Inferring Transposable Element Haplotypes Markers from Population Genomics Data using Hierarchical Clustering

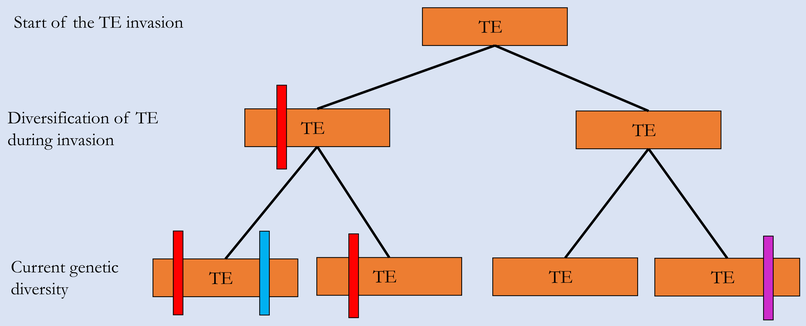
Iskander Said*, Michael McGurk*, Clayton Hughes, Andrew Clark, Daniel Barbash

Department of Molecular Biology and Genetics, Cornell University, Ithaca NY

*these authors contributed equally to this work

Contact me:
@ is_a_biologist
iskander.said@gmail.com

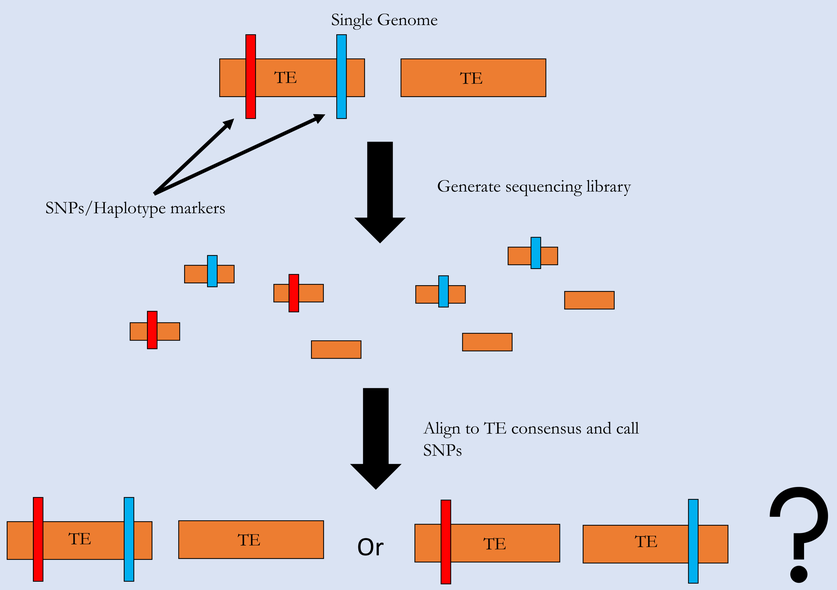
TEs are genetically diverse genomic parasites



Potential causes of genetic diversity:

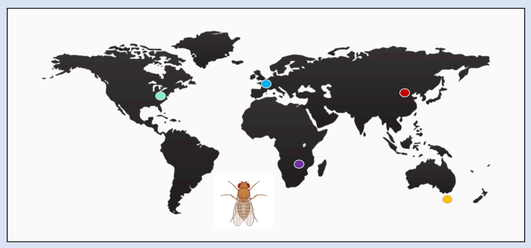
- **Neutral processes**
- **Competition**
  - TEs have limited genomic space and must compete for this resource.
- **Repression**
  - The host genome immune system (piRNAs) act to silence TEs, and they must escape repression.

Short-read data loses linkage phase of SNPs on TE sequences



- What are the correct TE haplotypes?
- How can we retain the SNP/haplotype marker linkage?

Analyzed 41 active *D. melanogaster* TEs from the Global Diversity Lines



Clustering cutoff parameters chosen from empirical null distribution

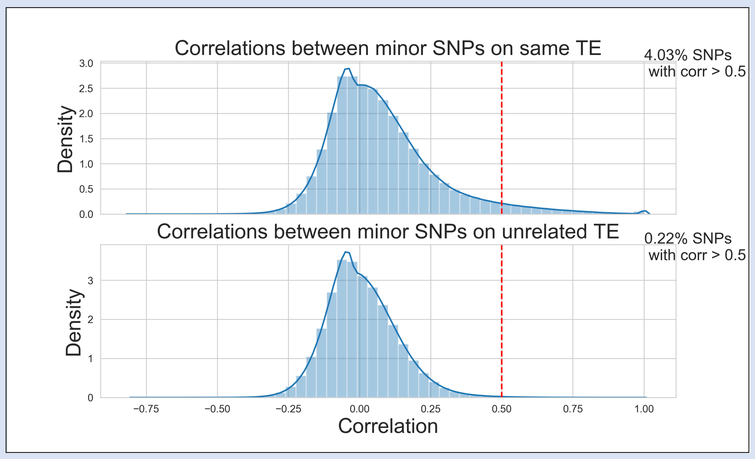
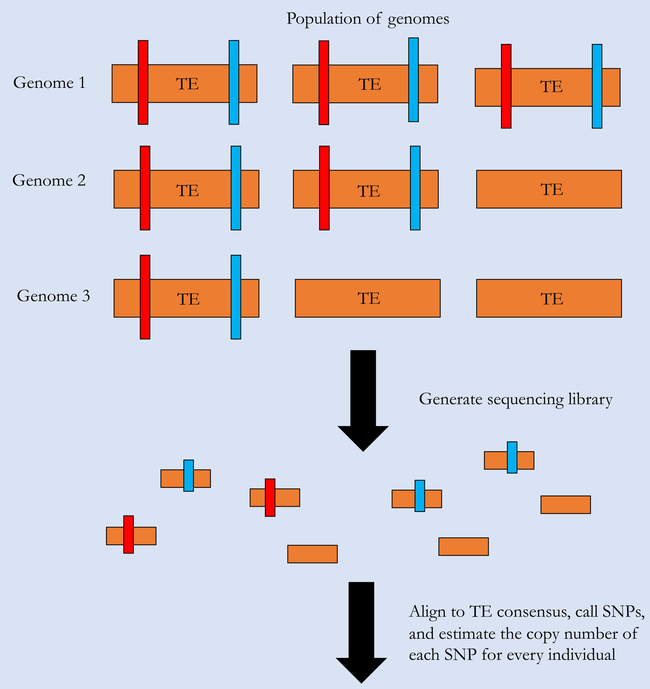
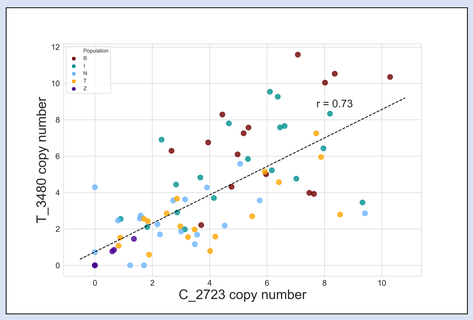


Fig 3. Correlations between SNPs within the same TE (test), and correlations of SNPs between two different TEs (empirical null). We reason that the rate of spurious correlations for a given correlation cutoff could be calculated by computing pairwise correlations of SNPs between unrelated TEs. We find that 0.22% of the correlations from the null were greater than 0.5, while in the test distribution this number was 4.03%. This provides a proxy for a false positive rate.

SNP correlations recover linkage phase





Are copy number of SNPs correlated?

Fig 1. Correlation in copy number between two SNPs from the Jockey element. These two SNPs show strong correlation in copy number (r = 0.73). We therefore infer that these SNPs segregate on the same element.

Hierarchical Clustering creates clusters of highly correlated SNPs

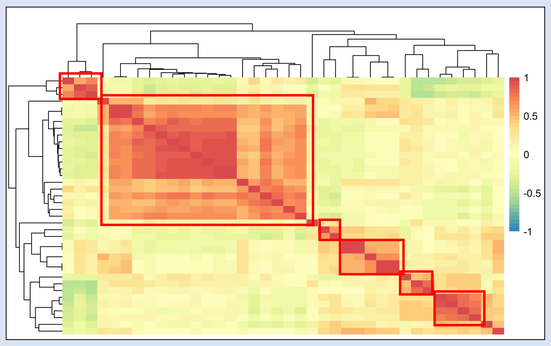
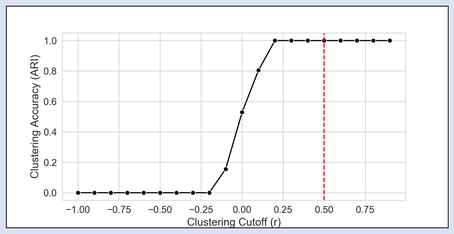


Fig 2. Seriated heatmap of correlations in copy number between all SNPs from the Jockey element. Each red box denotes a cluster of correlated SNPs (haplotype markers) called via Hierarchical Clustering (Average Linkage, r = 0.5).

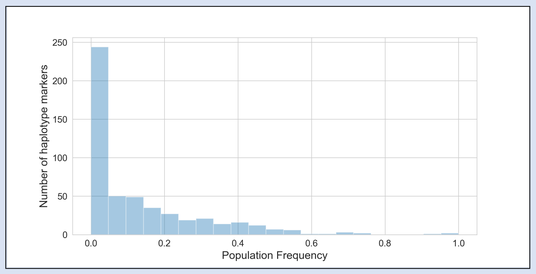
Simulations of TE haplotypes benchmark clustering performance



- Simulated short-read data from simulated TE haplotypes in a population.
- More in depth simulations will come soon!

Fig 4. Adjusted Rand Index (ARI) versus clustering cutoff parameter used for the haplotype marker inference of five unique, and unrelated simulated TE haplotype. ARI is stable at wide range of clustering cutoff parameter values.

Used PacBio genomes to detect presence of inferred haplotypes



- Aligned TE consensus to 19 PacBio genomes (5 GDL, 14 DSPR).
- Queried each TE alignment for haplotype markers discovered from GDL short-reads.
- ~66% of haplotype markers detected in PacBio genomes.

Fig 5. Frequency of haplotype markers in DSPR, and GDL PacBio genomes. I queried the PacBio genomes for the haplotype markers discovered in the GDL. The population frequency of a haplotype marker is the frequency of that haplotype marker in the entire PacBio genome dataset.

Population structure of TE variants is common

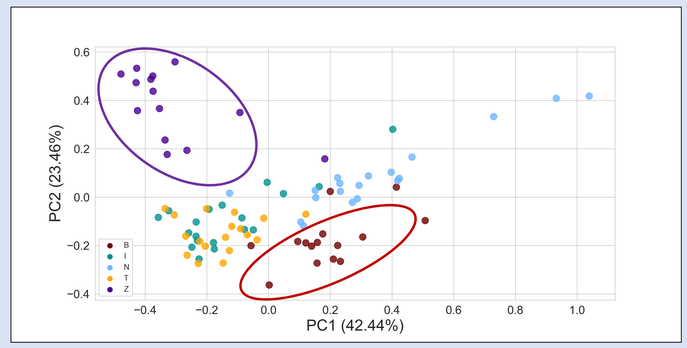


Fig 6. PCA on the allele frequencies of every SNP for Jockey elements for each GDL line. Red and purple circles denote clustering of Beijing and Zimbabwe variants, respectively.

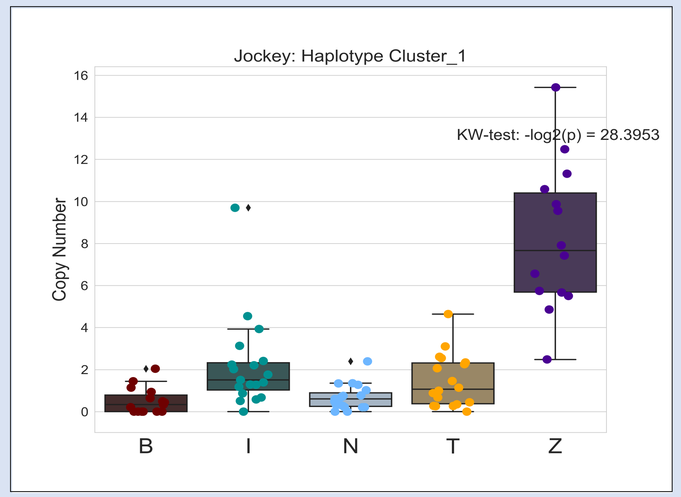


Fig 7. Copy number of a single haplotype marker cluster drives population structure of Jockey variants in Zimbabwe strains.

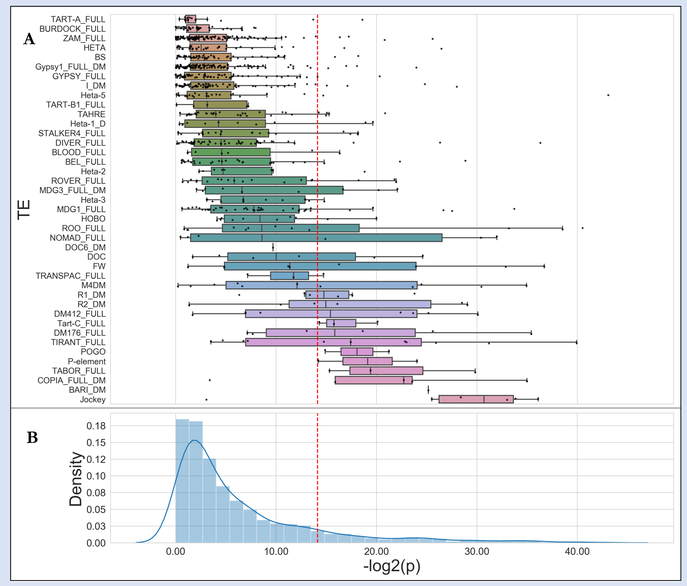


Fig 8. Kruskal-Wallis test on haplotype marker copy number between populations. (A) Boxplots of -log2(p) scores from K-W tests for each haplotype marker for each active TE. (B) Density plot of all –log2(p) scores from K-W tests across all TEs. Red dashed line is Bonferroni corrected critical value. 13% of haplotype markers passed corrected critical value and show population structure.

Preliminary Results: Can SNPs help TEs evade piRNAs?

Fig 9. Haplotype markers over laid with percent of total piRNA reads at each position for Hobo element. 29 ovarian piRNA libraries were aligned to TE consensus sequences (From GDL, DGRP, Misy and Paris). The number of reads at each position was normalized by calculating percent of reads that aligned to TE for sense, and antisense strands. Shown is the median of those values at each position from the 29 libraries. Haplotype markers were tested for enrichment of antisense piRNA read depth via permutation testing. Haplotype marker cluster 3 showed enrichment, while other haplotype markers did not.

Thank you Barbash lab, and Clark lab for intellectual contributions, and productive coffee breaks!