9-1-2016

The Genetic Basis of Natural Variation in Caenorhabditis elegans Telomere Length

Daniel E. Cook
Northwestern University

Stefan Zdraljevic
Northwestern University

Robyn E. Tanny
Northwestern University

See next page for additional authors

Follow this and additional works at: http://escholarship.umassmed.edu/neurobiology_pp

Part of the Neuroscience and Neurobiology Commons

Repository Citation
Cook, Daniel E.; Zdraljevic, Stefan; Tanny, Robyn E.; Seo, Beomseok; Riccardi, David D.; Noble, Luke M.; Rockman, Matthew V.; Alkema, Mark J.; Braendle, Christian; Kammenga, Jan E.; Wang, John; Kruglyak, Leonid; Felix, Marie-Anne; Lee, Junho; and Andersen, Erik C., "The Genetic Basis of Natural Variation in Caenorhabditis elegans Telomere Length" (2016). Neurobiology Publications and Presentations. 191.
http://escholarship.umassmed.edu/neurobiology_pp/191

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in Neurobiology Publications and Presentations by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
The Genetic Basis of Natural Variation in Caenorhabditis elegans Telomere Length

Authors

Keywords
Caenorhabditis elegans, QTL, shelterin, telomere length, whole-genome sequence

Rights and Permissions
Available freely online through the author-supported open access option.
The Genetic Basis of Natural Variation in Caenorhabditis elegans Telomere Length


*Department of Molecular Biosciences, †Interdisciplinary Biological Science Program, ‡‡‡‡Robert H. Lurie Comprehensive Cancer Center, ††††Chemistry of Life Processes Institute, and §§§Northwestern Institute on Complex Systems, Northwestern University, Evanston, Illinois 60208, ‡Department of Biological Sciences, Institute of Molecular Biology and Genetics, Seoul National University, 08826, Korea, ††††Department of Biology, and **Center for Genomics and Systems Biology, New York University, New York 10003, ††Department of Neurobiology, University of Massachusetts Medical School, Worcester, Massachusetts 01605, ‡‡Centre National de la Recherche Scientifique, Institut National de la Santé et de la Recherche Médicale, Institut de Biologie Valrose, Université Nice Sophia Antipolis, 06100 Nice, France, ‡‡‡‡Laboratory of Nematology, Wageningen University, 6708 PB, Netherlands, ***Biodiversity Research Center, Academia Sinica, Taipei 115, Taiwan, ††††Departments of Human Genetics and Biological Chemistry, University of California, Los Angeles, California 90095, †††Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, §§§Institut de Biologie de l’Ecole Normale Supérieure, Centre National de la Recherche Scientifique, Institut National de la Santé et de la Recherche Médicale, 75005 Paris, France, and ****Department of Biophysics and Chemical Biology, Seoul National University, 08826, Korea

ORCID IDs: 0000-0003-3347-562X (D.E.C.); 0000-0003-2883-4616 (S.Z.); 0000-0001-6492-8906 (M.V.R.); 0000-0003-0229-9651 (E.C.A.)

ABSTRACT Telomeres are involved in the maintenance of chromosomes and the prevention of genome instability. Despite this central importance, significant variation in telomere length has been observed in a variety of organisms. The genetic determinants of telomere-length variation and their effects on organismal fitness are largely unexplored. Here, we describe natural variation in telomere length across the Caenorhabditis elegans species. We identify a large-effect variant that contributes to differences in telomere length. The variant alters the conserved oligonucleotide/oligosaccharide-binding fold of protection of telomeres 2 (POT-2), a homolog of a human telomere-capping shelterin complex subunit. Mutations within this domain likely reduce the ability of POT-2 to bind telomeric DNA, thereby increasing telomere length. We find that telomere-length variation does not correlate with offspring production or longevity in C. elegans wild isolates, suggesting that naturally long telomeres play a limited role in modifying fitness phenotypes in C. elegans.

KEYWORDS Caenorhabditis elegans; QTL; shelterin; telomere length; whole-genome sequence

Genome-wide association (GWA) studies, in which phenotypic differences are correlated with genome-wide variation in populations, offer a powerful approach to understand the genetic basis of complex traits (McCarthy et al. 2008). GWA requires accurate and quantitative measurement of traits for a large number of individuals. Even in organisms that are studied easily in the laboratory, the measurement of quantitative traits is difficult and expensive. By contrast, the rapid decrease in sequencing costs has made the collection of genome-wide variation accessible. From Drosophila (Mackay et al. 2012; Lack et al. 2015) to Arabidopsis (Weigel and Mott 2009) to humans (The 1000 Genomes Project Consortium 2012), the whole genomes from large populations of individuals can be analyzed to identify natural variation that is correlated with quantitative traits. Because the genome itself can vary across populations, whole-genome sequence data sets can be mined for traits without measuring the physical organism. Specifically, large numbers of sequence reads generated from individuals in a
species can be analyzed to determine attributes of genomes, including mitochondrial- or ribosomal-DNA copy numbers. Another such trait is the length of the highly repetitive structures at the ends of linear chromosomes called telomeres (Blackburn 1991).

Telomeres are nucleoprotein complexes that serve as protective capping structures to prevent chromosomal degradation and fusion (O’Sullivan and Karlseder 2010). The DNA component of telomeres in most organisms consists of long stretches of nucleotide repeats that terminate in a single-stranded 3’ overhang (McEachern et al. 2000). The addition of telomeric repeats is necessary because DNA polymerase is unable to completely replicate the lagging strand (Watson 1972; Levy et al. 1992). The length of telomeres can differ among cell populations (Samassekou et al. 2010), from organism to organism (Fulcher et al. 2014), and within proliferating cellular lineages (Frenck et al. 1998). Two antagonistic pathways regulate telomere length. In the first pathway, the reverse transcriptase telomerase adds de novo telomeric repeats to the 3’ ends of chromosomes. In the second, telomere lengthening is inhibited by the shelterin complex. Shelterin forms a protective cap at telomere ends, presumably through the formation of lariat structures known as t-loops (Griffith et al. 1999). The t-loops are hypothesized to inhibit telomerase activity by preventing access to the 3’ tail. Additionally, because uncapped telomeres resemble double-stranded DNA breaks, shelterin association with telomeric DNA represses endogenous DNA-damage repair pathways, preventing chromosomal fusion events, and preserving genome integrity (De Lange 2010).

Variation in telomere length has important biological implications. In cells lacking telomerase, chromosome ends become shorter with every cell division, which eventually triggers cell-cycle arrest (Harley et al. 1992). In this way, telomere length sets the replicative potential of cells and acts as an important tumor-suppressor mechanism (Harley et al. 1992; Deng et al. 2008). In populations of nonclonal human leukocytes, telomere lengths have been shown to be highly heritable (Broer et al. 2009). Quantitative trait loci (QTL) identified from human GWA studies of telomere length implicate telomere-associated genes, including telomerase (TERT), its RNA template (TERC), and OBFC1 (Levy et al. 2010; Jones et al. 2012; Codd et al. 2013). QTL underlying variation in telomere length have been identified in Arabidopsis thaliana, Saccharomyces paradoxus, and S. cerevisiae using both linkage and association approaches (Gatbonton et al. 2006; Liti et al. 2009; Kwan et al. 2011; Fulcher et al. 2014). In S. paradoxus, natural variation in telomere lengths is mediated by differences in telomerase complex components. In S. cerevisiae, natural telomere lengthening is caused by a loss of an amino acid permease gene. Thus far, no studies in multicellular animals or plants have been able to identify specific genes responsible for population telomere-length differences. Recent advances in wild-strain genotypes and sequences (Andersen et al. 2012) in Caenorhabditis elegans make it a powerful model to address natural variation in telomere length and its fitness consequences.

Like in humans, telomerase and shelterin activities regulate C. elegans telomere length (Malik et al. 2000; Cheung et al. 2006; Meier et al. 2006; Shtessel et al. 2013). The TRT-1-containing telomerase complex is hypothesized to add TTAGGC repeats to the ends of chromosomes and prevents chromosome shortening (Meier et al. 2006), and the shelterin complex regulates access of the telomerase complex to chromosome ends (Raices et al. 2008; Cheng et al. 2012; Shtessel et al. 2013). The length of telomeres in the laboratory strain N2 is variable and ranges between 2 and 9 kb (Wicky et al. 1996; Raices et al. 2005). The telomere lengths in wild isolates of C. elegans are largely unexplored. Previous studies examined variation in telomere length using a small number of wild strains (Cheung et al. 2004; Raices et al. 2005). However, several of the supposed wild strains have since been determined to be mislabeled versions of the laboratory strain N2 (McGrath et al. 2009). Thus, it is not known if and how telomere lengths vary among C. elegans natural strains. Additionally, the fitness consequences of telomere-length variation have not been defined.

Here, we collected a new set of whole-genome sequences from 208 wild C. elegans strains and used these strains to investigate natural variation in telomere length across the species. Computational estimates of telomere lengths were confirmed using molecular measurements, indicating that this technique can be applied across a large number of wild strains. Using association mapping, we found that variation in the gene protection of telomeres 2 (pot-2) is correlated with differences in C. elegans telomere lengths. Natural variation in pot-2 affects gene function and causes longer than average telomeres in some wild strains. Additionally, we examined whether population differences in telomere length connect to differences in fitness traits, including brood size and longevity. Our results indicate that variation in pot-2 does not correspond to variation in fitness as measured in the laboratory and does not show strong signatures of selection in nature. These data suggest that telomere length beyond a basal threshold is of limited consequence to C. elegans. Our results underscore how traits obtained from sequence data can be used to understand the dynamic nature of genomes within populations.

Materials and Methods

Strains

C. elegans strains were cultured using bacterial strain OP50 on a modified nematode growth medium [NGM with 1% agar, 0.7% agarose (NGMA)] to prevent burrowing of wild isolates (Andersen et al. 2014). Strain information is listed in Supplemental Material, File S1. The following strains were scored for the molecular telomere assays described below: AB4 (CB4858 isotype), CB4856, CX11285, CX11292, DL238, ECA248, ED3012, EG4349, JT11398, JU311, JU1400, JU2007, KR314, N2, NIC2, NIC3, NIC207, PB303, and QX1212.
Library construction and sequence acquisition

DNA was isolated from 100–300 μl of packed animals using the Blood and Tissue DNA isolation kit (QIAGEN, Valencia, CA). The provided protocol was followed with the addition of RNase (4 μl of 100 mg/ml) following the initial lysis for 2 min at room temperature (RT). DNA concentration was determined using the Qubit dsDNA Broad Range Assay Kit (Invitrogen, Carlsbad, CA). Libraries were generated using the Illumina Nextera Sample Prep Kit and indexed using the Nextera Index Kit. A total of 24 uniquely-indexed samples were pooled by mixing 100 ng of each sample. The pooled material was size selected by electrophoresing the DNA on a 2% agarose gel and excising the fragments ranging from 300 to 500 bp. The sample was purified using the QIAGEN MinElute Kit and eluted in 11 μl of Buffer EB. The concentration of the purified sample was determined using the Qubit dsDNA High Sensitivity Assay Kit. Sequencing was performed on the Illumina HiSeq 2500 platform. To increase coverage of some strains, we incorporated data from two separate studies of wild strains (Thompson et al. 2013; Noble et al. 2015).

Trimming and demultiplexing

When necessary, demultiplexing and sequence trimming were performed using fastx_barcode_splitter.pl (version 0.0.14) (Gordon and Hannon 2010). Sequences were trimmed using trimmomatic (version 0.32) (Bolger et al. 2014). Nextera libraries were trimmed using the following parameters:

NexteraPE-PE.fa:2:80:10 MINLEN:45
TruSeq libraries were trimmed using:

TruSeq2-PE.fa:2:80:10 TRAILING:30 SLIDINGWINDOW:4:30 MINLEN:30

The full details of the preparation, source, and library are available in File S2.

Alignment, variant calling, and filtering

FASTQ sequence data has been deposited under National Center for Biotechnology Information Bioproject accession PRJNA318647. Sequences were aligned to WS245 (http://www.wormbase.org) using the Burrows–Wheeler Aligner (BWA) (version 0.7.8-r455) (Li and Durbin 2009). Optical/PCR duplicates were marked with PICARD (version 1.111). BAM and CRAM files are available at http://www.elegansvariation.org/Data. To determine which type of single-nucleotide-variant (SNV) caller would perform best on a genomically diverse population to each other. After grouping, bcftools has two different calling methods: a consensus caller (specified using -c), and a more recently developed multi-allelic caller (specified using -m) (Li 2011). We performed variant calling using all four combinations of individual/joint calling and the consensus/multiallelic parameters. Because of the hermaphroditic life cycle of C. elegans, heterozygosity rates are likely low. Occasionally, heterozygous variants will be called despite skewed read support for reference or alternative alleles. To account for these likely erroneous calls, we performed ‘heterozygous polarization’ using the log-likelihood ratios of reference to alternative genotype calls. When the log-likelihood ratio was < −2 or > 2, heterozygous genotypes were polarized (or switched) to reference genotypes or alternative genotypes, respectively. All other SNVs with likelihood ratios between −2 and 2 were called NA. Following variant calling and heterozygous polarization on resulting calls, we observed increased rates of heterozygous calls using joint methods and decreased true positive (TP) rates using our simulation data set (File S3, File S4). Given C. elegans’ predominantly self-fertilizing mode of reproduction, we decided to focus on the individual-based calling method that performed better. Next, we determined the optimal filters to maximize TP rates and minimize false positive (FP) and false negative (FN) results using our simulated data (File S4). After implementing different combinations of filters, we found that depth (DP), mapping quality (MQ), variant quality (QUAL), and the ratio of high-quality alternative base calls (DV) over DP filters worked well (Figure S1). Variants with DP ≤ 10, MQ < 40, QUAL < 30, and DV/DP < 0.5 were called NA. Using these filters, we called 1.3-million SNVs across 152 isotypes. This data set is available at http://www.andersenlab.org/Research/Data/Cooketal.

Validation of SNV-calling methods

In addition to performing simulations to optimize SNV-calling filters, we compared our whole-genome sequence variant calls with SNVs identified previously in CB4856 (Wicks et al. 2001). Out of 4256 sites we were able to call in regions that were sequenced using Sanger sequencing, we correctly identified 4223 variants (99.2% of all variants) in CB4856. One TP was erroneously filtered and two FPs were removed using our filters, and we failed to call the nonreference allele for 30 variants (FNs).

Additionally, we examined sequence variants with poor parameter values in terms of depth, quality, heterozygosity, or modification by our heterozygous polarization filter. We used primer3 (Rozen and Skaltsky 2000) to generate a pair of primers for performing PCR and a single forward primer for Sanger sequencing. We successfully sequenced 73 of 95 sites chosen from several strains. Comparison of variant calls after imputation and filtering yielded 46 TPs and 14 true negatives. We successfully removed 3 out of 11 FPs and erroneously filtered two sites that should have been called as nonreference
Some strains in our original collection were isolated from the same or nearly identical locations. Therefore, we determined if these strains share distinct genome-wide haplotypes or isotypes. To determine strain relatedness, we sequenced and called variants from sequencing runs independently (e.g., individual FASTQ pairs) to ensure that strains were properly labeled before and after sequencing. We then combined FASTQ files from sequencing runs for a given strain and examined the concordance among genotypes. Comparison of variants identified among sequenced strains were used to determine whether the strains carried identical haplotypes. We observed that some strains were highly related to each other as compared with the rest of the population. Strains that were >99.93% identical across 1,589,559 sites identified from sequencing runs of individual strains were classified as isotypes (Figure S2). The disparity between the final 1.3-million SNV set as compared to the 1.5-million SNV set comes from the different levels and types of filters applied for strain concordances to make isotypes or from SNV calling across the JU360 and JU363 were previously thought to be separate, wide genotype but exhibit distinct phenotypes (Sterken strains with the same names but stored at the 2000 had different genome-wide haplotypes compared to wide level and at high depths of coverage, these strains are but highly related, isotypes. We found that, at the genome- Genes Genetics Center (CGC). CB4851 from the CGC had a different name from Cambridge, United Kingdom. We renamed the CB4851 strain from the CGC as ECA243. By contrast, the CB4855 from Cambridge, United Kingdom is differ- ent from the CGC version of CB4855. We gave this strain the name ECA248 to avoid confusion. CB4858 from Cambridge, United Kingdom is nearly identical to N2 and not studied further. CB4855 from the CGC has a genome-wide haplotype that matches CB4858, which has a different history and isolation location. Therefore, we cannot guarantee the fidelity of this strain, and it was not studied further. CB4855 from Cambridge, United Kingdom is differ- ent from the CGC version of CB4855. We gave this strain the name ECA248 to avoid confusion. CB4858 from CGC has a different genome-wide haplotype than CB4858 from Cambridge, United Kingdom. Therefore, we renamed CB4858 from Cambridge, United Kingdom to ECA252, and it is a separate isotype. The CB4858 from the CGC was renamed ECA251 and is the reference strain from the CB4858 isotype.

**Imputation and variant annotation**

Following SNV calling and filtering, some variant sites were filtered. Next, we generated an imputed SNV set using beagle (version r1399) (Browning and Browning 2016). This imputed variant set is available at http://www.andersenlab.org/Research/Data/Cooketal. We used SnpEff (version 4.1g) (Cingolani et al. 2012) on this SNV set to predict functional effects.

**Telomere-length estimation**

Telomere lengths were estimated using TelSeq (version 0.0.1) (Ding et al. 2014) on BAM files derived from wild isolates or Million Mutation Project (MMP) strain sequencing. To estimate telomere lengths, TelSeq determines the reads that contain greater than seven telomeric hexamer repeats (TTAGGC for C. elegans). Compared to most hexamers, telomeric hexamers can be found tandemly repeated within sequenced reads. TelSeq calculates the relative proportion of reads that ap- pear to be telomerically derived among all sequenced reads and transforms this value into a length estimate using the formula $l = t_k sc$ where $l$ is the length estimate and $t_k$ is the abundance of reads with a minimum of $k$ telomeric repeats. The value of $s$ is the fraction of all reads with a GC composition similar to the telomeric repeat (48–52% for C. elegans). The value of $c$ is a constant representing the length of 100-bp windows within the reference genome where GC content = GC content of the telomeric repeat / total number of telomere ends. By default, TelSeq provides length estimates applicable to humans. We found the number of 100-bp windows with a 50% GC content in the WS245 reference genome to be 58,087. We calculated $c$ for C. elegans as 58,087 kb/12 telomeres = 484. This value was used to transform human length estimates to length estimates appropriate to C. elegans.

Notably, telomere-length estimates are averaged across all chromosomes, as no specific data about any one particular telomere is determined. To assess how well the TTAGGC hexamer distinguishes telomeric reads from nontelomeric reads, we examined the frequencies of noncyclical permutations of the C. elegans hexamer in the N2 laboratory strain using TelSeq (Figure S3). We observe that the majority of hexamers examined were not present in more than six copies in a high frequency of reads. By contrast, the reads possessing the telomeric hexamer with seven copies or more were more abundant than any other hexamer. Tandem repeats of the telomeric hexamer are present within the reference genome at the ends of each chromosome and occasionally internally within chromosomes at between 2 and 71 copies (File S5). After running TelSeq on our wild isolates, we removed eight sequencing runs (out of 868 total) that possessed zero reads with 15 or more copies of the telomeric hexamer. These se- quencing runs provide additional support for SNV calling but had short read lengths that would underestimate telomere length. We used the weighted average of telomere-length es- timates for all runs of a given strain based on total reads to calculate telomere-length estimates. File S6 details telomere-length estimates for every sequencing run.

**Quantitative PCR assays for telomere-length measurements**

Telomere lengths were measured by quantitative PCR (qPCR) as described previously with some modifications (Cawthon...
was then treated with 10 mM buffer containing 10 mM Tris (pH 7.5) and 1 mM EDTA. DNA extraction and ethanol precipitation. DNA was eluted with 0.1 mg/mL proteinase K. DNA was isolated by phenol Tris-Cl (pH 8.5), 0.1 M NaCl, 50 mM EDTA (pH 8.0), 1% SDS, fraction.

again isolated with phenol extraction and ethanol precipitation. DNA was quantified with TFL-TELO software (Dr. Peter Lansdorp, Terry Fox Laboratory, Vancouver, BC) (Poon et al. 1999). Because embryos were scored for fluorescent foci, it is possible that the telomere-length estimates were quantified from both telomeres and internal repeat elements sharing homology to telomeric DNA.

GWA mapping

GWA mapping was performed on marker genotype data and telomere-length estimates using the rrBLUP package (version 4.3) (Endelman 2011) and GWAS function. rrBLUP requires a kinship matrix and an SNV set to perform GWA. We generated a kinship matrix using our imputed SNV set with the A.mat function within rrBLUP. Genomic regions of interest were determined empirically from simulating a QTL that explained 20% of the phenotypic variance at each marker in our mapping data set. All simulated QT were mapped within 100 markers (50 markers to the left and 50 markers to the right) of the simulated marker positions. To generate a SNV set for mapping, we again used our imputed SNV set. However, we filtered the number of SNVs to a set of 38,688 markers. This set was generated by lifiting over (from WS210 to WS245) a set of 41,888 SNVs previously used for GWA mapping (Andersen et al. 2012) and filtering our imputed SNVs to those sites.

MMP analysis

Whole-genome sequence data from mutagenized strains within the MMP was obtained from the sequence read archive (SRA) (project accession number SRP018046). We removed 59 strains that were contaminated with other strains. We were also unable to locate the sequence data for 12 MMP strains on SRA, leaving us with 1,936 mutagenized strains. Within the MMP project, read lengths varied among sequencing runs, being either 75 bp or 100 bp. We ran TelSeq on all sequencing runs assuming 100-bp reads. To use 75-bp sequencing runs, we took the 448 strains that were sequenced at both 75 and 100 bp lengths and used those estimates to develop a linear model. Then, this model was used to transform 75-bp length estimates to 100-bp estimates (Figure S5). We then used the weighted average of telomere-length estimates for all runs of a given strain based on total reads to calculate telomere-length estimates. Because telomeric reads resemble PCR duplicates, TelSeq uses them in calculating telomere length. However, we observed very low PCR and optical duplicate rates among MMP sequence data, likely due to differences in library preparation in contrast to wild isolate sequence data. These differences likely account for shorter telomere lengths in our data.

Long-telomere strains from the MMP were classified as strains with telomere lengths greater than the 98th quantile of
all MMP strains (6.41 kb). Mutation data were obtained from the MMP website (http://genome.sfu.ca/mmp/mmp_mut_strains_data_Mar14.txt). A hypergeometric test was performed to identify which genes were enriched for mutations from long-telomere strains (File S7) using the phyper function in R (R Development Core Team 2013). FX1400 was propagated for 10 generations prior to whole-genome sequencing. Telomere length was estimated using TelSeq.

Statistical analyses

Statistical analyses were performed using R (version 3.2.3). Plots were produced using ggplot2 (version 2.0.0).

Longevity assays

At least 80 L4 animals were plated onto each of three separate 6-cm NGMA plates in two independent assays and viability assessed each day until all animals were scored as dead or censored from the analysis as a result of bagging or missing animals. Animals were scored as dead in the absence of touch response and pharyngeal pumping. Animals were transferred to fresh plates every day from the initiation of the assay until day 7 of adulthood to remove progeny, and transferred every other day until the completion of the assay. The following short telomere strains were scored: EG4349, JU2007, NIC1, and NIC3. The following long-telomere strains were scored: KR314, NIC207, QX1212, and RC301. Additionally, N2 and CB4856 were scored.

High-throughput fecundity assays

Assays were performed similarly to those previously reported (Andersen et al. 2015) with the following differences: Animals were bleached, synchronized, and grown to L4 larvae in 96-well plates. From the L1–L4 stage, animals were fed 5 mg/ml of a large-scale production HB101 lysate in K medium (Boyd et al. 2010) to provide a stereotyped and constant food source. Then, three L4 larvae from each of the 152 genotypes were dispensed using a COPAS BIOSORT instrument to wells containing 10 mg/ml HB101 lysate in K medium, and progeny were counted 96 hr later. Fecundity data were calculated using 12 samples—triplicate technical replicates from four biological replicates. The data were processed using COPASutils (Shimko and Andersen 2014) and statistically analyzed using custom R scripts.

Clustering of relatedness

Variant data for dendrogram comparisons were assembled by constructing a FASTA file with the genome-wide variant positions across all strains and subsetting by regions as described. Multiple sequence comparison by log-expectation (MUSCLE, version v3.8.31) (Edgar 2004) was used to generate neighbor-joining trees. The R packages ape (version 3.4) (Paradis et al. 2004) and phyloseq (version 1.12.2) (McMurdie and Holmes 2013) were used for data processing and plotting.

Data availability

All data necessary for confirming the conclusions presented in the article are represented fully within the article.

Results

Whole-genome sequencing of a large number of wild C. elegans strains identifies new isotypes and highly diverged strains

Previous genome-wide analyses of C. elegans population diversity used single-nucleotide variants (SNVs) ascertained from only two strains (Rockman and Kruglyak 2009), from reduced representation sequencing that only studied a fraction of the genome (Andersen et al. 2012), or from a small set of wild strains (Thompson et al. 2013). To address these limitations, we sequenced the whole genomes of a collection of 208 wild strains (File S1). Because C. elegans reproduction occurs primarily through the self-fertilization of hermaphrodites, highly related individuals proliferate and disperse, often in close proximity to one another (Barrière and Félix 2005; Félix and Braendle 2010). As a result, strains isolated in nature are frequently identical and share genome-wide haplotypes or isotypes. Sequencing data generated from strains belonging to the same isotype can be combined to increase depth of coverage and to improve downstream analyses. To identify which strains shared the same genome-wide haplotypes, we compared all of the variation identified in each of the 208 strains to each other in pairwise comparisons. The 208 strains reduce to 152 unique genome-wide haplotypes or isotypes (File S1). The combination of sequence data from all strains that make up an isotype led to a 70-fold median depth of coverage (Figure S6), enabling the discovery of SNVs and other genomic features. The number of SNVs in
comparisons of each isotype to the reference strain N2 ranged from strains highly similar to N2 with few SNVs, to highly diverged strains with 402,436 SNVs (Figure S7), and the density of SNVs across the genome matched previous distributions with more variants on chromosome arms than centers (Andersen et al. 2012) (Figure S8). An analysis of relatedness among these 152 isotypes recapitulated the general relationships previously identified among a set of 97 wild isotypes (Andersen et al. 2012) (Figure S9) with the addition of 55 new isotypes. Past studies identified one highly diverged strain isolated from San Francisco, CA, QX1211, which had divergence almost three times the level of other wild C. elegans strains (Andersen et al. 2012). Among the 55 new isotypes, one additional strain, ECA36 from New Zealand, is equally diverged, suggesting that wider sampling will recover additional diversity for this species. Altogether, our considerably expanded collection of whole-genome sequence data serves as a powerful tool to interrogate how natural variation gives rise to differences among individuals in a natural population.

**C. elegans wild strains differ in telomere lengths**

Our collection of high-depth whole-genome sequence data samples a large number of strains in the C. elegans species. The recent development of TelSeq, a program designed to estimate telomere length using short-read sequence data (Ding et al. 2014), allowed us to examine natural variation in telomere lengths computationally across wild C. elegans strains. We detected considerable natural variation in the total length of telomeric DNA in a strain (Figure 1), ranging from 4.12 kb to a maximum of 83.7 kb with a median telomere length of 12.25 kb (File S8). The TelSeq telomere-length estimate for N2 from our study was 16.97 kb, which is higher than previous estimates of 4–9 kb (Wicky et al. 1996) and 2–9 kb (Raices et al. 2005). This discrepancy likely arises from computational as compared to molecular assays, as we will discuss below. We found that the distribution of telomere lengths in the C. elegans population approximated a normal distribution with a right tail containing strains with longer than average telomeres. We found that our computational estimates of telomere length from Illumina sequence data were significantly influenced by library preparation, possibly driven by the method of DNA fragmentation (Figure S10). However, we were able to control for these differences using a linear model. We also observed a weak correlation between depth of coverage and TelSeq length estimates, but adjustments for library preparation eliminated this relationship (Figure S11).

TelSeq length estimates have been shown to give similar results as molecular methods to measure human telomere length (Ding et al. 2014). As of now, no studies have used TelSeq to examine C. elegans telomeres, so we investigated how well TelSeq estimates correlated with molecular methods, including terminal restriction fragment (TRF) Southern blot analyses, qPCR of telomere hexamer sequences, and FISH analyses. Using 20 strains, we found that the results from these molecular assays correlated well ($p = 0.445$ TRF, $0.815$ FISH, $0.699$ qPCR; Spearman’s rank correlation) with computational estimates of telomere lengths (Figure 2; File S9). These molecular results validated our computational estimates of telomere lengths and indicate that we can use TelSeq estimates to investigate the genetic causes underlying telomere variation.

**Species-wide telomere-length differences correlate with genetic variation on chromosome II**

To identify the genes that cause differences in telomere length across the C. elegans population, we used a GWA mapping approach as performed previously (Andersen et al. 2012) but taking advantage of the larger collection of wild strains. We treated our computational estimates of telomere length as a quantitative trait and identified one significant QTL on the right arm of chromosome II (Figure 3; File S10). To identify

---

**Figure 2** Telomere-length estimates correlate with alternative molecular measurement methods. Scatterplot of TelSeq telomere-length estimates (y-axis) plotted against alternative methods of telomere-length measurement on the x-axis. Alternative methods plotted on the x-axis and their associated Spearman’s rank correlation are (A) qPCR measurements normalized by N2 qPCR telomere-length estimate and scaled relative to the TelSeq N2 telomere-length estimate ($p = 0.445, P = 0.049$), (B) TRF ($p = 0.699, P = 8.5 \times 10^{-4}$), and (C) FISH measurements normalized by N2 FISH telomere-length estimate and scaled relative to the TelSeq N2 telomere-length estimate ($p = 0.815, P = 1.03 \times 10^{-5}$). Gray lines represent the regression lines between TelSeq and each method. Dashed diagonal lines represent identity lines.
the variant gene(s) that underlie this QTL, we investigated the SNVs within a large genomic region (12.9–15.3 Mb) surrounding the most significant marker on chromosome II. This region contains 557 protein-coding genes (File S11), but only 332 of these genes contained variants that are predicted to alter the amino acid sequences among the 152 strains. We examined genes with predicted protein-coding variants that could alter telomere length by correlating their alleles with the telomere-length phenotype. Nine genes possessed variation that was most highly correlated with telomere length ($p \geq 0.39$; File S11). The chromosome II QTL explains 28.4% of the phenotypic variation in telomere length. Three additional suggestive QTL on chromosomes I, II, and III were detected close to but below the significance threshold. Taken together, the four QTL explain 56.7% of the phenotypic variation in telomere length.

**Variation in pot-2 underlies differences in telomere length**

One of the nine genes in the chromosome II large-effect QTL is pot-2, a gene that was implicated previously in the regulation of telomere length (Raices et al. 2008; Cheng et al. 2012; Shtessel et al. 2013). A quantitative complementation test could be used to confirm that wild strains have the same functional effect as a pot-2 deletion. However, differences in telomere length caused by mutations in genes that encode telomere-associated proteins often do not have observable telomere defects for a number of generations (Vulliamy et al. 2004; Armanios et al. 2005; Marrone et al. 2005). It is technically not feasible to keep the genome heterozygous during long-term propagation. Given the large number of genes present within our confidence interval and challenges associated with examining telomere length using traditional genetic approaches, we sought alternative methods to confirm that variation in pot-2 could cause long telomeres. The ability to computationally estimate telomere length allowed us to further validate our approach using data from the MMP (Thompson et al. 2013) and examine whether the equivalent of a mutant screen for telomere length would provide insight into our result examining wild isolate genomes.

The MMP generated >2000 mutagenized strains using the laboratory N2 background. After each strain was passaged by self-mating of hermaphrodites for 10 generations, the strains were whole-genome sequenced to identify and predict the effects of induced mutations. The MMP data set can be used to identify correlations of phenotype and mutant genes in the laboratory strain background. We obtained whole-genome sequence data from 1936 mutagenized N2 strains, each of which has a unique collection of mutations. Importantly, 10 generations of self-propagation of these mutagenized strains prior to sequencing likely allowed telomere lengths to stabilize in response to mutations in genes that regulate telomere length, enabling us to observe differences. TelSeq returned telomere-length estimates for this population, which had a right long-tailed distribution (Figure 4A). The median telomere length among the mutagenized strains was 4.94 kb, which is much shorter than the length estimate of N2 in our collection (16.97 kb). This disparity likely arises due to differences in library preparation, sequencing platform, and data processing. We classified 39 of 1936 strains within the population as long-telomere strains with telomere lengths greater than 6.41 kb (98th percentile). Reasoning that certain mutant genes would be overrepresented in these 39 strains compared to the others, we performed a hypergeometric test to identify if enrichment for particular genes in long-telomere strains existed. After adjusting for multiple statistical tests, we identified pot-2 as the only gene highly enriched for mutations in six of the 39 long-telomere strains.
potential of individual cells contributes to organismal longevity, including longevity or cellular senescence (Harley et al. 1992; Heidinger et al. 2012; Soerensen et al. 2012). Our results with natural variation in telomere lengths provided a unique opportunity to connect differences in the length of telomeres with effects on organismal fitness. We measured offspring production for our collection of 152 wild strains and found no correlation with telomere length ($p = 0.062$; Figure 6A). Long telomeres allow for increased replicative potential of cells (Harley et al. 1992), but it is unclear how the replicative potential of individual cells contributes to organismal longevity phenotypes (Hornsby 2007). We chose nine strains covering the range of telomere-length differences and found no correlation with longevity ($p = 0.05$; Figure 6B, Figure S12). Taken together, these results suggest that the long telomeres found in some wild C. elegans strains do not have significant fitness consequences in these laboratory-based experiments.

Because we did not observe a strong effect on organismal fitness, we investigated the population genetics of pot-2 to test whether that locus had any signature of selection.

**Natural variants in pot-2 do not have detectable fitness consequences**

We connected genetic variation in the gene pot-2 with telomere-length differences across C. elegans wild strains. Specifically, an F68I variant in the putative telomere-binding OB-fold domain might cause reduction of function and long telomeres. A variety of studies have observed a relationship between telomere length and organismal fitness, including longevity or cellular senescence (Harley et al. 1992; Heidinger et al. 2012; Soerensen et al. 2012). Our results with natural variation in telomere lengths provided a unique opportunity to connect differences in the length of telomeres with effects on organismal fitness. We measured offspring production for our collection of 152 wild strains and found no correlation with telomere length ($p = 0.062$; Figure 6A). Long telomeres allow for increased replicative potential of cells (Harley et al. 1992), but it is unclear how the replicative potential of individual cells contributes to organismal longevity phenotypes (Hornsby 2007). We chose nine strains covering the range of telomere-length differences and found no correlation with longevity ($p = 0.05$; Figure 6B, Figure S12). Taken together, these results suggest that the long telomeres found in some wild C. elegans strains do not have significant fitness consequences in these laboratory-based experiments.

Because we did not observe a strong effect on organismal fitness, we investigated the population genetics of pot-2 to test whether that locus had any signature of selection.
Examination of Tajima’s D at the *pot-2* locus yielded no conspicuous signature, though the characteristic high linkage disequilibrium of *C. elegans* makes gene-focused tests challenging in this species (Figure S13). Furthermore, the haplotypes that contain this variant are rare (Figure S14) and not geographically restricted (Figure S15). Like the measurements of organismal fitness and lack of correlation with telomere-length differences, the population genetic test of neutrality indicates that the observed variation in *pot-2* is not under strong selective pressure. Together, these results suggest that natural variation in telomere length plays a limited role in modifying whole-organism phenotypes in *C. elegans*.

**Discussion**

In this study, we report the identification of a QTL on the right arm of chromosome II containing a variant within the gene *pot-2* that contributes to differences in telomere length of *C. elegans* wild isolates. To date, no connection of *pot-2* to natural variation in telomere lengths has been described. Several lines of evidence support the F68I allele of *pot-2* as the variant modulating telomere lengths. First, others have shown previously that loss of *pot-2* results in progressive telomere lengthening in the laboratory strain background (Raices et al. 2008; Shtessel et al. 2013). Second, the F68I variant is the only SNV in *pot-2* that correlates with long telomeres. This variant falls within the OB fold of *POT-2*, and our examination of strain telomere lengths within the MMP shows enrichment of mutations from long-telomere strains found within the OB-fold domain. Third, OB folds are known to interact with single-stranded nucleic acids, and TelSeq telomere-length estimates of wild isolates and randomly mutagenized laboratory strains show that mutation or variation of the OB-fold domain reduces function and causes long telomeres, as we also observed in the

**Figure 5** Variation within *pot-2* in wild isolate and MMP strains. Natural variation and induced mutations that alter codons across *pot-2* are shown along with the telomere-length estimates for all strains. (A) A schematic illustrating the *pot-2* genomic region is shown. The dark gray region represents the part of the genome encoding the OB-fold domain. Purple regions represent untranslated regions. (B) Strains that harbor the alternative (non-reference) allele are plotted by telomere length on the y-axis and genomic position on the x-axis. Both synonymous and nonsynonymous variants are labeled. Variants resulting in a nonsynonymous coding change are bolded. The blue line indicates the median telomere-length value for wild isolates. The color of boxplots and markers indicates variants from the same haplotypes. (C) Boxplot of natural isolate distribution of telomere lengths. Blue lines within the center of each box represent the median while the box represents the IQR from the 25th–75th percentile. Whiskers extend to 1.5× the IQR above and below the box. Plotted points represent individual strains. (D) Telomere length is plotted on the y-axis as in (B), but strains do not share mutations because strains harbor unique collections of induced alleles. The blue line indicates median telomere length for the MMP population. (E) Boxplot of the distribution of telomere lengths in the MMP is shown. Boxplot follows same conventions as in (C). N2 telomere length in our population was estimated to be 16.9 kb, whereas median telomere length in MMP was estimated to be 4.94 kb. This disparity is likely caused by differences in library preparation, sequencing platform, and data processing.
pot-2(tm1400) deletion strain. Moreover, this amino acid change could plausibly alter the function of the OB fold within POT-2. Nucleic acid recognition of OB folds occurs through a variety of molecular interactions, including aromatic stacking (Theobald et al. 2003). A change from phenylalanine to isoleucine would eliminate a potential aromatic stacking interaction and presumably reduce the binding affinity and function of POT-2. It would be interesting to determine if natural variation in either of the OB-fold domains of hPOT1 contribute to the natural variation in telomere lengths among diverse individuals in the human population.

We wondered why additional genes involved in the regulation of telomeres were not identified from our study of telomere lengths across wild isolates and mutagenized laboratory strains. Homologs for both telomerase and shelterin complex components are found in C. elegans (Stein et al. 2001). We identified natural variation in trt-1 but only in the highly diverged strains ECA36 and QX1211. These rare alleles are removed from the GWA mapping because we require allele frequencies to be greater than 5%. Laboratory mutants in trt-1 have short telomeres (Cheung et al. 2006; Meier et al. 2006), but we do not see enrichment of trt-1 mutations in the MMP collection for short or long telomeres. C. elegans contains orthologous genes for two of the six shelterin complex members, hPOT1 and RAP1 (Harris et al. 2009). Four C. elegans genes with homology to hPOT1 have been identified (mrt-1, pot-1, pot-2, and pot-3) (Raices et al. 2008; Meier et al. 2009), and C. elegans rap-1 is homologous to human RAP1 (Raices et al. 2008; Meier et al. 2009). The genes rap-1 and pot-3 had no variants or only rare variants, respectively. All of the other homologous genes contained variants in 5% or more of the wild isolates. None of these genes mapped by GWA besides pot-2, and none of the mutations in these genes were enriched in short- or long-telomere strains from the MMP collection. Perhaps shorter telomere strains are less fit and do not survive well in the wild or during the growth of mutant MMP strains. These results suggest that long telomeres are likely of limited consequence compared to short telomeres in natural settings. Additionally, because TelSeq provides an average estimate of telomere length, it is possible for the variance of telomere lengths to increase without affecting average length estimates. For this reason, we might not detect a QTL at pot-1, which has been previously reported to result in longer but more heterogeneous telomeres (Raices et al. 2008).

Our observation that considerable telomere-length variation in the wild isolate population exists allowed us to directly test whether variation in telomere length contributes to organismal fitness. We did not see any correlation between telomere length and offspring production, suggesting that fitness in wild strains is not related to telomere length. In contrast to findings in human studies, we did not identify a relationship between telomere length and longevity. Our results confirm past findings that telomere length is not associated with longevity in a small number of C. elegans wild isolates or laboratory mutants (Raices et al. 2005). Although the effects of telomere length on longevity have been observed in a well-controlled study of the gene hrp-1 on isogenic populations in the laboratory (Joeng et al. 2004), this study differs from our results in wild isolates. The different genetic backgrounds of wild isolates could further complicate a connection of telomere length to longevity because of variable modifier loci and definitively noisy longevity assays. Even though we did not identify a correlation between telomere length and either longevity or offspring production under laboratory conditions, our study suggests a limited role for telomeres in postmitotic cells. Furthermore, the population genetic results do not strongly support evidence of selection on pot-2 variants.

In summary, this study demonstrates that a variant in pot-2 likely contributes to phenotypic differences in telomere length among wild isolates of C. elegans. The absence of evidence for selection on the alternative alleles at the pot-2 locus and the lack of strong effects on organismal fitness traits suggest that differences in telomere length do not substantially affect individuals at least under laboratory growth conditions. Additionally, our study demonstrates the ability to extract and use phenotypic information from sequence data. A number of approaches can be employed to examine other dynamic components of the genome, including mitochondrial and
ribosomal DNA copy numbers, the mutational spectrum, or codon biases. These traits present a unique opportunity to identify how genomes differ among individuals and the genetic variants underlying those differences.

**Acknowledgments**

We thank Joshua Bloom and members of the Andersen laboratory for critical comments on this manuscript. We also thank M. Barkoulas, T. Bélcard, D. Bourchis, N. Callemeyn-Torre, S. Carvalho, J. Dumont, L. Frézal, C.-Y. Kao, L. Lokmane, I. Ly, K. Ly, A. Paaby, J. Røksen, and G. Wang for isolating new wild C. elegans strains. The National Bioresource Project provided the FX1400 strain, and Wormbase data made a variety of analyses possible. This work was supported by a National Institutes of Health R01 subcontract to E.C.A. (GM-107227), the Chicago Biomedical Consortium with support from the Searle Funds at the Chicago Community Trust, and an American Cancer Society Research Scholar grant to E.C.A. (127313-RSG-15-135-01-DD), along with support from the Cell and Molecular Basis of Disease training grant (T32GM008061) to S.Z. and from the National Science Foundation Graduate Research Fellowship (DGE-1324585) to D.E.C.

**Literature Cited**


both with susceptibility to colorectal cancer and with longer telomeres. Gut 61: 248–254.
Vulliamy, T., A. Marrone, R. Szydlo, A. Walne, P. J. Mason et al., 2004 Disease anticipation is associated with progressive telomere shortening in families with dyskeratosis congenita due to mutations in TERC. Nat. Genet. 36: 447–449.

Communicating editor: V. Reinke
The Genetic Basis of Natural Variation in *Caenorhabditis elegans* Telomere Length

Figure S1. Rescuing properties of intestinally expressed PITR-1::GFP. A) Expression pattern of PITR-1::GFP in *C. elegans* intestine. B) Intestinally expressed PITR-1::GFP fails to rescue YP170::RFP overexpression phenotype. Scale bars (A) 20μm, (B) 100μm.
Figure S2. Expression of YP170 in N2 and *pitr-1* mutant worms. Analysis of YP170 levels in WT and *pitr-1* mutant worms by SDS-PAGE and Western Blotting. d1 and d5 indicate days of adulthood. Actin was used as a loading control.
<table>
<thead>
<tr>
<th></th>
<th>N2</th>
<th>ppw-1 (pk2505)</th>
<th>ppw-1 (pk1425)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L4440</td>
<td>RNAi#1</td>
<td>RNAi#2</td>
<td>L4440</td>
</tr>
<tr>
<td>RNAi#1</td>
<td>RNAi#1</td>
<td>RNAi#2</td>
<td>L4440</td>
</tr>
<tr>
<td>RNAi#2</td>
<td>RNAi#1</td>
<td>RNAi#2</td>
<td>L4440</td>
</tr>
</tbody>
</table>

---

**Figure S3.** SDS-PAGE and WB analysis of YP170 expression in control and *pitr-1* RNAi worms. RNAi was performed as described in the methods and F1 adults of N2 and *ppw-1* mutant strains were analysed for YP170 expression. Actin was used as a loading control.
Figure S4. *pitr-1* is broadly expressed in *C. elegans*. Expression of GFP driven by the *pitr-1* promoter is indicated in A, pharynx and nerve ring; B, body-wall muscle; C, intestine, and nerve cord; D, ventral nerve cord; E, vulval muscle; F, sheath cells; G, spermatheca; H, tail neurons. Scale bar, 10μm.