# An Algorithmic Analysis of the Role of Unequal Crossover in Alpha-Satellite DNA Evolution

**Can Alkan**[1] cxa27@eecs.cwru.edu
**Jeffrey A. Bailey**[2] jab@po.cwru.edu
**Evan E. Eichler**[2] eee@po.cwru.edu
**S. Cenk Sahinalp**[12] cenk@eecs.cwru.edu
**Eray Tuzun**[1] ext29@eecs.cwru.edu

[1] Department of EECS, Center for Computational Genomics, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH 44106, USA
[2] Department of Genetics, Center for Computational Genomics, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH 44106, USA

**Abstract**

Human DNA consists of a large number of tandem repeat sequences. Such sequences are usually called satellites, with the primary example being the centromeric alpha-satellite DNA. The basic repeat unit of the alpha-satellite DNA is a 171bp monomer. However, with the exception of peripheral alpha-satellite DNA, monomers can be grouped into blocks of $k$-monomers ($4 \leq k \leq 20$) between which the divergence rate is much smaller (e.g. 5%). Perhaps the simplest and best understood mechanism for tandem repeat array evolution is the unequal crossover. Although it is possible that the alpha-satellite sequence developed as a result of subsequent unequal crossovers only, no formal computational framework seems to have been developed to verify this possibility. In this paper we develop such a framework and perform experiments which seem to indicate that pericentromeric alpha-satellite segments (which are devoid of higher-order structure) are evolutionarily distinct from the higher-order repeat segments. It is likely that the higher order repeats developed independently in distinct regions of the genome and were carried into their current locations through an unknown mechanism of transposition.

## 1 Introduction

A considerable portion of the human DNA sequence consist of tandemly repeated sequences which are generally called satellites. The primary example of satellite sequences is the alpha-satellite DNA which is located in the centromeric regions of human chromosomes. Alpha-satellite sequences are composed of tandemly repeated *monomers*, basic repeat units of size approximately 171bp. Arbitrary pairs of alpha-satellite monomers usually exhibit considerable sequence divergence (up to 40%). However it is usually possible to partition the alpha-satellite sequence into blocks of some $k$ monomers ($4 \leq k \leq 20$) between which the sequence divergence is much lower (5% or less) [16].

In addition to higher-order repeats, large tracts of alpha-satellite DNA that are devoid of any higher order repeat structure have been observed recently [11, 7] at the periphery of human centromeric DNA [12, 7, 11]. These are usually called *monomeric alpha-satellite DNA*; subsequently the alpha-satellite segments with higher order repeat structure are usually called *higher order alpha-satellite DNA* (see Figure 1 for the composition of the alpha-satellite DNA).

### 1.1 Satellite DNA Evolution

One possible explanation for the amplification (replication/duplication) of satellite DNA is through random unequal crossover events, either between sister chromatid pairs during meiosis, or between

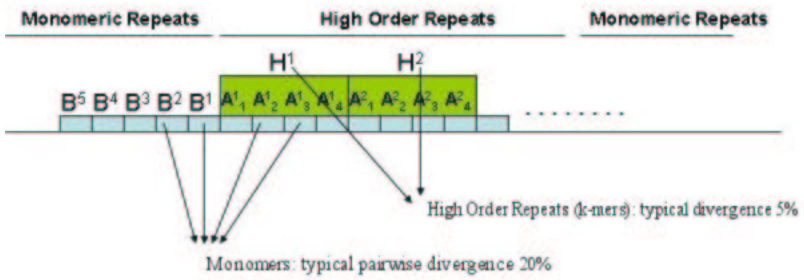

Monomeric Repeats
High Order Repeats
Monomeric Repeats
H1
H2
B5 B4 B3 B2 B1
A1 1 A1 2 A1 3 A1 4 A2 1 A2 2 A2 3 A2 4
.........
High Order Repeats (k-mers): typical divergence 5%
Monomers: typical pairwise divergence 20%

Figure 1: The composition of human centromeric DNA. Here, A and B represent monomers, A being in higher-order, and B in monomeric structure. The higher-order monomers marked with the same subscript are closer to each other (divergence $\leq 5\%$).

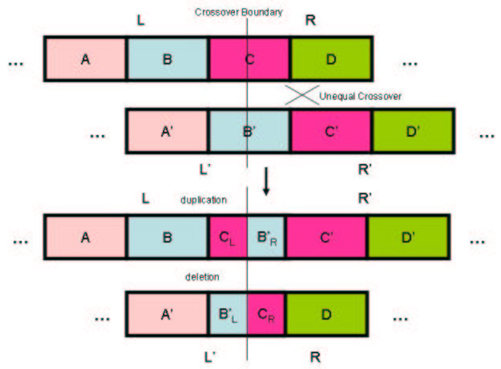


Figure 2: Unequal crossover leading to tandem repeat structure.

homologous chromosomes [5]. The potential role of unequal crossover in tandem DNA array amplification was investigated by [13] and others; it can be argued that DNA segments that are not maintained by natural selection may acquire short segments which are highly similar at nearby locations due to random mutations, and random unequal crossover events between regions containing such segments will result in deletion or tandem replication of these segments.

Subsequent unequal crossovers between pairs of tandem array blocks either tandemly duplicate or delete an integral number of blocks. The number of duplicated or deleted blocks $\ell$ is simply equal to the number of unpaired blocks at either end of the tandem array (see Figure 2). Typically duplication events occur with $\ell = 1$ but if a duplication with some $\ell > 1$ blocks occurs, the next duplication or deletion event may have a better chance of involving exactly $\ell$ blocks again, giving rise to a *higher order* repeat structure.

Although it is possible that some of the higher-order repeat segments in the alpha-satellite DNA appear as a result of unequal crossover events (see for example [11]), recent studies [1] suggest the possibility of an unknown mechanism, complementing unequal crossover in this task. More specifically, it is suggested that the higher order alpha-satellite DNA in certain chromosomes emerged elsewhere in the genome and was transposed into the already existing monomeric repeat sequence by an unknown mechanism, overtaking the function of the existing monomeric structure associated with the centromere. After the establishment of higher-order arrays, the monomeric arrays became inactive.

### 1.2 Summary of Contributions

The main goal of our paper is to develop an algorithm to assess whether unequal crossover is solely responsible for the evolution of alpha-satellite DNA or was complemented by an unknown mechanism as suggested in [1]. Our experimental approach, as per [14, 17], starts with the construction of the phylogenetic trees of the monomers which are extracted from sequenced clones from Human Genome Project (HGP) [8] databases that involve large tracts of alpha-satellite DNA. Because many of the monomers are extracted from unfinished draft sequences, we make no assumptions on the relative ordering of the monomers on the respective clone. Therefore we only associate with each monomer the accession number of the clone it is extracted from. Thus the phylogenetic trees we build involve tandem arrays which have large gaps, whose repeat units are unordered. This limits the applicability of available methods such as [14, 17] in our studies.

We build phylogenetic trees of monomers extracted from two or more clones and try to reach conclusions based on the evolutionary relationships between monomeric and higher-order repeats. Many of the phylogenetic trees we built exhibit a strong separation in the evolutionary history of monomers from higher-order alpha-satellite DNA and monomers from monomeric tracts as per [1]. We also observed that the monomers from different monomeric clones *mix well* [1].

One of our contributions is a simple probabilistic framework for measuring how *surprising* it is that monomers from arbitrary clones remain evolutionarily distinct (or mix well). Within this framework, we obtain exact expressions for the probability of *evolutionary distinctness* for pairs of tandemly generated DNA segments. Our analysis is based on the *restricted tandem repeat history* (RTRH) model used in many other studies of tandem DNA evolution [3, 14, 4, 6, 9, 17]. We thus present a method for checking whether the monomers involved in the phylogenetic studies were derived from each other via tandem duplications prescribed by the RTRH model, practically testing the validity of the model itself. Calculations based on our analysis seem to indicate that unequal crossover is not the only mechanism responsible for alpha-satellite DNA evolution.

## 2 Algorithmic Reconstruction of Evolutionary History

Evolution of tandem DNA arrays, especially via unequal crossovers have recently attracted considerable attention from the computational biology community [3, 14, 4, 6, 9, 17]. A significant portion of available literature is devoted to reconstructing the duplication history of the tandem DNA arrays. Introduced in [3] the *tandem repeat history* (TRH) problem can be described as follows. Given a tandem DNA array the TRH problem aims to iteratively contract pairs of subsequent equal length DNA segments involving one or more repeat units/monomers. The goal is to minimize the total cost of contractions where cost is usually defined in terms of the sequence divergence between contracted copies.

TRH problem involves two key parameters: (1) the length of a duplicated segment (i.e. how many basic repeat units are duplicated in a single unequal crossover event), and (2) the location of the crossover boundary (see Figure 2). The general TRH problem with no restrictions on these two parameters is NP-Complete [9]. A greedy heuristic for this version of the problem is described in [3] and also in [14]. It is possible to impose restrictions on the TRH problem so that (1) only one repeat unit can be duplicated at a time and (2) the crossover boundaries coincide with repeat unit boundaries; we will call this problem the restricted TRH (RTRH) problem as per [3]. Unfortunately the RTRH problem is NP-hard as well [9] although polynomial time algorithms providing approximate solutions are available (a 2-approximation algorithm is given in [3] and polynomial time approximation schemes are given in [9, 14],

A recent related work [4] aims to reconstruct the evolutionary history of *mini-satellite* sequences which are much shorter and less divergent. This work further imposes restrictions on the RTRH

[1] Monomers from different clones may occur as siblings in the phylogenetic tree.

problem by assuming that the contraction cost of two subsequent repeat units is independent of their sequence divergence - provided that the two copies are not identical. A smaller cost is assumed for contracting two identical repeat units. If the root sequence of the phylogenetic tree is also specified, a polynomial time ($O(n^4)$) algorithm is available for solving this further restricted version of the RTRH problem [4].

## 3 Phylogenetic Studies of Tandem DNA Array Evolution

As explained earlier our main concern in this paper is not directly inferring the evolutionary history of the alpha-satellite DNA. Methods for solving this problem *assume* that tandem duplications are the only source of tandem DNA evolution. We rather focus on how likely it is that the alpha satellite sequences evolve solely through unequal crossover events resulting in tandem duplications and deletions. One recent work [14] investigates a similar *reconstructing a duplication model from phylogeny* (RDMP) problem. Here, given as the input (1) the complete sequence of a tandem DNA array and (2) the (binary) phylogenetic tree of its basic repeat units, the goal is to check whether there exists a sequence of tandem duplication events that can produce the complete tandem array from a single repeat unit. A quadratic time algorithm for this problem is given in [14], and an optimal linear time algorithm is given in [17].

Unfortunately the nature of the sequence data we use and the specific problem we need to address limits the applicability of the available methods for our purposes. In particular, they usually require the tandem array sequence in question to be complete and correctly assembled. Furthermore the basic question we consider is different. For example the algorithms in [14], and [17] aim to find whether it is *possible* that tandem duplication can be the sole mechanism behind the evolution of a tandem DNA array. We would rather like to compute how *probable* it is that the DNA array was produced via tandem duplications only.

One more issue that we had to consider is related to input data reliability. Although heterochromatic regions have not been a target of HGP [8] there is now a large number of sequenced clones that contain large blocks of contiguous alpha-satellite DNA. Unfortunately many of the clones that contain these monomers are unfinished draft sequences. They contain large gaps and the order of the monomers is suspect. Furthermore the relative locations of these clones in their respective chromosomes are largely unknown. These issues are not unique to alpha-satellite sequences, however input data reliability has never been considered in the context of tandem repeat history reconstruction.[2]

## 4 Probabilistic Testing of the Role of Unequal Crossover in Tandem DNA Array Evolution

As mentioned earlier, our main goal is to derive information about the evolutionary history of the alpha-satellite DNA, whose sequence information is known only indirectly as follows: We are given a collection of monomers whose relative locations in the tandem array are unknown. The only information we have about the monomers is that each of them comes from one of a number of distinct clones extracted from the tandem array and it is known to which clone each monomer belongs.

In the following section we show that one can derive clues about the evolutionary history of the alpha-satellite DNA sequence if clones from both the higher-order region and the monomeric tracts are considered: Under the assumption that the alpha-satellite sequences were generated by uniformly random tandem duplications (as per the RTRH model), we calculate the probability that the monomers from any two clones have distinct ancestors. Our experiments indeed indicate that

[2] For example, the experimental study in [14] examines only a few (16) repeat units with considerable divergence and thus input reliability was not really an issue; the study in [3] is performed on synthetic data. Other studies [9, 17] focus on performance evaluation under the implicit assumption that the sequence data is reliable.

monomers from higher-order regions of the alpha-satellite DNA and those from monomeric tracts have distinct ancestors. Within the parameters associated with our experimental setting we show that the likeliness of this observation is quite low; this leads us to the conclusion that unequal crossover is not exclusively responsible from alpha-satellite sequence evolution.

### 4.1 Evolutionary Distinctness Problem

We first describe a probabilistic framework to analyze how likely it is that monomers extracted from two long substrings of a tandem array have distinct ancestors.

Let $S$ and $R$ be two substrings of a tandem repeat array $A$. It is assumed that (1) $A$ has been generated by restricted tandem duplications as per the RTRH model (i.e. each unequal crossover event results in the duplication of exactly one monomer and the crossover boundaries coincide with monomer boundaries) and (2) the probability of which monomer gets duplicated (and to which direction) in any event is uniform over all monomers and both directions, independent of previous events. Consider the actual evolutionary tree $T$ of the monomers in $A$; let the lowest common ancestor (LCA) of all monomers in $S$ be $a(S)$ and the LCA of all monomers in $R$ be $a(R)$. The evolutionary distinctness problem is as follows: given the distance between $S$ and $R$, what is the probability that $a(S)$ and $a(R)$ are two distinct nodes in $T$ such that none is an ancestor of the other. In other words, what is the probability that the lowest common ancestor of any pair of monomers $m_S$ from $S$ and $m_R$ from $R$ is identical for all pairs $m_S, m_R$.

We provide an exact expression for evolutionary distinctness of $S$ and $R$ as follows: Let the lengths (in terms of monomers) of $S$ and $R$ be $w$ and $v$ respectively. Assuming that $S$ and $R$ do not overlap in $A$, let $P$ be the substring of $A$ that stretches between $S$ and $R$; let $k$ be the length (in terms of monomers) of $P$. Given any $0 \leq h \leq k$ in $P$ let $P_h(R)$ be the length $h$ substring of $P$ that is closest to $R$; similarly let $P_h(S)$ be the length $h$ substring of $P$ that is closest to $S$. Clearly the concatenation of $P_h(S)$ and $P_{k-h}(R)$ gives $P$ itself. During the construction of array $A$ through uniformly random duplication events let $E$ be the event that $a(S)$ and $a(R)$ are distinct and none is an ancestor of the other; thus $Pr(E|k, v, w)$ is what we want to evaluate.

**Lemma 1** $Pr(E|k, v, w) = \frac{k+1}{k+v+w-1}$.

**Proof.** In order for $E$ to take place there must exist an $0 \leq h \leq k$ for which the following event $E_h$ must take place. Given $h$, let $a_h(S) = a(S, P_h(S))$ be the LCA of all monomers in $S$ and $P_h(S)$, and $a_h(R) = a(R, P_{k-h}(R))$ be the LCA of all monomers in $R$ and $P_{k-h}(R)$. $E_h$ is the event that $a_h(S)$ and $a_h(R)$ are distinct and none is an ancestor of the other (See Figure 3).

Notice that for any two $h \neq h'$, the events $E_h$ and $E_{h'}$ can not take place simultaneously. Thus $Pr(E_h \cap E_{h'} | h \neq h', k, v, w) = 0$. Furthermore $E = \cup_{0 \leq h \leq k} E_h$, thus $Pr(E|k, v, w) = \sum_{0 \leq h \leq k} Pr(E_h|k, v, w)$, which reduces our job to the computation of $Pr(E_h|k, v, w)$.

$E_h$ is the event that given $k+v+w$ subsequent monomers, exactly $h+w$ subsequent monomers are descendants of $a_h(S)$ and the remaining $k-h+v$ subsequent monomers are descendants of $a_h(R)$. Consider the reverse random process of building an evolutionary tree out of these $k+v+w$ monomers in iterations. In the first iteration the probability of any two neighbor monomers being contracted is uniform. Only a contraction between the two boundary monomers between the sets of $h+w$ and $k-h+v$ monomers will contradict with $E_h$. Similarly, in each subsequent iteration only one potential contraction will contradict with $E_h$. Thus $Pr(E_h|k, v, w) = (1 - \frac{1}{k+v+w-1}) \cdot (1 - \frac{1}{k+v+w-2}) \cdots (1 - \frac{1}{2}) = \frac{1}{k+v+w-1}$. Because $E_h$'s are independent $Pr(E|k, v, w) = \frac{k+1}{k+v+w-1}$.

**Lemma 2** *If $k$ is not fixed but is determined uniformly at random from the range $[0 \ldots (m-w-v)]$, then* $Pr(E|v, w) \cong 1 - (\frac{v+w-2}{m-v-w+1} \cdot \ln \frac{m-1}{v+w-1})$. [3]

[3] ignoring some small order ($o(\frac{1}{m-v-w} + \frac{1}{v+w})$) additive terms.

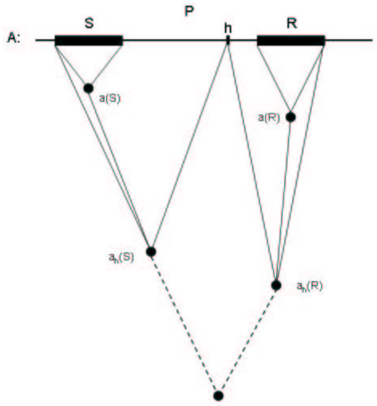


Figure 3: Distinct evolution of substrings $R$ and $S$.

**Proof.** By definition,

$$\begin{aligned} Pr(E|v,w) &= (\tfrac{1}{m-v-w+1}) \cdot \textstyle\sum_{k=0}^{m-v-w} \tfrac{k+1}{k+v+w-1} = (\tfrac{1}{m-v-w+1}) \cdot \textstyle\sum_{k'=v+w-1}^{m-1} \tfrac{k'-v-w+2}{k'} \\ &= (\tfrac{1}{m-v-w+1}) \cdot \textstyle\sum_{k'=v+w-1}^{m-1} (1 - \tfrac{v+w-2}{k'}) \cong (\tfrac{m-v-w}{m-v-w+1}) - (\tfrac{v+w-2}{m-v-w+1}) \cdot \ln \tfrac{m-1}{v+w-1} \\ &\cong 1 - (\tfrac{v+w-2}{m-v-w+1} \cdot \ln \tfrac{m-1}{v+w-1}). \end{aligned} \tag{1}$$

## 5 The Experiments

Our experimental framework involves constructing the phylogenetic trees of monomers extracted from various clones from HGP [8] databases that involve (higher-order and monomeric) alpha-satellite DNA. In this section we first describe how we obtain the monomer data sets and which data sets we chose to construct the phylogenetic trees. Based on the calculations presented earlier, we then try to compute the probability that random unequal crossover events can result in the evolutionary trees we obtained.

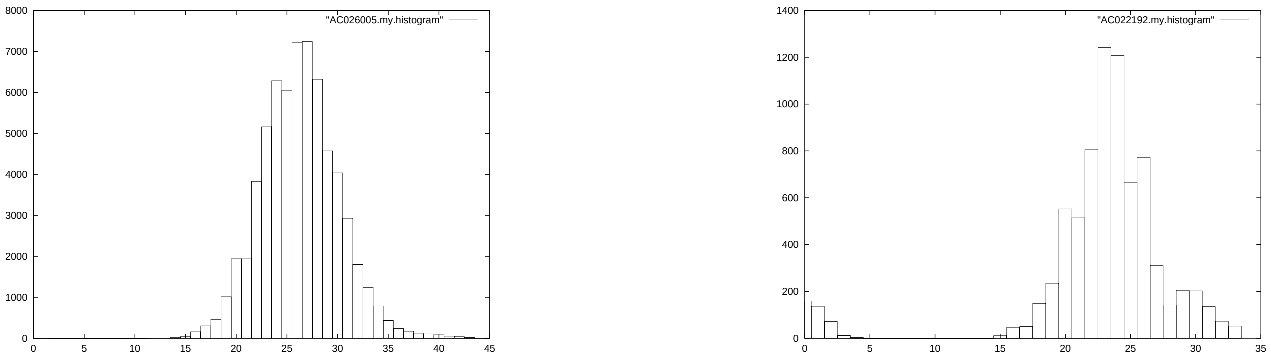
### 5.1 Data Extraction and Classification

As a first step, we established a library of higher-order repeat sequences that have been identified in the literature for each human chromosome [1, 10]. These sequences were searched in the entire Human Genome Project [8] sequence data so as to identify large-insert clones that include tracts of higher order and monomeric alpha-satellite DNA. As a result, a number of large-insert clones were selected whose map location on specific human chromosomes were confirmed to overlap with centromeric DNA.

In the second step we extracted all monomers from each selected clone through the `repeatmasker` program [19] using the higher order monomer library.[4]

As a third step, we classified each clone as higher-order or monomeric based on known sequence similarity distributions between monomers in higher order and monomeric tracts [1, 10, 11]. Accordingly, a clone was identified to be *higher order* if each of its monomers were highly similar ($> 95\%$) to at least one other monomer extracted from that clone. We obtained only four such clones. Similarly, a clone was identified to be *monomeric* if its monomers exhibited significant divergence against all other monomers in the same clone ($> 10\%$). Eleven monomeric clones were obtained as a result of our search. (Figure 4 demonstrates typical distributions of pairwise divergence between monomers from both higher-order and monomeric clones.)

[4] Existing tandem repeat identification and extraction programs such as [2] could not be trivially employed for our purposes. As mentioned earlier this is due to the fact that many of the clones we analyzed were unfinished draft sequences whose sequence information is not always reliable. Furthermore there was a need for filtration of common repeat sequences in the clones we extracted.



AC026005: monomeric

AC022192: higher order

Figure 4: Classification of clones according to pairwise monomer divergence.The range of percentage divergence between monomers (x-axis) is plotted against the number of monomer pairs with given divergence levels (y-axis).

### 5.2 Phylogenetic Tree Construction

Our primary goal in phylogenetic analysis was to understand the evolutionary relationships between the monomeric tracts and the high order tracts of alpha-satellite DNA evolution. Because very few monomeric and higher-order clones were extracted from the same chromosome we applied the following strategy. For each monomeric clone, we built a phylogenetic tree of its monomers together with the monomers from the higher-order repeat library. This library includes a reasonably comprehensive set of monomers from higher-order alpha-satellite DNA. Our premise is that because the percentage divergence between repeats from the same higher-order tandem array is 5% or less, every potential higher-order clone will have a number of representative monomers in the library. To confirm this, we first constructed the phylogenetic trees of monomers from each higher-order clone against those from our high order repeat library. A sample phylogenetic tree involving monomers from clone AC022192 from chromosome 5 is given in Figure 5. In all of the trees built, we observed that monomers from the higher-order clone clustered into distinct families as expected. Moreover most of these families appeared to be strongly related to monomers from the repeat library, which seems to indicate that the library is fairly comprehensive.

Our main tests involved construction of a phylogenetic tree for each monomeric clone identified. Each experiment involved all monomers from a given monomeric clone as well as all monomers from the higher-order repeat library. In addition, a sample of monomers from higher-order alpha satellite sequences of *Old World* (macaque and baboon) and *New World* (brown capuchin, titi monkey etc.) monkey species were added for outgrouping.

The phylogenetic trees are constructed as follows. We used `Clustal-W` program (based on neighbor join (NJ) method) to construct first a *guide tree* (i.e. *dendogram tree*) and then the multiple alignment of the monomers involved. We used the multiple alignment of monomers to reconstruct their phylogenetic tree with a bootstrap support of *100*. In order to do this efficiently we modified the `Clustal-W` code (v 1.82) [15] so as to reduce its space usage by a factor of 4. On a DELL Pentium-IV PC with 2Ghz speed and 1Gbyte memory, the construction of a tree took between 36 hrs (for 700 monomers) to 6 days (1000 monomers). To visualize the phylogenetic trees obtained, we modified the `ATV` (a tree viewer) program [18] to color-code specified nodes.

Two of the phylogenetic trees involving clones AC022482 from chromosome 10 and AC026005 from chromosome 8 are given in Figures 6 and 7 respectively. In both figures one can observe that the evolutionary relationship between monomers extracted from the respective clone are quite weak suggesting a lack of recent duplication activity. Nevertheless, these relationships are much stronger than those between the monomers extracted from the clone and the higher-order repeats. This suggests the possibility that the higher-order repeats did not emerge from the existing monomeric structure of the respective chromosome, but rather were brought in through some unknown mechanism as suggested in [1]. Among the eleven phylogenetic trees we constructed five of them displayed such

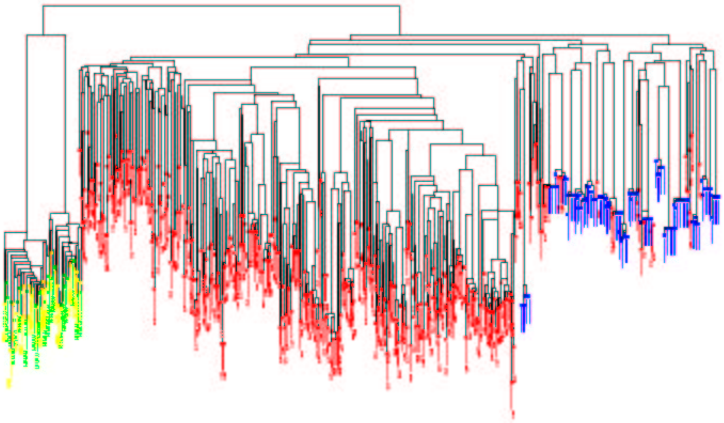


Figure 5: The phylogenetic analysis of all monomers from higher-order clone AC022192. Color coding: (1) monomers extracted from the clone: *blue*, (2) higher order monomer families from our library: *red.* (3) Old World monkey sequences: *yellow* (baboon) and *green* (macaque). Although the monomers are highly clustered among themselves they also mix well with some of the monomers from the higher order repeat library.

characteristics. The division of monomers into higher-order and monomeric clusters were supported by 70%-90% of bootstrap replicates. We also built a phylogenetic tree that involves a large sample of monomers from each of these five clones as well as monomers from higher-order repeat library and the monkey sequences (see Figure 8). Note that although these monomers are extracted from five different chromosomes (2,8,9,10,19) they still seem to mix well while staying distant from the higher-order repeats. We observed some level of mixing between monomeric and higher-order repeats in the phylogenetic trees we built for the remaining six monomeric clones.

It is possible to compute the probability that 5 out of 11 monomeric clones stay evolutionarily distinct from the higher-order monomers by using the simple expressions we provided earlier. Although estimates on the length of monomeric alpha-satellite DNA vary, 200Kbp seems to provide an agreeable upper bound. All the monomeric clones used in our experiments were longer than 150Kbp - which we take to be the size of the clone. Our calculations make the following assumptions: (1) unequal crossover is solely responsible from the evolution of the alpha-satellite DNA, (2) once the higher-order tracts emerged, the monomeric repeat units became inactive, and (3) our higher-order repeat library involves at least one higher-order monomer from the tandem array involving each monomeric clone. To obtain a relatively generous estimate on the evolutionary distinctness of monomeric tracts and higher-order tracts, we set $m = 200K/171$, $w = 150K/171$ and $v = 1$. This gives a probability of $1 - \frac{w}{m-w} \ln \frac{m}{w} = 1 - \frac{150}{200-150} \ln \frac{200}{150} = 0.14$. The frequency of evolutionary distinctness in the 11 phylogenetic trees we constructed were $\frac{5}{11} > 0.45$; this is three times our estimate, indicating that the distinctness of the two classes of repeats may not be by chance.

Another interesting observation is that the monomers from different chromosomes *mix well* with each other while remaining distinct from the higher-order monomer set, as in Figure 8. This clustering further supports the hypothesis that the monomeric tracts and the higher-order tracts have distinct origins and thus unequal crossover may not be the only mechanism responsible from the evolution of alpha-satellite DNA.

## Conclusion

The phylogenetic trees we obtained seem to support a distinct evolutionary relationship between monomeric and higher order alpha-satellite DNA. The division between monomeric and higher-order sequence sets in the phylogenetic trees constructed were supported by 70-90% of bootstrap replicates - which suggests no derivative phylogenetic relationship between higher-order and monomer repeat classes. This implies that unequal crossover may not be the sole mechanism determining the compo-

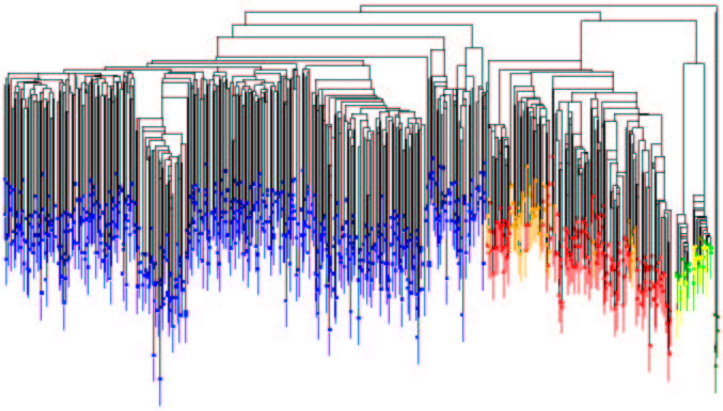


Figure 6: The phylogenetic analysis of all monomers from monomeric clone AC022482. Color coding: (1) monomers extracted from clone: *blue*, (2) higher order monomer families from our library: *red/orange* (3) monomers from Old World monkey sequences: *yellow* (baboon) and *green* (macaque), and (4) New World monkey monomers: *dark green*. Note that this tree involves a richer library of higher-order sequences compared to the one in Figure 5; additional sequences are colored orange. Still, there is strong disconnection between the monomeric and high order tracts (from our -richer-higher-order repeat library).

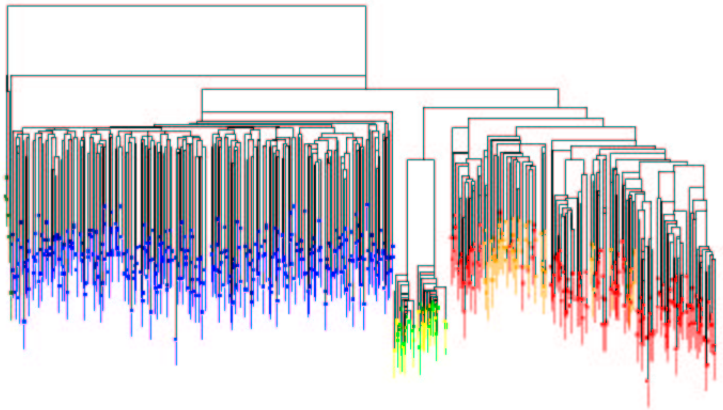


Figure 7: The phylogenetic analysis of all monomers from another monomeric clone AC026005. Color coding: (1) monomers extracted from clone: *blue*, (2) higher order monomer families from our library: *red/orange*, (3) monomers from Old World monkey sequences: *yellow* (baboon) and *green* (macaque), and (4) New World monkey monomers: *dark green*. Once again, the phylogenetic relationship among the monomers from the clone are weak; yet they are still well separated from the monomers from our higher order repeat library.

sition of large tracts of alpha-satellite sequence within the vicinity of the human centromeric DNA.

## References

[1] Alexandrov, I., Kazakov, A., Tumeneva, I., Shepelev, V., and Yurov, Y., Alpha-satellite DNA of primates: Old and new families, *Chromosoma*, 110(4):253–266, 2001.

[2] Benson, G., Tandem repeats finder - a program to analyze DNA sequences, *Nucleic Acids Res.*, 278:573–580, 1999.

[3] Benson, G. and Dong, L., Reconstructing the duplication history of a tandem repeat, *Proc. Seventh International Conference on Intelligent Systems for Molecular Biology*, 44–53, 1999.

[4] Berard, S. and Rivals, E., Comparison of minisatellites, *Proc. Sixth Annual International Conference on Computational Biology*, 67–76, 2002.

[5] Charlesworth, B., Sniegowski, P., and Stephan, W., The evolutionary dynamics of repetitive DNA in eukaryotes, *Nature*, 371:215–220, 1994.

[6] Elemento, O., Gascuel, O., and LeFranc, M.P., Reconstructing the duplication history of tandemly repeated genes, *Molecular Biology and Evolution*, 19(3):278–288, 2002.

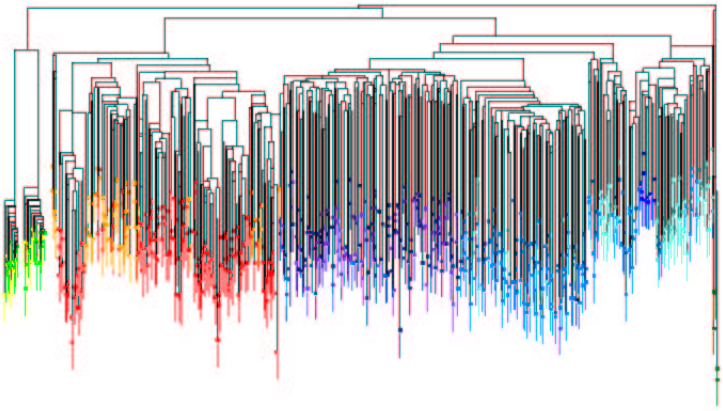


Figure 8: The phylogenetic analysis of monomers from five monomeric clones. Color coding: (1) monomers extracted from the clones: various tones of *blue*, (2) higher order monomer families from our library: *red/orange*, (3) monomers from Old World monkey sequences: *yellow* (baboon) and *green* (macaque), and (4) New World monkey monomers: *dark green*. The monomers (colored blue) from different clones do mix well, however they are well separated from the monomers from the higher-order repeat library (with 31% bootstrap support).

[7] Horvath, J.E., Viggiano, L., Loftus, B.J., Adams, M.D., Archidiacono, N., Rocchi, M., and Eichler, E.E., Molecular structure and evolution of an alpha satellite/non-satellite junction at 16p11, *Human Molecular Genetics*, 9(1):113–123, 2000.

[8] IHGSC, International Human Genome Sequencing Consortium

[9] Jaitly, D., Kearney, P., Lin, G., and Ma, B., Methods for reconstructing the history of tandem repeats and their application to the human genome, *Journal of Computer and System Sciences*, (in press).

[10] Lee, C., Wevrick, R., Fisher, R.B., Ferguson-Smith, M.A., and Lin, C.C., Human centromeric DNAs, *Human Genetics*, 100:291–304, 1997.

[11] Mashkova, T., Oparina, N., Alexandrov, I., Zinovieva, O., Marusina, A., Yurov, Y., Lacroix, M.H., and Kisselev, L., Unequal crossover is involved in human alpha-satellite DNA rearrangements on a border of the satellite domain, *FEBS Letters*, 441(3):451–457, 1998.

[12] Schueler, M.G., Higgins, A.W., Rudd, M.K., Gustashaw, K., and Willard, H.F., Genomic and genetic definition of a functional human centromere, *Science*, 294(5540):109–115, 2001.

[13] Smith, G.P., Evolution of repeated DNA sequences by unequal crossover, *Science*, 191:528–535, 1976.

[14] Tang, M., Waterman, M., and Yooseph, S., Zinc finger gene clusters and tandem gene duplication, *Journal of Computational Biology*, 9(2):429–446, 2002.

[15] Thompson, J.D., Higgins, D.G., and Gibson, T.J., Clustal-W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, 22(22):4673–4680, 1994.

[16] Willard, H.F. and Waye, J.S., Chromosome-specific subsets of human alpha-satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat, *J. Mol. Evol.*, 25(3):207–214, 1987.

[17] Zhang, L., Ma, B., and Wang, L., Efficient methods for inferring tandem duplication history, *Proc. 2nd International Workshop on Algorithms in Bioinformatics*, 97–111, 2002.

[18] Zmasek, C.M. and Eddy, S.R., ATV: Display and manipulation of annotated phylogenetic trees, *Bioinformatics*, 17:383–384, 2001.

[19] Smit, A.F.A. and Green, P., `http://ftp.genome.washington.edu/RM/RepeatMasker.html`