RESEARCH

Transposable Element Subfamily Annotation is Unreliable in Biological Replicates

Kaitlin Carey[1,†], Gilia Patterson[1,2,†] and Travis J Wheeler[1*]

*Correspondence:
travis.wheeler@umontana.edu
[1]Department of Computer Science, University of Montana, 32 Campus Drive, Missoula, MT
Full list of author information is available at the end of the article
†Equal contributors

Abstract

Background: Transposable element (TE) sequences are classified into families based on the reconstructed history of replication, and into subfamilies based on more fine-grained features that in some cases capture family history, and in other cases are simply intended to improve annotation sensitivity. We evaluate the reliability of annotation with common subfamilies by assessing the extent to which subfamily annotation is reproducible in replicate copies created by segmental duplications in the human genome, and in homologous copies shared by human and chimpanzee.

Results: We find that standard methods annotate over 10% of replicates as belonging to different subfamilies, despite the fact that they are expected to be annotated as belonging to the same subfamily. Point mutations and homologous recombiniation appear to be responsible for some of this discordant annotation (particularly in the young Alu family), but are unlikely to fully explain the annotation unreliability.

Conclusions: The surprisingly high level of disagreement in subfamily annotation of homologous sequence highlights a need for further research into definition of TE subfamilies, methods for representing subfamily annotation confidence of TE instances, and approaches to better utilizing such nuanced annotation data in downstream analysis.

Keywords: Transposable Elements; Interspersed Repeats; Subfamilies; Segmental Duplications

Introduction

Transposable elements (TEs) are usually annotated within a genome using a tool, such as RepeatMasker [1], that compares a genome to a library of known TEs, such

as Repbase [2]. In such a library, TE remnants are classified into families and subfamilies based on the history of replication and divergence reconstructed from their sequences. This history can be complex, with numerous replication bursts leading to clusters of related TEs [3, 4]. Standard practice is to reconstruct and define subfamilies based on shared diagnostic sequence variation [5, 6] within such bursts. Because annotation with these subfamilies is believed to give some indication of a sequence's historical context, it is important that such annotation be reproducible. We evaluate the reliability of subfamily annotation, focusing attention on the two families with the largest distribution of subfamilies found in the human genome: Alu and L1. Alus are young and short, and carry a significant risk of cross-annotation due to straightforward mechanisms such as random point mutation and gene conversion; L1s are older, longer elements with complex histories, and discordance is likely due to more complex mechanisms such as recombination and incomplete cataloging of subfamilies.

Adjudication of subfamily annotation candidates

When annotating TEs, the common strategy is to use sequence alignment software to compare genomic sequence to each family in the database of TE elements. When a collection of TE elements within this database are similar to each other, they will all tend to align well to the same genomic sequence, so that one genomic region may attract many competing annotations. The common strategy for selecting which annotation is prefered (a process that we call *adjudication*) is to select a single highest-scoring alignment. In this work, we do not consider how subfamily selection impacts downstream analysis; we do, however, work under the assumption that it is desirable that assignment to a particular subfamily is reliable.

Biological replicates to assess subfamily annotation reliability

We demonstrate that annotations produced by a common tool and TE library pairing (RepeatMasker and Repbase) are not entirely reproducibile. We evaluate the reliability of subfamily annotations using biological replicates, in which two identical copies of a TE have independently and randomly accumulated mutations. If subfamily annotation is reliable, both copies should be annotated as belonging to the same subfamily. We use two sources of biological replicate: (i) segmental duplications in the human genome and (ii) human-chimp species divergence. In both

data sets, we find that >10% of Alu and L1 TE pairs are classified into different subfamilies; we call this *discordant annotation*. Though our analysis is performed with one tool/database combination, these results are expected to generalize to any annotation tool with similarly organized subfamilies.

The role of point mutations and homologous recombination in annotation discordance

In the case of younger TE families (e.g. Alus), subfamilies are usually differentiated based on a small number of diagnostic nucleotides. We find that, under a simple substitution model, nearly half of observed discordance may simply be the result of accumulated point mutations. This raises concerns that the subfamilies may be too similar to reliably distinguish instances of one from the other.

A complicating factor in counting discordant annotations is that TEs are hot spots for non-allelic homologous recombination [7, 8], due to the presence of many highly similar cousin sequences belonging to either the same or similar subfamilies. A common scenario is that a double-stranded break in one chromosome (the acceptor) is repaired using a similar sequence from another location (the donor) as a template [9, 10]. If the break occurs in one TE of a replicate pair, the donor sequence may be one of many cousin TE instances, possibly one from a different related subfamily. In this case the annotation system would be correct in assigning the pair of TEs to different subfamilies. We demonstrate that this form of recombination is likely responsible for some (but probably not all) observed annotation discordance.

Implications

The high level of discordance in annotation of biological replicates points to a general lack of reliability in annotation of genome sequence into subfamilies. This observation guides our recommendation that TE researchers should be cautious in developing subfamily libraries, and in their application to genome annotation. When developing subfamily libraries, we recommend that TE researchers use measures of subfamily reliability to decide how to split families into subfamilies. When using subfamily libraries for annotation, we recommend that annotation software provide a measure of annotation confidence, and that tools used in downstream analyses account for variability in annotation confidence.

Results

TE subfamily annotation shows high level of discordance in biological replicates

To understand the reliability of subfamily annotation, we have analyzed two human datasets that serve as biological replicates: duplicates due to segmental duplication, and duplicates due to speciation. We call pairs that are classified into different subfamilies despite being derived from a common TE insertion event *discordant annotation*, and found that more than 10% of TE pairs are discordantly annotated in both datasets.

TE annotation discordance in segmental duplications

One source of biological replicates is segmental duplication [11], in which long (>1000 base pair) regions of DNA have been duplicated one or more times. When a TE is present within a region that is duplicated, the TE instance in the original segment and the instance in the duplicate segment are biological replicates (Figure 1).

We identified all instances where a TE was copied as part of a segmental duplication, restricting our analysis to segments that duplicated only once. We excluded instances in which one TE remnant was less than 50 base pairs long, a TE in one segment overlapped multiple TEs in the other segment, or a TE in one segment overlapped less than 80% of a TE in the other segment. There were 16,962 instances of these straightforward TE duplications. A pair was labeled as discordant if RepeatMasker's annotation placed the TEs into different subfamilies. The extent of discordant annotation varied among families (Table 1), but was high in both younger TE families (Alu, 12.4%) and older families (L1, 14.1%). This table focuses on Alu and L1 because these are families for which subfamilies are intended to provide some insight into biology; in these cases, it is particularly important that subfamily annotation be reliable. Some other families (e.g. MIR, L2, MLT in human) may consist of subfamilies created simply to improve annotation sensitivity; these subfamilies also demonstrate high discordant annotation within our segmental duplication analysis (e.g. MIR=10.4%, L2=15.7%, and MLT=6.7%).

Table 2 shows pair-annotation relationships for the three main types of Alu subfamilies: AluJ, AluS, and AluY. In all cases, the large majority of pairs are anno-

tated with matching subfamilies; but when discordant annotation is observed, it is common that the mis-matched pair crosses from one subfamily type to another.

TE annotation discordance in TEs shared by humans and chimps

Another source of biological replicates is TEs present in the common ancestor of human and chimpanzee. The pair of homologous TEs diverging from the common ancestor should be identified as belonging to the same subfamily. We correlated RepeatMasker annotation of TEs with homologous segments in the whole genome alignment of human (hg38) and chimp (panTro4) from the UCSC Genome Browser [12], using the UCSC liftOver tool. The rates of discordant annotation in homologous pairs of Alu and L1 TEs, summarized in Table 3, are similar to the rates in segmental duplications.

Drift via point mutations may explain some discordance in younger subfamilies

In cases where two TE subfamilies are distinguished by only a small number of diagnostic nucleotide substitutions, it is possible that a TE instance belonging to one subfamily will accumulate random point mutations at those diagnostic sites, leading to a change in annotated subfamily. Specifically, if half of the diagnostic sites switch from agreeing with the consensus for one subfamily to agreeing with the consensus for a single other subfamily, annotation may shift.

We quantified the expected frequency of such random subfamily drift, using a simple point mutation model. The model assumes that, after initial insertion of a TE instance in the genome (Fig 1b), the probability of a diagnostic site mutating away from the subfamily's diagnostic nucleotide is simply the observed percent divergence between the subfamily consensus and individual instances of that subfamily; further, assuming a mutation occurs, the model assumes all resulting nucleotides are equally likely. We focused on Alu families, since these are young and the most apt to endure subfamily adjustment due to point mutations of diagnostic sites; see Methods for details. Probability of subfamily change due to drift was computed for each subfamily drifting to each other subfamily, and a weighted average was computed for each of AluJ, AluS, and AluY, based on the expected frequency of initial subfamily membership.

The results in Table 4 show that $\sim 7-8\%$ of inserted AluS and AluY subfamily instances are expected to mutate such that they agree with other Alu subfamilies, i.e.

are expected to produce discordant annotation; these may explain $\sim \frac{1}{2}$ of observed discordance. These changes are essentially always expected to occur within-type, so do not explain between-type changes (e.g. from AluY-type to AluS-type). The large majority of changes are due to (i) promiscuous interchange within a small clique of nearly-identical AluS subfamilies (AluSg, AluSz, AluSx, AluSx1, and AluSx3) and (ii) within another small AluY clique (AluYc, AluYf1, and AluYm1, and the AluY subfamily). Older AluJ subfamilies appear to be unlikely to convert due to point mutations. The subfamiles in these cliques are also responsible for much of the observed discordance.

The possible role of homologous recombination in discordant annotation

We also considered the possibility that homologous recombination (as, for example, in gene conversion) might be responsible for observed discordance. For a segmental duplication, the alignment of the segments shows some divergence since the duplication event, and it is expected that percent identity should be fairly consistent across the entire segment. After a recombination event, the whole-segment alignment is expected to show a reduced pairwise identity at the recombined region relative to the surrounding segmental duplication.

To gain some insight into the frequency of recombination, we applied a simple test to identify these significant dips in percent identity. We computed the percent identity for each full duplicated segment pair and for non-overlapping length-100 windows within TEs in those pairs. To identify windows with significantly lower identity than the background identity of the segment, we computed a the binomial CDF and adjusted for multiple testing (due to multiple windows) by Bonferroni correction. Table 5 presents the proportion of TE pairs containing at least one low-identity window ($P < 0.001$), and shows that discordant pairs are much more likely than concordant pairs to manifest this signal of recombination. Results are presented for L1 pairs and for Alu pairs broken out to the three types (AluJ, AluS, and AluY; in order from oldest to youngest[13]). Even among the most recombination-rich subfamilies (AluY), fewer than 30% of discordant pairs show a signal of having endured recombination. Though the precise percent of recombination is likely wrong due to the simplified model, the test highlights the much higher apparent recombination

in discordantly-annotated pairs, suggesting that recombination may be the cause of some observed discordance.

Subfamily annotation confidence can be quantified, reflects reliability

When a TE family is represented by several highly similar subfamily sequences, an instance of the family belonging properly to one subfamily may align with high score to many or all of the subfamily elements. An annotation pipeline must pick from among these high-scoring candidate annotations. When scores of these competing annotations are similar, the standard annotation-based-on-highest-score strategy overstates confidence. An extreme example is in the case where two alignments supporting competing annotation have the same score: confidence in assigning the sequence to one subfamily or the other should be no greater than 50%, since either one is an equally good option.

Using a simple calculation of annotation confidence based on the ensemble of competing annotations (see Methods), we find that discordant Alu annotations in the segmental duplication dataset show significantly lower confidence in at least one of the pair of annotations than is seen in concordant annotations (Figure 2; P-value $9.4 * 10^{-116}$ according to the Kolmogorov-Smirnov test). For each pair, the maximum-confidence annotation was identified for both elements, and the element with the lower confidence was selected; among concordant pairs, the median of these less-confident elements showed 73.4% confidence, while the median for discordant pairs was 52.8%. These results show that discordant pairs are likely to be the result of uncertain annotation.

Discussion

Our study highlights problems with reliability in subfamily annotation. The causes of the high levels of unreliability are multi-faceted. Previous work [14] has raised concern that the subfamilies incorporated into standard repeat libraries do not accurately reflect the complex histories of the families. This likely contributes some of the inconsistent subfamily annotation. Consider, for example, the case in which a master element produces a number of instances, but a subfamily is not created for that set; in this case, all those instances will be annotated as belonging to some subfamily that arose in time *close* to the true subfamily. If there are two *close* subfamilies, then two duplicated copies may be assigned to those subfamilies

based on random non-discriminatory mutations. This would be expected to produce reduced-confidence and possibly-discordant annotations, so may explain much of our observed discordance. We suggest that biological replicates such as those described here can serve as useful markers for identifying challenging subfamily definitions.

We have described a simple method for identifying possible instances of recombination, but caution that this method should not be used to quanitify recombination rates - it is primarily useful as a method for showing differences in apparent recombination rates in discordant and concordant pairs. Recombination is most likely to occur when the donor sequence is highly similar to the acceptor sequence [15], so that many cases of recombination are expected to leave little trace in the form of sequence identity shifts. We also highlight that our analysis does not include recombination events that lead to the deletion of a TE instance, as such events would by definition not leave a pair that could be discordantly annotated. For a similar reason, our analysis only captures recombination events involving two break points surrounding a single region within the TE, since a recombination with a single breakpoint would split a segmental duplicate region in half, and thus escape our analysis pipeline.

Here, we have aimed to quantify unreliability through discordance in biological replicates, but most TE instances are not part of a biological-replicate pair. We caution that subfamily annotations are not infallible, and recommend two mitigation strategies. First, we suggest that TE subfamily library creation methods should explicitly make use of biological replicates to identify subfamily sets with high interchange rates, and consider collapsing those into a single subfamily to improve reliability. Second, we believe that annotation uncertainty should be explicitly represented within genome TE annotation software, so that downstream analyses can optionally make informed decisions about how to utilize proposed annotations.

Finally: our analysis has focused on subfamilies designed to represent the biology of TE instances (age, species-specificity, evolutionary history, etc.), specifically on human Alu and L1 subfamilies. Some subfamilies (such as MIR and L2 subfamilies) in databases like RepBase are used simply to increase annotation coverage by representing different regions of sequence space. In these cases, subfamily assignments should also be used with caution, but for the more mundane reason that those sub-

families only communicate something about the search mechanism, not about the biology of the sequence element.

Methods

Discordant annotation of TEs in segmental duplications

Our segmental duplicate analysis incorporated two independent databases: (1) segmental duplications from [16, 17] and (2) transposable elements from RepeatMasker [1]. Both datasets annotated human genome hg19. RepeatMasker results were from RepeatMasker open-4.0.5, using RepeatMasker Repbase Library 20140131 [2]. The database of segmental duplications consists of 25,800 pairwise alignments. All duplications are greater than 1,000 base pairs long and at least 90% identical, so the duplications probably occurred in the last 40 million years [16]. In the entire genome, RepeatMasker identified 5,467,457 TE remnants classified into 1,183 different subfamilies.

Segmental duplications cataloged by Bailey et al. [16, 17] (hg19) were correlated with TE annotations from RepeatMasker [1] (also hg19) based on Repbase subfamily consensus sequences [2]. TE duplicate pairs were identified based on the sequence alignment captured in the segmental duplication data from [16], which include segment context beyond the length of paired TE instances. To avoid TEs within segments with complicated histories, we restricted our analysis to segmental duplicates with only two copies, and found on canonical chromosomes 1-22, X, and Y. We took several steps to filter TE pairs that might be the result of independent insertions and so are not biological replicates. We considered only TE pairs in which both copies were longer than 50 nucleotides long, and at least 80% of the length of each copy was covered by the other, to avoid cases in which one copy is differentially classified based solely on being much shorter than the other. Further, we retained only TE pairs in which the pair are related in one contiguous alignment, to avoid cases of nuanced annotation due to, for example, a large insertion or deletion in one element of the pair following duplication. Finally, we ignored pairs in which at least one element was labeled ambiguously, with no specific subfamily (i.e. *Alu*). There were 16,962 instances of these straightforward TE duplications. When two aligned TEs were assigned by RepeatMasker to different subfamilies, we labeled the TE pair as discordantly annotated.

Comparison of TEs in human and chimp

We analyzed TEs annotated by RepeatMasker in the human genome (hg38) and the chimpanzee (*Pan troglodytes*) genome (panTro4). To find homologous pairs, we downloaded BED files of the annotations from the UCSC Table Browser [18] and used the liftOver tool [12] (downloaded on 14 April 2017) to convert the coordinates of TEs in the human genome into coordinates in the chimpanzee genome. We then used BEDTools [19] to find overlapping TEs and identified discordant annotation as before.

Subfamily conversion due to point mutations

For each subfamily, we estimate the probability that a nucleotide remains unchanged after a TE instance is inserted in the genome, $P(A_i)$, as the mean percent identity between the subfamily consensus and individual TE instances annotated by RepeatMasker. The probability that a specific site will change from the diagnostic (subfamily-specific) nucleotide is then $(1 - P(A_i))$. Assuming uniform chance of mutating to each of the other three nucleotides, the probability that a diagnostic site for subfamily i will change to the value associated with another subfamily j is $P(B_{ij}) = (1 - P(A_i))/3$; the remaining probability $P(O_{ij}) = 2(1 - P(A_i))/3$ is that chance that the diagnostic site will mutate away from the diagnostic value for subfamily i to a nucleotide other than the one that is diagnostic for subfamily j. Note that these mutation probabilities are since-insertion, not since-duplication, because the inserted element may accumulate some mutations suggestive of subfamily j prior to duplication.

Consider an Alu instance S belonging to subfamily i, and suppose that at the moment of insertion (at some point prior to duplication), it agreed with the consensus for i at all n diagnostic sites that differentiate the consensus of i from the consensus of subfamily j. One history that would cause one copy of S to be identified as belonging to subfamily j is for at least $n/2$ of those diagnostic sites to mutate to agree with family j, and that no diagnostic sites mutate to some other value that disagrees with both i and j. The probability of this occurring is the product of (i) the probability of no mutations of a diagnostic site to an *other* value, and (ii) the probability that fewer than $n/2$ diagnostic sites do not change from the value for i (the cumulative probability from the Binomial distribution):

$$P(S : i \to j, other = 0) = (1 - P(O_{ij}))^n \cdot B\left(\left\lfloor \frac{n-1}{2} \right\rfloor, n, P(c : i \to j)\right) \quad (1)$$

where the probability of a diagnostic site not changing from i to j, given that it also did not change to an *other* value is:

$$P(c : i \to j) = \frac{p(A_i)}{(1 - P(O_{ij}))} = \frac{3p(A_i)}{1 + 2p(A_i)} \quad (2)$$

and the Binomial CDF is:

$$B(x, n, p) = \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i} \quad (3)$$

More generally, if some number k of the i-diagnostic sites mutate to a non-informative *other* state, then only $(n-k)/2$ sites need to change to agree with j, so that the overall probability of S being identified as belonging to j based on diagnostic sites is:

$$P(S : i \to j) = \sum_{k=0}^{n-1} \binom{n}{k} (1 - P(O_{ij}))^{n-k} P(O_{ij})^k \cdot B\left(\left\lfloor \frac{n-k-1}{2} \right\rfloor, n-k, P(c : i \to j)\right) \quad (4)$$

Equation 4 was used to compute the probability of converting an instance of subfamily i to be recognized as belonging to subfamily j, for each pair of subfamilies. Then for each subfamily, a weighted average of these probabilities was computed for each type (J,S,Y), based on the observed frequency of each subfamily in the human genome (from http://repeatmasker.org).

Computing subfamily annotation confidence

We compute a measure of confidence that the annotated sequence belongs to a subfamily i by leveraging the probabilistic underpinnings of alignment scores.

Suppose we have $Q = q_1, q_2, ..., q_n$ competing subfamily annotations of genomic sequence t. If we define $P(q_i|t)$ as the probability that the true label of t is q_i, then the confidence that q_i is the correct label is

$$\text{Conf}(q_i|t) = \frac{P(q_i|t)}{\sum_j P(q_j|t)} \quad (5)$$

Assuming a uniform distribution over Q, $P(q_i|t) \propto P(t|q_i)$, so that

$$\text{Conf}(q_i|t) = \frac{P(t|q_i)}{\sum_j P(t|q_j)} \tag{6}$$

Under scoring matrices such as those used in RepeatMasker (based on crossmatch [20], the score for aligning a pair of letters is based on a log odds ratio [21], where the ratio is "the probability of the two letters aligning if the sequences are homologous" vs "the probability of two letters aligning if the sequences are not homologous". Typically, the real-valued log odds values are scaled by factor λ then rounded to the nearest integer value:

$$\text{score}(a,b) = \text{int}\left(\lambda \log \frac{P(a,b)}{P(a)P(b)}\right) \tag{7}$$

In an alignment with no insertions, the overall alignment score corresponds to a scaled log of the ratio of the probability of observing t if it is homologous to q_i vs the probability of observing t under a random (non-homology) model:

$$\text{score}(t,q_i) = \lambda \cdot \log \frac{P(t|q_i)}{P(t|R)} \tag{8}$$

Though typically these scores are integer-rounded, and alignment gap penalties are ad hoc (read: not derived from probabilities), we accept a simplifying approximation that they map to feasible probabilities [22], and utilize equation 5 in computing confidence values. This implies that

$$P(t|q_i) = P(t|R) \cdot 2^{\text{score}(t,q_i)/\lambda} \tag{9}$$

and after straightforward algebraic manipulation following substitution into equation5,

$$\text{Conf}(q_i|t) = \frac{2^{(\text{score}(t,q_i)/\lambda)}}{\sum_j 2^{(\text{score}(t,q_j)/\lambda)}} \tag{10}$$

This approach is admittedly simplistic, in that it assumes that all competing sequence alignments cover the same genomic range (what we've called t). Even so,

it allows us to inspect the relationship between confidence in subfamily annotation and the risk of discordance due to accumulation of point mutations.

Alignments used for annotation with RepeatMasker are produced using cross-match with custom scoring matrices based on regional GC content. For each segmental duplicate Alu pair (t_1, t_2), we first infer the λ value for the region-specific scoring matrix using the esl_scorematrix executable available via special compilation of the Easel sequence analysis library (http://bioeasel.org, implementing the method of [21]). Using this and alignment scores, we used equation 10 to compute estimates of the confidence for the best-scoring annotation for both t_1 and t_2, then captured the lower confidence value for that pair: $m = \min(\text{Conf}(\hat{q}|t_1), \text{Conf}(\hat{q}|t_2)$. We binned all such per-pair-minimum-confidence values and presented binned frequencies in Figure 2. The Kolmogorov-Smirnov test was used to determine if the two distributions were significantly different.

Computing recombination estimates

To estimate the rate of recombination in the segmental duplicate TE pairs, we compared the identity of pairs of TEs to the identity of the segments containing them. Each segmental duplication is described by an alignment of the sequence of the original segment and the sequence of the duplicated segment; these can be many kilobases in length, and can contain multiple TE instances. For each segmental duplication alignment, we computed the percent identity as the number of columns containing identical nucleotides in both sequence, divided by the number of non-gap columns in the alignment. Then for each TE pair p identified via the previously-described filtering process, we split the alignment of p into non-overlapping windows of 100 non-gap columns, starting at the first aligned position. We counted the number of identical columns c among these 100, and computed the binomial CDF (the probabilitiy of observing c or fewer identical columns out of 100, given the overall percent identity of the entire segmental duplication alignment). For each TE pair, we captured the smallest identity count among all windows, then subjected the corresponding binomial CDF value to Bonferroni correction to account for possibly-multiple windows. We reported TE pairs with $P < 0.001$, for both discordant and concordant pairs. We selected windows of length 100 because gene conversion events are typically at least 50 base pairs long [9]. Because we captured non-overlapping

windows, the final ($n \mod 100$) columns of an n column alignment are not used for the recombination estimate; this likely results in an under-estimate of recombination frequency

Availability of data and materials

The datasets analyzed and generated for this study, along with the scripts used for analysis, are available at http://wheelerlab.org/pubs/2020-discordant-CareyPatterson/CareyPatterson_suppl.tar.gz

Acknowledgements

We thank Thomas Jones for suggesting the segmental duplication data set as an appropriate biological replicate, and for more generally starting us down the path of thinking about subfamily reliability and annotation confidence. We also thank Robert Hubley and Arian Smit for providing a modified version of RepeatMasker's ProcessRepeats script to support confidence analysis, as well as insightful comments and suggestions during the course of our analysis. This work was supported by NIH grants U24 HG010136 (NHGRI) and P20 GM103546 (NIGMS).

Author's contributions

Gilia Patterson registered RepeatMasker annotation data with segmental duplications and whole genome human chimp alignments, developed the methods for identifying discordant annotations, and implemented the initial experiments to assess the role of recombination. Kaitlin Carey extended recombination analysis, developed experiments to assess the role of point mutations, and implemented software for computing and analyzing alignment confidence. Travis Wheeler conceived and coordinated the study, and performed some data analysis. All authors reviewed results, contributed to writing, and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Computer Science, University of Montana, 32 Campus Drive, Missoula, MT. [2]Institute of Ecology and Evolution, University of Oregon, 272 Onyx Bridge, Eugene, OR,.

References

1. Smit, A., Hubley, R., Green, P.: Repeatmasker open-4.0. available from http://www.repeatmasker.org
2. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J.: Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research **110**(1-4), 462–467 (2005)
3. Shen, M.R., Batzer, M.A., Deininger, P.L.: Evolution of the master Alu gene(s). Journal of Molecular Evolution **33**(4), 311–320 (1991)
4. Deininger, P.L., Batzer, M.A., Hutchison, C.A., Edgell, M.H.: Master genes in mammalian repetitive DNA amplification. Trends in Genetics **8**(9), 307–311 (1992)
5. Willard, C., Nguyen, H.T., Schmid, C.W.: Existence of at least three distinct Alu subfamilies. Journal of Molecular Evolution **26**(3), 180–186 (1987)
6. Price, A.L., Eskin, E., Pevzner, P.A.: Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. Genome Research **14**(11), 2245–2252 (2004)
7. Harpak, A., Lan, X., Gao, Z., Pritchard, J.K.: Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. Proceedings of the National Academy of Sciences **114**(48), 12779–12784 (2017)
8. Fawcett, J.A., Innan, H.: The role of gene conversion between transposable elements in rewiring regulatory networks. Genome biology and evolution **11**(7), 1723–1729 (2019)
9. Chen, J.-M., Cooper, D.N., Chuzhanova, N., Férec, C., Patrinos, G.P.: Gene conversion: mechanisms, evolution and human disease. Nature Reviews Genetics **8**(10), 762–775 (2007)
10. Sung, P., Klein, H.: Mechanism of homologous recombination: mediators and helicases take on regulatory functions. Nature reviews Molecular cell biology **7**(10), 739–750 (2006)

11. Cheung, J., Estivill, X., Khaja, R., MacDonald, J.R., Lau, K., Tsui, L.-C., Scherer, S.W.: Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biology **4**(4), 25 (2003)
12. Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M., *et al.*: The UCSC genome browser database: 2015 update. Nucleic Acids Research **43**(D1), 670–681 (2015)
13. Deininger, P.L., Daniels, G.R.: The recent evolution of mammalian repetitive DNA elements. Trends in Genetics **2**, 76–80 (1986)
14. Wacholder, A.C., Cox, C., Meyer, T.J., Ruggiero, R.P., Vemulapalli, V., Damert, A., Carbone, L., Pollock, D.D.: Inference of transposable element ancestry. PLoS Genet **10**(8), 1004482 (2014)
15. Mansai, S.P., Kado, T., Innan, H.: The rate and tract length of gene conversion between duplicated genes. Genes **2**(2), 313–331 (2011)
16. Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., Eichler, E.E.: Recent segmental duplications in the human genome. Science **297**(5583), 1003–1007 (2002)
17. Eichler, E.: Segmental duplication database. available from http://humanparalogy.gs.washington.edu/
18. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., Kent, W.J.: The UCSC table browser data retrieval tool. Nucleic Acids Research **32**(suppl 1), 493–496 (2004)
19. Quinlan, A.R., Hall, I.M.: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**(6), 841–842 (2010)
20. Green, P.: Cross_match, University of Washington, Seattle, USA (1994)
21. Yu, Y.-K., Altschul, S.F.: The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. Bioinformatics **21**(7), 902–911 (2005)
22. Frith, M.C.: How sequence alignment scores correspond to probability models. Bioinformatics **36**(2), 408–415 (2020)

Figures

Tables

Table 1 Rate of discordant subfamily annotation in human segmental duplications. TE pairs within segmental duplications were identified as described in the text (hg19 segmental duplicates from [17], TEs based on RepearMasker+Repbase annotation on hg19, filtered for length and quality of overlap between segmental duplicates). Among these, TEs belonging to the Alu and L1 families were considered, because these subfamilies are intended to represent biological history. Discordant annotations are those in which one element in a TE pair is assigned to one subfamily in Repeatmasker, while the other element in the pair is assigned to a different subfamily.

Family	# Subfamilies	# in genome	# pairs in filtered seg. duplications	# discordant	% discordant
Alu	47	1196725	10347	1290	12.4
L1	131	951429	6615	933	14.1

Figure 1 Segmental duplications generate replicate, divergent TE copies. (a) A TE (grey rectangle) is integrated into a segment of an ancestral genome (white rectangle). (b) Over time, this TE instance accumulates random mutations (black bars), diverging from the original inserted sequence. (b) A segment around the TE is duplicated, resulting in two identical copies of the TE instance. (c) Over time, each TE copy accumulates random mutations, so that they are no longer identical.

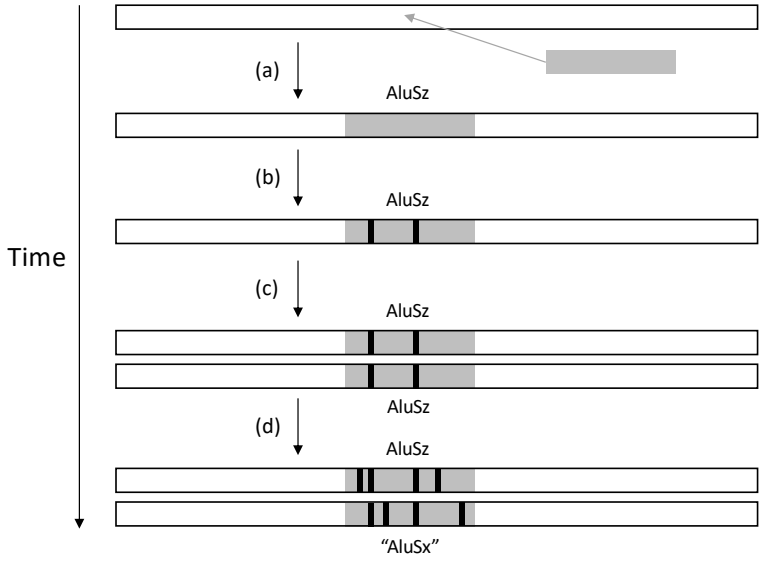

Table 2 Instances of pairs of TEs in segmental duplications for the three types of Alu subfamilies. In this table, each cell tallies the number of cases where one of the elements of a segmental duplicate TE pair belongs to the type specified by the row (from among AluJ, AluS, and AluY), and the other element belongs to type specified by the column. The first column captures concordant pairs (both entries share the same subfamily). In the next three columns, the diagonal captures the count of cases where one element belongs to one subfamily, and the other element shares the same type, but changes subfamily. Other cells represent a change in type. Mismatch percents (final column) exceed those in the previous table, because each discordant pair is double counted - the table is intended to highlight the differences in between-type discordance rates.

	concordant	non-match AluJ	non-match AluS	non-match AluY	Other, e.g. FRAM/FLAM	mismatch percent
AluJ	2308	254	69	10	134	16.8%
AluS	5776	69	629	77	24	12.2%
AluY	973	10	77	89	4	15.5%

Table 3 Subfamily counts and rates of discordant annotation based on homologous TEs in humans and chimps. RepeatMasker annotions of the hg19 human genome and panTro4 chimp genomes were paired using the UCSC liftover tool. Discordant annotation was identified as that in which lifted-over annotations differed at the subfamily level.

Family	# in human genome	#in chimp genome	# homologous pairs	% discordant
Alu	1209213	1118920	1093387	14.95 %
L1	1093751	960213	1050856	17.60 %

Figure 2 Annotation confidence tends to be lower in discordant TE pairs. For each segmental duplicate Alu pair identified as in the text, the confidence for the best-supported annotation of each element was captured, and the element with the smaller confidence was identified. Pairs were binned according to this smallest-maximum-confidence value. Each bin represents the fraction of all discordant- or concordant- pairs with that bin's smallest-maximum-confidence value.

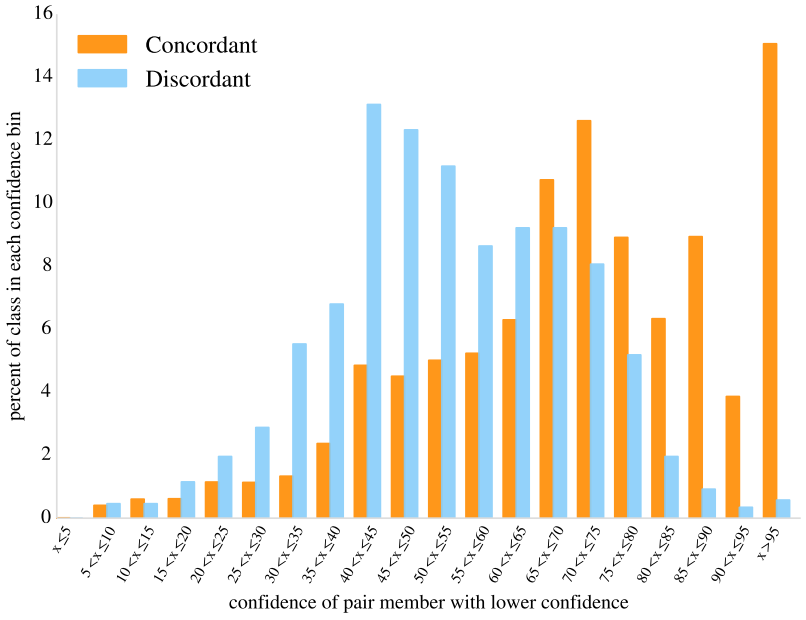

Table 4 Expected conversion between Alu subfamiles based on a simple model of substitution mutations. For each pair of subfamilies, we computed the probability of switching from one subfamily to another, based on the probability of changing the necessary number of diagnostic sites. Subfamily pairs were clustered by type, capturing the probability of converting from one of the types either within type (diagonal) or between types (off-diagonal). The final column is the sum of all probabilities of converting from the row header to any other subfamily.

	non-match AluJ (%)	non-match AluS (%)	non-match AluY (%)	combined mismatch
AluJ	0.72	2e-5	8e-10	0.73%
AluS	2e-3	7.76	4e-3	7.77%
AluY	3e-5	0.02	6.69	6.71%

Table 5 Discordant TE pairs show higher rates of apparent recombination. For each segmental duplicate TE pair, average segment percent identity was computed over the length of the segment. Then percent identity was computed for non-overlapping length-100 windows for each TE pair. We identified TEs containing windows with significantly reduced identity relative to the containing segmental duplication ($P < 0.001$, Bonferonni correction applied to account for possibly-multiple windows per TE). We quantified the observed rates for Alu and L1 subfamilies, computing apparent recombination in both discordant and concordant pairs.

	% of pairs w/ evidence of recombination		Mean pct id
	concordant	discordant	to consensus
L1	1.3%	2.8%	79.9%
AluJ	1.6%	5.8%	87.4%
AluS	3.7%	19.0%	93.1%
AluY	5.1%	27.6%	96.2%