The genetic basis of natural variation in *C. elegans* telomere length

Cook, D.C.[1,2], Zdraljevic, S.[1,2], Tanny, R.E.[1], Seo, B.[3], Riccardi D.D.[4,5] Noble L.M.,[4,5] Rockman M.V.[4,5], Alkema, M.J.[6], Braendle, C[7]., Kammenga, J.E.[8], Wang, J.[9], Kruglyak L.[10,11], Félix, MA.[12], Lee, J.[3,13], and Andersen, E.C.[1,14,15,16,*]

1. Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208
2. Interdisciplinary Biological Science Program, Northwestern University, Evanston, IL 60208
3. Department of Biological Sciences, Institute of Molecular Biology and Genetics, Seoul National University, Seoul 08826, Korea.
4. Department of Biology, New York University, New York, NY 10003
5. Center for Genomics and Systems Biology, New York University, New York, NY 10003
6. Department of Neurobiology, University of Massachusetts Medical School, Worcester, MA 01605
7. Univ. Nice Sophia Antipolis, CNRS, Inserm, IBV, Parc Valrose, 06100 Nice, France
8. Laboratory of Nematology, Wageningen University, Wageningen, The Netherlands
9. Biodiversity Research Center, Academia Sinica, Nankang Taipei, Taiwan
10. Departments of Human Genetics and Biological Chemistry, University of California at Los Angeles, Los Angeles, CA 90095
11. Howard Hughes Medical Institute, Chevy Chase, MD 20815
12. Institut de Biologie de l'École Normale Supérieure, CNRS-Inserm, Paris, France
13. Department of Biophysics and Chemical Biology, Seoul National University, Seoul 08826, Korea.
14. Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Chicago, IL 60611
15. Chemistry of Life Processes Institute, Northwestern University, Evanston, IL 60208
16. Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208

* Corresponding Author

Corresponding author:

Erik C. Andersen

Dept. of Molecular Biosciences

Northwestern University

2205 Tech Dr.

Evanston, IL 60208

erik.andersen@northwestern.edu

847-467-4382

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 oligosaccharide/oligonucleotide-binding fold of 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*.

Introduction

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 (Project *et al.* 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 that 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 Jr. *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 non-clonal human leukocytes, telomere lengths have been shown to be highly heritable (Broer *et al.* 2013). Quantitative trait loci (QTL) identified from human genome-wide association (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 *Saccharomyces 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 2-9 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 the computational technique can be applied across this large number of wild strains. Using association mapping, we found that variation in the gene *pot-2* is correlated with differences in *C. elegans* telomere length. 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 fitness traits, including brood size and longevity. Our results indicate that variation in *pot-2* does not correspond with 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 utilized to understand the dynamic nature of genomes within populations.

104

105 # Results

106

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

109

110 Previous genome-wide analyses of *C. elegans* population diversity used single-
111 nucleotide variants (SNVs) ascertained from only two strains (Rockman and Kruglyak
112 2009), reduced representation sequencing that only studied a fraction of the genome
113 (Andersen *et al.* 2012), or were limited to a small set of wild strains (Thompson *et al.*
114 2013). To address these limitations, we sequenced the whole genomes of a collection of
115 208 wild strains (Supplementary File 1). Because *C. elegans* reproduction occurs
116 primarily through the self fertilization of hermaphrodites, highly related individuals
117 proliferate in close proximity to one another (Barrière and Félix 2005; Félix and Braendle
118 2010). As a result, strains isolated from similar locations in nature are frequently
119 identical and share genome-wide haplotypes or isotypes. Sequencing data generated
120 from strains belonging to the same isotype can be combined to increase depth of
121 coverage and to improve downstream analyses. To identify which strains shared the
122 same genome-wide haplotypes, we compared all of the variation identified in each of
123 the 208 strains to each other in pairwise comparisons. The 208 strains reduce to 152
124 unique genome-wide haplotypes or isotypes (Supplementary File 1). The combination of
125 sequence data from all strains that make up an isotype led to a 70-fold median depth of
126 coverage (Supplementary Figure 1), enabling the discovery of single nucleotide variants
127 (SNVs) and other genomic features. The number of SNVs in comparisons of each
128 isotype to the reference strain N2 ranged from strains highly similar to N2 to 402,436
129 SNVs (Supplementary Figure 2), and the density of SNVs across the genome
130 matched previous distributions with more variants on chromosome arms than
131 centers (Supplementary Figure 3). A clustering analysis of these 152 isotypes
132 recapitulated the general relationships previously identified among a set of 97 wild

isotypes (Andersen *et al.* 2012) (Supplementary Figure 4) with the addition of 55 new isotypes. Past studies identified one highly diverged strain isolated from San Francisco, 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 and a median telomere length of 12.25 kb (Supplementary File 2). The Telseq telomere length estimate for N2 was 16.97 kb, which is higher than previous estimates (Wicky *et al.* 1996). 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 (Supplementary Figure 5). 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 (Supplementary Figure 6).

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, quantitative PCR (qPCR) of telomere hexamer sequences, and fluorescence *in situ* hybridization (FISH) analyses. Using twenty strains, we found that the results from these molecular assays correlated well (*rho* = 0.445 TRF, 0.815 FISH, 0.699 qPCR, Spearman's rank correlation) with computational estimates of telomere lengths (Figure 2, Supplementary File 3). 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 quantitative trait loci (QTL) on the right arm of chromosome II (Figure 3, Supplementary File 4). To identify the variant gene(s) that underlie this QTL, we investigated the SNVs within a large genomic region (12.9 to 15.3 Mb) surrounding the most significant marker on chromosome II. This region contains 557 protein-coding genes (Supplementary File 5), 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. Thirty-four genes possessed variation that was most highly correlated with telomere length (*rho* ≥ 0.4, Supplementary File 5). The chromosome II QTL explains 28.4% of the phenotypic variation in telomere length. Three additional suggestive QTL on chromosomes I, II, and

192 III were detected close to but below the significance threshold. Taken together, the four
193 QTL explain 56.7% of the phenotypic variation in telomere length.
194
195 **Variation in *pot-2* underlies differences in telomere length**
196
197 One of the 34 genes in the chromosome II large-effect QTL is *pot-2* (Protection Of
198 Telomeres 2), a gene which was previously implicated in regulation of telomere length
199 (Raices *et al.* 2008; Cheng *et al.* 2012; Shtessel *et al.* 2013). Given the large number of
200 genes present within our confidence interval and challenges associated with examining
201 telomere length using traditional genetic approaches, we sought alternative methods to
202 confirm that variation in *pot-2* could cause long telomeres. A quantitative
203 complementation test could be used to confirm that wild strains have the same
204 functional effect as a *pot-2* deletion. However, differences in telomere length caused by
205 mutations in genes that encode telomere-associated proteins often do not have
206 observable telomere defects for a number of generations (Vulliamy *et al.* 2004;
207 Armanios *et al.* 2005; Marrone *et al.* 2005). It is technically not feasible to keep the
208 genome heterozygous during long-term propagation. Fortunately, the ability to
209 computationally estimate telomere length allowed us to further validate our approach
210 using data from the Million Mutation Project (MMP) (Thompson *et al.* 2013) and
211 examine whether the equivalent of a mutant screen for telomere length would provide
212 insight into our result examining wild isolate genomes.
213
214 The MMP generated over two thousand mutagenized strains using the laboratory N2
215 background. After each strain was passaged by self-mating of hermaphrodites for ten
216 generations, the strains were whole-genome sequenced to identify and to predict the
217 effects of induced mutations. The MMP data set can be used to identify correlations of
218 phenotype and mutant genes in the laboratory strain background. We obtained whole-
219 genome sequence data from 1,936 mutagenized N2 strains, each of which has a unique
220 collection of mutations. Importantly, ten generations of self-propagation of these
221 mutagenized strains prior to sequencing likely allowed telomere lengths to stabilize in

222 response to mutations in genes that regulate telomere length, enabling us to observe
223 differences. TelSeq returned telomere length estimates for this population, which had a
224 right long-tailed distribution (Figure 4A). We classified 39 of 1936 strains within the
225 population as long-telomere strains with telomere lengths greater than 6.41 kb (98th
226 percentile). Reasoning that certain mutant genes would be overrepresented in these 39
227 strains compared to the others, we performed a hypergeometric test to identify if
228 enrichment for particular genes in long-telomere strains existed. After adjusting for
229 multiple statistical tests, we identified *pot-2* as highly enriched for mutations in six of the
230 39 long-telomere strains ($p = 2.69e^{-11}$, Bonferroni corrected, Figure 4B). No other genes
231 within any of the QTL intervals or any other part of the genome were enriched for
232 mutations among long-telomere strains. This approach was different from association
233 mapping and identified the same locus regulating telomere length. Additionally, we
234 computationally examined telomere length from whole-genome sequencing of a *pot-2*
235 knockout strain. This strain possesses a large deletion that spans the first and second
236 exons of *pot-2* likely rendering it nonfunctional. We propagated this mutant strain for ten
237 generations prior to whole-genome sequencing and TelSeq analysis. The telomere
238 length of *pot-2(tm1400)* mutants was calculated to be 30.62 kb. Given these data, we
239 have three independent tests that indicate that variation in *pot-2* likely underlies natural
240 differences in telomere lengths across the *C. elegans* species.
241
242 Our results are consistent with the established role of *pot-2* as an inhibitor of telomere
243 lengthening (Shtessel *et al.* 2013). However, no connection of *pot-2* to natural variation
244 in telomere lengths has been described previously. POT-2 contains an OB-fold thought
245 to interact with telomeric DNA. OB-folds are involved in nucleic acid recognition (Flynn
246 and Zou 2010), and the OB-fold of the human POT-2 homolog (hPOT1) binds telomeric
247 DNA (Lei *et al.* 2004). We investigated the variant sites altered in the *C. elegans*
248 species along with the mutations found in the MMP mutagenized strains (Figure 5). We
249 found that the natural variation in *pot-2* resulted in a putative phenylalanine-to-isoleucine
250 (F68I) change in the OB-fold domain of 12 strains. Strains with the POT-2(68I) allele
251 have long telomeres on average, whereas strains with POT-2(68F) allele have normal

252 length telomeres on average. Synonymous variants or variation outside of the OB-fold
253 domain were rarely found in strains with long telomeres. Because loss of *pot-2* is known
254 to cause long telomeres (Raices *et al.* 2008; Cheng *et al.* 2012; Shtessel *et al.* 2013),
255 the F68I variant likely reduces or eliminates the function of *pot-2*. Additionally, six out of
256 the 39 long-telomere MMP strains had mutations in *pot-2*, including five strains that had
257 mutations within or directly adjacent to the OB-fold and an additional strain with a
258 nonsense mutation outside the OB-fold domain that likely destabilizes the transcript.
259 These data support the hypothesis that *pot-2* is the causal gene underlying variation in
260 telomere lengths across the *C. elegans* species.
261
262 **Natural variants in *pot-2* do not have detectable fitness consequences**
263
264 We connected genetic variation in the gene *pot-2* with telomere-length differences
265 across *C. elegans* wild strains. Specifically, an F68I variant in the putative telomere-
266 binding OB-fold domain might cause reduction of function and long telomeres. A variety
267 of studies have observed a relationship between telomere length and organismal
268 fitness, including longevity or cellular senescence (Harley *et al.* 1992; Heidinger *et al.*
269 2012; Soerensen *et al.* 2012). Our results with natural variation in telomere lengths
270 provided a unique opportunity to connect differences in the length of telomeres with
271 effects on organismal fitness. We measured offspring production for our collection of
272 152 wild strains and found no correlation with telomere length (*rho*=0.062, Figure 6A).
273 Long telomeres allow for increased replicative potential of cells (Harley *et al.* 1992), but
274 it is unclear how the replicative potential of individual cells contributes to organismal
275 longevity phenotypes (Hornsby 2007). We chose nine strains covering the range of
276 telomere-length differences and found no correlation with longevity (*rho*=-0.008, Figure
277 6B, Supplementary Figure 7). Taken together, these results suggest that the long
278 telomeres found in some wild *C. elegans* strains do not have significant fitness
279 consequences in these laboratory-based experiments.
280

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 (Supplementary Figure 8). Furthermore, the haplotypes that contain this variant are rare (Supplementary Figure 9) and not geographically restricted (Supplementary Figure 10) Like the measurements of organismal fitness and lack of correlation with telomere-length differences, the population genetic measures for selection indicate 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. 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). Additionally, 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. Importantly, 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 *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

310 interactions, including aromatic stacking (Gatzeva-topalova *et al.* 2011). A change from
311 phenylalanine to isoleucine would eliminate a potential aromatic stacking interaction and
312 presumably reduce the binding affinity and function of POT-2. Somatic mutations in
313 individuals with chronic lymphocytic leukemia were found to be concentrated within the
314 OB-folds of hPOT1 (Ramsay *et al.* 2013).
315
316 We wondered why additional genes involved in the regulation of telomeres were not
317 identified from our study of telomere lengths across wild isolates and mutagenized
318 laboratory strains. Homologs for both telomerase and shelterin complex components
319 are found in *C. elegans* (Stein *et al.* 2001). We identified natural variation in *trt-1* but
320 only in the highly diverged strains ECA36 and QX1211. These rare alleles are removed
321 from the GWA mapping, because we require allele frequencies to be greater than 5%.
322 Laboratory mutants in *trt-1* have short telomeres (Cheung *et al.* 2006; Meier *et al.* 2006),
323 but we do not see enrichment of *trt-1* mutations in the MMP collection for short or long
324 telomeres. *C. elegans* contains orthogolous genes for two of the six shelterin complex
325 members, hPOT1 and RAP1 (Harris *et al.* 2009). Four *C. elegans* genes with homology
326 to hPOT1 have been identified (*mrt-1*, *pot-1*, *pot-2*, and *pot-3*) (Raices *et al.* 2008; Meier
327 *et al.* 2009), and *C. elegans rap-1* is homologous to human RAP1 (Raices *et al.* 2008;
328 Meier *et al.* 2009). The genes *rap-1* and *pot-3* had no variants or only rare variants,
329 respectively. All of the other homologous genes contained variants in 5% or more of the
330 wild isolates. None of these genes mapped by GWA besides *pot-2*, and none of the
331 mutations in these genes were enriched in short- or long-telomere strains from the MMP
332 collection. Perhaps shorter telomere strains are less fit and do not survive well in the
333 wild or during the growth of mutant MMP strains. These results suggest that long
334 telomeres are likely of limited consequence compared to short telomeres in natural
335 settings. Additionally, because TelSeq provides an average estimate of telomere length,
336 it is possible for the variance of telomere lengths to increase without affecting average
337 length estimates. For this reason, we might not detect a QTL at *pot-1*, which has been
338 previously reported to result in longer but more heterogeneous telomeres (Raices *et al.*
339 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 background effects of wild isolate variation along with telomere-length variation could obfuscate a direct connection to longevity. 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 post-mitotic 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 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 to utilize 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.

Materials and Methods

Strains

C. elegans strains were cultured using bacterial strain OP50 on a modified nematode growth medium (NGMA, 1% agar, 0.7% agarose) to prevent burrowing of wild isolates (Andersen *et al.* 2014). Strain information is listed in Supplementary File 1. 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). The provided protocol was followed with the addition of RNAse (4 µl of 100 mg/ml) following the initial lysis for two minutes at room temperature. DNA concentration was determined using the Qubit dsDNA BR Assay Kit (Invitrogen). Libraries were generated using the Illumina Nextera Sample Prep Kit and indexed using the Nextera Index Kit. Twenty-four 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 300-500 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 HS 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 Supplementary File 6.

Alignment, variant calling, and filtering

FASTQ sequence data has been deposited under NCBI Bioproject accession PRJNA318647. Sequences were aligned to WS245 (http://www.wormbase.org) using 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 www.elegansvariation.org/Data. To determine which type of SNV caller would perform best on our dataset and to set appropriate filters, we simulated variation in the N2 background. We used bamsurgeon (github.com/adamewing/bamsurgeon), which modifies base calls to simulate variants at specific positions within aligned reads and then realigns reads to the reference genome using BWA. We simulated 100,000 SNVs in 10 independent simulation sets. Of the 100,000 sites chosen in each simulation set, bamsurgeon successfully inserted an average of 95,172.6 SNVs. Using these 10 simulated variant sets, we tested two different methods of grouping our strains for

variant calling: calling strains individually (comparing sequences from a single strain to the reference) or calling strains jointly (comparing all strains in a population to each other). After grouping, bcftools has two different calling methods: a consensus caller (specified using -c), and a more recently developed multiallelic 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 less than -2 or greater than 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 rates using our simulation data set (Supplementary File 7, Supplementary File 8). Given *C. elegans* predominantly asexual mode of reproduction, we decided to focus on the individual-based calling method that performed better. Next, we determined the optimal filters to maximize true positive (TP) rates and minimize false positive (FP) and false negative (FN) results using our simulated data (Supplementary File 8). 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 (Supplementary Figure 11). Variants with DP <= 10, MQ <= 40, QUAL < 30, and DV/DP < 0.5 were called NA. Using these filters, we called 1.3M SNVs across 152 isotypes. This data set is available at 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 4,256 sites we were able to call in regions that were sequenced using Sanger sequencing, we correctly identified 4,223 variants (99.2% of all variants) in CB4856. One true positive was erroneously filtered and two false positives were removed using our filters, and we failed to call the non-reference allele for 30 variants (false negatives).

Additionally, we examined sequence variants with poor parameters in terms of depth, quality, heterozygosity, or modification by our heterozygous polarization filter. We used primer3 (Rozen and Skaletsky 1998) 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 true positives (TP) and 14 true negatives (TN). We successfully removed 3/11 false positives (FP) and erroneously filtered two sites that should have been called as non-reference (FN). We also validated the variant responsible for the F68I change in JT11398.

Identification of clonal sets

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 sequencing 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 greater than 99.93% identical across 1,589,559 sites were classified as isotypes (Supplementary Figure 12). Because LSJ1 and N2 share a genome-wide genotype but exhibit distinct phenotypes (Sterken *et al.* 2015), we treated each strain as a separate isotype. We found the following isotype differences from the

previous characterization of a large number of strains. JU360 and JU363 were previously thought to be separate, but highly related, isotypes. We found that, at the genome-wide level and at high depths of coverage, these strains are from the same isotype. Several wild strains isolated before 2000 had different genome-wide haplotypes compared to strains with the same names but stored at the *Caenorhabditis* Genetics Center (CGC). CB4851 from the CGC had a different genome-wide haplotype compared to a strain with the same name from Cambridge, UK. We renamed the CB4851 strain from the CGC as ECA243. By contrast, the version from Cambridge, UK was 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, UK is different 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, UK. Therefore, we renamed CB4858 from Cambridge, UK 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. Therefore, we generated an imputed SNV set using beagle (version r1399) (Browning and Browning 2016). This imputed variant set is available at 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 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 appear 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 is equal to the GC content of the telomeric repeat divided by the 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 non-telomeric reads, we examined the frequencies of non-cyclical permutations of the *C. elegans* hexamer in the N2 laboratory strain using TelSeq (Supplementary Figure 13). 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 2-71 copies (Supplementary File 9). 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 sequencing 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 estimates for all runs of a given strain based

548 on total reads to calculate telomere-length estimates. Supplementary File 10 details
549 telomere length estimates for every sequencing run.
550
551 **Quantitative PCR assays for telomere-length measurements**
552
553 Telomere lengths were measured by qPCR as described previously with some
554 modifications (Cawthon 2009). Primer sequences were modified from the vertebrate
555 telomere repeat (TTAGGG) to use the *C. elegans* telomere repeat (TTAGGC):
556
557 telG: 5'-ACACTAAGCTTTGGCTTTGGCTTTGGCTTTGGCTTAGTCT-3'
558 telC: 5'-TGTTAGGTATGCCTATGCCTATGCCTATGCCTAT**GCCTAA**GA-3'
559
560 The internal control, *act-1*, was amplified using the following primer pair:
561
562 forward: 5'-GTCGGTATGGGACAGAAGGA-3'
563 reverse: 5'-GCTTCAGTGAGGAGGACTGG-3'
564
565 Two primer pairs were amplified separately (singleplex qPCR). All the samples were run
566 in triplicate. qPCR was performed using iQ SYBR green supermix (BIORAD) with
567 iCycler iQ real-time PCR detection system (BIORAD). After thermal cycling, C_t (cycle
568 thresholds) values were exported from BIORAD iQ5 software.
569
570 **Terminal restriction fragment (TRF) Southern blot assay**
571
572 Animals were grown on 100 mm Petri dishes with NGM seeded with OP50.
573 Synchronized adult animals were harvested and washed four times with M9 buffer.
574 Pelleted animals were lysed for four hours at 50℃ in buffer containing 0.1 M Tris-Cl (pH
575 8.5), 0.1 M NaCl, 50 mM EDTA (pH 8.0), 1% SDS, and 0.1 mg/mL proteinase K. DNA
576 was isolated by phenol extraction and ethanol precipitation. DNA was eluted with buffer
577 containing 10 mM Tris (pH 7.5) and 1 mM EDTA. DNA was then treated with 10 μg/mL

boiled RNase A. DNA was again isolated with phenol extraction and ethanol precipitation. Five µg of DNA was digested with *Hinf*I at 37℃ overnight. Telomere restriction fragment was blotted as described previously (Seo *et al.* 2015) (Supplementary Figure 14). Digoxigenin-labeled $(TTAGGC)_4$ oligonucleotides were used as probes. Digoxigenin probes were detected with DIG nucleic acid detection kit (Roche). Blots were imaged with ImageQuant LAS4000 (GE healthcare).

Fluorescence *in situ* hybridization assays

Fluorescence *in situ* hybridization (FISH) was performed as previously described (Seo *et al.* 2015). Embryos were isolated by bleaching synchronized adult animals using standard methods (Stiernagle 2006). Isolated embryos were fixed in 2% paraformaldehyde (PFA) for 15 minutes at room temperature (RT) on a polylysine treated glass slide. The slide was put on dry ice and freeze-cracked. The embryos were permeabilized in ice-cold methanol and acetone for 5 minutes each. The slides were washed with 1X phosphate buffered saline containing 0.1% Tween-20 (PBST) three times for 15 minutes each at room temperature. 10 *µ*l of hybridization buffer (50 nM Cy3-$(TTAGGC)_3$ peptide nucleic acids probe (PANAGENE), 50% formamide, 0.45 M sodium chloride, 45 mM sodium citrate, 10% dextran sulfate, 50 µg/mL heparin, 100 µg/mL yeast tRNA, 100 µg/mL salmon sperm DNA) was added on the slide. The samples were denatured on a heat block at 85℃ for three minutes. After overnight incubation at 37℃, the samples were washed in the following order: 1X PBST once for five minutes at room temperature, 2X SSC (0.3 M sodium chloride, 30 mM sodium citrate) in 50% formamide once for 30 minutes at 37℃, 1X PBST three times for 10 minutes each at room temperature. The samples were incubated in DAPI and mounted in anti-bleaching solution (Vectashield). The samples were imaged with a confocal microscope (LSM700, Zeiss). Telomere spots were quantified with TFL-TELO software (Dr. Peter Lansdorp, Terry Fox Laboratory, Vancouver) (Poon SS1, Martens UM, Ward RK 1999).

607

608 **Genome-wide association (GWA) mapping**

609

610 GWA mapping was performed on marker genotype data and telomere-length estimates
611 using the rrBLUP package (version 4.3) (Endelman 2011) and GWAS function. rrBLUP
612 requires a kinship matrix and a SNV set to perform GWA. We generated a kinship
613 matrix using our imputed SNV set with the A.mat function within rrBLUP. Genomic
614 regions of interest were determined empirically from simulating a QTL that explained
615 20% of the phenotypic variance at each marker in our mapping data set. All simulated
616 QTL were mapped within 100 markers (50 markers to the left and 50 markers to the
617 right) of the simulated marker position. To generate a SNV set for mapping, we again
618 used our imputed SNV set. However, we filtered the number of SNVs to a set of 38,688
619 markers. This set was generated by lifting over (from WS210 to WS245) a set of 41,888
620 SNVs previously used for GWA mapping (Andersen *et al.* 2012) and filtering our
621 imputed SNVs to those sites.

622

623 **Million Mutation Project analysis**

624

625 Whole-genome sequence data from mutagenized strains within the Million Mutation
626 Project (MMP) was obtained from the sequence read archive (SRA, project accession
627 number SRP018046). We removed 59 strains that were contaminated with other strains.
628 We also were unable to locate the sequence data for 12 MMP strains on SRA, leaving
629 us with a total of 1,936 mutagenized strains. Within the MMP project, read lengths
630 varied among sequencing runs, being either 75 bp or 100 bp. We ran TelSeq on all
631 sequencing runs assuming 100 bp reads. To utilize 75 bp sequencing runs, we took the
632 448 strains that were sequenced at both 75 and 100 bp lengths and used those
633 estimates to develop a linear model. Then, this model was used to transform 75 bp
634 length estimates to 100 bp estimates (Supplementary Figure 15). We then used the
635 weighted average of telomere length estimates for all runs of a given strain based on
636 total reads to calculate telomere-length estimates. Because telomeric reads resemble

PCR duplicates, TelSeq utilizes 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 estimates from the MMP sequence 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 was 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 (Supplementary File 11) using the *phyper* function in R (R Core Team 2013). FX1400 was propagated for ten 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 fourth larval stage 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 seven 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.

667

668 **High-throughput fecundity assays**

669

670 Assays were performed similar to previously reported (Andersen *et al.* 2015) with the
671 following differences. Animals were bleached, synchronized, and grown to L4 larvae in
672 96-well plates. From the L1 to L4 stage, animals were fed 5 mg/mL of a large-scale
673 production HB101 lysate in K medium (Boyd *et al.* 2010) to provide a stereotyped and
674 constant food source. Then, three L4 larvae from each of the 152 genotypes were
675 dispensed using a COPAS BIOSORT instrument to wells containing 10 mg/mL HB101
676 lysate in K medium and progeny were counted 96 hours later. Fecundity data were
677 calculated using 12 samples – triplicate technical replicates from four biological
678 replicates. The data were processed using COPASutils (Shimko and Andersen 2014)
679 and statistically analyzed using custom R scripts.

680

681 **Clustering of relatedness**

682

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

689

690 ## Acknowledgements

691

692 We would like to thank Joshua Bloom and members of the Andersen laboratory for
693 critical comments on this manuscript. We also thank M. Barkoulas, T. Bélicard, D.
694 Bourc'his, N. Callemeyn-Torre, S. Carvalho, J. Dumont, L. Frézal, C.-Y. Kao, L.
695 Lokmane, I. Ly, K. Ly, A. Paaby, J. Riksen, G. Wang for isolating new wild *C. elegans*

696 strains. The National Bioresource Project provided the FX1400 strain, and Wormbase
697 data made a variety of analyses possible. This work was supported by an NIH R01
698 subcontract to E.C.A. (GM107227), the Chicago Biomedical Consortium with support
699 from the Searle Funds at the Chicago Community Trust, and an American Cancer
700 Society Research Scholar Grant to E.C.A. (127313-RSG-15-135-01-DD), along with
701 support from the National Science Foundation Graduate Research Fellowship to D.E.C.
702 (DGE-1324585).

703 # References

704

705 Andersen E. C., Gerke J. P., Shapiro J. A., Crissman J. R., Ghosh R., Bloom J. S., Félix
706 M.-A., Kruglyak L., 2012 Chromosome-scale selective sweeps shape
707 Caenorhabditis elegans genomic diversity. Nat. Genet. **44**: 285–290.
708 Andersen E. C., Bloom J. S., Gerke J. P., Kruglyak L., 2014 A Variant in the
709 Neuropeptide Receptor npr-1 is a Major Determinant of Caenorhabditis elegans
710 Growth and Physiology. PLoS Genet. **10**.
711 Andersen E. C., Shimko T. C., Crissman J. R., Ghosh R., Bloom J. S., Seidel H. S.,
712 Gerke J. P., Kruglyak L., 2015 A Powerful New Quantitative Genetics Platform,
713 Combining Caenorhabditis elegans High-Throughput Fitness Assays with a Large
714 Collection of Recombinant Strains. G3 (Bethesda). **5**: 911–20.
715 Armanios M., Chen J.-L., Chang Y.-P. C., Brodsky R. a, Hawkins A., Griffin C. a,
716 Eshleman J. R., Cohen A. R., Chakravarti A., Hamosh A., Greider C. W., 2005
717 Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in
718 autosomal dominant dyskeratosis congenita. Proc. Natl. Acad. Sci. U. S. A. **102**:
719 15960–15964.
720 Barrière A., Félix M.-A., 2005 *Natural variation and population genetics of*
721 Caenorhabditis elegans*. WormBook (2005). doi:10.1895/wormbook.1.43.1*.
722 Blackburn E. H., 1991 Structure and function of telomeres. Nature **350**: 569–573.
723 Bolger A. M., Lohse M., Usadel B., 2014 Trimmomatic: A flexible trimmer for Illumina
724 sequence data. Bioinformatics **30**: 2114–2120.
725 Boyd W. A., McBride S. J., Rice J. R., Snyder D. W., Freedman J. H., 2010 A high-
726 throughput method for assessing chemical toxicity using a Caenorhabditis elegans
727 reproduction assay. Toxicol. Appl. Pharmacol. **245**: 153–159.
728 Broer L., Codd V., Nyholt D. R., Deelen J., Mangino M., Willemsen G., Albrecht E., Amin
729 N., Beekman M., Geus E. J. C. de, Henders A., Nelson C. P., Steves C. J., Wright
730 M. J., Craen A. J. M. de, Isaacs A., Matthews M., Moayyeri A., Montgomery G. W.,
731 Oostra B. A., Vink J. M., Spector T. D., Slagboom P. E., Martin N. G., Samani N. J.,
732 Duijn C. M. van, Boomsma D. I., 2013 Meta-analysis of telomere length in 19,713

733 subjects reveals high heritability, stronger maternal inheritance and a paternal age
734 effect. Eur. J. Hum. Genet. **21**: 1163–8.
735 Browning B. L., Browning S. R., 2016 Genotype Imputation with Millions of Reference
736 Samples. Am. J. Hum. Genet. **98**: 116–126.
737 Cawthon R. M., 2009 Telomere length measurement by a novel monochrome multiplex
738 quantitative PCR method. Nucleic Acids Res. **37**: 1–7.
739 Cheng C., Shtessel L., Brady M. M., Ahmed S., 2012 Caenorhabditis elegans POT-2
740 telomere protein represses a mode of alternative lengthening of telomeres with
741 normal telomere lengths. Proc. Natl. Acad. Sci. **109**: 7805–7810.
742 Cheung I., Schertzer M., Baross A., Rose A. M., Lansdorp P. M., Baird D. M., 2004
743 Strain-specific telomere length revealed by single telomere length analysis in
744 Caenorhabditis elegans. Nucleic Acids Res. **32**: 3383–3391.
745 Cheung I., Schertzer M., Rose A., Lansdorp P. M., 2006 High incidence of rapid
746 telomere loss in telomerase-deficient Caenorhabditis elegans. Nucleic Acids Res.
747 **34**: 96–103.
748 Cingolani P., Platts A., Wang L. L. L., Coon M., Nguyen T., Wang L. L. L., Land S. J., Lu
749 X., Ruden D. M., 2012 A program for annotating and predicting the effects of single
750 nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila
751 melanogaster strain w 1118; iso-2; iso-3. Fly (Austin). **6**: 80–92.
752 Codd V., Nelson C. C. P. C., Albrecht E., Mangino M., Deelen J., Buxton J. L., Hottenga
753 J. J., Fischer K., Esko T., Surakka I., Broer L., Nyholt D. R., Mateo Leach I., Salo
754 P., Hägg S., Matthews M. K., Palmen J., Norata G. D., O'Reilly P. F., Saleheen D.,
755 Amin N., Balmforth A. J., Beekman M., Boer R. a de, Böhringer S., Braund P. S.,
756 Burton P. R., Craen A. J. M. de, Denniff M., Dong Y., Douroudis K., Dubinina E.,
757 Eriksson J. G., Garlaschelli K., Guo D., Hartikainen A.-L., Henders A. K., Houwing-
758 Duistermaat J. J., Kananen L., Karssen L. C., Kettunen J., Klopp N., Lagou V.,
759 Leeuwen E. M. van, Madden P. a, Mägi R., Magnusson P. K. E., Männistö S.,
760 McCarthy M. I., Medland S. E., Mihailov E., Montgomery G. W., Oostra B. a, Palotie
761 A., Peters A., Pollard H., Pouta A., Prokopenko I., Ripatti S., Salomaa V., Suchiman
762 H. E. D., Valdes A. M., Verweij N., Viñuela A., Wang X., Wichmann H.-E., Widen

763 E., Willemsen G., Wright M. J., Xia K., Xiao X., Veldhuisen D. J. van, Catapano A.
764 L., Tobin M. D., Hall A. S., Blakemore A. I. F., Gilst W. H. van, Zhu H., Consortium
765 C., Erdmann J., Reilly M. P., Kathiresan S., Schunkert H., Talmud P. J., Pedersen
766 N. L., Perola M., Ouwehand W., Kaprio J., Martin N. G., Duijn C. M. van, Hovatta I.,
767 Gieger C., Metspalu A., Boomsma D. I., Jarvelin M.-R., Slagboom P. E., Thompson
768 J. R., Spector T. D., Harst P. van der, Samani N. J., 2013 Identification of seven loci
769 affecting mean telomere length and their association with disease. Nat. … **45**:
770 422–7, 427e1–2.
771 Deng Y., Chan S. S., Chang S., 2008 Telomere dysfunction and tumour suppression:
772 the senescence connection. Nat. Rev. Cancer **8**: 450–458.
773 Ding Z., Mangino M., Aviv A., Spector T., Durbin R., 2014 Estimating telomere length
774 from whole genome sequence data. Nucleic Acids Res. **42**: 1–4.
775 Edgar R. C., 2004 MUSCLE: Multiple sequence alignment with high accuracy and high
776 throughput. Nucleic Acids Res. **32**: 1792–1797.
777 Endelman J. B., 2011 Ridge Regression and Other Kernels for Genomic Selection with
778 R Package rrBLUP. Plant Genome J. **4**: 250–255.
779 Félix M.-A., Braendle C., 2010 The natural history of Caenorhabditis elegans. Curr. Biol.
780 **20**: R965–R969.
781 Flynn R. L., Zou L., 2010 Oligonucleotide/oligosaccharide-binding fold proteins: a
782 growing family of genome guardians. Crit. Rev. Biochem. Mol. Biol. **45**: 266–275.
783 Frenck Jr. R. W., Blackburn E. H., Shannon K. M., Frenck R. W., Blackburn E. H.,
784 Shannon K. M., 1998 The rate of telomere sequence loss in human leukocytes
785 varies with age. Proc. Natl. Acad. Sci. U. S. A. **95**: 5607–10.
786 Fulcher N., Teubenbacher A., Kerdaffrec E., Farlow A., Nordborg M., Riha K., 2014
787 Genetic Architecture of Natural Variation of Telomere Length in Arabidopsis
788 thaliana. Genetics **199**: 625–635.
789 Gatbonton T., Imbesi M., Nelson M., Akey J. M., Ruderfer D. M., Kruglyak L., Simon J.
790 A., Bedalov A., 2006 Telomere length as a quantitative trait: Genome-wide survey
791 and genetic mapping of telomere length-control genes in yeast. PLoS Genet. **2**:
792 0304–0315.

793 Gatzeva-topalova P. Z., Warner L. R., Pardi A., Carlos M., 2011 Nucleic Acid
794 Recognition by Ob-Fold Proteins. **18**: 1492–1501.
795 Gordon A., Hannon G. J., 2010 Fastx-toolkit. FASTQ/A short-reads pre-processing
796 tools. Unpubl. http//hannonlab. cshl. edu/fastx_ toolkit.
797 Griffith J. D., Comeau L., Rosenfield S., Stansel R. M., Bianchi A., Moss H., Lange T.
798 De, 1999 Mammalian telomeres end in a large duplex loop. Cell **97**: 503–514.
799 Harley C. B., Vaziri H., Counter C. M., Allsopp R. C., 1992 The telomere hypothesis of
800 cellular aging. Exp. Gerontol. **27**: 375–382.
801 Harris T. W., Antoshechkin I., Bieri T., Blasiar D., Chan J., Chen W. J., Cruz N. D. La,
802 Davis P., Duesbury M., Fang R., Fernandes J., Han M., Kishore R., Lee R., Mu H.,
803 Nakamura C., Ozersky P., Petcherski A., Rangarajan A., Rogers A., Schindelman
804 G., Schwarz E. M., Tuli M. A., Auken K. Van, Wang D., Wang X., Williams G., Yook
805 K., Durbin R., Stein L. D., Spieth J., Sternberg P. W., la Cruz N. De, Davis P.,
806 Duesbury M., Fang R., Fernandes J., Han M., Kishore R., Lee R., M??ller H. M.,
807 Nakamura C., Ozersky P., Petcherski A., Rangarajan A., Rogers A., Schindelman
808 G., Schwarz E. M., Tuli M. A., Auken K. Van, Wang D., Wang X., Williams G., Yook
809 K., Durbin R., Stein L. D., Spieth J., Sternberg P. W., 2009 Wormbase: A
810 comprehensive resource for nematode research. Nucleic Acids Res. **38**: 463–467.
811 Heidinger B. J., Blount J. D., Boner W., Griffiths K., Metcalfe N. B., Monaghan P., 2012
812 Telomere length in early life predicts lifespan. : 1–6.
813 Hornsby P. J., 2007 Telomerase and the aging process. Exp. Gerontol. **42**: 575–81.
814 Joeng K. S., Song E. J., Lee K.-J., Lee J., 2004 Long lifespan in worms with long
815 telomeric DNA. Nat. Genet. **36**: 607–611.
816 Jones a. M., Beggs a. D., Carvajal-Carmona L., Farrington S., Tenesa a., Walker M.,
817 Howarth K., Ballereau S., Hodgson S. V., Zauber a., Bertagnolli M., Midgley R.,
818 Campbell H., Kerr D., Dunlop M. G., Tomlinson I. P. M., 2012 TERC
819 polymorphisms are associated both with susceptibility to colorectal cancer and with
820 longer telomeres. Gut **61**: 248–254.
821 Kwan E. X., Foss E., Kruglyak L., Bedalov A., 2011 Natural polymorphism in bul2 links
822 cellular amino acid availability with chronological aging and telomere maintenance

823 in yeast. PLoS Genet. **7**.
824 Lack J. B., Cardeno C. M., Crepeau M. W., Taylor W., Corbett-Detig R. B., Stevens K.
825 a., Langley C. H., Pool J. E., 2015 The Drosophila Genome Nexus: A Population
826 Genomic Resource of 623 Drosophila melanogaster Genomes, Including 197 from
827 a Single Ancestral Range Population. Genetics **199**: 1229–1241.
828 Lange T. De, 2010 How shelterin solves the telomere end-protection problem. Cold
829 Spring Harb. Symp. Quant. Biol. **75**: 167–177.
830 Lei M., Podell E. R., Cech T. R., 2004 Structure of human POT1 bound to telomeric
831 single-stranded DNA provides a model for chromosome end-protection. Nat. Struct.
832 Mol. Biol. **11**: 1223–1229.
833 Levy M. Z., Allsopp R. C., Futcher A. B., Greider C. W., Harley C. B., 1992 Telomere
834 end-replication problem and cell aging. J. Mol. Biol. **225**: 951–960.
835 Levy D., Neuhausen S. L., Hunt S. C., Kimura M., Hwang S.-J., Chen W., Bis J. C.,
836 Fitzpatrick A. L., Smith E., Johnson A. D., Gardner J. P., Srinivasan S. R., Schork
837 N., Rotter J. I., Herbig U., Psaty B. M., Sastrasinh M., Murray S. S., Vasan R. S.,
838 Province M. A., Glazer N. L., Lu X., Cao X., Kronmal R., Mangino M., Soranzo N.,
839 Spector T. D., Berenson G. S., Aviv A., 2010 Genome-wide association identifies
840 OBFC1 as a locus involved in human leukocyte telomere biology. Proc. Natl. Acad.
841 Sci. U. S. A. **107**: 9293–8.
842 Li H., Durbin R., 2009 Fast and accurate short read alignment with Burrows-Wheeler
843 transform. Bioinformatics **25**: 1754–1760.
844 Li H., 2011 A statistical framework for SNP calling, mutation discovery, association
845 mapping and population genetical parameter estimation from sequencing data.
846 Bioinformatics **27**: 2987–2993.
847 Liti G., Haricharan S., Cubillos F. a., Tierney A. L., Sharp S., Bertuch A. a., Parts L.,
848 Bailes E., Louis E. J., 2009 Segregating YKU80 and TLC1 alleles underlying
849 natural variation in telomere properties in wild yeast. PLoS Genet. **5**.
850 Mackay T. F. C., Richards S., Stone E. a., Barbadilla A., Ayroles J. F., Zhu D., Casillas
851 S., Han Y., Magwire M. M., Cridland J. M., Richardson M. F., Anholt R. R. H.,
852 Barrón M., Bess C., Blankenburg K. P., Carbone M. A., Castellano D., Chaboub L.,

853 Duncan L., Harris Z., Javaid M., Jayaseelan J. C., Jhangiani S. N., Jordan K. W.,
854 Lara F., Lawrence F., Lee S. L., Librado P., Linheiro R. S., Lyman R. F., Mackey A.
855 J., Munidasa M., Muzny D. M., Nazareth L., Newsham I., Perales L., Pu L.-L., Qu
856 C., Ràmia M., Reid J. G., Rollmann S. M., Rozas J., Saada N., Turlapati L., Worley
857 K. C., Wu Y.-Q., Yamamoto A., Zhu Y., Bergman C. M., Thornton K. R., Mittelman
858 D., Gibbs R. a., 2012 The Drosophila melanogaster Genetic Reference Panel.
859 Nature **482**: 173–178.
860 Malik H. S., Burke W. D., Eickbush T. H., 2000 Putative telomerase catalytic subunits
861 from Giardia lamblia and Caenorhabditis elegans. Gene **251**: 101–108.
862 Marrone A., Walne A., Dokal I., 2005 Dyskeratosis congenita: Telomerase, telomeres
863 and anticipation. Curr. Opin. Genet. Dev. **15**: 249–257.
864 McCarthy M. I., Abecasis G. R., Cardon L. R., Goldstein D. B., Little J., Ioannidis J. P. a,
865 Hirschhorn J. N., 2008 Genome-wide association studies for complex traits:
866 consensus, uncertainty and challenges. Nat. Rev. Genet. **9**: 356–369.
867 McEachern M. J., Krauskopf A., Blackburn E. H., 2000 Telomeres and their control.
868 Annu. Rev. Genet. **34**: 331–358.
869 McGrath P. T., Rockman M. V., Zimmer M., Jang H., Macosko E. Z., Kruglyak L.,
870 Bargmann C. I., 2009 Quantitative Mapping of a Digenic Behavioral Trait Implicates
871 Globin Variation in C. elegans Sensory Behaviors. Neuron **61**: 692–699.
872 McMurdie P. J., Holmes S., 2013 Phyloseq: An R Package for Reproducible Interactive
873 Analysis and Graphics of Microbiome Census Data. PLoS One **8**.
874 Meier B., Clejan I., Liu Y., Lowden M., Gartner A., Hodgkin J., Ahmed S., 2006 trt-1 is
875 the Caenorhabditis elegans catalytic subunit of telomerase. PLoS Genet. **2**: 187–
876 197.
877 Meier B., Barber L. J., Shtessel L., Boulton S. J., Gartner A., Ahmed S., 2009 The MRT-
878 1 nuclease is required for DNA crosslink repair and telomerase activity in vivo in
879 Caenorhabditis elegans. EMBO J. **28**: 3549–3563.
880 Noble L. M., Chang A. S., McNelis D., Kramer M., Yen M., Nicodemus J. P., Riccardi D.
881 D., Ammerman P., Phillips M., Islam T., Rockman M. V., 2015 Natural Variation in
882 plep-1 Causes Male-Male Copulatory Behavior in C. Elegans. Curr. Biol. **25**: 2730–

2737.

O'Sullivan R. J., Karlseder J., 2010 Telomeres: protecting chromosomes against genome instability. Nat. Rev. Mol. Cell Biol. **11**: 171–181.

Paradis E., Claude J., Strimmer K., 2004 APE: Analyses of phylogenetics and evolution in R language. Bioinformatics **20**: 289–290.

Poon SS1, Martens UM, Ward RK L. P., 1999 Telomere length measurements using digital fluorescence microscopy. Cytometry. **278**: 267–278.

Project G., Project G., Asia E., Africa S., Figs S., Tables S., 2012 An integrated map of genetic variation from 1,092 human genomes. Nature **135**: 0–9.

R Core Team, 2013 R Core Team. R A Lang. Environ. Stat. Comput. R Found. Stat. Comput. Vienna, Austria.: ISBN 3–900051–07–0, URL http://www.R–project.org/.

Raices M., Maruyama H., Dillin A., Kariseder J., 2005 Uncoupling of longevity and telomere length in C, elegans. PLoS Genet. **1**: 295–301.

Raices M., Verdun R. E., Compton S. A., Haggblom C. I., Griffith J. D., Dillin A., Karlseder J., 2008 C. elegans Telomeres Contain G-Strand and C-Strand Overhangs that Are Bound by Distinct Proteins. Cell **132**: 745–757.

Ramsay A. J., Quesada V., Foronda M., Conde L., Martínez-Trillos A., Villamor N., Rodríguez D., Kwarciak A., Garabaya C., Gallardo M., López-Guerra M., López-Guillermo A., Puente X. S., Blasco M. a, Campo E., López-Otín C., 2013 POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. Nat. Genet. **45**: 526–30.

Rockman M. V., Kruglyak L., 2009 Recombinational landscape and population genomics of caenorhabditis elegans. PLoS Genet. **5**.

Rozen S., Skaletsky H. J., 1998 Primer3. Bioinforma. Methods Protoc. Methods Mol. Biol. **3**: 1–41.

Samassekou O., Gadji M., Drouin R., Yan J., 2010 Sizing the ends: Normal length of human telomeres. Ann. Anat. **192**: 284–291.

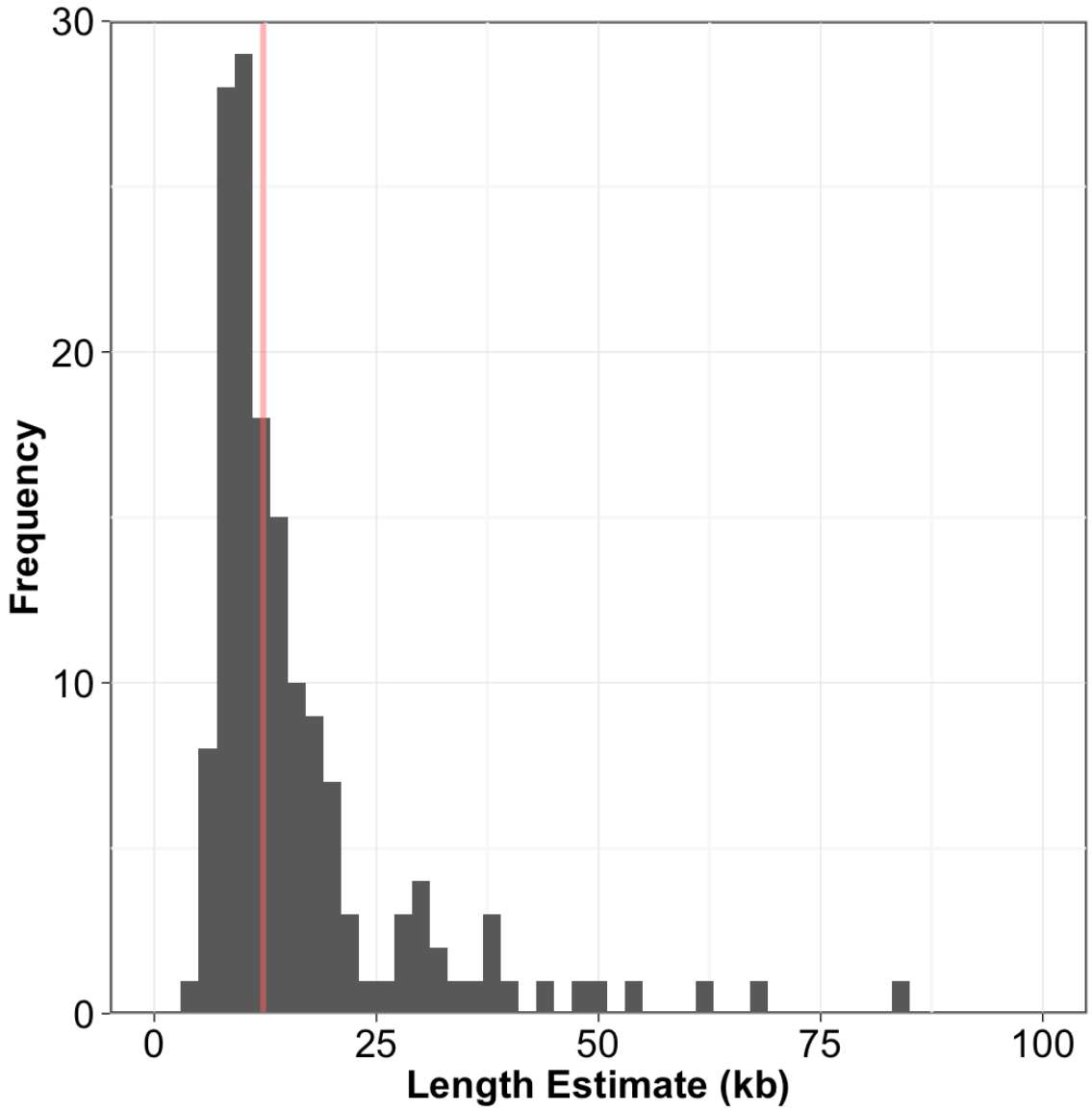
Seo B., Kim C., Hills M., Sung S., Kim H., Kim E., Lim D. S., Oh H.-S., Choi R. M. J., Chun J., Shim J., Lee J., 2015 Telomere maintenance through recruitment of internal genomic regions. Nat. Commun. **6**: 8189.

913 Shimko T. C., Andersen E. C., 2014 COPASutils: An R Package for Reading,
914 Processing, and Visualizing Data from COPAS Large-Particle Flow Cytometers.
915 PLoS One **9**: e111090.
916 Shtessel L., Lowden M. R., Cheng C., Simon M., Wang K., Ahmed S., 2013
917 Caenorhabditis elegans POT-1 and POT-2 repress telomere maintenance
918 pathways. G3 (Bethesda). **3**: 305–13.
919 Soerensen M., Thinggaard M., Nygaard M., Dato S., Tan Q., Hjelmborg J., Andersen-
920 Ranberg K., Stevnsner T., Bohr V. A., Kimura M., Aviv A., Christensen K.,
921 Christiansen L., 2012 Genetic variation in TERT and TERC and human leukocyte
922 telomere length and longevity: A cross-sectional and longitudinal analysis. Aging
923 Cell **11**: 223–227.
924 Stein L., Sternberg P., Durbin R., Thierry-Mieg J., Spieth J., 2001 WormBase: network
925 access to the genome and biology of Caenorhabditis elegans. Nucleic Acids Res.
926 **29**: 82–86.
927 Sterken M. G., Snoek L. B., Kammenga J. E., Andersen E. C., 2015 The laboratory
928 domestication of Caenorhabditis elegans. Trends Genet.: 1–8.
929 Stiernagle T., 2006 Maintenance of C. elegans. WormBook: 1–11.
930 Thompson O., Edgley M., Strasbourger P., Flibotte S., Ewing B., Adair R., Au V.,
931 Chaudhry I., Fernando L., Hutter H., Kieffer A., Lau J., Lee N., Miller A., Raymant
932 G., Shen B., Shendure J., Taylor J., Turner E. H., Hillier L. W., Moerman D. G.,
933 Waterston R. H., 2013 The million mutation project: A new approach to genetics in
934 Caenorhabditis elegans. Genome Res. **23**: 1749–1762.
935 Vulliamy T., Marrone A., Szydlo R., Walne A., Mason P. J., Dokal I., 2004 Disease
936 anticipation is associated with progressive telomere shortening in families with
937 dyskeratosis congenita due to mutations in TERC. Nat Genet **36**: 447–449.
938 Watson J. D., 1972 Origin of concatemeric T7 DNA. Nat. New Biol. **239**: 197–201.
939 Weigel D., Mott R., 2009 The 1001 genomes project for Arabidopsis thaliana. Genome
940 Biol. **10**: 107.
941 Wicks S. R., Yeh R. T., Gish W. R., Waterston R. H., Plasterk R. H., 2001 Rapid gene
942 mapping in Caenorhabditis elegans using a high density polymorphism map. Nat.

943 Genet. **28**: 160–164.

944 Wicky C., Villeneuve a M., Lauper N., Codourey L., Tobler H., Müller F., 1996
945 Telomeric repeats (TTAGGC)n are sufficient for chromosome capping function in
946 Caenorhabditis elegans. Proc. Natl. Acad. Sci. U. S. A. **93**: 8983–8988.

947 Figures



948

949 **Fig 1. Distribution of telomere-length estimates** A histogram of telomere-length
950 estimates weighted by the number of reads sequenced per run is shown. Bin width is 2.
951 The red line represents the median telomere-length estimate of 12.2 kb.

952

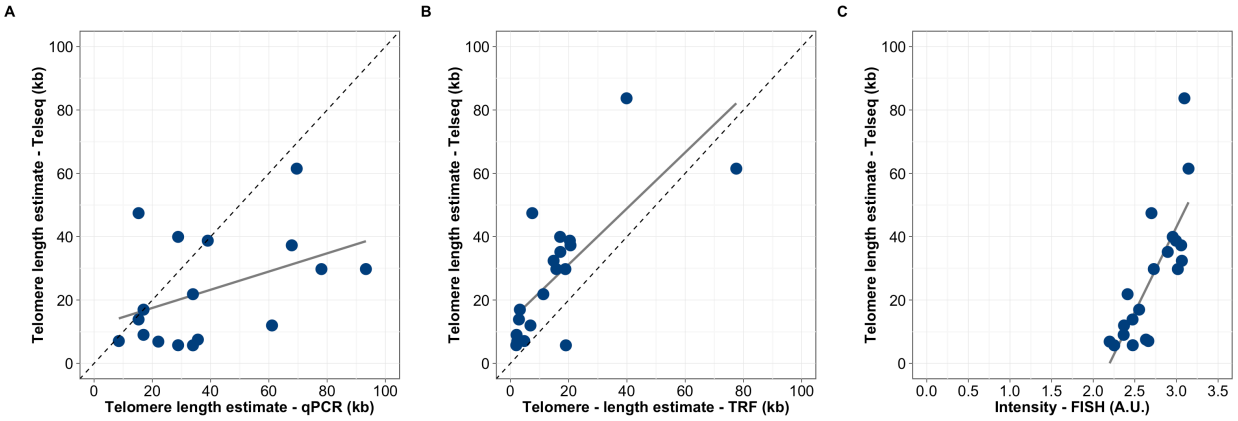


Fig 2. Telomere-length estimates correlate with alternative molecular measurement methods Scatterplot of Telseq Telomere-length estimates (y-axis) plotted against alternative methods of telomere length measurement on the x-axis. Alternative methods plotted on the x-axis and their associated Spearman's rank correlation are (A) qPCR measurements normalized by N2 qPCR value and scaled relative to the TelSeq N2 telomere length estimate (*rho* = 0.445, p = 0.049), (B) TRF (*rho* = 0.699, p = $8.5e^{-4}$) and (C) FISH (*rho* = 0.815, p = $1.03e^{-5}$). Grey lines represent the regression lines between Telseq and each method. Dashed diagonal lines represent identity lines.

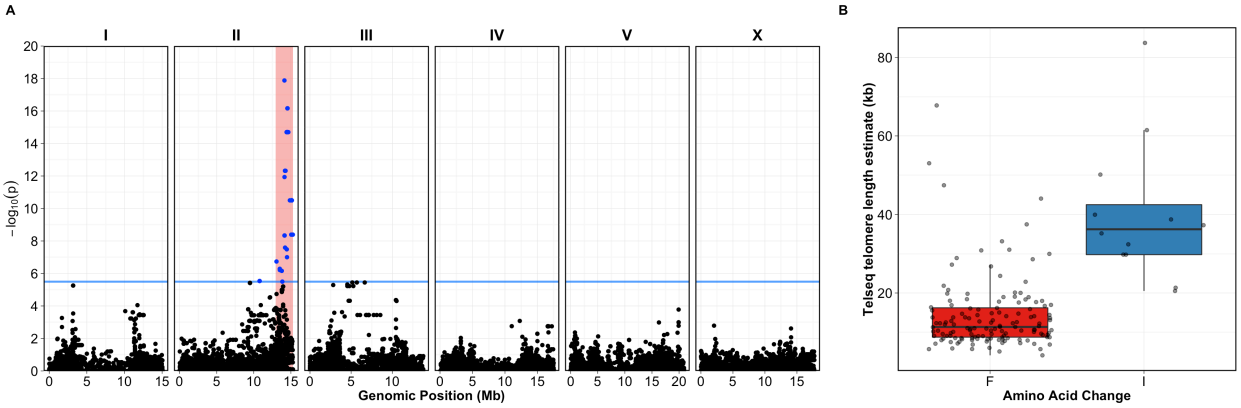


Fig 3. Genome-wide association of telomere length (A) Genome-wide association of telomere length residuals (conditioned on DNA library) is visualized using a Manhattan plot. Genomic coordinates are plotted on the x-axis against the negative of the log-transformed *p*-value of a test of association on the y-axis. The blue bar indicates the Bonferroni-corrected significance threshold (α = 0.05). Blue points represent SNVs above the significance threshold whereas black points represent SNVs below the significance threshold. Light-red regions represent the confidence intervals surrounding significantly associated peaks. (B) Shown is the split between TelSeq estimated telomere lengths (y-axis) by genotype of *pot-2* at the presumptive causative allele as boxplots (x-axis). The variant at position 14,524,396 on chromosome II results in a putative F68I coding change. Horizontal lines within each box represent the median, and the box represents the interquartile range (IQR) from the 25th-75th percentile. Whiskers extend to 1.5x the IQR above and below the box. The plotted points represent estimates beyond 1.5x the IQR.

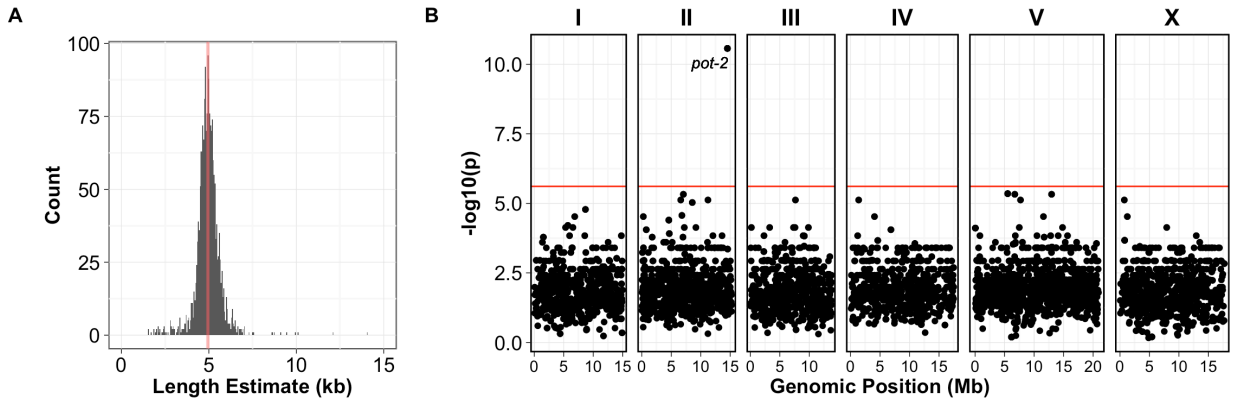


Fig 4. Mutations in *pot-2* are more often found in strains with long telomeres than in strains with short telomeres (A) A histogram of telomere-length estimates among the 1,936 mutagenized strains from the Million Mutation Project. Median telomere length is 4.94 kb. (B) Plot of significance from a hypergeometric test for every C. elegans protein-coding gene. The red line represents the Bonferroni ($\alpha = 0.05$) threshold set using the number of protein coding genes (20,447). Each point represents a gene plotted at its genomic position on the x-axis, and the log transformed *p*-value testing for enrichment of mutations in long-telomere strains.

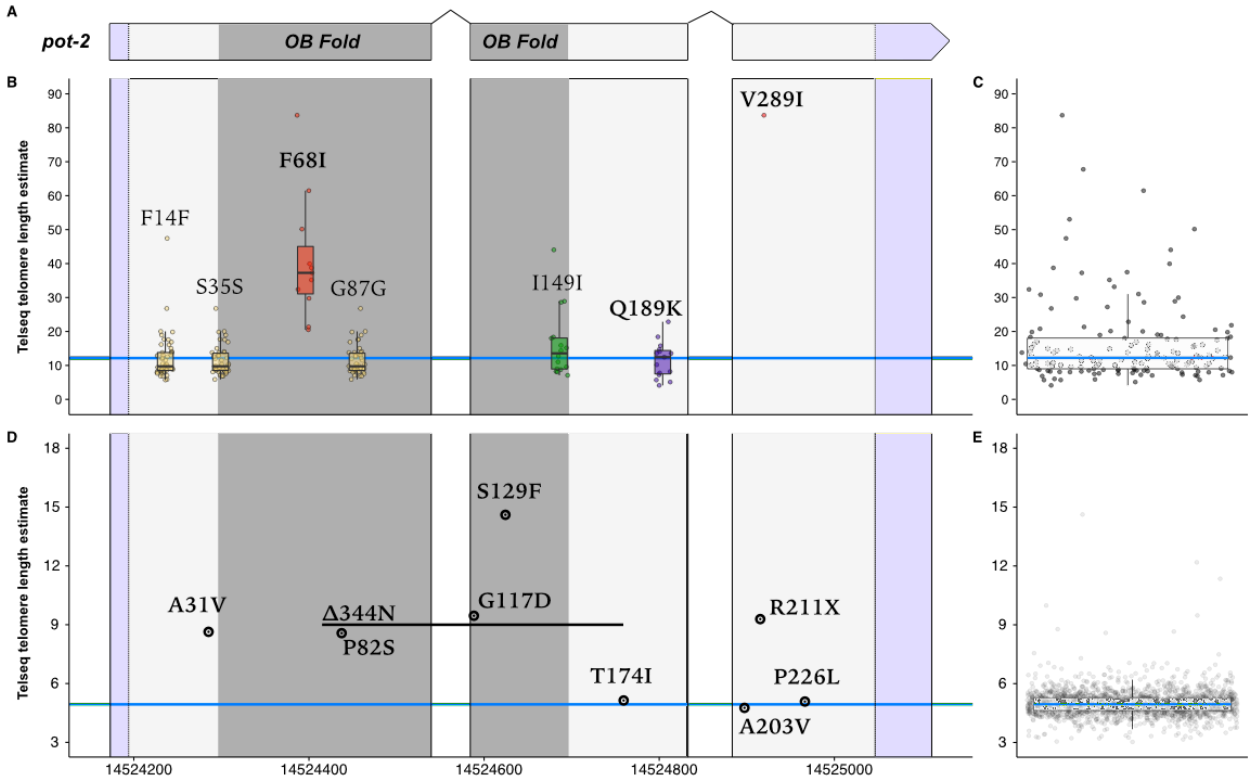


Fig 5. Variation within *pot-2* in wild isolate and Million Mutation Project strains Natural variation and induced mutations that alter codons across *pot-2* are shown along with the telomere-length estimates for all strains. In panel (A), a schematic illustrating the *pot-2* genomic region is shown. 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 interquartile range (IQR) from the 25th – 75th percentile. Whiskers extend to 1.5x the IQR above and below the box. All plotted points represent estimates beyond 1.5x the IQR. (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 wild isolate population was estimated to be 16.9 kb whereas median telomere length in MMP was 4.94 kb. Differences are likely caused by library preparation method and/or sequencing method.

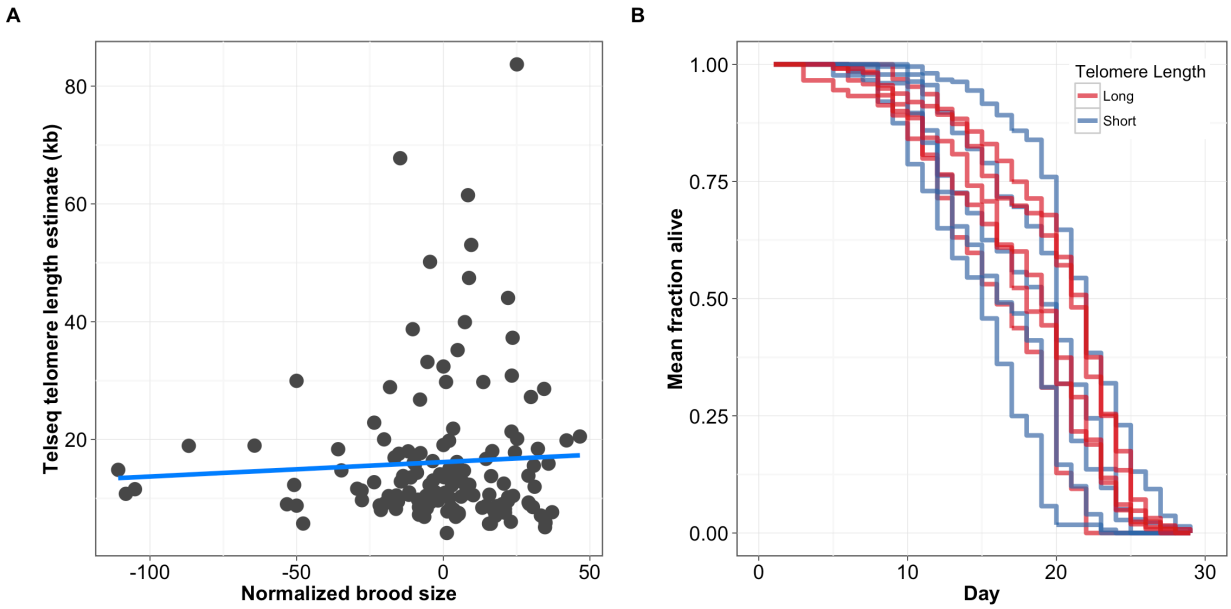


Fig 6. Fitness traits are not associated with telomere length (A) Normalized brood sizes (x-axis) of 152 wild isolates are plotted against the telomere-length estimates from those same strains (y-axis). The blue line indicates a linear fit of the data. However, the correlation is not significant (*rho* = -0.062, p = 0.463). (B) Survival curves of nine wild isolates with long and short telomeres. Lines represent aggregate survival curves of three replicates. Survival among long and short telomere length strains is not significantly different (p = 0.358; Mantel-Cox analysis).