# SHORT TERM SCIENTIFIC MISSION (STSM)
# SCIENTIFIC REPORT

**This report is submitted for approval by the STSM applicant to the STSM coordinator**

**Action number:** CA17108
**STSM title:** Population genomics of insect vector
**STSM start and end date:** 01/07/2019 to 31/08/2019
**Grantee name:** Laura Vavassori

**PURPOSE OF THE STSM:**

(max.200 words)

The general purpose of the STSM was to broaden my knowledge on population genomics of insect vectors. Specifically I envisioned to acquire expertise on next generation sequencing data analysis, focusing on double-digest restriction site-associated DNA sequencing (ddRADseq)[1] data. I aimed at learning about bioinformatics processes for identification of SNPs as genetic markers from ddRAD data. Using my own dataset, I aimed at familiarizing with software used to infer population histories based on ddRAD sequencing data to reveal the origin of the invasion of the Asian tiger mosquito (*Ae.albopictus*) in Switzerland and at evaluating the levels of gene flows between recent and established populations of *Ae. albopictus* in Switzerland and its neighbouring countries.

**DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS**

(max.500 words)
My training focused on population genomics analysis of double-digestion restriction site-associated DNA sequencing (ddRAD) data. I have conducted the laboratory work and sequencing of the data before the start of the STSM, as part of my PhD project.

As part of the STSM, I have acquired knowledge on bioinformatics processes using different software to process my own raw data. I have learned how to evaluate the quality and consistency of sequencing data using FASTQC. I have used the software STACKS [2] to examine raw sequencing, check barcodes and demultiplex my raw data. With the alignment programme bwa mem [3], I aligned the paired-end sequencing reads to the reference genome kindly provided my the STSM hosts. I could inspect the quality of the alignment with programmes such as Samtools[4] and Picardtools [5]. In order to identify the variants in the sequencing data, I have used two variant callers, namely STACKS [2] and mpileup (BCFtools) and kept as tags only the overlap called by the two programmes. This conservative approach was chosen as a variety of software are available for variant calling, and choosing one software rather then another one could have an impact on the downstream population genetic analysis results[6]. Before the SNP calling, I applied a series of filtering with programmes such as PLINK [7] and VCFtools[8] to remove missingness and poor quality reads. Finally, I have obtained the SNPs and started to evaluate the population structures based on these genetic markers. I familiarized with both model-based Bayesian methods such as ADMIXTURE[9] and model-free based method such as PCA (Principal component analysis) to investigate genetically distinguishable populations in my dataset.

I have used different R packages for the analysis of genomics data, from data exploration and cleaning to downstream analysis. Example of these packages are SimRAD to estimate the number of loci expected based on genome information prior to the bioinformatics analysis, or the package adegenet and hierstat to describe and identify genetic cluster and calculate standard population genetic parameters.
I am now a competent user of the bioinformatics software UNIX and I have acquired experience to work with large amount of data.

**DESCRIPTION OF THE MAIN RESULTS OBTAINED**

This STSM refers to my PhD project, where I am investigating the passive dispersal of the Asian tiger mosquito, *Aedes albopictus*, in Switzerland and beyond. I have collected 120 individuals covering the majority of sites in Switzerland where there are either established populations or this species has been intercepted. Next, neighbouring countries have been sampled focusing on regions bordering Switzerland with a more extensive sampling of Italian populations, as this is the most likely source for the introduction into Switzerland. Out of these 120 samples, I had to exclude 10 samples due to poor sequencing quality. In total, I obtained 648 million reads, with an average of 5 million reads per sample. A total of 48 million reads aligned to the reference genome, with 4 million reads in average per sample (76%). The variant callers STACKS and mpileup identified a different number of variants, namely 160,265 and 33,483 respectively. The overlap of variants identified by both variant callers are 26,649 variants. Therefore the final filtered dataset consists of 4993 SNPs and 110 individuals.

**FUTURE COLLABORATIONS (if applicable)**

As the downstream analyses are not completed yet, future discussion on the final results are planned. From the work conducted during the STSM, at least one mutual publication is envisioned in 2019. Outside of the framework of this current project, the collaboration between Yale University and the Swiss Tropical and Public Health Institute has now been established.

**References**


1. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. Orlando L, editor. PLoS ONE. 2012;7(5):e37135.

2. Rochette NC, Catchen JM. Deriving genotypes from RAD-seq short-read data using Stacks. Nat Protoc. 2017 Dec;12(12):2640–59.

3. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

4. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

5. Broad Institute. Picard Toolkit [Internet]. [cited 2019 Sep 16]. Available from: https://broadinstitute.github.io/picard/

6. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015 Dec;5(1):17875.

7. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience [Internet]. 2015 Feb 25 [cited 2019 Sep 16];4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4342193/

8. variant call format and VCFtools | Bioinformatics | Oxford Academic [Internet]. [cited 2019 Sep 16]. Available from: https://academic.oup.com/bioinformatics/article/27/15/2156/402296

9. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64.