

# Inferences About the Distribution of Dominance Drawn From Yeast Gene Knockout Data

Aneil F. Agrawal<sup>\*,1</sup> and Michael C. Whitlock<sup>†</sup>

<sup>\*</sup>Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario M5S 3B2, Canada and

<sup>†</sup>Department of Zoology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada

Manuscript received October 25, 2010

Accepted for publication November 10, 2010

## ABSTRACT

Data from several thousand knockout mutations in yeast (*Saccharomyces cerevisiae*) were used to estimate the distribution of dominance coefficients. We propose a new unbiased likelihood approach to measuring dominance coefficients. On average, deleterious mutations are partially recessive, with a mean dominance coefficient  $\sim 0.2$ . Alleles with large homozygous effects are more likely to be more recessive than are alleles of weaker effect. Our approach allows us to quantify, for the first time, the substantial variance and skew in the distribution of dominance coefficients. This heterogeneity is so great that many population genetic processes analyses based on the mean dominance coefficient alone will be in substantial error. These results are applied to the debate about various mechanisms for the evolution of dominance, and we conclude that they are most consistent with models that depend on indirect selection on homeostatic gene expression or on the ability to perform well under periods of high demand for a protein.

AS Mendel observed, the effects of an allele in a homozygous state may be difficult to predict from its effects in heterozygous individuals. Such genetic dominance is a common phenomenon, and dominance affects many important evolutionary processes in diploid organisms. Dominance may predict the fate of beneficial alleles, through a process called Haldane's sieve, because recessive beneficial mutations are less likely to fix when started from a low frequency than beneficial alleles (HALDANE 1927), but this difference may disappear depending on the starting conditions in the population (ORR and BETANCOURT 2001). More importantly, if deleterious alleles are maintained in populations mainly due to the balance between selection removing them and recurrent mutation to the deleterious state, then the frequency of such alleles should be inversely proportional to their dominance coefficient when paired with a fully functional allele (HALDANE 1937). While dominance does not affect the mean fitness in panmictic populations except for completely recessive alleles (MULLER 1950), dominance is a key determinant of mutation load in populations with some nonrandom mating, either because of inbreeding (MORTON 1971) or because of population structure (WHITLOCK 2002). Moreover, the amount of inbreeding depression associated with segregating deleterious alleles depends quite

strongly on the dominance coefficient (CHARLESWORTH and CHARLESWORTH 1987). In a medical genetic context, the degree of dominance of deleterious mutations affects the impact of each new mutation on human health.

In spite of the importance of dominance for fitness, we have remarkably little reliable information about the magnitude of such dominance. Most of our information about dominance comes from a small number of species (*Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, and yeast), and many of the estimates of dominance are mutually inconsistent (see Table 1). Moreover, we have very little information about the variation in dominance among genes, and what exists is biased in a number of ways (CABALLERO *et al.* 1997; GARCÍA-DORADO and CABALLERO 2000; FRY and NUZHIDIN 2003; FRY 2004).

In this article we adopt the usual population genetic conventions of describing dominance and fitness. When discussing deleterious alleles, the selection coefficient  $s$  is the proportional decline in fitness of an individual homozygous for a mutant allele, relative to individuals that are homozygous for the wild-type allele. Thus, the homozygote mutant individual will on average have fitness that is  $1 - s$  as great as that of an average wild-type homozygote. We use  $h$  to indicate the dominance coefficient of the allele, the proportion of fitness effect expressed in the homozygote. Thus, a heterozygote would have fitness on average equal to  $1 - hs$  times the fitness of the wild-type homozygote. If  $h = 0$ , the heterozygote has fitness equal to that of the wild-type heterozygote, while if  $h = 1$ , then the heterozygote has

Supporting information is available online at <http://www.genetics.org/cgi/content/full/genetics.110.124560/DC1>.

<sup>1</sup>Corresponding author: Department of Ecology and Evolutionary Biology, University of Toronto, 25 Willcocks St., Toronto, Ontario M5S 3B2, Canada. E-mail: a.agrawal@utoronto.ca

TABLE 1

Previous dominance estimates for new mutations affecting fitness components

Species	Method of mutation <sup>a</sup>	$\bar{h}$	Source
<i>C. elegans</i>	EMS	0.1	PETERS <i>et al.</i> (2003)
<i>C. elegans</i>	MA	0.55	VASSILIEVA <i>et al.</i> (2000)
<i>D. melanogaster</i>	MA	-0.22 <sup>b</sup>	MUKAI <i>et al.</i> (1964)
<i>D. melanogaster</i>	MA	0.09 <sup>b</sup>	MUKAI <i>et al.</i> (1965)
		0.13 <sup>b</sup>	
<i>D. melanogaster</i>	MA	0.34 <sup>b</sup>	MUKAI and YAMAZAKI (1968)
		0.35 <sup>b</sup>	
		-0.32 <sup>b</sup>	
<i>D. melanogaster</i>	MA	-0.09 <sup>b</sup>	MUKAI (1969)
		-0.15 <sup>b</sup>	
		-0.09 <sup>b</sup>	
<i>D. melanogaster</i>	MA	0.49 <sup>b</sup>	OHNISHI (1977a,b)
		0.40 <sup>b</sup>	
<i>D. melanogaster</i>	P elements	0.01	MACKAY <i>et al.</i> (1992)
<i>D. melanogaster</i>	MA	0.12	HOULE <i>et al.</i> (1997)
<i>D. melanogaster</i>	MA	0.3	CHAVARRIAS <i>et al.</i> (2001)
<i>D. melanogaster</i>	MA	0.17	FRY and NUZHIDIN (2003)
<i>D. melanogaster</i>	Transposable elements	0.5	FRY and NUZHIDIN (2003)
<i>S. cerevisiae</i>	EMS	0.2	SZAFRANIEC <i>et al.</i> (2003)

<sup>a</sup> MA, mutation accumulation; EMS, ethyl methanesulfonate mutagenesis.

<sup>b</sup> Serious concerns have been raised about these estimates; see GARCIA-DORADO and CABALLERO (2000) and FRY (2004).

fitness equal to that of the mutant homozygote. In practice  $h$  can be negative (indicating overdominance if the mutant allele is deleterious), between 0 and 1 as discussed above or even  $>1$  with underdominance.

To model evolution with new mutations, it is essential to know not only what the average dominance coefficient of new mutations might be, but also the variation among loci in this average and whether the dominance coefficient  $h$  is correlated in any way with the strength of selection against deleterious alleles  $s$ . If alleles of large effect are recessive but alleles of small effect are additive, then inbreeding depression will largely be caused by alleles of large effect. In this case, purging through repeated inbreeding may strongly affect the inbreeding depression of a population. If, on the other hand, there are a large number of deleterious recessive alleles of small effect, purging will be less effective, and inbreeding depression cannot evolve as easily (LANDE and SCHEMSKE 1985; CHARLESWORTH and CHARLESWORTH 1987; CHARLESWORTH *et al.* 1990; HEDRICK 1994). More generally, many simulation models of evolution assume a relationship between dominance and the strength of selection or of the variation in the dominance coefficient

(*e.g.*, CABALLERO and KEIGHTLEY 1994); the accuracy of these models may depend on the accuracy of their assumptions about this relationship.

Variation in  $h$  is interesting for another reason, for it points to the potential variety of mechanisms for, and the evolution of, dominance. A deleterious allele may be recessive because a single copy of a functioning allele may be sufficient to meet the needs of the organism, or it may be partially recessive or dominant if two functional alleles are required for full function. In some cases, a mutated allele can directly cause problems for the organism, over and above a loss of function. We discuss these mechanisms in greater depth later in the DISCUSSION. Many researchers have hypothesized that some types of proteins—such as structural proteins—may be more likely to have dominant deleterious mutations, while others, for example enzymes, are more likely to have recessive deleterious mutations (BOURGUET *et al.* 1997; VEITIA 2005). Some empirical work has supported such variation by functional type (KONDRASHOV and KOONIN 2004; PHADNIS and FRY 2005).

The distribution of dominance coefficients—and its possible relationship with correlates such as functional type and selection coefficient—may inform the search for the mechanism for evolution of dominance, a long-standing question in evolutionary biology. Some models of dominance suggest that it is purely a by-product of protein function and not subject to effective selection. This is essentially the perspective of Sewall Wright who argued for a physiological basis of dominance. In contrast, some models suggest that dominance often evolves due to selection on heterozygous individuals; this was championed by R. A. Fisher in his famous conflict with Wright over dominance that drove the two seminal population geneticists apart (PROVINE 1986). Many other models of the evolution of dominance have appeared in the years since, and these models sometimes make different predictions about the relationship between  $s$  and  $h$  or between  $h$  and functional types. For example, CHARLESWORTH (1979) pointed out that a pattern of negative correlation between  $s$  and  $h$  is not predicted under Fisher's hypothesis. However, the widespread view that there is an inverse relationship between dominance coefficient and the selection coefficient is based on largely two studies, one a review article by SIMMONS and CROW (1977) and the other an analysis of yeast deletion data by PHADNIS and FRY (2005). In the DISCUSSION, we return to various models of the evolution of dominance, to compare them in the light of the results of our investigation of dominance coefficients for growth rate in yeast deletion mutations.

SIMMONS and CROW (1977) summarized the *Drosophila* studies of the 1960s and 1970s and concluded that the dominance coefficient of mildly deleterious mutations averaged something close to  $\sim 0.3$ , while the mean  $h$  for alleles of large homozygous effect was closer to 0.02. However, experimental protocols, including most types

of mutation accumulation experiments and mutagenesis screens, are typically biased toward excluding dominant or semidominant mutations of large effect, leading to an underestimate of  $h$  for large-effect mutations. Moreover, while Simmons and Crow were careful to distinguish between alleles sampled from standing variation and new mutations in their analysis, some of the subsequent quotation of their article obscures the difference. (Alleles of large effect that have been exposed to selection will on average have much lower dominance coefficients than new mutations, because recessive mutations are more protected from the purifying effects of selection. Hence, standing variation is an unreliable guide to the nature of new mutations.) Some studies of new mutations suggested much higher dominance coefficients for new strong mutations (say,  $s > 0.1$ ), on the order of 0.07 (SIMMONS and CROW 1977). On the other hand, the average  $h$  over all the studies of new mildly deleterious mutations reported in Simmons and Crow (in their Table 5) was only 0.08. Looking at even smaller values of  $s$  (say,  $0 < s < 0.03$ ), there is almost no information on weak-effect mutations in the Simmons and Crow summary.

Recent work by PHADNIS and FRY (2005) looked at the effects of heterozygous and homozygous expression of a very large collection of deletions of open reading frames in yeast. From this analysis, they concluded that the dominance coefficient on average decreases with increasing selection coefficient. Their results indicated that this relationship is qualitatively similar for all functional types of proteins, although mutations of structural proteins appeared to have higher dominance for a given  $s$  than other functional types.

In this article, we revisit the issue of the joint distribution of selection coefficient and dominance coefficient. We use a large data set on the effect on growth rate of yeast protein-coding region deletions (STEINMETZ *et al.* 2002). This is the same data set used by PHADNIS and FRY (2005), but we perform a more comprehensive analysis that allows us to study the distribution of dominance coefficients and its relationship to selection in considerably more detail. Using more recent information, we are able to better identify wild-type genotypes in the original data. This allows us to distinguish between beneficial and deleterious mutations and estimate the dominance of each type separately. Moreover, we are able to quantify the variance and skew in  $h$  as well as test different types of relationships between  $h$  and  $s$ . In a separate (regression) analysis, we find that the degree of dominance varies between protein functional types. We end by discussing the relative credibility of various models of evolution of dominance, in light of the conclusions from this analysis.

#### YEAST KNOCKOUT DATA

We analyze the fitness of strains of the yeast *Saccharomyces cerevisiae* with gene deletions, expressed in either hetero-

zygous or homozygous states. In this section, we describe some of the general approaches for analysis of these data. The *Saccharomyces* Genome Deletion Project, using the methods developed by SHOEMAKER *et al.* (1996), created strains of yeast that collectively contained deletions of most open reading frames in the yeast genome (see, *e.g.*, GIAEVER *et al.* 2002). Each deletion is marked by a unique pair of barcodes, which were inserted into the genome at the location of the deletion as part of the deletion protocol. These bar codes allow quantification of the relative numbers of individuals present in a population according to genotype, via microarray analysis.

STEINMETZ *et al.* (2002) measured the growth rates of 4706 of the homozygous strains and 5791 of the heterozygous strains. Most of these strains were measured in two replicates for each of the homozygous and heterozygous strains. Growth rate was calculated by a regression through five time points, averaging over the four reads from sense and antisense tags of both upstream and downstream barcodes. (In a few cases the downstream barcodes were not assessed.) The results are available at [http://www-deletion.stanford.edu/YDPM/YDPM\\_index.html](http://www-deletion.stanford.edu/YDPM/YDPM_index.html).

One limitation of the Steinmetz *et al.* data is that they do not include any wild-type genotypes. Ideally, one would like to have the ancestral strain that had bar codes added to nonfunctional DNA, to measure the fitness effects of a wild-type strain including whatever effect the bar code insertion may have had. Fortunately, it turns out that a large number of bar codes and deletions were made in the *Saccharomyces* Genome Deletion Project at sites that turn out to have no evidence of having open reading frames or any other known functional elements (HILLENMEYER *et al.* 2008). To find a list of such genes that may act as suitable wild-type markers, we searched the *Saccharomyces* Genome Database (SGD PROJECT 2009) for all open reading frames (ORFs) that are annotated as “dubious,” meaning that the annotated ORF has insufficient evidence of including a true ORF. Of the dubious ORFs, we examined the phenotype information in SGD and discarded all ORFs that suggested partial overlap with another ORF or transcription region or that showed any known phenotypic effect. This left a list of 182 dubious ORFs. We then examined data about homozygous knockouts of these ORFs in the multiple-environment growth assays of HILLENMEYER *et al.* (2008), and we selected the 120 lines that showed the least variance in growth rate across environments. These 120 lines are used below as wild-type strains. In contrast, PHADNIS and FRY (2005) used the knockout strains with highest fitness as putative wild types; we show later in this article that these high-fitness strains are likely to actually carry beneficial mutations and therefore overestimate wild-type fitness.

The recorded growth rate values for each line represent the slope of the regression of log population

size on time, based on five time points, with one added to each for convenience. We converted these values to a measure of relative fitness by raising  $e$  to the power of the recorded value. These values are reported in growth rates per hour; to calculate growth rates per cell division all growth rate values reported should be divided by  $\sim e^t$ , where  $t$  is the average length of a cell generation in hours for the wild type. This value is not given by STEINMETZ *et al.* (2002) for their conditions, and so we leave the growth rates expressed per hour rather than calculate per-generation value.

These data are unusually detailed on an important topic, but there are some limitations to the data that need to be addressed. For example, the yeast knockout data are extraordinarily broad in their coverage of the genome, but this breadth comes at the expense of replication per genotype. The growth rate of each genotype is measured only twice typically, and the measurements are error-prone. As a result, the data for each gene are not reliable. The data become useful when combined across the whole genome, and the likelihood function that we used accounts for some of the measurement error collectively.

In addition, many loci are missing from the assay. In some cases, this is because both the heterozygote and the homozygote have a low growth rate and could not be sustained in the collection. For many of these cases, the dominance coefficient for growth rate would be high, because both the heterozygote and the homozygote have no growth in YPD. Thus the data set is unavoidably biased against alleles with large heterozygous effect and therefore of high dominance coefficient. Such mutations are unlikely to greatly affect evolution, however, as they would be immediately removed by selection within a generation.

Furthermore, loci with large homozygous effects are excluded from the data set, meaning that we have little information from these data about mutations of large homozygous effect. Our estimates of the overall average dominance coefficient are therefore potentially biased, if these loci have dominance coefficients different from that of the typical locus, which seems likely given the results. In general, we have little certainty about the dominance of genes with large  $s$  (say,  $s > 0.3$ ) from this data set.

Some of the gene deletion lines in STEINMETZ *et al.* (2002) may contain small aneuploid regions with duplications or deletions of genes (HUGHES *et al.* 2000). Such aneuploid duplications, if they include the genes in question, may buffer or compensate for the effect of loss of functional alleles. Moreover, it is possible that heterozygous and homozygous strains may differ in their genotypes at other loci due to insertions and deletions (or indeed other mutations) during the strain creation process. Such new mutations will create error in the estimates of the fitness effects of the target loci.

The strains are grown at sufficiently large population sizes that it is possible that new beneficial mutations may arise during the strain creation or fitness measures. If

these mutations arise after the creation of separate heterozygous and homozygous lines but before assessment in the STEINMETZ *et al.* (2002) study, then they can confound the estimation of the relative fitness of the two genotypes. SLIWA and KORONA (2005) show that many of the apparently more fit deletion strains have high fitness due to beneficial mutations at other loci. If, however, a mutation appears during the fitness measurement process it is unlikely to affect our conclusions, because such mutations are unlikely to arise in both replicates and we trim the data of any genes that show strong differences in growth rate between replicates (as described below).

The loci included in the STEINMETZ *et al.* (2002) growth rate screen show large variation in dominance coefficients, yet there is reason to believe that this is an underestimate of the true variation in  $h$ . WILKIE (1994) reviewed reasons for dominant deleterious mutations. These include haploinsufficiency (where two functional copies are necessary for normal function), dominant negatives (such as resulting from the disruption of a dimer or multimer or from competition for substrate), increased instability in structural elements (such as cell walls), toxic proteins, and ectopic expression. Interestingly, all but the first of these depend on the fact that the deleterious allele actively causes a problem, and they would not be expected necessarily to show large heterozygous effects with a knockout allele. As such, many types of dominant mutations would not be present in the knockout data set, where all of the mutations are deletions and do not produce organisms that express a deleterious allele. As a result, the dominance coefficients predicted by analysis like ours, based on knockout lines alone, are probably biased downward relative to  $h$  for all types of mutations.

These data contain many point estimates for which there is almost no homozygous effect with a much larger heterozygous effect. In these cases, the heterozygous effect is much larger than would be predicted by the measurement error predicted by a normal distribution on the basis of the variance between replicates. These estimates may be valid, but they suggest that there may be an alternative source of large errors not well described by a typical error function using a normal distribution. Such errors seemed plausible to us, which is the reason that we used a trimmed mean approach in our method (described below). These unexpected results for a small fraction of the genes in the study suggest that there may be some types of experimental artifacts that have not been taken into account in this study, although we use a trimming process that should have removed most of these artifacts from the data analyzed here.

For each replicate, the data from all genes were collected simultaneously. Consequently, there are several reasons to expect that data from different genes may be correlated in these data. Genes that are nearby



(or otherwise share qualities such as edge effects) on the expression analysis chips may have positively correlated measurement error. Thus, some of the aspects of the distributions we infer may result from correlated errors across multiple loci. With only two replicates per gene (and with a similar chip design from both replicates) it is difficult or impossible to assess the extent of this correlated error.

Finally, the data were collected in an artificial laboratory environment that was not the same as the conditions under which this species has evolved. As a result, the fitness effects of a gene deletion in either heterozygous or homozygous state may be quite different from those in the natural state.

## THE DISTRIBUTION OF DOMINANCE

**Overview of the likelihood model:** To estimate the distribution of dominance and the relationship between dominance and the strength of selection, we compared a series of likelihood models using the yeast growth rate data. Our full likelihood model involves 18 parameters required to describe (i) the distribution of homozygous selection coefficients  $s$ , (ii) the distribution of dominance coefficients  $h$  (and its relationship with  $s$ ), and (iii) measurement error. Each of these three components is summarized below and model parameters are defined in Table 2. In the main text we sketch the assumptions used in the likelihood model; greater details are provided in supporting information, [File S1](#).

**The distribution of  $s$ :** After considerable preliminary examination of the homozygous knockout data, we settled on the compound distribution described below for modeling the distribution of selection coefficients. Using only the data from the homozygous knockout lines, we compared the Akaike's information criterion (AIC) scores of numerous models of the distribution of homozygous deleterious effects, including combinations of the exponential, uniform, and gamma distributions. The distribution of deleterious effects that we use, which is a linear combination of a uniform distribution and two exponential distributions, has enough degrees of freedom to fit the data well, and yet it still can be integrated to give a likelihood model that can be calculated in a reasonable amount of computer time.

We assume that a fraction  $p_{\text{ben}}$  of mutations are beneficial. The effects of these beneficial mutations are assumed to follow an exponential distribution with mean  $\mu_{\text{ben}}$ , because this provided a reasonable match to a visual inspection of the distribution and such a distribution is expected by theory (ORR 2003). The remainder of the mutations are assumed to be deleterious. Because visual inspection of the data indicates that the distribution of deleterious effects is not well described by any one simple distribution, we use a compound distribution consisting of three parts. Of the deleterious mutations, a fraction

$p_{\text{uniform}}$  have deleterious effects with magnitudes drawn from a uniform distribution from 0 to  $u_{\text{max}}$ . The remaining deleterious mutations have effects with magnitudes chosen from either of two exponential distributions with means  $\mu_{\text{exp1}}$  and  $\mu_{\text{exp2}}$  with a fraction  $p_{\text{exp1}}$  from the former and the remainder from the latter. Figure 1 shows the match of the resulting distribution to the data for the homozygous fitnesses. Moreover, simulated data generated using this model reproduce various other features of the real data (not shown).

**The distribution of  $h$ :** For deleterious mutations, we sought to identify the mean value of the dominance coefficient as a function of the selection coefficient,  $\mu_{h(\text{del})}(s)$ . We varied the possible relationship between  $\mu_{h(\text{del})}(s)$  and  $s$ , comparing models with no relationship with a linear relationship ( $\mu_{h(\text{del})}(s) = \beta_1 + \beta_2 s$ ) or with an asymptotic relationship between  $h$  and  $s$  ( $\mu_{h(\text{del})}(s) = \beta_1 / (1 + \beta_2 s)$ ). We approximated both of these continuous relationships by step functions of  $s$ , to obtain tractable likelihood functions. Seventeen intervals of  $s$  values were used for these steps, with finer steps in the range of small  $s$  because this is where most of the data lie. The boundaries of the intervals are given by {0, 0.02, 0.04, 0.06, 0.08, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}. The actual relationship between  $\mu_{h(\text{del})}$  and  $s$  was based on the functions above but, instead of using  $s$ , the middle value of the interval containing  $s$  was used (see [File S1](#) for details).

In addition, we considered the variation in  $h$  around its mean value within an interval, examining models with no variation, with normally distributed variation, or with asymmetric variation around the mean represented by a displaced gamma distribution. The displaced gamma distribution has three parameters: the shape parameter, the scale parameter, and a displacement value. As explained in [File S1](#), the displacement parameter allows the shape of the distribution (*e.g.*, its variance and skew) to be determined independently of the mean. The distribution of  $h$  around its mean was also calculated as a step function, over 25 small intervals of possible  $h$  values, determined by the corresponding quantiles of the basic distribution. More details are given in [File S1](#).

Beneficial mutations are rare enough and of small enough effect (as it turns out) that we did not feel that there was enough power to distinguish a relationship between  $s$  and  $h$  for beneficial mutations; we therefore fit the data to a constant mean  $h$  value.

**Measurement error:** There is considerable measurement error in the individual estimates of fitness. Moreover, preliminary inspection of the data revealed some peculiarities to this error that we have attempted to model. First, there appears to be more measurement error in measures of heterozygous fitness than in those of homozygous fitness, evidenced by the variance between the two replicate measurements of each genotype. In addition, measurement error seems to be greater for

**TABLE 2**  
**Parameters in the likelihood model**

Parameter	Description
$W_{\text{wt,Hom}}$	Fitness of the wild-type genotype in the homozygote pool
$W_{\text{wt,Het}}$	Fitness of the wild-type genotype in the heterozygote pool
$p_{\text{ben}}$	Proportion of mutations that have beneficial homozygous effects
$p_{\text{uniform}}$	Proportion of mutations that come from the uniform component of the compound distribution for deleterious mutations
$p_{\text{exp1}}$	Of those deleterious mutations that do not come from the uniform component of the compound distribution, a fraction $p_{\text{exp1}}$ come from the first (of two) exponential distributions
$\mu_{\text{ben}}$	Mean of the exponential distribution describing the distribution of homozygous fitness effects of beneficial mutations
$u_{\text{max}}$	Upper limit for the uniform distribution that is a component of the compound distribution of the magnitude of deleterious mutations
$\mu_{\text{exp}X}$	Mean of the $X$ th exponential distribution that is a component of the compound distribution of the magnitude of deleterious mutations ( $X \in \{1, 2\}$ )
$\mu_{h(\text{ben})}$	Mean dominance of beneficial mutations
$\beta_1$	Intercept (in linear model) or numerator (in asymptotic model) relating $\mu_{h(\text{del})}(s)$ to $s$
$\beta_2$	Slope (in linear model) or denominator term (in asymptotic model) relating $\mu_{h(\text{del})}(s)$ to $s$
$\sigma_h^2$	Variance of the distribution of $h$ around its mean within an $s$ interval
$\delta$	Mean of the gamma distribution used to determine the shape of the variation around the mean dominance, $\mu_h$
$V_{\text{me,Hom}}$	Baseline level of variance due to measurement error for homozygotes
$V_{\text{me,Het}}$	Baseline level of variance due to measurement error for heterozygotes
$v_{\text{Hom}}$	Coefficient measuring the extent to which error variance increases for deleterious effects in homozygotes: $V_2(s) = (1 + v_{\text{Hom}}s) V_{\text{me,Hom}}$
$v_{\text{Het}}$	Coefficient measuring the extent to which error variance increases for deleterious effects in heterozygotes: $V_1(s) = (1 + v_{\text{Het}}hs) V_{\text{me,Het}}$

low-fitness genotypes than for high-fitness ones. To account for these observations, we have modeled measurement error as follows. We assumed that measurement error is normally distributed with mean zero and variance  $V$ . The baseline measurement error variance is  $V_{\text{me,Hom}}$  and  $V_{\text{me,Het}}$  for homozygotes and heterozygotes, respectively. This is the error variance applied to wild types and genotypes with fitness greater than wild type. For deleterious heterozygotes, the measurement error is  $V_1(hs) = (1 + v_{\text{Het}} |hs|) \times V_{\text{me,Hom}}$  and, for deleterious homozygotes, the measurement error is  $V_2(s) = (1 + v_{\text{Hom}} |s|) \times V_{\text{me,Hom}}$ , to account for the possible increase in error for large-effect mutations.

**Trimming the data:** The strength of the data set is in the large number of genes measured rather than the quality of individual measures. For our analysis, we include only genes for which complete data are available, *i.e.*, two measures of fitness of the mutation in both homozygous and heterozygous states. Still, some estimates are too deviant between replicates to come from the expected normal distribution of measurement error that characterizes most of the data set, evidenced by a very poor fit to normality of the differences between replicate measurements. These may represent various experimental artifacts. We attempted to eliminate these suspect estimates by trimming the data in two ways. First, we compared the difference between replicates. We eliminated genes from our data set if the difference between replicates, for either homozygotes or hetero-

zygotes, was too extreme on the basis of the error variance. More precisely, we first estimated the error variance by the among-replicate variance obtained from an ANOVA of the full data set (calculating the variance components of replicate nested within gene, separately for heterozygotes and homozygotes). Then, assuming that measurement error was normally distributed with a mean of zero and using the estimated error variance, we calculated what discrepancy between replicates would be observed with the probability  $\leq 1/2n$ , where  $n$  is the number of genes. Loosely, this value can be thought of as the maximum difference we would likely see if we had twice as many genes (a factor of 2 was included to be conservative) under a normal measurement error structure. Thirty-five (of 4102) genes had bigger differences than this maximum and so were eliminated from further analysis.

Even after this initial trim, the data set contained a substantial fraction of genes for which the point estimate indicates very small effects on homozygotes, but strongly negative heterozygous effects. Such apparently strong underdominance may be real, but a plausible alternative is that some fraction of the genotypes are measured in error with a distribution of error not well described by the Gaussian. Such errors seemed likely, so we further trimmed the data set. First, we sorted genes into deciles with respect to homozygous fitness estimates. We then eliminated genes in the top 2.5% and bottom 2.5% with respect to heterozygous fitness estimates *within* each

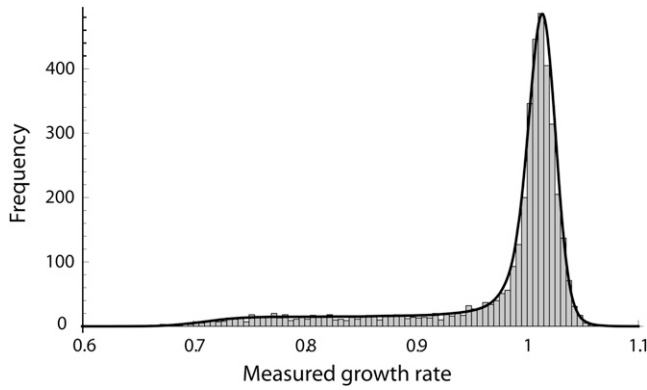


FIGURE 1.—The distribution of observed growth rates observed from the homozygote yeast knockout data. The solid thick line indicates the distribution of observed growth rates predicted by the model described in the text with the maximum-likelihood values of its parameters.

decile, leaving us with a total of  $T = 3867$  mutants. This procedure should eliminate the worst experimental artifacts that could bias our inferences of  $h$  without spuriously creating a correlation between  $h$  and  $s$ . We believe that these trimmed data provide more meaningful insights into  $h$ , so we focus our analysis on this trimmed data set. To the extent that this trimming process may have eliminated valid data points, we have underestimated the variance, skew, and mean (due to the skew) of the distribution of dominance coefficients.

**Maximum-likelihood results:** We focus our discussion primarily upon our inferences regarding the dominance coefficients of deleterious mutations because the different models we evaluated were constructed to examine various aspects of this distribution. Other aspects of joint  $h$  and  $s$  distribution will be considered after identifying the best model. Complete maximum-likelihood results for the trimmed data set are presented in Table 3. In our most constrained model (model 1), there is one value of  $h$  for all deleterious mutations. In this model, we estimate  $h = 0.039$ . However, the AIC score improves dramatically if we allow for variation in  $h$ , even without postulating a relationship between  $h$  and  $s$  (*i.e.*,  $\beta_2 = 0$ ), as shown in our models 2 and 3. In model 2, the distribution of  $h$  values is based on a normal distribution whereas, in model 3, the distribution of  $h$  values is based on a displaced gamma distribution. Both of these models have much better AIC scores than model 1, indicating substantial variation in  $h$ . Moreover, the displaced gamma distribution fits the data much better than the normal distribution (model 3 *vs.* model 2), indicating that the distribution of  $h$  values is strongly positively skewed. From model 3, we estimate that mean, variance, and skewness of  $h$  are  $E[h] = 0.037$ ,  $V[h] = 0.011$ , and  $S[h] = 2.99$ , respectively.

Next, we examine the relationship between  $h$  and  $s$ . Model 4 fits a linear relationship between  $h$  and  $s$ , whereas model 5 fits an asymptotic relationship; both models assume gamma variation in  $h$  for a given value of

$s$ . Both models 4 and 5 have vastly better AIC scores than the comparable model with no relationship between  $h$  and  $s$  (model 3). Comparing the linear and asymptotic models, we find that the latter is much better. We examined this asymptotic model using a Gaussian for the  $h$  distribution (model 6) and confirmed our earlier finding that the gamma distribution provides a much better fit (compare model 5 to model 6).

On the basis of the results above, the asymptotic relationship with gamma variance in  $h$  (model 5) is superior to any of the other options previously discussed. We extended this model to ask whether the variance in  $h$  depended on the sign and/or strength of selection. In model 7, we allowed the distribution of  $h$  for beneficial mutations to vary from the distribution of  $h$  for deleterious mutations by a factor  $\alpha^2$  (see File S1 for details).

This model yields the best AIC score and the maximum-likelihood estimate of the new parameter is  $\alpha = -5.79$  (other parameter estimates are given in Table 4). The negative sign indicates a reversal in skew for the beneficial mutations from our earlier models, so that the skew in  $h$  is positive for deleterious mutations but negative for beneficial ones. Note that in both cases fitness is skewed toward less fit values. The magnitude of this  $\alpha$ -estimate implies more variance in  $h$  for beneficial mutations than for deleterious mutations of a single  $s$  interval. This could reflect the fact that we did not attempt to account for a relationship between  $h$  and  $s$  for beneficial mutations whereas we have done so for deleterious mutations.

Finally, we built upon model 7 to ask whether the variance in  $h$  for the deleterious mutations depends on the strength of selection. However, we found no support for a relationship between the variance in  $h$  and the magnitude of  $s$ , as this model had only a slightly improved likelihood ( $\log(L) = 36,211.5$ ) despite the additional degree of freedom relative to model 7, thus resulting in a worse AIC score.

Because model 7 is the most strongly supported of all the models we considered, we discuss these estimates in more detail. Figure 1 shows the distribution of homozygous fitness described by these parameters. The model predicts that  $\sim 8\%$  of deletions have beneficial effects in homozygotes. However, most of these are predicted to have very weak effects; 70% of the supposed beneficial mutations have a homozygous effect of  $< 1\%$ . In reality, many of these may be neutral, at least in laboratory conditions. Other studies of yeast deletion lines have also indicated that some deletions are beneficial, at least under the lab conditions (BELL 2010), while other studies suggest that the proportion of beneficial deletions is somewhat smaller (SLIWA and KORONA 2005). The average dominance of beneficial mutations is estimated as  $E[h_{\text{ben}}] = 0.40$ .

The remaining 92% of mutations are considered deleterious. For deleterious mutations, the mean selec-

tion is  $E[s] = 0.045$ . There are no extremely deleterious mutations in this data set, by experimental design, so this number significantly underestimates the true value. The distribution of  $h$  values for both beneficial and deleterious mutations, predicted by model 7, is shown in Figure 2. In calculating the average dominance for deleterious mutations from this model, we need to account for the distribution of  $s$  values because of the negative relationship between  $h$  and  $s$ . Doing so, we find  $E[h_{\text{del}}] = 0.77$ . This high value occurs because the model predicts very high  $h$  for weakly deleterious alleles and most of the genes fall in the weakest class. For some purposes (*e.g.*, predicting the change in mean fitness during periods of inbreeding or purging; see the APPENDIX for calculations) it may be more useful to calculate an  $s$ -weighted average dominance ( $E_s[h_{\text{del}}] = \sum hs / \sum s$ ); we then find  $E_s[h_{\text{del}}] = 0.205$ .

However, as Figure 2 illustrates, the mean is not very informative because of the variation in  $h$ . Very weakly deleterious mutations tend to be dominant or underdominant ( $h > 1$ ) for growth rate whereas more strongly selected genes are quite strongly recessive. Moreover, even the mean value of  $h$  for a given  $s$  interval is not very informative because of the considerable variation and skew in the distribution.

The model finds that very weakly deleterious alleles are predicted to be very slightly underdominant ( $h > 1$ ); however, it is difficult to assess how much confidence to place on these observations. The power to measure such weak effects in homozygotes or heterozygotes for any particular gene is extremely low as the magnitude of the signal is very low relative to the measurement error. However, many of the data lie in this realm, so considerable power comes from the number of genes being measured. We examined a model in which the mean  $h$  within an  $s$  interval was constrained to be  $\leq 1.0$ . This model had a score that was only 0.26 log-likelihood units less than that of model 7. Thus, there is no strong evidence that weakly deleterious alleles are underdominant on average. Similarly, we examined a model in which the mean  $h$  within an  $s$  interval was constrained to be  $\leq 0.5$ . This model was worse by 25 log-likelihood units, indicating that weakly deleterious alleles tend to be dominant, rather than additive or recessive. However, it is important to remember some of the limitations of the data set when interpreting the average dominance values. The fitness of each heterozygous line was assessed in competition with all other heterozygotes, whereas homozygous lines were measured in competition with all other homozygotes. If the heterozygous pool was more competitive, and thus more selective, than the homozygous pool, then estimates of average dominance values may be upwardly biased (PHADNIS and FRY 2005).

To confirm that the results do not result from autocorrelations caused by measurement error (as suggested by BELL 2010), we conducted a simulation study of our statistical approach. We used a modified version of the

likelihood model in which the distribution of  $h$  values was based on 9 points, rather than 25, for each  $s$  interval to speed calculations; the 9-point version gives similar results to the 25-point version when analyzing the real data. Despite having only 1 d.f. more than model 3 (no relationship between  $h$  and  $s$ ), model 5 is better by 81.4 log-likelihood units in the real data. This is strongly significant by a traditional likelihood-ratio test ( $\chi^2 = 162.8$ , d.f. = 1,  $P < 10^{-15}$ ). To ensure that the  $h$  and  $s$  relationship we observed was not a statistical artifact, we simulated data sets using a model without any such relationship; specifically, we used the maximum-likelihood estimates of model 3. For each simulated data set, we obtained a likelihood score using a constrained model (equivalent to model 3) and a likelihood score using an unconstrained model (equivalent to model 5). As expected under the null hypothesis, the difference in log-likelihood scores between constrained and unconstrained models in most of the simulated data sets was  $< 2$  log-likelihood units. Of 550 simulated data sets, only one gave a greater difference in log-likelihood scores than the observed difference of 81.4 from the real data; all other simulated data sets showed much smaller differences (the second highest was 40.7). Thus, our analysis provides strong evidence for a negative relationship between  $h$  and  $s$ .

#### DOMINANCE VARIES BY PROTEIN FUNCTION

We wish also to know whether the degree of dominance depends on the functional type of the gene. We assorted the genes in the yeast knockout data set into five of the high-level ratings in the gene ontology (GO) scheme (from [www.geneontology.org](http://www.geneontology.org)), describing molecular function categories. We used the five largest classes for yeast, following PHADNIS and FRY (2005): “binding” (GO:0005488), “catalytic” (or enzymes) (GO:0003824), “structural” (GO:0005198), “transcription regulation” (GO:0030528), and “transporter” (GO:0005255). Some genes are annotated to more than one category, and we kept all genes in all GO categories that apply. Many genes are not annotated to any of these five categories, leaving only 1820 gene annotations to analyze.

We used a model II regression method with the line forced to have an intercept at zero, regressing heterozygous growth rate on homozygous growth rate, implemented with `slope.test` in the `smatr` package in R (R DEVELOPMENT CORE TEAM 2009). We based the analysis on the trimmed data sets derived in the last section, but this trimming does not have qualitative effects on the results. This regression estimator ignores the relationship between  $h$  and  $s$ , but as we will see the pattern based on protein function is the opposite of the overall pattern. The GO categories vary strongly in their dominance coefficients, with structural genes having much higher  $h$  values than other GO categories (Table 4). This is in spite of the fact that the structural proteins on average have



**TABLE 3**  
**Maximum-likelihood results**

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Model Structure							
<i>h</i> and <i>s</i> relationship	None	None	None	Linear	Asymptotic	Asymptotic	Asymptotic
<i>h</i> ( <i>s</i> ) distribution	Not variable	Normal	Gamma	Gamma	Gamma	Normal	Gamma
No. of free parameters	16	17	18	19	19	18	20
Model parameters							
$W_{wt,Hom}$	1.016	1.015	1.016	1.017	1.020	1.020	1.019
$W_{wt,Het}$	1.005	1.005	1.005	1.006	1.014	1.014	1.013
$V_{me,Hom}$	1.74e-4	1.75e-4	1.75e-4	1.72e-4	1.63e-4	1.65e-4	1.63e-4
$V_{me,Het}$	4.79e-4	4.73e-4	4.88e-4	4.84e-4	4.67e-4	3.93e-4	4.45e-4
$v_{Hom}$	24.1	24.0	23.8	24.1	24.78	23.9	24.73
$v_{Het}$	86.6	59.4	15.8	13.0	2.32	19.7	5.54
$p_{ben}$	0.211	0.232	0.215	0.167	0.071	0.049	0.078
$p_{uniform}$	0.273	0.280	0.761	0.261	0.233	0.228	0.235
$p_{exp1}$	0.761	0.754	0.985	0.783	0.885	0.898	0.886
$\mu_{ben}$	0.0070	0.0068	0.0069	0.0071	0.0086	0.0090	0.0082
$u_{max}$	0.301	0.301	0.301	0.301	0.304	0.305	0.304
$\mu_{exp1}$	0.0055	0.0054	0.0056	0.0064	0.0089	0.0094	0.0090
$\mu_{exp2}$	0.028	0.029	0.029	0.029	0.040	0.0416	0.040
$\mu_{h(ben)}$	0.976	1.00	0.985	0.935	0.4548	-0.105	0.401
$\beta_1$	0.039	0.033	0.037	0.113	4.46	2.81	4.23
$\beta_2$	—	—	—	-0.323	310.0	181.8	297.4
$\sigma_h^2$	—	0.003	0.011	0.010	0.011	0.005	0.010
$\delta$	—	—	0.042	0.037	0.037	—	0.038
$\alpha$	—	—	—	—	—	—	-5.79
Model results							
Log( <i>L</i> )	35,992	36,063	36,115	36,126	36,204	36,122	36,211
Relative AIC	431	290	188	168	12	174	0

The distribution of dominance values for a given *s* [*i.e.*, *h*(*s*) distribution] was modeled in one of three ways: not variable, based on a normal distribution, or based on a displaced gamma distribution. The relative AIC of model *i* is defined as the  $AIC_i - AIC_7$ , where  $AIC_7$  is the AIC score of the best model (model 7).

much higher selection coefficients than other categories, making the relationship between dominance and selection the opposite to the general pattern. A comparison of slopes using major axis regression [using slope.com in smatr (WARTON and WEBER 2002)] shows a strong interaction between GO class and *s* in predicting *h**s* ( $P < 0.0001$ ), but this entirely disappears when the structural proteins are removed from the analysis ( $P = 0.75$ ). Structural proteins have an unusual relationship between dominance and selection coefficient, even for deletion mutations. The elevated dominance of structural proteins was first noted by PHADNIS and FRY (2005), although they did not compare GO categories statistically.

## DISCUSSION

Despite the importance of the dominance coefficient for many evolutionary processes, there is remarkably little information about dominance coefficients of new mutations taken widely from the genome. Despite its limitations, the yeast deletion project provides unique and unusually valuable data on the effects of heterozygous and homozygous mutations spread over most of

the genome (PHADNIS and FRY 2005; BELL 2010). Because of the limitations of the data, our estimates need to be regarded with some caution. Nonetheless, we believe our analysis yields some important insights into the distribution of dominance coefficients. With these data, we have been able to confirm that the mean dominance coefficients for alleles with very small effects are high, and the dominance coefficients for alleles with mild to strongly deleterious effects on average are in the range of just under additivity to a few percent. Alleles are more likely to be recessive if they have strong homozygous effects. The mean dominance coefficient over all mutations that have a deleterious effect on fitness from these data is  $\sim 0.8$ . This high average dominance is due to the estimated high *h* values for very weakly selected loci, which constitute the majority of the data set. When the average *h* value is weighted by *s*, the mean shrinks to  $\sim 0.2$ . This latter number is qualitatively similar to the mean values obtained over a larger range of studies (Table 1). The typical mutation that affects fitness is moderately, but not completely, recessive.

More novel is the observation of a great deal of variation masked by these averages. A broad range of

TABLE 4

Mean selection coefficient and dominance coefficient by gene ontology

Gene ontology	No. of genes	$E[s]$	$h$	95% confidence limit for $h$
Binding	478	0.05	0.051	(0.035, 0.066)
Catalytic	816	0.04	0.046	(0.034, 0.059)
Structural	167	0.14	0.127	(0.099, 0.155)
Transcription regulation	162	0.05	0.062	(0.037, 0.087)
Transporter	197	0.04	0.048	(0.021, 0.076)

dominance coefficients are observed for all values of the selection coefficient. Part of this variation relates to the function of the gene. As has been suggested from anecdotal data (WILKIE 1994; KONDRASHOV and KOONIN 2004), the dominance coefficients of structural proteins are on average much higher than those for other gene ontology classes. These data are unusual in their breadth across the genome, and they are also unusual for the insight that they lend into the effects of weakly selected mutations, which have typically been ignored. To our knowledge this is the first study that attempts to quantify higher moments of the  $h$  distribution. One of the surprises of this analysis was the strong positive skew, indicating that the modal value of  $h$  is lower than the mean.

On the basis of a comparatively small number of transposable element insertions in *Drosophila*, CABALLERO and KEIGHTLEY (1994) made some inferences about the joint distribution of  $h$  and  $s$ . They proposed that  $h$ , for a given  $s$ , was uniformly distributed between 0 and  $\exp(-Ks)$ . Our results differ from this model in two ways with respect to the variance in  $h$ . First, our results indicate that, for a given  $s$ , values of  $h$  are not uniformly distributed but rather strongly skewed. Second, the Caballero and Keightley model predicts a decline in the variance in  $h$  with increasing  $s$ ; we found no support for such a decline in the yeast data set.

Our findings have strong implications about the evolution of inbreeding depression. Inbreeding depression can be caused by either deleterious recessive alleles or overdominance, and the majority of opinion suggests that recessive alleles are more important (CHARLESWORTH and CHARLESWORTH 1987). The contribution of a locus to a population's inbreeding depression depends strongly and nonlinearly on its dominance coefficient (CHARLESWORTH and CHARLESWORTH 1987); inbreeding depression is caused almost entirely by alleles with dominance coefficients close to 0 or by overdominance (Figure 3). Because of the nonlinear relationship between inbreeding depression and  $h$ , the observed variance and positive skew in  $h$  would result in more inbreeding depression than expected on the basis of average dominance.

If one ignored the variation in  $h$  and simply used the arithmetic average  $h$ ,  $E[h_{\text{del}}] = 0.77$ , then one would

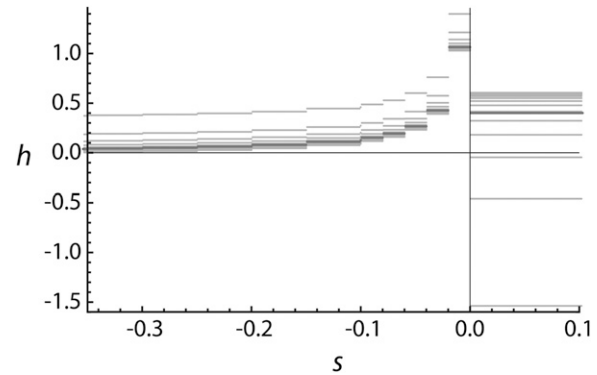


FIGURE 2.—Prediction of distribution of dominance values,  $h$ , from model 7 as a function of selection,  $s$ . Negative values of  $s$  represent deleterious mutations. The two discrete approximations used in the likelihood model are shown. For deleterious mutations, the distribution of  $h$  is constant within short intervals of  $s$ , but varies among  $s$  intervals. Within each  $s$  interval, the distribution of  $h$  values is represented by 25 equally probable values of  $h$ , shown here as vertically stacked segments with light shading; the mean,  $\mu_h$ , is shown with dark shading.

wrongly predict negative inbreeding depression (inbred offspring are more, rather than less, fit than outbred offspring). In fact, the equilibrium inbreeding depression is a function of the reciprocal of  $h$  (CHARLESWORTH and CHARLESWORTH 1987). This means that it is necessary to use the harmonic mean of  $h$  to properly account for the variation in  $h$  when predicting the effects of inbreeding across the genome. Our analysis suggests that the harmonic mean of  $h$  is  $\sim 0.18$ , thus predicting reasonably strong positive inbreeding depression.

Prior inbreeding in a population purges the population of inbreeding depression most effectively against deleterious alleles with  $h$  close to zero (WHITLOCK 2002; see dashed line in Figure 3). If the previous bout of inbreeding is of short duration, purging works only against alleles of large effect, *i.e.*, when  $s$  is large. Therefore we expect purging to be important only for strongly deleterious alleles with small dominance coefficients. The results of this study suggest that a considerable fraction of the genome is in this category, reinforcing the expectation that purging can be partially effective. In the APPENDIX, we derive equations for the short-term changes in fitness that occur as result of a single generation of purging. The correlation between  $h$  and  $s$  is strong enough that the predicted change in mean fitness during purging is  $\sim 3.5$  times greater than that predicted using approximations based on the mean  $h$  and mean  $s$ . In the APPENDIX, we also calculated the expected change caused by purging in the magnitude of inbreeding depression. The expected change in inbreeding depression is of much larger magnitude and of opposite sign when accounting for the relationship between  $h$  and  $s$  than would be erroneously predicted using the mean values of these quantities.

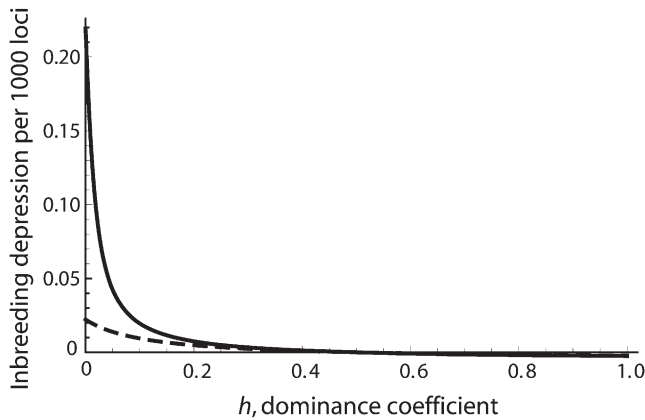


FIGURE 3.—Inbreeding depression contributed by 1000 loci, with fixed selection and dominance parameters. The inbreeding depression expected by the equations in WHITLOCK (2002) is shown, in which inbreeding depression depends strongly on the dominance coefficient, but is not affected by the selection coefficient, provided  $s > 0$  and  $h > 0$ . The inbreeding coefficient of the measured individuals is shown as  $\frac{1}{4}$ , corresponding to full-sib mating. The solid line shows the inbreeding depression expected at mutation–selection balance in a panmictic population, and the dashed line shows inbreeding depression in a population that was previously at mutation–selection equilibrium with  $F_{ST} = 0.1$ . Inbreeding depression is strongly affected by the dominance coefficient, and purging by previous inbreeding is also greatest at low  $h$ .

The data also show great variation in  $h$ , such that a large fraction of deleterious alleles do not have low dominance coefficients, even among the loci with large  $s$  values, suggesting that a large number of deleterious alleles in a population would not be strongly affected by purging through inbreeding.

In our survey of the yeast data, we find a strong relationship between  $h$  and  $s$ , as first reported by PHADNIS and FRY (2005). Our estimated relationship between  $h$  and  $s$  is similar to that inferred by DENG and LYNCH (1996) on the basis of rather limited *Drosophila* data. Deng and Lynch assumed an exponential relationship between  $h$  and  $s$  for deleterious alleles,  $\mu_{h(\text{del})} = \frac{1}{2} \exp(-Bs)$ . They inferred  $B = 13$  on the basis of a single pair of values, estimates of the average  $h$  and  $s$  from mutation accumulation studies. Surprisingly, their inferred function qualitatively matches the one measured here, at least for deleterious alleles of moderate effect,  $s > 0.04$ . Deng and Lynch point out that measurements of mutation parameters based on genomic mutation accumulation experiments can be extremely biased if the relationship between  $h$  and  $s$  is not properly included.

This result has strong implications for models of the evolution of dominance. Many models of the evolution of dominance of deleterious alleles have been proposed (BOURGUET 1999). The models are of basically three types, alternatively concluding (1) that dominance arises from direct selection to modify the fitness effect of heterozygotes, (2) that dominance arises as an indirect

result of selection on another aspect of the biology of organisms, or (3) that dominance is a natural result of metabolic pathways, without respect to the selective process that refines those pathways.

The earliest evolutionary model of dominance was of the first type. FISHER (1928) proposed that dominance was the result of selection modifying the fitness effects of heterozygotes that were present in a population by recurrent deleterious mutation to the deleterious allele. This model has been rejected by the majority of population geneticists as a result of WRIGHT's (1929, 1934) demonstration that the strength of selection on such modifiers would be too weak (because the frequency of the heterozygotes would be proportional to the mutation rate) and therefore not capable of overcoming genetic drift or mutation at the modifier locus itself. However, other models of direct selection have been proposed (CLARKE 1964; WAGNER and BÜRGER 1985; OTTO and BOURGUET 1999), but these are typically applicable only in special cases.

Several models have proposed that dominance has evolved as a result of indirect selection of various sorts. HALDANE (1939) proposed that dominance may result from a buffering process, whereby alleles are ordinarily overexpressed to buffer for stressful times. Similarly, HURST and RANDERSON (2000) predicted that weak beneficial effects may accrue from higher expression even in homozygous individuals [if the benefits of increased flux through a pathway are monotonically related to fitness, as has been assumed in simple models of metabolic control theory (KACSER and BURNS 1981)]. Such constitutive higher expression means that heterozygotes may have sufficient quantities of gene product to do relatively well. Similarly, OMHOLT *et al.* (2000) proposed that dominance may be a by-product of the feedback mechanisms between loci. Under this model, dominance results from the up-regulation of a locus that has some, but not sufficient, activity. By this view, a heterozygote may be up-regulated to produce enough active enzyme, but a homozygote with two defective copies would not be able to complete its function even after up-regulation. The importance of regulatory feedback for selectively important genes is consistent with the greater extent of transcriptional compensation of deletion heterozygotes for highly expressed genes (MCANALLY and YAMPOLSKY 2010).

The preeminent mode of physiological dominance, the metabolic control theory of KACSER and BURNS (1981; see also KEIGHTLEY and KACSER 1987; KEIGHTLEY 1996), predicts that dominance is an epiphenomenon of the biochemistry of metabolic pathways. In this model, variation among enzymes in their catalytic capacity means that some enzymes will be more critical to the overall function of a pathway than others and that for most enzymes the relationship between dosage (or capacity) of that enzyme and the rate of flux through the pathway shows diminishing returns. As a result, most

enzymes ought to show recessive deleterious mutations. This model is perhaps the most broadly accepted theory of dominance currently discussed.

CHARLESWORTH (1979) argued that a negative relationship between  $h$  and  $s$  precluded FISHER's (1928) model of the evolution of dominance via modifiers on standing variation. By Fisher's model, as elucidated by WRIGHT (1929,1934), the strength of selection on a modifier should be proportional to its allele frequency times its effect. At an equilibrium between mutation and selection, the frequency of an allele should be inversely proportional to its heterozygous effect. In other words, modifiers should evolve equally often for alleles of weak effect or of strong effect, because the weak-effect alleles occur more often in the population exactly in inverse proportion to their weaker effects.

Metabolic control theory (KACSER and BURNS 1981) makes no explicit prediction about the relationship across loci between  $h$  and  $s$  for null mutations. However, some models derived from metabolic control theory (e.g., HURST and RANDERSON 2000) do make this prediction. Most models of dominance via indirect selection predict a relationship between  $h$  and  $s$ . In both Haldane's and Hurst and Randerson's models, there are fitness advantages to a higher level of expression than is necessary for normal survival; under both models one would predict that such expression would be increased more when selection on the effects of that locus was greater. As a result, both models predict that loci with potentially strong selection should be more likely to produce recessive alleles. Similarly, in the OMHOLT *et al.* (2000) model, dominance results from the up-regulation of a locus that has some, but not sufficient, activity. If one assumes that the strength of selection for an effective feedback mechanism is proportional to the strength of selection acting against defective homozygotes at a locus, then this model also predicts a relationship between  $h$  and  $s$ . All of these models that hold that dominance is an indirect response to selection on wild-type homozygous expression are consistent with the observed pattern of a negative relationship between  $s$  and  $h$ .

The dominance coefficient is a crucial parameter for many important evolutionary processes. Despite a fair number of studies measuring  $h$ , we still have very little reliable information even about the mean value of this parameter. We still know very little about the variability of  $h$ , and we very much require more information about the relationship between  $h$  and other features of mutations, like the selection coefficient. These yeast data are extremely valuable, and we can learn a great deal about dominance from them. However, evolutionary genetics requires more data of this sort, from other species and from a more representative spectrum of mutations.

We are very grateful for useful comments by Deborah Charlesworth, Thomas Lenormand, an anonymous reviewer and, especially, Jim Fry. This article was produced with support from Discovery Grants from the

Natural Science and Engineering Research Council (Canada) (to M.C.W. and A.F.A.).

## LITERATURE CITED

- BELL, G., 2010 Experimental genomics of fitness in yeast. *Proc. R. Soc. B* **277**:1459–1467.
- BOURGUET, D., 1999 The evolution of dominance. *Heredity* **83**: 1–4.
- BOURGUET, D., T. LENORMAND, T. GUILLEMAUD, V. MARCEL, D. FOURNIER *et al.*, 1997 Variation of dominance of newly arisen adaptive genes. *Genetics* **147**: 1225–1234.
- CABALLERO, A., and P. D. KEIGHTLEY, 1994 A pleiotropic nonadditive model of variation in quantitative traits. *Genetics* **138**: 883–900.
- CABALLERO, A., P. D. KEIGHTLEY and M. TURELLI, 1997 Average dominance for polygenes: drawbacks of regression estimates. *Genetics* **147**: 1487–1490.
- CHARLESWORTH, B., 1979 Evidence against Fisher's theory of dominance. *Nature* **278**: 848–849.
- CHARLESWORTH, B., and D. CHARLESWORTH, 1987 Inbreeding depression and its evolutionary consequences. *Annu. Rev. Ecol. Syst.* **18**: 237–268.
- CHARLESWORTH, B., and D. CHARLESWORTH, 1998 Some evolutionary consequences of deleterious mutations. *Genetica* **102/103**: 3–19.
- CHARLESWORTH, D., M. T. MORGAN and B. CHARLESWORTH, 1990 Inbreeding depression, genetic load, and the evolution of outcrossing rates in a multilocus system with no linkage. *Evolution* **44**: 1469–1489.
- CHAVARRÍAS, D., C. LÓPEZ-FANJUL and A. GARCÍA-DORADO, 2001 The rate of mutation and the homozygous and heterozygous mutational effects for competitive viability: a long-term experiment with *Drosophila melanogaster*. *Genetics* **158**: 681–693.
- CLARKE, C. A., 1964 Frequency-dependent selection for the dominance of rare polymorphic genes. *Evolution* **18**: 364–369.
- DENG, H.-W., and M. LYNCH, 1996 Estimation of deleterious-mutation parameters in natural populations. *Genetics* **144**: 349–360.
- FISHER, R. A., 1928 The possible modification of the response of the wild type to recurrent mutations. *Am. Nat.* **62**: 115–126.
- FRY, J. D., 2004 On the rate and linearity of viability declines in *Drosophila* mutation-accumulation experiments: genomic mutation rates and synergistic epistasis revisited. *Genetics* **166**: 797–806.
- FRY, J. D., and S. V. NUZHGIN, 2003 Dominance of mutations affecting viability in *Drosophila melanogaster*. *Genetics* **163**: 1357–1364.
- GARCÍA-DORADO, A., and A. CABALLERO, 2000 On the average coefficient of dominance of deleterious spontaneous mutations. *Genetics* **155**: 1991–2001.
- GIAEVER, G., A. M. CHU, L. NI, C. CONNELLY, L. RILES *et al.*, 2002 Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**: 387–391.
- HALDANE, J. B. S., 1927 A mathematical theory of natural and artificial selection, part V: selection and mutation. *Proc. Camb. Philos. Soc.* **23**: 838–844.
- HALDANE, J. B. S., 1937 The effect of variation in fitness. *Am. Nat.* **71**: 337–349.
- HALDANE, J. B. S., 1939 The theory of the evolution of dominance. *J. Genet.* **37**: 365–374.
- HEDRICK, P. W., 1994 Purging inbreeding depression and the probability of extinction: full-sib mating. *Heredity* **73**: 363–372.
- HILLENMEYER, M. E., E. FUNG, J. WILDENHAIN, S. E. PIERCE, S. HOON *et al.*, 2008 The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320**: 362–365.
- HOULE, D., K. A. HUGHES, S. ASSIMACOPOULOS and B. CHARLESWORTH, 1997 The effects of spontaneous mutation on quantitative traits. II. Dominance of mutations with effects on life-history traits. *Genet. Res. Camb.* **90**: 27–34.
- HUGHES, T. R., C. J. ROBERTS, H. DAI, A. R. JONES, M. R. MEYER *et al.*, 2000 Widespread aneuploidy revealed by DNA microarray expression profiling. *Nature Genetics* **25**: 333–337.
- HURST, L. D., and J. P. RANDERSON, 2000 Dosage, deletions and dominance: simple models of the evolution of gene expression. *J. Theor. Biol.* **205**: 641–647.
- KACSER, H., and J. A. BURNS, 1981 The molecular basis of dominance. *Genetics* **97**: 639–666.



- KEIGHTLEY, P. D., 1996 Metabolic models of selection response. *J. Theor. Biol.* **182**: 311–316.
- KEIGHTLEY, P. D., and H. KACSER, 1987 Dominance, pleiotropy and metabolic structure. *Genetics* **117**: 319–329.
- KONDRASHOV, F. A., and E. V. KOONIN, 2004 A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet.* **20**: 287–291.
- LANDE, R., and D. W. SCHEMSKE, 1985 The evolution of self-fertilization and inbreeding depression in plants. I. Genetic models. *Evolution* **39**: 24–40.
- MACKAY, T. F. C., R. F. LYMAN and M. S. JACKSON, 1992 Effects of P element insertions on quantitative traits in *Drosophila melanogaster*. *Genetics* **130**: 315–332.
- MCANALLY, A. A., and L. Y. YAMPOLSKY, 2010 Widespread transcriptional autosomal dosage compensation in *Drosophila* correlates with gene expression level. *Genome Biol. Evol.* **2**: 44–52.
- MORTON, N. E., 1971 Population genetics and disease control. *Soc. Biol.* **18**: 243–251.
- MUKAI, T., 1969 The genetic structure of natural populations of *Drosophila melanogaster*. VI. Further studies on the optimum heterozygosity hypothesis. *Genetics* **61**: 479–495.
- MUKAI, T., and T. YAMAZAKI, 1968 The genetic structure of natural populations of *Drosophila melanogaster*. V. Coupling-repulsion effect of spontaneous mutant polygenes controlling viability. *Genetics* **59**: 513–535.
- MUKAI, T., S. CHIGUSA and I. YOSHIKAWA, 1964 The genetic structure of natural populations of *Drosophila melanogaster*. II. Overdominance of spontaneous mutant polygenes controlling viability in homozygous genetic background. *Genetics* **50**: 711–715.
- MUKAI, T., S. CHIGUSA and I. YOSHIKAWA, 1965 The genetic structure of natural populations of *Drosophila melanogaster*. III. Dominance effect of spontaneous mutant polygenes controlling viability in heterozygous genetic backgrounds. *Genetics* **52**: 493–501.
- MULLER, H. J., 1950 Our load of mutations. *Am. J. Hum. Gen.* **2**: 111–176.
- OHNISHI, O., 1977a Spontaneous and ethyl methanesulfonate-induced mutations controlling viability in *Drosophila melanogaster*. II. Homozygous effect of polygenic mutations. *Genetics* **87**: 529–545.
- OHNISHI, O., 1977b Spontaneous and ethyl methanesulfonate-induced mutations controlling viability in *Drosophila melanogaster*. III. Heterozygous effect of polygenic mutations. *Genetics* **87**: 547–556.
- OMHOLT, S. W., E. PLAHTÉ, L. ØYEHAUG and K. XIANG, 2000 Gene regulatory networks generating the phenomena of additivity, dominance, and epistasis. *Genetics* **155**: 969–980.
- ORR, H. A., 2003 The distribution of fitness effects among beneficial mutations. *Genetics* **163**: 1519–1526.
- ORR, H. A., and A. J. BETANCOURT, 2001 Haldane's sieve and adaptation from the standing genetic variation. *Genetics* **157**: 875–884.
- OTTO, S. P., and D. BOURGUET, 1999 Balanced polymorphisms and the evolution of dominance. *Am. Nat.* **153**: 561–574.
- PETERS, A. D., D. L. HALLIGAN, M. C. WHITLOCK and P. D. KEIGHTLEY, 2003 Dominance and overdominance of mildly deleterious induced mutations for fitness traits in *Caenorhabditis elegans*. *Genetics* **165**: 589–599.
- PHADNIS, N., and J. D. FRY, 2005 Widespread correlations between dominance and homozygous effects of mutations: implications for theories of dominance. *Genetics* **171**: 385–392.
- PROVINE, W. R., 1986 *Sevall Wright and Evolutionary Biology*. University of Chicago Press, Chicago.
- R DEVELOPMENT CORE TEAM, 2009 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. <http://www.R-project.org>.
- SGD PROJECT, 2009 *Saccharomyces Genome Database*. Accessed July 7, 2009. <http://www.yeastgenome.org/>
- SHOEMAKER, D. D., D. A. LASHKARI, D. MORRIS, M. MITTMAN and R. W. DAVIS, 1996 Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Genet.* **14**: 450–456.
- SIMMONS, M. J., and J. F. CROW, 1977 Mutations affecting fitness in *Drosophila* populations. *Annu. Rev. Genet.* **11**: 49–78.
- SLIWA, P., and R. KORONA, 2005 Loss of dispensable genes is not adaptive in yeast. *Proc. Natl. Acad. Sci. USA* **102**: 17670–17674.
- STEINMETZ, L. M., C. SCHARFE, A. M. DEUTSCHBAUER, D. MOKRANJAC, Z. S. HERMAN *et al.*, 2002 Systematic screening for human disease genes in yeast. *Nat. Genet.* **31**: 400–404.
- SZAFRANIEC, K., D. M. WLOCH, P. SLIWA, R. H. BORTS and R. KORONA, 2003 Small fitness effects and weak genetic interactions between deleterious mutations in heterozygous loci of the yeast *Saccharomyces cerevisiae*. *Genet. Res. Camb.* **82**: 19–31.
- VASSILIEVA, L., A. HOOK and M. LYNCH, 2000 The fitness effects of spontaneous mutations in *Caenorhabditis elegans*. *Evolution* **54**: 1234–1246.
- VEITIA, R. A., 2005 Gene dosage balance: deletions, duplications and dominance. *Trends Genet.* **21**: 33–35.
- WAGNER, G., and R. BÜRGER, 1985 On the evolution of dominance modifiers II: a non-equilibrium approach to the evolution of genetic systems. *J. Theor. Biol.* **113**: 475–500.
- WARTON, D. I., and N. C. WEBER, 2002 Common slope tests for bivariate errors-in-variables models. *Biom. J.* **44**: 161–174.
- WHITLOCK, M. C., 2002 Selection, load, and inbreeding depression in a large metapopulation. *Genetics* **160**: 1191–1202.
- WILKIE, A. O. M., 1994 The molecular basis of genetic dominance. *J. Med. Genet.* **31**: 89–98.
- WRIGHT, S., 1929 Fisher's theory of dominance. *Am. Nat.* **63**: 274–279.
- WRIGHT, S., 1934 Physiological and evolutionary theories of dominance. *Am. Nat.* **67**: 24–53.

Communicating editor: D. CHARLESWORTH

## APPENDIX: TRANSIENT EFFECTS OF PURGING

Consider a single locus that mutates from the wild-type allele  $A$  to the deleterious mutant allele  $a$  at rate  $\mu$ . The fitnesses of the three genotypes  $AA$ ,  $Aa$ , and  $aa$  are 1,  $1 - hs$ , and  $1 - s$ . Assuming  $hs \gg \mu$  and panmixis, the equilibrium frequency of the mutant allele is  $q_{eq} \approx \mu / hs$ .

The mean fitness of outcrossed progeny is

$$E_{eq}[W_{out}] = (1 - q_{eq})^2 + 2(1 - q_{eq})q_{eq}(1 - hs) + q_{eq}^2(1 - s) \approx 1 - 2\mu.$$

For offspring produced by experimental inbreeding at rate  $f$ , the mean fitness is

$$E_{eq}[W_{in}] = (1 - f)E_{eq}[W_{out}] + f((1 - q_{eq}) + q_{eq}(1 - s)) \approx 1 - \left(2(1 - f) + \frac{f}{h}\right)\mu.$$

Inbreeding depression is given by

$$\delta_{\text{eq}} = 1 - \frac{E_{\text{eq}}[W_{\text{in}}]}{E_{\text{eq}}[W_{\text{out}}]} \approx \frac{f(1-2h)}{h} \mu.$$

The results above were previously derived (HALDANE 1937; CHARLESWORTH and CHARLESWORTH 1998; WHITLOCK 2002).

We now consider changes that occur as a result of a single generation of purging by experimental inbreeding at rate  $f^*$ . It is important to emphasize that the changes reported below are transient and refer to values in the generation immediately following the purging event. The change in the mutant frequency after one generation of inbreeding and selection is

$$\Delta q_{\text{purge}} \approx -\frac{f^* + h(1-f^*)}{h} \mu.$$

Following mutation, the frequency of the mutant at the start of the next generation is  $q' \approx q_{\text{eq}} + \Delta q_{\text{purge}} + \mu$ . Using this allele frequency, we can calculate mean fitness of outbred and inbred offspring after a generation of purging, which we denote as  $E[W_{\text{out}}]$  and  $E[W_{\text{in}}]$ , respectively. The change in mean fitness of outbred offspring due to a single round of purging is

$$\Delta E[W_{\text{out}}] = E'[W_{\text{out}}] - E_{\text{eq}}[W_{\text{out}}] \approx 2f^*(1-h)s\mu.$$

Therefore, when considering the effects of purging across multiple loci, the relevant measure of dominance is the  $s$ -weighted average of  $h$ ,  $E_s[h_{\text{del}}] = \sum hs / \sum s$  (the summation is across all genes). From our data, we estimate  $E_s[h_{\text{del}}] = 0.205$  whereas the arithmetic average  $h = 0.77$ . The former value is more appropriate for predicting the effect of purging on mean fitness, and it results in a  $\sim 3.5$ -fold larger effect than predicted by the erroneous use of  $h = 0.77$ .

We can also ask how inbreeding depression changes as a result of a single generation of purging. Here we assume that fitnesses of inbred and outbred individuals are measured in which the experimental inbreeding coefficient is  $f$ . The experiment is performed twice, once before and once after a generation of purging, where purging involves a generation of inbreeding at rate  $f^*$ . The change in inbreeding depression due to purging is given by

$$\Delta \delta = \delta' - \delta_{\text{eq}} = \frac{E_{\text{eq}}[W_{\text{in}}]}{E_{\text{eq}}[W_{\text{out}}]} - \frac{E'[W_{\text{in}}]}{E'[W_{\text{out}}]} \approx -f \times f^* s \phi \mu,$$

where  $\phi = (1 - 3h + 2h^2)/h$  is the relevant dominance term. Therefore, when considering the mean change in inbreeding depression summed over multiple loci, we need an  $s$ -weighted average of  $\phi$ ,  $E_s[\phi_{\text{del}}] = \sum \phi s / \sum s$ . From our data set, we estimate  $E_s[\phi_{\text{del}}] = 20.2$ , indicating that purging would cause a substantial reduction in inbreeding depression. If we calculate  $\phi$  using the average arithmetic  $h$ ,  $E[h_{\text{del}}] = 0.77$ , we find  $\phi = -0.16$ , which leads to a dramatically different prediction about the effect of purging on inbreeding depression.

# GENETICS

## **Supporting Information**

<http://www.genetics.org/cgi/content/full/genetics.110.124560/DC1>

## **Inferences About the Distribution of Dominance Drawn From Yeast Gene Knockout Data**

**Aneil F. Agrawal and Michael C. Whitlock**

Copyright © 2011 by the Genetics Society of America  
DOI: 10.1534/genetics.110.124560

## FILE S1

**Likelihood model:** The total log-likelihood can be decomposed into two parts representing observed estimates of fitness from wild types and mutants, respectively. Using  $L$  to represent the likelihood with subscripts to distinguish parameters and data for different subsets of genotypes, the log-likelihood of the full model can be written as:

$$\log(L_{Total}[\text{parameters} \mid \text{all data}]) = \log(L_{wt}[\text{parameters} \mid \text{wild type data}]) + \log(L_{mutant}[\text{parameters} \mid \text{mutant data}])$$

where  $wt$  in the subscript indicates wild type genotypes and  $mutant$  indicates knockout lines.

The log-likelihood function for wild types can be further decomposed into two parts representing independent information from the heterozygous and homozygous screens:

$$\log(L_{wt}[\text{parameters} \mid \text{wild type data}]) = \log(L_{wt,Het}[\text{parameters} \mid \text{heterozygous wild type data}]) + \log(L_{wt,Hom}[\text{parameters} \mid \text{homozygous wild type data}])$$

Here, “heterozygous” wild-type data and “homozygous” wild-type data refer to estimates of fitness from genotypes that are heterozygous or homozygous for putatively neutral deletions. In principle, the fitness of these two types should not differ ( $W_{wt,Het} = W_{wt,Hom}$ ), and indeed we obtain very similar estimates for these parameters. (In practice, these measures could differ for at least two reasons. First, if other random mutations have occurred in the genome since the deletion lines were created, then lines homozygous for the neutral deletion are more likely to be homozygous for other mutations. Second, fitness of “heterozygous” and “homozygous” wild-types is measured by growth rate with heterozygous or homozygous deletion pools, respectively, with the former expected to be more competitive than the latter [Phadnis and Fry 2005].)

We treated each of the two replicates of each of the 120 putatively neutral deletions as an independent measure of wild-type fitness, yielding 240 estimates of wild type fitness for each of the heterozygous and homozygous states. We denote the true growth rate of a genotype by a capital letter  $W$ , and we use lower case  $w$  to refer to the observed value of that growth rate. Assuming measurement error is normally distributed around zero, then

$$\log(L_{wt,Het}[W_{wt,Het}, V_{me,Het} \mid \text{heterozygous wild type data}]) = \sum_{i=1}^{240} \log(N(0, V_{me,Het}, w_{wt,Het,i} - W_{wt,Het}))$$

where  $N(0, v, x)$  is the probability density function for a normally distributed variable with mean zero and variance  $v$  evaluated at  $x$ , and  $w_{wt,Het,i}$  is the  $i^{th}$  measure of heterozygous wild type fitness. The log-likelihood function for homozygous wild type fitness is defined analogously.

The log likelihood function for the fitness of mutants for a given distribution of dominance and selection is given by



$$\log(L_{mutant}[\text{parameters} \mid \text{mutant data}]) = \sum_i^T \log \left( \begin{array}{l} p_{ben} L_{ben}[\text{parameters} \mid \text{data}_i] \\ + (1 - p_{ben}) p_{uniform} \times L_{uniform}[\text{parameters} \mid \text{data}_i] \\ + (1 - p_{ben})(1 - p_{uniform}) p_{exp1} \times L_{exp1}[\text{parameters} \mid \text{data}_i] \\ + (1 - p_{ben})(1 - p_{uniform})(1 - p_{exp1}) \times L_{exp2}[\text{parameters} \mid \text{data}_i] \end{array} \right)$$

where the summation is over all  $T$  mutant genes and  $\text{data}_i$  represents the two replicate measures of homozygous and heterozygous fitnesses for gene  $i$ . For each gene, the four terms in parentheses account for the likelihood of the data if the gene comes from any of the four components that comprise the compound  $s$ -distribution. Below we define each of these four likelihood sub-functions in the spirit in which they were intended by representing them as continuous functions. Because the full equations do not integrate across the distributions of both  $s$  and  $h$ , it was necessary to use discrete approximations and these will be explained later.

If the homozygous effect of the mutation is beneficial, then the likelihood of the data would, ideally, be given by

$$L_{ben}[\text{parameters} \mid \text{data}_i] = \iint \left[ \begin{array}{l} \frac{1}{\mu_{ben}} \exp(-s/\mu_{ben}) H_{ben}(h \mid \mu_{h(ben)}, \sigma_h^2, \delta) \\ \times \prod_{j=1}^2 N(0, V_1(hs), w_{mut, Het, i, j} - (1 + hs)W_{wt, Het}) \\ \times \prod_{j=1}^2 N(0, V_2(s), w_{mut, Hom, i, j} - (1 + s)W_{wt, Hom}) \end{array} \right] ds dh$$

where  $w_{mut, Hom, i, j}$  is the  $j^{th}$  estimate ( $j \in \{1, 2\}$ ) of homozygous fitness for gene  $i$  and  $w_{mut, Het, i, j}$  is the corresponding values with respect to heterozygous fitness for the same gene.  $H_{ben}(h \mid \mu_{h(ben)}, \sigma_h^2, \delta)$  is the probability density function of  $h$  values for beneficial mutations. This distribution has mean  $\mu_{h(ben)}$  and variance  $\sigma_h^2$  and the shape follows that of either a normal or gamma distribution. If the model is employing a gamma distribution, then it is based on the shape of a gamma distribution with mean  $\delta$  and variance  $\sigma_h^2$ , which corresponds to a shape parameter of  $\delta^2 / \sigma_h^2$  and a scale parameter of  $\sigma_h^2 / \delta$ . This distribution is then shifted to have the desired mean of  $\mu_{h(ben)}$ , rather than  $\delta$ . We refer to  $\delta$  as the displacement parameter; its function is to allow the skew of the gamma distribution to be determined independently from the mean. (The distribution of  $h$  is explained in greater detail below.)

The likelihood sub-functions for genes with deleterious homozygous effects are given by

$$L_{uniform}[\text{parameters} \mid \text{data}_i] = \iint \left[ \begin{array}{l} \frac{1}{u_{max}} H_{del}(h \mid s, \beta_1, \beta_2, \sigma_h^2, \delta) \\ \times \prod_{j=1}^2 N(0, V_1(hs), w_{mut, Het, i, j} - (1 - hs)W_{wt, Het}) \\ \times \prod_{j=1}^2 N(0, V_2(s), w_{mut, Hom, i, j} - (1 - s)W_{wt, Hom}) \end{array} \right] ds dh$$

and

$$L_{expX}[\text{parameters} | data_i] = \iint \left[ \frac{1}{\mu_{expX}} \exp(-s/\mu_{expX}) H_{del}(h | s, \beta_1, \beta_2, \sigma_h^2, \delta) \right. \\ \times \prod_{j=1}^2 N(0, V_1(hs), w_{mut, Het, i, j} - (1 - hs)W_{wt, Het}) \\ \times \left. \prod_{j=1}^2 N(0, V_2(s), w_{mut, Hom, i, j} - (1 - s)W_{wt, Hom}) \right] s dh$$

where the subscript  $expX$  is used to generically depict either of the exponential components of the distribution of deleterious effects.  $H_{del}(h | s, \beta_1, \beta_2, \sigma_h^2, \delta)$  is the probability density function of  $h$  values for deleterious mutations. Although,  $H_{del}$  and  $H_{ben}$  are similar, they are not quite the same. Dominance values do not depend on  $s$  for beneficial mutations in the model. This was assumed because of lack of statistical power to make detailed estimation of patterns of dominance for beneficial mutations; most mutations are not beneficial and those that are seem to have very weak effects. For  $H_{del}$ , the mean value of  $h$  for a given  $s$  was modeled with either a linear ( $\mu_{h(del)}(s) = \beta_1 + \beta_2 s$ ) or an asymptotic ( $\mu_{h(del)}(s) = \beta_1 / (1 + \beta_2 s)$ ) relationship between  $h$  and  $s$ . By setting  $\beta_2 = 0$ , we can calculate the likelihood of no relationship.

In order to obtain a tractable likelihood function we discretized the  $h$  distribution in two ways. First, we assumed that the distribution of  $h$  was constant across small intervals of  $s$ , but variable among these intervals. For deleterious mutations, we used 17 intervals of  $s$ ; the boundary points of these intervals are given by the elements of the vector  $\mathbf{S} = \{0, 0.02, 0.04, 0.06, 0.08, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ ; narrower intervals were used for smaller  $s$  because this is where most of the data lie and also where the greatest change in  $h$  seems to occur. We use  $s_j = (S_j + S_{j+1})/2$  to represent the median value of the  $j^{th}$   $s$ -interval, where  $S_j$  is the  $j^{th}$  element of  $\mathbf{S}$ . Second, within each  $s$ -interval, we approximated a continuous distribution (either normal or displaced gamma) for  $h$  as a discrete distribution with 25 equally probable values for  $h$ .

The likelihood sub-functions for deleterious mutations then become

$$L_{uniform}[\text{parameters} | data_i] \\ = \sum_{j=1}^{17} \sum_{k=1}^{25} \frac{1}{25} \int_{S_j}^{S_{j+1}} \left[ \frac{1}{u_{max}} \times \prod_{j=1}^2 N(0, V_1(hs_j), w_{mut, Het, i, j} - (1 - h_{del, k}[s_j, \beta_1, \beta_2, \sigma_h^2, \delta] s)W_{wt, Het}) \right. \\ \times \left. \prod_{j=1}^2 N(0, V_2(s_j), w_{mut, Hom, i, j} - (1 - s)W_{wt, Hom}) \right] ds$$

and

$$L_{expX}[\text{parameters} | data_i] = \sum_{j=1}^{17} \sum_{k=1}^{25} \frac{1}{25} \int_{s_j}^{s_{j+1}} \left[ \begin{aligned} & \frac{1}{\mu_{expX}} \exp(-s/\mu_{expX}) \\ & \times \prod_{j=1}^2 N(0, V_1(hs_j), w_{mut, Het, i, j} - (1 - h_{del, k}[s_j, \beta_1, \beta_2, \sigma_h^2, \delta] s) W_{wt, Het}) \\ & \times \prod_{j=1}^2 N(0, V_2(s_j), w_{mut, Hom, i, j} - (1 - s) W_{wt, Hom}) \end{aligned} \right] ds$$

where the first  $\Sigma$  represents summation over the 17  $s$ -intervals, the second  $\Sigma$  represents summation over the 25 discrete values of  $h$  within each  $s$ -interval, and the integration is over all values of  $s$  within each  $s$ -interval.  $h_{del, k}[s, \beta_1, \beta_2, \sigma_h^2, \delta]$  is the  $k^{th}$  value (of 25) of  $h$  for deleterious mutations. (Note that the appropriate adjustments were applied in using  $L_{uniform}$  so that integration did not occur beyond its maximum  $u_{max}$ ).

Because  $H_{ben}$  was assumed to be independent of  $s$ , it was not necessary to use multiple  $s$ -intervals for beneficial mutations. The likelihood sub-functions for beneficial mutations is given by

$$L_{ben}[\text{parameters} | data_i] = \sum_{k=1}^{25} \frac{1}{25} \int \left[ \begin{aligned} & \frac{1}{\mu_{ben}} \exp(-s/\mu_{ben}) \\ & \times \prod_{j=1}^2 N(0, V_1(hs), w_{mut, Het, i, j} - (1 + h_{ben, k}[\mu_{h(ben)}, \sigma_h^2, \delta] s) W_{wt, Het}) \\ & \times \prod_{j=1}^2 N(0, V_2(s), w_{mut, Hom, i, j} - (1 + s) W_{wt, Hom}) \end{aligned} \right] ds$$

where  $h_{ben, k}[\mu_{h(ben)}, \sigma_h^2, \delta]$  is the  $k^{th}$  value (of 25 intervals) of  $h$  for beneficial mutations.

**Modeling the distribution of  $h$ :** After specifying the average within the  $j^{th}$   $s$ -interval,  $\mu_{h(del), j}$ , we also considered variation in  $h$  for each value of  $s$ . We tested two alternative distributions of  $h$  values within each interval. First, we used a normal distribution as the basis for generating the 25 values for  $h$  within each interval. Specifically, the  $k^{th}$  value (of 25) of  $h$  within the  $j^{th}$   $s$ -interval was given by

$$h_{del, k}[s_j, \beta_1, \beta_2, \sigma_h^2, \delta] = \mu_{h(del), j} + Q_N(0, \sigma_h^2, \frac{1}{25}(k - \frac{1}{2}))$$

where  $Q_N(\bar{x}, \sigma_x^2, x)$  is the quantile (or inverse cumulative density) function of a normal distribution with mean  $\bar{x}$  and variance  $\sigma_x^2$  evaluated at  $x$ .

As an alternative, we used a displaced gamma distribution as the basis for generating the 25 values for  $h$  within each interval:

$$h_{del, k}[s_j, \beta_1, \beta_2, \sigma_h^2, \delta] = \mu_{h(del), j} + Q_G(\delta^2 / \sigma_h^2, \sigma_h^2 / \delta, \frac{1}{25}(k - \frac{1}{2})) - d$$

where  $Q_G(a, b, x)$  is the quantile function of a gamma distribution with shape parameter  $a$  and scale parameter  $b$  evaluated at  $x$ . In both the normal and gamma models, the parameter  $\sigma_h^2$  represents the variance of the distribution. In the displaced gamma model,  $\delta$  is the mean of the gamma distribution, but the  $h_{del,k}$  values are displaced to have mean  $\mu_{h(del),j}$  because  $d = (1/25)\sum_k Q_G(\delta^2/\sigma_h^2, \sigma_h^2/\delta, \frac{1}{25}(k - \frac{1}{2})) \approx \delta$ .

The dominance coefficients for beneficial mutations were generated in a similar manner except that the mean dominance value,  $\mu_{h(ben)}$ , was assumed to be independent of  $s$ . For normal-based models, the dominance values are given by

$$h_{ben,k}[\mu_{h(ben)}, \sigma_h^2] = \mu_{h(ben)} + Q_N(0, \sigma_h^2, \frac{1}{25}(k - \frac{1}{2})).$$

For gamma-based models, the dominance values are given by

$$h_{ben,k}[\mu_{h(ben)}, \sigma_h^2, \delta] = \mu_{h(ben)} + Q_G(\delta^2/\sigma_h^2, \sigma_h^2/\delta, \frac{1}{25}(k - \frac{1}{2})) - d.$$

In Model 7, we allowed variance in  $h$  of beneficial mutations to differ from the variance in  $h$  of deleterious mutations by a factor  $\alpha^2$ . This was accomplished by using a modified version of the equation for the dominance of beneficial mutations

$$h_{ben,k}[\mu_{h(ben)}, \sigma_h^2, \delta] = \mu_{h(ben)} + \alpha(Q_G(\delta^2/\sigma_h^2, \sigma_h^2/\delta, \frac{1}{25}(k - \frac{1}{2})) - d).$$

Finally, we built upon Model 7 to ask whether the variance in  $h$  for the deleterious mutations depends on the strength of selection. To do so, we modified the equation for the dominance of deleterious mutations by using the parameter  $\tau$  as follows

$$h_{del,k}[s_j, \beta_1, \beta_2, \sigma_h^2, \delta] = \mu_{h(del),j} + (1 + \tau s_j)(Q_G(\delta^2/\sigma_h^2, \sigma_h^2/\delta, \frac{1}{25}(k - \frac{1}{2})) - d).$$

Because we have used discrete approximations, the realized moments of the  $h_k$  values are not exactly equal to the moments of the continuous distributions from which they were generated, though they are reasonably close. When we report means and variances of the  $h$  distribution, we calculate these from the realized  $h_k$  values; when we are referring to the parameters of the model from which these were derived, we specify those parameters, i.e.,  $\beta_1$ ,  $\beta_2$ ,  $\sigma_h^2$ , and  $\delta$ .

**Finding maximum likelihood values:** Because of the number of parameters involved, we used a multi-step approach to help ensure that we were finding global maximum likelihood values rather than local maxima. For each model, we began with a random set of parameter values, then using the optimization function *optim* in *R* (R Development Core Team 2009), we performed repeated optimizations using the Nelder-Mead algorithm. This algorithm is fairly good at moving out of local maxima. When no further improvement could be gained using the Nelder-Mead algorithm, we then performed a final optimization using the quasi-Newton method, BFGS. For each model, we repeated this entire



process from 100 different initial random starting points. Typically, the best 30-40 results for each model were very similar, suggesting convergence on the global optimum.

### **Trimming the data**

As described in the main text, the data were filtered in an attempt to eliminate dubious values. In the final stage of filtering, genes were sorted into deciles with respect to homozygous fitness (averaged across both replicates). Within each decile, genes in the top 2.5% and bottom 2.5% with respect to heterozygous fitness were eliminated. The figure below shows heterozygous fitness plotted against homozygous fitness (in both cases, averaged across replicates). Points in red were eliminated from data set by the criteria of this final filter. The dashed line is the 1 to 1 line. As intended, this filter eliminated a number of points for which heterozygotes appeared to be strongly deleterious yet their corresponding homozygotes appeared to be nearly wild-type.

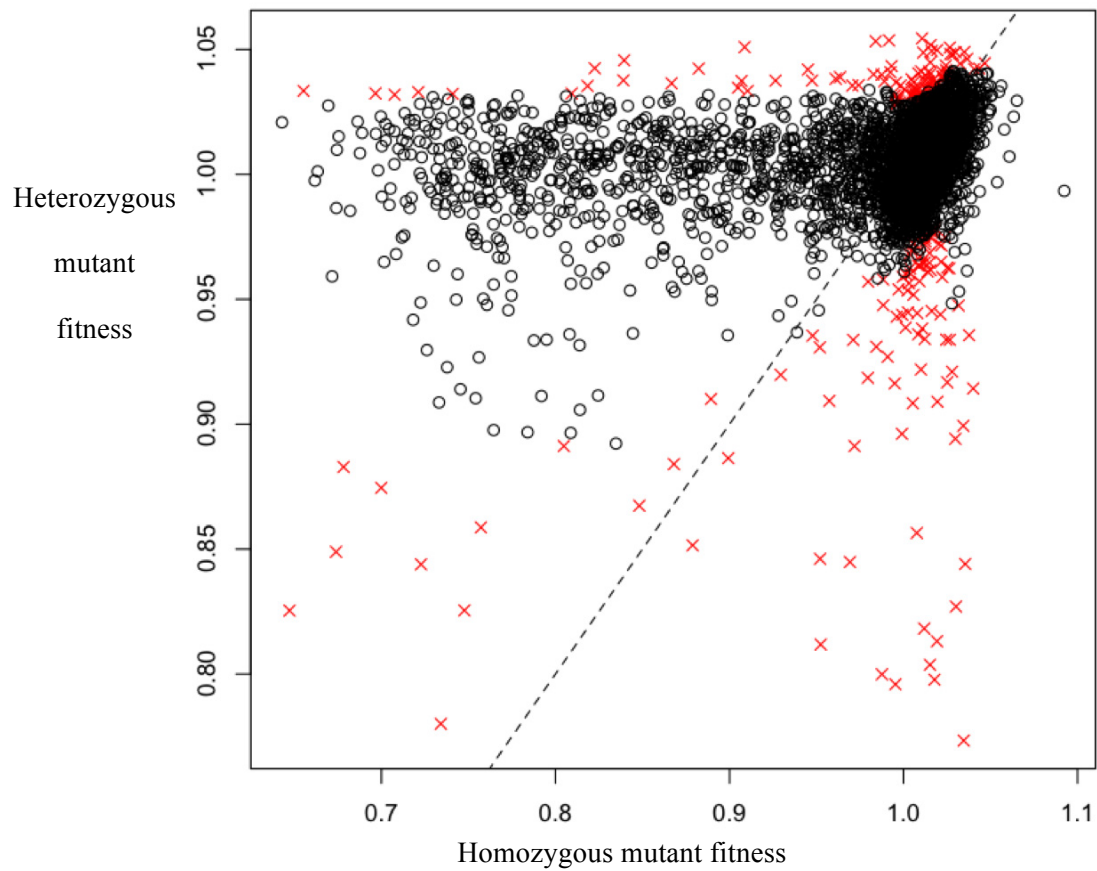


FIGURE S1.—Average heterozygous fitness is plotted against average homozygous fitness. Genes for which the discrepancy between replicates was much higher than expected based on Gaussian measurement error (i.e., the first data filter) are not shown. Points shown in red represent genes that were excluded by our second filter (i.e., removal of top 2.5% and bottom 2.5% with respect to heterozygous fitness within each decile of homozygous fitness).