Bornberglab
Molecular Evolution and Bioinformatics

Genome assembly and annotation of 7 populations of *Drosophila melanogaster*

Background: The mechanisms that allow the appearance of new genes are mainly the gene duplication and the insertion of novel elements. However, over the last 10 years, t became clear that additional genetic mechanisms underlie the creation of new protein coding genes. Among them, one of the more fascinating is the DE NOVO appearance of new genes (*1*). Strong evidence demonstrated that many of these orphan genes can be involved in primordial functions (*2*, *3*)). The de novo genes have already been well investigated in some model organisms, but their mechanisms of appearance remain unclear, as their appearance has not been yet studied in the populations of a unique species.

Objectives: 7 Populations of Drosophila melanogaster distributed in Europe and in Zambia have been sequenced. In order to study the genes appeared de novo, the genomes have to be assembled, and then annotated. For each of the genomes, different scripts in bash will be used to follow the different steps allowing a genome assembly and it annotation. If time is remaining, the de novo genes will be detected and studied.

Requirements:

- Interest in genomics and study of the genomes contents
- Interest in learning and/or practicing linux and bash scripts

Methods: Detecting of repetitive elements, paired-end assembly, statistics on the genome contents, Genome annotation (GeneMapper, Exonerate, Gnomon).

Supervision: Anna Grandchamp, a.grandchamp@uni-muenster.de, Molecular evolution and bioinformatics Group (http://bornberglab.org/).

References

1. C. L. Brown, I. A. Aksay, D. A. Saville, M. H. Hecht, *Journal of the American Chemical Society* **124**, 6846–6848 (2002).
2. K. Khalturin, G. Hemmrich, S. Fraune, R. Augustin, T. C. Bosch, *Trends in Genetics* **25**, 404–413 (2009).
3. A. M. Gubala *et al.*, *Molecular biology and evolution* **34**, 1066–1082 (2017).

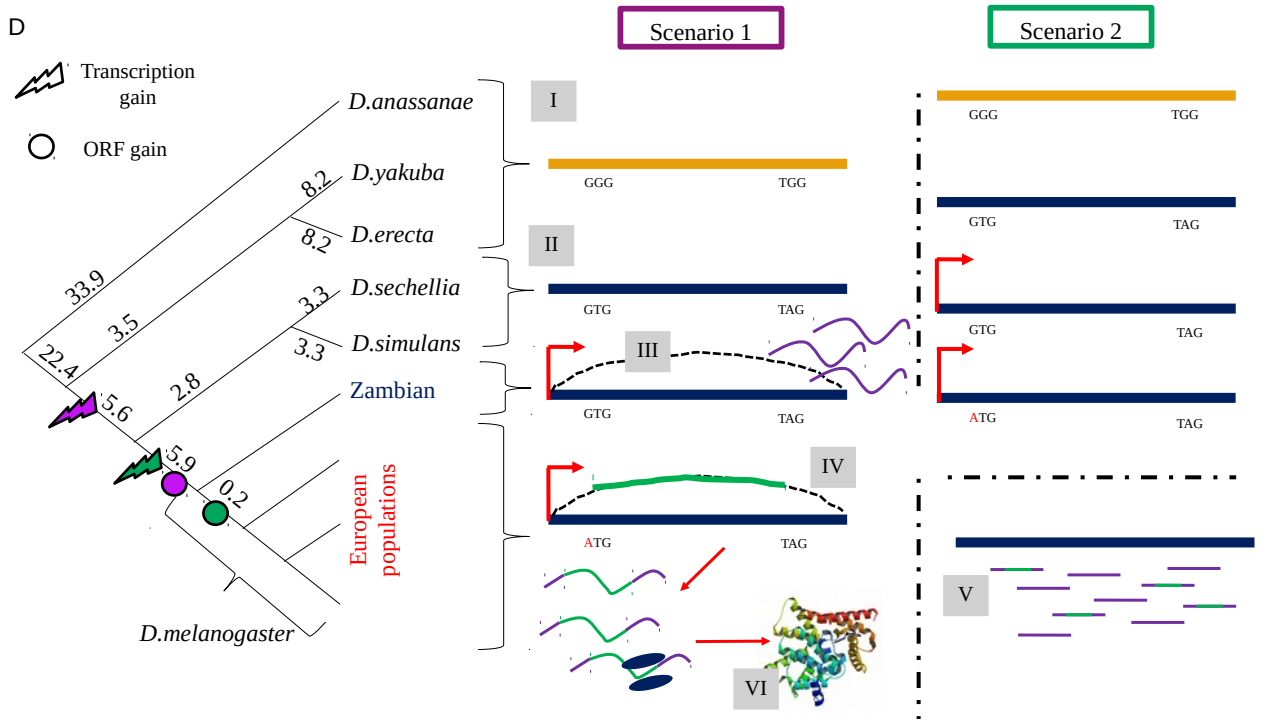

Figure 1: Schematics of how denovo protein coding genes may emerge along the drosophila tree:Tree on left follows data given by time tree with numbers along branches giving the approximate times elapsed in million years. Cross and circle represent possible emergence events of first transcription and an ORF formation. The broad arrow symbolises that transcript/gene flow between European populations is possible. The proteins form less and more secondary structural elements in Scenario 1 and 2 respectively as they are of different age. I : "no sign of homology detectable at the DNA level". II: "non-coding outgroup". III: "transcribed ingroup, but not yet translated". IV: "ribosome binding and translation into a de novo protein". V: "RNAseq of transcripts". VI: "Mapping reads on genome".