The Share of Human Genomic DNA under Selection Estimated from Human–Mouse Genomic Alignments

F. CHIAROMONTE,* R.J. WEBER,† K.M. ROSKIN,† M. DIEKHANS,† W.J. KENT,† AND D. HAUSSLER‡

** Department of Statistics and Department of Health Evaluation Sciences, Pennsylvania State University, University Park, Pennsylvania 16803; †Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064; ‡Howard Hughes Medical Institute, University of California, Santa Cruz, California 95064*

Draft sequences covering most euchromatic parts have recently become available for two mammalian genomes, human (Lander et al. 2001; Venter et al. 2001) and mouse (Waterston et al. 2002). This raises the possibility of using comparative genomics to estimate what fraction of the human genome evolves under purifying selection. Lacking genomes of other mammals, this comparative exercise is still in its preliminary stages. However, a rough estimate has been made that ~5% of the human genome is in short segments that appear to be under selection based on comparison with mouse (Waterston et al. 2002). Here, as a basis for future refinements, we present the computational strategy that led to this estimate, providing details on scoring functions, data preparation, and statistical techniques. We also describe stability analyses, control experiments, and tests for the effects of artifacts that were performed to establish robustness of our results, and discuss possible alternate interpretations.

Our strategy hinges on three elements: (1) the construction of various collections of short aligned windows of the human genome (e.g., 50 bp)—in particular, a large collection of such windows that are very likely to have evolved neutrally since the divergence of human and mouse ("ancestral repeats," relics of transposons that were present in the genome of our common ancestor with mouse); (2) the development of a score function quantifying conservation in short aligned windows, and providing a satisfactory "template" for neutral behavior when computed on windows in ancestral repeats; and (3) statistical techniques to estimate and compare the score distributions for genome-wide and ancestral repeat windows, and thus infer an upper bound on the share of genome-wide windows that are compatible with the neutral template. The remaining share of the genome is populated by windows that are too conserved to be modeled by the neutral template, and hence are either evolving under purifying selection, or are evolving neutrally but are experiencing fewer substitutions than nearby windows in ancestral repeats for some unknown reasons.

Because ancestral transposons have been inactive since their insertion in the genome of the common ancestor of human and mouse, they are one type of human DNA that is most likely to have evolved free of any selective pressure. The rate of substitution in these sites between human and mouse is similar to, but slightly less than, that observed in fourfold degenerate sites from codons, and covaries regionally with that rate (Waterston et al. 2002; Hardison et al. 2003). This suggests that both of these types of sites provide reasonable models to evaluate the rate of neutral substitution, and that this rate depends on some local properties of the chromosome where it is measured, as was known from previous studies on the effects of GC content on substitution rates (Bernardi 1993). Because ancestral repeats constitute 22% of the human genome and are still reliably alignable to mouse, they allow us to construct a very large number of short aligned windows of neutrally evolving DNA (Waterston et al. 2002; Schwartz et al. 2003).

There are many ways to measure conservation in short aligned windows, even with just two species. The aim here is to provide a simple template for neutral behavior that allows, in comparison, a satisfactory separation of aligned sequence that is undergoing purifying selection. To this end, we further explore the *normalized percent identity* score introduced in Roskin et al. (2002), and used in Waterston et al. (2002). Definitions of several variants of this score, and a brief discussion of the first crude estimates of the share under selection, which were made with one of these variants, can be found in Roskin et al. (2002, 2003). Yet more possible scores are analyzed in Elnitski et al. (2003), with a particular focus on separating regulatory elements from neutral DNA.

The normalized percent identity score involves no assumptions on the characterization of DNA functions that might be under purifying selection, except that they result in a higher degree of conservation. It has a straightforward definition, and the advantage of being relatively easy to compute. The score is obtained by calculating the fraction of aligned bases in the window that are identical between human and mouse, and then subtracting a mean and dividing by a standard deviation estimated under neutrality. The only subtlety comes from estimating the neutral mean and standard deviation. The neutral mean for a window is estimated locally, using only aligned bases from ancestral repeats in a region surrounding the window. Local estimation of the neutral mean percent identity allows the conservation score to compensate for regional variations in the rate of neutral evolution (Roskin et. al. 2002, 2003; Waterston et al. 2002; Hardison et al. 2003). This includes variations induced by changes in GC content and other features. The standard deviation estimate is derived from the mean estimate using a simple binomial model, and thus is also local.

We make no parametric assumptions in estimating the normalized percent identity score distribution for either genome-wide or ancestral repeat windows. With approximately two million data points in the smaller data set of ancestral repeat windows, there is no need for such assumptions. However, we do use Gaussian kernel smoothing to estimate a continuous nonparametric score distribution from these empirical data. We decompose the continuous genome-wide distribution as a mixture of a neutral component and a component that appears to be under selection.

METHODS

Data preparation. Our collections of short aligned windows were constructed using a fixed grid of locations along the human sequence. The grid is such as to always guarantee nonoverlapping windows for the sizes we consider. For a given window size (W) and alignment filtering threshold (T), the genome-wide collection is constructed first extending windows of W bases at each location, and then discarding all windows with less than T bases aligned with mouse. For the same window size and filtering threshold, the collection of windows relative to a particular feature type (ancestral repeats, coding regions) is constructed in a similar fashion, first extending windows of size W at grid locations, and then discarding windows whose overlap with aligned features of that type is less than T bases. Table 1 gives coverage provided by genome-wide windows for the $W = 50$, $T = 40$ case presented in our main analysis, as well as other combinations of window size and filtering threshold.

Ancestral repeats were repeats identified by RepeatMasker (available at http://ftp.genome.washington.edu/RM/RepeatMasker.html; Smit and Green 1999) and present at orthologous sites. A list of specific families of ancestral repeats is given in the Methods web-available compendium to Waterston et al. (2002).

Known coding region annotation was obtained by aligning the RefSeq (Pruitt and Maglott 2001) human mRNAs from GenBank release 130.0 to the human genome with BLAT (Kent 2002; Kent et al. 2002). We selected annotations that had an aligned mouse position and met the following criteria: (1) CDS appeared complete in both human and mouse, beginning with a start codon, and ending with a stop codon. The mouse stop codon was allowed up to 20 codons before the human stop codon. (2) There were no in-frame stop codons. (3) Introns in human CDS had splice sites in the form *GT..AG*, *GC..AG*, or *AT..AC*. This resulted in 11,718 gene alignments.

Further details on data preparation can be found in the Methods web-available compendium to Waterston et al. (2002), and in Schwartz et al. (2003).

Eliminating pseudogenes. The initial BLASTZ alignment contained numerous processed and nonprocessed pseudogenes that could artificially inflate our estimate of the share under selection. To remove these pseudogenes, we apply a filter that only keeps each reciprocal best pair of alignments between human and mouse: If a segment of mouse sequence aligns to multiple human genome locations, we only keep the region that aligns back to that same region in mouse and gives the highest alignment score. Pseudogenes are clearly under different selective pressure than the genes they are duplicated from, so they should not align as well in both directions as the genes themselves. Applying this filter removes ~14% of the initial alignment, and whereas the initial alignment covers 89% of RefSeq genes, the filtered one only covers 83%. Therefore, our filter errs on the side of caution, likely removing more highly conserved sequence than needed to eliminate pseudogenes' effects, but this is acceptable in an attempt to produce a conservative, lower bound estimate of the share under selection. In other experiments, we used the chaining method described in Kent et al. (2003) in place of this reciprocal best filtering method and obtained similar results, with slightly higher estimates of the share under selection (not shown in this paper).

Normalized percent identity. The normalization presented in Equation 1 centers the fraction of aligned bases in a window ($m(w)$) by an estimated regional expectation under neutrality (m_o), given by the average fraction of identical aligned base pairs in ancestral repeats in a region surrounding the window, but not containing it. The region is chosen to contain $K = 6{,}000$ aligned bases that are believed not to be under selective pressure, including those in the window itself (for instance, when creating the neighborhood of an ancestral repeat window of size $W = 50$ with at least $T = 40$ aligned bases, this corresponds to between 5,950 and 5,960 bases once the window itself is removed). The average size of the regions constructed in this way is 379,079 bp. The parameter 6,000 was chosen to reduce the variance among normalized scores of ancestral repeat windows. The results are not very sensitive to this parameter: For instance, using $K = 600$ leads to an estimate of 5.11% for the share under selection, $K = 3{,}000$ gives 5.19%, and $K = 12{,}000$ gives 5.08%. As K grows, the estimated local mean m_o approaches the global mean. In the limit, for infinitely large K, we obtain an estimate 4.84%. This shows that we apparently do lose a bit in the estimate of the share under selection if we do not try to account for local evolutionary rate variation, but the numbers we obtain are still in the same ballpark.

Gaussian kernel density estimation. Gaussian kernel smoothing (see Eq. 2) was implemented using the *R* language (Ihaka and Gentlman 1996) routine *density(x, n, window, bw, na.rm=T, from, to)* where

- x is the vector of observations (e.g., the vector of S-scores for 50-bp ancestral repeat WA-windows for estimating the neutral density, and that of 50-bp genome-wide WA-windows for estimating the genome-wide density).
- n determines the number of equispaced abscissa values between *from* and *to* on which the smooth curve ordinate values are computed. We fixed the same n (10,000) *from* and *to* (minimum and maximum observed scores for genome-wide windows) for all estimations, to have density values on exactly the same abscissa grid.
- *window* determines the type of kernel to be employed. We used "g" for Gaussian.

- *bw* is the parameter defining the degree of smoothing. We used $bw = 0.5$ (which according to the routine specifications corresponds to a Gaussian kernel standard deviation of 0.5) when considering 50-,100-, and 200-bp WA-windows, and $bw = 0.75$ (kernel standard deviation = 0.75) when considering 30-bp WA-windows.

The *na.rm = T* is a technical argument ensuring that missing values, if any, be discarded from the calculation.

Mixture estimation. Based on the upper bound expressed by Equation 5, we approximate the neutral weight from above using the empirical minimum of a ratio: $p_o = \min_S[f_{genome}(S)/f_{neutral}(S)]$. Under-smoothing in the density estimates may translate in ragged fluctuations of this ratio, especially for extreme values of S where very few observations are available and thus both densities are very close to 0. These fluctuations complicate a reliable assessment of the empirical minimum. This problem, whose potential effect was evaluated through some control experiments, can be satisfactorily mitigated by selecting an appropriate degree of smoothing in the Gaussian kernel procedure, and implementing an additional "trimming" procedure for ratio fluctuations on extreme S values.

Trimming the neutral density. The small fluctuations in the estimated neutral density cause $p_o\, f_{neutral}(S) > f_{genome}(S)$ for some values of S, but according to the mixture in Equation 3, this cannot happen. As a consequence, the estimate for p_o must be decreased until all the fluctuations are below the genome-wide density. This causes $(1 - p_o)\, f_{genome}(S)$ to increase and makes some known neutral windows appear selected. Alternatively, we can explicitly model the error in the neutral density so

$$f_{neutral}(x) = f^*_{neutral}(x) + \varepsilon$$

where $f_{neutral}(S)$ is the density estimated from the data, $f^*_{neutral}(S)$ is the true neutral density, and ε is a positive constant error term. The amount of trimming is set to $\alpha = 0.01$ where $\int \varepsilon\, dx \le \alpha$ and therefore $\iint f^*_{neutral}\, dx > 1 - \alpha$. With this error term the estimate of p_o becomes

$$p_o = \frac{f_{genome}(x)}{\max(0, f_{neutral}(x) - \varepsilon)}$$

Even this simple constant error model has a dramatic effect in reducing the number of neutral windows incorrectly labeled selected, as results of the control experiment described below illustrate (see Fig. 5).

Probability of selection estimation. The equality in Equation 5 is derived from the mixture in Equation 3 as follows:

$$\begin{aligned}\Pr(w \text{ selected} | S(w) = S) &= 1 - \Pr(w \text{ neutral} | S(w) = S) \\ &= 1 - \frac{\Pr(w \text{ neutral} \cap S(w) = S)}{\Pr(S(w) = S)} \\ &= 1 - \frac{\Pr(w \text{ neutral}) \Pr(S(w) = S | w \text{ neutral})}{f_{genome}(S)} \\ &= 1 - p_o \frac{f_{neutral}(S)}{f_{genome}(S)}\end{aligned}$$

Thus, the probability of selection as a function of normalized percent identity is conservatively estimated with the curve $1 - p_o\, [f_{neutral}(S)/f_{genome}(S)]$.

RESULTS

Main Analysis

Our main analysis uses the collection of all 50-bp nonoverlapping windows of the human genome with at least 40 bases aligned to mouse, referred to as "well-aligned windows" or *WA-windows* below, plus the subset of these windows within aligned ancestral repeat sequence (see Methods, and Table 1 for coverage statistics). The average numbers of bases aligned in these WA-windows are 47.5 and 46.94, respectively. To score a window w, we compute the fraction $m(w)$ of aligned base pairs in w that are identical between human and mouse and subtract from it an estimated regional expectation under neutrality, m_o. This estimate is the average fraction of identical aligned base pairs in $K = 6{,}000$ aligned ancestral repeat sites in a region surrounding the window w, a regional size of about 400 kb, determined so as to optimize a tradeoff between sample variance in the estimate of m_o and regional fluctuations in m_o. The results are not greatly sensitive to the choice of K (see Methods). We then rescale $(m(w) - m_o)$ to take into account differences in fluctuation magnitude due to m_o and to the number of aligned positions in the window, $n(w)$. This results in the normalized percent identity score

$$S(w) = \frac{(m(w)\, n(w) - m_o\, n(w))}{\sqrt{m_o(1 - m_o) n(w)}} = \sqrt{\frac{n(w)}{m_o(1 - m_o)}}\,(m(w) - m_o) \qquad (1)$$

As shown in the red curve in Figure 1, for ancestral repeat WA-windows, the empirical distribution of S is tight and symmetric about 0 (mean = – 0.119, S.D. = 1.208, median = – 0.126). It is bell-shaped, but its tails are too heavy for a Gaussian. On the other hand, for genome-wide WA-windows (blue curve in Fig. 1), the empirical distribution is broader and asymmetric, with a heavier right tail (mean = 0.367, S.D. = 1.541, median = 0.239).

Table 1. Estimates of the Share of the Human Genome under Selection for Different Window Sizes (W) and Required Number of Aligned Bases (T)

W	T	$p_1 = (1 - p_O)$	Coverage	α_{sel} (%)
30	20	0.15	846472K (30.4%)	4.51
	25	0.17	743308K (26.7%)	4.50
	30	0.23	439501K (15.8%)	3.65
50	40	0.19	756051K (27.1%)	5.19
	45	0.22	623286K (22.4%)	4.90
	50	0.31	292506K (10.5%)	3.31
100	80	0.23	739836K (26.6%)	6.15
	90	0.29	550530K (19.8%)	5.8
	100	0.52	122437K (4.4%)	2.29
200	160	0.31	708701K (25.4%)	7.92
	180	0.40	467954K (16.8%)	6.68
	200	0.81	328668K (1.2%)	0.96

The table reports the estimated mixture coefficient for the selected component, $p_1 = 1 - p_0$, coverage of the human genome (in terms of number of bases and percentage), and estimated share of the genome contained in windows under selection, α_{sel}.

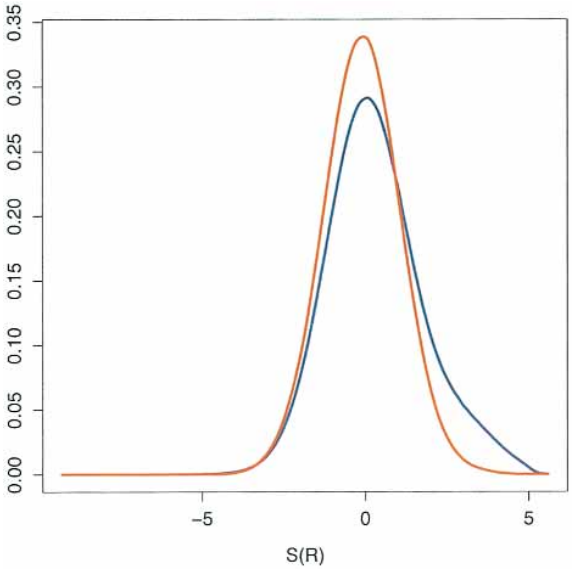

Figure 1. Smoothed densities of normalized percent identity for ancestral repeats and genome-wide WA-windows (50 bp, at least 40 aligned). $f_{neutral}(S)$ and $f_{genome}(S)$ are depicted in red and blue, respectively. They are obtained through Gaussian kernel smoothing, a technique that employs the convolution of a Gaussian density with the discrete distribution placing equal mass on each observed value.

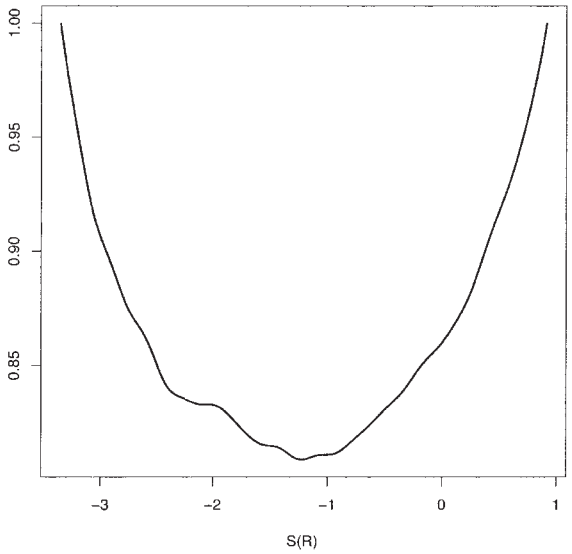

Figure 2. Ratio between the smoothed densities of normalized percent identity for genome-wide and ancestral repeat WA-windows (50 bp, at least 40 aligned). The minimum of this curve, $p_o = 0.808$, estimates an upper bound for the neutral weight in the mixture (i.e., the share of genome-wide windows compatible with the neutral template provided by ancestral repeats).

We employed Gaussian kernel smoothers to produce the estimated density functions $f_{neutral}(S)$ and $f_{genome}(S)$ depicted by the blue and red curves in Figure 1. A Gaussian kernel smoother (Wegman 1972; Silverman 1986) estimates the density of a variable X, for which observations $\{x_1,...x_N\}$ are available, by convolving the density of a normal $N(0,\sigma^2)$ with a distribution placing mass $1/N$ on each observed value:

$$f(X) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(X-x_i)^2}{2\sigma^2}\right\} \quad (2)$$

We decompose the distribution of S for genome-wide WA windows as a mixture of a neutral component (the score distribution for WA-windows in ancestral repeats) and a component that appears to be under selection, with weights p_o, and $p_1 = (1 - p_o)$, respectively:

$$f_{genome}(S) = p_o f_{neutral}(S) + (1 - p_o) f_{selected}(S) \quad (3)$$

(For background on mixtures, see Lindsay 1995; McLachlan and Peel 2000; for an approach similar to the one used here, see Efron et al. 2001.) Thus, a WA-window is assumed to be neutral (have conservation consistent with $f_{neutral}$) with probability p_o, and undergoing selection (have conservation consistent with $f_{selected}$) with probability $(1 - p_o)$.

We have estimated $f_{neutral}$ and f_{genome} from our data, and will use Equation 3 to estimate p_o, which will then determine $f_{selected}$. Although the parameter p_o is not univocally determined by Equation 3, non-negativity of densities implies that

$$p_o \le \frac{f_{genome}(S)}{f_{neutral}(S)} \quad (4)$$

for all scores S. Thus, we estimate an *upper bound* to the neutral weight as $p_o = min_S[f_{genome}(S)/f_{neutral}(S)]$, which gives a value of 0.808 . This is illustrated in Figure 2. (In practice, additional steps are taken to ensure that inaccuracies in the estimated density ratio $f_{genome}(S)/f_{neutral}(S)$ do not affect the result; see Control Experiments below and Methods.) Figure 3 summarizes the corresponding "conservative" mixture decomposition: the blue curve de-

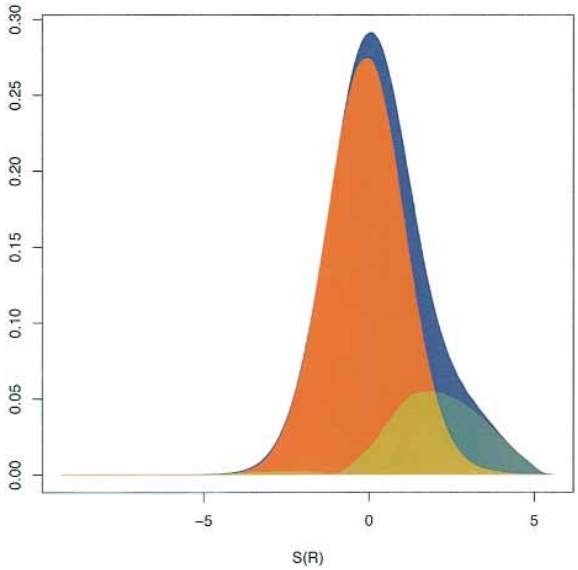

Figure 3. Mixture decomposition of the distribution of normalized percent identity for genome-wide WA-windows (50 bp, at least 40 aligned) into a neutral component and a component under selection. This is a "conservative" decomposition that uses the estimated upper bound p_o. The blue curve depicts $f_{genome}(S)$, the red curve depicts $p_o f_{neutral}(S)$, and the green curve depicts the difference $f_{genome}(S) - p_o f_{neutral}(S) = (1 - p_o) f_{selected}(S)$.

picts $f_{genome}(S)$, the red curve depicts $p_o f_{neutral}(S)$, and the green curve depicts the difference $f_{genome}(S) - p_o f_{neutral}(S) = (1 - p_o) f_{selected}(S)$ (the estimated score distribution for WA-windows under selection, rescaled by its weight). Note that no parametric assumptions are used in this decomposition. The density of the scores in the selected component captures the empirical structure of all observed conservation levels in 50-base windows beyond those that can be explained by the neutral model; we don't assume that the amount of "selection" follows any particular parametric model.

This calculation suggests that, at most, 80.8% of the genome-wide WA-windows are consistent with neutral evolution, with the remainder (at least 19.2%) appearing to be under selection, or neutral but accumulating substitutions at a slower rate than those in ancestral repeats. Because ~27.1% of all human bases are covered by WA-windows, under the additional conservative assumption that no regions outside these well-aligned windows are under selection, this result implies that a fraction $a_{selected}$ of at least 0.192*0.271 = 0.0520 (about 5.2%) of the human genome is contained in 50-bp windows that appear to be under selection by this test.

Window Size and Alignment Threshold: Separating Selected and Neutral Behaviors

Fully investigating stability of the above results with respect to different choices of alignment and score functions is beyond the scope of this paper. However, we note that very similar results were obtained on another alignment, using a somewhat different score function (taking into account base composition and adjacent bases' effects on neutral evolution), and a cruder mixture modeling method (Roskin et al. 2002, 2003). In addition to the fact that ancestral repeats accumulate substitutions slower than fourfold degenerate sites (Waterston et al. 2002; Hardison et al. 2003), this is further evidence against the hypothesis that DNA in ancestral repeats accumulates substitutions faster than other types of neutral DNA, perhaps due to some lingering base-compositional property of the ancient relics, and hence that the analysis above overestimates the share of the genome under selection. We discuss other tests for this type of "biased neutral model" effect below. However, to get a general feel for the stability of the results, we first investigate the effect of window size (W) and threshold number of aligned bases (T) on our "conservative" estimate of the mixture coefficient p_o, and subsequent lower bound estimate of the share under selection.

Outcomes for various choices of W and T are reported in Table 1. The estimated fraction of WA-windows under selection increases with increasing window size and required number of aligned bases, while the total fraction of the genome covered by WA-windows decreases. The variation in the estimated share under selection, $a_{selected}$, reflects a tradeoff between these two effects.

As the window size and/or the alignment threshold decrease, neutral and genome-wide distributions of normalized percent identity become more similar, making it more difficult to statistically separate neutral and selected components. This is reflected in the results given in Table 1. When the neutral and selected distributions are highly overlapping, and thus the neutral and genome-wide distributions more similar, the lower bound we produce is very weak, which in turn leaves room for gross underestimation of the apparent share under selection (in the extreme case of two identical score distributions for neutral and selected windows, our conservative estimation of p_o would be 1, and thus our lower bound estimate of the share under selection 0, although the actual neutral weight and share under selection could be anywhere between 0 and 1). This effect becomes more severe for smaller window sizes; the smaller the size, the less neutral and selected windows separate in terms of normalized percent identity.

To see this, we considered windows we have good reason to believe are under selection; namely, windows entirely contained in the coding regions of known genes in the RefSeq database (Pruitt and Maglott 2001). For window sizes W = 30, 50, 100, and 200 bp, we set the alignment threshold to T = 25, 40, 80, and 160, respectively, and compared the score distribution for well-aligned coding windows (*WAC-windows*) to the distribution for neutral WA-windows, i.e., WA-windows in ancestral repeats. We found a substantial overlap for 30-bp windows, but much less overlap for windows of 50 bp and larger (Fig. 4).

Using the mixture decomposition for a fixed window size, say W = 50 bp, we can estimate the probability that a generic 50-bp window w is under selection given its normalized percent identity:

$$\Pr(w \text{ selected } | S(w) = S) = 1 - p_o \frac{f_{neutral}(S)}{f_{genome}(S)} \quad (5)$$

For any collection C containing N 50-bp windows, we

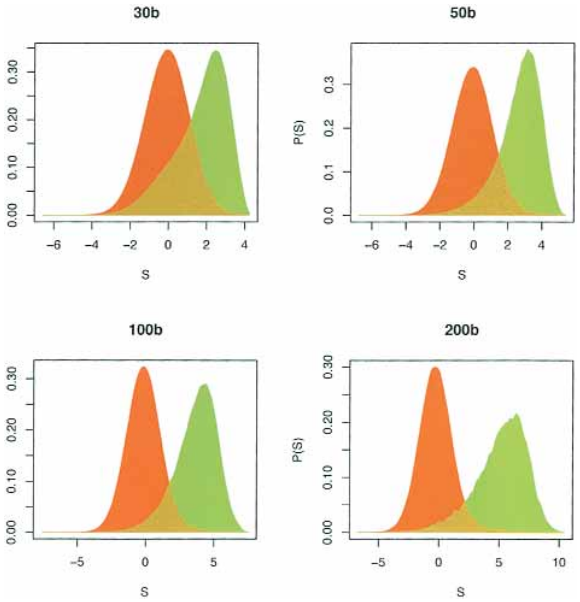

Figure 4. Gaussian Kernel smoothing of normalized percent identity distributions for WAC (well-aligned coding) windows (*green*) and WA-windows in ancestral repeats (red), for window sizes 30 bp (alignment threshold = 25), 50 bp (alignment threshold = 40), 100 bp (alignment threshold = 80), and 200 bp (alignment threshold = 160).

can use this formula to calculate the expected fraction of windows in C that are under selection as

$$\frac{1}{N}\sum_{w\in C}\Pr(w \text{ selected}|S(w)) \quad (6)$$

If, for example, we apply Equation 6 with C defined as all $W = 50$-bp WA-windows (alignment threshold $T = 40$), we recover the mixture coefficient $p_1 = 0.192$ discussed above, because p_1 is the fraction of these WA-windows that are estimated by the mixture decomposition to be under selection, and this must be the same as the expected fraction of windows under selection. Here Equation 6 merely provides another way of calculating the same number, and hence a nice test for our software. However, if we apply Equation 6 with C defined as $W = 50$-bp WAC-windows, then we can calculate something more interesting; namely, the expected fraction of well-aligned coding windows that are under selection. We performed this calculation for various window sizes.

For 200-bp windows we obtained 86%, for 100-bp windows 78%, for 50-bp windows 65%, but for 30-bp windows, we obtained only 48%. This further indicates how our mixture decomposition method produces a very conservative lower bound for the share under selection when applied to the normalized percent identity distribution of small windows.

A Tighter Lower Bound: Splitting Well-aligned Windows

Our computational strategy requires enough separation between the neutral and selected distribution of normalized percent identity for the mixture to reliably detect the difference. In fact, the definition of well-aligned windows ($T = 25$ for $W = 30$ bp, $T = 40$ for $W = 50$ bp, $T = 80$ for $W = 100$ bp, $T = 160$ for $W = 200$ bp) and choice of window size for the main analysis ($W = 50$ bp) stemmed from separation considerations; see also the Discussion section below. However, if ancestral transposon relics are a good neutral model, our figure of 5.2% may still represent a fairly conservative lower bound for the share under selection. As a means to tighten this lower bound, we can further isolate extremely well-aligned genome-wide and neutral windows, splitting WA-windows into a high and a low alignment range. We tried, respectively, 20–24 and 25–30 aligned bases for $W = 30$, 40–44 and 45–50 for $W = 50$, 80–94, and 95–100 for $W = 100$, and 160–194 and 195–200 for $W = 200$.

We repeated our calculations (estimating smooth densities for neutral and genome-wide scores, decomposing the genome-wide score distribution into a neutral and a selected component, computing a share under selection based on the mixture weight estimate and coverage) separately for high- and low-range WA-windows, and added the results. As shown in Table 2, this consistently produces slightly higher share figures.

The reason for the tighter lower bound is that neutral and genome-wide normalized percent identity distributions are more dissimilar within each of the two groups than they are for WA-windows as a whole; that is, the split increases separation between neutral and selected behavior. From a purely theoretical point of view, splitting could either increase or decrease separation (this represents an interesting area for further theoretical study), but if it increases separation, then still finer partitions of WA-windows may lead to even higher share estimates. However, finer partitions lead to the compounding of errors in the calculations performed for each group, and this limits their utility. We address the issue of statistical error next.

Control Experiments

As a control for the error associated with our Gaussian smoothing and mixture decomposition, using 50-bp windows with a threshold of 40 aligned bases, we divided the WA-windows in ancestral repeats into two sets, A and B, at random. Set A was used to estimate the neutral score distribution. Set B was used to estimate a genome-wide distribution under a "null" scenario of no selection. Since both data sets contain neutral windows, one expects a near 0 estimate for the fraction under selection: If $f_{neutral}(S) = f_{genome}(S)$ exactly for all scores S, we would have $p_o = \min_S[f_{genome}(S)/f_{neutral}(S)] = 1$, and hence $1 - p_o = 0$. However, random differences between $f_{\text{neutral}}(S)$ and $f_{genome}(S)$ do occur, especially for extreme values of S where very few observations are available and thus both densities are very close to 0. These differences between small density values can generate fairly wide fluctuations in the ratio, resulting in a minimum sizably smaller than 1 (on some control experiments the minimum was <0.9).

The magnitude of this error can be greatly reduced by selecting an appropriate degree of smoothing in the Gaussian kernel procedure and implementing an additional "trimming" procedure for ratio fluctuations on extreme S values (see Methods). With these steps, the control experiments resulted in ratio minima above 0.985. Figure 5

Table 2. Estimates of the Share of the Human Genome under Selection Obtained Splitting WA-windows into a High and a Low Alignment Range, for Various Window Sizes (W)

		Low		High		Summed	WA-windows
W	T	range	$\alpha_{sel,L}$ (%)	range	$\alpha_{sel,H}$ (%)	$\alpha_{sel,+}$ (%)	α_{sel} (%)
30	20	20–24	0.22	25–30	4.5	4.72	4.51
50	40	40–44	0.344	45–50	4.955	5.30	5.15
100	80	80–94	1.53	95–100	4.9	6.43	6.15
200	160	160–194	4.7	195–200	3.45	8.15	7.92

The table reports estimated share of the genome contained in windows under selection for low range ($\alpha_{sel,L}$) and high range ($\alpha_{sel,H}$), and the overall estimate obtained as their sum ($\alpha_{sel,+}$). The last column contains the estimate obtained without partitioning WA-windows (α_{sel}).

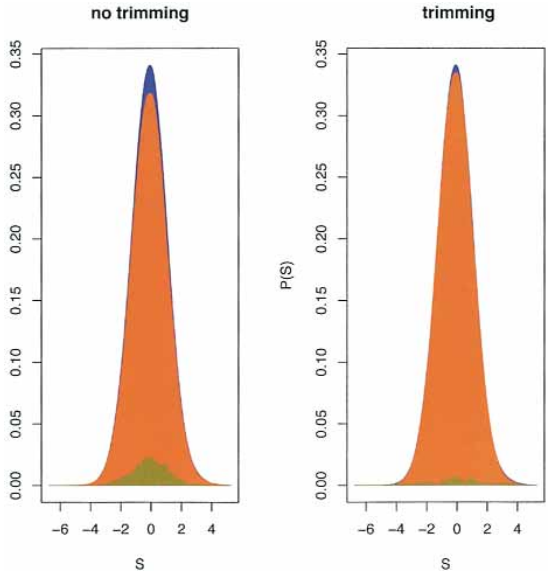

Figure 5. Results of a control experiment for 50-bp WA-windows (alignment threshold = 40). The set of ancient repeat windows is randomly divided into two subsets, A and B, of equal size. A is used to estimate the neutral density (*red*), B is used to estimate a genome-wide density (*blue*) under a "null" scenario of no selection, and the usual procedure is applied to estimate p_o and the density under selection (*green*). Except for error affecting the Gaussian kernel estimation and mixture decomposition, the red and blue curves should be almost coincident, and the green curve negligible. The two panels show the decompositions obtained without (*left*) and with (*right*) "trimming."

illustrates the effectiveness of trimming on a control experiment.

Tests for Alignment Artifacts

A concern with the use of ancestral repeats as a model of neutral substitutions between human and mouse is the reliability of their cross-species alignments. One problem is the possibility that nonorthologous repeats are aligned. This risk is effectively minimized by the BLASTZ alignment procedure used to obtain human–mouse whole-genome alignments: The procedure very carefully first seeds all alignments off unique DNA matches between the two genomes, and only after this extends these matches into the adjacent repetitive regions (Schwartz et al. 2003). Estimates of the amount of nonorthologous DNA that was aligned by this method are quite small (Waterston et al. 2002).

To further ensure that we were not getting nonorthologous alignment, we additionally refined these alignments using a reciprocal filtering method. This method removes nonorthologous alignments by selecting, among BLASTZ alignments, those that can be aligned in both directions (human to mouse and mouse to human) with the highest score—reciprocal best alignments. In our earlier analyses, alignments of mouse genes to human processed pseudogenes had been occasionally included, and as a result, human pseudogenes had appeared largely as if they were under selection. When we switched to reciprocal best alignments, we saw a reduction in our share under selection estimate of about 0.33% (e.g., from 5.33% to 5.0%)—mostly because of the removal of the alignments to processed pseudogenes. Consequently, all the results in this paper use reciprocal best alignments. We note that this filtering eliminates more alignments than the recently proposed chaining (Kent et al. 2003), thereby leading to potentially more conservative lower bounds on the share under selection (see Methods). In fact, we did some experiments recomputing our estimate on chained, syntenic alignments, and obtained results very similar to those obtained on reciprocal best alignments.

Another artifact could derive from failure to correctly align at the base-by-base level some human–mouse orthologous pairs of ancestral repeats. A bias in our estimates could be introduced by an inability to find the most diverged pairs of orthologous repeats, causing them to be absent from our data set, or because after finding these diverged pairs, the relatively large distance between them tempts the optimization method used in the detailed pairwise alignment to find more base identities between them than are actually there in the evolutionarily correct alignment. Similar effects have been observed in simulation studies of pair-wise alignments on synthetic sequences derived by too many substitutions from a synthetic ancestral sequence (Holmes and Durbin 1998). Both of these types of bias would cause the observed level of conservation in ancestral repeats to be greater than it should be, and hence make the true share under selection even larger than what we are estimating. This analysis further reinforces our claim that our estimates produce a lower bound on the share under selection, provided the model of neutral DNA by ancestral repeats is adequate. The additional weakness in the lower bound caused by these potential alignment artifacts is not great because, as mentioned above, the observed levels of conservation between human and mouse in fourfold degenerate sites of codons is similar to that in ancestral repeats, and the former are not subject to alignment artifacts.

Inadequacies of Ancestral Repeats as a Model of Neutral Evolution

The final issue is whether or not the relics of ancestral transposons provide an unbiased model of neutral evolution. We cannot completely resolve this with the data we have, but we have done some tests, in addition to the use of alternate score functions that compensate for base compositional biases, and effects of flanking bases, including dinucleotide effects like bias in substitution rates for CpGs (Roskin et al. 2003).

First, any property of genomic DNA that causes rates of neutral substitution to vary from region to region, such as GC content, should be compensated for by the way we compute our score function relative to the neutral rate estimated from a surrounding window of DNA, rather than on an absolute scale. This prevents one class of DNA from standing out as apparently richer in elements under selection just because it is in a general region of the genome that is accumulating changes at a slower rate.

However, we could still have a bias in the estimate of the share under selection if there are properties that cannot be corrected for by a score function that accounts for compositional biases, and that affect the neutral substitution rate differently in relics of ancestral transposons than they do in other types of neutral DNA.

One possibility is that relics of ancestral transposons, because of their similarity to each other, are more apt to undergo gene conversion (or ectopic conversion) events than other neutral DNA. If we are looking at a pair of orthologously placed transposon relics in the human and mouse genomes, but one of them, say the one in the mouse, has undergone a gene conversion in the rodent lineage, then additional substitutions may have been introduced by that event, making the human–mouse aligned elements less conserved than would be a pair that did not undergo lineage-specific gene conversion. If this were quite common, it would cause us to overestimate the share under selection using our neutral model. However, we note that all gene conversion events that occurred to transposon relics before the human–mouse split would have no effect: They would merely replace the DNA that is inherited by both species. Only primate or rodent lineage-specific gene conversion events can introduce bias. Since triggering a gene conversion event requires a reasonably high degree of sequence identity between the two copies, this means that relics from ancestral transposon families that were only active long before the human–mouse split, and hence whose copies were highly dissimilar to each other at the time of the split, are much less likely to have had a lineage-specific gene conversion event after the split. Basing the neutral model on these "most ancient" relics would then remove the bias.

We divided ancestral repeats into 130 subfamilies as defined by RepeatMasker (Smit and Green 1999), computed the average number of mismatches between each repeat and the consensus sequence for its subfamily, and eliminated from the analysis all repeats belonging to the 65 least diverged subfamilies, representing those transposon relics that were present in the ancestral genome the least amount of time before the human–mouse split. This eliminated roughly 60% of the bases in the original collection of ancestral repeats. For $W = 50$-bp WA-windows, the estimated share under selection obtained with this restricted set of ancestral repeats was very similar to that obtained with the full set of ancestral repeats as a neutral model (ratio of the two estimates was between 0.95 and 1 in different tests of this type with various data sets and alignments). This argues against gene conversion being a source of bias.

Another possibility is that after transposons are inserted they undergo a more rapid substitution rate. Possible causes for this may include mechanisms to suppress transposon transcription, or some holdover from another ancient cellular defense mechanism against insertions of transposons, as well as rapid adaptation of their GC content to the GC content of the surrounding DNA (Bernardi 1993). This would also cause an overestimate in the share under selection using transposon relics as neutral model. However, it seems plausible that such an increased rate of substitutions would diminish as the transposon relics age, so that a very old transposon relic which has been accumulating substitutions in the genome for 50 to 100 million years would behave more like typical neutral DNA. Consequently, we would have expected to see more of a change in the estimate of the share under selection in the above-mentioned experiments, in which we eliminated "younger" ancestral transposon relics from the neutral model. The largest effect we saw when experimenting with different subsets of ancestral transposons to define the neutral model was when we used only SINEs (both "young" and "old" ancestral SINEs). Here the estimate dropped to 4.67%, possibly indicating some bias, but still not as big a fluctuation as we saw by varying the window size and alignment threshold (Table 1).

Finally, one type of neutral DNA that may affect our results because it evolves in a distinct way, different from that of transposon relics, is DNA in simple repeats. However, we found that there are only 2 million bases of human simple repeats aligned with mouse, less than 1/1000th of the genome, so these by themselves could not substantially affect our estimate of the share under selection. Essentially, even rejecting all simple repeats as a priori not being under selection, we would not reduce our estimate of the share of the genome under selection.

DISCUSSION

Ultimately, one would like to identify individual bases of the human genome that are under selection. However, with only one alignment available (to the mouse genome), there is insufficient information to do so at this time. Even a global statistical estimate of the share under selection cannot be made from data relative to single bases, because the neutral and genome-wide distributions will be very similar for any reasonable conservation score computed on two-species comparisons of individual bases.

To illustrate this, consider the simple conservation function that has score = 1 if the aligned bases are identical in human and mouse, and score = 0 otherwise. The estimated probability of score = 1 is 0.667 for aligned neutral bases (from ancestral repeats) and 0.699 for aligned bases genome-wide. Applying our basic mixture estimation method to these data gives an upper-bound estimate of $p_o = 0.904$. This is the maximum fraction of the genome-wide aligned sites that is compatible with the neutral score distribution, and is (implicitly) obtained by assuming that bases under selection are *always* identical between human and mouse. We cannot do any better without assuming prior knowledge of the score distribution for selected sites, something we have avoided in our approach. When we then convert p_o into a lower bound on the fraction of sites in the human genome undergoing selection, we obtain $a_{selected} = 0.35*(1–0.904) = 0.0336$, or about 3.4% (35% of the bases in the human genome are aligned to mouse). In light of the implicit assumption that all selected bases are exactly conserved, this is clearly a very weak lower bound. The problem here is the strong overlap between the neutral distribution and the (unob-

served) selected distribution of the score, which makes them hard to separate. As we have seen, this strong overlap is also present in conservation scores computed on windows when those windows are small, e.g., only 30 bp. This casts doubt on the stringency of lower bounds obtained from very small windows: Technically, they are valid lower bounds, but they might yield significant underestimates of the share under selection.

Using larger windows (instead of single-base positions or very small windows) to estimate the share under selection carries other limitations. We can produce a more stringent lower bound, and thus a better estimate for the fraction of windows that appear to be under selection and the fraction of the human genome that is contained in such windows. However, because the score applies to each window as a whole, we need to restrict attention to well-aligned windows, and recall that our share estimate does not automatically equate to an estimate of the fraction of individual bases under selection. This may not be a severe limitation, because bases are not selected entirely independently from their neighbors; altogether, it may make more sense to consider small regions under selection than individual bases under selection.

From a certain point onward, increasing the window size appears to cause an inflation in the total estimated share of the genome under selection beyond what we can attribute to better separation of the neutral and selected score distributions; see Table 1. However, this is a misinterpretation of the results. For instance, in attempting to compare "on a base-by-base level" the estimated fraction of the genome in 50-bp windows under selection to the estimated fraction in 100-bp windows, we are implicitly converting the probability that a window is selected, $\Pr(w \text{ selected} \mid S(w) = S)$ defined above, into the expected number of bases in the window w that are under selection, which is not legitimate. In fact, the estimates for 50-bp windows and for 100-bp windows are not directly comparable in this fashion: They are estimates of two different underlying quantities, one measuring evolution of smaller (50 bp) segments and the other larger (100 bp) segments.

From a biological perspective, we would like to reliably detect the effects of purifying selection on as small a unit as makes sense; ideally at most a few tens of bases. Given the limitations posed by employing only the human–mouse alignment, the best we can do at this point is to use 50-bp WA-windows: About 5% of the human genome is contained in 50-bp WA-windows that are more conserved than neighboring neutral windows (modeled by ancestral repeats) and thus appear to be under selection. As discussed above, we cannot eliminate the possibility that mechanisms other than purifying selection explain the data we see. In particular, some unidentified specialized types of molecular evolution within ancestral repeats could be causing some kinds of neutral windows to be significantly more conserved than neighboring neutral windows from ancestral repeats, which would artificially inflate our estimate of the share under selection. Tests with alternate score functions that compensate for compositional effects (Roskin et al. 2002, 2003) and tests with different subsets of ancestral repeats as neutral models provide some evidence against the existence of an extreme bias of this type, but cannot eliminate this possibility. Additional evidence will be required to positively prove that the effect we are seeing is due to selection.

The estimate, if valid, leads to the question of what function these elements under selection may possess. 5% is considerably more than can be accounted for by the estimated fraction of the genome that is coding, which is about 1.5%. Note that including all 50-bp windows that contain any coding bases typically adds only about 25 bp on either end of a 200-bp coding exon, increasing the 1.5 coding percentage by a factor of only 5/4. Moreover, this is a considerable overassessment of the effect of coding bases on our estimate of the share under selection because, as we have seen, only about 70% of fully coding 50-bp WA-windows (WAC-windows) are contributing to the estimate as it is, and we expect the fraction to be less for partially coding 50-bp WA-windows. Hence, the bulk of the "selection signal" we are detecting is likely to be coming from noncoding bases, possibly performing regulatory or other important functions.

With multiple alignments to several mammals, it should be possible to develop better score functions based on more accurate models of molecular evolution. These will allow us to separate neutral and selected windows more effectively, and thus to further investigate the properties of small regions of the human genome that are under selection.

ACKNOWLEDGMENTS

We thank all researchers involved in the International Mouse Genome Sequencing Consortium for help and data sharing. In particular, we are grateful to E. Lander, M. Zody, R. Waterston, F. Collins, P. Green, N. Goldman, A. Smit, and W. Miller for their suggestions and comments. We also thank T. Pringle for comments on the research. F.C. was supported by National Human Genome Research Institute grant HG-02238. R.J.W., K.M.R., M.D., W.J.K., and D.H. were supported by NHGRI grant 1P41HG-02371. D.H. was also supported by the Howard Hughes Medical Institute.

REFERENCES

Bernardi G. 1993. The isochore organization of the human genome and its evolutionary history: A review. *Gene* **135:** 57.

Efron B., Tibshirani R., Storey J., and Tusher V. 2001. Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **96:** 1151.

Elnitski L., Hardison R., Li J., Yang S., Kolbe D., Eswara P., O'Connor M., Schwartz S., Miller W., and Chiaromonte F. 2003. Distinguishing regulatory DNA from neutral sites. *Genome Res.* **13:** 64.

Hardison R., Roskin K., Yang S., Diekhans M., Kent W., Weber R., Elnitski L., Li J., O'Connor M., Kolbe D., Schwartz S., Furey T.S., Whelan S., Goldman N., Smit A., Miller W., Chiaromonte F., and Haussler D. 2003. Co-variation in frequencies of substitution, deletion, transposition and recombination during eutherian evolution. *Genome Res.* **13:** 13.

Holmes I. and Durbin R. 1998. Dynamic programming alignment accuracy. *J. Comput. Biol.* **5:** 493.

Ihaka R. and Gentlman R. 1996. R: A language for data analysis and graphics. *J. Comput. Graph. Stat.* **5:** 299.

Kent W.J. 2002. BLAT: The BLAST-like alignment tool. *Genome Res.* **12:** 656.

Kent W.J., Baertsch R., Hinrichs A., Miller W., and Haussler D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci.* **100:** 11484.

Kent W.J., Sugnet C., Furey T., Roskin K., Pringle T., Zahler A., and Haussler D. 2002. The human genome browser at UCSC. *Genome Res.* **12:** 996.

Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J.P., Miranda C., Morris W., and Naylor J., et al. (International Human Genome Sequencing Consortium). 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860.

Lindsay B.G. 1995. *Mixture models: Theory, geometry and applications* (NFS-CBMS Regional Conference Series in Probability and Statistics), vol. 5. American Statistical Association, Alexandria, Virginia.

McLachlan G.J. and Peel D. 2000. *Finite mixture models*. Wiley, New York.

Pruitt K. and Maglott D. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29:** 137.

Roskin K.M., Diekhans M., and Haussler D. 2003. Scoring two-species local alignments to try to statistically separate neutrally evolving from selected DNA segments. In *Proceedings of the 7th Annual International Conference on Research in Computational Molecular Biology* (RECOMB 2003), p. 257.

Roskin K., Diekhans M., Kent W., and Haussler D. 2002. Score functions for assessing conservation in locally aligned regions of DNA from two species. UCSC Technical Report CRL-02-30, September 14, 2002. Center for Biomolecular Science and Engineering, Baskin Engineering, University of California, Santa Cruz.

Schwartz S., Kent W., Smit A., Zhang Z., Baertsch R., Hardison R., Haussler D., and Miller W. 2003. Human-mouse alignments with Blastz. *Genome Res.* **13:**103.

Silverman B. 1986. *Density estimation for statistics and data analysis*. Chapman and Hall, London, United Kingdom.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., Gocayne J.D., Amanatides P., Ballew R.M., Huson D.H., Wortman J.R., Zhang Q., Kodira C.D., Zheng X.H., Chen L., Skupski M., Subramanian G., Thomas P.D., Zhang J., Gabor Miklos G.L., and Nelson C., et al. 2001. The sequence of the human genome. *Science* **291:** 1304.

Waterston R.H., Lindblad-Toh K., Birney E., Rogers J., Abril J.F., Agarwal P., Agarwala R., Ainscough R., Alexandersson M., An P., Antonarakis S.E., Attwood J., Baertsch R., Bailey J., Barlow K., Beck S., Berry E., Birren B., Bloom T., Bork P., Botcherby M., Bray N., Brent M.R., Brown D.G., and Brown S.D., et al. (Mouse Genome Sequencing Consortium). 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520.

Wegman E. 1972. Nonparametric probability density estimation. *Technometrics* **14:** 533.

WEB SITE REFERENCE

Smit A.F. and Green P. 1999. RepeatMasker. (http://ftp.genome.washington.edu/RM/RepeatMasker.html)