Genomic and transcriptomic evolution of *D. Suzukii*

Project manager:
Cristina VIEIRA-HEDDI
cristina.heddi@univ-lyon1.fr

Superviser:
Arnaud MARY
mary.univ.lyon1@gmail.com

Students:

Chloé AJOULAT
chloe.aujoulat@etu.univ-lyon1.fr

Tommaso BARBERIS
tommaso.barberis@etu.univ-lyon1.fr

Bertrand HUGUENIN-BIZOT
bertrand.huguenin-bizot@etu.univ-lyon1.fr

Marie VERNERET
marie.verneret@etu.univ-lyon1.fr

Contents

Abbreviations

BAM Binary SAM file. 6

CNRS Centre National de la Rercherche Scientifique. 3

DE differentially expressed. 4, 5, 7

dnaPipeTE De Novo Assembly and Annotation Pipeline for Transposable Elements. 8

G Generation. 3–9

GATK Genome Analysis ToolKit. 6, 8

LBBE Biometry and Evolutionary Biology Laboratory. 3, 4

NGS Next Generation Sequencing. 8

SNP Single Nucleotide Polymorphism. 4–9

TE Transposable Elements. 4–6, 8, 10

UCBL Université Claude Bernard Lyon 1. 3, 9

VCF Variant Call Format. 6–8

1 Introduction

1.1 Laboratory

The Biometry and Evolutionary Biology Laboratory (LBBE) is located in Lyon and is part of the Centre National de la Rercherche Scientifique (CNRS) and the Université Claude Bernard Lyon 1 (UCBL). Created in 1966, the LBBE is today composed of 4 departments gathering a total of 14 research teams. Cristina **Vieira-Heddi**, our project manager and lecturer at UCBL, works within the multi-scale coevolution department of the LBBE and more specifically in a team focused on Genetics and Evolution of Interaction. The main goal of the team is to understand better evolutionary implications of interactions between multiple organisms' components whether it is genes or transposable elements.

1.2 Context

Since its recent invasion in the European and American continents, *Drosophila suzukii*, has become a major pest of berry crops. Contrary to other *Drosophila* species that develop themselves on damaged or rotten fruit, *D. suzukii* is able to lay its eggs in healthy fruit before harvest, using its sclerotinized ovipositor. The colonizing individuals' ability to adapt to new environmental conditions makes biological invasions often used to study rapid adaptation [2]. In order to control the *D. suzukii* population it is therefore required to improve our knowledge about its ability to adapt to a local environment outside its origin areas.

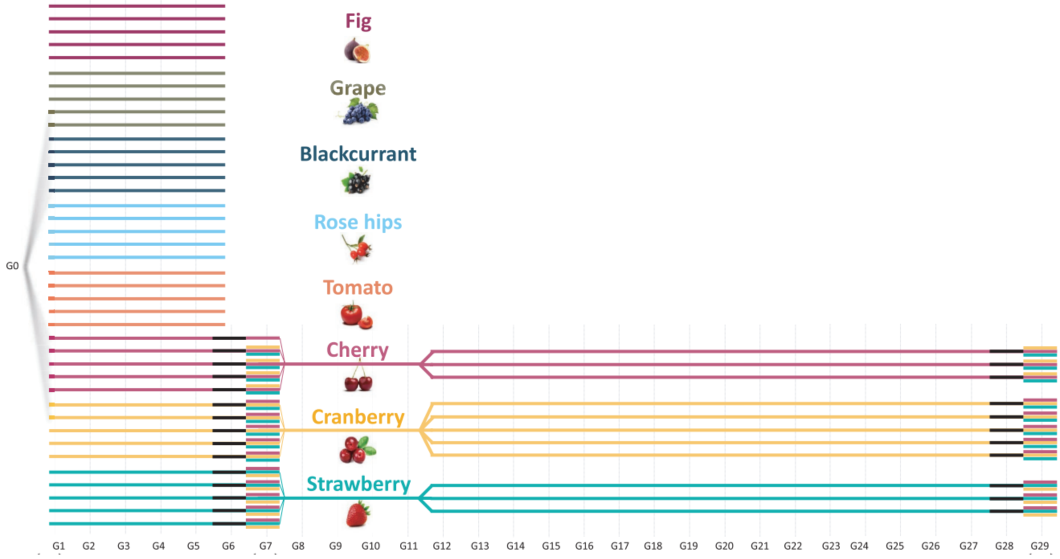

Figure 1: Experimental evolution design showing the different fruit media [1].

A study of this pest adaptation based on a selection experiment has already been conducted to see the flies' evolution at the phenotypic level [1]. They have bred 9 generations of *D. suzukii* in a laboratory without any environmental factors. The population obtained corresponds to the G0 generation and has been separated on several fruit purees media in order to have 400 individuals per population.

As we can see above on the Figure 1, by the fifth generation of experimental evolution, populations on blackcurrant, fig, grape, rose hips or tomato were either extinct or close to extinction, with fewer than 30 individuals per population, so they had been excluded from the study. Population replicates for each cherry, strawberry and cranberry media has been pooled together on G7 in order to avoid inbreeding depression. On G11, populations had recovered, so they have split them again in 3 replicates for cherry and strawberry and in 5 for cranberry.

The fitness was measured on G1, G7 and G20 flies. The experiment demonstrated a temporal adaptation of *D. suzukii* populations in the three different selective environments (cherry, strawberry and cranberry). This process produces a fitness improvement which corresponds to a greater ability to reproduce and could explain the adaptation abilities of *D. suzukii.*
Based on the same experiment, a second study has been started by the LBBE to see if the *D. suzukii* genome and transcriptome are affected by the environment.
The project is focused on both genes and Transposable Elements (TE) expression in order to find the variants explaining the local environment adaptation of *D. suzukii.* The first results showed that the evolution of *D. suzukii* on fruit media induces several gene and TE expression changes. A third of the changes seems to be depending on time generation and not on the fruit. Moreover, genes differentially expressed seem to be related to reproduction.

2 Project objectives

2.1 Problematics

Whereas the phenotypic study gave very interesting results on the flies' fitness improvement, the first results obtained in the genomic study remain to be confirmed. Indeed, a significant number of differentially expressed genes across generations has been found with RNAseq analysis [12] between G0 and G7 but not between the different environments of G7.
Moreover, in differentially expressed (DE) genes an increase of the number of Single Nucleotide Polymorphism (SNP) was found in G7 compared to G0, but this result was not expected. Natural selection generally keeps alleles which allow the best adaptation, and it is synonymous with a decrease in SNP number. Thus why was this difference found? Is it linked to DE genes? Does it mean that a real selection process happened in only 7 generations?

2.2 Main goal

Our role is to contribute to understanding better the evolutionary process that allows flies to adapt to local environments.
Since we have DNAseq data available, we will perform a variant calling on the whole genome in order to count the number of different genomic positions between G0 and G12, and between fruits media. In addition, we would take into account the proportion of called bases for every position to study if a selection of predominant alleles has happened. In other words, we would analyse polymorphism across generations and between the fruit media.

After that, we would like to confirm the previous SNP analysis on differentially expressed genes (observed with RNAseq data) with the results obtained in the DNA poolseq analysis.
This confirmation could help to understand how and why the number of SNP increase across generations in the RNAseq analysis.

In this way, we would like to produce a clear representation of the SNP proportion between RNA et DNAseq data for all the genes, intergenic regions and then only in DE genes with some tables and Euler diagrams. We might also see more specifically SNP in upregulated and downregulated DE genes or in the different gene families. Finally, it could be interesting to calculate several genetic diversity indexes such as synonymous and non-synonymous mutation rates or other genetic parameters such as the Watterson's θ [9] or the Tajima's D [13]. These estimators are often used to describe the genetic diversity in a population by counting the number of polymorphic sites. We could thus estimate genetic drift studying neutral markers which are in intergenic regions and estimate selection studying gene polymorphisms.

2.3 Secondary goal

The study is not only focused on genes, but also on TE. In the previous RNA seq analysis some TE has been found with significant expression changes. However, they can have a functional effect on its host and could be important in the adaptation process, interfering in gene regulation during evolution [14]. So it could be interesting to analyse more specifically each TE family to see if they have an impact and if they are correlated with DE genes. In this context, we would like to determine TEs age on our data and compute their abundance. It will allows us to compare the G0 and G12 TEs with abundance graphs. At first sight, no differences are expected between the two generations.

3 State-of-the-art

Using RNA in order to detect genomic variations in expressed regions like DNA-seq has been already described by Piskol et al. [11].

Jehl et al. [5] confirmed how RNAseq is a reliable method for SNP calling comparing results from RNAseq data with results obtained from DNAseq.

Kapin et al. [6] have already done a genomic analysis of *D. melanogaster* and used different standard population genetic parameters: π [7], Watterson's θ and Tajima's D on poolseq data.

About TE analysis, V. Merel et al. [8] already performed a study about TE in the *D. suzukii* assembly genome, so here we want to do the same but with our data and perform also an abundance estimation of TE [4]. The only problem is that we only have non assembled reads so we will also base our analysis on C. Goubert et al. [3] who used a *de novo* assembly and annotation method on TEs.

4 Material and methods

4.1 Dataset

In order to perform our project, several data obtained by L. Olazcuaga et al. will be used [1]. Paired-end RNAseq data (from ovipositor of 20 *D. suzukii*) are available of two biological replicates forG0 and G7 (for cherry and strawberry media) including two different population replicates and paired-end DNAseq data from a poolseq (40 individuals) are available for G0 and G12 (cherry, cranberry and strawberry media). For the RNAseq data, we already have the SNP calling results from the *HaplotypeCaller* tool (Genome Analysis ToolKit (GATK)) in the `.vcf` (Variant Call Format) format.

We also have an annotated genome of *D. suzukii* in the `.gff` format with the correspondent file in the `.fasta` format for the nucleic sequences [10]. In regards to TE, we have a fasta file that contains TE sequences as well as text file for their annotations [8].

4.2 Work Environment

All the project will be conducted on a Linux environment on the master's server, pedago-ngs. We will setup a Git reposity on pedago-ngs and also a Gitub page online.
The Git Hub repository will also contain at least one file written in markdown, precising the commands we will use and their corresponding options. This file will also include the different tables and graphs which will be produced. This will constitue our deliverable.
The different tools will mainly be excecuted in `bash`.

4.3 Tools and steps

4.3.1 Steps

The first step before any analysis is the DNAseq data pre-processing. It is the mandatory first phase that must precede all variant discovery.
It involves pre-processing the raw sequence data (provided in FASTQ) to produce analysis-ready BAM files. This consists in doing a quality control of the raw sequences (trim the data if quality control is not satisfying) and doing an alignment on a reference genome that will be executed with the same aligner than the one used for *D. suzukii* RNAseq mapping. Some data cleanup operations will also be necessary to correct for technical biases and make the data suitable for the next analysis. For the SNP calling two different tools will be used to confirm the results. The both of them will produce VCF files containing a SNP list for each sample. These files will be the base to compare the *D. suzukii* genome at the G0 and G12 generations but also to compare DNAseq results with the RNAseq ones. To decrease as much as possible the bias, the closer pipeline than the one used for the RNA SNP calling [12] will be performed. Finally, to compute the genetic diversity indexes the reading frame of each gene will have to be defined. Indeed, it is very important to know where each gene starts in order to know the position of the nucleotides in the codons and determine synonymous and non synonymous substitutions. These analyses would be used in order to determine if the presence of variants could be explained by heterozygosity or polymorphism. This point would concern polymorphism and SNP, as well as the analyses of the genes which are differentially expressed.

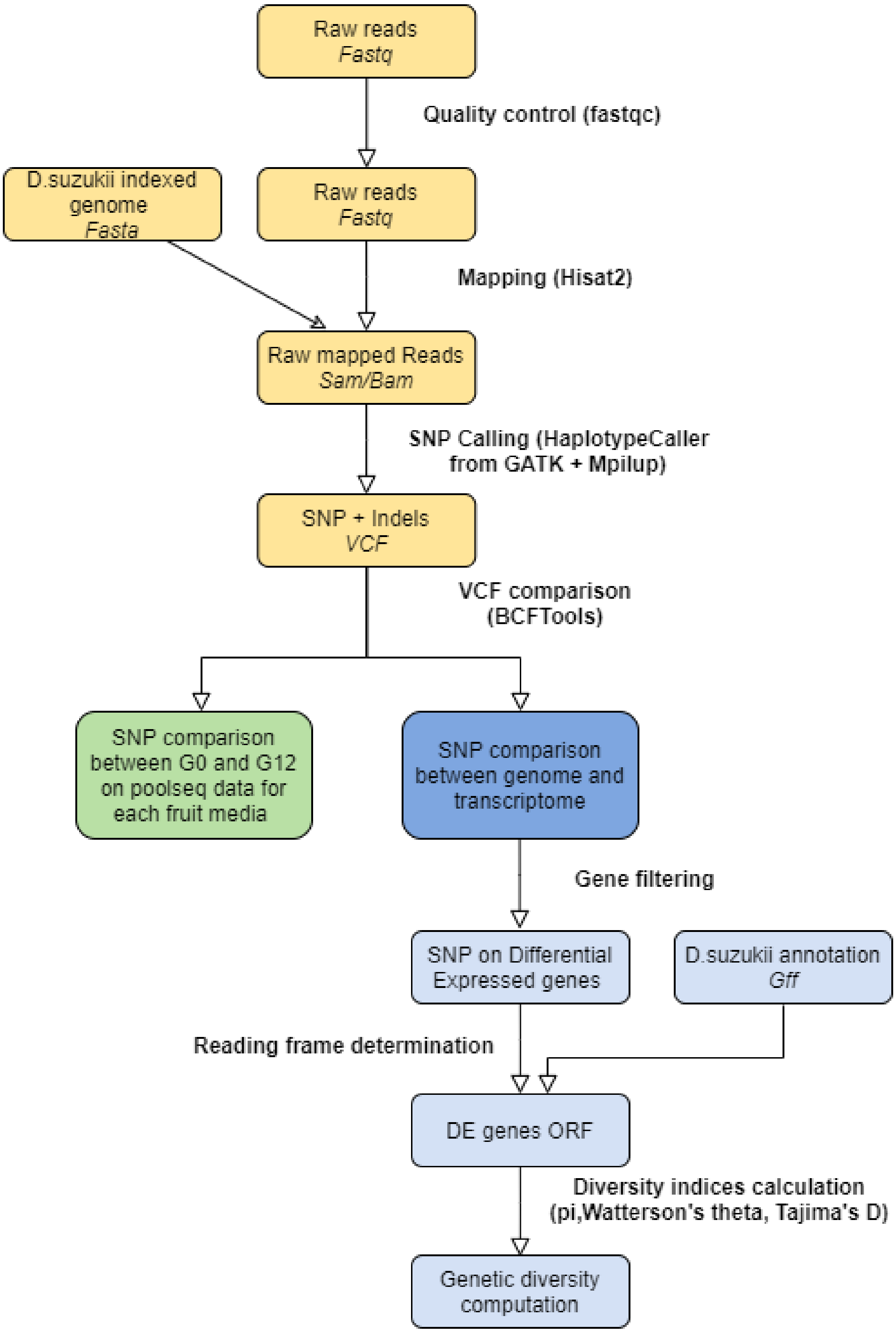
Raw reads
Fastq
Quality control (fastqc)
D.suzukii indexed genome
Fasta
Raw reads
Fastq
Mapping (Hisat2)
Raw mapped Reads
Sam/Bam
SNP Calling (HaplotypeCaller from GATK + Mpilup)
SNP + Indels
VCF
VCF comparison (BCFTools)
SNP comparison between G0 and G12 on poolseq data for each fruit media
SNP comparison between genome and transcriptome
Gene filtering
SNP on Differential Expressed genes
D.suzukii annotation
Gff
Reading frame determination
DE genes ORF
Diversity indices calculation (pi,Watterson's theta, Tajima's D)
Genetic diversity computation

Figure 2: Diagram representing the main steps of the SNP analysis.

The first step is to produce Variant Call Format (VCF) files containing the SNP (yellow part). The second is the comparison of different VCF files : between G0 and G12 to see if there is a temporal difference (green part) and between transcriptome and genome to see which SNP and thus which alleles are expressed (blue part) and more particularly for de DE genes.

About the analysis of *D.suzukii*'s TEs, we will start from the poolseq raw reads. The first step is to assembly the reads and then to identify the transposable elements doing a TE calling. Once indentification completed, we will have to annotate each sequence. These sequences will then be counted in order to have a TE frequency which will allow to estimate TE abondance in the *D.suzukii* genome. Abondance graphs will be produced for each G0 and G12 samples and will inform us about TE relative diversity. Indeed, if recent TEs are observed only in G12 data it will mean that a change of diversity happened during the experiment.

4.3.2 Tools

Tools that we will use:

- ☛ **FASTQC v.0.11.8**: is a quality control tool used for high-throughput sequencing data. In this study, we will perform FastQC on all the DNA data we have: the data for G0, and the other one for G12 (cranberry, strawberry and cherry).
- ☛ **HiSAT2 v.2.1.0**: is an alignment program used for mapping next-generation sequencing reads on a reference genome here the one of *D. suzukii.* This tool can be used for both DNA seq and RNA seq data.
- ☛ **GATK v.4.2.2.0**: is a set of tools which allow the detection of variants in high-throughput sequencing data, variant discovery and genotyping. Here, in particular, we will mainly use the GATK `haplotypeCaller`.
- ☛ **SAMtools v.1.9** is a tool for analyzing the coverage of mapped reads on the reference genome. We will use more particularly the Samtools `mpilup` tool.
- ☛ **BCFTools v.1.13**: is a set of utilities allowing VCF files comparison. In this context, we will use BCFTools to study the SNP differences between G0 and G12 in the poolseq experiment and between genome and transcriptome analysis.
- ☛ **Packages R v.3.6.3**: In order to get the visualization of the corresponding results, we will use `R` software, Rstudio interface, and different packages such as `ggplot`.
- ☛ **dnaPipeTE v.1.3**: fully automated pipeline designed to assemble and quantify repeats from genomic NGS reads.
- ☛ **Trinity v.2.5.1**: *De novo* assembler used in the dnaPipeTE pipeline in several iteration in order to improve the repeat assembly.
- ☛ **RepeatMasker**: Used in the dnaPipeTE pipeline to annotate the assembled contigs of TE.
- ☛ **TEanalysis v.4.6**: to generate an abundance graph of TE in fonction of their divergence. It allows to test enrichment.

5 Constraints

An important constraint for the project will be the time. Indeed, the deadline for the project is on the 15th of December. We will need to organize our work in order to learn how to manipulate the tools presented above and sum up the biological notions related to the bioinformatic work for this deadline. In addition, we will have to schedule several meetings with our supervisors to verify if everything is correct and help us to figure the technical problems out.

One of the other obstacles we will have to face is to handle correctly the amount of tools that we will use. We will have to make choices about settings of software programs, but also about the acceptance criteria to determine which SNP is significant in the evolution process of *D. suzukii*. Maybe we will have to install and use even more software during the global analysis, that would be in intermediate steps.

5.1 Technical constraints

For our project, we have to start from other people's work and in order to produce comparable results we have to use the same version for software programs/libraries, and sometimes it will be difficult to find the right version if the author lacks to note it. For the computational part, we have chosen to use the pedagogical server from the UCBL, but this server has been recently hosted on a new machine, so there could be problems related to it, such as the lack of some tools that are not installed yet, for example. In addition, in our group we don't have the same university schedule so, sometimes it can be difficult to organize team meetings. Finally, some tools need specific packages which are accessible only on a paid subscription basis, so we will have to see if we can use it or not.

5.2 Scientific constraints

About the experimental protocol adopted to obtain data and about their types, we have several doubts:

- in the DNAseq data, we have 3 types of fruits against 2 type for the RNAseq (no data for the cranberry population);
- we have to compare SNP between RNAseq and DNAseq from two different generations (G7 for RNA and G12 for DNA). Even if, 5 generations are probably not enough to see new mutations, it would have been easier to compare the same generation;
- the way *D. suzukii* breed is a central point of its evolutionary advantage on other *Drosophila*, because it has a sexual apparatus that is strong enough to lay in ripe fruits, but the protocol has been done using fruit puree;
- we also could analyse the difference between fruits in G5 because several populations were either extinct or close to extinction, so we find that it can be interesting if a genetic adaptation happened in other fruits at this moment;
- we may not have enough power for our statistical tests with only two replicas for RNAseq data.

6 Project progress

We have a total of 26 days (two days every week) to complete this document and the project. Globally, the project spans from 16th of September to 10th of December. The first week, waiting for the meeting with our project director, we took advantage to explore the bibliography and search complementary information about the subject. After that, we have spent two weeks on the production of this document. We installed the necessary tools, imported the data on the server and began the first analysis according to the needs.

We will then divide the tasks in three main groups:

☛ variant calling analysis on the DNAseq data (deadline: end of October);

☛ genetic diversity computation (deadline: begin of December);

☛ TE analysis (deadline: begin of December).

In order to have the three steps independant, we could take a fake vcf file as an example and test the further analysis to see how genomic diversity measures can be calculated to progress on the second step without having the results of the first one. We will leave the last week and a half free to produce and organize the oral presentation of the project.

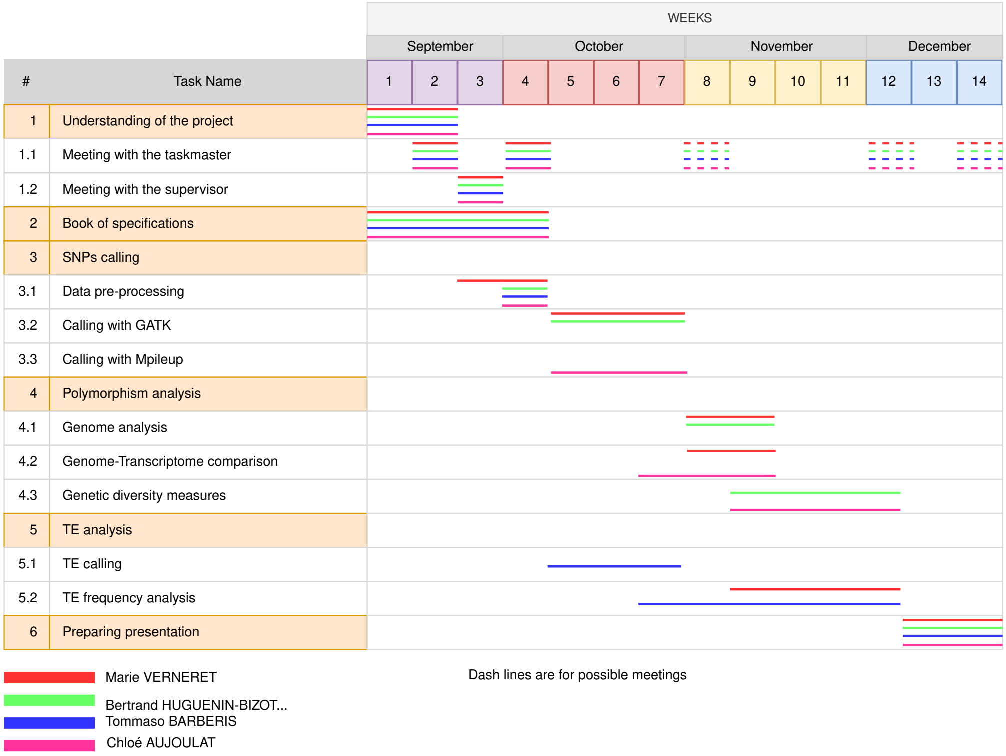

Figure 3: Planning of the project

References

[1] *Adaptation and correlated fitness responses over two time scales in Drosophila suzukii populations evolving in different environments - Olazcuaga - 2021 - Journal of Evolutionary Biology - Wiley Online Library.* URL: https://onlinelibrary.wiley.com/doi/full/10.1111/jeb.13878.

[2] *Deciphering the Routes of invasion of Drosophila suzukii by Means of ABC Random Forest — Molecular Biology and Evolution — Oxford Academic.* URL: https://academic.oup.com/mbe/article/34/4/980/2953204.

[3] Clément Goubert et al. "De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti)". In: *Genome Biology and Evolution* 7.4 (Mar. 11, 2015), pp. 1192–1205. ISSN: 1759-6653. DOI: 10.1093/gbe/evv050. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419797/ (visited on 10/08/2021).

[4] A. Hua-Van et al. "Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences". In: *Cytogenetic and Genome Research* 110.1 (2005). Publisher: Karger Publishers, pp. 426–440. ISSN: 1424-8581, 1424-859X. DOI: 10.1159/000084975. URL: https://www.karger.com/Article/FullText/84975 (visited on 10/07/2021).

[5] Frédéric Jehl et al. "RNA-Seq Data for Reliable SNP Detection and Genotype Calling: Interest for Coding Variant Characterization and Cis-Regulation Analysis by Allele-Specific Expression in Livestock Species". In: *Frontiers in Genetics* 12 (2021), p. 1104. ISSN: 1664-8021. DOI: 10.3389/fgene.2021.655707. URL: https://www.frontiersin.org/article/10.3389/fgene.2021.655707 (visited on 10/07/2021).

[6] Martin Kapun et al. *Genomic analysis of European Drosophila melanogaster populations on a dense spatial scale reveals longitudinal population structure and continent-wide selection.* Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article. May 4, 2018, p. 313759. DOI: 10.1101/313759. URL: https://www.biorxiv.org/content/10.1101/313759v2 (visited on 10/07/2021).

[7] *Mathematical model for studying genetic variation in terms of restriction endonucleases.* URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC413122/.

[8] Vincent Mérel et al. *The worldwide invasion of Drosophila suzukii is accompanied by a large increase of transposable element load and a small number of putatively adaptive insertions.* Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article. Nov. 7, 2020, p. 2020.11.06.370932. DOI: 10.1101/2020.11.06.370932. URL: https://www.biorxiv.org/content/10.1101/2020.11.06.370932v1 (visited on 10/07/2021).

[9] Masatoshi Nei. *Molecular Evolutionary Genetics.* Publication Title: Molecular Evolutionary Genetics. Columbia University Press, Mar. 2, 1987. ISBN: 978-0-231-88671-0. DOI: 10.7312/nei-92038. URL: https://www.degruyter.com/document/doi/10.7312/nei-92038/html.

[10] Mathilde Paris et al. "Near-chromosome level genome assembly of the fruit pest Drosophila suzukii using long-read sequencing". In: *Scientific Reports* 10.1 (July 8, 2020), p. 11227. ISSN: 2045-2322. DOI: 10.1038/s41598-020-67373-z.

[11] Robert Piskol, Gokul Ramaswami, and Jin Billy Li. “Reliable identification of genomic variants from RNA-seq data”. In: *American Journal of Human Genetics* 93.4 (Oct. 3, 2013), pp. 641–651. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2013.08.008.

[12] Marie Tabourin et al. “Analyse de l’expression des gènes et des éléments transposables chez Drosophila suzukii suite à une expérience d’évolution expérimentale”. In: (), p. 35.

[13] F. Tajima. “Statistical method for testing the neutral mutation hypothesis by DNA polymorphism”. In: *Genetics* 123.3 (Nov. 1989), pp. 585–595. ISSN: 0016-6731. DOI: 10.1093/genetics/123.3.585.

[14] *Transposable elements drive rapid phenotypic variation in Capsella rubella — PNAS*. URL: https://www.pnas.org/content/116/14/6908.