# Genome-Wide Association Study of Extreme Longevity in *Drosophila melanogaster*

Molly K. Burke*, Elizabeth G. King, Parvin Shahrestani[1], Michael R. Rose, and Anthony D. Long

Department of Ecology and Evolutionary Biology, University of California, Irvine

[1]Present address: Department of Entomology, Cornell University, Ithaca, NY

*Corresponding author: E-mail: burkem@uci.edu.

**Accepted:** November 13, 2013

**Data deposition:** This project has been deposited at Sequence Read Archive under the accession SRP022566. SNP tables and source code are available from authors on request.

## Abstract

Human genome-wide association studies (GWAS) of longevity attempt to identify alleles at different frequencies in the extremely old, relative to a younger control sample. Here, we apply a GWAS approach to "synthetic" populations of *Drosophila melanogaster* derived from a small number of inbred founders. We used next-generation DNA sequencing to estimate allele and haplotype frequencies in the oldest surviving individuals of an age cohort and compared these frequencies with those of randomly sampled individuals from the same cohort. We used this case–control strategy in four independent cohorts and identified eight significantly differentiated regions of the genome potentially harboring genes with relevance for longevity. By modeling the effects of local haplotypes, we have more power to detect regions enriched for longevity genes than marker-based GWAS. Most significant regions occur near chromosome ends or centromeres where recombination is infrequent, consistent with these regions harboring unconditionally deleterious alleles impacting longevity. Genes in regions of normal recombination are enriched for those relevant to immune function and a gene family involved in oxidative stress response. Genetic differentiation between our experimental cohorts is comparable to that between human populations, suggesting in turn that our results may help explain heterogeneous signals in human association studies of extreme longevity when panels have diverse ancestry.

**Key words:** GWAS, synthetic populations, ageing.

## Introduction

Average human life expectancy in developed nations ranges from about 80 to 85 years, and human twin studies suggest that only 20–30% of the variation in survival within this range is determined by genetic variation (Herskind et al. 1996). Extreme longevity, often defined as surviving beyond 100 years, may be a distinct subphenotype with a larger genetic component (Perls and Terry 2003). Survival to extreme ages is rare but clusters in families (Perls et al. 2000), and relatives of centenarians have a marked delay in age-related diseases (Perls et al. 2007). In light of these observations, candidate gene studies (Ewbank 2007; Flachsbart et al. 2009) and genome-wide scans (Puca et al. 2001; Newman et al. 2010; Sebastiani et al. 2012) have been carried out on panels of extremely long-lived individuals. Association studies of survival to very old age tend to be underpowered, because sample sizes are small, typically fewer than 500 individuals, and also because panels of long-lived individuals tend to be made up of both nonagenarians and centenarians (Tan et al. 2008). These studies do not usually implicate the same loci in panels with different ancestral backgrounds, with two exceptions: 1) APOE, a known risk factor for cardiovascular and Alzheimer's disease (Schächter et al. 1994; Ewbank 2007; Sebastiani et al. 2012); and 2) FOXO3a, which encodes a regulator of the insulin-IGF1 signaling pathway (Anselmi et al. 2009; Flachsbart et al. 2009; Pawlikowska et al. 2009).

According to recent US census data, about 0.9% of males and 2.8% females survive to age 100 (U.S. Social Security Administration, 2009 Census [http://www.ssa.gov/oact/STATS/table4c6.html, last accessed November 30, 2013]). Here, we conduct a genome-wide case–control study in *Drosophila melanogaster* that compares the oldest 2% of female flies with a similar number of adult females from the same cohort sampled at a younger age. This experiment mimics human centenarian studies, providing a well-defined model for evaluating the prospects of the study of aging using

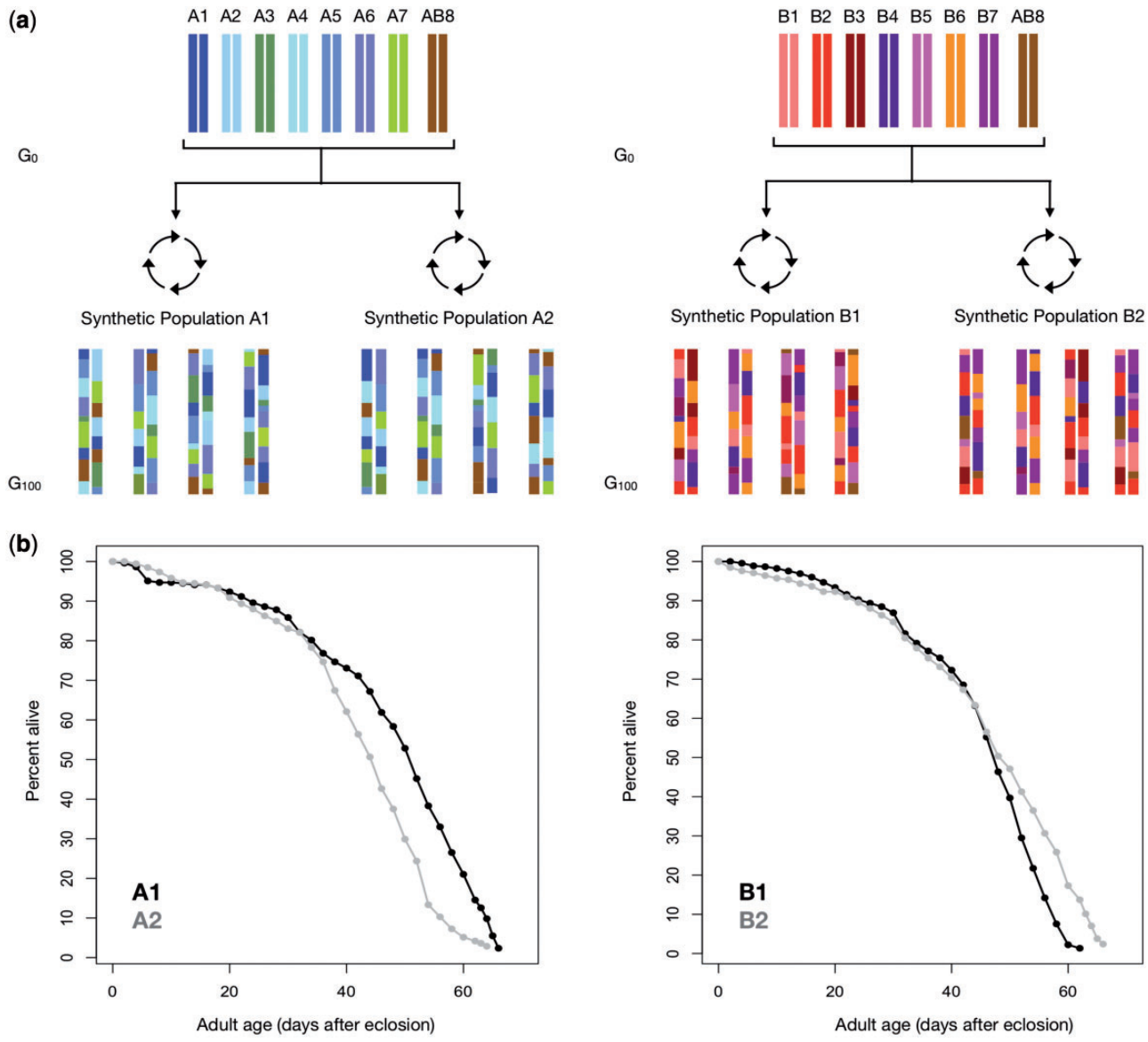


**Fig. 1.**—Overview of sampled flies. (*a*) Schematic depicting the establishment of synthetic recombinant populations used in this study (cf. King, Macdonald et al. 2012; King, Merkes et al. 2012). A1, A2, B1, and B2 were sampled for this work after approximately 100 generations of random mating. (*b*) Mortality data from females in the four "census" cages. Dead flies were sexed, counted, and removed from each population cage every other day. When only ~10 surviving females remained, these were collected alive and used for DNA extraction (thus the survivorship curve does not reach zero at the end of the assay).

relative enrichment or deficit of alleles in extremely old cohorts. We used "synthetic" recombinant populations of *Drosophila* for which extensive haplotype information is available (King, Macdonald et al. 2012; King, Merkes et al. 2012) (fig. 1*a*) to establish large cohorts (fig. 1*b*) from which we collected pools of young control adult females and extremely old female flies. We replicated the experiment four times, with replicates initiated from two pairs of synthetic populations, each pair derived from an independent set of founders (the "A" and "B" populations), with the individual populations within a pair initiated from the same founders but maintained apart for ~100 generations (A1 vs. A2 or B1 vs. B2). We then used next-generation deep resequencing of these control and old panels to compare genetic differentiation between them.

## Materials and Methods

### Experimental Populations

We used the synthetic recombinant populations of the *Drosophila* Synthetic Population Resource (DSPR) (King, Macdonald et al. 2012; King, Merkes et al. 2012) as

four aging populations from which to gather genomic data. Two independent sets of seven inbred *Drosophila* lines (the "founders") were crossed to initiate two synthetic recombinant populations A and B; and an additional inbred line was used in the founding of both (fig. 1*a*). Both A and B have been then maintained as two independent large populations (A1/A2, B1/B2) and, for this experiment, were sampled after approximately 100 generations. The founder strains were sampled from various geographic locations and are representative of genetic variation present in natural *Drosophila* populations worldwide. All genetic material for each synthetic recombinant population is thus derived from just eight founder haplotypes, and each of the 15 inbred founder strains have been resequenced to ~50× coverage. After 100 generations of maintenance, the genome of any given recombinant individual used to initiate the current experiment is a mosaic of the founder chromosomes. As part of the DSPR project, given medium-density genome-wide single-nucleotide polymorphism (SNP) genotypes for a collection of recombinant inbred lines (RILs) (or individuals) obtained from the synthetic populations, hidden Markov models have been developed which can determine at any given chromosomal position which of the eight founder haplotypes define that segment of the genome. By virtue of the founder haplotypes being completely resequenced, we thus have the ability to accurately infer the complete genome sequence of that line/individual; such in silico genome sequences are available for ~1,700 inbred lines derived from generation 50 of the synthetic population.

## Longevity Assay

The four synthetic recombinant populations, which normally are kept in large milk bottles (1.89 l), were expanded and standardized using 8-dram culture vials for three generations prior to the assay. During the standardization generations, flies spent the first 14 days of life developing in vials with cornmeal–dextrose medium. On day 14, adult flies were moved into plexiglass cages and fed media supplemented with live yeast paste to stimulate oviposition. Eggs were collected within 12-h oviposition windows to ensure that individuals in the subsequent generation were as close to the same age as possible. During the cohort assay, 120 14-day-old females (from egg) were collected from each cohort and their DNA extracted in bulk for later genomic library preparation ("control" libraries). Approximately 12,000 14-day-old individuals were then transferred from vials and equally divided into 12 plexiglass cages in which to age. Previous work has shown that at densities of 1,000 flies/cage, fly density itself has negligible impacts on mortality (Shahrestani P, unpublished data). Flies were fed fresh media every other day, and before feeding, dead flies were removed and counted. For 1 of the 12 cages per population, flies were removed, counted, and sexed every day, for the purpose of generating detailed population survivorship curves (fig. 1*b*). For the remaining 11 of the 12 cages per population, flies were removed and counted every other day; this was to ensure that excess dead flies did not accumulate in each cage and also to verify the total number of flies allocated to each cage. The last surviving ~2% of females in each cage were retained (supplementary table S1, Supplementary Material online) for genomic DNA (gDNA) library preparation ("old" libraries). We chose to ignore the oldest living male flies for the purposes of this experiment, so downstream results should be interpreted as female-specific.

## Genome Sequencing and SNP Identification

gDNA was extracted from both control and old female pools from each population using the Qiagen/Gentra Puregene kit, following the manufacturer's protocol for bulk DNA purification. The resulting eight gDNA pools were prepared as standard paired-end adapter libraries and each run as single PE75 lanes on an Illumina HiSEQ 2000. Raw reads were aligned to the reference genome sequence of *D. melanogaster* using bwa (Li and Durbin 2010). We then used samtools mpileup and the open-source code PoPoolation (version 2) to generate an allele count for each population and site (Kofler et al. 2011). All subsequent SNP-level analysis using these data tables was carried out in R (www.r-project.org, last accessed November 27, 2013).

To identify informative SNPs, we first generated a list of positions at which we expect bi-allelic SNPs to be segregating among the eight founder haplotypes contributing to a given population; all other sites were ignored. We further excluded sites in the observed data where there was some evidence of the third allele (frequency of a third allele $>0.05$ given coverage $>10$), where coverage was very high ($>2{,}000$ in control and old samples combined in A or B), or where we obtained zero coverage in either the control or old sample. Ultimately, we ended up with ~1.2 M SNPs in the A populations and ~1.1 M SNPs in the B populations. We note that as we know all the SNPs potentially segregating in the populations by virtue of their founders being known and sequenced, SNP ascertainment is straightforward.

## Quantifying Differences

### *SNP-Level Analysis*

To quantify SNP frequency differences in the data, we first calculated the absolute differences in minor allele frequencies at all of our identified SNPs. We define the minor allele as the least common allele across our eight founder haplotypes, so the minor allele is not necessarily the least common allele per position in the observed data. To take linkage into account, we calculated sliding-window averages of the absolute differences (SWAD) with a window size of 200 consecutive SNPs and a step size of 50 SNPs. As average coverage per population is high (supplementary fig. S1,

Supplementary Material online), we did not weigh the allele frequency difference by coverage.

### *Haplotype-Level Analysis*

We used our SNP data and the consensus sequences of the founders to estimate the most likely set of founder haplotype frequencies in the control and old populations across the genome. We estimated these frequencies using a 201 SNP sliding window with a step of 50 SNPs. At a given genomic position, we used the set of 100 SNPs on either side of that position to determine the most likely set of founder haplotype frequencies that would produce the observed set of SNP frequencies. We considered the set of founder haplotype frequencies that minimized the following quantity to be the most likely set of founder haplotype frequencies at the focal position:

$$\left(\sum_{i=1}^{201}\sqrt{C_i}\cdot\left(\mathrm{MAF}_i-\sum_{j=1}^{8}\left(f_{i,j}\cdot h_{i,j}\right)\right)^2\right)+\left(100\cdot 201\cdot\left(\left(\sum_{j=1}^{8}h_{i,j}\right)-1\right)^2\right) \quad (1)$$

where $C_i$ is the sequence coverage at the $i$th SNP in the window, $\mathrm{MAF}_i$ is the minor allele frequency at the $i$th SNP, $f_{i,j}$ is the allelic state (1 = minor allele, 0 = major allele) of the $j$th founder at the $i$th SNP, and $h_{i,j}$ is the haplotype frequency of the $j$th founder at the $i$th SNP. The set of eight haplotype frequencies ($h$) for the window are the quantities being optimized. Optimization was achieved using the optim function in R. The second part of the above equation serves to constrain the set of haplotype frequencies such that the sum of the eight founder haplotype frequencies cannot exceed 1 and it reduces to near zero when the founder haplotype frequencies sum to 1. Individual haplotype frequencies were bounded by 0 and 0.95. Individual haplotype frequencies were bounded by 0.95 rather than 1 because bounding by 1 occasionally led to local convergence with a single haplotype frequency equaling 1. The term preventing the sum of the haplotype frequencies exceeding 1 severely limits the search space when one haplotype frequency is equal to 1, leading to these local convergences. This phenomenon is completely prevented by bounding the individual haplotype frequencies at 0.95.

We found that the haplotype estimator was not able to accurately estimate the haplotype frequencies if two or more founder haplotypes are highly correlated (i.e., cannot be distinguished from one another over the 201 SNP window). In this case, the haplotype estimator can produce large isolated differences in $D$ between control and old pools, as for the two indistinguishable haplotypes only their sum is constrained. To prevent these spurious differences from arising, when two or more haplotypes were highly correlated with one another (>0.9), we instead estimated a single combined frequency of the correlated haplotypes using the average allelic states of the correlated haplotypes in our haplotype estimator as $f_i$ for that set of haplotypes. At these positions, our haplotype estimator produces fewer than eight founder frequencies depending on the number of correlated haplotypes.

To identify positions with overall divergent haplotype frequencies in the control versus old pools, we calculated the squared difference in haplotype frequencies between them. This measure of Euclidean distance between two vectors is a general approach but widely used in distance-based inference in biology; for example, to analyze gene expression data from microarrays (Shannon et al. 2003). Our test statistic, hereafter $D$, is essentially the average percent distance between haplotypes at a given position:

$$100\cdot\sqrt{\frac{\sum_{j=1}^{n}\left(h_{O,j}-h_{Y,j}\right)^2}{n}} \quad (2)$$

where $h_{O,j}$ is the haplotype frequency of the $j$th founder in the old population, $h_{Y,j}$ is the haplotype frequency of the $j$th founder in the younger control pool, and $n$ is the number of haplotypes estimated at that position. Typically $n$ is 8, but as noted earlier, when haplotypes are correlated, they are combined and fewer than eight frequencies are estimated. We then used the loess smoothing function in R with a span of 0.01 to smooth across genetic distance. This process tempers any highly localized fluctuations, which are not expected based on our expected size distribution of nonrecombined haplotypes at generation 100. Autosomal segment sizes are expected to follow an exponential distribution with an expected median size of 1.4 cM at generation 100, although segment sizes at generation 50 were slightly larger than the theoretical expectation.

## Determining Statistical Significance

A large set of RILs generated from the synthetic populations allowed us to generate a null distribution of both $D$ and the SWAD for the purpose of determining statistical significance. Briefly, ~500 RILs were created from each synthetic subpopulation at generation 50 via 25 generations of full-sib mating (for details on the creation of these RILs, see King, Merkes et al. 2012). The complete underlying founder haplotype structure of these RILs is known. King, Macdonald et al. (2012) describe the implementation of a hidden Markov model incorporating dense genotype data for the RILs and the founder genome sequences to infer the founder haplotype at each position in each RIL. Because the haplotype structure is known, we can infer the genotype of each RIL at each SNP by using the haplotype assignments and the founder consensus sequences. For a given RIL and position, the hidden Markov model results in a probability assignment for each founder haplotype indicating the likelihood of the genetic material at that position is derived from that founder

haplotype. The probability a given RIL harbors the minor allele ($m_z$) at a given position is

$$m_z = \sum f_j \cdot p_j \quad (3)$$

where $f_j$ is the allelic state (1 = minor allele, 0 = major allele) of the $j$th founder and $p_j$ is the probability the ancestry of the RIL is the $j$th founder at that position. The estimated minor allele frequency in a given population of RILs at a given position is the mean of $m_z$ across RILs. Therefore, for any given set of RILs, we can estimate the SNP frequency at every position and then use these SNP frequencies to calculate SWAD or $D$ as described earlier.

To generate null distributions of SWAD and $D$, we performed 5,000 iterations of a Monte Carlo simulation by sampling with replacement two sets of 200 RILs (corresponding to the young and old pools, with 200 haploid genomes the minimum number of copies in each sample, cf. supplementary table S1, Supplementary Material online) for each subpopulation. Then, we calculated the SNP frequencies in each set and ran the haplotype estimator on both sets and calculated our SWAD and $D$ statistics. Generally, the idea is to create null distributions of SWAD and $D$ that result from random sampling of two sets of 100 individuals alone (i.e., no true difference in frequency). Supplementary figure S2, Supplementary Material online, provides a visual overview of this strategy. The vectors of coverages observed in the real data were also used to estimate frequencies in the simulated data for consistency; in other words, we drew $C_i$ (the observed coverage at the $i$th SNP, in either the old or young pool) alleles with replacement from the corresponding old or young simulated pool in each Monte Carlo iteration. Thus, we model stochasticity in the random draws of genome-wide haplotypes from each synthetic population, as well as the variation in sequence coverage commonly observed in Next-Gen data sets.

Although this strategy of using Monte Carlo-based sampling of in silico sequenced RILs may not appear intuitive, we find it the most appropriate method generating appropriate null sampling distributions. For example, tests based on permuting reads between old and young pools are overly liberal when individuals are not barcoded. A test designed to detect differences between two finite pools drawn from a large population must detect differences in haplotype (or SNP) frequencies above and beyond that generated purely by the sampling process, and permuting reads between samples loses this sampling information. In addition, permuting reads between pools can potentially destroy longer range information about linkage disequilibrium, and this information is important in a study such as this one where the long-range haplotype information can be important. By contrast, the Monte Carlo RIL-based approach allows the sampling of entire haplotypes (i.e., entire RILs) and thus controls for the finite sample of individuals used to make the pools and retains important associations that should be factored into the generation of a null distribution of a test statistic. Had pooled gDNA libraries had been constructed of barcoded individuals whose genomes could be parsed out, the sequences of these entire individuals could be permuted between control and old pools. This would likely be the most appropriate strategy for determining significance, although it presents a considerable practical challenge, specifically the costs involved in making the ~1,000 libraries implied by the ~1,000 individuals examined in this study. As the technology for highly multiplexed Next-Gen sequencing projects continues to improve, this could become a viable approach in the near future.

### *SNP-Level Analysis (SWAD)*

The mean SWAD of the Monte Carlo iterations varied substantially across the genome, suggesting that the use of a genome-wide significance threshold would be overly conservative for much of the genome. Instead, we used the 5,000 SWAD values generated at each position to calculate position-specific $P$ values for each subpopulation as follows:

$$\frac{k_i+1}{N+1} \quad (4)$$

where $k_i$ is the number of iterations exceeding the observed SWAD value at position $i$ and $N$ is the total number of iterations (5,000). Computation time limited the number of iterations we were able to perform and the number of iterations places a limit on our $P$ values. The lowest $P$ value we can obtain is when $k_i$ equals zero, producing a $P$ value of 0.0002. This $P$ value is greater than what would be obtained as a 5% significance threshold after correcting for multiple tests, preventing us from obtaining individually genome-wide significant positions. Because we could not use individual $P$ values to determine significance, we identified regions enriched for low $P$ values ($<0.005$). To do this, we broke the genome into 0.5-cM intervals and determined the proportion of $P$ values below 0.005 within each interval (fig. 2).

### *Haplotype-Level Analysis*

We were able to determine genome-wide significance thresholds for the haplotype level analysis. First, we smoothed each Monte Carlo iteration in the same way as the observed data, using the loess smoothing function in R with a span of 0.01, smoothing across genetic distance. The resulting mean of the Monte Carlo iterations is quite stable across the genome, as is the per position 0.995 quantile. We then used the peak finder function msPeakSimple from the msProcess library in R with a span of 50 and a signal-to-noise threshold of 2 to identify distinct peaks across the genome. For a wide range of $D$ values, we quantified the number of distinct peaks per genome scan for each Monte Carlo iteration that exceeded that $D$. We could then calculate the number of observed peaks exceeding a given $D$ threshold per Monte Carlo genome scan. The $D$ that corresponds to 0.05 peaks per

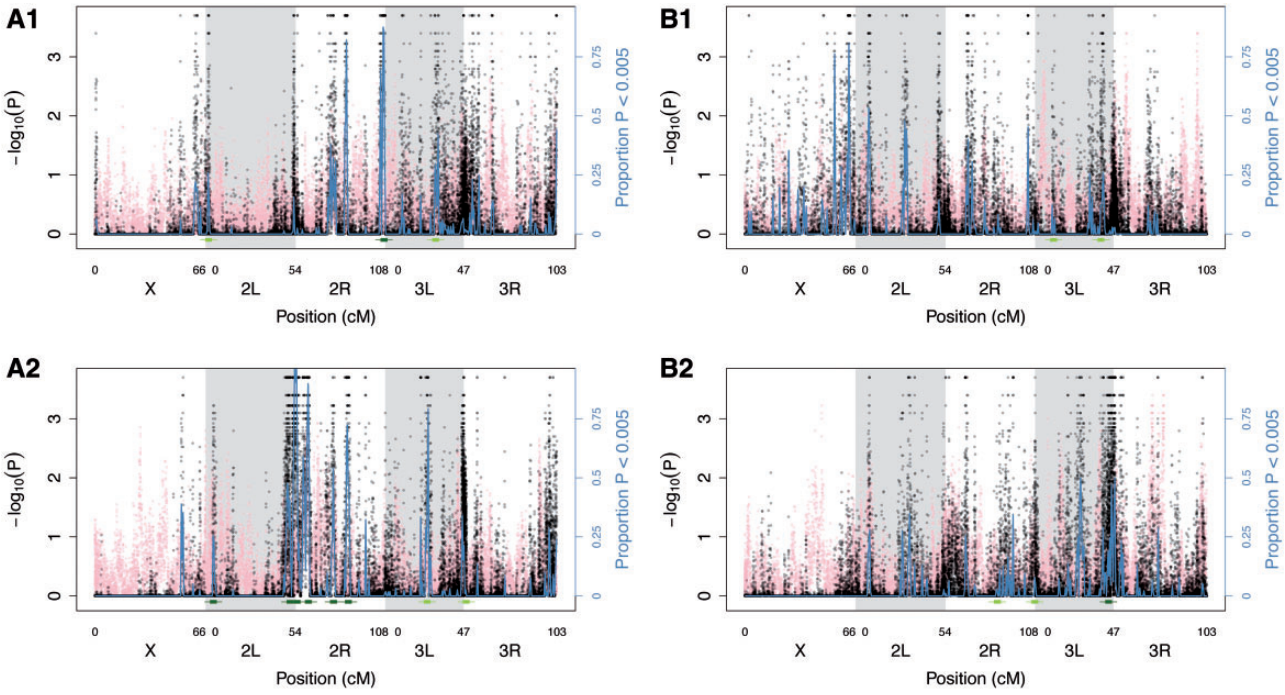


**Fig. 2.**—Empirical *P* values from the SNP-level analysis. Black points are the observed empirical $-\log_{10}(P$ values) plotted with some transparency such that multiple points plotted in the same location appear darker. Pink points are the empirical $-\log_{10}(P$ values) obtained from a single Monte Carlo iteration to show the variation obtained by chance alone. The blue line is the proportion of *P* values below 0.005 in 0.5-cM intervals across the genome. Green lines at the base of the plot denote the region of interest for any peaks exceeding the 50% threshold, as determined by the haplotype analysis (for comparison with fig. 3). Light green lines denote peaks exceeding the 50% threshold, whereas dark green lines denote peaks exceeding the 5% threshold. The thick portion of the line denotes a 2-cM interval and the thin portion denotes a 5-cM interval.

genome scan is our threshold corresponding to a 5% false-positive rate (FPR). We also present a more liberal 50% threshold (i.e., one peak every other genome scan). To localize a region of interest for each identified peak, we used a standard genetic distance of 1 cM on either side of the highest *D* value (interval spanning 2.5 cM on either side also shown; supplementary figs. S3*a–h* and S4*a–h*, Supplementary Material online).

## Results and Discussion

### Longevity Assay

The synthetic recombinant populations (fig. 1) were synchronized and used to establish cohorts of 12,000 individuals per mortality. Approximately 98% of the females in each cohort died within 62–66 days posteclosion (fig. 1*b* and supplementary table S1, Supplementary Material online), and the remaining 2% of surviving females were sacrificed alive and used to create four pooled old gDNA libraries. These pools were sequenced alongside four control gDNA libraries consisting of the same number of females sacrificed young, within 24 h of eclosion. We obtained an average coverage of approximately 80 × at the positions we considered SNPs per each of the four populations and two time-points examined (supplementary fig. S1, Supplementary Material online). The synthetic populations were founded from a total of 15 highly inbred and completely resequenced founder strains; the SNPs of this study thus correspond to those SNPs known to be segregating in the populations based on their being identified in the founders. Although many SNPs are shared between populations, alleles private to the A or B populations are common as the A and B populations are derived from different founders.

As the choice to retain the oldest 2% of surviving females was based on human population data, and not informed by *Drosophila* biology, a cursory analysis of aging trajectories in the assay populations is merited. Evolutionary theory predicts that mortality should increase in an approximately exponential fashion through the first part of adult life and then plateau at some advanced age (Mueller and Rose 1996). This plateau is called the "late-life" phase of adulthood, and it is a feature of all aging populations. The age at which this plateau starts is considered the transition between aging and late-life and is termed the "breakday" (Mueller and Rauser 2011). In human populations in developed nations, estimates of breakday range from 90 to 110, suggesting that centenarian individuals in genome-wide association studies (GWAS) panels are near the transition from aging to late-life

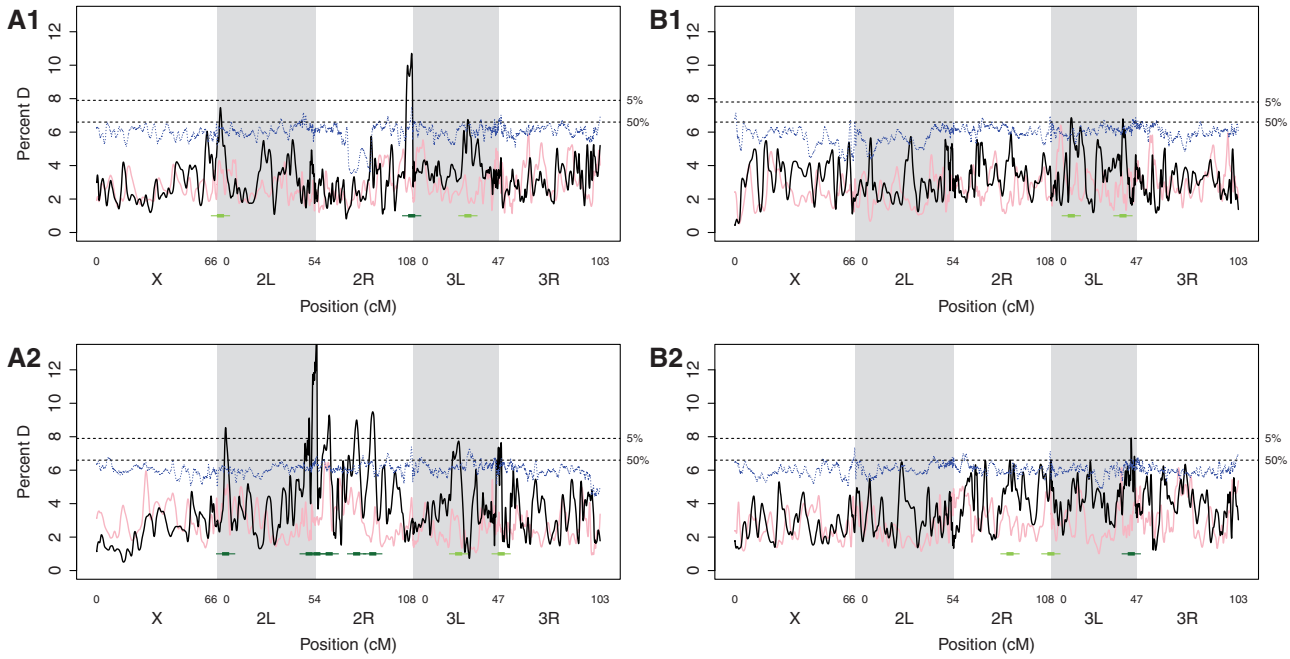


**Fig. 3.**—*D* across the genome. Different chromosome arms are denoted by different background shades (white/gray). The solid black line is the *D* for the observed data, the dotted blue line is the 99.5% per position quantile, and the dashed black lines are the genome-wide threshold for an alpha of 0.05 and 0.5 (noted as 5% and 50% on the right axis). The pink line is the *D* for a single Monte Carlo iteration to show the variation obtained by chance alone. Dark green lines at the base of the plot denote the region of interest for any peaks exceeding the 5% threshold, and light green lines denote the region of interest for any peaks exceeding the 50% threshold only. The thick portion of the line denotes a 2-cM interval and the thin portion denotes a 5-cM interval.

(Greenwood and Irwin 1939; Mueller et al. 2011). We estimate that all four of our populations were in late-life at the end of the mortality assay when individuals were sampled for sequencing and thus old enough to merit their use in an association study of extreme old age. We fit our mortality data to the evolutionary model of late-life using maximum-likelihood estimates (Shahrestani et al. 2012). We found that for all four of our populations, the estimated breakday occurred just before the time of "centenarian" sampling (supplementary fig. S5, Supplementary Material online).

## Genomic Regions of Differentiation

We identify 10–30 genomic regions, per population, of localized high SNP frequency differentiation between the control and old pools (fig. 2). Patterns of allele frequency differentiation in each population are evidently unique, although we do observe generally high SNP frequency differentiation near centromeres and at the ends of chromosomes. We used observed SNP frequencies in the pools to estimate the frequencies of the each founder haplotype at different locations across the genome (fig. 3). We developed a statistic, *D* (the Euclidean distance between the two haplotype frequency vectors), that summarizes total haplotypic divergence between control and old flies within each population. At a genome-wide alpha of 5% ($D > 7.9$), we find one significant peak in A1, six in A2, zero in B1, and one in B2 (table 1). At a more liberal genome-wide alpha of 50% ($D > 6.6$), we find two additional peaks per population (supplementary table S2, Supplementary Material online). Peaks in *D* are local and isolated, with elevated *D* values spanning <2 cM from each peak's center (table 1 and supplementary table S2, Supplementary Material online). Alternatively, analyzing SNP frequency differences alone leads to narrower peaks roughly half this size. Although this is potentially valuable in terms of localizing putatively causative variants, we frequently observe large SNP differentiation in regions without apparent high values of *D* (most notably in population B1). We regard these peaks as likely false positives, consistent with comparisons of SNP versus haplotype divergence in the context of "collaborative cross" genetic mapping studies (Valdar et al. 2005; Aylor et al. 2011). As haplotype peaks should therefore be more reliable than the peaks generated by individual SNP frequencies, we chose to only identify genes/functional groups under the more inclusive haplotype peaks.

We find a total of 1,654 genes occurring under the eight peaks significant at a 5% FPR (table 1). Five of these haplotype peaks occur in regions of suppressed recombination, near centromeres or chromosome ends (cf. Fiston-Lavier et al. 2010, fig. 1). Population genetics theory predicts that genomic regions of suppressed recombination should harbor a greater number of unconditionally deleterious alleles at higher

**Table 1**

Summary of "Peaks" Exceeding Our 5% FPR or Genomic Regions with Values of $D > 7.9$

| Pop | Chr | Peak Position (Physical) | Peak Position (cum. cM) | cM/MB | 2 cM Centered on Peak Position | *D* | PropP at Nearest 0.5 cM | No. of Genes under Peak | Supp. Figure |
|---|---|---|---|---|---|---|---|---|---|
| A2 | 2R | 2R:11,845,275 | 142.997 | 4.24 | 11,606,503–12,082,950 | 8.99 | 0.543 | 53 | 3*a* and 6*a* |
| A2 | 2R | 2R:14,325,000 | 151.864 | 3.01 | 13,980,492–14,681,997 | 9.49 | 0.477 | 90 | 3*b* and 6*b* |
| A2 | 2L | 2L:2,053,196 | 70.916 | 2.95 | 1,695,797–2,384,371 | 8.55 | 0.255 | 98 | 3*c* and 6*c* |
| A1 | 2R | 2R:20,757,872 | 173.280 | 2.00 | 20,297,228–21,100,988 | 10.71 | 0.000 | 100 | 3*d* and 6*d* |
| A2 | 2R | 2R:6,016,304 | 127.894 | 1.94 | 5,501,046–6,534,300 | 9.29 | 0.898 | 110 | 3*e* and 6*e* |
| A2 | 2L-2R | 2R:228,561 | 121.116 | 1.84 | 19,958,857–2,789,898[a] | 14.61 | 0.648 | 723 | 3*f* and 6*f* |
| A2 | 2L | 2L:15,038,696 | 116.880 | 1.17 | 14,257,961–16,026,513 | 9.12 | 0.406 | 137 | 3*g* and 6*g* |
| B2 | 3L | 3L:16,750,434 | 218.032 | 0.79 | 15,623,323–18,194,769 | 7.92 | 0.190 | 343 | 3*h* and 6*h* |

NOTE.—The peak locations in this table correspond to supplementary figs. S3*a–h* and S6*a–h*, Supplementary Material online (in order listed here). Shaded peaks are in regions of normal recombination, and nonshaded peaks are in regions of reduced combinations near centromeres or chromosome ends.

[a]Peak spans centromere.

equilibrium frequencies, because purifying selection operates less effectively in such regions due to Muller's ratchet (Muller 1964). Thus, our observation of elevated *D* values occurring in low-recombination regions of the genome is perhaps intuitive. This observation is also notable in the context of the mutation accumulation evolutionary theory of ageing, which posits that unconditionally deleterious alleles can accumulate if their fitness effects are confined to postreproductive periods of life when selection against them is weak (Medawar 1952; Charlesworth 1994). Investigators have garnered substantial empirical evidence in support of mutation accumulation, in *Drosophila* as well as other taxa (Hughes 2010). Genomic regions exhibiting high *D*, but low recombination rates, are therefore good candidates for harboring variants implicated under the mutation accumulation theory of ageing. Unfortunately, as these regions also implicate a huge number of genes, they are not amenable to singling out individual candidates. For this reason, we restrict further discussion to the three peaks located in regions of relatively high rates of recombination, >2 cM/MB (table 1, shaded region).

Notably, half of our observed significant regions occur in a single population (A2), although it is difficult to imagine a biological reason for this observation. We examined the raw genotype data from A2 for evidence of contamination from feral flies that accidentally became incorporated to the A2 sample during the assay preparation or DNA extraction. At least two lines of evidence suggest that there is no such contamination. All synthetic populations were assayed for the presence of *P* elements in the generation preceding the experiment; all were confirmed to be *P*-element free, suggesting no contamination by feral, *P*-element harboring flies. Also, less than 0.1% of the observed polymorphic positions in the A2 genome were predicted to be monomorphic from the founder consensus sequence data (i.e., site at which all founders share the same allele at that locus), and this percentage was similarly low across the four synthetic populations. This suggests that all four synthetic populations remain uncontaminated by flies of different founder ancestry.

## Genes under Significant Peaks

We find 270 genes total in the three 2-cM windows corresponding to the 5% FPR peaks in regions of normal recombination (supplementary fig. S3*a–c*, Supplementary Material online). Among the half of these 270 genes that have been annotated with Gene Ontology (GO) terms based on experimental evidence, the most common GO Biological Process term is "defense response," shared by ten genes. This term appears to be modestly enriched in the data set (10/270 genes = 3.7% vs. a genome-wide instance of 206/16,085 genes = 1.2%). The second most common term among our 270 genes is "glutathione metabolic process," but this is a special consequence of the nine genes of the glutathione transferase family that are clustered together on chromosome 2R. Although we do not think it appropriate to rely on GO term enrichment as a conclusive approach for the identification of candidate genes, we do find it useful in a general descriptive sense. Both defense response genes and the glutathione transferase family could have relevance for longevity. Most of the identified defense response genes have specific known immune functions following exposure to pathogenic bacteria. As insects age, they accumulate injuries (Burkhard et al. 2002) and are therefore more susceptible to opportunistic infection by pathogens. Variation in immune response genes could therefore causally affect longevity. Glutathione transferases inactivate damaging secondary metabolites, such as hydroperoxides, formed during oxidative stress (Hayes and Flanagan 2005). Free radical-scavenging enzymes such as superoxide dismutase (*Sod*) and catalase have long been implicated in aging (Harman 1956). Experimentally, these enzymes have been shown to increase lifespan when overexpressed in *Drosophila* (Orr and Sohal 1994), and laboratory-selected populations of flies with different evolved longevity phenotypes exhibit different frequencies of

allozymic variants with different levels of activity at such loci (Deckert-Cruz et al. 1997). Glutathione transferases fall into this category of antioxidant enzymes, and thus the genes that encode them could have consequences for lifespan.

Although our gene list is ripe for the validation of candidate longevity genes, perhaps the list's most interesting feature is an absence of genes previously described in the literature. There are 130 *D. melanogaster* genes with a priori GO terms associated with aging (www.flybase.org), and only one of these, *Atg7*, occurs in our gene list. *Atg7* is an autophagy gene known to decrease in expression with age, leading to an accumulation of damaged proteins in tissues (Demontis and Perrimon 2010). Notably, oft-cited "longevity" genes, such as *mth* (Lin et al. 1998) and *indy* (Rogina et al. 2000), shown to increase lifespan when mutated, are not associated with regions of haplotype or SNP divergence in control versus old flies. It is possible that naturally occurring variants in these genes contribute to longevity in a subtle manner and cannot be detected via our approach, or that natural variation in the expression of these genes is largely due to *trans*-regulatory genes. It is also possible that longevity-increasing mutations of large effect in these genes have pleiotropic consequences for early-life fitness components such as fecundity, preventing them from contributing to segregating variation in our test populations.

## Replication

The three regions of high differentiation in regions of normal recombination all occur in population A2. Thus, none of our candidate genes are implicated in other synthetic populations, a pattern analogous to that seen in human centenarian studies, where associations seem to be limited to single populations. However, it is fair to note that when considering regions identified at lower levels of statistical stringency (FPR $>$ 50%), we frequently observed regions of high differentiation within 2–5 cM of each other in two different populations (table 1 and supplementary table S2, Supplementary Material online). This is the case for two of the three highly significant regions identified in A2: one hit on chromosome 2R (supplementary fig. S3*b*, Supplementary Material online) is located 0.5 cM away from a less significant peak in the B2 population (supplementary fig. S4*d*, Supplementary Material online), and another hit on chromosome 2L (supplementary fig. S3*c*, Supplementary Material online) is located less than 3 cM away from a less significant peak in the A1 population (supplementary fig. S4*e*, Supplementary Material online). On one hand, this replication can be considered additional evidence for alleles affecting longevity in these genomic regions. On the other hand, the observation that regions are not implicated at the same high levels of significance in replicate populations suggests a limit to this study's ability to detect such alleles. The degree to which associations are limited to single populations due to truly independent causative variants versus a lack of power therefore remains unclear.

To address the issue of exactly how differentiated our replicate populations are in general, we calculated $F_{st}$ values for pairwise comparisons between each of our four synthetic populations using our SNP data in a standard $F_{st}$ equation (Hartl and Clark 2007). We find the pairwise comparisons between the "pure" replicates from the same original cross to be smaller than the comparisons between populations from different ancestral crosses: A1/A2 = 0.09; B1/B2 = 0.07; average A/B = 0.14 (supplementary table S2, Supplementary Material online). Although this is expected, it is notable that the $F_{st}$ values between some human populations exceed those measured in our pure replicates (Nelis et al. 2009) (supplementary table S3, Supplementary Material online). This provides some context for the general failure to replicate "extreme longevity" loci in human association studies. If we cannot implicate the same regions between *Drosophila* populations that have diverged for approximately 100 generations, it similarly may not be reasonable to expect human centenarian studies of populations with comparable degrees of shared ancestry to produce replicable results.

The degree to which we expect to see replicable differences among populations remains unclear. There are two salient considerations: 1) replication is also rare among human genome-wide association study panels of sizes comparable to those used here; and 2) the populations used in this study are not true replicates. To address the first issue, in association studies where power is low (e.g., with sample sizes of $<$500 individuals), simulation studies have shown that replication is often difficult to achieve, even when attempting to replicate associations in the *same* population (Long and Langley 1999). It is generally acknowledged that genome-wide associations typically only become replicable in panels of ~2,000 cases versus controls (Wellcome Trust Case Control Consortium 2007), and of course this sample size is difficult to achieve in studies of the extremely old.

The second issue to consider is the appropriateness of our populations as biological replicates. Although our population pairs do share ancestry, this is no guarantee that they are currently segregating the same causative alleles. To illustrate this idea, we queried a locus (chr3L:11105723) at which a nonsynonymous SNP affects the activity of *Sod*, an enzyme with known effects on aging from both transgenic experiments (Orr and Sohal 1994) and studies of selectively bred populations (Deckert-Cruz et al. 1997). This position is polymorphic in both founder populations, with a minor allele count of two of eight in the A founders and one of eight in the B founders. Despite this nonsynonymous SNP initially segregating in all four populations, we only detect this position as polymorphic in the single A2 synthetic population examined in our experiment after 100 generations of selection and drift. That is, the minor allele of *Sod* has been stochastically lost in the other three replicate populations, not

an uncommon occurrence in the DSPR (cf. King, Merkes et al. 2012, fig. 2). Although we do not observe this nonsynonymous SNP to be significantly differentiated in the A2 cohort (15% in the control vs. 23% in the old pool), it nevertheless prompts the question of how many potentially functional alleles have been lost from one of our two population pairs. To visualize this, we went through the most significant peaks (from table 1) and looked at individual founder haplotype frequencies contributing to change between the young and old pools (supplementary fig. S6*a*–*h*, Supplementary Material online). In most of these peak regions, it is very clear that certain haplotypes are estimated at near-zero frequencies in one of the control pools versus the control pool of the corresponding replicate population. For example, the peak centered on chr2R:11,845,275 is significant in A2 but not in A1. Supplementary figure S6*a*, Supplementary Material online, shows that founder haplotype 5 occurs at intermediate frequency in the control A2 pool and at a somewhat higher frequency in the A2 old pool, suggesting that this haplotype likely contributes to the significant *D* value in the A2 population. However, this haplotype is virtually not present in the A1 control pool. As haplotype 5 has been lost (or is rare) in the A1 pool, it cannot increase as the pool ages, which is perhaps why we do not estimate a significant *D* value in this replicate population. This phenomenon appears to be occurring in the majority of cases (supplementary fig. S6*a*–*g*, Supplementary Material online; for a possible counterexample, see supplementary fig. S6*h*, Supplementary Material online).

### SNP Differentiation at a Finer Scale Than Haplotype Differentiation

Although regions of high SNP differentiation are more numerous than regions with high *D* values, they tend to span smaller physical distances along the chromosome. These peaks could thus potentially be informative, when used alongside *D* peaks, for localizing putative causative regions. To evaluate this, we looked at our three major regions of interest and evaluated them for evidence of smaller SWAD peaks contained within the haplotype peak that might further localize candidate regions. For two of these peaks (supplementary fig. S3*a* and *c*, Supplementary Material online), the region implicated by SNP frequency difference is nearly identical to that implicated by *D*. However, for the third peak (supplementary fig. S3*b*, Supplementary Material online), the SNP frequency differences implicate a 1-cM region, rather than the entire 2-cM region spanned by the haplotype peak. To evaluate how this localization impacts our previous conclusions vis-à-vis our candidate genes, we removed the 1-cM region of relatively low SNP frequency differences from our analysis. This region resolves to 60 genes, and omitting these genes from our list of 270 does not change the most common GO terms observed. However, the single a priori longevity gene we observed, *Atg7*, is one of the 60 genes removed. This weakens the case for this gene being important in our study and perhaps lends more credence to the idea that so-called "aging" genes, which traditionally have been identified on the basis of mutant screens, are not reliable candidates for studies of longevity in outbred flies.

## Conclusions

So what is the relevance of this work for human association studies of centenarians? The number of our cases, ~120 extremely old individuals per population, is not as large as what we see in the best human studies. That being said, the strengths of this study are 1) four cohorts handled identically, 2) a controlled environment in which to measure the phenotype, and 3) appropriate controls collected from within the same cohort. These experimental design features are unachievable in human studies. In addition, the synthetic population resource used here provides haplotype information that allowed us to identify putative candidate regions that SNP-level analysis alone may not have identified. In our data set, the SNP-level analyses appeared to result in many more differentiated regions than the haplotype-level analysis. Previous studies contrasting haplotype-level analyses with marker-level analyses have shown that marker-level analyses can lead to spurious results through simple random sampling (Aylor et al. 2011) and that marker-level analyses are prone to larger mapping location errors (Valdar et al. 2005). Thus, there is value in applying the GWAS approach to populations derived from a limited number of founders and mapping effects back to founder haplotypes as opposed to SNPs. By focusing on a limited pool of founder haplotypes, it is apparent that different haplotypes have different effects that SNPs alone cannot enumerate. This observation may explain some of the population heterogeneity seen in human longevity GWAS panels.

Our observation that five out of eight regions with significant effects on longevity are in regions of suppressed recombination, which are much more likely to harbor unconditionally deleterious alleles of large effect with elevated minor allele frequencies than regions of normal recombination, lends support to the mutation accumulation hypothesis for variation in longevity. This observation furthermore suggests that telomeric and centromeric regions may be fruitful places to look for genes that impact longevity in humans. Finally, we associate bacterial defense and glutathione transferase genes with extreme longevity, suggesting that standing variation impacting longevity in outbred populations may have a different genetic basis than genes identified via forward screens for mutants of large effect, perhaps due to early-life trade-offs.

## Supplementary Material

Supplementary figures S1–S6 and tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

## Acknowledgments

The authors are grateful to the UCI High-Throughput Genomics Core Facility for technical help with sequencing and to the authors of PoPoolation2 for their open-source code. They thank S. Nguyen and E. Pham for help with the mortality assay. D. Rand provided valuable feedback on a previous version of this manuscript. This work was supported by National Institutes of Health (NIH) grants R01 RR024862 and R01 GM085251 (to A.D.L.) and F32 GM099382 (to E.G.K.).

## Literature Cited

Anselmi CV, et al. 2009. Association of the FOXO3A locus with extreme longevity in a southern Italian centenarian study. Rejuvenation Res. 12: 95–104.

Aylor DL, et al. 2011. Genetic analysis of complex traits in the emerging Collaborative Cross. Genome Res. 21:1213–1222.

Burkhard DU, Ward PI, Blanckenhorn WU. 2002. Using age grading by wing injuries to estimate size-dependent adult survivorship in the field: a case study of the yellow dung fly *Scathophaga stercoraria*. Ecol Entomol. 27:514–520.

Charlesworth B. 1994. Evolution in age-structured populations. Cambridge: Cambridge University Press.

Deckert-Cruz DJ, Tyler RH, Landmesser JE, Rose MR. 1997. Allozymic differentiation in response to laboratory demographic selection of *Drosophila melanogaster*. Evolution 51:865–872.

Demontis F, Perrimon N. 2010. FOXO/4E-BP signaling in *Drosophila muscles* regulates organism-wide proteostasis during aging. Cell 143: 813–825.

Ewbank DC. 2007. Differences in the association between apolipoprotein E genotype and mortality across populations. J Gerontol A Biol Sci Med Sci. 62:899–907.

Fiston-Lavier A-S, Singh ND, Lipatov M, Petrov DA. 2010. *Drosophila melanogaster* recombination rate calculator. Gene 463:18–20.

Flachsbart F, et al. 2009. Association of FOXO3A variation with human longevity confirmed in German centenarians. Proc Natl Acad Sci U S A. 106:2700–2705.

Greenwood M, Irwin JO. 1939. The biostatistics of senility. Hum Biol. 11: 1–23.

Harman D. 1956. Aging: a theory based on free radical and radiation chemistry. J Gerontol. 11:298–300.

Hartl DL, Clark AG. 2007. Principles of population genetics, 4th ed. Sunderland (MA): Sinauer Associates.

Hayes JD, Flanagan JU. 2005. Glutathione transferases. Annu Rev Pharmacol Toxicol. 45:51–88.

Herskind AM, et al. 1996. The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870–1900. Hum Genet. 97:319–323.

Hughes KA. 2010. Mutation and the evolution of ageing: from biometrics to system genetics. Philos Trans R Soc Lond B Biol Sci. 365:1273–1279.

King EG, Macdonald SJ, Long AD. 2012. Properties and power of the *Drosophila* Synthetic Population Resource for the routine dissection of complex traits. Genetics 191:935–949.

King EG, Merkes CM, et al. 2012. Genetic dissection of a model complex trait using the *Drosophila* Synthetic Population Resource. Genome Res. 22:1558–1566.

Kofler R, Pandey RV, Schlötterer C. 2011. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27:3435–3436.

Li H, Durbin RR. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595.

Lin YJ, Seroude L, Benzer S. 1998. Extended life-span and stress resistance in the *Drosophila* mutant methuselah. Science 282: 943–946.

Long AD, Langley CH. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. 9:720–731.

Medawar PB. 1952. An unsolved problem of biology. London: H.K. Lewis.

Mueller LD, Rauser CL, Rose MR. 2011. Does aging stop? New York: Oxford University Press.

Mueller LD, Rose MR. 1996. Evolutionary theory predicts late-life mortality plateaus. Proc Natl Acad Sci U S A. 93:15249–15253.

Muller HJ. 1964. The relation of recombination to mutational advance. Mutat Res. 106:2–9.

Nelis M, et al. 2009. Genetic structure of Europeans: a view from the North-East. PLoS One 4:e5472.

Newman AB, et al. 2010. A meta-analysis of four genome-wide association studies of survival to age 90 years or older: the Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium. J Gerontol A Biol Sci Med Sci. 65:478–487.

Orr WC, Sohal RS. 1994. Extension of life-span by overexpression of superoxide dismutase and catalase in *Drosophila melanogaster*. Science 263:1128–1130.

Pawlikowska L, et al. 2009. Association of common genetic variation in the insulin/IGF1 signaling pathway with human longevity. Aging Cell 8: 460–472.

Perls T, Terry D. 2003. Genetics of exceptional longevity. Exp Gerontol. 38: 725–730.

Perls T, et al. 2000. Exceptional familial clustering for extreme longevity in humans. J Am Geriatr Soc. 48:1483–1485.

Perls T, et al. 2007. Survival of parents and siblings of supercentenarians. J Gerontol A Biol Sci Med Sci. 62:1028–1034.

Puca AA, et al. 2001. A genome-wide scan for linkage to human exceptional longevity identifies a locus on chromosome 4. Proc Natl Acad Sci U S A. 98:10505–10508.

Rogina B, Reenan RA, Nilsen SP, Helfand SL. 2000. Extended life-span conferred by cotransporter gene mutations in *Drosophila*. Science 290:2137–2140.

Schächter F, et al. 1994. Genetic associations with human longevity at the APOE and ACE loci. Nat Genet. 6:29–32.

Sebastiani P, et al. 2012. Genetic signatures of exceptional longevity in humans. PLoS One 7:e29848.

Shahrestani P, Quach J, Mueller LD, Rose MR. 2012. Paradoxical physiological transitions from aging to late life in *Drosophila*. Rejuvenation Res. 15:49–58.

Shannon W, Culverhouse R, Duncan J. 2003. Analyzing microarray data using cluster analysis. Pharmacogenomics 4:41–51.

Tan Q, Zhao JH, Zhang D, Kruse TA, Christensen K. 2008. Power for genetic association study of human longevity using the case-control design. Am J Epidemiol. 168:890–896.

Valdar W, Flint J, Mott R. 2005. Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. Genetics 172: 1783–1797.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678.

**Associate editor:** Marta Wayne