Report Short Visit Grant "ESF Research Networking Programmes"

Applicant: Valentina Mastrantonio

Purpose of the visit

Hybridization and introgression between animal species are very widespread processes in nature, as showed by several studies during the last few years. However, a strong debate remains about the role played by these processes in promoting and maintaining genetic diversity, which represents a key issue for applied sciences, such as conservation biology. Recently, Next-generation sequencing (NGS) technologies are offering the possibility to investigate genetic diversity at scales not previously possible.

In my PhD project, I'm investigating hybridization, introgression and its evolutionary consequences on animal species, monitoring an artificial sympatric area between the coastal mosquito species *Aedes mariae* and *Ae. zammitii* across about 25 years. The results thus far obtained showed that the two species have not evolved a complete reproductive isolation, but they still hybridize, even though F1 hybrid males are unfit. I'm currently investigating the pattern of this gene exchange at mitochondrial marker, but I would extend this analysis also to the nuclear genome. To this purpose, I would isolate Single Nucleotide Polymorphisms (SNPs) using NGS to analyse the pattern and the dynamic of nuclear gene introgression. SNPs are ideal markers for hybridization and introgression assessment because they are frequent in the genome and can be rapidly, reliably and cheaply isolated and genotyped using NGS. Therefore my purpose for this "short visit grant" (15 days) was visiting the laboratory of the Prof. Andrea Crisanti, which represents one of the most important centre in this context, with the aim to learn basic skills of NGS and discuss with him and his colleagues the best approach to reach my objective.

Description of the work carried out during the visit and of the main results

During the two weeks spent at the Imperial College, I have analysed what kind of experimental design could be the most appropriate for isolating SNPs in *Aedes mariae*/*Ae. zammitii* biological system. Very interesting and useful suggestions emerged that allowed me to delineate the best approach to reach my aim and the trade-offs that I should consider for an appropriate experimental design. Below I describe the principal outcomes and the main conclusions of my visit.

i) Whole-genome sequencing or reduced-representation methods?

Hybridization and introgression often occur in a complex spatial and temporal context and therefore they need detailed molecular tools to be investigated. This often means having a high number of markers distributed across the genome that allow a good coverage to accurately estimate interspecific gene flow. In this context, two alternatives could be possible: sequencing the whole genome or using reduced-representation methods. Two aspects pushed us to choose reduced-representation methods. To date, it is true that new genomes are sequenced regularly and are affordable for many groups of research, but this approach remains still high expensive for organisms having a big and complex genome, as the *Aedes* species. Genomic data showed a genome 5 times greater in *Aedes aegypti* than several other *Anopheles* mosquitoes (e.g. 1,376 and 278 million base pairs in *Ae. aegypti* and *An. gambiae*, respectively). The complete genome sequencing therefore does not seem the most feasible solution, considering also that it should be due for both *Ae. mariae* and *Ae. zammitii*. Furthermore, we have to consider that many biological questions do not require the sequencing of a single base of the genome to be answered, but they could need only the polymorphisms measured in a subset of genomic regions. Indeed the genome of many organisms is composed and organised in linked blocks and well-spaced markers can provide the sufficient coverage. The use of reduced-representation methods therefore seems the best solution, because it could allow us to produce markers directly, avoiding the cost of genome assembly, and use the economic resources in a higher coverage or sampling.

ii) What kind of reduced-representation method?

Reduced-representation methods use restriction enzymes to produce a reduced representation of the genome, allowing over-sequencing of nucleotides flanking the restriction site. According to these methods, genomic DNA is digested and cut by restriction enzymes to generate restriction fragments, that are further pooled, selected by size and sequenced by NGS. Genetic markers are then isolated using the polymorphisms occurring in these sequences.

Three main classes of these methods exist: reduced-representation sequencing (RRLs; CRoPS), Restriction-associated DNA (RAD-seq); and low coverage genotyping (MSG; GBS). Despite they involve similar key steps, two main differences could be pointed out between the three classes of methods, as shown in Table 1: the need of a reference genome and the possibility to use a paired-end sequencing approach. The only technique that allows me to overcome the need of an available reference genome and to use a paired-end sequencing approach is RAD-seq. These two aspects, together with other issues (see below), contributed to the choice of RAD-seq method to isolate SNPs in *Aedes mariae* and *Ae. zammitii*.

Table 1. Principal requisites and characteristics of the three classes of genotyping methods.

	Use of enzymes	Reference genome	Multiplex	Paired-end approach	Sequence
Reduced-representation sequencing (RRLs; CRoPS)	✓	✓	✓	✓	Ends of fragment
RAD-sequencing	✓	✗	✓	✓	All fragment
Low coverage genotyping (GBS; MSG)	✓	✓	✓	✗	Ends of fragment

iii) Why RAD-sequencing is appropriate for Ae. mariae/Ae. zammitii system?

RAD-sequencing is a method that allows to identify SNPs in the short regions flanking the restriction sites of a given enzyme and to compare them between individuals, regardless of the length of the restriction fragments. Using RAD-seq, therefore, it is possible to reduce the costs compared to whole-genome sequencing, indeed only part of the genome is sequenced, but maintaining a high genome-wide marker density. As the other methods above cited, RAD-seq begins with digestion of DNA by restriction enzymes. After digestion the restriction fragments are ligated to an Illumina adapter (P1 adapter) that matches with the end of the fragment and contains a MID (multiplex identifier sequences), that allows to uniquely identify the individual. All samples can therefore be pooled and randomly sheared to produce fragments of a few hundreds of base pairs. These fragments are then ligated to a second adapter (P2 adapter) and amplified by Polymerase Chain Reaction (PCR). The amplified fragments are therefore selected by size, approximately 200–500 base fragments, and RAD-seq library is constructed, sequencing the selected fragment by the Illumina platform (Figure1). To date this platform allows to sequencing about 300 nucleotides flanking the restriction site that can be screened to find SNPs. Some advantages with respect to the other reduced-representation methods above mentioned can be underlined. First, differently to reduced-representation sequencing and low coverage sequencing methods, it allows to use more than two restriction enzymes, giving the possibility to increase the amount of fragments to be analysed. Second, it includes an extra shearing step to capture all the restriction sites, thus having possible markers spread throughout the genome. Finally, it is cheaper and less laborious than other reduced-representation methods because it allows to pool samples before adaptor ligation and therefore to carry out all the remaining steps of the protocol on the pooled library.

One of the most important characteristics of RAD-seq, that makes it particularly suitable for my system and purpose, is the possibility to isolate SNPs without a reference genome. To date, no reference genome exists for *Aedes mariae* and *Ae. zammitii*, therefore no reference genomic data can be used to align sequence reads and call SNPs. RAD-seq allows to analysed tags *de novo*. Identical reads can be indeed brought together and considered as candidate alleles. By aggregating these unique sequences, differing by a small number of mismatches, SNPs can be identified for a given locus. Errors in SNPs calling will be identified using the reads counts: real homozygous and heterozygous will have relative high reads count, while errors will have low reads counts. In this context, another advantage of RAD-seq, as mentioned above, is the possibility to use the paired-end sequencing approach. This technique, allowing the sequencing of each side of the restriction site, can produce extended contigs that can be used as reference to which all reads can be aligned and SNPs can be called across the whole fragment (Figure1). This is intuitively very advantageous when you have not a well-assembled reference genome, as is the case of *Ae. mariae* and *Ae. zammitii*. All the above characteristics make RAD-sequencing the best solution to my purpose, because it summarizes all the advantages of reduced-representation methods, but allowing to accurately analyse individuals from wild populations at a large number of markers, even when no prior sequence data are available, promising the assessment of hybridization and introgression in my study-species.

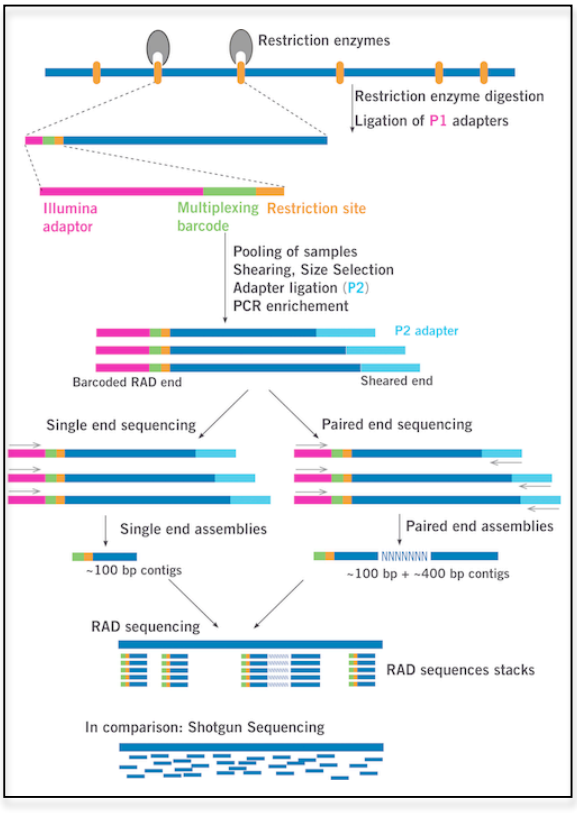

Figure 1. RAD-sequencing approach

Future collaboration with host institution and projected publications/articles resulting from the grant.

To date, I liaised with my host institution. After the first suggestions, they are helping me to resolve the basic trade-offs and delineate the best experimental design (i.e. what enzyme? How many tags do I need? How many samples should be multiplexed? How can I get the best depth of coverage?).

At least two publications in international peer-review journals would result from this grant:

- Description of SNPs discovery in *Aedes mariae* and *Ae. zammiti* mosquitoes
- Analysis of the introgression pattern between *Aedes mariae* and *Ae. zammiti* at nuclear genome across time and space in the artificial sympatric area.

All preliminary and published results will be further presented in international congresses.