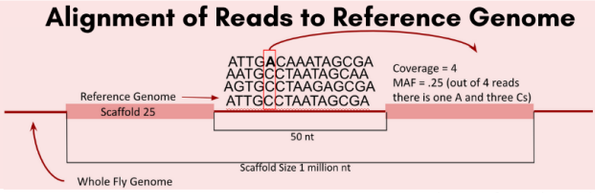
Identifying Sites of Sequence Variation Among Populations of *Drosophila sechellia* Using Bioinformatics Pipeline

Angela Sofia Colorado, Jiahui Candice Shi, and Joseph Coolon
Department of Biology, Wesleyan University, Middletown, CT 06459
Email: acolorado@wesleyan.edu jshi@wesleyan.edu
Website: http://coolonlab.research.wesleyan.edu

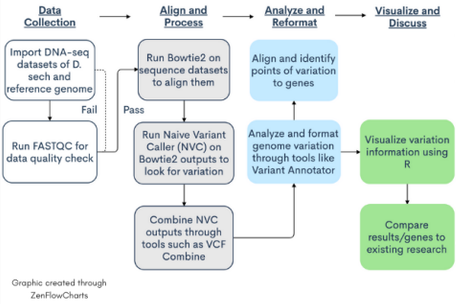
Introduction to *D. sechellia*

Drosophila sechellia is a species of Fruit Fly only found on the Seychelles Islands off the eastern coast of Africa. Upon colonization of this island chain, *D. sechellia* evolved to be a host specialist on the toxic fruit of *Morinda citrifolia*. A key adaptation central to *D. sechellia's* host specialization is evolved resistance to octanoic acid (OA) the primary toxic compound found in *M. citrifolia* . Previous research has identified genes including as *Est6* and 6 different *Osiris* family genes that may contribute to *D. sechellia's* ability to survive in high octanoic acid conditions (Lanno et al., 2017). Our work strives to identify the gene(s) important to the development of key adaptations in *D. sechellia* in a genome-wide and non-trait specific manner. To do this, we compared DNA sequence variation between 14 samples of *D. sechellia* genomes to determine areas of high and low variation. These regions may provide candidate genes responsible for the specialized traits of *D. sechellia* and other insights for the genetic basis of *D. sechellia* evolution.

Alignment of Reads to Reference Genome



Design of Bioinformatics Pipeline

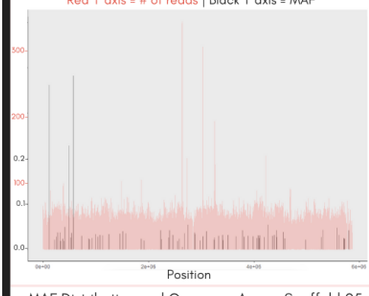


- Analysis performed in Galaxy, an open source data analysis platform
- Initial *D. sechellia* data is from Schrider et al., 2018
- Bowtie2 (Langmead et al., 2012) aligned DNA dataset to *D. sechellia* genome
- NVC tool (Blankenberg et al., 2014) identified points of variation in datasets by comparing sequences to reference genome
- Variant Annotator (Dickens et al., 2014) provided minor allele frequencies (MAF) and coverage for each variant

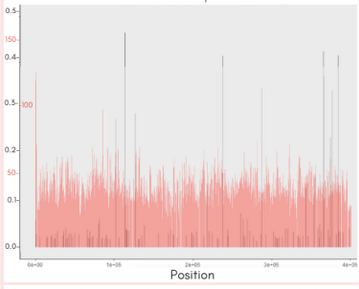
Sites of Sequence Variation

- The MAF and coverage of each scaffold, fragment of genome, is graphed. The Osiris gene family is located within scaffold 5, 25, 27, and 38.
- Minor allele frequency (MAF) calculates the frequency of the second most common allele at a position within the total number of reads/population.
- Coverage means the number of reads the program picked up at a given position. Coverage can impact MAF values, so having the same coverage within each scaffold is ideal.

MAF Distribution and Coverage Across Scaffold 5
Red Y axis = # of reads | Black Y axis = MAF



MAF Distribution and Coverage Across Scaffold 38
Red Y axis = # of reads | Black Y axis = MAF



MAF Distribution and Coverage Across Scaffold 25
Red Y axis = # of reads | Black Y axis = MAF

MAF Distribution and Coverage Across Scaffold 27
Red Y axis = # of reads | Black Y axis = MAF

Conclusions

- Quality control analyses indicate the data are high quality and similar coverage levels across genome demonstrate we have sufficient data to estimate sequence variation genome-wide.
- Some regions have higher or lower sequence variation suggesting these regions have differences in mutation rates and/or differences in selective pressures.
- Extreme deviations in coverage for some regions of the genome suggest duplication and deletion are important aspects of sequence variation that may be evolutionarily important.

Future Directions

- Filter out anomalies and map specific regions to genes to create a candidate pool of genes to investigate, with specific focus to:
  - Genes that have been shown to play important roles in *D. sechallia's* speciation (i.e. *Est6* and *Osiris6*)
  - Positions that show high amounts of duplication or deletion
- Determine the mechanisms for duplication at certain locations (i.e. tandem duplication) and what genes are impacted.

Acknowledgements

This work was funded by the Biology Department at Wesleyan University, the McNair Program, Grants and Support and Scholarship (GSOS). Thank you to the members of the Coolon lab for their expertise and a special thanks to Professor Joe Coolon for his continued guidance and teaching during this summer and beyond.Special thanks to Ronnie Hendrix and Erika Taylor from the McNair program, whose support was essential for this project. Lastly, thank you to all the staff who made the Research in the Sciences program possible.

References

Blankenberg, Daniel and Von Kuster, Gregory and Bouvier, Emil and Baker, Dannon and Afgan, Enis and Stoler, Nicholas and Taylor, James and Nekrutenko, Anton (2014). Dissemination of scientific software with Galaxy ToolShed. In Genome Biology, 15 (2), pp. 403. [doi:10.1186/gb4161]

Dickins, Benjamin and Rebolledo-Jaramillo, Boris and Su, Marcia Shu Wei and Paul, Ian M and Blankenberg, Daniel and Stoler, Nicholas and Makova, Kateryna D and Nekrutenko, Anton (2014). Controlling for contamination in re-sequencing studies with a reproducible web-based phylogenetic approach. In BioTechniques, 56 (5), pp. 154--141. [doi:10.2144/000114146]

Langmead, Ben and Salzberg, Steven L (2012). Fast gapped-read alignment with Bowtie 2. In Nature Methods, 9 (4), pp. 357ð359. [doi:10.1038/nmeth.1923]

Lanno, S. M., S. M. Gregory, S. J. Shimshak, M.K. Alverson, K. Chiu et al., 2017 Transcriptomic analysis of octanoic acid response in Drosophila sechellia using RNA-sequencing. G3 Genes, Genomes. Genet. 7: 3867-3875.

Schaeffer, S. W., Bhutkar, A., McAllister, B. F., Matsuda, M., Matzkin, L. M., O'Grady, P. M., Rohde, C., Valente, V. L. S., Aguadé, M., Anderson, W. W., Edwards, K., Garcia, A. C. L., Goodman, J., Hartigan, J., Katoaka, E., Lapoint, R. T., Lazovsky, E. R., Machado, C. A., Noor, M. A. F., ... Kaufman, T. C. (2008). Polytene Chromosomal Maps of 11 Drosophila Species: The Order of Genomic Scaffolds Inferred From Genetic and Physical Maps. Genetics, 179(5), 1601-1655. https://doi.org/10.1534/genetics.107.086074

Schrider, D. R., Ayroles, J., Matute, D. R., & Kern, A. D. (2018). Supervised machine learning reveals introgressed loci in the genomes of Drosophila simulans and D. sechellia. PLoS genetics, 14(4), e1007341. https://doi.org/10.1371/journal.pgen.1007341