THE WILHELMINE E. KEY 2012 INVITATIONAL LECTURE

Background Selection 20 Years on

BRIAN CHARLESWORTH

From the Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Ashworth Laboratories, King’s Buildings, West Mains Road, Edinburgh EH9 3JT, UK

Address correspondence to Professor Brian Charlesworth at the address above, or e-mail: Brian.Charlesworth@ed.ac.uk

Brian Charlesworth is a Senior Honorary Professiorial Fellow in the Institute of Evolutionary Biology at the University of Edinburgh. He obtained his PhD in genetics at Cambridge in 1969, and was a postdoctoral fellow with Richard Lewontin at the University of Chicago (1969–1971). He subsequently worked at the Universities of Liverpool, Sussex, and Chicago, moving to Edinburgh as a Royal Society Research Professor (1997–2007). He is a Fellow of the Royal Society, Fellow of the Royal Society of Edinburgh, and Honorary Fellow of the American Academy of Arts and Sciences. He was awarded the Darwin Medal of the Royal Society in 2000. His current research interests are in population genetics, molecular evolution, and genome evolution. He has published over 300 papers and three books (two co-authored with Deborah Charlesworth).

Abstract

The mutation process continually produces new deleterious variants at sites throughout the genome, which are then mostly eliminated by selection. This causes a reduction in variability at linked neutral or nearly neutral sites, as well as distortions of the genealogies of samples of alleles from a population. In regions of the genome where recombination is frequent, the effects of selection against deleterious mutations on variability and evolution at linked sites can be predicted under the assumption that most deleterious mutations have such large effects that their behavior in the population is effectively deterministic—this is background selection in the strict sense. But in genomic regions with little or no recombination, such as the Y chromosome, large departures from the predictions using deterministic models may occur, because of interference between different sites under selection. Evidence from Drosophila and human populations is discussed, which suggests that these processes play a major role in shaping patterns of DNA sequence variation and evolution, including the relative levels of variation on X chromosomes and autosomes, and the highly reduced variability seen in regions that lack crossing over.

Key words: background selection, genetic drift, Hill-Robertson interference, mutations, natural selection

It has long been known that most new mutations with phenotypic effects are deleterious. Current evidence on mutation rates and mutational effects on fitness suggests that higher organisms like humans and Drosophila receive an average of one or more new, deleterious mutations per individual per generation, most of which are destined for elimination from the population by selection (Drake et al. 1998; Lynch 2010; Keightley 2012). This continual input of deleterious mutations has many important consequences. Here, I will focus on just one of these—the process that has come to be known as “background selection,” abbreviated as BGS (Charlesworth et al. 1993). The basic idea was originally described by Fisher (1930), when discussing the fate of a new, beneficial mutation in an asexual population. Only the portion of a non-recombining population that currently carries the smallest number of deleterious mutations contributes to the ancestry of future generations, because (in the absence of back mutation or recombination) the
The Properties of Deleterious Mutations

This section provides a sketch of current knowledge of the frequency of occurrence of deleterious mutations, and the sizes of their effects on fitness—see Keightley (2012) for a more comprehensive survey. A crucial parameter is the per-genome deleterious mutation rate, \( U \). This is defined as the mean number of mutations that appear per generation in a newly formed zygote, and that reduce the fitness of their carriers sufficiently to ensure their eventual elimination from the population. Comparisons of sets of homologous DNA sequence drawn from a pair of related species provide an estimate of the proportion of mutations that are sufficiently deleterious that they are certain to be eliminated from the population by selection (Kondrashov and Crow 1993). For selectively neutral sequences, the between-species divergence per neutral nucleotide site, \( K_{\text{neu}} \), is proportional to the product of the mutation rate and the time separating the two sets of sequences. Using the divergence value, \( K_{\text{neu}} \), for sequences that are subject to purifying selection, such as nonsynonymous sites or functional noncoding sequences, the “level of selective constraint”, \( c \), can be estimated as \( 1 - K_{\text{sel}}/K_{\text{neu}} \) (Kondrashov and Crow 1993; Keightley 2012). This measure is an underestimate of the true fraction of mutations that are deleterious, because some slightly deleterious mutations can be fixed by drift, and some mutations are fixed by positive selection.

Mutation rates in model organisms can be estimated from DNA sequence alterations in lines in which mutations have accumulated under conditions when selection is weak or absent; in humans, comparisons of whole genome sequences of trios of parents and an offspring can be carried...
out, allowing the direct detection of new mutations (Drake et al. 1998; Lynch 2010; Keightley 2012; Kong et al. 2012). By multiplying the total mutation rate by the average value of \( c \) across the genome (weighting by the number of sites in each category), the deleterious mutation rate \( U \) can be estimated. The current estimates of \( U \) for \( D. \) melanogaster and humans are 1.2 and 2.1 per diploid zygote per generation, respectively (Keightley 2012).

In order to make predictions about BGS, we also need estimates of the frequency distribution of selection coefficients against newly arising deleterious mutations. Such estimates are becoming available from data on the frequencies within a population of variants at nucleotide sites subject to purifying selection (Charlesworth 2012a; Keightley 2012). On the assumption that most autosomal mutations in randomly mating populations are selected against when carried as heterozygotes with wild type, these methods estimate the distribution of the reduction in fitness of the heterozygous carriers of a new mutation, measured relative to the fitness of the wild-type homozygote; this reduction is denoted by \( t \).

The results for nonsynonymous mutations in humans, mice, plants, and Drosophila suggest that the distribution of \( t \) is very wide, so that gamma or log-normal distributions generally provide a much better fit than a normal distribution (reviewed by Charlesworth 2012a; Keightley 2012). For example, in \( Drosophila \) pseudoobscura the coefficient of variation of \( t \) exceeds one, and the mean \( t \) for mutations segregating in the population is very small, such that its product with the effective population size, \( N_e \), is of the order of 20 (Haddrill et al. 2010). Given that silent site diversity values indicate an \( N_e \) of around a million for this species, this implies a corresponding mean \( t \) of about \( 10^{-3} \); the mean \( t \) for new mutations is harder to estimate accurately and is likely to be at least an order of magnitude higher, because the most strongly selected mutations are rapidly removed from the population (Haddrill et al. 2010). In Drosophila, it seems clear that there is only a small proportion (5% or less) of new nonsynonymous mutations that behave as effectively neutral \((N_{te} < 0.5)\). Approximately 90% are so strongly selected \((N_{te} > 5)\) that they have a negligible chance of fixation by drift, which means that their dynamics are effectively the same as in an infinite population. In humans, the proportion of effectively neutral mutations is closer to 25%, reflecting their much smaller \( N_e \) (Eyre-Walker and Keightley 2009).

These results imply that the mean number of deleterious nonsynonymous mutations that are carried by a typical individual is approximately 5,000 for Drosophila (Haddrill et al. 2010) and 800 for humans (Charlesworth and Charlesworth 2010, p. 295). Accordingly, the estimated mean number of deleterious, nonsynonymous mutations present in a randomly chosen gene sampled from a population is surprisingly large for organisms like Drosophila with a high \( N_e \)—around 0.18 per gene in \( D. \) pseudoobscura (Haddrill et al. 2010).

The continual input of deleterious mutations and their elimination by selection can thus affect variation and evolution at nearby sites in the genome. Furthermore, the relatively strong selection against new nonsynonymous mutations (and a portion of mutations in noncoding sequences) means that this process can be approximately described using deterministic models of mutation and selection—at least in freely recombining parts of the genome. The next part of this paper describes the current status of our theoretical understanding of BGS; this is followed by a discussion of the relevance of the theory to data on patterns of DNA sequence variability in Drosophila and humans.

**The Theory of BGS**

**Introduction**

The possible types of effects of selection against deleterious mutations on linked sites span a continuum, divided up according to the effectiveness of selection against deleterious mutations relative to genetic drift (Charlesworth et al. 2010). At one extreme, sites are subject to strong purifying selection, with \( N_{te} \gg 1 \), where \( N_{te} \) is the effective population size in the absence of selection. This means that their average frequencies are close to their equilibrium values under mutation and selection (Nordborg et al. 1996)—BGS in the strict sense. This ignores any complications due to genetic drift and non-random associations between linked sites and also assumes that mutations affect fitness independently of each other.

For non-recessive, autosomal mutations affecting a strongly selected nonsynonymous or noncoding site, the equilibrium frequency \( q_i \) of the mutant variant at the \( i \)-th site in question in a randomly mating population is determined by the ratio of the mutation rate to the deleterious variant, \( u_i \), and the heterozygous selection coefficient against the variant, \( t_i \); that is, \( q_i = n_i / t_i \) (Haldane 1927). With a dominance coefficient \( h_i \) the effective selection coefficient against the mutation is \( s_i = h_i q_i \) (providing that \( h_i > 0 \) as is likely to be the case for most mutations (Crow 1993)), where \( q_i \) is the selection coefficient against homozygotes. For X-linked mutations with equal selection on the 2 sexes, the corresponding effective selection coefficient against a mutation is \( s_i = 2h_i + 1 \) (Charlesworth and Charlesworth 2010, pp. 160–161).

A basic question concerns the effect of BGS on the mean coalescence time at a given neutral site, \( T_{2p} \), for a pair of alleles sampled from the population, that is, the mean time it takes for a pair of homologous sequences to trace their ancestry back to a common ancestral sequence (Hudson 1990). Under the standard assumptions of coalescent theory for a panmictic population with constant effective population size \( N_{pe} \) in the absence of BGS, we have \( T_{2p} = 2N_{pe} \). The mean pairwise nucleotide site diversity (\( \pi \)) for neutral mutations under the infinite sites model commonly used in the interpretation of data on DNA sequence variability is directly proportional to \( T_{2p} \) (Hudson 1990), so that the value of \( T_{2p} \) tells us about the effect of BGS on the expected level of DNA sequence variability. The ratio of \( T_{2p} \) under BGS to \( 2N_{pe} \), denoted by \( B \), is a useful measure of the effect of BGS on variability.
The Effect of BGS on Neutral Diversity in a Recombining Genomic Region

For sites with \( N_0 t_i >> 1 \) that are distributed along a single recombining chromosome or chromosomal region, the following formula (Nordborg et al. 1996) provides a good approximation for \( B \) for a given neutral site:

\[
B = \exp \left( -\sum_i \frac{n_i t_i}{t_i + \epsilon_i \left[ 1 - t_i \right]^2} \right) \quad (1)
\]

Here, \( r_i \) is the frequency of recombination between the focal neutral site under consideration and the \( i \)th site subject to purifying selection; \( n_i \) and \( t_i \) were defined above.

Equation (1), or the related formula of Hudson and Kaplan (1995), can be used to obtain an overview of the effect of BGS on a neutral site in a defined region of the genome such as a single chromosome, assuming that BGS effects on this site from mutations elsewhere in the genome are negligible (Hudson and Kaplan 1995; Nordborg Charlesworth and Charlesworth 1996).

For example, assume for simplicity that the sites subject to mutation and selection are distributed uniformly across the region, with the same effective selection coefficient \( t \) at each site, and a diploid deleterious mutation rate \( U \) for the region. The map length of the region is \( M \) Morgans, and the rate of recombination is linearly related to map distance (implying that there is a negligible frequency of double crossovers), with a constant rate of occurrence of recombination events across the regions. Assume also that a proportion \( x \) of the region is located to the left of the focal site, and \( 1 - x \) to the right. The following expression is then obtained by approximating the sum in equation (1) by integration (Nordborg et al. 1996):

\[
B = \exp \left( -\frac{U (t + 2xM[1 - t])}{2(t + \infty M[1 - t])(t + [1 - x]M[1 - t])} \right) \quad (2)
\]

If there is a probability distribution of \( t \) values, of the type discussed above, we can integrate the term in the exponent over the distribution over the portion of the distribution for which \( t \) is sufficiently large that the deterministic approximation holds, with only mutations that meet this condition being considered (Nordborg et al. 1996; Charlesworth 2012b).

But further useful approximations can be obtained that remove the need to specify this distribution. Provided that \( t << M \), so that recombination is more powerful than selection, and provided that the focal site is not too close to the end of the region, so that \( x(1 - x)M >> t \), equation (2) becomes:

\[
B = \exp \left( -\frac{U}{2M} \right) \quad (3)
\]

That is, the effect of BGS on the mean coalescent time in this case depends only on the density of deleterious mutations per unit map length, as pointed out by Hudson and Kaplan (1994) and Barton (1995).

For a site at either end of the region \((x = 0 \) or \( 1)\), this approach gives the following:

\[
B = \exp \left( -\frac{U}{2M} \right) \quad (4)
\]

The effect of BGS in this case is thus smaller than for a focal site in the middle of the region, by a factor of \( \exp (-U/2M) \) (Nordborg et al. 1996).

BGS is thus most effective in the middle of a recombining region, and least effective at either end of it, as one would intuitively expect from the fact that the density of sites under selection that are close to the focal site is greatest in the center of the region. Equations (3) and (4) provide a useful rough guide to the expected overall effects of BGS on coalescent time, which are independent of the distribution of selection coefficients. Some applications of these results to the interpretation of data are described later.

The Effect of BGS in a Non-recombining Genomic Region

The case of a genome region with no recombination will now be considered; this is relevant to Y and W chromosomes in species with chromosomal sex determination, and to the “dot” chromosome of Drosophila (Ashburner et al. 2005, Chapter 17), as well as to asexual species or highly self-fertilizing species where recombination is effectively absent throughout the genome (Charlesworth et al. 1993). In the complete absence of recombination, equation (1) yields the expression:

\[
B = \exp \left( -\frac{U}{2\hat{t}} \right) \quad (5)
\]

Here, \( \hat{t} \) is the harmonic mean of the \( t_i \) and \( B \) is the first term of a Poisson distribution whose mean is equal to the average number of mutations per haploid genome, \( U/(2\hat{t}) \), at equilibrium under mutation and selection (Charlesworth et al. 1993), that is, \( B \) is the equilibrium frequency of mutation-free haplotypes, \( f_0 \). The intuitive basis for this formula was described in the introductory section.

This is, however, merely an approximation, which implicitly assumes that the coalescence of a pair of alleles sampled from the population can take place only within the section of the population that is free of mutations, as well as assuming that all sites under selection are held at mutation-selection equilibrium. Equations describing the more general case, when coalescence can also occur in haplotypes carrying 1 or more mutations, were developed by Hudson and Kaplan (1994), using a “structured coalescent” model, in which differences in coalescence probabilities among classes of haplotypes carrying different numbers of deleterious mutations are taken into account (Figure 2). The principle used is that coalescence of a pair of alleles can only occur when they are both in the same class with respect to the number of deleterious mutations, just as two alleles in a subdivided population can only coalesce if they are in the same local population (Charlesworth and Charlesworth 2010, Chapter 7).
The relevant equations determine $T_2$, which approaches the value given by equation (5) when selection is sufficiently strong (Hudson and Kaplan 1994; Gordo et al. 2002). When selection is sufficiently weak, however, the reduction in $T_2$ caused by BGS is overestimated by equations (1) and (5), as is observed in simulations with and without recombination (Charlesworth et al. 1993, 1995; Nordborg et al. 1996; Zeng and Charlesworth 2010, 2011).

Effects of BGS on the Shape of Gene Genealogies

As mentioned in the introductory section, the reduction in pairwise coalescent time caused by BGS, described by equations (1–5), can usefully be regarded as equivalent to a reduction in effective population size, such that $N_e = BN_0$. However, this is only part of the picture. It was early recognized that BGS also distorts the shape of gene genealogies compared with the standard neutral coalescent processes, causing a greater relative contribution from terminal branches of the gene tree for a set of alleles sampled from a population than would be expected from a standard coalescent process with $N_e = BN_0$ and a corresponding excess of rare variants; this effect was found to be greatest when selection is weak (Charlesworth et al. 1993; Hudson and Kaplan 1994; Charlesworth et al. 1995). The number of neutral sites in a sample that are segregating for polymorphic variants (5) is, therefore, not reduced by BGS to the same extent as the pairwise diversity, $\pi$, resulting in negative expected values of the frequently used Tajima’s $D$ statistic, although the effect is small when selection is strong (Charlesworth et al. 1993; Hudson and Kaplan 1994; Charlesworth et al. 1995; Tachida 2000; Gordo et al. 2002; Williamson and Orive 2002; O’Fallon et al. 2010; Seger et al. 2010; Walczak et al. 2012).

For the case of complete linkage, when the mutation-selection balance approximation applies, these effects can be predicted approximately from the dependence of the rate of coalescence on the age $\tau$ of a pair of alleles that are ancestral to a set of alleles sampled from a population (Nicolaisen and Desai 2012), where $\tau$ is measured backwards from the time of sampling (Figure 2). Assuming a fixed selection coefficient $s$, and a mutation rate that is sufficiently low that at most only a single mutation occurs per haploid genome per generation, the mean coalescent time for ancestral alleles of age $\tau$ decreases with $\tau$. This dependence can be represented using equation (5), multiplying $U/(2\tau)$ by $[1 - \exp(-\tau\tau)]^2$. This allows the complete probability distribution of coalescent times over a gene genealogy to be calculated, as well as statistics derived from it such as $S$ and Tajima’s $D$ (Nicolaisen and Desai 2012). $T_2$ approaches the value corresponding to equation (5) as $\tau$ tends to infinity. With strong selection, this value is approached very quickly, and hence dominates the entire process, so that the genealogy for a sample of $k$ alleles is close to that for the standard neutral coalescent process with $N_e = j_k N_0$ (the top box in Figure 2). With weak selection, the longer coalescent times associated with small $\tau$ contribute more to the overall picture, because the deleterious mutations persist longer in the population, distorting the genealogy in favor of longer external branches than expected for the standard coalescent. A semi-analytic approach to this problem for the case when $N_j$ is around 1, so that the deterministic assumption is invalid, has been developed by O’Fallon et al. (2010).

Similar effects of selection on the gene genealogy should also occur with recombination, but its effect has been little studied analytically, except for Santiago and Caballero (1998). A structured coalescent procedure for modeling BGS that incorporates recombination and variation in selection coefficients across sites has recently been developed, which shows that some degree of distortion of the genealogy and inaccuracy of equation (1) occurs even with recombination (Zeng and Charlesworth 2011; Zeng 2012). This approach enables the development of tests of significance for the fit to data of the predictions of BGS models, and can also allow for changes in population size.
The Effects of Stochastic Departures from Mutation-Selection Equilibrium

When there is little or no recombination, the condition $N_d h_i > 1$ for mutation-selection balance is necessary but not sufficient. If $h_i N_d$ is sufficiently small, deleterious mutations are fixed in the population by genetic drift at a roughly constant rate, which is larger than with recombination; nonetheless, nearly all sites under selection remain close to their equilibria under mutation and selection if $N_d h_i > 1$, so that back mutations from mutant to wild type can be ignored, at least in the initial stages of the process before many mutations become fixed. This is the process of Muller’s ratchet (Muller 1964); it is called a ratchet because of its property of a steady rate of loss from the population of the “least-loaded” haplotype that carries the current minimum number of deleterious mutations (Felsenstein 1974). This process represents the second part of the continuum of the behavior of mutation-selection models, and has been intensively studied theoretically (see Stephan and Kim 2002 and Charlesworth 2012a) for reviews.

With a fixed selection coefficient $t$, the ratchet moves at an evolutionarily significant rate when $h_i^* N_d$ is of the order of 30 or less (Gordo et al. 2002; Jain 2008). Its effects on neutral variability over much of the parameter space turn out to be quite accurately represented by the structured coalescent model, despite the violation of the mutation-selection balance assumption (Gordo et al. 2002). For a given mutation rate, as $t$ is decreased while holding $N_d$ fixed (or vice versa), both forward simulations and structured coalescent calculations show that $B$ reaches a minimal value, which coincides with the condition for the ratchet to operate; $B$ then starts to increase again, reaching unity when $t = 0$. The minimum coincides approximately with a maximal distortion of neutral genealogies, but the degree of distortion stays fairly flat until $t$ approaches zero (Gordo et al. 2002).

The ratchet model assumes one-way mutation from good to bad. This assumption works reasonably well for the initial process of departure from the deterministic equilibrium under mutation and selection, when the wild-type variant is much more frequent than the mutant at each site and any reverse mutations can be neglected, and is also well-suited to the representation of irreversible mutations such as insertions and deletions (Dolgin and Charlesworth 2006; Kaiser and Charlesworth 2010). For nucleotide substitutions, however, back mutations become increasingly frequent as deleterious mutations become fixed by the ratchet. Ultimately, the fixation of deleterious mutations is balanced by the fixation of reverse mutations, with a steady flux of mutations in both directions. Unless the population has gone extinct because of the reduced mean fitness associated with the accumulation of deleterious mutations, a new equilibrium will eventually be established (Wagner and Gabriel 1990; Gordo and Charlesworth 2001; Kaiser and Charlesworth 2009; Charlesworth et al. 2010).

When both the condition for the ratchet to move is met for a region with little or no recombination, and $N_d h_i$ is of order 1 or less, deleterious mutations at nonsynonymous or functionally significant noncoding sites can reach high frequencies or fixation, as a result of genetic drift, with a much higher probability than when there is free recombination. This represents the third part of the continuum of behavior—“weak selection Hill–Robertson interference” (WSHR) (McVean and Charlesworth 2000; Comeron and Kreitman 2002; Comeron et al. 2008). This term refers to the fact that, if many sites under selection are located in a non-recombinating genome or genomic region, Hill–Robertson interference (Hill and Robertson 1966; Felsenstein 1974) among them weakens the effective strength of selection that acts on each individual site, to such an extent that they are subject to effects of drift that are of the same order as the effects of selection.

Simulations with realistic distributions of selection coefficients show that this process is associated with an increased rate of substitution of variants at the sites under selection, and a reduction in DNA sequence variability at both selected and neutral sites located in the same genomic region (Kaiser and Charlesworth 2009; Charlesworth et al. 2010). The reduction in variability can be on the order of 10-fold in regions with 1 megabase or more of coding sequence, and is approximately 10-fold for a region of the size of the Drosophila dot chromosome (with about 80 kilobases of coding sequence), in agreement with data on sequence variability in large genomic regions that lack recombination (Figure 3). The distortion in gene genealogies at neutral sites embedded in the region, with its accompanying excess of rare variants compared with neutral expectation, is much larger than with BGS or the ratchet (Kaiser and Charlesworth 2009; Zeng and Charlesworth 2010).

The effects of adding more sites under selection into a non-recombinating region become weaker as the number of sites increases, and the level of neutral diversity approaches an asymptotic value (Figure 3), presumably because the interference among the selected sites increasingly undermines the effective strength of selection acting on them (Kaiser and Charlesworth 2009). These effects of WSHRI with large numbers of selected sites imply that organisms such as many bacteria, with low genome-wide levels of recombination, will have much lower effective population sizes (and hence levels of silent site variability and codon usage bias), than would be expected from their very large numbers of individuals (McVean and Charlesworth 2000).

Applications to Data on DNA Sequence Variability

Chromosome-wide Levels of Variability

The utility of the approximate description of the overall level of neutral variability in a specified genome region, provided by equations (2) and (3), can be illustrated with the examples of Drosophila and human chromosomes. The map length in females for an arm of one of the 2 major autosomes of D. melanogaster is about 0.5 Morgans (Ashburner et al. 2005, Chapter 10). Given that there is no crossing over in males, and that an autosomal gene spends half of its time in each sex, the effective map length ($M_e$) to be used in the equations is 0.25, because recombination rates in males and females are each given a weight of one-half (Charlesworth 2012b). Using a conservative estimate of 1.0 for the deleterious mutation
rate for \textit{D. melanogaster}, and the fact that a chromosome arm is on average about 20\% of the genome, the \textit{U} value for a chromosome arm is 0.20. Equation (2) then gives $B = 0.45$ for a single autosomal chromosome arm—a large reduction below 1. It cannot, therefore, be assumed that observed levels of variability at putatively neutral sites in regions of normal recombination in organisms like \textit{Drosophila} with compact genomes are free of the effects of selection at linked sites, as has also been argued for the effects of selective sweeps (Gillespie 2001; Weissman and Barton 2012).

In humans, there are 23 chromosomes with a sex-averaged mean map length per autosome of approximately 1.6 Morgans (Jensen-Seaman et al. 2004); \textit{U} is approximately 2.1 (see above), corresponding to a total value of 1.99 for the autosomes, given their relative contribution to the genome (Jensen-Seaman et al. 2004). This yields an estimate of $B = 0.94$ for a centrally located site on a typical autosome. This neglects, however, the substantial variation in physical size and map length among chromosomes (Jensen-Seaman et al. 2004), and so should be taken as a crude approximation; it is to be expected that the larger chromosomes, with their higher gene content and lower rates of crossing over, will have smaller \textit{B} values than this, and the smaller chromosomes will have higher values. Even larger differences among chromosomes are to be expected in birds and reptiles, with their extreme size and recombination rate differences between micro- and macro-chromosomes; there are indeed differences among bird chromosomes with respect to levels of variability, in the direction expected from hitchhiking effects (Huynh et al. 2010). Nonetheless, the prediction for humans is close to an estimate of an average reduction in neutral diversity on a chromosome of about 6\% from hitchhiking effects—obtained from an analysis of the relation between variability and proximity to functionally important sequences (Cai et al. 2009).

\textit{X} chromosomes in mammals spend two-thirds of their time in females (where they can recombine) and one-third of their time in males, where they do not recombine with most of the \textit{Y}, whereas autosomes can recombine in both sexes. This implies a stronger overall effect of \textit{BGS} on the \textit{X} chromosome in mammals. Similar effects are to be expected for the \textit{Z} chromosome in birds (Ellegren 2009). The physical size and map length of the \textit{X} (Jensen-Seaman et al. 2004) suggest that $B = 0.91$ for humans, about 97\% of the autosomal value.

The data on overall ratios of \textit{X} to autosomal within-population diversities in humans are somewhat conflicting. Ratios of 0.73 and 0.61 (after divergence corrections) were obtained from African and European whole genome resequencing data by Gottipatti et al. (2011); values greater than 0.75 and 0.71 for Africans and Europeans were obtained by Hammer et al. (2010) from smaller datasets. The ratios are significantly higher for sites close to genes than for sites remote from genes, strongly suggesting the action of hitchhiking effects caused by selection on coding or flanking regulatory sequences, as was also found in another study of 4 human populations (Hernandez et al. 2011). In Africans, but not Europeans, the \textit{X} to autosome diversity ratio is greater than 0.85 far from genes (Gottipatti et al. 2011). These differences between populations may reflect nonequilibrium demography in European populations, and differences in the variance in male mating success due to sexual competition, which is expected to inflate the \textit{X} to autosome diversity ratio (Charlesworth 2009; Evans 2013).

In \textit{Drosophila}, the lack of crossing over in males means that the \textit{X} has about 4/3 times the autosomal effective
recombination rate, for pairs of sites with comparable frequencies of recombination in females on the X and the autosomes (Charlesworth 2012b). In D. melanogaster, it also has a slightly longer map than an autosomal arm (0.6 Morgans). Equation (2) with $U = 0.20$ yields $B = 0.61$ for the X chromosome, giving an $X$/autosome ratio of $B$ of 1.36. The corresponding ratio of diversity values for an equilibrium population is $(3 \times 1.36) / 4 = 1.02$, in the absence of sexual competition effects or mutation rate differences between X and autosomes, for which there is little evidence in Drosophila (Bauer and Aquadro 1997; Keightley et al. 2009; Zeng and Charlesworth 2010; Campos et al. 2013). There should thus be similar or even slightly higher levels of silent site diversity for the X chromosome than the autosomes, as is true for the recombining regions of the genome in the presumptively ancestral East African populations of D. melanogaster (Andolfatto 2001; Hutter et al. 2007; Singh et al. 2007; Campos et al. 2013). Furthermore, in these populations, silent diversity on the X after correction for the relation between the recombination rate and diversity (see below) is about three-quarters of the corrected value for the autosomes (Vicoso and Charlesworth 2009; Campos et al. 2013); this strongly suggests that some form of hitchhiking effect is responsible for the elevated diversity on the X in these populations.

In contrast, a similar calculation for D. pseudoobscura, which has map lengths that are more than twice those in D. melanogaster (Kulathinal et al. 2008; Stevison and Noor 2010; McGaugh et al. 2012), yields a predicted X to autosome diversity ratio of 0.83. This is close to the values obtained from estimates of silent site variability at X-linked and autosomal loci in D. pseudoobscura and its relative D. miranda (Haddrill et al. 2010). BGS seems, therefore, to be capable of explaining patterns of chromosome-wide levels of variability in at least 2 Drosophila species.

It should be noted, however, that the assumption of a uniform density of sites subject to selection used to obtain these results is unrealistic, because selected sites are clustered into coding sequences and groups of functional noncoding sequences, separated by blocks of sequence (intronic and intergenic), portions of which are likely to be under weak or no selection, especially in organisms with large genomes such as mammals (Keightley 2012). More realistic models of a chromosome arm of Drosophila, with alternating weakly selected and strongly selected noncoding and coding sequences, have recently been analyzed (Charlesworth 2012b), which show that the X to autosome diversity ratio is in fact fairly insensitive to the presence of weakly selected intergenic sequences.

**Effects of Recombination on Diversity**

This broad-brush approach neglects the interaction between recombination and BGS. A good starting point for examining this empirically is provided by genomic regions of low or zero crossing over, which consistently show highly reduced levels of sequence variability in a variety of organisms (Frankham 2012). The small dot chromosome of Drosophila is classic material for this problem, because it represents an isolated region of the genome with an almost complete lack of crossing over in all species that have been studied (Ashburner et al. 2005, Chapter 17), and contains about 0.6% of the total coding sequence.

Using a conservative estimate of $U$ of 0.5 per diploid genome, and a $t_\text{DOT}$ (substantially higher than the estimate mentioned earlier), $f_\text{DOT} = \exp (-0.25 \times 0.06 \times 10^{-3}) = \exp (-15) = 3.1 \times 10^{-7}$. This predicts that no variability should be found on the dot chromosome in a sequence variability study of reasonable size. In fact, the dot is generally found to have about 10% as much variability as a typical region of the Drosophila neo-Y chromosome of D. miranda (Figure 3). The neo-Y has about 1500 functional genes (Bachtrog et al. 2008; Zhou and Bachtrog 2012) and exhibits a silent site diversity value that is about 1% of the genome-wide average (Bartolome and Charlesworth 2006). In addition, the alternative hypothesis of a recent selective sweep seems to be inconsistent with the pattern of sequence variation seen on the dot chromosome of Drosophila americana (Betancourt et al. 2009).

When there is a moderate amount of recombination, simulations show that equation (1) should work well as a description of the effect of BGS for a single chromosome with $N_\text{e} t > 3$ or so (Nordborg et al. 1996). The well-established positive correlation between the nucleotide site diversity of a locus in Drosophila and the local rate of genetic recombination that it experiences, as determined by its location on the chromosome (Begun and Aquadro 1992; Comeron et al. 2012; Langley et al. 2012; McGaugh et al. 2012), can be tested against the predictions for this equation, or the related equation of Hudson and Kaplan (1995). This has met with some success (Hudson and Kaplan 1995; Charlesworth 1996; Hamblin and Aquadro 1999).

However, the D. melanogaster material used by Hudson and Kaplan (1995) and Charlesworth (1996) came largely from European and North American samples, which are depauperate in variability compared with the ancestral East African populations, because of population bottlenecks associated with the movement of flies from Africa in association with humans (Begun and Aquadro 1993; Andolfatto 2001; Haddrill et al. 2005; Hutter et al. 2007). It remains to be seen if the more recent estimates of selection and mutation parameters discussed earlier, together with the use of genome-wide data from East African populations and high-density genetic maps, and from other species of Drosophila such as D. pseudoobscura (McGaugh et al. 2012), produce similarly good fits, and to what extent selective sweeps also contribute to these patterns (Sella et al. 2009; Stephan 2010); the processes are not, of course, mutually exclusive.
Attempts have also been made to investigate the fit of BGS models to patterns of human genetic diversity (Paysseur and Nachman 2002; Reed et al. 2005; Hellmann et al. 2008; Cai et al. 2009; McVicker et al. 2009). The results, together with those of Hammer et al. (2010), Hernandez et al. (2011) and Gottipatti et al. (2011), show that diversity at neutral or nearly neutral sites is reduced close to selectively constrained coding and non-coding sequences, where allele frequency distributions also deviate more from neutral expectation, and that diversity is significantly correlated with local recombination rates. Although BGS models apparently provide a good fit to the data, it remains unclear to what extent they provide a unique explanation, or whether the effects of selective sweeps also contribute (Hellmann et al. 2008; Cai et al. 2009); whole human genome resequencing data seem, however, to provide little evidence for frequent, recent selective sweeps, so that BGS may be the main cause of the effects seen in humans (Hernandez et al. 2011; Alves et al. 2012).

Conclusions

The results described above show that we now possess a solid body of theory that makes testable predictions about the effects of BGS and related processes such as weak selection Hill–Robertson interference. Some patterns of variability in human and Drosophila seem to be quantitatively consistent with the predictions of the models. There is reason to be optimistic that new genome-wide datasets on variability, estimates of mutation and selection parameters, and detailed recombination maps, all of which are becoming available from the application of next generation sequencing data, will considerably advance our ability to test the models much more rigorously, and to assess the respective contributions of BGS and selective sweeps to genome-wide patterns of sequence variation and evolution.

Funding

US National Science Foundation, the Royal Society, the Leverhulme Trust, the UK Biotechnology and Biological Sciences Research Council, the UK Natural Environment Research Council and the School of Biological Sciences, University of Edinburgh.

Acknowledgments

I am very grateful for their contributions to this work to Deborah Charlesworth, and the following past and present members of my research group: Peter Andolfatto, Andrea Betancourt, Carolina Bartolomé, Doris Bachtrg, José Campos, Elie Dolgin, Isabel Gordo, Penny Haddrill, Vera Kaiser, Lawrence Loewe, Xulio Maside, Gil McVean, Martin Morgan, Magnus Nordborg, Beatrix Vicoso, and Kai Zeng. I thank the American Genetic Association, especially its President, Mohamed Noor, for honoring me as the 2012 Wilhelmine Key Lecturer, and for inviting me to write this review based on my lecture.

References


Reed FA, Alcey JM, Aquadro CF. 2005. Fitting background-selection predictions to levels of nucleotide variation and divergence along the human autosomes. Genome Res. 15:1211–1221.


Revised November 29, 2012; Accepted December 4, 2012

Corresponding Editor: Mohamed Noor