Title: Mutational load and the functional fraction of the human genome

Authors: Benjamin Galeota-Sprung[1,a], Paul Sniegowski[1], Warren Ewens[1]

[1] University of Pennsylvania

[a] Corresponding author

Word count: 6630

Sept 30, 2019

Abstract: This paper considers the whole-genome mutational load in humans caused by the occurrence of deleterious mutations. It is claimed that the load is less, often far less, than has been recently calculated and, given realistic genetic parameter values, is well within the capacity of the human population. This implies that the proportion of the human genome that is functional can be larger than that which has recently been claimed possible.

Keywords: mutational load; functional proportion of the human genome; population genetics

Introduction

The total proportion of the human genome that is functional has been a question of intense interest, as has the same question for other organisms. It has long been known that genome size does not bear a close a relationship with apparent complexity: for example, the lungfish genome is 60 times larger than that of the human genome, and there is a three-order-of-magnitude range of genome sizes in angiosperms. This is the C-value paradox (Thomas 1971; reviewed in Gregory 2005).

A natural definition of "functional", at the organismal level, is "selected for", which implies the possibility of deleterious mutation (Graur 2013). A variety of evolutionary studies examining divergence from related organisms (reviewed in Ponting and Hardison 2011) suggest that 3-15% of the human genome is subject to purifying selection, with methods that account for rapidly evolving yet still constrained sequences (e.g. Meader *et al.* 2010) falling on the higher end of this range. A more recent estimate is ~8% (Rands *et al* 2014), and methods that utilize intraspecies in addition to interspecies variation also fall within the aforementioned range (Gulko *et al.* 2015, Ward and Kellis 2012).

The ENCODE project was a large-scale effort to systematically map regions of transcription, transcription factor association, chromatin structure and histone modification in the human genome. Regions assigned to any of these mappings were considered by the ENCODE authors to be functional, leading to a total estimate of 80% of the human genome as functional (Dunham *et al.* (ENCODE) 2012). This estimate is roughly an order of magnitude greater than divergence- and polymorphism-based estimates. The discordance between the ENCODE estimate and those of other studies, together with ENCODE's expansive definition

of functionality—one seemingly divorced from an evolutionary approach—led to criticism (Doolittle 2013; Graur *et al.* 2013) of the ENCODE estimate. Indeed, such a high fraction of functionality would be difficult to reconcile with the fact that one half to two thirds of the human genome consists of inactivated transposable elements (International Human Genome Sequencing Consortium 2001, de Koning *et al.* 2011). Nor does a high estimate for the proportion of the human genome that is functional help to resolve the C-value paradox (Doolittle 2013).

Graur (2017) used an argument from population genetics theory which cast further doubt upon the ENCODE estimate of the proportion of the genome that is functional. Using what is known about the human per-base mutation rate and a certain concept of mutational load, he posited a low upper limit for this proportion of the genome potentially subject to deleterious mutation—with likely parameter values, Graur's approach establishes a limit almost certainly no higher than ~5%.

Here we present a different approach. We argue that the influence of both the rate and the fitness effects of deleterious mutations upon the variance in fitness in a stably persisting population composed of real individuals should be considered. Using this argument we suggest that the upper limit for the proportion of the human genome that could plausibly be functional is higher, probably far higher, than the Graur estimate.

Theoretical development

The mutational load

There are various definitions of a genetic load "L" in the literature. Perhaps the most frequently used definition is

$$L = \frac{w_{\max} - \bar{w}}{w_{\max}}. \tag{1}$$

Agrawal and Whitlock (2012, their page 116) and Lesecque *et al.* (2012, their Equation (1)) use this load definition. Here $\bar{w}$ is the mean population fitness and $w_{\max}$ is the fitness of individuals of the most fit possible genotype. The load defined in this way is a quantification of the proportion by which mean fitness is depressed by some population genetic process. While there are many kinds of loads; in this case we are concerned with the load due to mutation. Although Graur gives the load definition (1), we will see later that he, wisely in our view, does not use it. We will show that Graur is in effect concerned with the quantity

$$L = \frac{w_{\max}}{\bar{w}}, \tag{2}$$

that is, the factor by which the fitness of individuals of the most fit possible genotype is greater than the mean fitness.

Fitness models and parameter values

We begin by formalizing the model of Graur (2017) as follows. We define a functional site as a site at which a deleterious mutation is possible. These are the only sites that we (and Graur) consider. For a single functional site, we define A as the normal nucleotide at this site and M as the mutant nucleotide that imparts a reduced fitness when homozygous. We

assume a mutation rate of u from A to M and that, relative to a fitness of 1 for the AA genotype, the fitness of the MM genotype is $1-s$, where $s > 0$. u is identical to Graur's μ_{del}, but for typographical convenience we use the notation u. The value of u is very small and following common practice we and Graur ignore terms of order u^2. Again following common practice we and Graur ignore the exceedingly small effect of back mutations from M to A.

Empirical estimates of human mutation rates include all mutations that occur. However, not all mutations, even at a functional site, are necessarily deleterious. Graur defines p as the probability that a mutation is deleterious, and we use v to denote the empirically estimated rate of mutation per base pair in the human genome. Then $u = vp$ is the probability of a deleterious mutation at a functional site.

Assumptions Although we start our analysis by considering one site only in the genome, it is necessary to make some assumptions in carrying this analysis to the whole-genome level. Here we make the same assumptions as those made by Graur (2017), namely that all functional sites share the same values for u and s described above, that there is a multiplicative fitness relationship among functional sites, and that there is no linkage disequilibrium between functional sites.

Notation and numerical values. Apart from the notation for the parameters v, u, p and s defined above, we define g as the number of base pairs in the diploid human genome and f as the proportion of sites in the genome that are functional. This implies that n, the number of base-pair functional sites in the genome, is $\frac{1}{2}fg$, with the factor $\frac{1}{2}$ accounting for the fact that a site in these models is a diploid site. We consider only these n functional sites. Graur assumes that $g = 6.114 \times 10^9$. We use this value for comparing our calculations

with those given in his Table 1, and otherwise will adopt the slightly different recent estimate $g = 6.4 \times 10^9$ (Schneider *et al.* (2017)). Graur (2017) claims that 0.4 is a reasonable value for p, the probability that a mutation at any functional site is deleterious, and we also adopt this value. Lesecque *et al.* (2012) give an average value of v of 1.1×10^{-8} derived from three studies. Besenbacher *et al.* (2014) estimate a range 1.27×10^{-8} to 1.5×10^{-9} for v and Milholland *et al.* (2016) give an estimate around 1.2×10^{-8}, which is similar to that in other recent studies. We adopt the value $v = 1.2 \times 10^{-8}$ which, given $p = 0.4$, implies that $u = 4.8 \times 10^{-9}$ or about 5×10^{-9}, a value considered by Graur.

It is harder to choose a value for s, first because published estimates of s rely on indirect methods with large resulting standard errors, and second because the value of s differs from site to site, probably substantially. Lesecque *et al.* (2012) recognize the site-to-site variation in s and illustrate their analysis by considering the cases $s = 10^{-3}, s = 10^{-2}$ and $s = 10^{-1}$. This corresponds to the Eyre-Walker *et al.* (2006) and Eyre-Walker and Keightley (2007) estimate that in humans most functional sites have a value of s in the range 10^{-3} to 10^{-1}. We choose an average value $s = 10^{-2}$ for some of our calculations.

We will later assume a reproducing generation population size of $N = 10^9$, an extremely conservative (that is, large) value for size of the human population in all but very recent times, and the one least favorable to our arguments. To summarize, we will sometimes use the following parameter values in our calculations:

$$g = 6.4 \times 10^9, \quad u = vp = 5 \times 10^{-9}, \quad s = 10^{-2}, N = 10^9. \tag{3}$$

The quantity nus will enter often into our calculations below. With the parameter values in

(3) and with $f = 0.05$, a value we consider below, $nus = 0.008$.

The additive case

We first consider the additive fitness, or no dominance, case, since other authors appear to focus on this case. For any given functional site the relative, or proportional, genotype fitnesses are of the following additive form:

$$\begin{array}{llll} \text{Genotype} & AA & AM & MM \\ \text{Fitness} & 1 & 1-\frac{1}{2}s & 1-s. \end{array} \tag{4}$$

The equilibrium frequency of A at each such site is $1 - \frac{2u}{s}$ so that population mean fitness at each site is the well-known value $1 - 2u$. (Graur models these fitnesses as survival probabilities, a consideration that we set aside for now, but address in section 4.4.)

Graur notes that historically the size of the human population did not change for tens of thousands of years, suggesting that the population mean fitness $\bar{w}$ must be approximately 1. Haldane (1937) made the same suggestion, which we follow by multiplying the fitnesses by a common factor such that the mean fitness is 1. This normalization leads to the following well-known genotype fitnesses and frequencies:

$$\begin{array}{llll} \text{Genotype} & AA & AM & MM \\ \text{Fitness} & (1-2u)^{-1} & (1-2u)^{-1}(1-\frac{1}{2}s) & (1-2u)^{-1}(1-s) \\ \text{Frequency} & (1-\frac{2u}{s})^2 & \frac{4u}{s}(1-\frac{2u}{s}) & (\frac{2u}{s})^2. \end{array} \tag{5}$$

Under the multiplicative fitness assumption and with these fitnesses and frequencies the whole-genome mean absolute fitness (*i.e* the expected number of progeny for a given geno-

type) is 1. Before discussing this further we consider the argument made by Graur for finding a limiting value for the proportion of the genome that is functional.

The Graur argument: a limit for f

The values in (5) show that the maximum fitness possible, that of an "optimal" individual who is AA at all n functional sites in the genome, is $w_{\max} = (1-2u)^{-n} \approx e^{2nu}$. Since the mean fitness is 1, this value is a measure of the load $w_{\max}/\bar{w}$ as defined in (2):

$$L = \frac{w_{\max}}{\bar{w}} = e^{2nu} \tag{6}$$

We confirm this calculation with two made by Graur (2017) for the case $f = 0.05$. Graur assumes that $g = 6.114 \times 10^9$ so that $n = \frac{1}{2}fg = 0.15284 \times 10^9$. For the case $u = 5 \times 10^{-9}$, a value considered by Graur, $e^{2nu} = e^{1.5285} = 4.6$, in agreement with Graur's tabled value. When $f = 0.05$, $g = 6.114 \times 10^9$ and $u = 5 \times 10^{-9}$, $e^{2nu} = 98$, in close agreement with Graur's tabled value 100, minor differences being introduced by the approximation to exponentiation. (Note that Graur's value for the maximum possible number of functional sites appears to omit the factor $\frac{1}{2}$, but his definition $1-u$ of the single-site mean fitness $1-2u$ does not include the factor 2, so that these differences cancel and we arrive at the same result.)

Graur argues that the load as defined in (6) cannot be cannot be higher than some reasonable value for humans. For reasons which we later explore (section 4.4), he interprets this quantity as the mean fertility, that is, the average number of offspring per adult conditional upon survival to reproduction. He sets this maximum value at 1.8, based on historical data, which corresponds to 3.6 offspring per mating pair. This limits $2nu$ to about 0.6. Since $n = \frac{1}{2}gf$

this sets a limit on the value of f. With the values g and u given in (3) this limiting value for f is about 0.01875. Graur's overall conclusion that 10-15% is an upper limit for f is therefore extremely conservative; as the calculation given here shows, using the most likely parameter values, his approach yields a limit of ~2% for f.

A revised analysis

As shown above, the fitness of an individual who is AA at all n functional sites is $(1-2u)^{-n} \approx e^{2nu}$. The main point of this paper is that no individual of this genotype will ever exist, so the load calculation (6) is not empirically relevant. The probability that a randomly chosen individual is AA at all n functional sites exists is $(1-\frac{2u}{s})^{2n} \approx e^{-4nu/s}$. With the parameter values in (3) and with $f = 0.05$ this is about e^{-320}. Similar calculations arise with other plausible parameter values. No individual who is AA at anything approaching n sites will ever appear in practice. But Graur's (2017) load computations for the additive case are (implicitly) based on the fitness of this idealized individual.

This point is not new. It was made by Wright (1977, page 481) in his discussion of Haldane's (1957) evolutionary, or substitutional, load concept, which was also based on a non-existent "optimal" individual. Wright states that "if many loci are involved, the genotype that combines the [optimal] genotypes at all loci is in general so rare theoretically that neither it nor anything approaching it exists in a finite population". Dobzhansky (1957), in response to Muller (1950), noted that for both flies and humans, "perhaps [individuals with no deleterious mutations] would be a superfly and a superman, but the fact is that such have never existed on earth." Mayr (1970) makes the same point in stating that "the whole approach [to Haldane-based load calculations] is misleading. It is based on a set of assumptions that have no real validity, primarily that of the existence of an optimal homozygous genotype".

Lesecque *et al.* (2012) describe an individual of optimal homozygous genotype as an "idealized " individual and Agrawal and Whitlock (2012) state that such an individual is unlikely to exist. Charlesworth (2013) states that "the mean fitness of a population relative to the fitness of a hypothetical optimal genotype that has a very low chance of being present in the population is essentially irrelevant". Henn *et al.* (2015) describe calculations involving $w_{\max}$ as idealized, and refer to the need to find the fitness of the most fit individual likely to exist in real populations. We find this fitness value below.

The Wright-Mayr view, and that of the authors cited above, is the one adopted here. Wright also noted that Crow's (1970) definition of load is flexible in that it relates to the"fitness requirement of actually or theoretically available [geno]types". At the whole-genome level, "actually available" concerns real populations and "theoretically available" concerns idealized populations. We believe that the appropriate choice is "actually available", which is the Wright-Mayr viewpoint.

With these considerations in mind, our aim in this work is to calculate the whole-genome distribution of fitnesses and establish a methodology for defining an upper limit to f based on the fitnesses of individuals who are likely to actually exist in a real population. To do this we employ the single-site model of (5) and adopt the simplifying assumptions described above. This requires some preliminary calculations.

We first find the distribution of the whole-genome fitness W of a randomly chosen individual when site fitness and frequency values are as given in (5). The mean and variance of W can be found from multinomial distribution formulae. Given the values in (5), the mean of W

is, exactly,

$$(1-2u)^{-n}[(1-\frac{2u}{s})^2+\frac{4u}{s}(1-\frac{2u}{s})(1-\frac{1}{2}s)+(\frac{2u}{s})^2(1-s)]^n \tag{7}$$

and the variance of W is, exactly,

$$(1-2u)^{-2n}[(1-\frac{2u}{s})^2+\frac{4u}{s}(1-\frac{2u}{s})(1-\frac{1}{2}s)^2+(\frac{2u}{s})^2(1-s)^2]^n-1. \tag{8}$$

The mean of W is $(1-2u)^{-n}(1-2u)^n=1$, as expected. If terms of order u^2 are ignored, the variance of W is

$$\text{Variance of } W \approx e^{nus}-1. \tag{9}$$

These calculations allow us to make two points. First, the idealized most fit possible individual has fitness of $(1-2u)^{-n}\approx e^{2nu}$, as calculated above, and with the parameter values in (3) this is approximately 4.95. If $s=0.01$ the standard deviation in fitness is found from (9) to be about 0.09, so that the fitness of this individual is about 44 standard deviations above the mean. Such an individual will never exist in the population, and the same conclusion is reached for other plausible parameter values. This is the Wright-Mayr argument.

Second, the fitness of an individual who is AA at x sites, AM at y sites and MM at z sites is $(1-2u)^{-n}(1-\frac{1}{2}s)^y(1-s)^z$. This fitness does not have a normal distribution. However the logarithm of this fitness, namely $\log w=-n\log(1-2u)+y\log(1-\frac{1}{2}s)+z\log(1-s)$ can be taken as having a normal distribution since for all practical purposes both y and z approximately have a normal distribution. (All logarithms are natural.) Therefore, to a sufficiently accurate approximation, the fitness W of an individual taken at random has a lognormal distribution.

The mean μ and the variance σ^2 of $\log W$ can be found from the known mean 1 of W and

known variance of W given in (9). These give

$$e^{\mu+\frac{1}{2}\sigma^2} = 1, \quad [e^{2\mu+\sigma^2}][e^{\sigma^2} - 1] = e^{nus} - 1.$$

From this,

$$\sigma^2 = nus \text{ and } \mu = -\frac{1}{2}nus. \tag{10}$$

With the parameter values in (3) the mean of $\log W$ is about -0.004 and the variance of $\log W$ is about 0.008.

The above calculation can be confirmed by that of Lesecque *et al.* (2012), who use different notation and a slightly different variance formula. Lesecque *et al.* consider the case $nu = 10$, $\frac{1}{2}s = 0.01$ and calculate the probability that W takes a value between 0.5 and 2 to be 97%. This is the probability that $\log W$ takes a value between $-\log 2 = -.69315$ and and $+\log 2 = +.69315$. With the values $nu = 10$ and $\frac{1}{2}s = 0.01$, $\log W$ has mean -0.05 and variance 0.1. The probability that $\log W$ takes a value between $-\log 2$ and and $+\log 2$ is found from normal distribution charts to be 97%, in agreement with the calculation made by Lesecque *et al.* (2012).

The fitness of the most fit individual likely to exist

We now calculate the fitness of the most fit individual who is likely to exist in a reproducing population of a plausible size. The mean number $n(\frac{2u}{s})^2$ of MM sites in a randomly chosen individual is of order u^2 and thus is very small, so that MM individuals can be ignored in the calculations. Lesecque *et al.* (2012) also do this. The focus is therefore on AA and AM individuals.

Ignoring terms of order u^2, the mean number of functional sites in a randomly-chosen AA individual is $n - \frac{4nu}{s} = n - m$ and the mean number of functional sites in a randomly-chosen AM individual is m, where $m = \frac{4u}{s}$. We assume that the most fit individual likely to appear in the population is AA at $n - m + r$ functional sites and is AM at $m - r$ functional sites, where the value of r has to be determined. To calculate r we assume that the number of functional sites at which a randomly chosen individual is AM has a Poisson distribution with mean m. (Lesecque *et al.* (2012) also make this Poisson distribution assumption.) This assumption implies that the standard deviation of this number of sites is $\sqrt{m}$. As in the recessive case we very conservatively assume a population of reproducing individuals of size 10^9. Using the normal approximation to the Poisson, the most fit individual who is likely to appear in a reproducing population of size 10^9 will therefore be AA at about $n - m + 5\sqrt{m}$ sites and AM at about $m - 5\sqrt{m}$ sites. Thus $r = 5\sqrt{m} = 5\sqrt{4nu/s}$.

It follows that when the mean population fitness is 1 the fitness requirement for the most fit individual who is likely to appear in a population of size 10^9 is

$$(1 - 2u)^{-n} \times (1 - \frac{1}{2}s)^{\frac{4nu}{s} - 5\sqrt{4nu/s}}. \tag{11}$$

Since $(1 - 2u)^{-n} \approx e^{2nu}$, $(1 - \frac{1}{2}s)^{\frac{4nu}{s}} \approx e^{-2nu}$ and $(1 - \frac{1}{2}s)^{-5\sqrt{4nu/s}} \approx e^{5\sqrt{nus}}$, this is approximately

$$e^{5\sqrt{nus}}. \tag{12}$$

This is the fitness requirement that Henn *et al.* (2015) require in the additive case. With the numerical values in (3) and with $f = 0.05$ the expression in (12) is about 1.17, sharply lower than the 4.95 that results from the use of (6). When $s \leq 10^{-2}$ similarly small values

arise for other plausible parameter values. As discussed below, with the numerical values in (3) a value of f as high as 0.25 or even higher is quite possible.

On pure viability selection

It might be objected that in our development so far, we have not addressed the features of fitness of Graur's model. In order to achieve a population mean absolute fitness of 1, Graur does not normalize the relative fitnesses in (4) as we have done to create (5). Instead, Graur in effect assumes that fitness is the product of viability, the probability of survival to reproduction, and fertility, the expected number of offspring produced conditional upon survival to reproductive age. Because the population mean absolute fitness is 1, the mean fertility must be the reciprocal of the mean viability. The fitnesses in (4) are interpreted as viabilities, and since these are probabilities, they cannot be normalized.

As noted by Gillespie (1998), the fitness of a genotype in nature must have many components such as viability, fertility, developmental time, mating success, and others. These components are understood to be incorporated into relative fitnesses such as those given in (4). In assuming that selection operates only on survival probabilities, Graur is in effect employing a model of "pure viability selection" (Nagylaki 1991). One objection to this particular model for fitness is that in the quantity $\frac{1}{2}gfpu$, every deleterious mutation per individual per generation (on average) is accounted for. It seems strange that the genetics of a natural population would be constructed such that all deleterious mutations affect viability, and none affect fertility. In reality, it must be the case that deleterious mutations affect both in unknown ratios.

Another objection to the pure viability selection model has to do with the assignment of

viabilities to a single site. To say that the fitness at a single site is 1 when the genotype is AA and $1 - s$ when the genotype is MM is really to say that the fitness of the MM genotype is reduced by the factor $1 - s$ when compared to the AA genotype. That is, the fitnesses are relative. When these quantities are instead treated as absolute viabilities, they are considered to be probabilities, and cannot then be rescaled. But it is then rather unclear what it really means for an individual to have a viability of 1, or $1 - s$, or any number, at a single site in a genome.

One could remove the grounds for the preceding objection by allowing the single site viabilities to be relative, such that, for example, the value $1 - s$ means very simply that viability is reduced by the factor $1 - s$. In that case, allowing that the maximum whole-genome viability should be set to 1, the main thrust of our argument applies: the maximum viability should be assigned to the fittest genotype that is likely to actually appear in a finite population. Then the viability of an individual who is AA at all sites would be greater than 1, a situation that may appear paradoxical because a probability cannot be greater than 1, but the evident resolution of this paradox is that such an individual will never exist.

The recessive case

We next discuss the recessive case, in which the single-site fitnesses differ from (4) in that the fitness of the heterozygote is 1. As in the additive case, we normalize so that the mean fitness is 1. This leads to the following fitness table.

Genotype	AA	AM	MM
Fitness	$(1-u)^{-1}$	$(1-u)^{-1}$	$(1-u)^{-1}(1-s)$
Frequency	$(1-\sqrt{\frac{u}{s}})^2$	$2\sqrt{\frac{u}{s}}(1-\sqrt{\frac{u}{s}})$.	$\frac{u}{s}$

(13)

We next calculate realistic whole-genome fitnesses making the same simplifying assumptions as those made for the additive case.

The whole-genome load and the distribution of fitness

We define W as the fitness of an individual taken at random. With site fitnesses as given in (13) and the various assumptions made above, the whole-genome mean of W is 1 and the the whole-genome variance of W is found from (13) and multinomial distribution formulae to be

$$(1-u)^{-2n}[1-\frac{u}{s}+\frac{u}{s}(1-s)^2]^n-1 \approx e^{nus}-1. \tag{14}$$

This formula allows us to make two points. The first is that, with the parameters in (3) and $f = 0.05$, the variance $e^{nus}-1$ of W is about 0.008 so that the standard deviation of W is about 0.09. The fitness of an individual who is either AA or AM at all functional sites is $(1-u)^{-n}$ or about $e^{1.08} \approx 2.94$, about 22 standard deviations above the mean. As stated above for the additive case, such an individual will never exist. The same conclusion holds for other plausible parameter values.

The second point requires more calculation. The fitness W of a randomly-chosen individual who is MM at x sites is $(1-u)^{-n}(1-s)^x$. Thus the fitness of this individual does not have a normal distribution. However to a close approximation $\log W = -n\log(1-u)+x\log(1-s)$

has a normal distribution since x has approximately a normal distribution. Therefore, to a close approximation, W has a lognormal distribution. The mean μ and variance σ^2 of $\log W$ can be found from the known mean (1) and known variance $(e^{nus} - 1)$ of W using standard formulas relating parameters in a normal distribution and the parameters in the corresponding lognormal distribution. It is found that

$$e^{\mu+\frac{1}{2}\sigma^2} = 1, \quad [e^{2\mu+\sigma^2}] \times [e^{\sigma^2} - 1] = e^{nus} - 1. \tag{15}$$

From this, $\sigma^2 = nus$ and then $\mu = -nus/2$.

With the parameter values in (3) and with $f = 0.05$ we obtain $\mu \approx -0.004$ and $\sigma^2 \approx 0.008$. The probability that W takes a value between 0.7 and 1.3 is the same as the probability that $\log W$ takes a value between -0.357 and +0.262, which from normal distribution charts is about 0.9985. Thus only about 0.15 % of individuals have a fitness outside the range 0.7 to 1.3. Calculations using other plausible parameter values lead to the same conclusion, namely that the variance of the fitness W is small and that the great majority of individuals have an easily achieved fitness. This conclusion once again quantifies the Wright-Mayr argument.

The fitness of the most fit individual likely to exist

We now find the fitness of the most fit individual likely to appear in the population. Since there are n functional sites and the probability that an individual is MM at a given functional site is u/s, the mean number of MM sites in a randomly chosen individual is $k = nu/s$. We assume that the actual number of MM sites carried by a randomly chosen individual has a Poisson distribution with parameter k. The standard deviation of this distribution is $\sqrt{k}$.

This distribution can be approximated by a normal distribution with mean k and standard deviation $\sqrt{k}$. The properties of the statistics of extreme values of normal random variables shows that in a population of 10^9 reproducing individuals, the individual with the smallest number of MM sites will have about $k - 5\sqrt{k}$ such sites, or $5\sqrt{k}$ fewer than the mean (of k). It follows that this individual is AA or AM at $5\sqrt{k}$ sites more than the mean number $n - k$ of these sites. From (13) the fitness of an individual having this number of AA or AM sites is

$$(1-u)^{-n}(1-s)^{k-5\sqrt{k}}. \tag{16}$$

This is approximately $e^{nu} \times e^{-sk} \times e^{5s\sqrt{k}}$. Since $k = \frac{nu}{s}, e^{-sk} = e^{-nu}$ and $e^{5s\sqrt{k}} = e^{5\sqrt{nus}}$, the expression in (16) becomes

$$e^{5\sqrt{nus}}. \tag{17}$$

This is the formula that Henn *et al.* (2015) claimed should be calculated for the recessive case. Note that this is exactly the same formula as in the additive case, so that the calculation and comments below (12) apply for the recessive case also.

In the calculations above we have in effect assumed that in the recessive case that the mean number of MM sites in a randomly chosen individual is $\frac{nu}{s}$. This calculation ignores stochastic effects in a population of finite size. Lesecque *et al.* (2012) show that when these effects are taken into account a slightly more accurate expression for this mean is $\frac{nu\pi}{4s}$. This does not materially change the main points of our analysis.

Parameter values and other considerations

The value of f. As the title of his paper indicates, Graur (2017) is most interested in the effect of f on the mutational load. He arrives at the view that f must be small by calculating the load as, in effect, e^{gfpu}, often arriving at extremely large load calculations when $f > 0.1$. As noted above, Graur assumes that $g = 6.114 \times 10^9$ and writes our vp as μ_{del}, and presents a table (his Table 1) of the value of the load for various combinations of values of f and μ_{del}. These can be very high. For example when $p = 0.4$ and $v = 1.25 \times 10^{-8}$, so that $\mu_{\text{del}} = 5 \times 10^{-9}$, the load as calculated by his formula is 22 when $f = 0.1$ and 100 when $f = 0.15$. Interpreted as the fitnesses of a mating pair, these values would imply 44 and 200 expected offspring, both obviously impossible.

We now calculate the load, defined equally in (12) and (17) for the recessive and the additive cases as the fitness requirement of the most fit individual likely to appear in the population. With the values in of g, u and s given in (3), so that $nus = 0.16f$, this load is $e^{5\sqrt{0.16f}}$. For $f = 0.1$ this load is 1.88 and for $f = 0.15$ it is 2.17. These loads are easily achieved. For $f = 0.2$ and $f = 0.25$ the loads are 2.44 and 2.72 respectively, again easily achieved, even assuming significant pre-reproductive mortality. Even higher values of f, such as $f = 0.5$, which leads to a fitness of the most fit individual of 4.11, are not impossible. However the values of the parameters p and s are not known with precision, so that these estimates must be taken with much caution. We now discuss these parameter values.

The value of s. The calculations in the preceding paragraph assume the arbitrary value $s = 10^{-2}$. The site-to-site distribution of s has been discussed by various authors. As a result of their estimation procedures Boyko *et al.* (2008) consider values of s ranging from

0.45×10^{-3} to 10^{-2}. As stated above, Eyre-Walker *el al.* (2006) and Eyre-Walker and Keightley (2007) claim that in humans most functional sites have a value of s in the range 10^{-3} to 10^{-1}. The conclusions that we reach about possible values of f hold for all values of $s \leq 10^{-2}$. For $s = 10^{-1}$ the values of (12) and (17) are larger than they are when $s = 10^{-2}$. When the values of g and v are as given in (3), the loads (12) and (17) with $s = 10^{-1}$ is $e^{5\sqrt{1.6f}}$. This is 7.38 when $f = 0.1$ and 11.6 when $f = 0.15$. These are large loads. However it is very unlikely that $s = 10^{-1}$ for all sites. If $s = 10^{-3}$ for one third of the sites, 10^{-2} for one third and 10^{-1} for one third, the generalization of the loads (12) and (17) when $f = 0.1$ is an easily achieved value. As stated above, since the parameters p and s are not known with precision, calculations of plausible values of f must be taken with some caution. In all cases, however, the plausible values are larger, sometimes much larger, than those calculated by Graur.

Agrawal and Whitlock (2012) define the load as in (1), so that for them the load in the additive case is $L = 1 - e^{-2nu}$, and make two comments about this formula. First, they state that the fact that this load formula is independent of s has led to a "misleading sentiment" among theoretical population geneticists who then feel that nothing need be known about the value of s or about ecological considerations in assessing loads. We agree, and believe that load calculations that ignore s are not realistic and have plagued population genetics theory for far too long. Second, they state that there is very little empirical evidence that Haldane's (1957) load theory, based on the formula $L = 1 - e^{-2nu}$, is even approximately correct. The theme of this paper is that the reason for this lack of empirical evidence is that Haldane's theory, being based on this formula, is not empirically relevant.

The value of N. Our choice $N = 10^9$ is intended to be extremely conservative. Of all the pa-

rameters involved, the value chosen for N in a population that has subdivided and increased substantially in size over hundreds of thousands of years is possibly the most problematic. All the calculations above, and many in the literature, assume random mating. Henn *et al.* (2015) consider in detail the fact that for many thousands of years random mating is an unreasonable assumption, given the division of the human population into different sub-groups, based largely on geographical dispersion. This dispersion also bears on the reasonable choice for N. Henn *et al.* also consider complications due to the effects of population-size bottlenecks and the "prodigious rate" of growth in the size of the human population, increasing from a few hundred thousand about 13,000 years ago to several millions four thousand years ago. The conclusions that we reach about possible values of f continue to hold, and are strengthened by, any reasonable choice for the various values of the human population size over the last two hundred thousand years.

Other fitness models

When h is positive a fitness model generalizing (5) is

$$\begin{array}{llll} \text{Genotype} & AA & AM & MM \\ \text{Fitness} & (1-2u)^{-1} & (1-2u)^{-1}(1-hs) & (1-2u)^{-1}(1-s) \\ \text{Frequency} & (1-\frac{u}{hs})^2 & \frac{2u}{hs}(1-\frac{u}{hs}) & (\frac{u}{hs})^2, \end{array} \tag{18}$$

The case $h = \frac{1}{2}$ corresponds to the additive model. The "Haldane load" $(1-2u)^{-n}$ is independent of h. It might then be expected that the realistic load generalizing (12) is also independent of the value of h, but this is not so. It is found after some algebra that with fitness and frequency values as given in (18), the mean of w is 1 but the the variance of w is

no longer as given in (9) but is, instead, $e^{2nuhs} - 1$. From this the mean μ and variance σ^2 of $\log w$ are no longer as given in (9) but are, instead,

$$\mu = -nuhs \text{ and } \sigma^2 = 2nuhs. \tag{19}$$

Henn *et al.* (2015) state that the average value of h in (18) across various non-human organisms is about 0.25. If the average value $h = 0.25$ applies also in humans, these amended values strengthen the conclusions reached above for the additive case when $s \leq 10^{-2}$ and slightly amend those for the case $s = 10^{-1}$.

Stochastic effects.

Two individuals with the same fitness do not necessarily have the same number of offspring: stochastic effects have to be taken into account. Lesecque *et al.* (2012) quantify this by discussing three models: an asexual model, a monogamous diploid model, a freely interbreeding diploid model, and for each model calculate the probability $P(0)$ that an individual has (in the monogamous diploid case, a couple have) no offspring. They consider the effect of the value of nu on $P(0)$ for a given value of s and produce very interesting graphs describing this effect (their Figure 4). Our interest is the effect of s on $P(0)$ for a given value of nu. The graphs in Figure 4 of Lesecque *et al.* (2012) show that value of $P(0)$ increases very slowly with s in all three cases. The number of offspring for an individual is determined more by stochastic effects than by the individual's fitness. The reason for this is the fact that the variance in fitness as given for example in (10) is very small.

Other kinds of loads

Load-based arguments seeking to limit the value of f need not remain limited to the mutational load. The substitutional load and the segregational loads also depend to some extent upon f and might be considered as well. The criticism of load arguments by Wright, Mayr and others referred to above were made with respect to one or both of these loads. If these loads were taken into account as well as the mutational load the possible values of f would be even smaller. However these load calculations are subject to the same criticisms that we have made for the mutational load.

Epistasis

The assumption of no (multiplicative) epistasis implied in the whole genome fitness model discussed above has received some attention in the literature. Agrawal and Whitlock (2012) state that empirical evidence for epistasis is mixed and provide references discussing evidence for both synergistic and non-syneregistic cases. From a theoretical point of view, as far back as 1966 Kimura and Maruyama (1966) discussed a model where the fitness of an individual having i deleterious genes in his/her genome is of the form $1 - h_1 i - h_2 i^2$ for some non-negative constants h_1 and h_2. This is a quadratic fitness model. This fitness can become negative if i is large enough, so Kimura and Maruyama assumed a zero fitness for $i \geq n$, where n is the smallest value of i for which $1 - h_1 i - h_2 i^2$ becomes negative. For the parameter values that they chose, individuals having a large enough number of deleterious genes so that $1 - h_1 i - h_2 i^2$ becomes negative are so rare that Kimura and Mauryama neglected them. This agrees with our main point: for realistic load calculations, individuals having genotypes that will never arise in practice can be ignored.

Quite apart from this, the very high loads calculated by Graur (2017) assume a completely

multiplicative fitness model and no linkage disequilibrium throughout the functional part of the genome. Models that relax the multiplicative assumption for extreme genotypes are plausible and can lead to a large decrease in the load. This idea is not new. For example, in the case of the segregational load, Sved, Reed and Bodmer (1967) state that under a multiplicative fitness model, individuals with fitness more than four standard deviations above the mean occur very rarely, and that replacing their fitness by a truncated value four standard deviations above the mean and compensating by a very minor upward adjustment in all fitnesses resolves the segregational load problem. The same argument can be made for the mutational load.

On the applicability of our results to organisms other than humans

Our argument, generalized, is that any load calculation involving the fitnesses of individuals unlikely to actually exist in a stably persisting human population is flawed. For example, consider a statement such as the following, of a type sometimes encountered in the literature: the number of *de novo* deleterious mutations per human individual (nu in our notation) cannot possibly be as high as x because e^{-x} is an impossibly low mean fitness. Implicit in the previous statement is the idea that $w_{\max} = 1$, and it is really the ratio $w_{\max}/\bar{w} = 1/e^{-x}$ that is found to be troubling. Yet if an individual possessing the genotype that would actually produce fitness equal to 1 will never exist, this quantity is not meaningful, in our view. However, the relation $w_{\max}/\bar{w} = e^{nu}$ is meaningful when individuals with the maximum theoretically possible fitness do actually exist. This can reasonably be expected when $nu << 1$, a condition that is likely the case for many microbes. For example, for *E. coli*, estimates of nu are on the order of 10^{-4} (Kibota and Lynch 1996). Assuming that $s = 0.01$, this suggests that ~99% of such a population would have no deleterious mutations. Hence our conclusions

should be applied with caution to species with parameter values very different from humans.

Summary and Conclusions

Summary

The per-nucleotide rate of mutation and the total size of the human genome appear to be fairly well established. The fraction f of the human genome that is functional, and hence the average number of *de novo* deleterious mutations per individual, remains uncertain. We have shown that when considering the likely maximum realized fitness in a finite population, the limit to f is by no means low. This result stands in contrast to arguments that depend upon the fitness of an individual who possesses the theoretical maxiumim fitness of the particular model employed. Such arguments appear to establish a low limit for f, but suffer from the flaw that such a human is only vanishingly likely to exist. Calculations that purport to establish a load should, in our view, be based on the distribution of actual fitnesses that are expected to exist in a real population. As we have shown, the properties of this distribution depend not just on nu (the number of *de novo* deleterious mutations per individual) but also on s, the selection coefficient against deleterious mutations.

Using the approach of Graur (2017) and adopting the most plausible value for the human per-base-pair deleterious mutation rate, the limit to f is ~2%. We have shown that when considering the likely maximum realized fitness in a finite, persisting human population, much higher values for f (as high as 0.25, and perhaps as high as 0.5 or even higher, with considerable uncertainty introduced by the parameter s) are plausible.

We stress that we, in this work, take no position on the actual proportion of the human genome that is likely to be functional. It may indeed be quite low, as the contemporary evidence from species divergence and intraspecies polymorphism data suggests. Many of the criticisms of the ENCODE claim of 80% functionality (e.g. Doolittle 2013, Graur 2013) strike us as well-founded. We wish only to point out that an argument from mutational load does not appear to be particularly limiting on f.

Conclusion

The main aim of is paper is to discuss reasonable estimates for the maximum possible proportion f of the human genome that is functional. The estimation procedure depends on various parameter values, some of which are not known accurately. Despite this, for many realistic parameter values, the possible values for f are higher, probably substantially higher, than have been estimated by what we consider to be inappropriate Haldane load calculations.

Literature cited

Agrawal, A., Whitlock M. 2012. Mutation load: the fitness of individuals in populations where deleterious alleles are abundant. *Ann. Rev. Ecol. Evol.* 43: 115-135.

Besenbacher S. *et al.* 2014. Novel variation and *de novo* mutation rates in population-wide *de novo* assembled Danish trios. *Nature Communications* DOI: 10.1038/ncomms6969.

Boyko, A. *et al.* 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4: e1000083.

Charlesworth, B. 2013. Why are we not dead one hundred times over? *Evolution* 67-11: 3354-3361.

Crow, J. 1970. Genetic loads and the cost of natural selection. In *Mathematical topics in population genetics*, ed. K. Kojima. Springer-Verlag.

Doolittle, W. F. 2013. Is junk DNA bunk? A critique of ENCODE. *PNAS* 110: 5294–5300.

Eyre-Walker, A., Woolfit, M., Phelps, T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173: 891-900.

Eyre-Walker, A., Keightley, P. 2007. The distribution of fitness effects of new mutations. *Nature Reviews* 8: 610-619.

ENCODE Project consortium 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57-74.

Graur, D. *et al.* 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution* 5(3):578-590.

Graur, D. 2017. An upper limit on the functional fraction of the human genome. *Genome Biology and Evolution* 9: 1880-1885.

Gurven, M. and Kaplan, H. 2007. Longevity among hunter-gatherers: a cross-cultural examination. *Population and Development Review.* 33:321-365.

International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822): 860-921.

de Koning, A. J. *et al* 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLOS Genetics*, 7(12), e1002384.

Gillespie, J. 1998. Population genetics: a concise guide. The Johns Hopkins University Press.

Gregory, T. R. 2005. Synergy between Sequence and Size in Large-Scale Genomics. *Nature Reviews Genetics* vol. 6:699–708.

Gulko, B. *et al* 2015. A method for calculating probabilities of fitness consequences for point

mutations across the human genome. *Nature genetics* 47(3): 276.

Haldane, J. 1937. The effect of variation on fitness. *Amer, Nat.* 71: 337-349.

Haldane, J. 1957. The cost of natural selection. *J. Genet.* 55:511-524.

Henn, B. *et al.* 2015. Estimating the mutation load in human genomes. *Nature Reviews: Genetics* 16: 333-343.

Kimura M., Maruyama T. 1966. The mutational load with epistatic gene interactions in fitness. *Genetics* 54: 1337-1351.

Lesecque Y., Keightley P., Eyre-Walker A. 2012. A resolution of the mutation laod in humans. *Genetics* 191: 1321-1330.

Mayr, E. 1970. *Populations, Species and Evolution.* Harvard University Press.

Meader, S. *et al* 2010. Massive turnover of functional sequence in human and other mammalian genomes. *Genome research* 20(10): 1335-1343.

Milholland B. *et al.* 2014. Differences between germline and somatic mutation rates in humans and mice. *Nature Communications*DOI: 10.1038/ncomms15183.

Nagylaki T. 1992. Introduction to theoretical population genetics. Springer-Verlag.

Ponting, C. and Hardison, R. 2011. What fraction of the human genome is functional? *Genome research* 21: 1769-1776.

Rands, C. *et al.* 2014. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLOS Genetics*, 10(7), e1004525.

Schneider V. *et al.* 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research* 27.5: 849-864.

Simons, Y., Turchin M., Pritchard J., Sella G. 2014. The deleterious mutation load is insensitive to recent population history. *Nature Genetics* 45: 220-226.

Sved, J., Reed T., Bodmer W. 1967. The number of balanced polymorphisms that can be maintained in a natural population. *Genetics* 55: 469-481.

Thomas, C.A. Jr. 1971. The genetic organization of chromosomes. *Annu. Rev. Genet.* 5: 237–256.

Ward, L.D. and Kellis, M., 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337:1675-1678.

Wright, S. 1977. *Evolution and the Genetics of Populations 3. Experimental results and evolutionary deductions.* University of Chicago Press.

650