Discordant Classification of Transposable Elements in Segmental Duplications Raise Concerns About Subfamily Definitions

Gilia R. Patterson
University of Montana - Missoula, gilia.patterson@umontana.edu

DISCORDANT CLASSIFICATION OF TRANSPOSABLE ELEMENTS IN SEGMENTAL DUPLICATIONS RAISE CONCERNS ABOUT SUBFAMILY DEFINITIONS

By

GILIA ROSE PATTERSON

Undergraduate Thesis
presented in partial fulfillment of the requirements
for the University Scholar distinction

Davidson Honors College
University of Montana
Missoula, MT

May 2016

Approved by:

Dr. Travis Wheeler, Faculty Mentor
Department of Computer Science

ABSTRACT

Patterson, Gilia, B.A., May 2016 Biology

Discordant Classification of Transposable Elements in Segmental Duplications Raise Concerns About Subfamily Definitions

Faculty Mentor: Travis Wheeler

Most of the human genome comes from transposable elements (TEs), sequences of DNA that can move and insert copies of themselves throughout the genome. TE sequences both inform and complicate analyses of genomes, so it is important that TEs are annotated completely and accurately. Remnants of TEs are annotated and classified into subfamilies based on their DNA sequences. A subfamily represents all the copies generated in a burst of replication by a few closely related TEs. Wacholder et al. (2014) suggested that the current methods for representing subfamilies are not accurate and should be reevaluated. We expand on this discussion and show that many TE sequences that should belong to the same subfamily are classified into discordant subfamilies. When a segment of genome with a TE remnant is duplicated, the TE remnants in each copy are replicates and so should be assigned to the same subfamily. We identified the location and subfamily of all TEs in known segmental duplications and found that a large fraction are assigned to different subfamilies, suggesting that the current method of classifying TEs splits them too much.

Discordant Classification of Transposable Elements in Segmental Duplications Raise Concerns About Subfamily Definitions

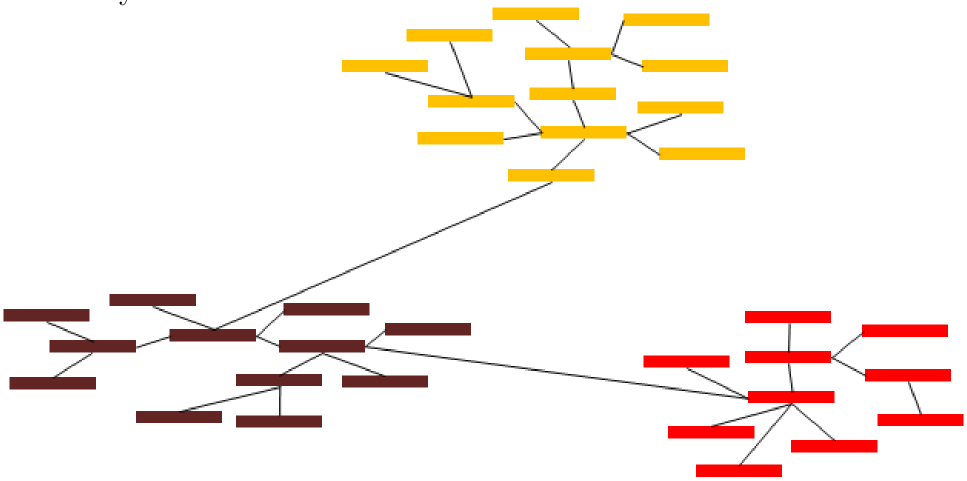
Background

Transposable elements (TEs) are sequences of DNA that can move and insert copies of themselves throughout the genome. TEs make up at least half of the human genome (Mills et al., 2007) and have shaped human evolution. For example, many fragments of TEs have been converted into genes or regulatory regions (Smit, 1999) and TEs have created new variants or caused disease when they inserted into genes or regulatory regions (Cordaux & Batzer, 2009). TE sequences also complicate many analyses of the human genome because they are repetitive and can cause false hits when searching a genome. Accurately annotating and classifying TEs is important for studying human evolution and for analyzing genomes.

Because TEs that insert into genes can be harmful, the human genome has evolved defense mechanisms, such as methylation, to prevent TEs from replicating (Yoder et al., 1997). An active TE may generate many copies in a burst of replication before these defense mechanisms stop it. If a sequence later escapes the defense, it may generate another burst of copies (Figure 1). TEs are classified into families of sequences that all come from one ancestor. Within families, sequences are classified into subfamilies that are meant to represent all the sequences generated by a few closely related TEs in one burst of replication.

A common method of annotating TEs is RepeatMasker (Smit et al., 2013). RepeatMasker uses a database of sequences each representing a family or subfamily and searches the genome for these sequences. Subfamilies are identified using CoSeg, a method that divides TE sequences into groups based on nucleotide similarity at diagnostic sites (Smit et al., 2013). Wacholder et al. (2014) evaluated CoSeg by reconstructing the evolutionary history of two families of TEs. They found evidence that subfamilies identified by CoSeg often do not represent the ancestry of TEs.

Figure 1: An example of the pattern of replication of a TE family. Each color represents sequences generated in the same burst of replication that should be classified into the same subfamily.



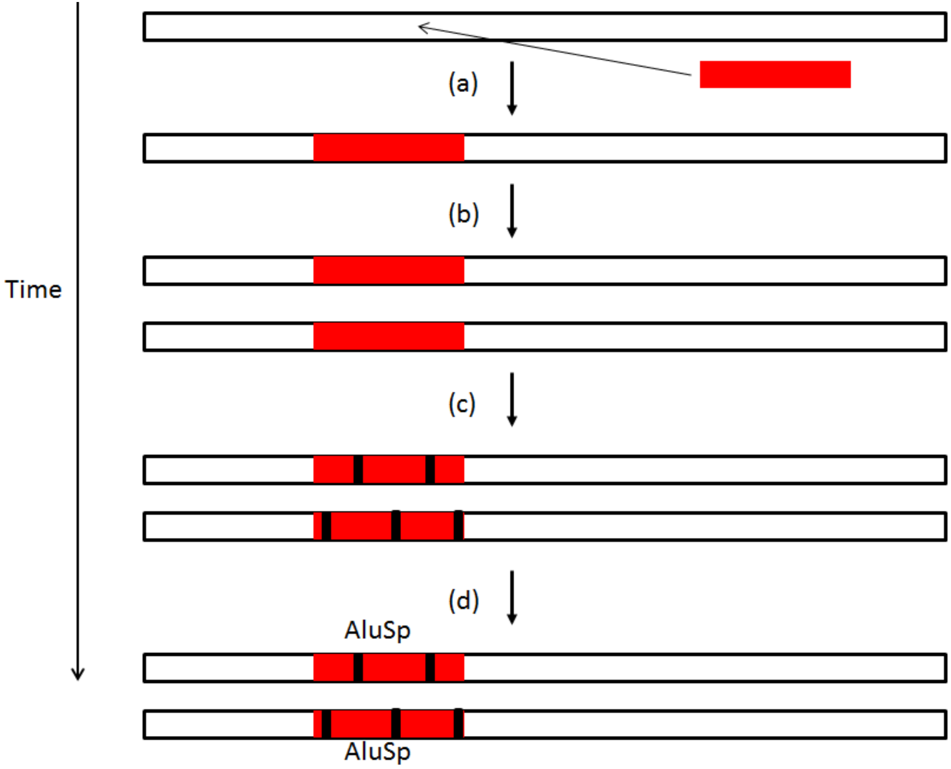
Another way to evaluate subfamilies is using replicates. If two copies of the same TE are placed in different locations in the genome and then annotated, they should be classified into the same subfamily. Segmental duplications provide replicates. Segmental duplications are long (>10 kb) regions of DNA that have been copied into another place in the genome. Often a TE is duplicated as part of the region, creating two replicate copies of a TE (Figure 2). Over time these copies will accumulate mutations, but they should still be annotated in the same subfamily.

We evaluate if the current method of dividing TE sequences into subfamilies is reproducible using a database of known segmental duplications in the human genome (Bailey et al., 2002).

Results

The current method of annotating and classifying TEs aligns representative sequences for each subfamily to the sequence of the genome. Representative sequences are generated by CoSeg (Smit et al., 2013). Using replicates of TEs provided by segmental duplications, we found that the current method has a surprising level of annotation that is not reproducible.

Figure 2: How segmental duplications generate replicate copies of transposable elements (TEs). (a) A TE (red rectangle) copies itself into a segment of the genome (white rectangle). (b) The region around the TE is duplicated; there are now two copies of the TE. (c) Over time, each copy accumulates random mutations (black rectangles). (d) RepeatMasker annotates each copy.



When a long segment of DNA containing a TE is duplicated, the TE in each copy of the duplication should be annotated in the same subfamily. We analyzed all known segmental duplications in the human genome and identified all instances where a TE was copied as part of the segment. We restricted our analysis to simple cases where it was clear that the TE was present before the segment duplicated and where the segment had only duplicated once. The rate of discordant classification is the number of instances where the two copies were annotated into different subfamilies divided by the total number of instances.

Table 1 shows rates of discordant classification for several common families. For example, in 15.22% of instances where an Alu was copied, the TEs were classified into different subfamilies. Alu is the most common family of TEs in the human genome. It is only 65 million years old (Batzer & Deininger, 2002) and is still actively generating new copies

(Mills et al., 2007). Older families of TEs showed similar rates of discordant classification. For example, in 10.49% of instances where a MIR was copied, the TEs were classified into different subfamilies. MIRs are 130 million years old (Murnane & Morales, 1995), much older than Alus. Over all families, 11.37% of TE copies were classified into different subfamilies.

Table 1: The number of subfamilies, number of TEs in the genome, number of instances of TE copies in the database of segmental duplications, percent of instances where TEs were classified into different subfamilies (% discordant), and age for six major families of TEs.

Family	# Subfamilies	# in genome	# copies in segmental duplications	% discordant	Age (million years)
Alu	47	1196725	16344	15.22	65 (Batzer & Deininger, 2002)
MIR	5	598863	5167	10.49	130 (Murnane & Morales, 1995)
L1	131	951429	10974	14.05	150 (Cordaux & Batzer, 2009)
L2	10	526188	4080	15.91	
MLT	70	273357	3183	5.84	
Tigger	45	125710	741	4.45	
Overall	1183	5467457	52777	11.37	

Tables 2 and 3 show the details of discordant classification for all MIR subfamilies (Table 2) and six of the most common Alu subfamilies (Table 3). We counted the number of duplicated copies where the two TEs were classified into each combination of subfamilies. Numbers in the diagonal of the table are instances where both copies were classified into the same subfamily and numbers off the diagonal are instances where TEs were classified into different subfamilies. We do not know which TE was classified correctly, so we place counts only in the upper triangle. For example, in 231 instances one copy was annotated as a MIR and the other copy was annotated as a MIRb. These 231 include both instances where the TEs should be MIRs and instances where the TEs should be MIRbs.

Table 2: Instances of TE copies in segmental duplications for the MIR family. Each cell is the number of instances where one copy of a TE was classified into the subfamily on the column and the other into the subfamily on the row. There are only numbers in the top triangle because we do not know which subfamily is correct. A MIR3 aligned to a MIRb is counted the same as a MIRb aligned to a MIR3.

	MIR	MIR1_Amn	MIR3	MIRb	MIRc
MIR	1371	4	16	231	39
MIR1_Amn		77	20	10	10
MIR3			651	21	37
MIRb				1818	149
MIRc					708

Table 3: Instances of TE copies in segmental duplications for six of the 47 Alu subfamilies. Each cell is the number of instances where one copy of a TE was classified into the subfamily on the column and the other into the subfamily on the row. are only numbers in the top triangle because we do not know which subfamily is correct. An AluJb aligned to a AluJr is counted the same as an AluJr aligned to an AluJb.

	AluSx	AluJr	AluSx1	AluSz	AluJb	AluY
AluSx	1612	4	98	102	14	23
AluJr		1485	4	6	199	1
AluSx1			1575	99	4	17
AluSz				1419	33	11
AluJb					1325	5
AluY						1145

Discussion

When a segment of genome with a TE remnant is duplicated, the TE remnants in each copy come from the same parent and should be annotated and classified as the same subfamily. Many TE remnants in segmental duplications are not annotated to the same subfamily (Table 1), suggesting that these remnants are misclassified.

We identified many subfamilies of TEs that are often misclassified as another subfamily (Tables 2 and 3). This suggests that these subfamilies do not represent different bursts of replication. The current method may split TEs that were generated in one burst of replication into multiple subfamilies because the sequences differ from random mutations.

Not all instances where two TEs were classified into different subfamilies are errors. Some cellular processes can cause replicate copies to appear to be misclassified. The most common process is gene conversion. Gene conversion occurs when a strand of DNA breaks and repair machinery recruits similar sequence from another location to repair the break (Chen et al., 2007). If the break occurs in a TE, the repair machinery may recruit a similar sequence from a TE in a different subfamily. The broken TE is then converted to a different subfamily. When a TE in a segmental duplication undergoes gene conversion with a TE outside of the duplication, the TEs may appear to be misclassified even though they are not.

To eliminate these instances, we are working on a program to identify all gene conversion in segmental duplication. Previous research has found that gene conversion occurs at low rates in segmental duplications (Sawyer, 2000), (Dumont & Eichler, 2013), so gene conversion is probably not responsible for most of the misclassifications.

Biologists use transposable elements to study human evolution and genetic diseases (Solyom & Kazazian Jr, 2012), so it is important that TEs are classified and annotated correctly. We find that the current method of classifying TEs into subfamilies splits too much. TE subfamilies are not reproducible and do not represent all TEs generated in one burst of replication. The current method of classifying TEs should be modified so that TE sequences are split into fewer subfamilies.

Methods

We analyzed two independent databases, (1) segmental duplications from Bailey et al. (2002) (http://humanparalogy.gs.washington.edu/) and (2) transposable elements from RepeatMasker (http://repeatmasker.org/), both annotating human genome GRCh37. The database of segmental duplications consists of 111,736 pairwise alignments. All duplications are greater than 1000 base pairs long and at least 90% identical, so the duplications probably occurred in the last 40 million years (Bailey et al., 2002). In the enire genome, RepeatMasker identified 5,467,457 TE sequences classified into 1,183 different subfamilies.

We developed software in Perl to find all instances where a TE was copied as part of a segmental duplication. Some segments of DNA have been duplicated more than once and so are included in multiple alignments. The software only analyzed segments that have duplicated once. It first found the location of every TE in each sequence of each alignment, then filtered these TEs to include only simple cases where it was clear that a TE in one sequence was a copy of a TE in the other sequence.

The software read through each pairwise alignment and used the output from RepeatMasker to identify the TEs in each sequence. For each TE, it recorded the location within the alignment and the subfamily. The program eliminated all TEs that had less than 50 base pairs in the alignment.

When one TE in a sequence of the alignment is a copy of a TE in the other sequence, these TEs will overlap. The software used the locations within alignments to identify all instances where a TE in one sequence overlapped at least 80% of a TE in the other sequence. It excluded complicated cases where a TE on one sequence overlapped multiple TEs in the other sequence.

Sometimes a TE inserted itself after the segment had been duplicated. To exclude these, for each pair of overlapping TEs the software counted the number of gaps and the percent identity in the aligned region of the other sequence. It eliminated all instances where the aligned region was more than 80% gaps or the TEs were less than 80% identical. After filtering, the program identified 52,777 instances of duplicated TEs.

The software compared the subfamilies of each duplicated copy. For every possible unordered pair of subfamilies, it recorded the number of times one TE copy was annotated in one of the subfamilies and the overlapping TE was annotated in the other.

For code, contact the author (gilia.patterson@umontana.edu).

References

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. (2002). Recent segmental duplications in the human genome. *Science*, *297*(5583), 1003–1007.

Batzer, M. A. & Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nature Reviews Genetics*, *3*(5), 370–379.

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., & Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics*, *8*(10), 762–775.

Cordaux, R. & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, *10*(10), 691–703.

Dumont, B. L. & Eichler, E. E. (2013). Signals of historical interlocus gene conversion in human segmental duplications. *PloS one*, *8*(10), e75949.

Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? *TRENDS in Genetics*, *23*(4), 183–191.

Murnane, J. P. & Morales, J. F. (1995). Use of a mammalian interspersed repetitive (mir) element in the coding and processing sequences of mammalian genes. *Nucleic acids research*, *23*(15), 2837–2839.

Sawyer, S. (2000). Geneconv: Statistical tests for detecting gene conversion (version 1.81). *Department of Mathematics, Washington University, St. Louis, Mo.*

Smit, A., Hubley, R., & Green, P. (2013). 2015. *RepeatMasker Open-4.0. Available from http://www. repeatmasker. org (accessed on 11 February 2014).*

Smit, A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current opinion in genetics & development*, *9*(6), 657–663.

Solyom, S. & Kazazian Jr, H. H. (2012). Mobile elements in the human genome: implications for disease. *Genome Med*, *4*(2), 12–12.

Wacholder, A. C., Cox, C., Meyer, T. J., Ruggiero, R. P., Vemulapalli, V., Damert, A., Carbone, L., & Pollock, D. D. (2014). Inference of transposable element ancestry. *PLoS Genet*, *10*(8), e1004482.

Yoder, J. A., Walsh, C. P., & Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends in genetics*, *13*(8), 335–340.