Identifying Syntenic Repeats Using Whole Genome Alignment

Julia Holz, Robert Hubley, Arian Smit
Institute for Systems Biology, Seattle Washington

ISB Science Transforming Health

Introduction

Transposable elements (TEs) are DNA sequences that are able to replicate themselves and move to new locations in the genome. They make up a large portion of many genomes, with the human genome consisting of over 47% transposable elements. Dfam is a database of transposable elements, which includes over 25,000 curated families, and over 3,000,000 total families [1]. One challenge in curating this TE database is determining if similar TE families in two related species should be considered as one family assigned to an ancestral taxon. This is a difficult problem, requiring judgment calls about the type and frequency of differences between the two families relative to what we expect from random mutations. However, the extent to which we find instances of these repeat families in syntenic locations in the respective genomes can inform our decision about whether to merge them.

Abstract

Our goal is to develop a pipeline that allows us to take the genomes of two species, and for any given repeat family found in the target genome at a set of locations, create a presence/absence table which indicates for each location whether the repeat is found at a corresponding location in the query genome. This should provide evidence as to whether the insertion events for this TE occurred in a common ancestor or after the two species diverged, thereby informing us about the evolutionary history of the TE family. We also hope to determine the level of evolutionary divergence between species our pipeline can handle in order to identify what species would be appropriate targets for this kind of analysis.

Preliminary Results

At this stage, we have begun evaluating AnchorWave for aligning both elephant and chimpanzee genomes to the human genome, and have observed large synteny blocks in the resulting alignment:

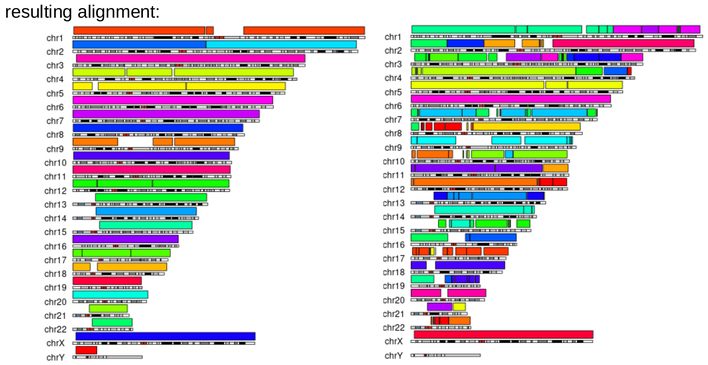


Figure 1: Aligned regions of chimpanzee (left) and elephant (right) genomes to the human genome, with human chromosomes below and aligned section lengths of query species above, colored by chromosomal origin in the query.

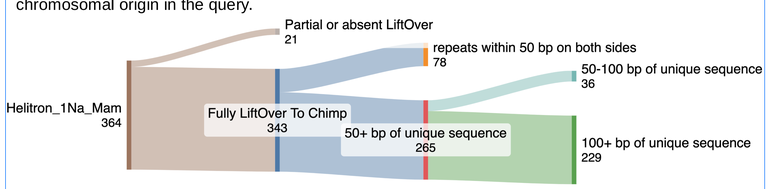
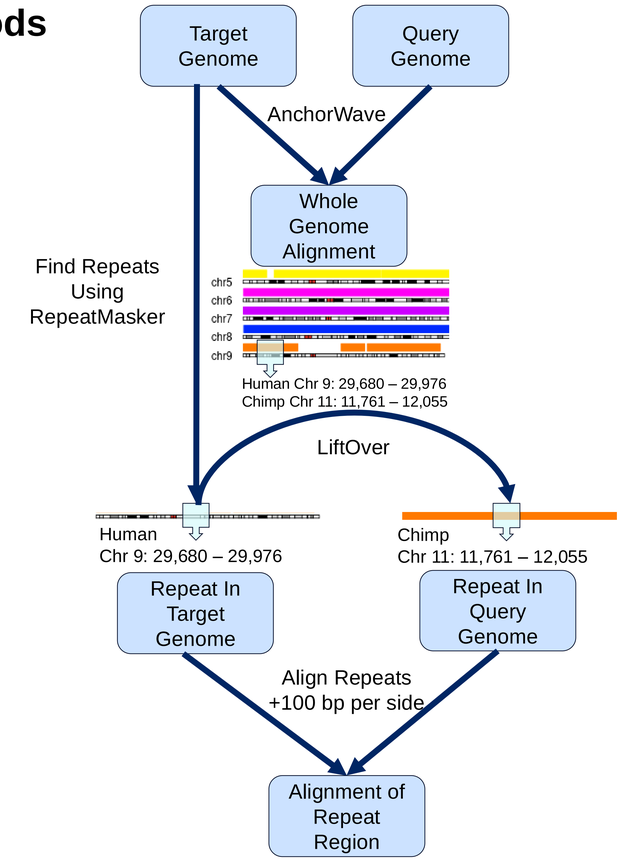


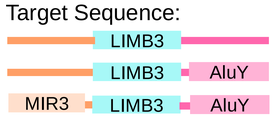
Figure 2: Number of occurrences of a Helitron repeat family in human that make it through the filtering steps in our pipeline. Unique sequence filters require non-repeat sequence on only one side. In general, for this family and others we have studied, a majority of TE instances are lifting over to the chimpanzee. However, a good portion of them do not have 50 base pairs of unique sequence on either side, which complicates synteny determination.

Methods



Possible situations on each side of the alignment

Target Sequence:



unique sequence on both sides → ideal case

unique sequence on one side → use unique side

repeats on both sides → apply stricter thresholds or discard

Query Sequence:

Syntenic Repeat Found?

repeat and adjacent sequences both align

repeat absent in query

repeat aligns well but adjacent sequences differ

Align all instances of family and use alignment thresholds to determine if each instance has a syntenic copy to create a presence/absence table

Position in Target:	Chr 1: 50921794-50921871	Chr 3: 53404519-53404914	Chr 23: 114561681-114562301	…
L1MB3 – Syntenic Copy Found?	✓	✗	✓	…

Assessment

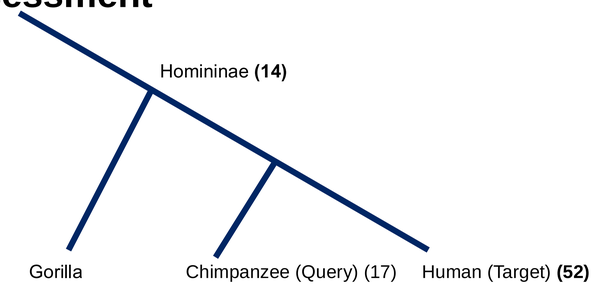
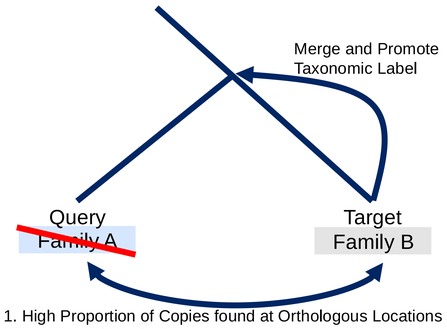


Figure 3: Number of Families for each group in our proposed verification dataset.

We plan to assess the success of our method using a dataset of curated human and chimpanzee TE libraries from Repbase [3], and Dfam[1], which contains repeat families that are known to be shared between human and chimpanzee, as well as human-exclusive and chimpanzee-exclusive families. If our method performs well, we expect to see many syntenic copies of the shared families, and no syntenic copies of the human-exclusive families when we use human as a target and chimpanzee as a query.

Goals and Significance

The goal of this pipeline is to accurately identify syntenic repeats, and therefore assess the ancestral state of the family. In some cases this would support the collapse of redundant families in Dfam, reducing the storage overhead in the database. Additionally, improving the taxonomic labels assigned to each family has a large impact on genome annotation tools such as RepeatMasker [2], which only search sequences against repeat families from a certain set of taxonomic groups based on the sequence's species of origin.



1. High Proportion of Copies found at Orthologous Locations
2. High Model Similarity

Figure 4: Possible application of our method to move families up the taxonomic tree. Our pipeline would essentially act as an additional check on a model-similarity based merging method such as SCULU [4], allowing us to merge families not only based on model similarity, but also, based on their occurrence in syntenic locations in multiple genomes, increasing our confidence in the appropriateness of the merge.

Citations

1. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J., & Smit, A. F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, *12*(1), 2.
2. Smit, A., Hubley, R., & Green, P. (2015). RepeatMasker Open-4.0. 2013-2015
3. Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1), 11.
4. Shingleton, A. (2022). Subfamily clustering using label uncertainty(For transposable element families). *Graduate Student Theses, Dissertations, & Professional Papers*.