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

Rele, Chinmay P.; Ellison, Chrisptopher E.

December 8, 2018

Abstract

Between 40% and 70% of eukaryotic genomes are composed of repetitive sequence, which influences genome organisation, gene expression, and genome evolution. 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. We needed to identify high repeat regions and identify how exactly they interact with the rest of the genome. 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. 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.

For documents and more data, please contact *chinmay.rele@rutgers.edu*

Contents

1 Introduction

1.1 Repeat Elements

Repeat sequences are abundant in every genome **de2011repetitive**. They can be beneficial from those such as the poly-A tail at the end of genes to prevent mRNA degradation **shapiro2005repetitive**, to micro-satellites, which have been characterised with protein- encoding genes **vieira2016microsatellite**.

Between 40% and 70% of eukaryotic genomes are composed of repetitive sequence, which influences genome organisation, gene expression, and genome evolution **biscotti2015repetitive**. There are three main type of repetitive sequences: (1) terminal repeats, (2) tandem repeats, and (3) interspersed repeats. Terminal repeats exist in the form of telomeric repeat regions, which prevent degrading the chromosome **o2010telomeres**. Tandem repeats are those that have been extended during replication, which can be the cause of many genetic diseases such as ASD. They may also fracture certain chromosomes due to hyper-long repeat expansion **sutherland1995simple**.

1.2 Transposons

By far, the most interesting form of repeat elements are transposable. Transposable Elements (TEs) or transposons were first identified by Barbara McClintock in the mid 1950s in maize **mcclintock1956controlling**. They move in two fashions by a Copy/Paste mechanism or by a Cut/Paste mechanism. In the former, Class I transposons, or retrotransposons, move via an RNA intermediate. They are first transcribed into an RNA, which is then reverse transcribed into DNA, which is then inserted into the genome at a new position **corces1991interactions**. Class II elements, DNA transposons have a Cut/Paste locomotive mechanism, i.e., they do not require an RNA intermediate. They are excised via a transposase, which binds to the local inverted repeats, excises the copy, and relocates it to another genomic locus **munoz2010dna**.

It is obvious by the mechanisms that Class I transposons double in copy number every time they move, whereas those Class II TEs transpose via a non-replicative mechanism, they must increase copy number based on indirect mechanisms that rely on host machinery **feschotte2007dna**. TEs

are usually much longer than simple repeat elements, and thus, need to be identified in a different way.

1.3 Sequencing Practices

The most common form of sequencing is Illumina sequencing. It consists of taking many short reads, and aligning them to a reference genome in order to make small contig sequences with a high amount of overlap or coverage (Quail et al. 2012). There are two wants of forming the genome of the organism from here: (1) by aligning each contig to a reference sequence (allowing for some error rate due to SNPs or other small mutations); or (2) by aligning the contigs to themselves to form the genomic sequence. However, both of these approaches cannot properly align repetitive sequences that are larger than the read length of Illumina (about 300bp) **nakamura2011sequence**. These repetitive sequences get clustered to a single locus and knowing their true location on the 2D chromosome is almost impossible solely using Illumina sequencing.

Oxford Nanopores MinION^{TM}, seems to overcome this problem. It has much longer read lengths (maxing out at about 15kb)**jain2016oxford**, but also has problems of its own. It tends to have a much higher error rate when calling bases. So, this is a method better suited for calling where the repeat elements are, and not the particular sequence of repeat elements (which can be sequenced properly using Illumina).

1.4 Variation Analysis

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 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.

2 Results and Discussion

2.1 Heatmap

We found a series of repeat elements that were highly correlated with each other. These included the telomeric repeats, ribosomal DNA, satellite sequences (with very high correlation of Responder elements), and 1.688 Satellite sequences.

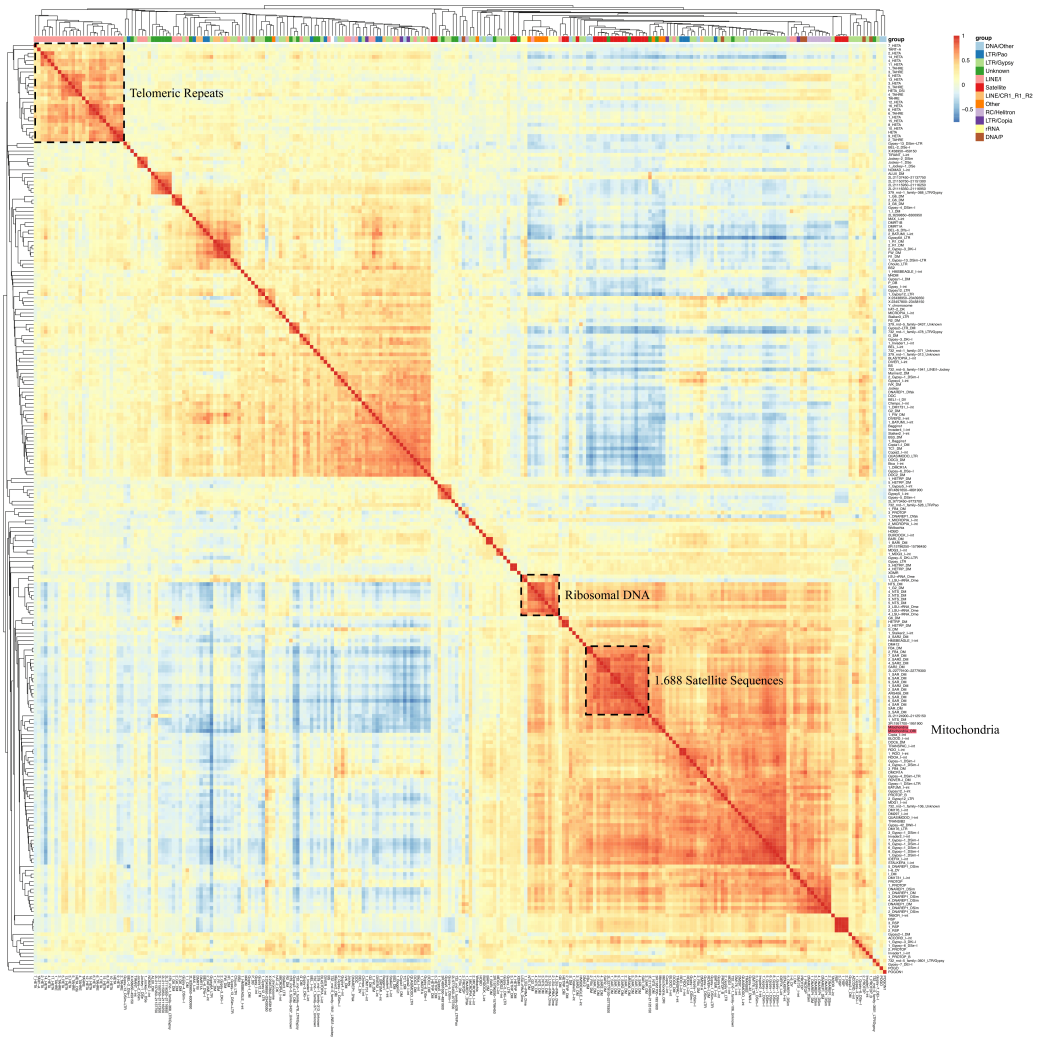

Figure 1: Heatmap of Spearman correlation matrix of repeat elements copy number vs. repeat element copy number averaged across all DGRP individuals.

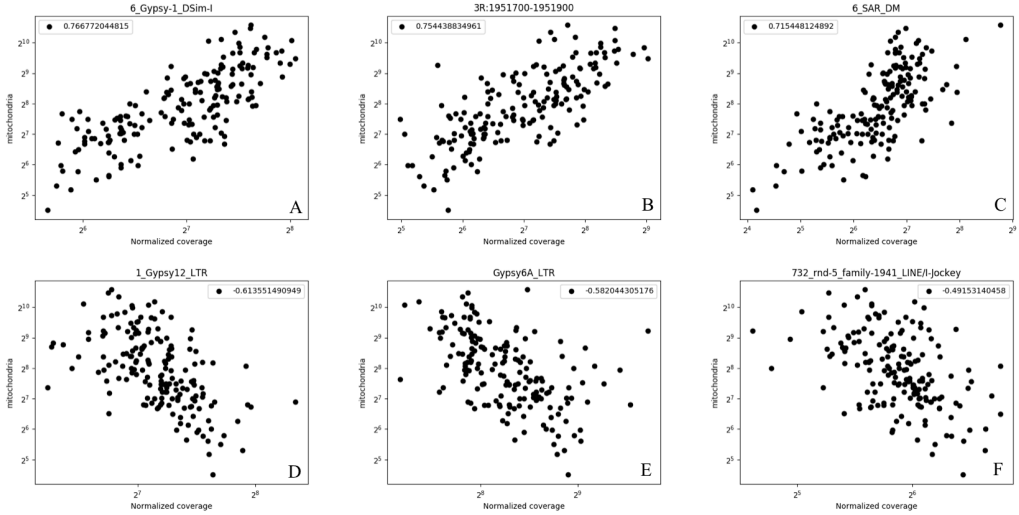

Figure 2: Coverage of mtDNA with the repeat (A) 6_Gypsy-1_DSim-I, (B) Unidentified repeat at 3R:1951700-1951900, (C) 6_SAR_DM, (D) 1_Gypsy12_LTR, (E) Gypsy6A_LTR, and (F) I-Jockey; with corresponding Spearman correlation coefficient (ρ). These repeats are emphasised to show high positive (A, B, C) or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

We assumed that there would be high correlation between repeat elements of the same class, and either low or regular correlation of repeat elements across classes. This did not hold true as there seems to be relatively high correlation between Satellite, Helitron and Gypsy repeat elements.

Focusing on mitochondrial repeats, we found that the copy number of a lot of genomic repeats correlated with the copy number of mtDNA; this should not be the case as mitochondria have their own genome as compared to the host. Most of these pairs were highly positively correlated [Table 1], while others showed more negative correlation [Table 2].

Table 1: Spearman correlation of mitochondrial copy number vs. copy number of following repeats across all DGRP strains having high positive correlation.

Correlation	**Name of Repeat Element**
0.791 661 9	Mitochondria_ORI
0.766 772 0	6_Gypsy-1_DSim-I
0.754 438 8	3R:1951700-1951900
0.753 285 3	5_Gypsy-1_DSim-I
0.730 613 6	8_Gypsy-1_DSim-I
0.715 533 1	7_Gypsy-1_DSim-I
0.715 448 1	6_SAR_DM
0.712 966 6	DM297_I-int
0.712 257 4	4_SAR_DM
0.704 863 9	3_SAR_DM

Table 2: Spearman correlation of mitochondrial copy number vs. copy number of following repeats across all DGRP strains having high negative correlation.

Correlation	**Name of Repeat Element**
−0.613 551 5	1_Gypsy12_LTR
−0.582 044 3	Gypsy6A_LTR
−0.491 531 4	732_rnd-5_family-1941_LINE/I-Jockey
−0.482 626 5	DIVER2_I-int
−0.477 631 9	QUASIMODO_LTR
−0.467 498 9	DMRT1A
−0.448 472 9	1_BATUMI_I-int
−0.447 137 9	2_BATUMI_I-int
−0.446 173 7	DOC3_DM
−0.433 512 1	BS3_DM

3 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.

4 Methods and Motivation

4.1 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 repeats.

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.

4.2 Protocol

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.

We ran *RepeatModeler*, which tries to find novel repeats, on the sequence data of the DGRP strains. We than ran a home-brew pipeline to isolate repeats with high coverage [A01-A07]. In this pipeline: we found windows of 50bp that do not have simple repeats (such as poly-A elements, and other tandem repeat elements with low complexity), filtered them and collapsed adjacent windows. We filtered them again based on collapsed window size. After isolating these high coverage windows, we intersected the coverage values of these windows across all individuals.

We then collapsed duplicate sequences using *UCLUST*, collapsed tandem repeats and filtered based on window size. We chose the minimum window

size to be 200bp. We then aligned them to Illumina sequences from each of the 173 DGRP strains and counted the coverage (the number of Illumina reads that aligned with the repeats deciphered above).

We calculated the Spearman coefficient of the coverage of each of these repeat elements to each other; after which we normalised these alignments to adjust for sequencing bias and correct for sequencing depth variation among individuals. deepTools was used to do this. We then made a heatmap of the data [Fig. 1].

5 Acknowledgements

5.1 Lab Members

We would also like to specially thank Weihuan Cao for constant input with improving procedures, preparing embryos, extracting nuclei and generally rearing the fly strains that have been used for the study.

5.2 Amarel

The authors acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here. URL: http://oarc.rutgers.edu.

6 Appendix

6.1 01_STRAIN_COV.SH

Run using: `sbatch 01_strain_cov.sh`

```
#!/bin/bash
#SBATCH --partition=genetics_1
#SBATCH --requeue
#SBATCH --job-name=01strain_cov
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=1
#SBATCH --mem=20G
#SBATCH --time=00:03:00
#SBATCH --output=slurm.%j.out
#SBATCH --error=slurm.%j.err
#SBATCH --export=ALL

cd /scratch/cpr74/2018Spring/coverage_per_strain/
    new_strains_2018_02_DGRP732_assembly
# Make small windows of 50 bp each.
bedtools makewindows -g DGRP732.final_pilon.genome -w 50 >
    ./ windows_50bp_732_pilon.genome
sleep 3
# Keep windows that do not overlap the simple repeats.
bedtools intersect -v -a windows_50bp_732_pilon.genome -b
    DGRP732.final_pilon.simple_repeats.bed > ./
    windows_without_low_complexity_repeats.bed
```

References