# Genome-Wide Linkage-Disequilibrium Profiles from Single Individuals

**Michael Lynch,* Sen Xu,* Takahiro Maruki,* Xiaoqian Jiang,* Peter Pfaffelhuber,† and Bernhard Haubold‡**
*Department of Biology, Indiana University, Bloomington, Indiana 47401, †Faculty of Mathematics and Physics, University of Freiburg, Freiburg 79104, Germany, and ‡Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Plön 24306, Germany

**ABSTRACT** Although the analysis of linkage disequilibrium (LD) plays a central role in many areas of population genetics, the sampling variance of LD is known to be very large with high sensitivity to numbers of nucleotide sites and individuals sampled. Here we show that a genome-wide analysis of the distribution of heterozygous sites within a single diploid genome can yield highly informative patterns of LD as a function of physical distance. The proposed statistic, the correlation of zygosity, is closely related to the conventional population-level measure of LD, but is agnostic with respect to allele frequencies and hence likely less prone to outlier artifacts. Application of the method to several vertebrate species leads to the conclusion that >80% of recombination events are typically resolved by gene-conversion-like processes unaccompanied by crossovers, with the average lengths of conversion patches being on the order of one to several kilobases in length. Thus, contrary to common assumptions, the recombination rate between sites does not scale linearly with distance, often even up to distances of 100 kb. In addition, the amount of LD between sites separated by <200 bp is uniformly much greater than can be explained by the conventional neutral model, possibly because of the nonindependent origin of mutations within this spatial scale. These results raise questions about the application of conventional population-genetic interpretations to LD on short spatial scales and also about the use of spatial patterns of LD to infer demographic histories.

THE analysis of linkage disequilibrium (LD) plays a central role in many areas of population genetics, including the determination of genetic maps, ascertainment of levels of recombination at the population level, and estimation of effective population sizes. For populations in approximate drift–mutation–recombination equilibrium, the latter becomes possible with neutral markers because the expected levels of allelic association across loci can be expressed in terms of the population parameters $\theta = 4N_e u$ and $\rho = 4N_e r$, where $N_e$ is the effective population size, $u$ is the mutation rate per nucleotide site, and $r$ is the rate of recombination between sites (Ohta and Kimura 1969; Hill 1975). In principle, if an estimate of $r$ is available, the effective population size can be extracted from $\rho$ (*e.g.*, Hill 1981; Hayes *et al.* 2003; Tenesa *et al.* 2007), or vice versa. Unfortunately, the stochasticity of evolutionary processes generates enormous evolutionary variance for two-locus measures (Hill and Weir 1988), so a very large number of LD estimates from independent pairs of sites is required for meaningful inferences on $\rho$. Moreover, it has become increasingly clear that gene conversion causes a nonlinear relationship between $\rho$ and physical distances between sites (Andolfatto and Nordborg 1998), raising questions about the very meaning of $\rho$.

With methods for whole-genome sequencing now well established, obtaining observations on large numbers of sites is no longer problematic. Although extending LD analysis to large numbers of individuals is still challenging (Maruki and Lynch 2014), it has been suggested that even a single individual, if sampled from a randomly mating population, can provide insight into genome-wide disequilibrium patterns (Lynch 2008). The proposed parameter, $\Delta$, denoted as the "correlation of zygosity," is a measure of the degree to which the spatial distribution of heterozygous sites deviates from the expectation under random segregation. Here, we consider the issues underlying this parameter in detail, first describing the observable single-individual measure in terms of the traditional population-level estimates of LD, then exploring the power of the proposed method, and finally summarizing the results of applications


doi: 10.1534/genetics.114.166843
Manuscript received June 3, 2014; accepted for publication June 17, 2014; published Early Online June 19, 2014.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.166843/-/DC1.
Corresponding author: Department of Biology, Indiana University, Bloomington, IN 47401. E-mail: milynch@indiana.edu

to the sequenced genomes of several vertebrate species. The empirical observations provide new insights into features of recombination that have been previously accessible with only a small number of model genetic organisms.

## Theory

### *The correlation of zygosity, Δ*

Consider all pairs of genomic sites separated by distance $d$ within a single individual, and let $\pi$ and $(1-\pi)$, respectively, be the fractions of heterozygous and homozygous sites within the individual. Provided two sampled sites are uncorrelated by descent, denoted by $(1-\Delta)$ below, the probabilities of jointly homozygous and jointly heterozygous states are $(1-\pi)^2$ and $\pi^2$, respectively. If the genotypes of the two sites are correlated, denoted by $\Delta$, they are either both homozygous or both heterozygous with respective probabilities $(1-\pi)$ and $\pi$. After some simplification, it follows that the frequencies of jointly homozygous, jointly heterozygous, and mixed homozygous/heterozygous pairs of sites are respectively:

$$\begin{aligned} H_0 &= (1-\Delta)(1-\pi)^2 + \Delta(1-\pi) \\ &= (1-\pi)^2 + \Delta\pi(1-\pi), \end{aligned} \tag{1a}$$

$$H_2 = \pi^2 + \Delta\pi(1-\pi), \tag{1b}$$

$$H_1 = 2\pi(1-\pi)(1-\Delta). \tag{1c}$$

From Equation 1c, it can be seen that

$$\Delta = 1 - \frac{H_1}{2\pi(1-\pi)} \tag{2}$$

is a measure of the deviation of the frequency of pairs of loci with mixed zygosities from the random expectation. In general, $\Delta$ is expected to be positive because recombination causes variance in coalescence times among chromosomal regions. Nonrecombining haplotype blocks that happen to have deep coalescent times have a higher probability of carrying multiple derived mutations than do pairs with shallow coalescent times, and hence of being doubly homozygous for such mutations. The mathematical bounds to $\Delta$ are therefore generally expected to be zero to one, and in numerous applications we have never seen otherwise.

### *Relationship to population-genetic parameters*

A more mechanistic understanding of $\Delta$ requires expressions for the expectations of the three two-site zygosity classes ($H_0$, $H_1$, and $H_2$) in terms of the underlying population determinants. In the following, we adhere to the standard Wright–Fisher population-genetic model, with biallelic loci (individual nucleotide sites), random mating, and discrete generations with consecutive episodes of mutation, recombination, and gamete sampling. Mutation is assumed to occur at rate $u$ per generation at each site, and the probability that a recombination event is initiated between sites is denoted as $r$.

Consider two hypothetical sites, and denote the respective frequencies of alleles **A** and **a** at the first site as $p$ and $1-p$ and of alleles **B** and **b** at the second site as $q$ and $1-q$. At the population level, the gamete frequencies can be represented in the usual manner as $P_{\mathbf{AB}} = pq + D$, $P_{\mathrm{ab}} = (1-p)(1-q) + D$, $P_{\mathrm{Ab}} = p(1-q) - D$, and $P_{\mathrm{aB}} = q(1-p) - D$, respectively, where $D$ is the disequilibrium coefficient. The expected frequency of homozygous–heterozygous pairs is then

$$\begin{aligned} E(H_1) &= 2E[(P_{\mathrm{AB}} + P_{\mathrm{ab}})(P_{\mathrm{Ab}} + P_{\mathrm{aB}})] \\ &= 2E\{[pq + (1-p)(1-q) + 2D] \\ &\quad \times [p(1-q) + q(1-p) - 2D]\} \\ &= 2E\big[2p(1-p) - 4p(1-p)q(1-q) \\ &\quad - 2D(1-2p)(1-2q) - 4D^2\big], \end{aligned} \tag{3}$$

where the latter expression assumes equal expected heterozygosities at both sites, *i.e.*, $E[2p(1-p)] = E[2q(1-q)]$. This expression further simplifies to

$$E(H_1) = 2(\pi - \delta), \tag{4a}$$

where $\pi = E[2p(1-p)]$, and $\delta = E[4p(1-p)q(1-q)] + E[2D(1-2p)(1-2q)] + E(4D^2)$. Likewise, it can be shown that the expected frequencies of double homozygotes and double heterozygotes are

$$E(H_0) = 1 - 2\pi + \delta, \tag{4b}$$

$$E(H_2) = \delta. \tag{4c}$$

Similar points have been made previously from the standpoint of an entire population (Sabatti and Risch 2002).

Under the assumption of neutrality, the expected values of $H_0$, $H_1$, and $H_2$ are functions of the underlying evolutionary forces of drift, mutation, and recombination. For a two-allele model, the expected heterozygosity for an equilibrium population is $\pi = \theta/(1+2\theta)$ (Kimura 1968), and expectations for the fourth-order moments can be obtained by simplification of diffusion approximations, given as in Ohta and Kimura (1969, Equations 17),

$$E[p(1-p)q(1-q)] \simeq \left(\frac{\theta}{1+2\theta}\right)^2 \left(\frac{A+2}{4A}\right), \tag{5a}$$

$$E[D(1-2p)(1-2q)] \simeq \frac{\theta^2}{(1+2\theta)A}, \tag{5b}$$

$$E(D^2) \simeq \frac{\theta^2(10+\rho+8\theta)}{8(1+2\theta)A}, \tag{5c}$$

where

$$\begin{aligned} A &= 9 + 6.5\rho + 0.5\rho^2 + 12\theta\big(4 + 5\theta + 2\theta^2\big) \\ &\quad + \rho\theta(17 + \rho + 8\theta). \end{aligned} \tag{5d}$$

Because $\theta$ is almost always $<0.05$ in diploid organisms (Lynch 2007; Leffler *et al.* 2012), terms involving $\theta^2$ and

$\theta^3$ that do not also involve $\rho$ can be ignored, and substitution into Equation 2 yields

$$E(\Delta) \simeq \frac{\theta(1+2\theta)(18+\rho)}{2(1+\theta)A}. \tag{6}$$

An alternative expression for $E(\Delta)$ presented in Haubold *et al.* (2010), which is closely approximated by

$$E(\Delta) = \frac{\theta(1+\theta)(18+\rho)}{18+13\rho+\rho^2+54\theta+\rho\theta(19+\rho+6\theta)}, \tag{7}$$

is obtained using expressions for the expectations of $H_0$ and $H_1$ given by Strobeck and Morgan (1978) for the infinite-allele model (in contrast to the infinite-sites model relied on above). Provided $\theta \ll 1$, the deviation in the numerator is trivial relative to the leading term $\theta(18 + \rho)$, and this is likely true in the denominator as well. Thus, focusing only on the leading terms that are similar

$$E(\Delta) \simeq \theta\left(\frac{18+\rho}{18+13\rho+\rho^2}\right). \tag{8}$$

The latter expression is also obtained with the infinite-sites model (where each newly arisen mutation occurs at a previously homozygous site) (Ohta and Kimura 1971; Hill 1975), demonstrating its generality with respect to the features of the mutation model. The quantity within parentheses has appeared previously in the literature, being equivalent to the probability that pairs of nucleotides at two sites are derived from the same common ancestor (Pluzhnikov and Donnelly 1996; Wiuf and Hein 2000; McVean 2002) and also to the correlation of coalescence times at two sites (Hudson 1991). Note that as $\rho \to 0$, $E(\Delta) \to \theta$; for $\rho \ll 1$, $E(\Delta) \simeq \theta[1 - (2\rho/3)]$; and as $\rho \to \infty$, $E(\Delta) \to \theta/\rho$.

As a check on the validity of the diffusion-theory approximation, we evaluated the average values of the two-site features derived from simulations of a Wright–Fisher model over the full range of per-site values of $\theta$ and $\rho$ observed in eukaryotes (Lynch 2007). The agreement between the expectations of the theory and simulated data was excellent.

#### *Relationship to population-level linkage disequilibrium*

How does $\Delta$ relate to the more conventional measures of LD extracted from population samples of multiple individuals, *e.g.*, $D^2$? Because $E[p(1-p)q(1-q)] \simeq 0.25\pi^2$ over the full range of allele frequencies, it follows from Equations 2 and 3 that

$$E(\Delta) \simeq \frac{2\{E[D(1-2p)(1-2q)] + 2E(D^2)\}}{\pi(1-\pi)}, \tag{9}$$

which using the expressions for the moments in the numerator reduces to

$$E(\Delta) \simeq \frac{2E(D^2)\left[2 + (1.25+0.125\rho)^{-1}\right]}{\pi(1-\pi)}, \tag{10}$$

when $\theta \ll 1$. Because the term within the square brackets is between 2.0 and 2.8 (for large and small $\rho$, respectively), as a rough approximation

$$E(\Delta) \simeq \frac{4E(D^2)}{\pi(1-\pi)}. \tag{11}$$

An alternative measure is the standardized linkage disequilibrium, defined by Ohta and Kimura (1969) as

$$r_d^2 = \frac{E(D^2)}{E[p(1-p)q(1-q)]}. \tag{12}$$

Although broadly used, this statistic is problematic in that it is not an unbiased estimator of the squared correlation, which is a function of the expectation of the ratio rather than the ratio of expectations. Because this difference in definition can lead to up to 100-fold differences in values of $r_d^2$ if allele frequencies are extreme, as is often the case with neutral alleles (McVean 2002; Song and Song 2007), most investigators restrict LD analyses to pairs of loci with intermediate allele frequencies. Such treatment is not possible, but also not necessary, with the proposed method. Again noting that the denominator is approximately $\pi^2/4$, we see that $r_d^2 \simeq 4E(D^2)/\pi^2$, so from Equation 11, provided $\pi \ll 1$, $E(\Delta) \simeq \pi r_d^2$.

#### *Interpretation of the recombination rate*

Recombination events involve heteroduplex formations between homologous chromosomes, where nonmatching sites must be resolved by mismatch repair, which causes gene conversion. Inclusion of this factor in the definition of the recombination rate is essential because although all recombination events result in gene conversion, not all conversion events are accompanied by crossovers.

Let $c$ be the total recombination rate per single nucleotide site (*i.e.*, the rate of recombination per site with or without crossing over), $x$ be the fraction of recombination events resulting in a crossover, and $d$ be the number of sites separating the two focal positions (with $d = 1$ for adjacent sites). From a two-site perspective, a gene-conversion event has consequences equivalent to a crossover if the conversion is restricted to a single heterozygous site, a point first made by Andolfatto and Nordborg (1998) under the assumption of a constant conversion tract length. Following Langley *et al.* (2000) and Frisse *et al.* (2001), and allowing for an exponential distribution of tract lengths with an average length $\bar{L}$ (in base pairs), the total recombination rate per site associated with conversions is $c\bar{L}(1-e^{-d/\bar{L}})$, with the term in parentheses being the fraction of conversion events covering only single sites. Because a single-site conversion accompanied by a crossover restores the original state, the total recombination rate is then

$$r \simeq c\left[xd + (1-x)\bar{L}\left(1-e^{-d/\bar{L}}\right)\right]. \tag{13a}$$

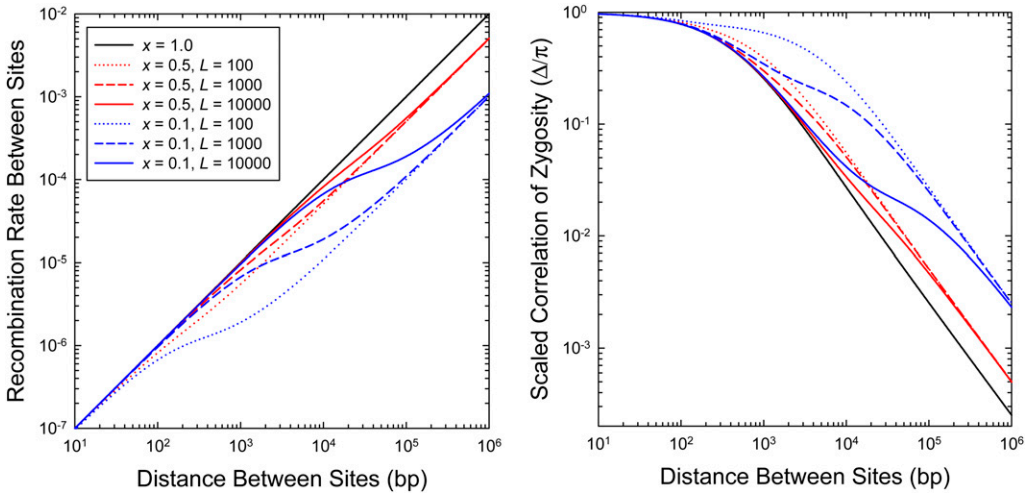


**Figure 1** (Left) The relationship between the recombination rate between pairs of sites separated by distance $d$, given for two values of $x$ (the fraction of recombination events resulting in crossing over), and three of $\bar{L}$ (the mean gene-conversion-tract length). (Right) The relationship between the expected value of the scaled statistic $\Delta/\pi$ and the distance between sites, for the same values of $x$ and $\bar{L}$. The results are obtained by using Equation 13a to define the distance-dependent recombination rate function and then applying this to Equation 6. In this particular example, $N_e = 10^5$ and $u = c = 10^{-8}$.

Thus, although it is commonly assumed that the recombination rate increases linearly with distance between sites, gene conversion results in two domains with different distant-dependent rates. For $d \ll \bar{L}$,

$$r \simeq cd, \tag{13b}$$

whereas for $d \gg \bar{L}$,

$$r \simeq cdx. \tag{13c}$$

Gene conversions unaccompanied by crossing over cause a lower rate of recombination at all distances, with a fractional reduction $x$ from very short to very long distances. Under the exponential-distribution model, the transition between the two asymptotic rates initiates at a distance approximately equal to $\bar{L}$ and is nearly complete at distance $10\bar{L}$. This effect leaves a characteristic signature in the expected relationship between $\Delta$ and $d$ (Figure 1).

### *Influence of population-size history*

Because the population-genetic interpretations of $\Delta$ (and indeed all prior measures of LD) noted above are based on the assumption of a constant long-term population size, it is prudent to consider the potential effects of past demographic changes. This problem can be readily examined by use of the recursion equations presented by Strobeck and Morgan (1978), whose equilibrium solutions were referred to above. As in Figure 1, it is useful here to simply evaluate the behavior of the scaled correlation $\Delta/\theta_0$, where $\theta_0$ denotes the current heterozygosity in the population, as this then leaves the influence of recombination as a separate term, *e.g.*, Equation 8. (Recall from above that $\Delta/\theta_0 \simeq r_d^2$.) To further simplify the presentation, we also assume $x = 1$ in the following examples.

With populations with constant $N_e$, the scaled correlation converges on values of 1.0 for all $N_e$ at short distances between sites, and then yields a set of parallel declining lines (on a logarithmic scale) at higher distances, with higher elevations associated with smaller $N_e$ (Figure 2). This asymptotic behavior, which arises because $\Delta \simeq 1/(4N_e cxd)$ at large $d$, implies an expected inverse relationship between $\Delta$ and $d$ at large distances.

For populations changing in size, we considered demographic shifts within the past $4N_c$ generations, where $N_c$ is the current effective population size. Because $4N_c$ is the mean coalescence time for a neutral mutation in a population of constant size, current patterns of LD will be negligibly influenced by demographic changes prior to this point. For populations declining in size progressively through time, the scaled correlation exceeds 1.0 at short distances between sites and then converges on the expected values for the current population size at large distances (Figure 2). In effect, continuously shrinking populations result in a situation in which LD at closely linked sites is elevated relative to the expectation for a constant-size population with the observed $\theta_c$. Such behavior is a reflection of the fact that the current population, which encourages greater buildup of LD, retains old variation reflective of a larger $N_e$. Populations expanding in size exhibit the opposite pattern—scaled correlations $<1.0$ at the shortest distances, but again with values converging on the expectations for the current population size at large $d$. Linear rates of change yield slightly less extreme behavior than exponential growth/decline (Figure 2).

An abrupt but transient transition to a novel population size was also investigated. In this case, both a sudden population expansion or bottleneck results in a reduction in the scaled correlation relative to expectations for the current population size, except in the case of a very distant expansion, which only trivially increases $\Delta/\theta_0$ (Figure 2). Again, large effects are noticeable only at short physical distances ($<1$ kb), with the results for larger distances reflecting the expectations for a constant population size $N_c$ with observed $\theta_c$, even though the latter is not the current equilibrium expectation $4N_c u$ because of the lingering influence of historical changes in $N_e$.

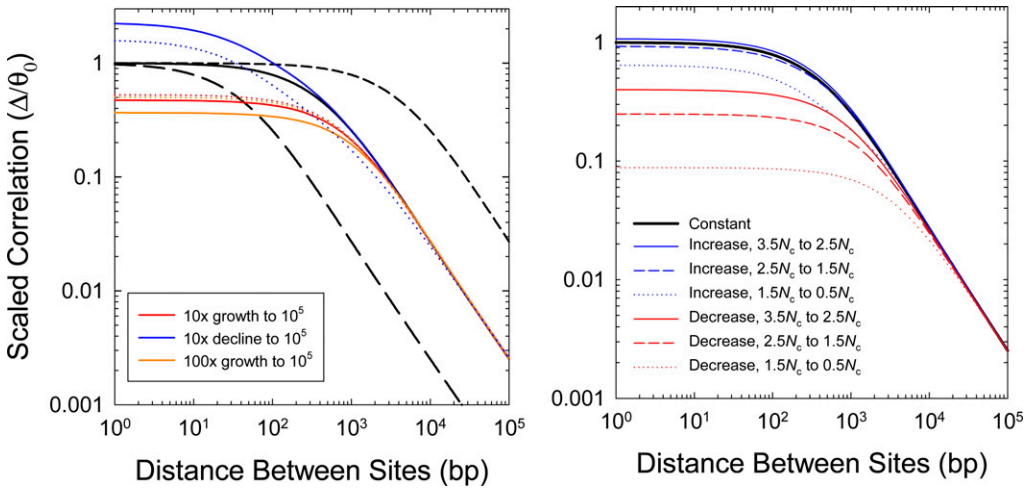


**Figure 2** Effects of long-term demographic changes on the behavior of Δ, here scaled to the level of heterozygosity in the current population. (Left) Cases of continuous growth or decline in population size. The denoted changes initiate $4N_c$ generations in the past, where $N_c$ is the current effective population size, starting with an ancestral population assumed to be in drift–mutation equilibrium using the expected values based on its size. Thick black lines denote three cases of constant population size, whereas colored lines denote cases of continuous population size change (solid for exponential change; dotted for linear change). (Right) Cases of abrupt but transient change in population size for intermediate periods of $N_c$ generations, for various periods in the past. The ancestral population is assumed to have the same size as the current population and to have been in drift–mutation equilibrium. In all cases, the current effective population size is $N_c = 10^5$, with per-site mutation ($u$) and recombination ($c$) rates both equal to $10^{-8}$.

These results suggest that estimates of Δ at very short distances can be influenced by past population-size changes, although the particular patterns observed are not necessarily very diagnostic of the historical nature of change unless the temporal pattern is largely continuous. For example, depending on the magnitude of change, transient population expansions and bottlenecks can lead to qualitatively very similar behaviors. On the other hand, even with extreme demographic changes, estimates of $\Delta/\pi$ for pairs of sites separated by >1 kb closely reflect the expectations of an equilibrium population with the current size. Hence, at appropriate physical distances, Equation 8 provides informative insight into the recombinational properties responsible for the pattern of decline of Δ with $d$.

### *Estimation procedures*

We now consider the practical issue of obtaining unbiased and minimum-sampling-variance estimates of Δ. Consider an idealized data set in which the raw reads for a single individual have depth of sequencing coverage $n$. With very high depths of coverage, the genotypic states at each locus can usually be determined essentially unambiguously from the consensus sequence (assuming proper alignment of sites). However, most high-throughput-sequencing methods are highly error prone, and most projects deploy a low enough level of sequence coverage per site that genotypic ascertainment is problematic. To deal with such problems, a maximum-likelihood (ML) procedure has been suggested for acquiring joint estimates of $\pi$ and Δ that best explain the read data at the full collection of genomic sites, while simultaneously correcting for biases associated with sequencing errors (Lynch 2008).

The general strategy is to condense the data at each pair of sites (say $a$ and $b$) down to quartets enumerating the numbers of times the four alternative nucleotides have been observed, ($n_{a\mathrm{A}}$, $n_{a\mathrm{C}}$, $n_{a\mathrm{G}}$, $n_{a\mathrm{T}}$) and ($n_{b\mathrm{A}}$, $n_{b\mathrm{C}}$, $n_{b\mathrm{G}}$, $n_{b\mathrm{T}}$). Letting $\mathbf{n}$ denote the octet for the pair of sites, the likelihood of the observed data can be expressed as

$$\ell(\mathbf{n}) = H_0\ell_{1a}\ell_{1b} + H_2\ell_{2a}\ell_{2b} + (1 - H_0 - H_2)(\ell_{1a}\ell_{2b} + \ell_{1b}\ell_{2a}), \tag{14}$$

where the four terms $\ell_{1a}$, $\ell_{1b}$, $\ell_{2a}$, and $\ell_{2b}$ describe the site-specific likelihoods of observations conditional on being homozygous ($\ell_{1a}$ and $\ell_{1b}$) or heterozygous ($\ell_{2a}$ and $\ell_{2b}$), defined below. The ML solution is the joint set of values of $\hat{H}_0$ and $\hat{H}_2$ (with $\hat{H}_1 = 1 - \hat{H}_0 - \hat{H}_2$) that maximizes the product of the likelihood of the observations over all pairs of sites separated by a specific distance $d$.

Each of the $\ell$ terms is a function of the genome-wide nucleotide composition, binomial allelic sampling (in the case of heterozygotes), and the error rate ($\varepsilon$, itself a parameter to be estimated). Expressions for these terms were given previously as Equations 4a and 4b in Lynch (2008), where issues of bias were noted. In the intervening time, we have formulated modified expressions that greatly reduce the bias in the derived parameter estimates, even at low coverage.

For each site, the probability of an observed quartet conditional on the site being homozygous is

$$\ell_1(n_1, n_2, n_3, n_4) = \sum_{i=1}^{4} p_i \cdot b(n - n_i; n, \epsilon) \cdot \binom{n - n_i}{n_j} \times \binom{n - n_i - n_j}{n_k} (1/3)^{n - n_i}, \tag{15a}$$

where $p_i$ is the frequency of nucleotide type $i$ in the genome (and hence the fraction of homozygotes expected to be of type $i$), $b(n - n_i; n, \varepsilon)$ is the binomial probability of $n - n_i$ errors in $n$ reads, conditional on the individual being a homozygote of type $i$ and given the error rate $\varepsilon$, and the final term defines the probability of the observed distribution of

the three hypothetical error types conditional on $i$ being the proposed homozygote type (the indices used are such that $i \neq j \neq k$, with $j$ and $k$ denoting two of the error types and the count for the third simply being defined as $n - n_i - n_j - n_k$). This final term, not used in Lynch (2008), provides information on the error rate not contained in the binomial term alone.

Conditional on the site being heterozygous, the likelihood of an observed quartet is

$$\ell_2(n_1,n_2,n_3,n_4) = \sum_{i=1}^{4}\sum_{j>i}^{4} 2p_ip_j \cdot b\left(n-n_i-n_j;n,2\epsilon/3\right) \cdot p\left(n_i;n_i+n_j,0.5\right)\cdot b\left(n_k;n-n_i-n_j,0.5\right)/S. \quad (15b)$$

Here, $2p_ip_j/S$ is the genome-wide expected frequency of $ij$ heterozygous genotypes among all possible heterozygotes, with $S = 1 - \sum_{i=1}^{4} p_i^2$ serving to normalize the sum of possible heterozygote frequencies to 1.0; $b(n - n_i - n_j; n, 2\varepsilon/3)$ is the probability of errors to nucleotides other than $i$ and $j$; $p(n_i; n_i + n_j, 0.5)$ is the probability of sampling the $i$th nucleotide $n_i$ times from the pool of hypothetically nonerroneous reads ($n_i + n_j$); and the final term (not included in Lynch 2008) is the probability of the observed distribution of read types interpreted as errors (not $i$ or $j$).

To determine the power and accuracy of the ML estimator, using the Wright–Fisher model described above, random pairs of two-locus genotypes were generated using biologically realistic values for the effective population size ($N_e$) and site-specific rates of mutation ($u$) and recombination ($r$) to define $\theta$ and $\rho$, which in turn defined the expected values of $H_0$ and $H_2$ based on Equations 4a–4c. Errors in sequences were generated by assuming constant rates per site per read with $\varepsilon = 0.01$, an unfortunately realistic situation. ML estimates of $\varepsilon$, $H_0$, and $H_2$ were obtained by maximizing the log-likelihood of the observed data (*i.e.*, the log of the product of Equation 14 for full sets of sites), and these were then extrapolated to yield

$$\hat{\pi} = 0.5\hat{H}_1 + \hat{H}_2 \quad (16a)$$

$$\hat{\Delta} = 1 - \frac{\hat{H}_1}{2\hat{\pi}(1-\hat{\pi})}. \quad (16b)$$

Some sets of results, with each independent simulation involving $10^5$ to $10^6$ pairs of sites, are compared to their expectations as defined in Figure 3 by Equation 6. On average, the estimates of $\Delta$ are independent of the level of coverage, provided the latter is $>4\times$. In some cases, the estimates are upwardly biased, but generally by no more than 25%, and such biases become asymptotically negligible with large numbers of sites.

In practical applications of this method, computational efficiency can be dramatically enhanced by performing a first pass through the data in a univariate analysis to estimate $\pi$ and $\varepsilon$ (Lynch 2008). This then enables the precomputation of values of $\ell_1$ and $\ell_2$ for different quartets, which can be repeatedly reused in analyses performed over different distances. The estimates of $\Delta$ for each distance of interest can then be rapidly acquired in a second round of analysis from the ML estimates of $H_0$ and $H_2$. These and other refinements are now implemented in the program mlRho v. 2.2 (available at http://guanine.evolbio.mpg.de/mlRho/). Strictly speaking, the overall approach is equivalent to a composite-likelihood procedure, as not all pairs of sites are genealogically independent (although those on different chromosomes are expected to be, so that for species with $n$ chromosomes, no more than a fraction $\sim 1/n$ of pairs of pairs will be nonindependent).

## Empirical Observations

To obtain insight into the recombinational features of natural populations, we applied the preceding methods to single-genome sequences of several noninbred vertebrate species, all of which harbor enough sites to avoid issues of bias (Supporting Information, File S1). As discussed in Haubold *et al.* (2010), to avoid artifacts that might arise from paralogous regions (including repetitive DNAs such as mobile elements), such applications require the initial mapping of the full set of raw sequencing reads to the genome assembly, followed by the removal of any reads that map to more than one genomic location and also masking paralogous regions and any other sites that attract unusually large numbers of reads.

To illustrate some general issues, we initially focus on results for four human genomes (one archaic), after which we summarize further observations across the vertebrate phylogeny. In accordance with the expectations outlined in Figure 1, the decline in $\Delta$ with physical distance is strongly nonlinear in all genomes, even on a logarithmic scale, and this nonlinearity remains apparent at all scales of physical distance (Figure 4). Particularly striking are the high values and very rapid decline of $\Delta$ over distances $<100$ bp or so. As noted above, for the standard neutral model, $\Delta$ is expected to converge on the standing level of heterozygosity ($\pi$) as distance $d \to 1$, but estimates of $\Delta$ at $d < 100$ are often an order of magnitude or more greater than the ML-derived estimates of $\pi$ (Table 1). Potential reasons for such behavior are discussed below.

The scaled correlations (obtained by dividing the ML estimates of $\Delta$ by the respective estimates of $\pi$) point to different past demographic histories of the sequenced humans (Figure 4). First, with the exception of the Watson sequence, the profiles are essentially parallel, with the estimates for the Chinese genome being elevated $\sim 7\times$ relative to that for the African genome and the estimates for the Denisovan genome being elevated a further fourfold. As noted above, such behavior is indicative of an $\sim$30-fold larger long-term $N_e$ in the recent ancestry of the African genome than in that of the Denisovan genome. The elevation of Denisovan estimates substantially $>1.0$ may imply

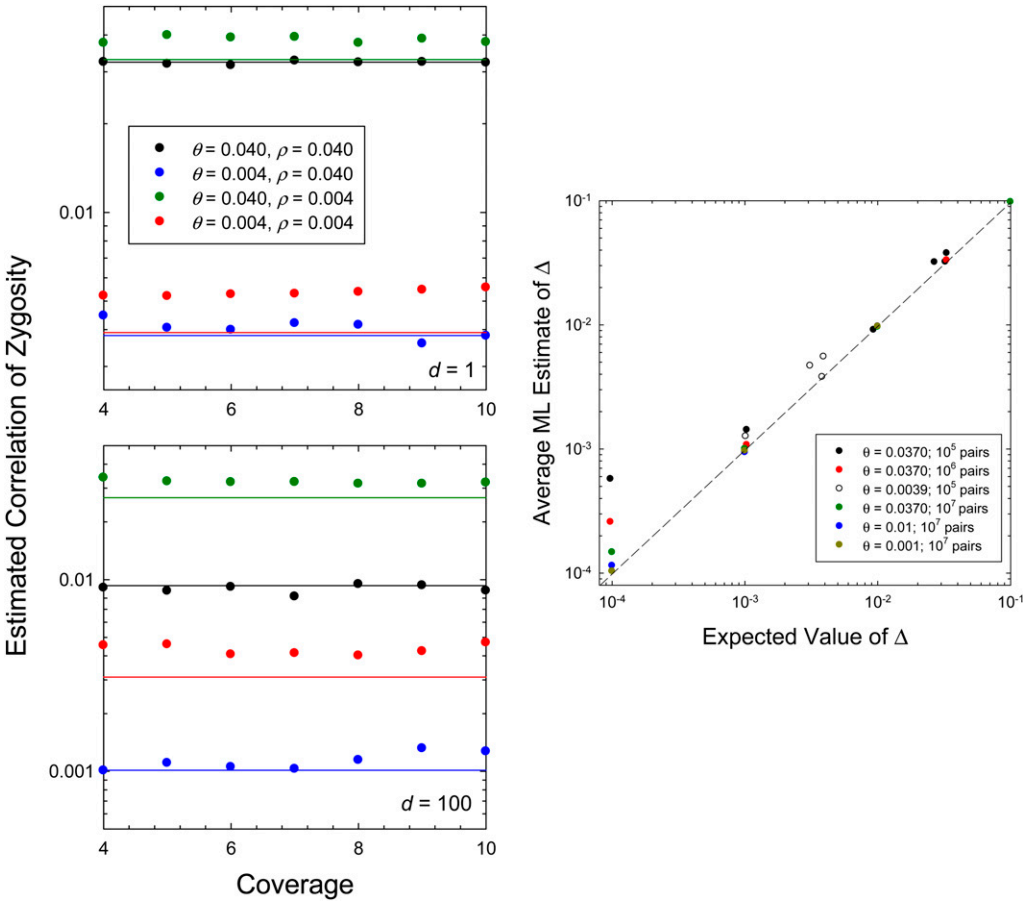


**Figure 3** Mean ML estimates for Δ obtained from simulated data. In all cases, the effective population size is $N_e = 10^6$. On the left, it is shown that the mean ML estimates of Δ are independent of the level of sequence coverage per site; 200 replicate runs were made per condition, with $10^5$–$10^6$ independent pairs of sites sampled in each case. The horizontal lines denote the expected values given by Equation 6. On the right, for 10× coverage, it is shown that mean estimates of Δ are asymptotically unbiased with increasing numbers of sites; as on the left, varying combinations of the per-site parameters $\theta$ and $\rho$ underlie the individual points.

a rapid ancestral decline in population size, whereas the estimates for the African genome are in better agreement with population stasis. The data for the Watson genome are qualitatively consistent with a current $N_e$ similar to that for the Chinese genome, but with a dramatic decline from a large ancestral population size, possibly a consequence of an out-of-Africa bottleneck. Finally, we note that the expected inverse relationship between Δ and $d$ is approached only at distances on the order of $10^5$, which is indicative of a substantial amount of noncrossover recombination events even at large distances.

To obtain estimates of the recombination parameters $\rho_1 = 4N_e c$, $x$, and $L$, we generated least-squares fits to Equation 8 as a function of $d$, with Equation 13a used to define the recombination rate as a function of distance $d$. Because of the nonlinear nature of the relationship between Δ and $d$, some compromises need to be made in deriving such fits, as overreliance on short *vs.* long distances will bias the fits to the ranges of distances containing the most data. Our analyses involved estimates of Δ at 1-bp intervals up to sites $10^3$ bp apart and pooled estimates for 10-bp windows for distances $10^3 < d \leq 10^4$ bp and for 100-bp windows for distances $10^4 < d \leq 10^5$ bp. The statistical analyses then sought the joint set of the four parameter estimates ($\theta$, $\rho_1$, $x$, and $L$) that minimized the mean-squared deviations of the logarithm of Δ from the model expectations. As the data for very short distances have features that apparently violate the assumptions of the model, these analyses were restricted to pairs of sites with distances in the range of 250 to $10^5$ bp. As can be seen in Figure 4, the resultant fits provide an excellent description of this range of $d$.

Several tentative conclusions can be drawn from the fitted parameter estimates for these human genomes (Table 1). First, the estimated fraction of human recombination events resulting in crossing over is in the range of $x$ = 0.066 to 0.235, with a mean of 0.171 (SE = 0.038). Second, estimates of $\rho_1$ are in the range of 0.00015 to 0.0015, with a mean of 0.00062 (0.00030). Third, estimates of average conversion-tract lengths are in the range 1389–3208 bp, with a mean of 2100 (390).

Qualitatively similar results were obtained with analyses of five other primate genomes (Figure 5). Again, the estimates of Δ are aberrantly high at short distances between sites, and again the expected asymptotic inverse relationship between Δ and $d$ is not attained even at a distance of 100 kb. When restricted to sites with $d > 250$ (>1000 in the case of chimpanzee, which exhibits aberrant behavior at smaller distances), the fit of Equations 8 and 13a to the data are again quite good. For this set of species, $x$ has an average value of 0.075 (0.027), whereas the average value of $\overline{L}$ is 4332 (976). Overall, the data on LD for distant sites suggest that recent effective population sizes are highest for *Pongo*

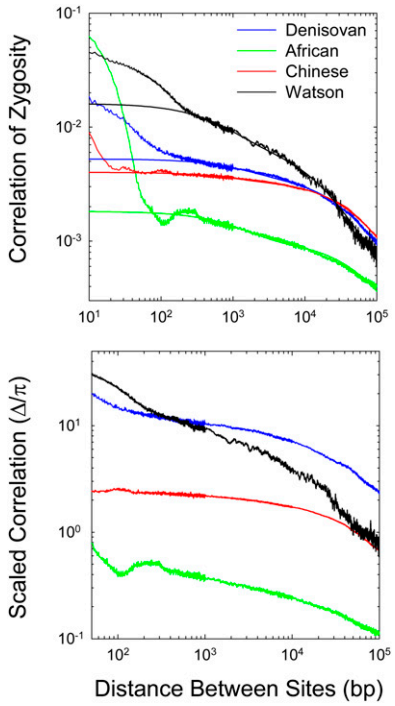


**Figure 4** The decline of the correlation of zygosity (Δ) with physical distance between sites, given for an archaic human (Denisovan) and three modern humans (African, Chinese, and Watson). (Top) The smooth lines are the fitted functions using sites separated by 250 to $10^5$ bp as described in the text (with the fitted parameters given in Table 1). (Bottom) Behavior of the correlation scaled by the estimates of heterozygosity within each genome.

**Table 1 Estimates of the features of recombination from the pattern of decline of Δ with physical distance between sites, derived from genomic sequences of single individuals of vertebrates (File S1)**

| Species | $\theta$ | $\theta'$ | $\rho_1$ | $x$ | $\bar{L}$ |
|---|---|---|---|---|---|
| Primates | | | | | |
| *Gorilla gorilla* | 0.0031 | 0.0031 | 0.00057 | 0.067 | 3792 |
| *Macaca mulatta* | 0.0026 | 0.0149 | 0.00218 | 0.018 | 1068 |
| *Pan troglodytes* | 0.0010 | 0.0024 | 0.00033 | 0.177 | 4286 |
| *Pongo abelii* | 0.0053 | 0.0036 | 0.00062 | 0.043 | 5662 |
| *P. pygmaeus* | 0.0027 | 0.0047 | 0.00064 | 0.069 | 6852 |
| *Homo sapiens* (Archaic Denisovan) | 0.0004 | 0.0051 | 0.00022 | 0.235 | 3208 |
| *H. sapiens* (African) | 0.0036 | 0.0018 | 0.00064 | 0.066 | 1970 |
| *H. sapiens* (Chinese) | 0.0016 | 0.0039 | 0.00015 | 0.216 | 1833 |
| *H. sapiens* (Watson) | 0.0010 | 0.0161 | 0.00146 | 0.168 | 1389 |
| Nonprimate mammals: | | | | | |
| *Ailuropoda melanoleuca* | 0.0013 | 0.0018 | 0.00023 | 0.250 | 16267 |
| *Canis familiaris* | 0.0009 | 0.0082 | 0.00503 | 0.020 | 1183 |
| *Loxodonta africana* | 0.0013 | 0.0221 | 0.00286 | 0.021 | 1084 |
| *Ornithorhynchus anatinus* | 0.0013 | 0.1709 | 0.28970 | 0.025 | 608 |
| Nonmammalian vertebrates: | | | | | |
| *Anolis carolinensis* | 0.0026 | 0.0094 | 0.00214 | 0.058 | 1927 |
| *Fugu rubripes* | 0.0032 | 0.0042 | 0.00128 | 0.025 | 7238 |
| *Petromyzon marinus* | 0.0044 | 0.0061 | 0.00118 | 0.052 | 2863 |

$\theta'$, an inferred estimate of $4N_e u$ from the fit to Equation 8, not necessarily equivalent to the independently derived ML estimate given as $\theta$); $\rho_1 = 4N_e c$, where $c$ is the recombination rate between adjacent sites; $x$, the fraction of recombination events leading to crossovers; and $\bar{L}$, the mean conversion-tract length (in base pairs).

*abelii*, lower for *P. pygmaeus* and *Gorilla gorilla*, still lower for *Pan troglodytes*, and lowest for *Macaca mulatta*. Although such estimates may seem discordant with the current day population sizes induced by recent human activity, the past few dozen generations are not expected to be sufficient to greatly influence standing levels of LD relative to the entire demographic history of the species. Results for additional vertebrates, summarized in Table 1, again consistently imply $x < 0.25$ and $\bar{L}$ on the order of one to several kilobases.

The approach taken above is computationally intensive, involving well over $10^9$ nucleotide sites for most analyses, which translates to orders of magnitude more pairs of sites depending on the sizes of the assembled contigs. Although not all pairs of sites are genealogically independent, the vast majority will be essentially so in vertebrates, which typically harbor a few dozen chromosomes approximately equal in size. Thus, the composite-likelihood approach appears to be justified statistically for individual estimates of Δ. The program mlRho computes the confidence interval around each estimate of Δ in the usual way by searching for positions of decline of two units on the likelihood surface; by solving Equation 8, this interval can then be translated into statistical bounds on the estimates of $\rho$. From the limited scatter of the data in Figure 4 and Figure 5 alone, it can be seen that with the magnitude of data available for vertebrate genomes, individual data points are quite reliable.

On the other hand, because estimates of Δ at close physical distances do share genealogical history, there is the issue of nonindependence along the distance profile. Such sampling covariance will be a function of not only the finite number of sites involved in an analysis, but also the evolutionary history of a population, and as a consequence no fully general statement can be made about the absolute magnitude of noise in a Δ profile that might be expected. However, to gain some insight into the matter of how variable Δ profiles might be among individuals, we consider two data sets.

First, we present empirical data on the Δ profile derived from 10 genomes of the microcrustacean *Daphnia pulex*, each derived from a different population (Figure 6). All 10 of these genotypes were extracted from populations known to be randomly mating on an annual basis and in Hardy–Weinberg equilibrium (Tucker *et al.* 2013). Despite the different population sources (ranging over several hundreds of kilometers), the Δ profiles are highly similar. Deviations become slightly larger at increasing physical distances because of declining sample sizes (and possibly recent differences in population-specific demographic histories).

The average estimate of $\theta = 4N_e u$ for these *D. pulex* genotypes is 0.0204 (0.0005), whereas the average estimate of $4N_e c$ is 0.00069 (0.00003), implying that on a per-site basis the mutation rate in this species is about thirty times the recombination rate. The estimated fraction of recombination events resolved as crossovers is $x = 0.099$ (0.020), and the average conversion tract length is $\bar{L} = 17791$ (2144). *Daphnia* is one of the few metazoan species in which conversion

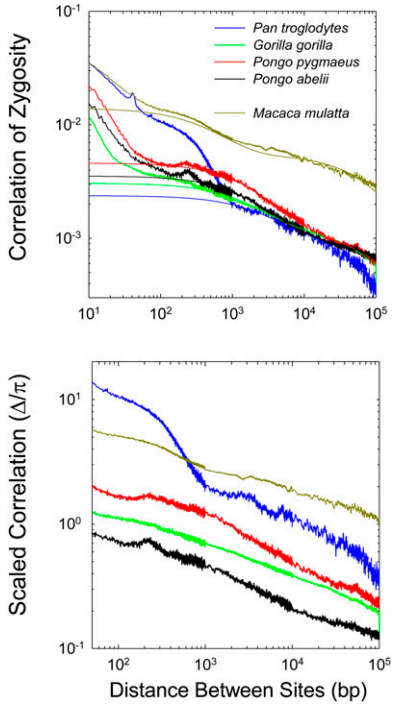


**Figure 5** The relationship between the correlation of zygosity (Δ) and physical distance between sites for five nonhuman primate species. Details as in Figure 4.

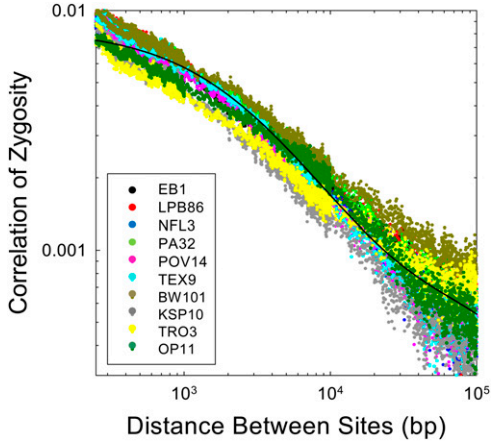


**Figure 6** Observed patterns of decline of the correlation of zygosity with physical distance between sites for 10 natural isolates of *Daphnia pulex*, each derived from a different population (data are derived from Tucker *et al.* 2013 for sexual genotypes). The fitted curve is based on the average parameter estimates over all 10 isolates, each applied to data for pairs of sites >250 bp in distance.

tracts can be directly observed by loss-of-heterozygosity in clonally propagated lines, and direct observations in this species (Omilian *et al.* 2007; Xu *et al.* 2011) are entirely consistent with these indirect estimates.

Second, to specifically investigate the variation in Δ profiles that can arise among individuals within a single population, we used the computer program ms (Hudson 2002) to generate independent diploid genotypes with the same input parameters (using the average values noted above for human sequences). As can be seen in Figure 7, when large numbers of sites are available per individual, the Δ profiles are very similar. Thus, there is little question that the variation observed among individuals in Figure 4 is a consequence of true variation in background parameters, and not a simple consequence of stochastic within-population variation.

## Discussion

Although numerous techniques have been proposed for estimating LD from population samples (Stumpf and McVean 2003; Slatkin 2008), most applications of these methods have involved small to moderate numbers of nucleotide sites harboring alleles with intermediate frequencies. In addition, almost all attempts to estimate the population recombination rate $\rho_1 = 4N_e c$ have assumed a standard neutral model with independently arising mutations, negligible gene conversion, and a linear relationship between the recombination rate and physical distance. The results presented above imply that all of these assumptions are violated in a wide range of species.

Rather than focusing on a large population sample with imprecise allele-frequency estimates, the approach presented here uses the information contained within the two chromosome sets of just a single diploid individual to estimate a function of LD that is directly related to the conventional population-level measure of LD, $E(D^2)$. Although estimation of the underlying recombinational parameters using this method does assume that the sampled individual is derived from a randomly mating population, the same assumption applies to extrapolations made with all prior LD methods involving multiple individuals.

Because a nonlinear regression involving four parameters has the capacity to yield a diversity of functions, an obvious question is whether the preceding empirical results are simply a curve-fitting procedure, as opposed to providing insights into the actual physical features of recombination. Several lines of evidence support the latter view. First, our conclusion that the fraction of recombination events resulting in crossing over ($x$) is generally on the order of 0.2 or less is quite consistent with previous observations. For example, $x$ has been estimated to be in the range 0.25–0.34 in budding yeast *Saccharomyces cerevisiae* (Malkova *et al.* 2004; Mancera *et al.* 2008) and $\sim$ 0.14 in humans (Frisse *et al.* 2001; Padhukasahasram and Rannala 2013). The latter is not significantly different from our estimate of 0.17 for humans nor from our average estimate of 0.08 for other primates. For gene-sized regions in *Drosophila melanogaster*, Langley *et al.* (2000) concluded that almost all recombination is a consequence of conversion events, and using their data set, Yin *et al.* (2009) inferred $x = 0.08$, which is not greatly different than a direct empirical estimate of 0.15 in same species (Hilliker *et al.* 1994). From a high-resolution cross in *Arabidopsis*, Yang *et al.* (2012) inferred $x \simeq 0.05$, and population sampling yields estimates of $x = 0.06$ and 0.16 for barley and maize, respectively (Morrell *et al.* 2006).

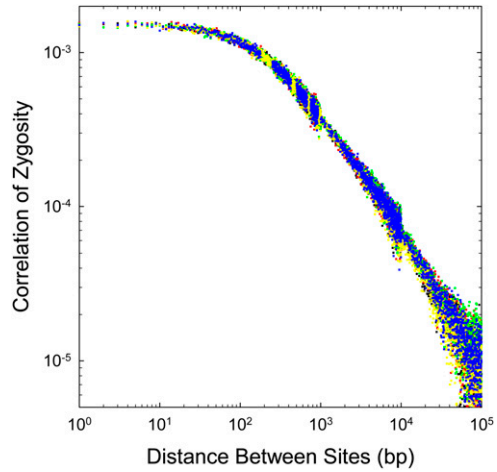


**Figure 7** Replicate Δ profiles for five simulated human genomes (different colored points), each using the same mutation and recombination parameters (averages from the four observed genomes): $\rho_1 = 4N_e c = 0.0062$, $x = 0.17$, $\bar{L} = 2100$, and $\theta = 0.0016$. Estimates beyond a distance of 1000 bp are pooled (bins of 10 bp for 1–10 kb, and bins of 100 bp for >10 kb). Discontinuities in the range of variation arise at 1 and 10 kb owing to this averaging and the fact that the sampling variance of LD increases with physical distance between sites. Each of the five sets of observations were generated by independent coalescents, with 10× coverage in the subsequently generated sequence, and Δ being estimated with the program mlRho.

Second, our conclusion that mean conversion-tract lengths are typically on the order of one to several kilobases is also consistent with previous observations. Estimates of $\bar{L}$ are on the order of 400 bp in *Eubacteria* (Santoyo and Romero 2005), in the range 500–4000 bp in *S. cerevisiae* (Ahn and Livingston 1986; Judd and Petes 1988; McGill *et al.* 1990), and 400–1400 bp in *D. melanogaster* (Hilliker *et al.* 1994; Preston and Engels 1996; Miller *et al.* 2012). Paigen *et al.* (2008) have observed conversion-tract lengths in the mouse ranging from 200 to 1200 bp, and Padhukasahasram and Rannala (2013) have inferred tract lengths of 100–900 bp on human chromosomes. Other observations reviewed in Chen *et al.* (2007) and Rukść *et al.* (2008) suggest mammalian conversion-tract lengths in the range of a few hundred to two thousand base pairs. Finally, we note that a variety of mechanisms associated with double-strand-break repair, including break-induced replication and synthesis-dependent strand annealing, can lead to loss-of-heterozygosity tract lengths exceeding several thousands of base pairs (Omilian *et al.* 2007; Llorente *et al.* 2008; Xu *et al.* 2011). Such effects will have the same influence on LD as conventional meiotic gene conversions and almost certainly contribute to the mean tract-lengths inferred in our analyses.

Taken together, these observations strongly suggest that the vast majority of meiotic recombination events are resolved as simple gene conversions without crossing over, with average tract lengths of many hundreds to thousands of base pairs. Assuming an approximately geometric distribution of tract lengths (Ahn *et al.* 1988; Hilliker *et al.* 1994), which may underestimate the actual range of lengths, this means that conversion-like events up to tens of kilobases commonly occur. Thus, rates of recombination between sites separated by less than many thousand kilobases are generally much greater (up to $1/x$ greater) than one would expect based on a linear extrapolation from estimates of crossover frequencies derived from genetic maps.

Although future work with a much wider range of taxa will be required to determine whether the known cytological differences in meiosis (Kohl and Sekelsky 2013) translate into quantitatively significant differences in the features of recombination (*e.g.*, $x$ and $\bar{L}$) among phylogenetic lineages, the consistency of our results with prior observations suggests a way forward. Aside from its applicability to the sequenced genomes of just single individuals, a substantial advantage of the proposed method is that it provides a genome-wide average view of recombinational features that is not dependent on the kinds of artificial constructs often deployed in laboratory experiments. In addition, analyses can readily be partitioned spatially to estimate regional features of chromosomes. Finally, because the method is agnostic with respect to allele frequencies, it provides a global estimate of LD that should be unaffected by biases that can arise when analyses are restricted to alleles that are able to rise to high frequencies, as has often been the focus of studies using $r_d^2$.

While the mechanisms remain unclear, our observations of anomalous patterns of high LD at short distances are consistent with results from previous analyses in *Drosophila* (Andolfatto and Przeworski 2000), humans (Przeworski and Wall 2001), sorghum (Hamblin *et al.* 2005), and *Arabidopsis* (Kim *et al.* 2007), all of which suggest problems with the standard neutral model at short (typically intragenic) distances. The central issue is that the absolute amount of LD at distances <200 bp or so is far greater than expected under models that assume independent mutational and recombinational events.

There are at least three reasons why genetic events on such small spatial scales should not be treated as independent. First, as pointed out by Schrider *et al.* (2011), new mutations arise in a significantly clustered manner on spatial scales <100 bp or so. Such effects might result simply from an occasional defective polymerase engaging at origins of replication. Second, Hicks *et al.* (2010) find in *S. cerevisiae* an ∼800-fold increase in the mutation rate in newly synthesized strands involved in gene conversions, and more generally, double-strand-break repair appears to be mutagenic (Malkova and Haber 2012). Such effects not only contribute to mutation clustering, but also render mutation and recombination nonindependent events. Third, nonhomologous gene conversion can introduce excess LD at individual sites (Walsh 1988; Mansai and Innan 2010). Although we attempted to avoid the latter effects by restricting our analyses to sequence reads mapping to single sites, it remains possible that more divergent stretches of DNA occasionally engage in illegitimate conversion events.

One additional factor might contribute to anomalously high estimates of LD at very short distances—nonindependence of

base-call errors at closely spaced sites. The likelihood functions employed in mlRho assume that errors at different sites arise independently, which seems entirely reasonable for distant sites contained in nonoverlapping sequence reads. However, it is possible that some individual reads will be subject to higher error rates throughout, causing an overestimate of LD on the spatial scale of individual reads. Localized mismapping might generate similar problems. Should they exist, such artifacts would be a concern for any method for estimating LD and are not unique to the method introduced herein.

Finally, we note that a number of studies have sought to exploit information on the relationship between population-level LD and physical distance to infer past population-size changes (*e.g.*, Hayes *et al.* 2003; Tenesa *et al.* 2007; Qanbari *et al.* 2010). With an adherence to the assumption that the recombination rate between sites increases linearly with distance, *i.e.*, $r_d = cd$, the general idea underlying such studies is that the distance-scaled recombination parameter $\rho_d/d$ (assumed to equal $4N_ec$) reflects population-size conditions $\sim 1/(2cd)$ generations in the past. A number of concerns with these methods have been raised (Corbin *et al.* 2012; Park 2012; MacLeod *et al.* 2013; Sheehan *et al.* 2013).

A related approach to inferring the history of population-size changes, which like our method uses sequence information from just single individuals, relies on inferences on the homozygosity tract-length distribution, *i.e.*, the genome-wide distribution of distances between heterozygous sites (Li and Durbin 2011; Harris and Nielsen 2013). It has been argued that this method, which can also be extended to a sample of individuals (Sheehan *et al.* 2013), allows for the testing of specific demographic hypotheses, although again the assumption of a linear recombination rate–physical distance relationship is adhered to. A potential limitation of such analyses (not shared by the $\Delta$ estimator) is the issue of missing data, which break up homozygosity tracts (but simply reduce sample sizes used in $\Delta$ estimation).

Our observations raise significant questions about the use of both types of methods to infer the past demographic history of populations. A primary issue is that the relationship between LD and physical distance depends not just on past demographic changes but on the scaling features of recombination. With a nonlinear mapping of the recombination rate and physical distance apparently being typical up to at least 100 kb between sites, and the recombination rate per base pair being on the order of 10× greater for close than distant pairs of sites, it appears that any attempt to convert spatial patterns of LD to temporal patterns in $N_e$ needs to employ some sort of nonlinear timescale transformation.

Without such correction, and adhering to a simple reliance on crossover-based genetic maps, biases in estimates of $N_e$ of up to an order of magnitude may result. In effect, by simply extrapolating from crossover rates, the actual amount of recombination will be increasingly underestimated for more closely spaced sites. This will lead to overestimates in both the distance into the past that might be represented by such sites and the associated $N_e$, the exact pattern of bias depending on $x$ and $\bar{L}$. We will leave to a future study an evaluation of the power of $\Delta$ profiles to infer past population history, but simply reemphasize that substantially different patterns of demographic change may also sometimes lead to quite similar distance-dependent patterns of LD (Figure 2). A similar point has been made previously with respect to interpretations based on the site-frequency spectrum of polymorphisms (Myers *et al.* 2008).

## Acknowledgments

We thank A. Thota, S. Michael, and R. Henschel (University Information Technology Services, Indiana University) for computational assistance and M. Hahn for helpful comments. This work was support by National Science Foundation (NSF) grants EF-0827411 and DEB-1257806 and National Institutes of Health grant GM101672 to M.L. The computational analyses were supported in part by NSF grants CNS-0723054 and CNS-0521433, which support IU computational facilities. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF grant OCI-1053575.

## Literature Cited

Ahn, B. Y., and D. M. Livingston, 1986 Mitotic gene conversion lengths, coconversion patterns, and the incidence of reciprocal recombination in a *Saccharomyces cerevisiae* plasmid system. Mol. Cell. Biol. 6: 3685–3693.

Ahn, B. Y., K. J. Dornfeld, T. J. Fagrelius, and D. M. Livingston, 1988 Effect of limited homology on gene conversion in a *Saccharomyces cerevisiae* plasmid recombination system. Mol. Cell. Biol. 8: 2442–2448.

Andolfatto, P., and M. Nordborg, 1998 The effect of gene conversion on intralocus associations. Genetics 148: 1397–1399.

Andolfatto, P., and M. Przeworski, 2000 A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. Genetics 156: 257–268.

Chen, J. M., D. N. Cooper, N. Chuzhanova, C. Férec, and G. P. Patrinos, 2007 Gene conversion: mechanisms, evolution and human disease. Nat. Rev. Genet. 8: 762–775.

Corbin, L. J., A. Y. Liu, S. C. Bishop, and J. A. Woolliams, 2012 Estimation of historical effective population size using linkage disequilibria with marker data. J. Anim. Breed. Genet. 129: 257–270.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. 69: 831–843.

Hamblin, M. T., M. G. Salas Fernandez, A. M. Casa, S. E. Mitchell, A. H. Paterson *et al.*, 2005 Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolor*. Genetics 171: 1247–1256.

Harris, K., and R. Nielsen, 2013 Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 9: e1003521.

Haubold, B., P. Pfaffelhuber, and M. Lynch, 2010 mlRho – A program for estimating the population mutation and recombination

rates from shotgun-sequenced genomes. Mol. Ecol. 19(Suppl. 1): 277–284.
Hayes, B. J., P. M. Visscher, and H. C. McPartlan, and M. E. Goddard, 2003 Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 13: 635–643.
Hicks, W. M., M. Kim, and J. E. Haber, 2010 Increased mutagenesis and unique mutation signature associated with mitotic gene conversion. Science 329: 82–85.
Hill, W. G., 1975 Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor. Popul. Biol. 8: 117–126.
Hill, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genet. Res. 38: 209–216.
Hill, W. G., and B. S. Weir, 1988 Variances and covariances of squared linkage disequilibria in finite populations. Theor. Popul. Biol. 33: 54–78.
Hilliker, A. J., G. Harauz, A. G. Reaume, M. Gray, S. H. Clark *et al.*, 1994 Meiotic gene conversion tract length distribution within the *rosy* locus of *Drosophila melanogaster*. Genetics 137: 1019–1026.
Hudson, R. R., 1991 Gene genalogies and the coalescent process. Oxford Surv. Evol. Biol. 7: 1–44.
Hudson, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.
Judd, S. R., and T. D. Petes, 1988 Physical lengths of meiotic and mitotic gene conversion tracts in *Saccharomyces cerevisiae*. Genetics 118: 401–410.
Kim, S., V. Plagnol, T. T. Hu, C. Toomajian, R. M. Clark *et al.*, 2007 Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat. Genet. 39: 1151–1155.
Kimura, M., 1968 Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. 11: 247–269.
Kohl, K. P., and J. Sekelsky, 2013 Meiotic and mitotic recombination in meiosis. Genetics 194: 327–334.
Langley, C. H., B. P. Lazzaro, W. Phillips, E. Heikkinen, and J. M. Braverman, 2000 Linkage disequilibria and the site frequency spectra in the *su(s)* and *su($w^a$)* regions of the *Drosophila melanogaster* X chromosome. Genetics 156: 1837–1852.
Leffler, E. M., K. Bullaughey, D. R. Matute, W. K. Meyer, L. Ségurel *et al.*, 2012 Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol. 10(9): e1001388.
Li, H., and R. Durbin, 2011 Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.
Llorente, B., C. E. Smith, and L. S. Symington, 2008 Break-induced replication: what is it and what is it for? Cell Cycle 7: 859–864.
Lynch, M., 2007 *The Origins of Genome Architecture*, Sinauer Assocs., Inc., Sunderland, MA.
Lynch, M., 2008 Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. Mol. Biol. Evol. 25: 2409–2419.
MacLeod, I. M., D. M. Larkin, H. A. Lewin, B. J. Hayes, and M. E. Goddard, 2013 Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. Mol. Biol. Evol. 30: 2209–2223.
Malkova, A., and J. E. Haber, 2012 Mutations arising during repair of chromosome breaks. Annu. Rev. Genet. 46: 455–473.
Malkova, A., J. Swanson, M. German, J. H. McCusker, E. A. Housworth, and F. W. Stahl, and J. E. Haber., 2004 Gene conversion and crossing over along the 405-kb left arm of *Saccharomyces cerevisiae* chromosome VII. Genetics 168: 49–63.
Mancera, E., R. Bourgon, A. Brozzi, W. Huber, and L. M. Steinmetz, 2008 High-resolution mapping of meiotic crossovers and noncrossovers in yeast. Nature 454: 479–485.
Mansai, S. P., and H. Innan, 2010 The power of the methods for detecting interlocus gene conversion. Genetics 184: 517–527.
Maruki, T., and M. Lynch, 2014 Genome-wide estimation of linkage disequilibrium from population-level high-throughput sequencing data. Genetics (in press).
McGill, C. B., B. K. Shafer, D. R. Higgins, and J. N. Strathern, 1990 Analysis of interchromosomal mitotic recombination. Curr. Genet. 18: 29–39.
McVean, G. A., 2002 A genealogical interpretation of linkage disequilibrium. Genetics 162: 987–991.
Miller, D. E., S. Takeo, K. Nandanan, A. Paulson, M. M. Gogol *et al.*, 2012 A whole-chromosome analysis of meiotic recombination in *Drosophila melanogaster*. Genes, Genomes. Genetics 2: 249–260.
Morrell, P. L., D. M. Toleno, K. E. Lundy, and M. T. Clegg, 2006 Estimating the contribution of mutation, recombination and gene conversion in the generation of haplotypic diversity. Genetics 173: 1705–1723.
Myers, S., C. Fefferman, and N. Patterson, 2008 Can one learn history from the allelic spectrum? Theor. Popul. Biol. 73: 342–348.
Ohta, T., and M. Kimura, 1969 Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. Genetics 63: 229–238.
Ohta, T., and M. Kimura, 1971 Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics 68: 571–580.
Omilian, A. R., M. E. Cristescu, J. L. Dudycha, and M. Lynch, 2007 Ameiotic recombination in asexual lineages of *Daphnia*. Proc. Natl. Acad. Sci. USA 103: 18638–18643.
Padhukasahasram, B., and B. Rannala, 2013 Meiotic gene-conversion rate and tract length variation in the human genome. Eur. J. Hum. Genet. 2013: 1–8.
Paigen, K., J. P. Szatkiewicz, K. Sawyer, N. Leahy, E. D. Parvanov *et al.*, 2008 The recombinational anatomy of a mouse chromosome. PLoS Genet. 4(7): e1000119.
Park, L., 2012 Linkage disequilibrium decay and past population history in the human genome. PLoS ONE 7: e46603.
Pluzhnikov, A., and P. Donnelly, 1996 Optimal sequencing strategies for surveying molecular genetic diversity. Genetics 144: 1247–1262.
Preston, C. R., and W. R. Engels, 1996 P-element-induced male recombination and gene conversion in *Drosophila*. Genetics 144: 1611–1622.
Przeworski, M., and J. D. Wall, 2001 Why is there so little intragenic linkage disequilibrium in humans? Genet. Res. 77: 143–151.
Qanbari, S., M. Hansen, S. Weigend, R. Preisinger, and H. Simianer, 2010 Linkage disequilibrium reveals different demographic history in egg laying chickens. BMC Genet. 11: 103.
Rukść, A., P. L. Bell-Rogers, J. D. Smith, and M. D. Baker, 2008 Analysis of spontaneous gene conversion tracts within and between mammalian chromosomes. J. Mol. Biol. 377: 337–351.
Sabatti, C., and N. Risch, 2002 Homozygosity and linkage disequilibrium. Genetics 160: 1707–1719.
Santoyo, G., and D. Romero, 2005 Gene conversion and concerted evolution in bacterial genomes. FEMS Microbiol. Rev. 29: 169–183.
Schrider, D. R., J. N. Hourmozdi, and M. W. Hahn, 2011 Pervasive multinucleotide mutational events in eukaryotes. Curr. Biol. 21: 1051–1054.
Sheehan, S., K. Harris, and Y. S. Song, 2013 Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics 194: 647–662.

Slatkin, M., 2008 Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet. 9: 477–485.

Song, Y. S., and J. S. Song, 2007 Analytic computation of the expectation of the linkage disequilibrium coefficient $r^2$. Theor. Popul. Biol. 71: 49–60.

Strobeck, C., and K. Morgan, 1978 The effect of intragenic recombination on the number of alleles in a finite population. Genetics 88: 829–844.

Stumpf, M. P., and G. A. McVean, 2003 Estimating recombination rates from population-genetic data. Nat. Rev. Genet. 4: 959–968.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17: 520–526.

Tucker, A., M. Ackerman, B. Eads, S. Xu, and M. Lynch, 2013 Population-genomic insights into the evolutionary origin and fate of obligately asexual *Daphnia pulex*. Proc. Natl. Acad. Sci. USA 110: 15740–15745.

Walsh, J. B., 1988 Unusual behaviour of linkage disequilibrium in two-locus gene conversion models. Genet. Res. 51: 55–58.

Wiuf, C., and J. Hein, 2000 The coalescent with gene conversion. Genetics 155: 451–462.

Xu, S., A. R. Omilian, and M. E. Cristescu, 2011 High rate of large-scale hemizygous deletions in asexually propagating *Daphnia:* implications for the evolution of sex. Mol. Biol. Evol. 28: 335–342.

Yang, S., Y. Yuan, L. Wang, J. Li, W. Wang *et al.*, 2012 Great majority of recombination events in *Arabidopsis* are gene conversion events. Proc. Natl. Acad. Sci. USA 109: 20992–20997.

Yin, J., M. I. Jordan, and Y. S. Song, 2009 Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data. Bioinformatics 25: i231–i239.

*Communicating editor: Y. S. Song*

GENETICS

**Supporting Information**
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.166843/-/DC1

# Genome-Wide Linkage-Disequilibrium Profiles from Single Individuals

**Michael Lynch, Sen Xu, Takahiro Maruki, Xiaoqian Jiang, Peter Pfaffelhuber, and Bernhard Haubold**


DOI: 10.1534/genetics.114.166843

**File S1**

**SUPPLEMENTAL MATERIAL**

**Analysis of data**. Because the interpretation of LD measures in terms of conventional population-genetic parameters relies on the assumption of drift-mutation-recombination equilibrium, we restricted our analyses to genomes from single diploid, non-intentionally inbred individuals. The raw reads (i.e., Sanger reads or Illumina short reads) for each genome were downloaded from the NCBI Short Read Archive. The latest repeat-masked reference assembly for each genome was downloaded from NCBI or Ensembl. The Sanger reads were trimmed based on sequence qualities using the program LUCY with default settings (Chou and Holmes 2001). The software BWA 0.7.5 (Li and Durbin 2009; Li and Durbin 2010) was used to map the Sanger raw reads against the reference assembly (bwasw command with default settings) and to align the Illumina short reads (sampe commands with default settings).

To eliminate potential problems that can arise from the mapping process, we applied three strict criteria for selecting sites that go into the final analyses. First, we removed all raw reads that mapped to multiple locations of the genome (which represent, for example, potential duplicate genes and transposons) as well as reads with a mapping quality score lower than 25. Second, sites located within paralogous genes with >90% sequence identity were excluded. Third, our analyses included only sites with coverages >4× but less than twice the genome-wide average coverage.

The software Samtools (Li et al. 2009) was used to generate a pileup of the mapped raw reads for every possible position in the genome. The pileup was then converted to a quartet profile using the software sam2pro (http://guanine.evolbio.mpg.de/mlRho/). For each site, the quartet profile describes the numbers of the four different nucleotide reads observed ($n_A$, $n_C$, $n_G$, $n_T$), with their sum ($n_A + n_C + n_G + n_T$) representing the coverage at each site.

The program mlRho 2.0 (Haubold et al. 2010) was used to calculate the maximum-likelihood estimates of the zygosity correlation coefficient ($\Delta$), genome-wide heterozygosity ($\theta$), and sequencing error rate ($\varepsilon$). The latter two parameters were estimated using all the available sites in the genomes and were subsequently used in the one-dimensional estimation of $\Delta$. We calculated $\Delta$ for all pairs of sites separated from 1 to 100,000 bp, with an incremental increase in window size for pooled analyses. For distances up to 1000 bp, $\Delta$ was analyzed for 1-bp distance increment (e.g., $\Delta$ was estimated for all pairs of sites that are 1-bp apart, then 2-bp, 3-bp, etc.), whereas for the distances between 1 and 10 kbp, these parameters were estimated with increments of 10 bp, pooling all site pairs within each 10-bp increment; and between 10 and 100 kbp, a window size of 100 bp was used.

**Literature Cited**

Chou, H. H., and M. H. Holmes. 2001. DNA sequence quality trimming and vector removal. Bioinformatics 17: 1093-1104.

Haubold, B., P. Pfaffelhuber, and M. Lynch. 2010. mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. Mol. Ecol. 19: 277-284.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078-2079.

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

**Table S1** Summary of single diploid genomes analyzed in this study. Theta: maximum-likelihood estimate of heterozygosity in a single genome; epsilon: maximum-likelihood estimate of sequencing error rate. All raw reads were downloaded from NCBI Short Read Archive.

| **Species** | **Coverage** | **No. of sites ($10^7$)** | **Theta** | **Epsilon** | **Reference** | **DNA source** |
|---|---|---|---|---|---|---|
| *Ailuropoda melanoleuca* | 26 | 10.798 | 0.00318 | 0.0150 | Li *et al.* (2010) | a single non-inbred female |
| *Anolis carolinensis* | 8 | 106.446 | 0.00260 | 0.0034 | Alfoldi *et al.* (2011) | A single female |
| *Canis familiaris* | 10 | 120.495 | 0.00094 | 0.0033 | Lindblad-Toh *et al.* 2005 | A female Boxer |
| Denisovan | 28 | 165.960 | 0.00042 | 0.0014 | Meyer *et al.* 2012 | DNA from phalanx of a Denisovan individual |
| *Fugu rubripes* | 6 | 7.535 | 0.00320 | 0.0014 | Aparicio *et al.* 2002 | A single fish |
| *Gorilla gorilla* | 32 | 122.182 | 0.00314 | 0.0090 | Scally *et al.* 2012 | A single female |
| *Homo sapiens* (African) | 18 | 162.557 | 0.00357 | 0.0066 | Schuster *et al.* 2010 | A single male (KB1) |
| *Homo sapiens* (Chinese) | 31 | 152.247 | 0.00164 | 0.0119 | Wang *et al.* 2008 | A single male |
| *Homo sapiens* (Watson) | 7 | 147.303 | 0.00103 | 0.0004 | Wheeler et al. 2008 | A single male |
| *Loxodonta africana* | 7 | 98.194 | 0.00134 | 0.0020 | http://www.broadinstitute.org/ | A single individual |
| *Macaca mulatta* | 8 | 81.724 | 0.00261 | 0.0033 | Gibbs *et al.* 2007 | A single female |
| *Ornithorhynchus anatinus* | 10 | 59.061 | 0.00131 | 0.0033 | Warren *et al.* 2008 | A single female |
| *Pan troglodytes* | 8 | 88.106 | 0.00101 | 0.0042 | Chimpanzee Sequencing and Analysis Consortium 2005 | A single male |
| *Petromyzon marinus* | 12 | 22.753 | 0.00443 | 0.0031 | Smith *et al.* 2013 | A single female |
| *Pongo abelii* | 4 | 27.204 | 0.00527 | 0.0087 | Locke *et al.* 2011 | A single female (SB550) |
| *Pongo pygmaeus* | 6 | 56.747 | 0.00272 | 0.0044 | Locke *et al.* 2011 | A single male (KB4204) |

**Literature Cited**


Chimpanzee Sequencing and Analysis Consortium, 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437**:** 69-87.

Alfoldi, J., F. Di Palma, M. Grabherr, C. Williams, L. Kong *et al.*, 2011 The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature 477**:** 587-591.

Aparicio, S., J. Chapman, E. Stupka, N. Putnam, J. Chia *et al.*, 2002 Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297**:** 1301-1310.

Gibbs, R. A., J. Rogers, M. G. Katze, R. Bumgarner, G. M. Weinstock *et al.*, 2007 Evolutionary and biomedical insights from the rhesus macaque genome. Science 316**:** 222-234.

Li, R., W. Fan, G. Tian, H. Zhu, L. He *et al.*, 2010 The sequence and de novo assembly of the giant panda genome. Nature 463**:** 311-317.

Lindblad-Toh, K., C. M. Wade, T. S. Mikkelsen, E. K. Karlsson, D. B. Jaffe *et al.*, 2005 Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438**:** 803-819.

Locke, D. P., L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth *et al.*, 2011 Comparative and demographic analysis of orang-utan genomes. Nature 469**:** 529-533.

Meyer, M., M. Kircher, M.-T. Gansauge, H. Li, F. Racimo *et al.*, 2012 A high-coverage genome sequence from an archaic Denisovan individual. Science 338**:** 222-226.

Scally, A., J. Y. Dutheil, L. W. Hillier, G. E. Jordan, I. Goodhead *et al.*, 2012 Insights into hominid evolution from the gorilla genome sequence. Nature 483**:** 169-175.

Schuster, S. C., W. Miller, A. Ratan, L. P. Tomsho, B. Giardine *et al.*, 2010 Complete Khoisan and Bantu genomes from southern Africa. Nature 463**:** 943-947.

Smith, J. J., S. Kuraku, C. Holt, T. Sauka-Spengler, N. Jiang *et al.*, 2013 Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. Nat. Genet. 45**:** 415-421.

Wang, J., W. Wang, R. Q. Li, Y. R. Li, G. Tian *et al.*, 2008 The diploid genome sequence of an Asian individual. Nature 456**:** 60-65.

Warren, W. C., L. W. Hillier, J. a. M. Graves, E. Birney, C. P. Ponting *et al.*, 2008 Genome analysis of the platypus reveals unique signatures of evolution. Nature 453**:** 175-183.

Wheeler, D. A., M. Srinivasan, M. Egholm, Y. Shen, L. Chen *et al.*, 2008 The complete genome of an individual by massively parallel DNA sequencing. Nature 452**:** 872-875.