zygotes and subsequent viable adult progeny produced during the first 24 h after mating between randomly selected male–female pairs. By the viable-adult assay, the sequenced inbred strain of *C. brenneri* (PB2801) had suffered a 23% loss of fitness compared with its ancestral strain LKC28 (Fig. 4A). Similarly, the inbred strain of *C. remanei* (PB4641) produced 55% fewer viable progeny than a recently acquired stock of EM464, the strain from which it was derived (Fig. 4B). We also found that F1 hybrids between PB4641 and an independent strain with equally low fitness benefit from substantial hybrid vigor (Supplemental Fig. S4), indicating that inbreeding depression is caused by deleterious recessive alleles.

Second, because the X chromosome is hemizygous in *Caenorhabditis* males, it is expected to have fewer recessive deleterious alleles and therefore less retained heterozygosity following inbreeding. Indeed, all 11 X-linked loci examined in *C. brenneri* were found to be homozygous, a significantly higher fraction than that found on autosomes (Fig. 1). Similarly, the X chromosome in *C. remanei* contained the fewest loci with independently assembled variants (Fig. 2). These variants had average read depths close to

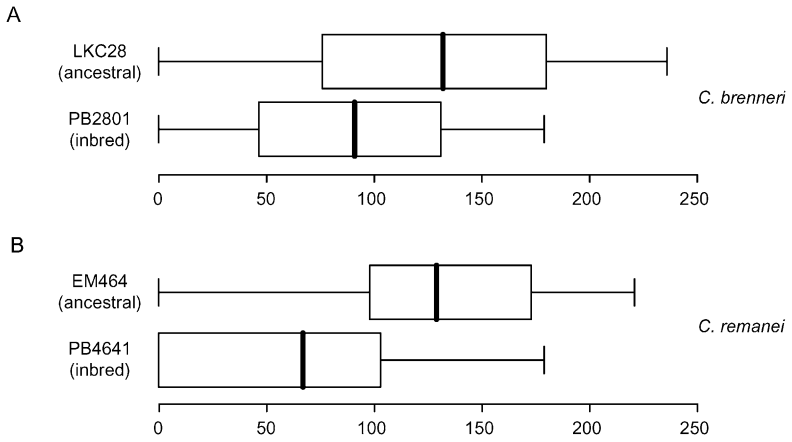

Figure 4. Persistent heterozygosity is associated with inbreeding depression. (*A*) Comparison of *C. brenneri* strains LKC28 (founder) with PB2801 (inbred for sequencing). The number of viable adults produced by PB2801 is significantly lower than those produced by LKC28 ($p < 0.04$; Kolmogorov–Smirnov test). (*B*) Comparison of *C. remanei* strains EM464 (founder) with PB4641 (inbred for sequencing). PB4641 has significantly lower fitness ($p < 0.001$, Kolmogorov–Smirnov test). Thick vertical lines indicate the median, boxes represent upper and lower quartiles, and whiskers the entirety of the distributions.

the roughly 6.9× coverage expected for homozygous X-linked loci (Fig. 2; Supplemental Table S2).

Sequenced strains are continuing to evolve

Because the sequenced strains were not completely homozygous, we expected to observe continuing changes in allele frequencies in both *C. brenneri* and *C. remanei*. Such changes may be cased by inadvertent selection during inbreeding or by genetic drift. For *C. brenneri*, we experimentally estimated allele frequencies at *fog-1* and *sur-2* loci by genotyping individuals of recently obtained stocks of the LKC28 and PB2801 strains. To estimate allele frequencies at the time of sequencing, we compared read depth coverage between the two identified alleles, assuming that it would reflect the relative frequency of alleles in the population of individuals from which the DNA was extracted. These results indicate that both alleles of *C. brenneri fog-1* are found in the wild isolate LKC28, but had changed substantially in frequency in the sequenced version of PB2801, and are now fixed for one of them in a recently obtained stock of PB2801 (Fig. 5A). The frequencies of the two alleles of *sur-2* were highly unequal in LKC28, but were nearly equal in PB2801 at the time of sequencing, and are currently 1:1 (Fig. 5B). Finally, the two alleles of *lin-7* evident from the genome sequence were at frequencies of 0.6 and 0.4 at the time of sequencing, whereas in the current strain PB2801 they are at 0.91 and 0.09. Similar shifts in allele frequencies were observed for *Cr-fem-3* in *C. remanei* (data not shown).

Finally, not only allele frequencies, but also specific phenotypes have evolved following inbreeding. In early experiments, we noticed that when *C. brenneri* females were given less than 8 h to mature after L4-to-adult molt prior to mating, the ancestral strain (LKC28) had a higher frequency of mating failure than did its inbred derivative, PB2801. When females were given less than 1 h to mature between the L4-to-adult molt and the onset of mating, 8/8 PB2801 crosses produced progeny, while only 2/8 LKC28 crosses had any offspring. It therefore appears that the females of PB2801 achieve sexual maturation faster than those of the ancestral strain. This could be a consequence of inadvertent selection for fast reproduction applied during inbreeding (Latter and Mulley 1995) and may represent one of many traits that distinguish the two strains. In the *C. brenneri* fitness tests reported here, all females were given sufficient time to mature from L4 to adult.

Discussion

Implications for *Caenorhabditis* genomics

We have presented evidence here that substantial portions of WGS assemblies of *C. brenneri* and *C. remanei* are heterozygous, despite intense inbreeding. We infer that the following sequence of events led to the observed situation. Highly polymorphic natural populations were brought into laboratory culture, resulting in partial loss of heterozygosity. During 20 generations of inbreeding prior to sequencing, some deleterious recessive alleles were rendered homozygous, which led to inbreeding depression. However, some of the deleterious alleles resided in *trans* with others in linked loci. This situation gave the heterozygotes a fitness advantage over homozygotes, and also kept linked loci lacking deleterious mutations in the heterozygous state as well. This heterozygote advantage via linkage to a deleterious allele, known as associative overdominance (Frydenberg 1963), retarded the accumulation of homozygosity for the remaining heterozygous regions. Eventual resolution of this balanced heterozygosity depends on the relative selective coefficients of the deleterious alleles in question, the distance between them, and the effects of genetic drift.

Because the DNA was extracted from populations, not individuals, the gonochoristic WGS assemblies represent complex mixtures of alleles reflecting their frequencies at the time of sequencing. Therefore, sequences representing retained alleles will not necessarily be found at a frequency of 1/2. Furthermore, because the strains were not completely homozygous at the time of sequencing, they are continuing to evolve. For example, the *fog-1* locus was biallelic in the *C. brenneri* strain PB2801 at the time of sequencing, but has since become homozygous in the version of this strain currently available from the *Caenorhabditis* Genetics Center (Fig. 5).

Our results are important for refinement of assembly and biological understanding of the *C. remanei* and *C. brenneri* genomes. Their sequences were sought in large part to facilitate the functional annotation of the *C. elegans* genome (Sternberg et al. 2003), as well as to develop *Caenorhabditis* as a system for the study of genome evolution. Fundamental to these goals is the establishment of accurate gene counts, gene orthologies, syntenic regions, and the genome size. We note that the total assembly size for *C. remanei* is 145 Mb, or about 14 Mb larger than the direct measurement based on DNA fluorescence (J.S. Johnston, pers. comm.). Our analysis indicates that all, or nearly all, of the excess sequence is due to independent assembly of alternative allelic forms, primarily on LGIV. This conclusion will produce a downward revision of the number of *C. remanei* genes and an upward adjustment of the number of genes with 1:1 *C. elegans* orthologs, with important

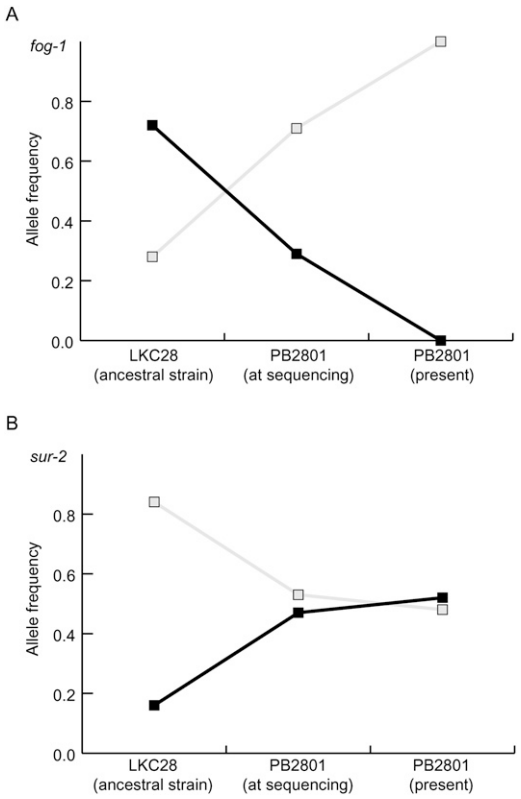

Figure 5. Changes in allele frequencies in (*A*) *fog-1* and (*B*) *sur-2* loci in *C. brenneri*. Gray represents allele "A" and black represents allele "B." Frequencies in LKC28 and "PB2801 present" were obtained by PCR genotyping of individual animals from these strains—44 and 76 for *fog-1* and 57 and 50 for *sur-2*. Allele frequencies in the strain PB2801 at the time of sequencing were inferred from the average number of sequence reads through each of the two alleles.

consequences for whole-genome multispecies comparisons. Similarly, since as much as 40% of the *C. brenneri* genome may have been heterozygous at the time of sequencing, the correct genome size may be over one-third smaller than assembly-based estimates.

Our results also indicate that experimental biologists accustomed to working with self-fertile species will need to pay particular attention to the abundant natural genetic variation that segregates in related gonochoristic species, even after lengthy laboratory culture. Not only do genotypes vary, but we have demonstrated that phenotypes as basic as reproductive fitness will vary greatly, even within the poorly defined populations known as "strains." A strain, therefore, has a more precise genetic meaning for hermaphroditic species than for gonochoristic ones. In light of these concerns, *Caenorhabditis* geneticists will benefit from establishing distinct stocks that are managed carefully to maintain genetic diversity. Such stocks are a mainstay of *Drosophila* research (Latter and Mulley 1995; Wu et al. 1995).

Limitations to computational inference of heterozygosity

We used a simple computational search to detect specific candidate heterozygous regions in the WGS assembly (Fig. 2). The key element of our approach was to identify genomic regions that have high proportions of dimorphic genes and also show a decrease in average sequence read coverage. Several features made our search conservative and therefore it likely underestimated the actual amount of retained heterozygosity in the sequenced genome of *C. remanei*. First, the set of single-copy *C. elegans* query genes used to identify potentially allelic gene predictions in *C. remanei* did not include the roughly one-third of genes that are part of families with recent duplications. Second, *C. remanei* sequences with complex orthology/paralogy relationships to their *C. elegans* counterparts were also excluded. Some of these are likely to be genuine bi-allelic genes. Third, WGS assemblies for species like the *Caenorhabditis* nematodes in this study are based on DNA samples prepared from thousands of individuals. Thus, allele frequencies at different loci can vary considerably. Rare alleles, represented by relatively few reads, may provide insufficient coverage to reconstruct the sequence of the entire locus (Supplemental Fig. S2). Finally, there is an added difficulty of distinguishing two alleles that have particularly similar sequences.

The above features of gonochoristic nematode WGS assemblies make it difficult to unequivocally identify all genes represented by two alleles with high-throughput computational methods alone. We examined the inferred heterozygous loci on *C. remanei* chromosome IV to estimate the fraction of computational "false-negatives." We found that in some regions nearly 50% of bi-allelic genes were not recognized as such in whole-genome analyses (data not shown). In all such cases however, several of their nearest neighbors were identified as heterozygous using manual searches. Because retained heterozygosity extends over closely linked loci (Fig. 1), this combined approach can be used to identify most heterozygous regions in genome sequences. We suggest that rapid computational searches like those described here can be used to identify the overall extent and likely localization of retained heterozygosity, which could be subsequently confirmed (or in some cases rejected) after more detailed computational and experimental investigation.

In addition to the gene-based method used for *C. remanei*, we also present a more generally applicable method for quantifying heterozygosity in newly assembled genomes that lack gene predictions and/or a closely related reference genome. This method combines a whole-genome estimate of sequence copy number distribution (Fig. 3) with a model for how heterozygous DNA content affects this distribution (Equations 1, 2). For *C. remanei*, the agreement between the gene-based method and the annotation-free method is good, suggesting that when heterozygosity is relatively low our simple model may provide an easy way to estimate its extent. However, application of the model to *C. brenneri* produced a value for h_2 that is substantially lower than that expected from the test set of 42 genes analyzed manually. As pointed out above, the two methods estimate different values: The former is the fraction of *nucleotides* that reside in heterozygous regions, whereas the latter is the fraction of heterozygous *genes*.

Potential causes for the above discrepancy may be sampling error in the gene-based analysis, or an unappreciated ascertainment bias imposed by the choice of genes included in the test set, most of which have developmental functions. Alternatively, the gene-based estimate may be closer to the true value, and our whole-genome computational analysis is underestimating h_2. Possible reasons for this might include underestimation of assembly size (A), overestimation of genome size (G), and/or overestimation of the extent to which heterozygous DNA can be recognized (f). A general caveat, then, is that the heterozygous fraction of a WGS assembly may become increasingly difficult to estimate as the extent of heterozygosity and the degree of allelic divergence become extreme.

Implications for other genome projects

Our results indicate that the presence of numerous deleterious recessive alleles is responsible for preserving heterozygosity in the genomes of inbred gonochoristic nematodes. Such mutations should exist in most gonochoristic species (Haag-Liautard et al. 2007), and both our work and that of others (Rumball et al. 1994; Richards et al. 2005) suggest that even after many generations, considerable heterozygosity will remain. This effect will be exacerbated if the DNA is extracted from a population. A major negative consequence of unrecognized heterozygosity is that it reduces the effective genome coverage (Fig. 2; Supplemental Fig. S2), which creates more frequent assembly gaps compared with a homozygous version of the same genome. Another is the inadvertent inflation of the apparent genome size and gene set, with detrimental effects on subsequent analysis.

One simple way to eliminate allelic variants is to relax assembly criteria so that they are placed into single contigs. However, this will have the unintended consequence of eliminating many bona fide gene duplicates. This will be particularly problematic in organisms with highly differentiated alleles, such as gonochoristic *Caenorhabditis*, as very recent (and therefore nearly identical) gene duplicates are also the most abundant (Lynch and Conery 2000). We therefore argue that strict assembly combined with post-assembly methods for recognizing heterozygosity like those described here will produce the most biologically accurate reference genomes. The independently assembled allelic forms of many genes can then be seen as a bonus source of information on polymorphism. As an example, our study revealed that *C. brenneri* alleles are more diverse than those of *C. remanei*.

Methods

Strains and sequence data

The genomes of *C. brenneri*, *C. remanei*, and *C. japonica* were sequenced by the Genome Sequencing Center at Washington University, St. Louis, and will be described in detail elsewhere. The inbred strains used for DNA extraction—PB2801 for *C. brenneri*, PB4641 for *C. remanei*, and DF5081 for *C. japonica*—were derived by inbreeding of the natural isolates LKC28, EM464 and DF5080, respectively. The inbreeding of *C. brenneri* and *C. remanei* was carried out by S. Baird (Wright State University), while inbreeding of *C. japonica* was conducted by K. Kiontke (New York University). Following 20 generations of single-female inbreeding, the DNA was extracted from mixed populations and 6–9× coverage WGS sequences were produced. For *C. brenneri* and *C. remanei*, this extraction was done on three separate occasions, while it was done twice for *C. japonica* (E. Schwarz and J. Spieth, pers. comm.). Individual sequence reads are publicly available at the NCBI trace archive, and the preliminary *C. remanei* PCAP assembly and gene predictions, Cr01, is available at http://ftp.wormbase.org/genomes/remanei. *C. elegans* data were obtained from WormBase genome release WS174 (ftp://ftp.wormbase.org/pub/wormbase/genomes/elegans/) and its cognate WormPep protein database (ftp://ftp.sanger.ac.uk/pub/databases/wormpep/wormpep).

Assembly of genomic contigs corresponding to individual genes

Individual sequence reads corresponding to the putative *C. brenneri*, *C. remanei,* and *C. japonica* orthologs of individual *C. elegans* genes were identified by BLAST searches of the NCBI trace archive (http://www.ncbi.nlm.nih.gov/BLAST/), using exon sequences as queries. Contigs were assembled using the STADEN software package. Whenever possible, contigs were extended up to 1-kb upstream of translation initiation signal and 1-kb downstream of the termination signal. The assemblies were manually inspected for discrepancies (e.g., errors in base calling), and cases of putative heterozygosity were identified and divided into variants using the "diploid graph" and "SNP candidates" options in STADEN (SOM).

Genome-wide assessment of heterozygous loci in *C. remanei*

We first created a data set of 20,099 unique *C. elegans* genes by eliminating all but the longest splice variants. We then defined a set of 14,530 single-copy *C. elegans* genes by all-against-all BLASTP and FASTA searches, retaining only genes that satisfied the single-copy criteria of either Stein et al. (2003) or Gu et al. (2002), respectively. This set was then used to query the 25,595 unique predicted *C. remanei* proteins in Cr01 via BLASTP. The highest-scoring hits were themselves used as queries, and resulting gene pairs with the following alignment attributes were retained: (a) be each other's reciprocal best hit; (b) have an overall BLASTP *E*-value of < -10; (c) have both \$e_component < -70 and (\$e_component $-$ \$minimum_ecomp < 30); (d) have an alignment length of at least 0.6× the query length; and (e) satisfy minimum sequence identity, based on a modification of the method of Gu et al. (2002) and Rost (1999), of 30% for alignments greater than 150 residues or, for shorter alignments, percent identity of at least $6 + 480L^{-0.32(1+e-L/1000)}$, where L is the length of the alignment. We tested these rules using known allelic genes to get the modified length and percentage identity requirement of empirical formula.

The above procedure yielded 1080 candidate allelic pairs of *C. remanei* peptides corresponding to a single-copy *C. elegans* gene. We then eliminated pairs with <90% amino acid sequence identity, pairs with overlapping membership, and those residing on the same genomic contig, which are likely to be tandem duplications. Assuming conserved synteny, these latter steps disproportionately eliminated pairs on chromosomes I, II, III, V, and X (Supplemental Table S2), resulting in 487, most of them on the central 80% of chromosome IV. For the plots in Figure 2, C. *remanei* read depths were determined from the assembly output Ace format files for all contig positions. The reads were aligned to form a contig using their clear (i.e., good) quality portions of the reads, and each gene's average read depth was calculated by dividing the total of read depths for all base pairs by the total gene length.

Genome-wide copy number analysis

For the estimates of the genome-wide copy number presented in Figure 3, we used the cross_match program (P. Green, unpubl.) with parameter values of −minmatch 20, −minscore 250, −masklevel 101, and −penalty − 3. To test for possible undercounting of highly diverged allelic DNA in the *C. brenneri* assembly, we repeated this analysis with a –minscore of 80. This only affected the copy number distributions slightly (e.g., d2 rose from 36% to 38%). Estimates of the parameter f for *C. remanei* were based on cross_match analyses of the 487 genes that were inferred to be heterozygous in the analysis shown in Figure 2. For *C. brenneri*, we estimated f from the heterozygous subset of the 42 manually assembled genes.

Analysis of sequence divergence between manually assembled variants

To eliminate artifacts from poor base calling during sequencing, we required potential variants to: (1) differ at several sites within

the length of a single sequence read, and (2) be supported by at least three independent reads. In practice, for most bi-allelic loci each variant was substantiated by many more than three independent reads. Depths of read coverage are shown in Supplemental Figure S1. In several heterozygous loci, regions of divergence between two variants are so distant that no individual sequence reads span both. Therefore, it was not always technically possible to reconstruct the linkage phase between clusters of divergent sites. Alignments of variants were generated using BIOEDIT software (SOM).

We used the intron-exons boundaries as annotated in *C. elegans* and *C. briggsae* genomes to designate orthologous exons in *C. brenneri*, *C. remanei*, and *C. japonica*. Divergence between pairs of variants was calculated independently for coding and noncoding regions, excluding contig ends for which confidence was low. Indels were not counted in these measures of divergence; however, large indels are documented independently (column "Comments" in Supplemental Table S1). Highly divergent and nonalignable regions were included in the calculations of divergence and documented in the "Comments" column. Calculations of K_a/K_s were performed using codeML (Yang 1997).

PCR genotyping

Genotyping of heterozygous *C. brenneri* genes was based on PCR assays for polymorphic indel polymorphisms (693 bp vs. 398 bp for *fog-1*, 416 bp vs. 156 bp for *sur-2*, 460 bp vs. 408 bp for *lin-7*). For *C. remanei*, primers EH31 and EH32 distinguish a 120 nt indel polymorphism in *Cr-fem-1*, and Rf3F1 (Haag et al. 2002) and EH30 allow amplification of a 450 nt product containing *EcoRV* restriction site polymorphism in *Cr-fem-3*. Primer sequences are given in SOM.

Single worm PCR was performed as described (Haag et al. 2002), with the modification of using 10 μL for lysis and a total reaction volume of 40 μL. Males were genotyped immediately after mating, whereas females were genotyped after 24 h of egg laying. Residual sperm from mating did not appear to interfere with the inference of the maternal genotype.

Fitness measurements

Worms were incubated at 20°C on standard NGM plates (Wood 1988). Fitness was measured as the number of zygotes and viable progeny generated within 24 h following mating. For each strain, male–female mating pairs were established from randomly picked young virgin adults, which had been allowed to mature for 6–24 h following the last larval (L4) stage. Males were removed after 1 h of mating, and females were allowed to lay eggs for 24 h in three 8-h (for *C. brenneri*) or two 12-h (for *C. remanei*) windows. Embryos and live progeny were counted for up to 24 h after removal of the female.

Acknowledgments

We are grateful to John Spieth for encouraging and facilitating this work; Thomas Nagylaki for generous help and advice; Spencer Johnston for permission to use unpublished data; Scott Baird, Erich Schwarz, and Michael Lynch for useful discussions; and Martin Kreitman and Chung-I Wu for helpful comments on the manuscript. This work was supported by grants from the National Institutes of Health (GM079414) and National Science Foundation (IBN0414512) (to E.S.H.), and by institutional funds from the University of Chicago (to I.R.).

References

Baird, S.E. 2002. Haldane's rule by sexual transformation in *Caenorhabditis*. *Genetics* **161:** 1349–1353.

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., et al. 2007. Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* **5:** e310. doi: 10.1371/journal.pbio.0050310.

Brenner, S. 1974. The genetics of *Caenorhabditis elegans*. *Genetics* **77:** 71–94.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Cutter, A.D. and Payseur, B.A. 2003. Selection at linked sites in the partial selfer *Caenorhabditis elegans*. *Mol. Biol. Evol.* **20:** 665–673.

Cutter, A.D., Baird, S.E., and Charlesworth, D. 2006. High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. *Genetics* **174:** 901–913.

Dolgin, E.S., Charlesworth, B., Baird, S.E., and Cutter, A.D. 2007. Inbreeding and outbreeding depression in *Caenorhabditis* nematodes. *Evolution* **61:** 1339–1352.

Drosophila 12 Genome Consortium. 2007. Evolution of genes and genomes on the *Drosophila phylogeny*. *Nature* **450:** 203–218.

Frydenberg, O. 1963. Population studies of a lethal mutant in *Drosophila melanogaster*. I. Behavior in populations with discrete generations. *Hereditas* **50:** 89–116.

Ghedin, E., Wang, S., Spiro, D., Caler, E., Zhao, Q., Crabtree, J., Allen, J.E., Delcher, A.L., Guiliano, D.B., Miranda-Saavedra, D., et al. 2007. Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317:** 1756–1760.

Graustein, A., Gaspar, J.M., Walters, J.R., and Palopoli, M.F. 2002. Levels of DNA polymorphism vary with mating system in the nematode genus Caenorhabditis. *Genetics* **161:** 99–107.

Gu, Z., Cavalcanti, A., Chen, F.C., Bouman, P., and Li, W.H. 2002. Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol. Biol. Evol.* **19:** 256–262.

Haag, E.S. and Ackerman, A.D. 2005. Intraspecific variation in *fem-3* and *tra-2*, two rapidly coevolving nematode sex-determining genes. *Gene* **349:** 35–42.

Haag, E.S. and Kimble, J. 2000. Regulatory elements required for development of *Caenorhabditis elegans* hermaphrodites are conserved in the *tra-2* homologue of *C. remanei*, a male/female sister species. *Genetics* **155:** 105–116.

Haag, E.S., Wang, S., and Kimble, J. 2002. Rapid coevolution of the nematode sex-determining genes *fem-3* and *tra-2*. *Curr. Biol.* **12:** 2035–2041.

Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D.L., Houle, D., Charlesworth, B., and Keightley, P.D. 2007. Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature* **445:** 82–85.

Hill, R.C., de Carvalho, C.E., Salogiannis, J., Schlager, B., Pilgrim, D., and Haag, E.S. 2006. Genetic flexibility in the convergent evolution of hermaphroditism in Caenorhabditis nematodes. *Dev. Cell* **10:** 531–538.

Hillier, L.W., Miller, R.D., Baird, S.E., Chinwalla, A., Fulton, L.A., Koboldt, D.C., and Waterston, R.H. 2007. Comparison of *C. elegans* and *C. briggsae* genome sequences reveals extensive conservation of chromosome organization and synteny. *PLoS Biol.* **5:** e167. doi: 10.1371/journal.pbio.0050167.

Istrail, S., Sutton, G.G., Florea, L., Halpern, A.L., Mobarry, C.M., Lippert, R., Walenz, B., Shatkay, H., Dew, I., Miller, J.R., et al. 2004. Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl. Acad. Sci.* **101:** 1916–1921.

Kim, J.H., Waterman, M.S., and Li, L.M. 2007. Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*. *Genome Res.* **17:** 1101–1110.

Kiontke, K., Gavin, N.P., Raynes, Y., Roehrig, C., Piano, F., and Fitch, D.H. 2004. *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc. Natl. Acad. Sci.* **101:** 9003–9008.

Kiontke, K., Barriere, A., Kolotuev, I., Podbilewicz, B., Sommer, R., Fitch, D.H., and Felix, M.A. 2007. Trends, stasis, and drift in the evolution of nematode vulva development. *Curr. Biol.* **17:** 1925–1937.

Kuwabara, P.E. and Shah, S. 1994. Cloning by synteny: Identifying *C. briggsae* homologues of *C. elegans* genes. *Nucleic Acids Res.* **22:** 4414–4418.

Latter, B.D. and Mulley, J.C. 1995. Genetic adaptation to captivity and inbreeding depression in small laboratory populations of *Drosophila melanogaster*. *Genetics* **139:** 255–266.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. 2007. The diploid

genome sequence of an individual human. *PLoS Biol*. **5:** e254. doi: 10.1371/journal.pbio.0050254.
Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.
Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Nagylaki, T. 1992. *Introduction to Theoretical Population Genetics*. Springer, New York.
Nayak, S., Goree, J., and Schedl, T. 2005. *fog-2* and the evolution of self-fertile hermaphroditism in *Caenorhabditis*. *PLoS Biol*. **3:** e6. doi: 10.1371/journal/pbio.0030006.
Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., et al. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. *Genome Res*. **15:** 1–18.
Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Eng*. **12:** 85–94.
Rumball, W., Franklin, I.R., Frankham, R., and Sheldon, B.L. 1994. Decline in heterozygosity under full-sib and double first-cousin inbreeding in *Drosophila melanogaster*. *Genetics* **136:** 1039–1049.
Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.
Sea Urchin Genome Sequencing Consortium. 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314:** 941–952.
Small, K.S., Brudno, M., Hill, M.M., and Sidow, A. 2007. A haplome alignment and reference sequence of the highly polymorphic *Ciona savignyi* genome. *Genome Biol*. **8:** R41. doi: 10.1186/gb-2007-8-3-241.
Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. 2003. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol*. **1:** E45. doi: 10.1371/journal.pbio.0000045.
Sternberg, P., Waterston, R., Spieth, J., Eddy, S., and Wilson, R. 2003. Genome sequence of additional *Caenorhabditis* species: Enhancing the utility of *C. elegans* as a model organism. [Genome sequencing white paper proposal to NHGRI.] www.genome.gov/Pages/Research/Sequencing/SeqProposals/C_remaneiSEQ.pdf.
Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Vinson, J.P., Jaffe, D.B., O'Neill, K., Karlsson, E.K., Stange-Thomann, N., Anderson, S., Mesirov, J.P., Satoh, N., Satou, Y., Nusbaum, C., et al. 2005. Assembly of polymorphic genomes: Algorithms and application to *Ciona savignyi*. *Genome Res*. **15:** 1127–1135.
Wood, W.B. 1988. *The Nematode Caenorhabditis elegans*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
Wu, C.I., Hollocher, H., Begun, D.J., Aquadro, C.F., Xu, Y., and Wu, M.L. 1995. Sexual isolation in *Drosophila melanogaster*: A possible case of incipient speciation. *Proc. Natl. Acad. Sci*. **92:** 2519–2523.
Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci*. **13:** 555–556.

Received June 4, 2008; accepted in revised form December 1, 2008.