www.sciencemag.org/cgi/content/full/330/6002/372/DC1

Supporting Online Material for

Selection at Linked Sites Shapes Heritable Phenotypic Variation in *C. elegans*

Matthew V. Rockman,* Sonja S. Skrovanek, Leonid Kruglyak*

*To whom correspondence should be addressed. E-mail: mrockman@nyu.edu (M.V.R.); leonid@genomics.princeton.edu (L.K.)

Published 15 October 2010, *Science* **330**, 372 (2010)
DOI: 10.1126/science.1194208

This PDF file includes:

Supporting Online Material

Materials and Methods

Microarray Analysis

We used Agilent 4x44k microarrays to measure transcript abundance in 208 RIAILs (recombinant inbred advanced intercross lines). Strains were grown at 20° following standard procedures (*S1*). Synchronization of the young adult hermaphrodites, isolation of RNA, labeling, and hybridization were performed as in Capra et al (*S2*). Dyes were assigned randomly to each sample and paired with an alternately labeled common reference (mixed stage, mixed N2:CB4856). We excluded from analyses probes that yielded fluorescence intensities near background or at saturation in more than 1/3 of the arrays, and we excluded probes that map to multiple dispersed genomic locations, determined by mapping probes with BLAT to the WS200 *C. elegans* reference genome sequence. Probes that map to multiple locations within a 20 kb window were retained. The resulting dataset included fluorescence intensities from 15,888 probes, with saturated or near-background measurements treated as missing data.

Linkage Mapping

We performed structured nonparametric interval mapping for each probe, using the marker data and methods described in Rockman and Kruglyak (*S3*). We analyzed 10 permuted datasets, each generated by permuting whole phenotype vectors to retain correlations among traits, and we used the distribution of lod scores from the permuted datasets to calculate an empirical False Discovery Rate as in references *S4* and *S5*. Analyses reported in the text used a gene-specific linear correction for dye effects. As checks on the robustness of our analysis, we analyzed each dye-class separately and summed the lod scores; results were qualitatively identical. We also tested the impact of treating low intensity measurements as missing data. When intensities indistinguishable from background were treated as low measured intensities (all set to identical low values below the lowest well-measured value), results were largely unchanged, with some exceptions for genes whose intensities cross into background range as a function of genotype. The result is that slightly more linkages are detected at a given threshold when near-background intensities are treated as low-intensity observations.

From the global analysis, detected linkages were partitioned into local and distant classes by two methods. Defining local linkages those with peak lod scores within a 1 Mb probe-centered window, 1,410 of the 2,309 linkages are local. Physical positions of QTLs were estimated by linear interpolation of physical on genetic position in intervals between genotyped SNPs. Defining local linkages as those whose 1-lod QTL support intervals encompass the probe position, 1,496 of the 2,309 linkages are local.

Hotspot Analysis

We identified linkage hotspots in two ways. First, we identified high-confidence distant linkages by performing interval mapping on the residuals of the abundance of each transcript regressed on genotype probabilities at its genomic locus; that is, we removed the effect of potential local linkage prior to interval mapping. This analysis identified 482 distant linkages at 5% FDR (lod > 4.85). We counted the number of

linkages in 5 cM bins across the genome and declared hotspots in bins containing more linkages than the largest number expected genome-wide under Poisson-distributed linkage (*S4*). The expected number of linkages per bin is 1.506, so a bin containing 8 linkages has $p < 0.05$ under a Poisson distribution. We observe 10 significant bins, but several of these are adjacent and were merged into five hotspots (following reference *S5*), representing 12 (II), 35 (IV-L), 35 (IV-R), 10 (V), and 127 (X) linkages. Because spurious linkage hotspots can arise from expression correlations (*S6*), we evaluated the presence of hotspots in permuted datasets. Two of ten permuted datasets contained a total of three hotspots (16, 16, and 58 linkages) at the specified lod threshold, but at higher significance thresholds for linkage (e.g., lod > 6, FDR < 0.1%), the permuted datasets include no hotspots while the four detected on chromosomes IV, V, and X remain significant. At such high lod scores, the number of linkages required for a bin to be a hotspot decreases, and several additional hotspots on chromosomes I, II, and V appear to be robust though they influence few genes (figure S1). The 12-linkage hotspot identified on chr II, however, appears to be very sensitive to the significance cutoff.

Next, we identified hotspots in our global analysis of expression data as above, using the set of 810 nuclear linkages classified as distant linkages on the basis of 1-lod confidence intervals. Hotspots of 10 linkages are unexpected under Poisson-distributed linkages, and we identify the four largest hotspots identified above, as well as four others (one near the hotspot on II, one on I, and two on V, neither coincident with the hotspot identified on that chromosome using the local residuals). However, at the 10-linkage threshold, the 10 permuted datasets exhibited 14 total hotspots. At a higher lod score threshold (lod > 6, FDR < 0.1%), the permuted datasets lack hotspots but the major hotspots on II and X remain, as does one on V. Additional minor hotspots also appear on V.

Results are all extremely similar if we use physical rather than genetic distance to define bins and if we vary the bin size.

Based on the analysis of permuted datasets and dependencies on the method for defining distant linkage, we consider three of the hotspots detected at the threshold required for a 5% FDR for single linkages to be robust: IV-L, IV-R, and X.

The **X hotspot** spans a much larger interval than the other hotspots, roughly 12 cM or 1.2 Mb from 4.2 to 5.4 Mb. Though *npr-1* is located within this interval at 4.77 Mb, many of the linkages to the hotspot have support intervals that do not overlap *npr-1*, and the region is likely to represent multiple causal variants.

Local linkages provide strong candidates for the hotspots on chromosome IV: *Y17G9B.8*, a histone acetyltransferase SAGA associated factor is a strong local linker at the peak of the **IV-L hotspot**, and *Y105C5A.15*, a zinc-finger transcription factor, links locally to the **IV-R hotspot** and is the only protein coding gene in the region, which otherwise is occupied by 21U small RNA genes. None of the hotspot-influenced gene sets exhibits detectable enrichment for any functional class of genes (*S7*).

We also investigated the four small hotspots that include few linkages with very strong lod scores.

The **I-L hotspot** falls in a physically large region with little recombination in our cross; the smallest confidence interval for these linkages spans 400kb from 4.24 to 4.64 Mb. The gene *ppw-1*, a candidate by virtue of its known functional polymorphism (*S8*), is nearby but is outside the CIs of the strongly linking genes. The linking transcripts have

little functional annotation. There are 4 strong local linkages to the same interval, including *met-1*, a histone methyltransferase and hence a strong functional candidate for trans-regulatory effects.

The **II-L hotspot** involves distant linkage between the hotspot and multiple transcripts that are very closely linked to one another. The CB4856 allele is associated with higher abundances of *AC8.3, AC8.4, AC8.7,* adjacent genes on the far left of X. One model to explain such clustered distant linkage is *trans*-regulation of chromatin (*S9*). An alternative model is that these clusters occur as segregating segmental duplications, with a duplicate copy of each cluster present in the genomic location of the hotspot. A third possibility is that the genes are similar to one another and exhibit cross hybridization: among the *AC8* probes, there are multiple perfect matches to sequence nearby, and in fact 30 of the 35 most distal probes on the far left tip of X were excluded from analysis because they have matches elsewhere in the genome (*AC8.3,4,7* are genes 32, 31, and 33 from the left tip).

The **V_La hotspot** includes genes scattered across the genome. There are at least three very strongly linking (lod >10) oxidoreductases. Other genes include *ins-37* and several nematode-specific genes, class *nspa*.

The **V-Lb hotspot** influences *F55G11.3, F5511.6, F5511.7,* and *dod-22,* adjacent genes on chr IV. As in the case of the II-L hotspot, this pattern could be due to trans-regulation of chromatin or segmental duplication. We investigated whether a genomic polymorphism present within *F55G11.3* exhibited segregation patterns consistent with segmental duplication; the fluorescence intensities for SNP CE4-230 (IV:12,971,747) suggested that it is present in the strains as a single bi-allelic locus. Further, two genes that are located between the strongly linking genes on IV (*F55G11.2* and *F55G11.8*) exhibit no linkage to II:70, suggesting that a single segmental duplication cannot account for the observed linkages.

Local Linkage Analysis

We tested for local linkage by calculating the lod score at the marker or pseudomarker nearest each probe, with genetic positions of probes estimated by linear interpolation. Pseudomarkers, ungenotyped genetic positions at which we have estimated genotype probabilities using the *calc.genoprob* function in *R/qtl* (*S10*), were spaced evenly at 1 cM intervals, so each probe was ≤ 0.5 cM from its nearest marker. An empirical FDR was calculated by repeating the analysis on each probe in each of the permuted datasets. Local linkages may be due to *cis*-regulatory variation or to *trans*-acting variation linked closely to the transcript whose abundance varies. The latter category includes protein-coding polymorphism that influences transcript abundance through a feedback mechanism (*S11*). For our purposes, the important feature of a local linkage is merely that its position in the genome is well defined.

Local linkage results are susceptible to hybridization artifacts if the CB4856 sequence differs from the N2 reference sequence represented on the microarrays. Indeed, we find that an excess of local linkages are associated with higher apparent expression from the reference strain, N2 (58.3%, two-tailed binomial $p < 10^{-15}$). As described in the main text and below, our analyses are robust to the exclusion of all genes that exhibit higher apparent expression for the N2 allele than for the CB4856 allele.

Genomic Heritability

To derive estimates of domain-specific genetic effects without having to detect and localize QTLs, we employed an approach that employs the variance in realized genotypic identity-by-descent among relatives. While relatives of equivalent rank have identical expected IBD proportions, the realized proportions vary due to random segregation. The degree of realized genotypic similarity can explain variation in the degree of phenotypic similarity in proportion to the genomic contribution to the relationship between genotype and phenotype; this mode of estimating heritability removes many constrictive assumptions required by other methods and has high accuracy (*S12, S13*). The approach can be applied to whole genomes and also to partitions of genomes (*S14- S16*). Our analysis included only the 8,973 probes with no missing data.

To accommodate the population structure of our RIAIL panel, we used a mixed model approach to estimate the variance components using *EMMA* (*S17*). We estimated the realized kinship matrix (*K*) from the RIAIL genotypes by first assigning genotypes to pseudomarkers every 1 kb across the 100,270 kb genome. We assigned each pseudomarker genotype using the *calc.genoprob* function in *r/qtl* and the physical position of each pseudomarker inferred using linear interpolation between the 1,454 genotyped nuclear SNPs. The genotype of the most distal marker on each chromosome end was assigned to all more distal positions. We then used the *K* matrix to estimate V_G and V_E (genetic and residual variances) for each trait by REML, and we calculated heritability as $H^2 = V_G/(V_G + V_E)$.

To test the significance of the estimated heritabilities, we repeated the analysis on 399 datasets with permuted strain labels. We estimated *p*-values from the position of the observed heritability among the heritabilities from permuted data for each trait and then used the distribution of *p*-values to estimate the False Discovery Rate using *qvalue* (*18*) with default settings. Using the entire genome to estimate realized relatedness, 1,191 of the probes exhibit significant heritability at FDR = 5% (Fig. S2). Much of the heritability is driven by strong local linkages, but analysis of the residuals of linear regressions of each trait on its genotype, designed to eliminate the effects of local linkage, yielded 232 probes with significant distant heritabilities at 5% FDR.

We defined arm and center partitions following the domain boundaries in Rockman and Kruglyak (*S3*), estimated *K* for each partition, and repeated the REML procedure for the real and permuted datasets. The variances among *K* matrix entries for the genome, the arms, and the centers are plotted in Figure S6. To accommodate any potential distortions of $V_{G,Arms} - V_{G,Centers}$ due to differences in power implied by Figure S6, we used the trait-specific empirical *p*-values for $V_{G,Arms} - V_{G,Centers}$ as our test statistic instead of the simple sign of the difference in V_G estimates. The *p*-values were strongly skewed toward both high and low values, indicating that the genetic basis of each trait is typically enriched in either arms or centers, consistent with contributions from large-effect or spatially clustered loci. We then tested whether the excess of traits with $p < 0.5$ (i.e., with genetic contributors to V_G enriched in arms) was more than expected. A straightforward application of a binomial probability is not appropriate because of correlations among traits. We therefore calculated the bias across traits in *p*-values for each of the permuted datasets, i.e., taking each dataset in turn from the 400 (399 permuted plus one experimental), treating it as the experimental dataset, and calculating the proportion of *p*-values greater and less than 0.5. These proportions formed the basis for calculating the

empirical two-tailed *p*-values for the degree of imbalance between arms and centers in contributing to trait heritability. These tests included only traits with genomic heritability, 1,191 traits in the full analysis and 232 traits in the distant analysis.

We verified that the results were not driven by the effects of the three robust QTL hotspots. We incorporated genotypes at these positions into the heritability analysis as fixed-effect covariates and re-estimated genomic heritability and $V_{G,Arms} - V_{G,Centers}$. The results are qualitatively identical. The genomic heritabilities with and without the hotspot covariates have a correlation coefficient of 0.968. The correlation coefficient for $V_{G,Arms} - V_{G,Centers}$ is 0.936. For the analysis that controls for local linkage, the genomic heritabilities with and without hotspots have a correlation of 0.916 and for $V_{G,Arms} - V_{G,Centers}$, 0.990. To assess whether the hotspots had a substantial effect on the overall pattern of arm bias, we calculated an arm bias index for each trait, with and without hotspot covariates and with and without controlling for local linkage. The index is $(V_{G,Arms} - V_{G,Centers})/(V_{G,Arms} + V_{G,Centers})$, which standardizes the bias according to the total explained genetic variance for each trait and yields a value between -1 and 1, with positive values indicating an excess of genetic variation derived from chromosomal arms. The distributions of arm bias index are almost identical across the four analyses (Figure S7).

Annotation Analysis

For most analyses, we counted each of the 15,888 probes as separate traits. 14,792 distinct WormBase WBGene identifiers are associated with 15,809 of the probes. The remaining probes map to regions annotated as intergenic, to genes annotated as 'retired' or 'transposon,' or to multiple closely linked recent duplicate genes. Of the 14,792 distinct genes, 13,922 are interrogated by a single probe, 758 by two probes, 90 by three probes, 10 by four probes, 3 by five probes and 1 by six. Most of the genes touched by multiple probes have diverse isoforms and may have independent genetic variation (*S19*). For example, of four *kin-1* probes, two exhibit no local linkage and two exhibit very strong local linkage.

For analyses in which genes were the unit of analysis, we used only the 14,415 nuclear genes assigned WBGene identifiers by WormBase for which results of RNAi experiments have been reported. For gene-level analyses we reduced multiple probe sets to a single observation; if any probes exhibited significant linkage, we scored that gene as having linkage. We calculated transcript lengths from the WormBase WS190 annotations and calculated gene interval sizes by adding the distance to the nearest exons in both 5' and 3' directions. Data on RNAi phenotypes comes from a WormMart query of WormBase WS190, collecting counts of phenotypes assayed and phenotypes observed for each gene. Point estimates of recombination rate are derived from reference *S3*. Sequence conservation is derived from the phastCons segmentation of the *C. elegans* genome into conserved and non-conserved sites (*S20*), downloaded from the UCSC Genome Bioinformatics Site (genome.ucsc.edu, database ce6, table phastConsElements6way).

As reported in the manuscript, each of these genic variables is associated in a univariate analysis with the presence or absence of local QTLs. We elaborate here on these results.

Genes with local QTLs are longer than those without local QTLs. Gene size includes the entire primary transcript, including introns, as well as the 5' and 3' flanking intergenic intervals. The median lengths of genes with and without local QTLs are 5,078 bp and 4,908 bp (*t*-test on log-transformed interval lengths $p = 0.004$). When we consider only the traits with higher transcript abundance associated with the CB4856 genotype, to control for potential hybridization artifacts, the median lengths are 4,907 and 5,272 bp ($p = 0.008$). Genes with and without local QTLs did not differ significantly in the lengths of the primary transcripts or in the lengths of the flanking intergenic intervals, only in the sum of these two lengths.

Genes that exhibit phenotypes when knocked down by RNAi are less likely to have local QTLs than genes with no RNAi phenotype (587/4497 genes with phenotypes, 1787/9918 genes without, $\chi^2 = 55.1$, $p < 2 \times 10^{-13}$). When only CB4856-high traits are considered, the effect remains (270/2216 vs 748/4869, $\chi^2 = 17.6$, $p = 4.7 \times 10^{-4}$).

Genes with local QTLs contain fewer evolutionarily constrained nucleotides than genes without (*t*-test on Box-Cox transformed values, $p < 4 \times 10^{-23}$; median conserved basepairs 1,176 vs. 1,504). The result holds for CB4856-high traits ($p < 8 \times 10^{-7}$; median conserved basepairs 1,290 vs. 1,573). These analyses used counts of conserved bases from the phastCons segmentation, Box-Cox transformed after adding one to each count, $\lambda = 1/3$.

Genes with and without local QTLs do not differ in the total amount of non-local phenotypic variance, defined as the total phenotypic variance minus the variance explained by the nearest marker or pseudomarker (*t*-test on log-transformed residual variances, $p = 0.93$). For CB4856-high traits only, there is a slight difference; traits with local QTLs have slightly lower residual phenotypic variance (median residual variances are 0.20 for traits without local QTLs and 0.18 for traits with local QTLs, $p = 0.02$).

Transcript abundance traits with local QTLs are more likely than traits without to also map to additional QTLs. These additional QTLs are distant QTLs detected from interval mapping on the residuals of trait values regressed on local genotypes. Of 2,374 traits with local QTLs, 132 (5.6%) also link elsewhere in the genome, versus 300 of 12041 traits (2.5%) that lack local QTLs ($\chi^2 = 63.2$, $p < 2 \times 10^{-15}$). For CB4856-high traits, the result is the same (76/1018 traits with local QTLs and 170/6067 traits without, $\chi^2 = 55.2$, $p < 2 \times 10^{-13}$)

Background selection model

We fit a background selection model using the standard assumption that chromosomes carry a Poisson-distributed number of mutations and the observed genetic variation (π_{BGS}) is that expected in the absence of background selection (π_0) scaled by the expected frequency of mutation-free chromosomes, the zero-mutation class from the Poisson distribution, i.e., $\pi_{BGS} = \pi_0 e^{-G}$, where G is the mean of the distribution of mutations per chromosome (*S21*). In a non-recombining chromosome at equilibrium, $G = U/2hs$, where U is the chromosomal mutation rate, s is the selection coefficient against mutations in the homozygous state, and h is the dominance coefficient. With recombination, the effect of background selection on a focal gene is a function is the integral over all linked sites of a weighted form of G, with weights determined by the

density of sites subject to deleterious mutation and the rate of recombination between each site and the focal gene (*S22*). Equation 15 from Hudson and Kaplan (*S22*) permits an approximation to the integral using a summation over discrete intervals.

We used equation 15 from Hudson and Kaplan (*S22*), with modifications to include the interval-specific density of conserved sites and a scaling factor to accommodate the reduction in effective recombination associated with partial selfing (*S23*). We estimated G for each gene *k* by summing over all linked pseudomarker intervals.

$$\hat{G}_k = \sum_i \frac{u_i s_d}{2(s_d + P\,|\,M_k - M_i\,|)(s_d + P\,|\,M_k - M_{i+1}\,|)}$$

In this formula, u_i is the fraction of U attributable to mutations in interval *i,* which we estimate from the number of evolutionarily conserved sites in the interval. We set $U = 0.3$ from the spontaneous mutation rate (~10^{-8} per site, reference *S24*) and the number of conserved bases in the genome (~29,640 kb), estimated by segmentation of the genome into conserved and non-conserved bases by the phylogenetic hidden Markov model phastCons (*S20*) . The conservation annotation was downloaded from the UCSC Genome Bioinformatics Site as described above. Background selection deals with mutations that are deterministically eliminated from the population, and hence the phastCons segmentation marks the susceptible sites within each interval.

M_k is the genetic position of the focal gene, estimated by linear interpolation from the genetic map. M_i is the genetic position of the left edge of interval *i*. We used the genetic map from reference *S3* after rescaling its expanded genetic map distances to yield 50 cM meiotic chromosomes.

s_d is a compound parameter representing the strength of selection against deleterious mutations incorporating dominance. We fixed s_d at a single value for each analysis, although real mutations exhibit an unknown distribution of deleterious effects. If the true distribution of selection coefficients is log-normally distributed, fixed s_d equal to the harmonic mean provides a reasonable approximation, with very little sensitivity to the exact shape of the distribution, for analysis of background selection (*S27*). In simulations, Loewe and Charlesworth (*S27*) found that the harmonic mean approximation recapitulated the spatial patterns found with a distribution of coefficients but slightly overstated the reduction in diversity due to background selection. Thus, given a distribution of deleterious selection coefficients, the best-fitting s_d identified in our analysis will therefore be slightly displaced (toward lower values) from the harmonic mean of the lognormal distribution that would best explain the data. For radically different forms of the distribution (e.g., multimodal), the effects of variance in s_d are unclear.

Formally, s_d in a partially selfing species is dependent on the outcrossing rate, r_{out}, if deleterious mutations are not dominant. Equation A2b from reference *S23* provides an approximation of the effective selection coefficient, given by $\tilde{s}_d = s(\hat{F} + h(1-\hat{F}))$, where the equilibrium inbreeding coefficient $\hat{F} = (1-r_{out})/(1+r_{out})$, s is the strength of selection against the homozygous mutant, and h is the dominance coefficient, ranging from 0 for recessivity to 1 for dominance. Across the range of values we consider for s_d and r_{out}, and for all values of h, this correction for inbreeding is negligible, with $\tilde{s}_d$ less than s_d by 18% in the most severe case, which is when the outcrossing rate is at its minimum and mutations are completely recessive.

The parameter *P*, the index of panmixis, rescales the genetic distances to account for the low rate of outcrossing in *C. elegans*, and is equal to 1 - $\hat{F}$ (*S23*, eq. A2). The value of this parameter depends on the frequency of outcrossing events, as defined above, but also on population structure and selection against outcross progeny (*S25*). For the case governed only by outcrossing rates, our range of *P* encompasses outcrossing rates ranging from 2.5×10^{-5} to 7.7×10^{-2}. Population genetic estimates of the effective outcrossing rate are on the order of 10^{-4} to 10^{-3} (*S26*).

There are many uncertainties associated with *P* and s_d. Consequently, we place little confidence in the best-fit parameter values; our analysis is aimed at demonstrating the robustness of background selection as an explanation across a large range of plausible values.

We performed a search of $P \times s_d$ space from 5×10^{-5} to 0.125 for each parameter at 25 intervals equally distributed in natural log space, using the likelihood ratio test (LRT) statistic to evaluate model fit. The LRT derives from a drop-one analysis of terms in a logistic regression model with local linkage as a binary dependent variable (Model 7, Table 2). The independent variables were gene interval length (log transformed), number of conserved bases in the interval (counts of conserved bases from the phastCons segmentation, Box-Cox transformed after adding one to each count, $\lambda = 1/3$), linkage to a distant locus (a binary factor), RNAi phenotype (a binary factor), and background selection effect ($\pi_{BGS}/\pi_0 = e^{-G_k}$). Distant linkage was defined as linkage at FDR $\leq$ 5% in the analysis of residuals from linear regressions on local genotypes. The LRT statistic is the difference between the residual deviance under the full model and the residual deviance under the model with the single background selection term dropped. Its significance was tested against a chi-squared distribution with one degree of freedom.

Robustness of background selection as an explanation for local linkage

We tested the robustness of our findings in several ways, detailed below. Results were extremely similar in all cases, with background selection explaining the domain effect on local linkage distribution. None of the alternative approaches to modeling or analyzing the data affected any claims about background selection. Plots of significance of background selection across $P \times s_d$ parameter space in each analysis are presented in Figure S5.

We repeated the analysis after excluding chromosomes I and X, which have distinctive map properties. Chromosome I suffered severe segregation distortion in our cross due to selection favoring the N2 allele of the paternal-effect locus *peel-1* (*S28*), and the effective recombination rate on the X chromosome has a different dependency on outcrossing rate than the autosomes. This dataset includes 10,065 genes.

We next repeated the analysis after excluding all genes that show higher expression in strains carrying the N2 allele of the locus, to control for potential hybridization (SNP-under-probe) effects. The resulting dataset includes 7,085 genes.

We then repeated the analysis excluding all N2-high genes and excluding chromosomes I and X. The resulting dataset includes 5,260 genes.

We also repeated the analyses for each of the above datasets using genomic $U = 1$ instead of $U = 0.3$.

We performed the analyses using untransformed values for interval length and number of conserved nucleotides; this approach gave uniformly poorer fit for these variables, increasing the deviance explained by background selection.

We repeated the analyses with alternative lod-score thresholds, corresponding to FDR = 0.001 and FDR = 0.2. In the former case, there are 1,616 genes with local linkage, and in the latter case, 3,427.

We also analyzed the data by multiple linear regression of log(lod scores), which are distributed normally according to a Kolmogorov-Smirnov test ($p = 0.72$). We favor the logistic regression approach because the majority of lod scores are low and reflect mostly noise. Nevertheless, the results of the linear regression were completely concordant with our logistic regression results.

The only result that was not robust is that the effect of RNAi phenotype on local linkage probability is non-significant in analyses that consider only the CB4856-high genes. This result may be a simple matter of lower power in these smaller datasets, as the RNAi effect is modest in all analyses.

All analyses were performed using *R* (*S29*).

Supporting References

S1. S. Brenner, The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71-94 (1974).

S2. E. J. Capra, S. M. Skrovanek, L. Kruglyak, Comparative developmental expression profiling of two *C. elegans* isolates. *PLoS ONE* **3**, e4055 (2008).

S3. M. V. Rockman, L. Kruglyak, Recombinational landscape and population genomics of *Caenorhabditis elegans*. *PLoS Genet* **5**, e1000419 (2009).

S4. R. B. Brem, G. Yvert, R. Clinton, L. Kruglyak, Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752-755 (2002).

S5. E. N. Smith, L. Kruglyak, Gene-environment interaction in yeast gene expression. *PLoS Biol* **6**, e83 (2008).

S6. M. Perez-Enciso, In silico study of transcriptome genetic variation in outbred populations. *Genetics* **166**, 547-554 (2004).

S7. G. F. Berriz, J. E. Beaver, C. Cenik, M. Tasan, F. P. Roth, Next generation software for functional trend analysis. *Bioinformatics*, (2009).

S8. M. Tijsterman, K. L. Okihara, K. Thijssen, R. H. Plasterk, *PPW-1*, a PAZ/PIWI protein required for efficient germline RNAi, is defective in a natural isolate of *C. elegans*. *Curr Biol* **12**, 1535-1540 (2002).

S9. X. Zhang, J. O. Borevitz, Global analysis of allele-specific expression in *Arabidopsis thaliana*. *Genetics* **182**, 943-954 (2009).

S10. K. W. Broman, H. Wu, S. Sen, G. A. Churchill, R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889-890 (2003).

S11. M. V. Rockman, L. Kruglyak, Genetics of global gene expression. *Nat Rev Genet* **7**, 862-872 (2006).

S12. P. M. Visscher, Whole genome approaches to quantitative genetics. *Genetica* **136**, 351-358 (2009).

S13. B. J. Hayes, P. M. Visscher, M. E. Goddard, Increased accuracy of artificial selection by using the realized relationship matrix. *Genet Res* **91**, 47-60 (2009).

S14. D. E. Goldgar, Multipoint analysis of human quantitative genetic variation. *Am J Hum Genet* **47**, 957-967 (1990).

S15. N. J. Schork, Genome partitioning and whole-genome analysis. *Adv Genet* **42**, 299-322 (2001).

S16. P. M. Visscher *et al.*, Genome partitioning of genetic variation for height from 11,214 sibling pairs. *Am J Hum Genet* **81**, 1104-1110 (2007).

S17. H. M. Kang *et al.*, Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709-1723 (2008).

S18. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-9445 (2003).

S19. C. D. Campbell, A. Kirby, J. Nemesh, M. J. Daly, J. N. Hirschhorn, A survey of allelic imbalance in F1 mice. *Genome Res* **18**, 555-563 (2008).

S20. A. Siepel *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050 (2005).

S21. B. Charlesworth, M. T. Morgan, D. Charlesworth, The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289-1303 (1993).

S22. R. R. Hudson, N. L. Kaplan, Deleterious background selection with recombination. *Genetics* **141**, 1605-1617 (1995).

S23. B. Charlesworth, M. Nordborg, D. Charlesworth, The effects of local seelction, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res*. **70,** 155-174 (1997).
S24. D. R. Denver *et al.*, A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proc Natl Acad Sci U S A* **106**, 16310-16314 (2009).
S25. A. Barriere, M. A. Felix, Temporal dynamics and linkage disequilibrium in natural *Caenorhabditis elegans* populations. *Genetics* **176**, 999-1011 (2007).
S26. A. D. Cutter, A. Dey, R. L. Murray, Evolution of the *Caenorhabditis elegans* genome. *Mol Biol Evol* **26**, 1199-1234 (2009).
S27. L. Loewe, B. Charlesworth, Background selection in single genes may explain patterns of codon bias. *Genetics* **175,** 1381-1393 (2007).
S28. H. S. Seidel, M. V. Rockman, L. Kruglyak, Widespread genetic incompatibility in *C. elegans* maintained by balancing selection. *Science* **319**, 589-594 (2008).
S29. R Development Core Team. (R Foundation for Statistical Computing, Vienna, Austria, 2008).

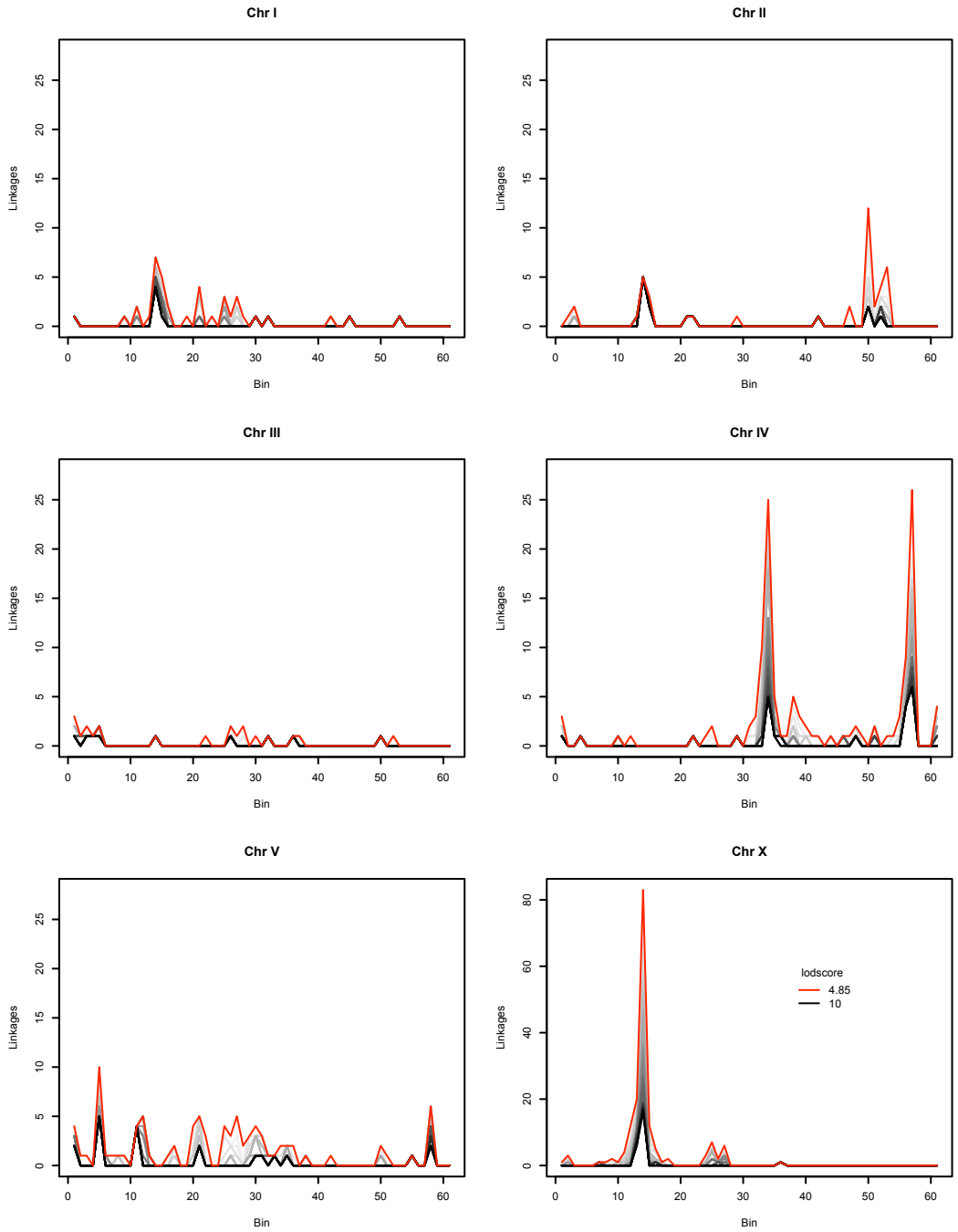
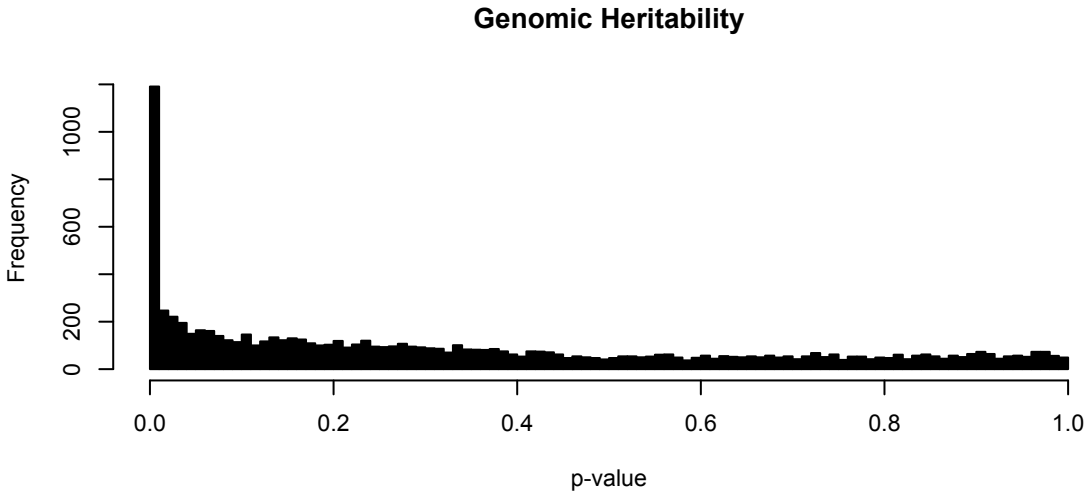


Figure S1. Hotspots of distant QTLs.
The number of distant linkages in each 5 cM bin is plotted across a range of significance thresholds from lod > 4.85 (FDR = 5%) to lod > 10 (FDR = 0). The distant linkages are based on interval mapping of the residuals of phenotypes on local genotypes. The genetic distances are based on the RIAIL cross and are expanded relative to meiotic genetic distances. Three locations exhibit an excess of linkages robust to lod-score thresholds.

A



B

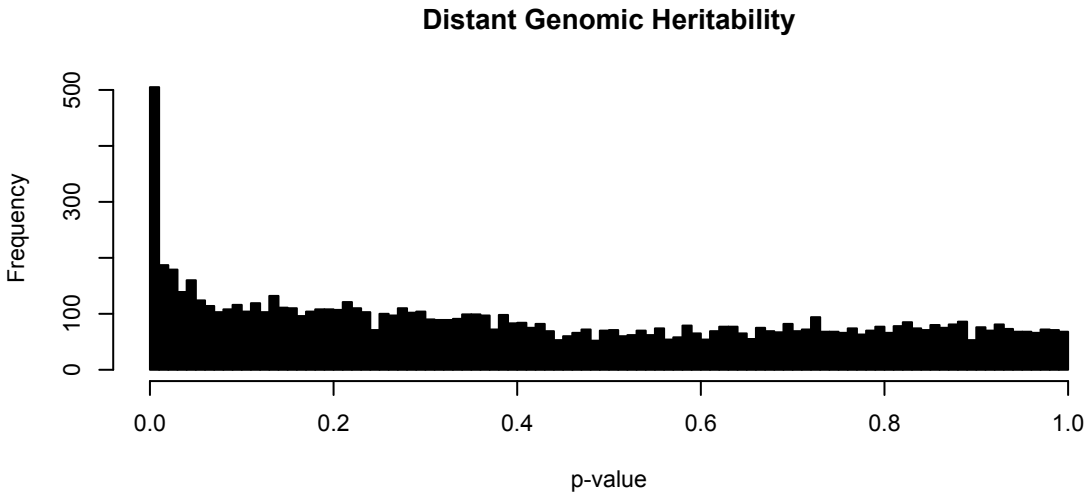


Figure S2. Significance of genomic heritability estimates.
Distributions of permutation-based empirical *p*-values for heritability $(V_G/(V_G + V_E))$ of transcript abundances estimated from genome-wide genotypic similarity. The top panel shows the distribution based on total phenotypic variation. The bottom panel shows the distribution based on distant variation only, the residuals of a regression of total phenotypic variation on genotype probability at the transcript locus.

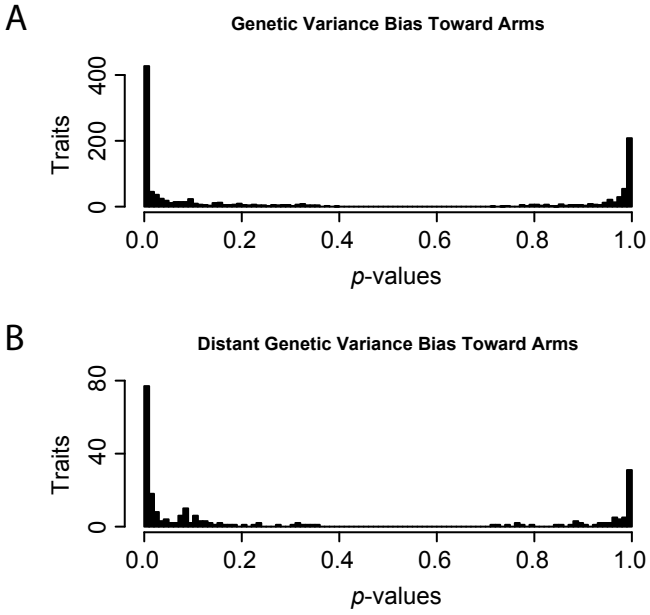


Figure S3. Arms contribute more than centers to the heritability of most traits.
A. For each heritable trait, we calculated the probability that the difference between the genetic variance contributed by arms and the genetic variance contributed by centers would be as large as that observed. Low p-values imply a larger than expected contribution by arms and high p-values imply a larger than expected contribution by centers. The departure from uniform p-values is heavily biased toward low p-values, representing traits whose heritable variation is largely attributable to variation in chromosome arms.
B. The result holds for total heritable variance and also for distant variance, which excludes contributions from variants tightly linked to the trait transcript.

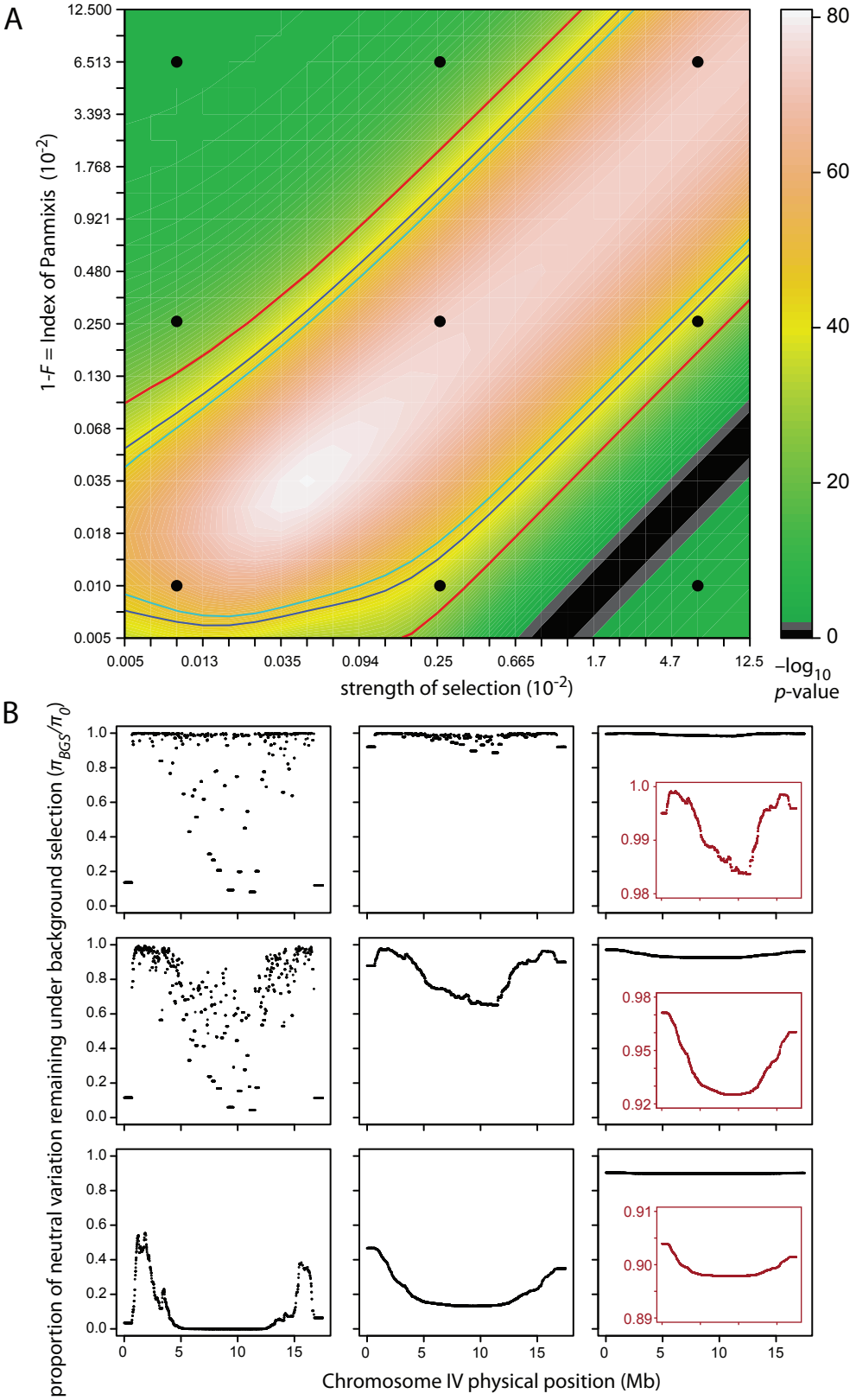


Figure S4. Background selection explains local linkage probability.
Caption continues on the following page.

Figure S4. Background selection explains local linkage probability.
A. The significance of background selection in a logistic regression model including gene-specific mutation and selection variables (Model 7) is plotted as a function of panmixis index and strength of selection against deleterious mutations; note the log scale. The best-fitting model has a panmixis index of 0.035% ($F = 0.99965$) and a strength of selection of 0.049%, but background selection is significant at $p < 0.01$ across all but a small slice of parameter space (black and grey). The red lines bracket the region of parameter space over which the loss of model-fit when background selection is dropped from the model exceeds that from dropping any other variable. The blue line surrounds the space in which a model featuring only background selection fits the data better than one including all gene-specific variables (Model 3), and the turquoise line surrounds the space in which background selection alone fits the data better than one including all gene-specific variables plus all interactions among them (Model 4).
B. The effects of background selection on Chromosome IV variation are shown for nine parameter combinations (corresponding black dots in A), illustrating the effects of variation in outcrossing rate and intensity of selection on expected levels of neutral variation along the chromosome. The strong-selection plots on the right are expanded along the y-axis in the brown insets.

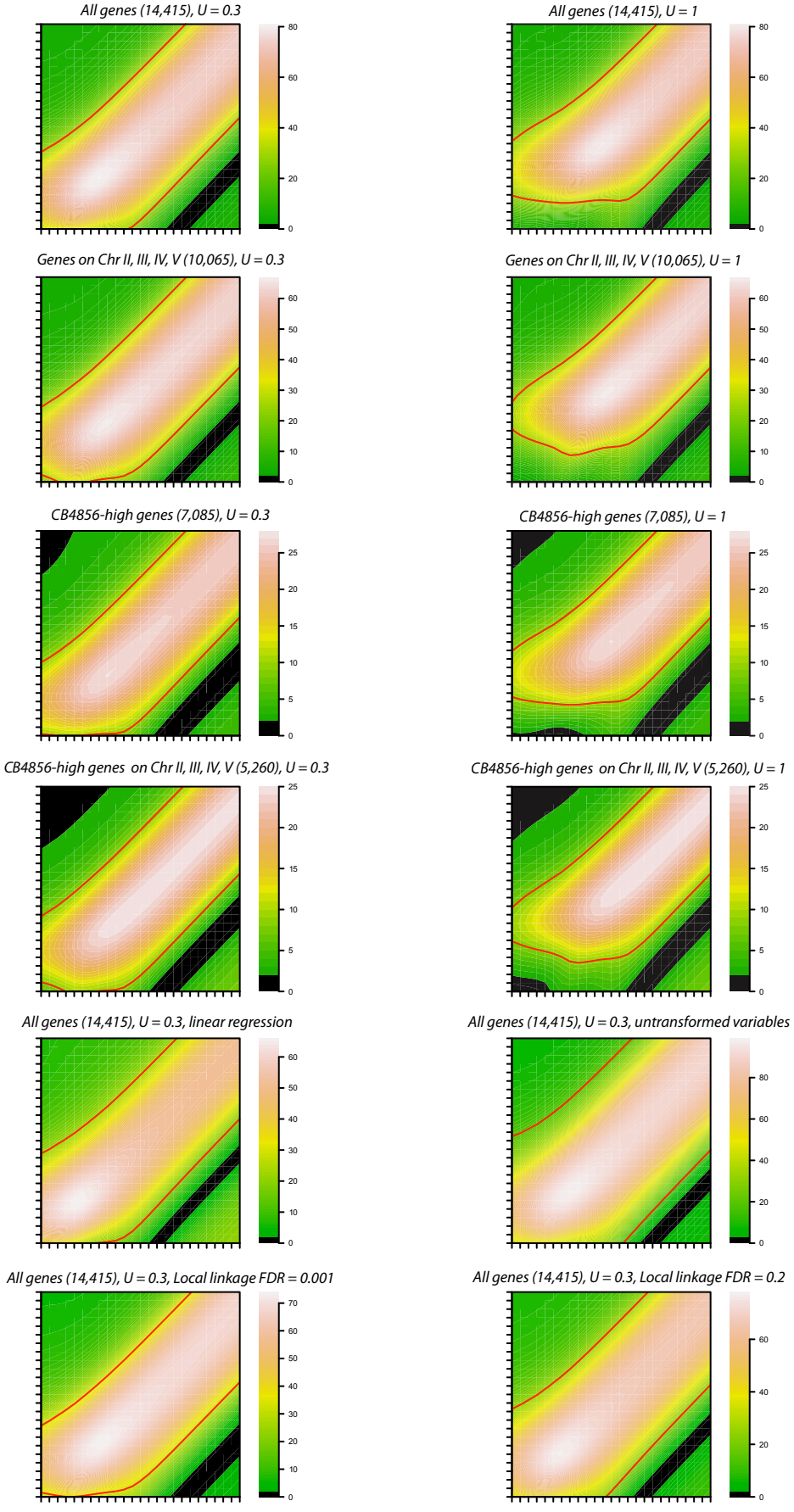


Figure S5. Background selection results are robust.
Caption continues on the following page.

Figure S5. Background selection results are robust.
The significance of background selection in several different models is plotted as a function of index of panmixis and strength of selection against deleterious mutations. In each case, the model includes gene-specific mutation and selection variables and the *p*-value tests the effect of dropping the background selection term from the model. Background selection is significant at $p < 0.01$ across all but a small slice of very low-outcrossing parameter space (black) in every case; note that these plots are log-scaled (axes as in Fig S4), with the bulk of parameter space in the upper right corner (see Fig. 2). The red lines bracket the region of parameter space over which background selection explains more of the local linkage probability than any other variable in the model. We calculated these *p*-values for two different values of the genomic mutation rate, *U*, and for four different datasets: all genes; only genes on chromosomes II-V, to eliminate possible map-distance complications for chromosomes I and X; only genes with higher expression in strains carrying the CB4856 allele of the gene's locus, to eliminate potential hybridization artifacts; and only these CB4856-high genes on chromosomes II-V. We also calculated *p*-values under a linear regression of the log of the local linkage lod scores; using raw (untransformed) values of two genic variables, gene interval length and number of conserved base pairs; and using a more stringent false discovery rate (0.001) and a less stringent false discovery rate (0.2).

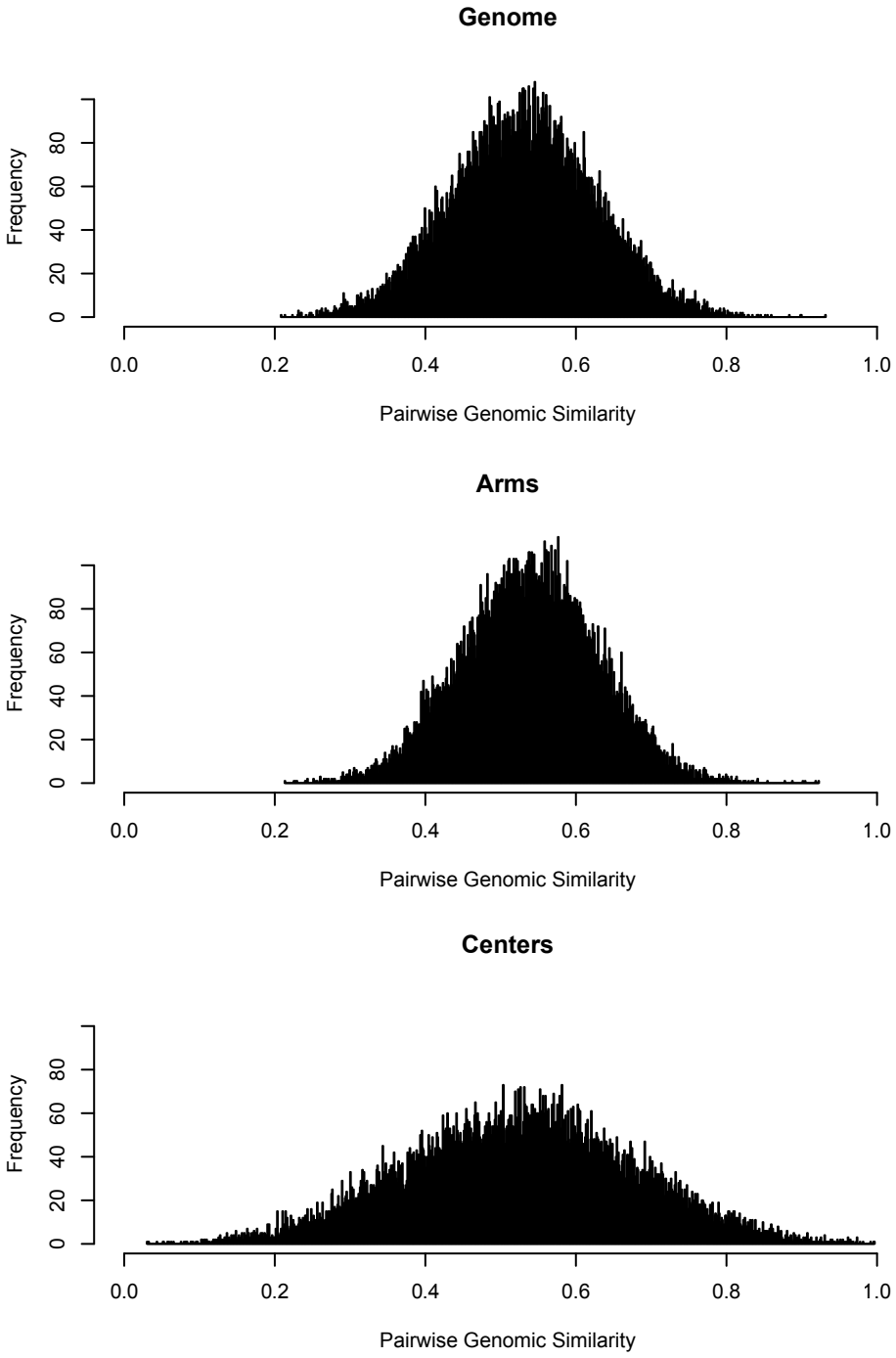


Figure S6. Genomic similarity for all pairs of RIAILs.
Genomic heritability is based on the relationship between pairwise genotypic and phenotypic similarity. These histograms show the distribution of genotypic similarities for all pairs of strains, using the entire genome (top), only the chromosome arms (middle), and only the chromosome centers (bottom). The means are greater than 0.5 because of the inclusion of RIAILs that share recent ancestors and because of segregation distortion on chromosome I. See reference S3 for details..

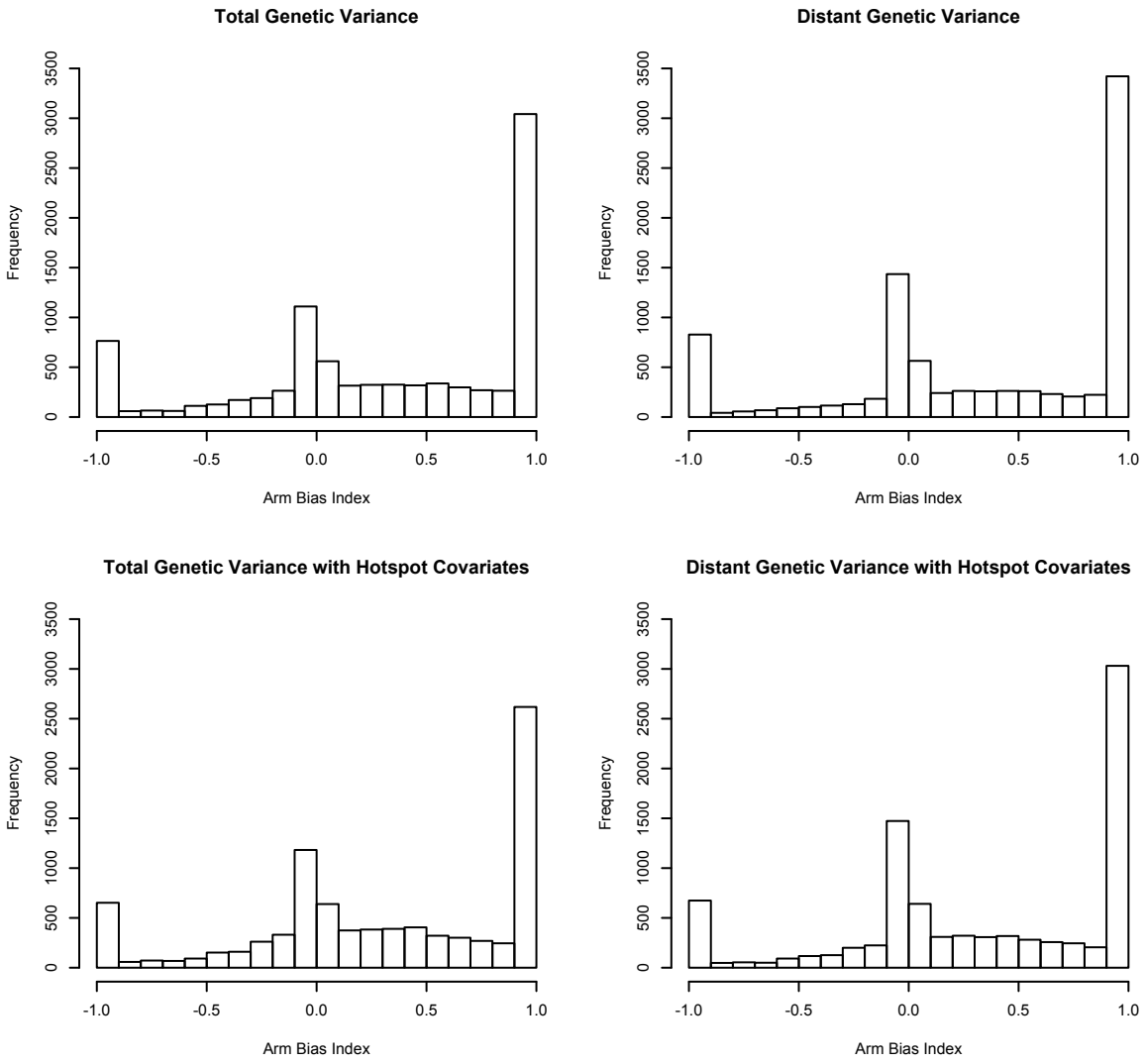


Figure S7. Distant linkage hotspots do not drive arm-biased patterns of genomic heritability. The arm bias index, $(V_{G,Arms} - V_{G,Centers})/(V_{G,Arms} + V_{G,Centers})$, calculated for each of 8,973 probes with no missing data, ranges from -1 to 1, with positive values indicating a greater contribution of chromosomal arms than centers to a trait's heritability. The overall pattern is not altered by controlling for the contribution of local QTLs or by controlling for the contribution of the three robust QTL hotspots.