Biased Clustered Substitutions in the Human Genome:
The Footprints of Male Driven Biased Gene Conversion

Timothy R. Dreszer\textsuperscript{1}, Gregory D. Wall\textsuperscript{2}, David Haussler\textsuperscript{1,3}, Katherine S. Pollard\textsuperscript{2,4}

1: Department of Biomolecular Engineering and 3: Howard Hughes Medical Institute, University of California, Santa Cruz, CA 95064; 2: Department of Statistics and 3: UC Davis Genome Center, University of California, Davis, CA 95616.

Supplementary Text S1 (corrected)
Supplemental Methods

1. Dataset Preparation.

Publicly available genome sequences were prepared into two distinct datasets: fixed substitutions and SNPs. Since many of the exact same analyses were performed separately upon the two types of single base pair changes in the human genome, we use the term “change” to refer generically to both types. Preparation of the fixed substitutions dataset involved the creation of a set of single nucleotide differences between human and chimp in regions of high quality chimp and macaque sequence. We lifted genome locations from the three species to a common set of assembly coordinates (hg17 or hg18, as specified in the main text). The dataset of human SNPs was combined with corresponding high quality chimp and macaque bases and lifted to hg17 genome assembly coordinates. Both datasets were stored in BED format.

For computations, BED format files were then converted into pairs of equal length arrays, one for location and one for base change. One such pair of arrays was generated for each chromosome. To minimize file size, each base change was reduced to a single 8-bit value, which distinguishes the following attributes:

1. Direction of Change: encodes the ancestral and derived bases. There are twelve unique combinations.
2. Ancestry: encodes which lineage the base change arose in. For SNPs, this is always the human line. However, the concept is still needed to establish the direction of mutation between two alleles.

The use of a pair of arrays provided several computational efficiency advantages. First, all substitutions from a single chromosome could easily fit into available computer memory. Second, finding all substitutions within a fixed region is simply a binary search for the array
locations nearest the region boarder. Third, analysis of whether a given substitution was a member of a cluster required a simple iteration over a range of location indices. Finally, by reducing the entire set of differences to a simple 8-bit value, bit mask analysis could rapidly determine a substitution to be weak to strong, without reconsidering either the ancestral or the derived base (e.g. AtoC:0001 AtoG:0011 TtoC:0101 TtoG:0111 → WtoS_Mask=binary 0001).

Fourth, subsets of the data could be easily generated based on array indexes.

All chimpanzee UBCS analyses were conducted on chimpanzee arrays containing chimpanzee genome assembly locations (pt0 or pt2, as specified in the main text). To compare human and chimp UBCS, chimp data were mapped onto the human genome. The distortion resulting from mapping chimpanzee substitutions using human coordinates proved minor. For instance, the mean difference between chimp-coordinate and human-coordinate based chimp UBCS across the autosomes is less than 2%, despite the slightly different chromosome lengths in the two species. We note that since the datasets originated from high confidence alignments, regions disrupted by genome reorganization are underrepresented.

2. Unexpected Bias Clustered Substitutions (UBCS).

UBCS is a statistic that measures excess bias among densely clustered substitutions in a genomic region. UBCS is computed independently for disjoint genomic regions. Consider a single genomic region containing $N$ substitutions. In order to compute UBCS we need two quantities: the observed number of biased clustered substitutions and the expected number under a null model where there is no association between bias and clustering. A substitution is defined as a clustered substitution (CS) if it is one of at least 5 substitutions within 300bp. This definition of cluster density was selected based on the observed patterns of clustering and bias across the human genome (Figure 1), but others could be used. A CS is further characterized as a biased
clustered substitution (BCS) if it belongs to a cluster with at least 80% weak-to-strong substitutions. Let B denote the set of BCS. Then, the observed number of BCS is

\[ O = \sum_{n=1}^{N} I(s_n \in B), \]

where \( \{s_n : n = 1,\ldots,N\} \) is the set of N substitutions in the region and I is the indicator function.

The expected number of BCS is computed under the null model with no association between clustering and bias. In this model, the probability that a substitution is biased does not depend on whether it is in a cluster or not. Let \( \hat{p} \) denote the proportion of all substitutions \( \{s_n : n = 1,\ldots,N\} \) (CS and non-CS) that are biased and \( p' \) denote the proportion of clusters that are biased. It should be clear that \( p' \) can be calculated as the binomial probability that a cluster contains at least 80% weak-to-strong substitutions, using \( \hat{p} \) and \( (1-\hat{p}) \). If all CS fell in exactly one cluster, then we could compute the expected number of BCS as

\[ E = p' \cdot \sum_{n=1}^{N} I(s_n \in C), \]

where \( C \) is the set of CS. However, our definition of clusters allows for a CS to lie in more than one cluster, any one of which can be biased (and hence make the CS a BCS). Therefore, we cannot compute the expected number of BCS by simple multiplication of \( p' \) times the number of CS. Instead, we must introduce a substitution-specific null probability of being a BCS, \( p'_n \), which depends on the number of clusters containing \( s_n \), the sizes of these clusters, and the overlap between these clusters. An example of how to compute \( p'_n \) is given below. Let \( p'_n = 0 \) when \( s_n \notin C \) (an non-clustered substitution). With the set of estimates \( \{p'_n : n = 1,\ldots,N\} \), we can compute the expected number of BCS when substitutions may lie in more than one cluster:

\[ E' = \sum_{n=1}^{N} p'_n. \]
Finally, UBCS can be computed as

\[ UBCS = O - E'. \]

UBCS is zero (O=E’) under a null model with no relationship between substitution density and bias. When UBCS>0, there is more bias among densely clustered substitutions than expected based on the overall bias of the region. When UBCS<0, there is less bias than expected. The magnitude of UBCS quantifies deviation from the null model of no association between clustering and bias.

**Example UBCS calculation:**
Consider a 1 Mbp region with \( N=6,000 \) human substitutions. Suppose \( M = \sum_{n=1}^{N} I(s_n \in C) = 1,000 \) of the substitutions are clustered substitutions (CS), \( O = \sum_{n=1}^{N} I(s_n \in B) = 150 \) are biased clustered substitutions (BCS), and 2400 are biased (not necessarily clustered) substitutions. Then, \( \hat{p} = 2400/6000 = 0.4 \). If every CS falls in exactly one cluster and each cluster has exactly 5 substitutions, we could compute \( p' = 1 - \left( \sum_{i=0}^{5} \hat{p}^i (1 - \hat{p})^{(5-i)} \right) = 0.913 = 0.087 \) and the expected number of BCS as \( E = p' \cdot M = 0.087 \cdot 1000 = 87 \) and hence \( UBCS = O - E = 150 - 87 = 63 \). Otherwise, we will need to compute \( p'_n \) for each substitution. If a substitution lies in exactly one cluster, then \( p'_n = p' \). If a substitution lies in two or more clusters, then the calculation of \( p'_n \) involves a combinatorial computation. For example, suppose \( s_n \) lies in exactly two clusters, \( C_1 \) and \( C_2 \). Suppose \( C_1 \) contains 5 substitutions, \( C_2 \) contains 6 substitutions, and the overlap \( C_1 \cap C_2 \) contains 2 substitutions. Let \( i \) denote the number of biased substitutions in \( C_1 \), \( j \) the number in \( C_1 \cap C_2 \), and \( k \) the number in \( C_2 - C_1 \). Then,
\[
\Pr(s_n \in B) = 1 - \Pr(C_1 \text{ not biased AND } C_2 \text{ not biased}) \\
= 1 - \Pr(C_1 \text{ not biased})\Pr(C_2 \text{ not biased} \mid C_1 \text{ not biased}) \\
= 1 - \Pr(i < 4)\Pr(j+k < 5 \mid i < 4) \\
= 1 - \left( \sum_{i=0}^{3} \hat{p}^i (1 - \hat{p})^{(5-i)} \sum_{j=0}^{2} \left( \begin{array}{c} i \\ j \end{array} \right) \binom{5-i}{2-j} \sum_{k=0}^{5-i} \hat{p}^k (1 - \hat{p})^{(4-k)} \right) \\
= 1 - (0.0778 + 0.2592*0.98976 + 0.3456*0.96672 + 0.2304*0.93088) \\
= 0.1171 \\
= p'_n \\
\]

After computing all the \( \{p'_n : n = 1, \ldots, N\} \), we can obtain \( E' \), the expected number of BCS. Suppose \( E' = 103 \). Then, \( UBCS = O - E' = 150 - 103 = 47 \).

3. Chromosome 2 Fusion Date.

Our estimate of the chromosome 2 fusion date is based on UBCS values from the distal 16Mps of chimp chromosomes 2a and 2b, the distal ends of human chromosome 2, and the internal region of human chromosome 2 that is orthologous to the distal ends of chimp chromosomes 2a.q and 2b.p. We refer to the two halves of the latter region as human chromosome 2a.q and 2b.p. In order to compute a 95% confidence interval for the fusion date, we also use UBCS from the distal ends of all autosomes with complete human and chimp substitution data over both 16Mbp distal regions (chromosomes 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 16, 17, and 20; Supplemental Table S2). We refer to these as the control chromosomes.

Our fusion date calculation is based on the assumption that the rate of accumulation of UBCS has been constant over the last 6 million years. The high level of similarity between human and chimp UBCS patterns along each of the autosomes suggests that the process generating UBCS is
quite stable and hence that this assumption is reasonable. Based on this stability, we further assume that the ratio of UBCS between the p and q arms of any chromosome will be similar for human and chimp. The Spearman rank correlation coefficient between the human and chimp ratios over all control chromosomes is 0.82, suggesting that our assumption is approximately correct.

We estimate the chromosome fusion date as follows. First, we use the observed UBCS from human chromosome 2.p plus the ratio of UBCS from the two ends of chimp 2a to estimate the expected UBCS for human chromosome 2a.q if it were still a distal region:

$$E(\text{Human 2a.q}) = \text{Human 2.p} \times \frac{\text{Chimp 2a.q}}{\text{Chimp 2a.p}},$$

where each term represents a human or chimp UBCS value for a particular 16Mbp distal (or previously distal) genomic region. We estimate $E(\text{Human 2b.p})$ similarly. Then, we combine these two estimates to get an expected UBCS value for the fusion region: $E(\text{Human 2ab}) = E(\text{Human 2a.q}) + E(\text{Human 2b.p})$. This quantity reflects the UBCS we would expect to see in the fusion region if it had been distal for the entire last 6 million years. Note that $E(\text{Human 2ab})$ will exceed Chimp 2a.q + Chimp 2b.p if Human 2p and/or Human 2q are large relative to the corresponding chimp values. Next, we compare $E(\text{Human 2ab})$ to the observed UBCS in the fusion region:

$$R = \text{Fusion Ratio} = \frac{\text{Human 2ab}}{E(\text{Human 2ab})},$$

where Human 2ab is the sum of the observed UBCS values in human chromosome 2a.q and 2b.p (the fusion region). Finally, we assume that the ratio $R$ represents the proportion of the last 6 million years that the fusion region has been distal. If $R$ is close to one, then the fusion would appear to be very recent. In contrast, if $R$ is small, then the UBCS pattern in the fusion region does not look like that of a distal region, indicating that the fusion was not recent. The fusion date can be estimated as 6 million years ago minus the number of distal years prior to the fusion:

$$\text{Fusion Date} = 6 \times (1-R) \text{ million years ago.}$$
**Example calculation:**

The observed smoothed UBCS values for each 16 Mbp region are:

Human: 2.p = 903.06, 2a.q = 791.73, 2b.p = 1185.57, 2.q = 1082.77

Chimp: 2a.p = 851.41, 2a.q = 1231.69, 2b.p = 807.34, 2b.q = 922.59

Recall that Human 2a.q and 2b.p are no longer distal. Next, compute the expected UBCS for the fusion region:

\[ E(\text{Human 2a.q}) = 903.06 \times \frac{1231.69}{851.41} = 1306.41 \]

\[ E(\text{Human 2b.p}) = 1082.77 \times \frac{807.34}{922.59} = 947.51 \]

\[ E(\text{Human 2ab}) = 1306.41 + 947.51 = 2253.92 \]

Then, compute the observed UBCS for the fusion region and the fusion ratio:

Human 2ab = 791.73 + 1185.57 = 1977.30

\[ R = \frac{1977.30}{2253.92} = 0.8772717 \]

Finally, use \( R \) to estimate the fusion date:

Fusion date = 6 * (1-0.8772717) = 0.7363696 million years ago = 736,370 years ago

In order to put a measure of reliability on the estimated fusion date of 0.74 million years ago, we developed a method to estimate the variance of the fusion ratio using the control chromosomes. We can simulate a fusion of any two human control chromosomes in either orientation (55 pairs of chromosomes or 110 possible fusions) and estimate the UBCS signal in the fusion region using the corresponding distal regions. Each estimated UBCS is then combined with the observed values to obtain a fusion ratio \( R \) for each of the 110 simulated fusions. The variance of these \( R \) values, denoted \( V \), is our estimate of the variance of the chromosome 2 fusion ratio. Since the distribution of \( R \) values is roughly Normal (assessed by histogram and qq-plot), we compute a
95% confidence interval for the fusion ratio as: \( R \pm 1.96\sqrt{V} = \left[ R_{\text{lower}}, R_{\text{upper}} \right] \). A simple linear transformation provides a 95% confidence interval for the fusion date:

\[
95\% \text{ CI} = \left[ \min(0, 6(1 - R_{\text{upper}})), 6(1 - R_{\text{lower}}) \right],
\]

where the minimum for the lower bound is used to prevent a negative date.

**Example calculation:**

\( R = 0.8773 \) (from above)

\( V = 0.03116 \) (from 110 simulated fusions), \( \sqrt{V} = 0.1765 \)

95% CI for fusion ratio = \([0.8773 - 1.96 \times 0.1765, 0.8773 + 1.96 \times 0.1765]\) = \([0.5309, 1.2228]\)

This CI means that with 95% confidence we estimate that the true proportion of the last 6 million years that the fusion region was distal is between 53% and 100% (the upper bound of 1.22 indicates that the fusion region could have even higher UBCS than expected if it were distal).

95% CI for fusion date = \([\min(0, 6(1 - 1.2228)), 6(1 - 0.5309)]\) = \([0, 2.93853]\)

So, with 95% confidence we estimate that the true fusion date is not more than 2.94 million years ago and could be very recent.