Understanding Genetic Variation in the Repetitive Sequence Content of *Drosophila melanogaster*

Chinmay P. Rele & Christopher E. Ellison Ph.D*
Department of Genetics, Rutgers University, New Brunswick; NJ 08854

* Corresponding author

RUTGERS
Aresty Research Center for Undergraduates

Background

Between 40% and 70% of eukaryotic genomes are composed of repetitive sequence, which influences genome organisation, gene expression, and genome evolution. Repetitive sequences can be in the form of repetitive genomic structures such as terminal repeats, tandem repeats or interspersed repeats. The latter differ from tandem repeats in that instead of coming directly after one another, they are non-adjacently dispersed throughout the genome. Most non-adjacent repeat sequences are in the form of transposons (TEs), as their primary function is to replicate and move around the genome in a pseudo-random fashion.

We chose to use the Drosophila Genome Reference Panel (DGRP), a collection of strains of *Drosophila melanogaster* caught in the wild and highly inbred in our analysis as hey have priorly been sequenced.

Illumina sequencing is the most common and cheapest form of sequencing available. However, despite it having high coverage values, making it ideal for SNP analysis, it suffers from having short read lengths, and thus, reads of repeat elements cannot be properly aligned to a contig and are usually ignored from genomic analyses. Understanding how these sequences vary at the population level is important for understanding their evolutionary dynamics.

Motivation

We needed to identify high repeat regions and identify how exactly they interact with the rest of the genome. The main motivation is to estimate the evolutionary biology of these repeat regions. This can be done by comparing the repetitive sequence content between individuals and seeing which repeats are correlated with the other.

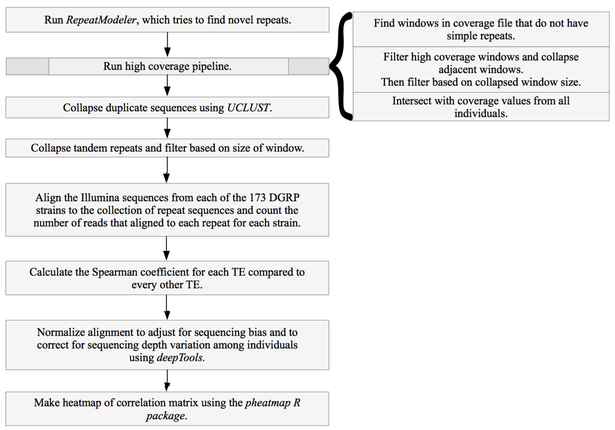
We still don't know how the repetitive portions of the genome vary among individuals as we cannot accurately align the short Illumina reads to the repetitive parts of a genome assembly.

This is because (1) each repeat is present at many loci in the genome, and (2) other individuals have additional repeat indels and expansion of tandem repeats.

We hypothesise that TEs and other repeats that use similar transposition/amplification mechanisms will have copy numbers that are correlated among individuals.

Methods

In order to compare repeat copy number among individuals, we needed to get the consensus sequence for every possible repetitive element in the *Drosophila* genome. We started analysis with sequences of known TEs.



Results

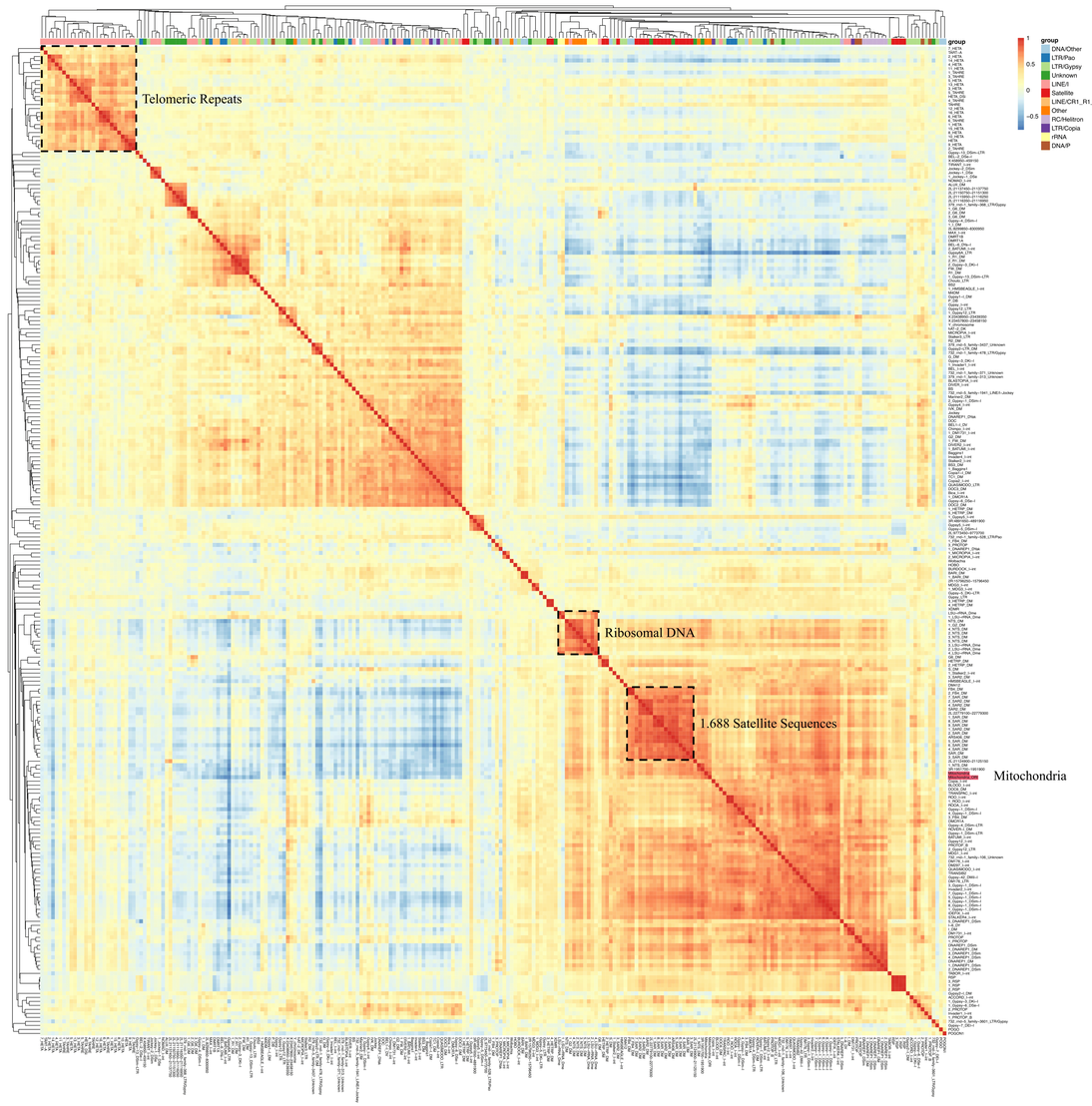


Fig. 1.: Spearman Correlation matrix showing correlation of repeat element copy number across all DGRP individuals

Though there was high correlation between telomeric repeats and satellite sequences, at the broadest level, our assumption that repeat elements of similar types would be similarly correlated did not hold true (based on topmost row of heatmap that show group allocations of the repeat elements.

Focusing on mitochondria, we found that the copy number of a lot of genomic repeats correlated with the copy number of mtDNA. Most of these pairs were highly positively correlated with each other, while a few had negative correlation.

Results

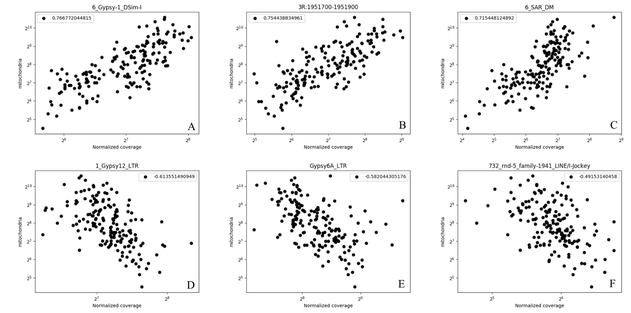


Fig. 2.: Coverage of mtDNA with the repeat (A) 6_Gypsy-1_DSim-I, (B) Unidentified repeat at 3R:1951700-1951900, (C) 6_SAR_DM, (D) 1_Gypsy12_LTR, (E) Gypsy6A_LTR, and (F) I-Jockey; with corresponding Spearman correlation coefficient (ρ). These repeats were chosen due to high positive or negative correlation.

Future Directions

We currently have the genome sequence of RAL-379 which has a high copy number of mitochondrial DNA (mtDNA). We need to sequence RAL-83, which has a low copy number of mtDNA and see what explains the pattern of correlation between genomic DNA and mtDNA.

We are currently creating *de novo* genome assemblies from the longer reads generated by the Nanopore MinION. This approach will allow us to identify the genomic location and copy number of each repeat element, which we will compare to the copy number inferred by aligning Illumina sequences to the consensus sequence of each repeat. With this study, we hope to learn about the evolutionary dynamics of repeat sequences and the forces that control their copy number in the genome.

Acknowledgements

We would like to thank the Aresty Program for funding this project, Amarel at OARC for computational resources, and Weihuan Cao for her continued support in our work.

References

deepTools: Fidel Ramírez, Friederike Dündar, Sarah Diehl, Björn A. Grüning, and Thomas Manke. deepTools: a flexible platform for exploring deep-sequencing data. Nucl. Acids Res. first published online May 5, 2014 doi:10.1093/nar/gku365

DGRP: Mackay, T. F. C., et. al. (2012). "The Drosophila melanogaster Genetic Reference Panel." Nature **482**(7384): 173-178.

USEARCH: Edgar,RC (2010) Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26(19), 2460-2461. doi: 10.1093/bioinformatics/btq461

RepeatModeler: http://www.repeatmasker.org/RepeatModeler/

Pheatmap: https://CRAN.R-project.org/package=pheatmap

eXpress: https://pachterlab.github.io/eXpress/overview.html

RUTGERS