ESF Research Networking Programme
Conservation Genomics: amalgamation of conservation genetics and ecological and evolutionary genomics (ConGenOmics)

Final report of the Short-visit Grant (Ref. Num.: SV/4985)

TITLE: Development of genomic markers for non-model species.

Applicant: Isa Pais, Instituto Gulbenkian de Ciência (IGC), Portugal

Host: Dr. Juan Montoya-Burgos, Université de Genève, Switzerland

I. Purpose of the visit

Since 2009, the Population and Conservation Genetics (PCG) Group led by Dr. Lounès Chikhi at the Instituto Gulbenkian de Ciência (IGC), has been studying the effects of habitat loss and fragmentation in Madagascar by looking at the patterns of genetic diversity of forest dwelling animals. During the last three years, biological samples have been collected from different lemur species with the main aim of performing comparative studies across regions and across species within the same regions. These studies aim to improve our understanding of how habitat fragmentation affects the different lemur populations, and if this is a consequence of anthropogenic activities in the area or a consequence of natural environmental changes that took place in the past.

In order to meet the aims proposed, we developed a strategy to identify new genomic markers (microsatellite and SNPs) specific of the species sampled (*Microcebus tavaratra* - mouse lemur, and *Lepilemur milanoii* – Daraina sportive lemur), as genetic markers are still not available for these species. High-throughput sequencing has already proven useful for the development of *de novo* molecular markers for non-model species (Davey *et al.*, 2011). Our group has recently applied such approaches together with shot-gun sequencing to the Bornean Elephant, a non-model and endangered species (Sharma *et al.*, 2012). While the project on the Bornean elephant was outsourced to a U.S. company (Floragenex, Inc.) and collaborator, the present aim of the PCG is to develop the skills required to identify genomic markers across the species studied.

RAD (Restriction-site Associated DNA) sequencing (RAD-seq or RAD-tag) seemed to be one of the best option to develop such new genomic markers as it requires little amounts of DNA, it provides the researchers with a large number of genomic markers in a relatively short time, and it allows to obtain genetic information by genotyping/sequencing from many individuals (Mardis, 2008). This seemed the appropriate technique to use as we have around 300 ear biopsies for several lemur species. Using RAD-sequencing with the appropriate restriction enzyme, we would expect to identify around 10,000 markers, out of which 1,000 should be polymorphic.

Currently, at the IGC, despite having all the material necessary for the library preparation, no experience or competence to develop a RAD-seq facility exists. As a lab technician of the project on Madagascar lemurs, I am responsible for most of the molecular work, and I was very much interested in learning about the library preparation techniques. Therefore, to develop an in-house protocol I first needed to gain theoretical knowledge and technical experience on RAD-seq library preparation, preferentially from a group performing such libraries in a routine-basis. This was the case of the lab of Dr. Juan Montoya-Burgos in Geneva.

II. Description of the work carried out during the visit

Previous to the practical steps of the protocol, *in silico* analyses were carried out in order to define the restriction enzyme that would provide us with the most favourable results in terms of the number of possible polymorphic sites discovered. As the genomes of *L. milanoii* and *M. tavaratra* species are not yet available, we had to use a sister species of the latter (*M. murinus*) to perform the *in silico* analyses. It turned out that when using the *NotI* enzyme, around 10,000 cutting sites were found, a number suggesting that at least 1,000 polymorphic sites should be obtained as a result.

Initially, the idea was to RAD-tag around 55 individuals from one of the two lemur species and to sequence them using the Illumina platform (Illumina HighSeq 2000) to obtain polymorphic markers. However, due to a restricted budget and the need to increase the per-individual sequencing depth, we decided that it would be more efficient to reduce the number of barcoded adaptors per sequencing lane, and to increase the number of lanes to be

sequenced. The strategy adopted during the short-visit consisted in barcoding 15 P1 adaptors (these are the tags that will identify the individuals during the bioinformatics analysis) and use it on individuals from the same lemur species distributed across 9 forest fragments from the Daraina region in northern Madagascar. We would then send three lanes to run in the sequencing platform, each containing 15 different individuals. This way we should obtain genetic information on 45 individuals for a much lower cost than in the initial plan and with a much deeper sequencing coverage.

This strategy was also designed to increase the probability of obtaining results faster and within the duration of my stay in Juan Montoya-Burgos' lab. For instance, by dividing the work in smaller groups of 15 individuals, we increased the chance of finishing one batch on time before sending it to the Illumina platform to be sequenced. However, due to technical problems we encountered during the library preparation process, I was only able to prepare one lane of 15 barcoded individuals in Switzerland, having to continue the rest of the work back in Portugal.

In order to apply the protocol for RAD-tag library preparation to our lemur species, we first needed to optimize it accounting for both the species and restriction enzyme chosen. Our protocol was based on a protocol originally described by Etter (Baird *et al.*, 2008) and used on stickleback, applying the *EcoR*I restriction enzyme. During the first day of my stay, we spent most of the time performing new calculations in order to determine the correct quantities of reagents to be used, such as the right amount of P1 adaptor that we should use according to the amount of cohesive-ends present in each digestion product.

As a starting point, the adaptors were ligated in order to have a double-stranded P1 barcoded adaptor per individual (15 different P1 adaptors) and one double-stranded P2 adaptor. All adaptors as well as all solutions/reagents were prepared to the right working concentrations.

The optimized workflow for each library preparation is as follows:

1. DNA extraction

This step was carried out in Portugal in order to save time during the short visit to the Swiss lab, where the aim was to prepare RAD-tag libraries for 45 *M. tavaratra* individuals.

2. ***Restriction endonuclease digestion***

Each *M. tavaratra* DNA was digested separately for 4hr at 37ºC, with 40U of *NotI* restriction enzyme. Despite several tests being performed in Portugal, significant work was still necessary to carry out in Switzerland, as digestion did not seem to provide the expected results.

3. ***P1 Adapter ligation***

Each individual sample was labelled with a unique barcoded P1 adapter in order to later differentiate DNA sequences and perform the post-sequencing analysis. On the first trial, 15 *M. tavaratra* individuals from 9 different forest fragments were labelled with a different P1 adaptor each.

4. ***Sample multiplexing***

In order to cut down the costs and time spent on the next steps of the protocol, we prepared a unique tube containing an equal amount of barcoded DNA from each individual. The fact that we combine and process all 15 samples as one also minimizes the differences in amplification efficiency that may arise between different library preparations when processing many at once.

5. ***DNA shearing***

At this point, the high molecular weight DNA fragments were broken down to an average size of 300-500 bp in order to create a library of P1 barcoded molecules with random variable ends, which will then be ligated to a second adaptor (P2), for future amplification.

6. ***Size selection/agarose gel extraction***

The sample containing all 15 individually barcoded DNA, was loaded on an agarose gel and the band between 300-500 bp was removed. The reason behind obtaining fragments of such a specific size is simply because this is the best size range of tags to be sequenced efficiently on an Illumina Genome Analyzer flow cell. This step also helps to remove any free un-ligated or concatemerized P1 adapters that may be present at an average size of 130bp.

7. ***Perform end repair***

Shearing the DNA may have caused the DNA ends of the fragments to have 5´ or 3´ overhangs which could interfere in the next steps. In order to convert these into phosphorylated blunt ends, we used the Quick Blunting Kit from NEB labs, which uses T4 DNA Polymerase and T4 Polynucleotide Kinase to perform the repairing of the DNA ends.

8. ***3´-dA overhang addition***

With the blunt-ends from the previous step being phosphorylated, it was then possible to add an 'A' base to each of the 3' ends. To do it, we used the dATP (Promega) and the Klenow Fragment (3´ to 5´ exo$^-$) (NEB labs). This step is essential for the ligation of the P2 adapter to every single fragment end, and is only possible due to the single 'T' base overhang at the 3´ end of its bottom strand.

9. P2 Adapter ligation

By incubating the sample with dATP and T4 DNA ligase, the P2 adapter that contains a 3´-dT overhang will be ligated to both ends of each DNA fragment. The divergent ends of the "Y" shaped P2 adapter will, in the following step, promote amplification of fragments that contain both P1 and P2 adapters. This will reduce the amount of fragments that will not bind to the flow cell (fragments ligated simply to P2 adapters), and increase the chances of producing a robust library.

10. RAD tag Amplification/Enrichment

In this final step, high-fidelity PCR amplification will take place using both P1 forward and P2 reverse adapters. This procedure will enrich for RAD tags containing both a P1 and P2 adapters, preparing them to be hybridized to an Illumina Genome Analyzer flow cell. Once we have the final product of the RAD library, the sample will be sent to the sequencing platform located at the University of Geneva.

III. Main results obtained

I should note that during my stay I did not have the time to finish all the RAD libraries initially proposed because I encountered several technical problems which slowed down the process. However, this period allowed me to learn all the techniques required for this type of library preparation, and I am currently finishing the libraries in Lisbon. In other words, the aim to transfer the skills to the IGC has been achieved to a great extent.

Throughout the protocol, several tests were performed in order to determine whether the experiment was providing us with the expected results or not. These tests allowed us to detect the errors in the optimized protocol for our species, and also helped us to understand the source of the errors and how we could solve it.

We met a major problem right at the beginning of my stay, relating to the digestion of the DNA. After several tests, we concluded that the TE buffer in which the DNA was being eluted was contained a high concentration of EDTA, which inhibited the restriction enzyme. As a result, we had to reduce the quantity of DNA digested to half of the amount recommended in Etter's (Baird *et al.*, 2008) protocol to ensure complete digestion and at the same time have enough product to perform the following steps.

Secondly, on the first attempt to produce a P1 library for 15 individuals, we realised that when reaching the last step of RAD-tag amplification, no DNA was present in the PCR product,

which meant that at some point we had lost most or all DNA. In order to find out the crucial step where DNA was lost, we prepared a second set of P1 barcoded multiplexed samples and determined that the purification columns that we were being using to clean the DNA at steps 7 and 8, had a default and were not yielding the expected amount of DNA. We then decided to replace the old columns by new ones that were known to work perfectly, and on the third attempt to prepare the RAD-tag library, we were successful to obtain a PCR product which has probably amplified the desired fragments.

In conclusion, after the short-visit I managed to get good results up to step 10 of the protocol, where I performed a PCR with small quantities of the library product just as a confirmation that the procedure was successful. I will now perform further RAD-tag amplification using all the material prepared in Switzerland, quantify the DNA post-amplification, clone and sequence it using the standard method. If the preliminary sequencing results prove that what amplified is actually lemur DNA and not the P1 adapter, we will send the final library to the Illumina platform in order to perform high-throughput sequencing. While we have achieved slightly less than we expected in terms of final work done during the grant period, I have actually learnt a great amount of new skills and I am confident to perform all steps and adapt them to the IGC lab in Lisbon.

4. Future collaboration with the host institution (if applicable)

This stay is part of collaboration on the development of markers for endangered species. The sequencing and part of the computational work will still be performed in collaboration with the host institution. After that it is difficult to determine whether further collaboration is expected in the near future. But we are optimistic (funding pending).

5. Projected publications / articles resulting or to result from the grant

The work performed during the stay will lead to at least one publication in collaboration with the host institution. This will be done in collaboration with a PhD student who will use the results obtained from the work carried out while in Geneva, and will learn the new skills on library preparation from me.

6. Other comments (if any)

As a lab technician who came from an ecological background I have often learnt on my own or with colleagues during the last couple of years. Going to the host Institution gave me the opportunity to work closely with a lab manager (Ilham Bahechar) who was very skilled and had a great experience. It was a excellent opportunity to change lab, see how different labs can be organised and to learn new technical and organizational skills.

References

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, *et al.* (2008) *Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers*. PLoS ONE 3(10): e3376. doi:10.1371/journal.pone.0003376

Davey JW, A Hohenlohe P, D Etter P, Q Boone J, M Catchen J, *et al.* (2011) *Genome-wide genetic marker discovery and genotyping using next-generation sequencing*. Nature Review Genetics 12: 499

Mardis ER (2008) *Next-Generation DNA Sequencing Methods*. Annu. Rev. Genomics Hum. Genet. 9:387–402

Sharma R, Goossens B, Kun-Rodrigues C, Teixeira T, Othman N, *et al.* (2012) *Two Different High Throughput Sequencing Approaches Identify Thousands of De Novo Genomic Markers for the Genetically Depleted Bornean Elephant*. PLoS ONE 7(11): e49533. doi:10.1371/journal.pone.0049533