Distribution of short, paired duplications in mammalian genomes

Elizabeth E. Thomas[1], Nathan Srebro[2], Jonathan Sebat[1], Nicholas Navin[1], John Healy[1], Bud Mishra[1,3], and Michael Wigler[1]

[1]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York, 11724, USA.
[2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, USA.
[3]Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, New York 10012, USA.

Classification: Biological Sciences, Evolution.

Corresponding Author:
Michael Wigler
1 Bungtown Rd
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724
Phone: 516-367-8376
Fax: 516-367-8381
wigler@cshl.edu

Mammalian genomes are densely populated with long duplicated sequences. In this paper, we demonstrate the existence of "doublets", short duplications between 25 and 100bp, distinct from previously described repeats. Each doublet is a pair of exact matches, separated by some distance. The distribution of these inter-match distances is strikingly non-random. An unexpectedly high number of doublets have matches either within 100 bp ("adjacent"), or at distances tightly concentrated around 1000 bp apart ("nearby"). We focus our study on these "proximate" doublets. First of all, they tend to have both matches on the same strand. By comparing "nearby" doublets shared in human and chimp, we can also see that these doublets seem to arise by an "insertion" event that produces a copy without markedly affecting the surrounding sequence. Most doublets in humans are shared with chimpanzee, but many new pairs arose after the divergence of the species. Doublets found in human but not chimp are most often composed of almost tandem matches, while older doublets (found in both species) are more likely to have matches spaced by about 1 kb, indicating that the nearly tandem doublets may be more dynamic. The spacing of doublets is highly conserved. So far, we have found clearly recognizable doublets in the following genomes: *Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana*, and *Caenorhabditis elegans*, indicating that the mechanism generating these doublets is widespread. A mechanism that generates short, local duplications while conserving polarity could have a profound impact on the evolution of regulatory and protein-coding sequences.

Genome expansion through duplication has been a prominent force in evolution. The human genome, in particular, is littered with signs of past duplication (*1*,*2*). Transposons (*3,4*), processed pseudogenes (*5*), and segmental duplications (*6*), are all known classes of repeats found in mammalian genomes. All of these types of duplications play an important role in gene and genome evolution (*7*), either through gene duplication and subsequent gene specialization, or through the creation of unstable genomic regions.

Duplication is also important on a smaller scale. Comparative studies of promoters such as the vertebrate growth hormone gene (*8*) make it clear that gene regulation often evolves by increasing the number of copies of a given *cis* regulatory motif. Similarly, protein function can also evolve by the addition of tandem copies of protein domains. These types of short, tandem, or nearly tandem, duplication events can have as striking an effect on gene evolution as whole gene or genome duplications. It is clear that once two or more copies of a given sequence are present at a locus, homologous recombination can further increase their number.

In this paper we present evidence that short, unique, sequences are being actively duplicated in mammalian genomes. These short duplications occur frequently, have a strong tendency toward proximity and conservation of polarity, and do not fit into any of the well-studied classes of interspersed repeats. Studying these short duplications will give us insight into the process by which a unique sequence is duplicated, an important first step in the creation of a tandem array, and potentially a key process in the evolution of gene regulation and protein function.

Methods

Identifying doublets. Human genome sequence is the April 2003 assembly from UCSC (*9*). We first identified "cores" of at least 25 bp which occur exactly twice in the genome. Genomic counts (number of occurrences of a substring within the genome) were determined using the mer-engine method (*10*). We further required that the 21 bp substrings of the cores occur nowhere else, and that at least one of the cores be flanked on either side by 21 bp of unique sequence. Each core is associated with the 100 bp immediately flanking it to the left and to the right (**Fig 1A**). The Needleman-Wunsch global alignment algorithm (*11*) was used to calculate the alignment scores between the flanks with match, mismatch, and gap scores set to 1, 0, and –1 respectively. Many of the flanks share a large degree of homology (**Fig 1B**), indicating that the cores are not independent short exact duplications, but small parts of a larger approximate duplication. To eliminate these we compared the observed alignment score to the distribution of alignment scores between "unrelated" sequences, determined by aligning one flank of one core and the reverse complement of the corresponding flank of the other core. We calculated the mean+2*standard deviation of the distribution of reverse complemented sequences, and used this number as an upper bound on the maximum allowable alignment score. About 86% of paired sequences were eliminated at this stage.

We anticipated that a small percentage of the remaining pairs would be the result of processed pseudogenes. If a gene occurs twice in the genome, once with introns and once without, then an exon of the complete gene will be a duplicated sequence immediately flanked by non-homologous sequence. To exclude this source of paired sequences, we next discarded any pairs in which the matched substring has homology to a sequence in NCBI's est_human database (expectation $\leq 10^{-4}$ using MegaBLAST with default parameters; http://www.ncbi.nih.gov/BLAST/). Approximately 20% of the pairs were eliminated at this stage.

To find doublets in other genomes, the same procedure was carried out, using matched pairs of genomic sequences and coding sequence databases. *Mus musculus* sequence is from NCBI's Build 30 (*12*), *Caenorhabditis elegans* sequence is WS110 from Wormbase (*13*), *Drosophila melanogaster* sequence is

release 3-1 from FlyBase (*14*), *Plasmodium falciparum* sequence is from the Sanger center (*15*), and *Arabidopsis thaliana* sequence is from NCBI (*16*).

Inter-core distance distribution. For a doublet with both cores on the same chromosome (intra-chromosomal doublets) the inter-core distance is the number of base pairs in the "spacer" (**Fig 1A**) between the two occurrences of the core. For each chromosome, we plotted the distribution of inter-core distances of all intra-chromosomal doublets on that chromosome (**Fig 1C depicts human chromosome 2**). We compared this distribution with two random models that take into account the overall number of intra-chromosomal doublets in each chromosome. The first model assumes each core location is independently and uniformly distributed along the chromosome, yielding an expected distance distribution of $P(\text{distance}<d)=2d-d^2$, where d is the inter-core distance normalized by chromosome length. If the distribution of core locations along the chromosome is non-uniform (some regions are core-rich and others core-poor), the distance distribution will deviate from this model. To account for such non-uniformities, the second model uses the true locations of all the intra-chromosomal cores on the chromosome, but assumes cores are randomly matched up into doublets, independent of their locations. The expected distance distribution based on this model was calculated by Monte Carlo simulation.

Comparison to Chimpanzee. Chimpanzee sequence (*Pan troglodytes*) is the December 2003 WIBR assembly from the Chimpanzee Genome Sequencing Consortium (http://www.genome.gov/11509418). In order to reduce our chances of finding paralogous rather than orthologous matches, we first screened out doublets in which either core was flanked by non-unique DNA. To do this, we determined the genomic counts for all 21 bp words in each of the flanks. If any of these 21 bp words occurred 10 or more times in the genome, or if the average genomic count was greater than 3, the corresponding doublet was eliminated from the analysis.

For each of the remaining doublets, we identified the outermost 100 bp flanks of the doublet and used MegaBLAST to find regions in the chimpanzee genome with at least 80% identity over 90 bp. We then extracted the intervening sequence. We created four different versions of the original doublet from the human or mouse genome: one with flanks, both cores and the inter-core sequence; one with both flanks, a single core, and the inter-core sequence; one with both flanks and a single core, but no inter-core sequence;

and one with flanks and inter-core sequence but no cores. The *needle* program from the EMBOSS suite (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/needle.html; gap open penalty = 10, gap extend penalty = 0.5, match score = 5, mismatch score = -4) was then used to align the genomic region from the alternate genome to all of these sequences, and the alignment with the best score was used to assign a label to the doublet: "two or more cores conserved", "one core and spacer conserved", "one core and no spacer conserved" or "no spacers conserved". The resulting alignments are viewable in **Supplemental Figure 1**, and the results are summarized in **Table 1**.

Comparison to Transposons. Alu annotations are from the UCSC genome browser (*9*) and consensus sequences are from RepBase Update database (*17*). For each of the 51 cases in which a doublet overlaps an annotated Alu, two versions of the annotated sequence were generated, one with the core (as found within the genome) and one with the core removed. Both versions were globally aligned to the consensus using *needle* (gap open penalty = 10, gap extend penalty = 0.5, match score = 5, mismatch score = -4), and if the core-excised version had a higher scoring alignment to the consensus, the core was classified as an "insertion".

Composition of Inter-core Sequences. For each of the 2,020 inter-core spacer sequences from "nearby" human doublets, we downloaded overlapping RepeatMasker, segmental duplication, and RefGene annotations from the UCSC genome browser (*9*). We did the same for five sets of randomly chosen genomic intervals with the same length distribution as the set of spacers. For each set of sequences and each type of annotation, we counted the number of sequences that overlapped a given type of annotation by at least 50%.

Segmental duplications have only been annotated if they are at least 1 kb in length. In order to limit any biases introduced by this length threshold, we also looked at "uniqueness" of the spacer sequences, compared to random genomic intervals and a random sampling of annotated segmental duplications. To determine "uniqueness" we annotated each sequence with the number of genomic occurrences of each of its constituent 18-mers (*10*). We then calculated what percentage of the 18-mers in any given sequence set were "unique" (found only once in the genome), "low count" (found between 2 and 5 times in the genome) or "high count" (found in more than five locations in the genome).

Results and Discussion

To find instances of short repeats, we searched the human genome for all exact matches (at least 25 bp in length) with dissimilar flanking sequences. We chose to look only at identically matching sequences, or "cores", with precisely two copies, to simplify both the definition of the sequences under consideration and the interpretation of the results. We required the flanking sequences to be unrelated in order to ensure that we were not looking at a small exact patch within a long approximately duplicated region. After filtering out sequences with homologous flanks or with homology to expressed sequences (see methods), we were left with 32,057 of these paired sequences or "doublets" in the human genome (**Fig 1A**). Although we set no maximum length on the core sequences, 99.9% are less than 100 bp in length. In fact, over half are 25 bp long and their length distribution decays rapidly (**Fig 2A, 2B, 2C**).

Doublets have several interesting characteristics. First, the distribution of their inter-core distances is strikingly non-random (**Fig 3A**). We observe three populations of doublets: those that are extremely close together ("adjacent"; cores at most 100 bp apart); those with distances distributed around 1 kb ("nearby"; cores more than 100 bp and at most 10 kb apart); and those with cores more than 10 kb apart or interchromosomal ("remote"). In addition, there is a bias towards conservation of polarity: the adjacent doublets are always direct repeats, and the nearby doublets have both cores in the same polarity about 70% of the time. Not surprisingly, the remote doublets show no bias

towards conservation of polarity. We made essentially the same observations in mouse (**Fig 3B, 2D**).

The numbers of "adjacent" and "nearby" doublets are significantly larger than what can be expected by chance, even considering the biases associated with the overall number of doublets (**Fig 1C**). The vast majority of such doublets, which we collectively call "proximate", are extremely unlikely to be coincidental matches. However, it is difficult to discern whether the large number of "remote" doublets is a result of biases in genome sequence composition, or represents a more specific phenomenon. We have therefore concentrated our attention on proximate doublets. This decision is supported by the observation that core lengths are shorter among remote doublets than proximate doublets (**Fig 2A, 2B, 2C**), and the observation that remote doublet cores tend to be more AT-rich than proximate (data not shown). Adjacent doublets are comprised of two identical sequences separated by a short spacer sequence of 1 to 100 bp. Since their polarity is preserved, they can be viewed as a subclass of tandem repeats, loosely defined as direct repeats of approximate matches with little or no spacer. Some of our adjacent doublets are clearly tandem repeats of two units that appear to have a spacer sequence only because the repeat has been partially eroded by point mutations. It is possible that all of our adjacent doublets are variants of this type, and more of this class would have been found if we loosened our strict ascertainment criteria.

Nearby doublets, with long inter-core distances between 100 bp and 10000 bp, cannot be classified as degenerate tandem repeats. To study the dynamics of both adjacent and nearby doublets, we compared them to the draft *Pan troglodytes*

(chimpanzee) sequence. Of 3,083 doublets with inter-core distances less than or equal to 10 kb, we found 2,589 in which the outer-most flanks have clear homologues in the chimpanzee assembly. In most cases, the cores themselves are also present in chimp. However, in 307 cases, one of the two cores is missing in chimp, implying either a gain of a new copy in the human lineage or a loss in the chimp lineage (**Fig 4A, and supplemental Fig 1**).

In one nearby doublet, we see that both one core and the inter-core sequence are missing in chimp relative to human. This particular doublet likely represents a recombination-mediated loss of core and inter-core sequence in the chimp lineage, rather than a gain in the human lineage (**see doublet 643 in supplemental figure 1**). However, this example is an exception: in the rest of the nearby doublets, the second core in humans appears to be an insertion relative to chimp.

To unequivocally discriminate between gains and losses of copies, we selected six nearby human doublets for further investigation and used PCR to detect the presence of both cores in chimpanzee, gorilla, orangutan, macaque, spider monkey, and lemur individuals, as well as a set of humans of diverse ethnicity. For each doublet, one core was always missing in non-human primates, while the other was always present (**data not shown**). In one of these six doublets, portrayed in figure 4A, we found the variable core to be polymorphic within the human population (**data not shown**). These data strongly indicate that the nearby doublets seen in human but not in chimp arise by gain of a new copy.

Gains giving rise to nearby doublets in humans are most easily visualized as a simple insertion of a copy of a core into a nearby site with minimal alteration to the surrounding sequence. To look for further examples of these structures, we compared paralogous regions within the human genome. To this end, we identified those doublets that overlap the Alu family of transposons, and examined the doublet sequences to determine whether the cores are insertions relative to the Alu consensus sequences. Of 51 nearby doublets with cores that overlap Alu annotations, we found 41 cases where the core appears to be an insertion relative to the transposon (**Fig 4B, supplemental Fig 2**).

One possible source of nearby doublet generation is segmental duplication. Nearby doublets could be the remnants of old segmental duplications, where only a short exact match remains. Although these should have been eliminated through our filtration process, we used several tests to determine whether they were still a source of doublets. Segmental duplications are preferentially located near centromeres and telomeres in humans (*18*), and so as a first test we compared the chromosomal distribution of segmental duplications to that of doublets. We did not find any positive correlation between proximate doublet location and these chromosomal structures (**supplemental Fig 3**). Furthermore, we did not observe any clustering of proximate doublets within any 100 kb partitions of the human genome, which strongly argues against their origin as remnants of larger segmental duplications (**supplemental Fig 4**). As a final test, we compared the length distribution of "young" doublet cores that are found in human but not chimp to the length distribution of conserved cores. If doublets originated in longer sequences, then younger doublets should be longer on average than old ones (**Fig 2E,**

2F). In fact, the distributions are very similar, with a slight length increase in young doublets presumably due to the decreased number of point mutations in young sequences.

To search for further clues of the origins of the doublets, particularly the nearby doublets, we compared the content of the inter-core sequence to randomly chosen genomic intervals of similar length, and also randomly chosen genomic intervals from annotated segmental duplications. We examined these intervals for the uniqueness of their constituent 18-mers, and overlap with the following types of genomic annotations: RepeatMasker, RefSeq genes, and segmental duplications. With respect to uniqueness, inter-core intervals are essentially indistinguishable from random genomic intervals, and clearly very distinct from segmental duplications (**supplemental Table 1**). Inter-core intervals are drastically reduced for annotations as "segmental duplications", and slightly under-represented for annotated repeats and genes (**supplemental Table 1**).

For clarity, we have examined a set of precisely defined short exact matches. Another group (Achaz et al., *19*), has studied more loosely defined approximate repeats in a range of organisms. For algorithmic simplicity, they too looked only at pairs of duplicated sequences. Although their data presumably encompass segmental duplications and pseudogenes as well as doublets, a bimodal distance relationship similar to what we have observed can be weakly discerned.

Achaz et al. postulate that all of the repeats they found were generated by direct tandem duplications, and that more distantly separated pairs were "spread" apart by later insertions. We can reject this model because our sequence comparisons suggest that in many cases the nearby doublets can be viewed as an insertion of a "core copy" into an

existing target sequence with minimal collateral damage to the target. Furthermore, we have compared the distances between pairs of cores conserved in chimp and human, and find that this distance is tightly conserved (**supplemental Fig 6**). Not only is there no evidence of spreading, but the inter-core spacers are, if anything, underrepresented for the agents that might cause spreading, such as transposons and segmental duplications.

Achaz et al. hypothesize that the closest pairs undergo high frequencies of recombination and are consequently unstable. Our data are consistent with this idea. Although there are roughly equal numbers of adjacent and nearby doublets in humans, the doublets that have changed since human-chimp divergence are mainly adjacent. Of the proximate doublets in humans that are conserved in chimp, 68% are "adjacent". Of the proximate doublets that are "new" since chimp, 96% are adjacent (**Table 1**). This is further supported by the observation that the inter-core sequences are often lost in adjacent doublets, implicating a deletion event in the transition from two cores to one.

Much more of the genome may have arisen by this duplication process than is immediately apparent. By requiring exact identity between the two cores, we have missed much older and more divergent short duplications present in the genome. In fact, since only 6% of the proximate doublets are new since the divergence of human and chimp, we expect that most doublets are ancient in origin. Moreover, more than half of exact doublets have cores no longer than our minimum length, so we expect we missed shorter duplications. In fact, smaller sequences, with a minimum length of 21 bp rather than 25, have a similar inter-core distance distribution and there are at least four times as many of these (**supplemental Fig 5**).

These findings are important because they suggest that the mammalian genome can expand and remodel by local random copying. The genomic forces giving rise to the events we have observed may be responsible for the duplication and shuffling of small functional motifs that have been preserved in vertebrate evolution. The comparisons of doublets orthologous between human and chimp suggest that the short adjacent duplications may be "reversible," providing an inexpensive way for the species to rapidly explore the functionality of its local sequence space. In future studies, it will be interesting to relax the requirement of exact identity between cores, to gain further insight into the mutational dynamics of doublets.

As we have already mentioned, the distribution and character of mouse doublets is similar to what we observe in humans. We repeated our analysis in the genomes of *Caenorhabditis elegans*, *Drosophila melanogaster*, *Plasmodium falciparum,* and *Arabidopsis thaliana*. In *D. melanogaster* and *P. falciparum* we see too few paired matches to conclude whether they have the same characteristics as human doublets. In the other two genomes we see very significant numbers of doublets (**Fig 3C, 3D**). In *A. thaliana* doublets are mainly "adjacent", whereas in *C. elegans* doublets are mainly "nearby". These observations suggest that the mechanisms that give rise to doublets are fairly widespread among eukaryotic genomes, but that unknown factors alter the relative contribution of these mechanisms to the evolution of different species.

A model involving double-stranded breaks leaving 5' overhangs, subsequently repaired by filling-in, followed by nonhomologous recombination, can explain the adjacent doublets. Although we do not know of documented cases of this type of repair,

the model seems plausible. The nearby doublets are not so readily explained. Since they too preserve polarity, we may surmise that they too reflect a repair event, but polarity is not absolutely preserved, and different classes of proximate doublets predominate in different genomes, suggesting different types of repair are at play. We offer no mechanism for the much more abundant remote doublets, and in fact cannot offer persuasive statistical arguments that remote doublets are not coincidental. A finished assembly of the chimp genome would help resolve this issue. In any case, breakage-repair seems a likely mechanism whereby genomes sample and replicate their own composition, which, over long time, can lead to the amplification and dispersion of small functional motifs.

References:

1. International Human Genome Sequencing Consortium. (2001) *Nature* **409,** 860–921.
2. 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., *et al.* (2001) *Science* **291,** 1304–1351.
3. Deininger, P.L. & Batzer, M.A. (2002) *Genome Research* **12**, 1455-1465.
4. Prak, E.L. & Kazazian, H.H. (2000) *Nature Rev. Genet.* **1**, 134-144.
5. Vanin, E. (1985) *Annual Review Genetics* **19**, 253-272.
6. Samonte, R.V. & Eichler, E.E. (2002) *Nature Rev. Genet.* **3**, 65-72.
7. Ohno, S. (1970) *Evolution by Gene and Genome Duplication* (Springer, Berlin)
8. Chuzhanova NA, Krawczak M, Nemytikova LA, Gusev VD, Cooper DN. (2000). *Gene* **254**, 9-18.
9. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, *et al*. (2003) *Nucleic Acids Res.* **31**, 51-54.

10. Healy, J., Thomas, E.E., Schwartz, J.T. & Wigler, M. (2003) *Genome Research* **13**, 2306-2315.
11. Needleman, S.B. & Wunsch, C.D. (1970) *J. Mol. Biol*. **48**, 443-453.
12. Mouse Genome Sequencing Consortium. (2002). *Nature*, **420**, 520-562.
13. *C. elegans* Sequencing Consortium. (1998). *Science*, **282**, 2012-2018.
14. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al.* (2000). *Science*, **287**, 2185-2195.
15. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, *et al*. (2002) *Nature*, **419**, 498-511.
16. Arabidopsis Genome Initiative. (2000). *Nature* **408**, 796-815.
17. Jurka, J. (2000) *Trends Genet.* 16, 418-420.
18. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. (2002) *Science* **297**, 1003-1007.
19. Achaz, G., Netter, P. & Coissac, E. *Mol Biol Evol* **18**, 2280-2288.

Acknowledgements:

We thank Evan Eichler, Mike Zody, Tarjei Mikkelsen, Eric Lander, and Jerzy Jurza, for their helpful critical reading of our paper. We thank Eric Siggia, Casey Bergman, Izik Pe'er, Dana Pe'er, Guillaume Achaz, and Ira Hall for interesting discussion. We thank Lakshmi Muthuswamy for help in determining mouse genomic counts. This work was supported by grants to M.W. from the NIH and NCI (2R01CA078544; 5P30CA45508; 5R01CA81152; 5R21HG02606) NYU/DARPA F5239. M.W. is an American Cancer Society Research Professor. E.E.T. is a Farish-Gerry Fellow of the Watson School of Biological Sciences and is a HHMI predoctoral fellow. J.S. is supported by NIH 5T32CA09311. This research was also supported by

grants to B.M. from NSF's Qubic program, NSF's ITR (medium) program, DARPA's BioComp/Biospice program, the US air force, and New York State Office of Science, Technology & Academic Research (NYSTAR) program.

Figure Legends:

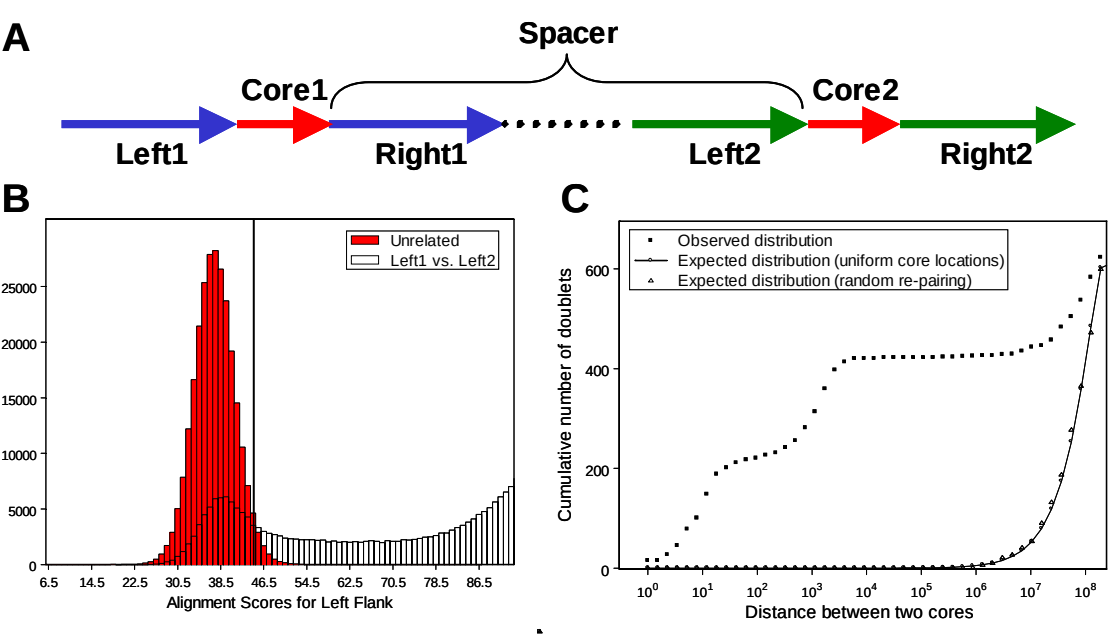
Figure 1. A – Anatomy of a doublet. Each doublet has two cores, which are identical (same polarity) or reverse complements (opposite polarity) and are at least 25 bp in length. Each core is associated with 100 bp of flanking sequence on either side (Left1, Right1, Left2, and Right2). These flanking sequences can overlap. The last 21 bp of Left1 and first 21 bp of Right1 must be unique in the human genome. If the two cores are on the same chromosome, the spacer is the sequence between the two cores, and its length is the inter-core distance. To be a doublet, the flanks cannot be homologous. B – Homology between flanks. The hatched histogram shows the distribution of alignment scores from comparing Left1 to Left2 (see part A). The red plot shows the distribution of alignment scores from comparing sequences that should be unrelated – Left1 and the reverse complement of Right2. The homology threshold is depicted as a vertical black line. C – Distribution of inter-locus distance of doublets on chromosome two. The observed inter-locus distribution (black squares) is compared with two random models for doublet occurrences: the same number of total doublets, each core occurring uniformly and independently at random across the chromosome (empty circles); Cores occurring at observed locations, but randomly re-paired into doublets (empty triangles).

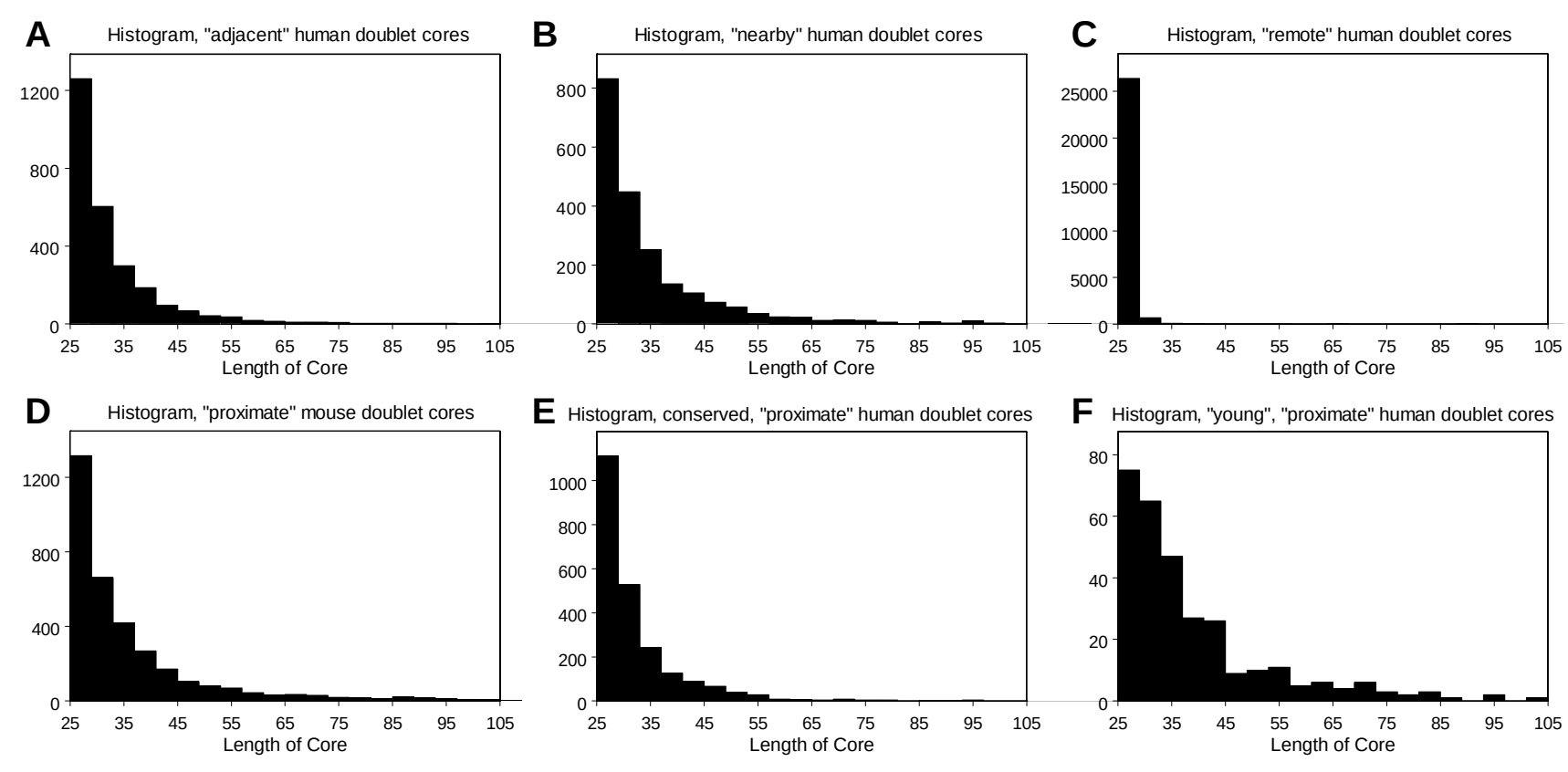
Figure 2. Core length distributions are shown for several different populations of doublets. For each of these populations, a bin size of 4 bp was used to bin the core lengths, and the distribution is plotted above. A – 2,696 "Adjacent" human doublets with spacer lengths ≤ 100 bp. B – 2,077 "Nearby" human doublets with spacer lengths > 100 bp and ≤10 kb. C – 29,013 "Remote" human doublets with spacer lengths > 10 kb. D – 3, 430 "Proximate" mouse doublets with spacer lengths ≤ 10 kb. E – 2,283 "Proximate" human doublets which are shared between the chimpanzee and human genomes. F – 306 "Proximate", "young" human doublets, where one of the two cores of the doublet is missing in chimp.

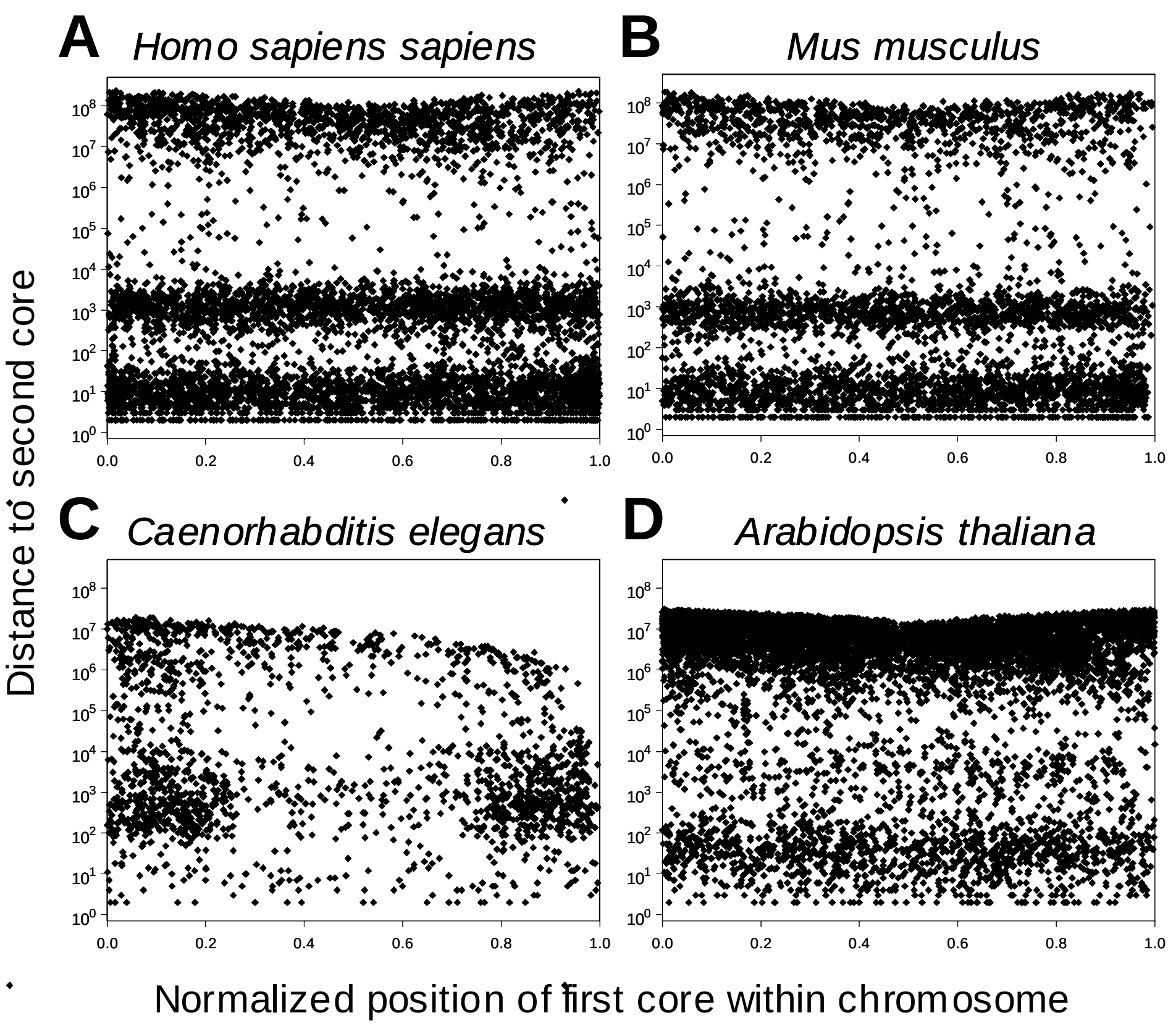
Figure 3. For four different organisms, the distance between the two cores of a doublet is plotted versus the chromosomal position of one of the cores. Doublets are only included if both cores are on the same chromosome. This graph represents merged data from all of a particular organism's chromosomes. Normalized positions are chromosomal position divided by chromosome length.

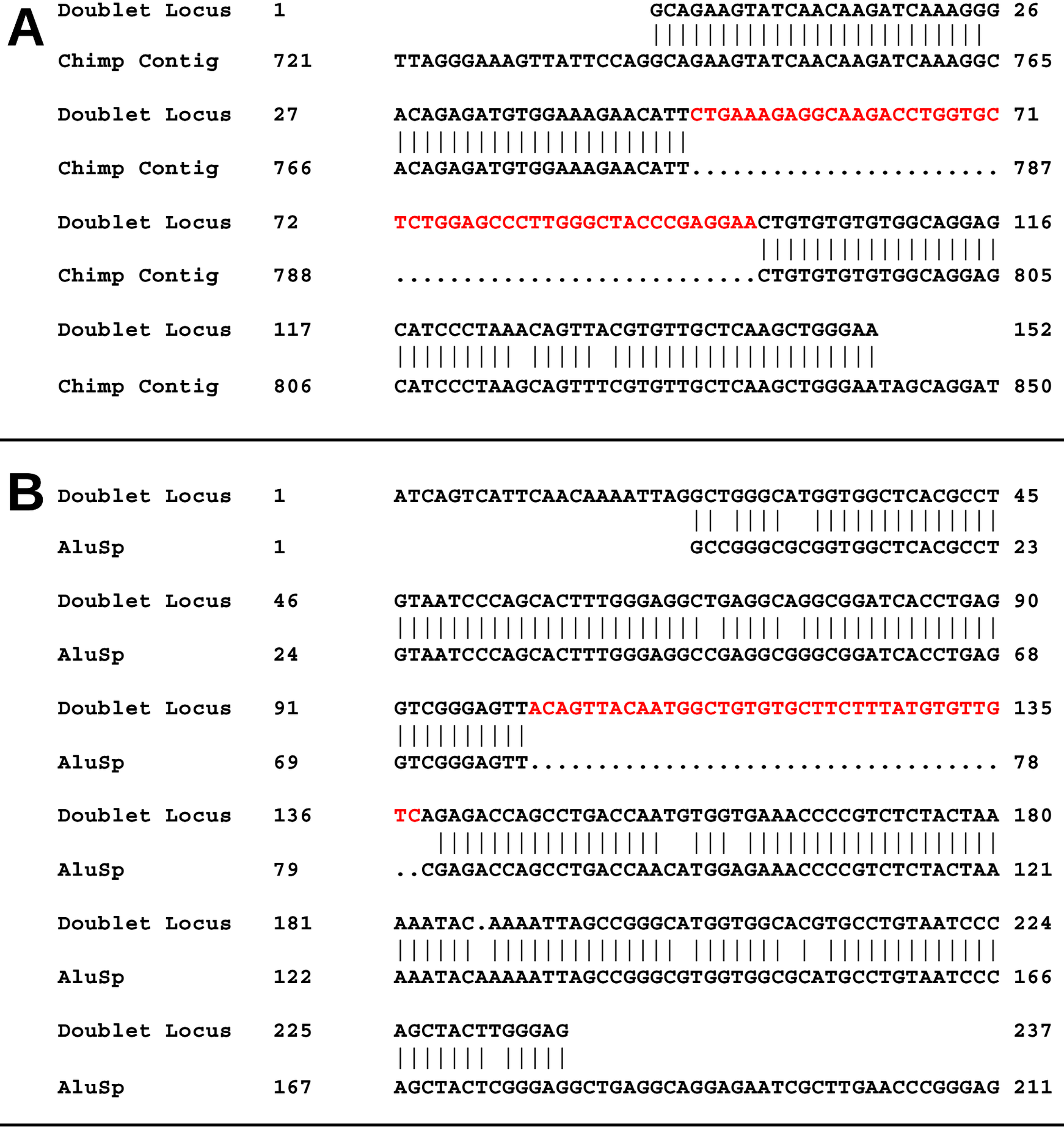
Figure 4. A – An alignment between one core of a human doublet and *Pan troglodytes* sequence. The core sequence (highlighted in red) is clearly missing in the orthologous region of chimpanzee. This sequence is polymorphic within human populations; the inserted core has an allele frequency of approximately 46%. B – An alignment between one locus of a different doublet and an *AluSp* transposon consensus sequence. The core is clearly an insertion relative to the consensus.

Table 1. Conservation of human proximate doublets in chimp. Proximate doublets are all those with inter-core distances between 1 bp and 10 kb. Nearby and adjacent are sub-classes of proximate doublets with inter-core distance ranges of 101-10,000 bp and 1-100 bp respectively. Adjacent doublets are more likely to be "young" than nearby doublets. However, young adjacent doublets are usually missing their inter-core spacer sequences in chimp, indicating that they are probably either tandem sequences (with an apparent spacer created by point mutation) or that they are the result of chimp-specific deletion rather than human-specific insertion. For more detailed alignments, see supplemental figure 1.


A
Spacer
Core1
Core2
Left1
Right1
Left2
Right2
B
Unrelated
Left1 vs. Left2
25000
20000
15000
10000
5000
0
6.5
14.5
22.5
30.5
38.5
46.5
54.5
62.5
70.5
78.5
86.5
Alignment Scores for Left Flank
C
Observed distribution
Expected distribution (uniform core locations)
Expected distribution (random re-pairing)
Cumulative number of doublets
600
400
200
0
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7
10^8
Distance between two cores


A
Histogram, "adjacent" human doublet cores
1200
800
400
0
25 35 45 55 65 75 85 95 105
Length of Core
B
Histogram, "nearby" human doublet cores
800
600
400
200
0
25 35 45 55 65 75 85 95 105
Length of Core
C
Histogram, "remote" human doublet cores
25000
20000
15000
10000
5000
0
25 35 45 55 65 75 85 95 105
Length of Core
D
Histogram, "proximate" mouse doublet cores
1200
800
400
0
25 35 45 55 65 75 85 95 105
Length of Core
E
Histogram, conserved, "proximate" human doublet cores
1000
800
600
400
200
0
25 35 45 55 65 75 85 95 105
Length of Core
F
Histogram, "young", "proximate" human doublet cores
80
60
40
20
0
25 35 45 55 65 75 85 95 105
Length of Core


A
Homo sapiens sapiens
B
Mus musculus
C
Caenorhabditis elegans
D
Arabidopsis thaliana
Distance to second core
Normalized position of first core within chromosome
10^0
10^1
10^2
10^3
10^4
10^5
10^6
10^7
10^8
0.0
0.2
0.4
0.6
0.8
1.0


A
Doublet Locus 1 GCAGAAGTATCAACAAGATCAAAGGG 26
Chimp Contig 721 TTAGGGAAAGTTATTCCAGGCAGAAGTATCAACAAGATCAAAGGC 765
Doublet Locus 27 ACAGAGATGTGGAAAGAACATTCTGAAAGAGGCAAGACCTGGTGC 71
Chimp Contig 766 ACAGAGATGTGGAAAGAACATT..... 787
Doublet Locus 72 TCTGGAGCCCTTGGGCTACCCGAGGAACTGTGTGTGTGGCAGGAG 116
Chimp Contig 788 .....CTGTGTGTGTGGCAGGAG 805
Doublet Locus 117 CATCCCTAAACAGTTACGTGTTGCTCAAGCTGGGAA 152
Chimp Contig 806 CATCCCTAAGCAGTTTCGTGTTGCTCAAGCTGGGAATAGCAGGAT 850
B
Doublet Locus 1 ATCAGTCATTCAACAAAATTAGGCTGGGCATGGTGGCTCACGCCT 45
AluSp 1 GCCGGGCGCGGTGGCTCACGCCT 23
Doublet Locus 46 GTAATCCCAGCACTTTGGGAGGCTGAGGCAGGCGGATCACCTGAG 90
AluSp 24 GTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACCTGAG 68
Doublet Locus 91 GTCGGGAGTTACAGTTACAATGGCTGTGTGCTTCTTTATGTGTTG 135
AluSp 69 GTCGGGAGTT..... 78
Doublet Locus 136 TCAGAGACCAGCCTGACCAATGTGGTGAAACCCCGTCTCTACTAA 180
AluSp 79 ..CGAGACCAGCCTGACCAACATGGAGAAACCCCGTCTCTACTAA 121
Doublet Locus 181 AAATAC.AAAATTAGCCGGGCATGGTGGCACGTGCCTGTAATCCC 224
AluSp 122 AAATACAAAAATTAGCCGGGCGTGGTGGCGCATGCCTGTAATCCC 166
Doublet Locus 225 AGCTACTTGGGAG 237
AluSp 167 AGCTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGGGAG 211

	2 or more cores in chimp		"young" 1 core with or without spacer		"young" with spacer in chimp		"young" without spacer in chimp	
Proximate	2283	*100.0%*	306	*100.0%*	27	*8.8%*	279	*91.2%*
Nearby	728	*31.9%*	13	*4.2%*	12	*92.3%*	1	*7.7%*
Adjacent	1555	*68.1%*	293	*95.8%*	15	*5.1%*	278	*94.9%*