|  |  |
| --- | --- |
| **esf_green_logo** |  **Research Networking Programmes**  |
|  |
|  |

**Short Visit Grant** [ ]  **or** **Exchange Visit Grant** [x]

***(please tick the relevant box)***

**Scientific Report**

**Scientific report (one single document in WORD or PDF file) should be submitted online within one month of the event. It should not exceed eight A4 pages.**

***Proposal Title****:* Population genomics of parallel adaptation in endemic and endangered Neotropical crater lake cihclid fishes

 ***Application Reference N°:*** 4458

1. **Purpose of the visit**

My stay at the Natural History Museum in Madrid aimed to support the research group of Dr. Marta Barluenga with my expertise on reference-free SNP (single-nucleotide polymorphism) data generation. During the last year the group generated genomic resources from incipient species in several endemic parallel radiations in Nicaraguan crater lakes through RAD-sequencing (restriction-site associated DNA sequencing). Seven RAD libraries were already prepared and sequenced with Illumina technology including trios of phenotypically similar ‘species’ from radiations of the Midas Cichlid species complex from six lakes. About 20 individuals per ‘species’ (294 in total) were included in this genome scan. The aim of the following computational analyses is to identify species-specific SNPs differentiating the incipient species and to construct a phylogeny from those SNPs. Secondly, the identification of phenotypic-specific SNPs that parallel species might have in common and which might be the basis of parallel evolution.

1. **Description of the work carried out during the visit**

During the two month of my stay in Madrid I mainly explored the data and adjusted my R scripts to the data. The steps for computational analyses were as follows: The first step is the demultiplexing and quality filtering of pooled samples. Before the quality-filtered and demultiplexed reads can be individually aligned, a pseudo-reference was constructed. After aligning sample for sample to that reference the genotypes for the two dominant alleles were called per individual. The last step before constructing a phylogeny and calculating fixation indexes (FST) the SNP calling across homologous loci over all individuals is performed.

DEMULTIPLEXING: First the reads were sorted according to their tagged barcodes and quality filtered. Each of the seven libraries contained 42 individuals, each individual tagged with a specific barcode. Each true RAD-tag sequence read starts with the 5-digit barcode followed by the 6 bp recognition site, TGCAGG, of the SbfI restriction enzyme that was used to prepare the RAD-libraries. Demultiplexing was conducted using the program process\_radtags which is contained in the Stacks software using the following settings: (-E) phred33 for the encoding of quality scores, (-e) SbfI for specifying the used restriction enzyme, (-i) fastq for the input format, (-r) rescues barcodes and RAD-Tags, (-c) removes all reads containing uncalled bases, and (-q) removes low quality score reads. The program corrects uncalled bases in the barcode as well as in the restriction site. If the quality score drops below a phred score of 10 within a sliding window (15 % of read length) the read is discarded.

BUILDING A PSEUDO-REFERENCE: For the diverse species of MIDAS cichlids no reference genome is available. To search homologous RAD-tags for SNPs, a pseudo-reference from the individual with the highest read number after demultiplexing and cleaning was used and assembled into stacks (reads which align perfectly). This was done using the program ustacks from the Stacks software package. ustacks also assembles stacks with a maximum number of allowed mismatches into loci and calls single-nucleotide polymorphisms (SNPs) within these loci. The program was run with the following settings: (-m) 4 for the minimal coverage of a single stack, (-H) disabled the use of reads which did not align in the first run, and (-r) for enabling an “Removal Algorithm” of highly repetitive reads as microsatellites. A fourth parameter (-M) that sets the maximum of mismatches when merging stacks was tested for two different settings: a stringent (-M 1) and a relaxed setting (-M 4). That was done in order to test impact of parameters on the final pseudo-reference and subsequent analyses.

For construction of the reference I used the produced output of the perfectly matching stacks called tags.tsv and a custom perl script to concatenate the consensus sequences into a single fasta file. The pseudo-reference for -M 1 (hereafter called M1 reference) had a total length of 7.07 Mb and for -M 4 (hereafter called M4 reference) 6.09 Mb. That the M1 reference is longer than the M4 is intuitive as stacks differing in two and more substitutions constitute a new stack, despite being possibly homologous.

ALIGNING READS TO THE PSEUDO-REFERENCE: The individual samples were aligned against the M1 and M4 pseudo-reference using Novoalign v2.08.03 (http://www.novocraft.com/). For that, the references were indexed using a step size (-s) of one with a calculated “genome size” of 7,072,877 bp and 6,090,618 bp in total (length of the pseudo-references M1 and M4), respectively. That implies a k-mer length (-k) of nine to be used for indexing. The actual alignment was run with options (-F) STDQF for quality scores of fastq files being standard Sanger coded, (-t) 180 allows for six high-quality mismatches, (-x) -15 sets the gap opening penalty, (-e) 30 stops the algorithm and alignment search for a read after it reached 30 best alignments, (-l) 30 defines the minimal number of good quality bases per read, (-o) SAM defines the output format and (-r) None excludes all alignments that map to more than one location. The resulting alignments for all 294 individuals were coded in sam files and were converted to bam using Samtools v0.1.18 (Li et al., 2009).

GENOTYPE CALLING PER INDIVIDUAL: To call genotypes at each RAD-tag nucleotide within a single specimen, an R-script was employed on the data set. The script makes use of the R package Rsamtools (Morgan & Herve Pages) for the import of bam formatted sequences into R. The input files contained multiple columns giving e.g. the position of an alignment in the reference sequence, the identifier of the read that matched the reference sequence and the mapping quality of the alignment. For each individual, the total read number was determined by counting all homologous reads of each RAD locus. All loci with a coverage greater than threefold the mean coverage were excluded as well as loci with a proportion of the first two dominant haplotypes relative to the total coverage of the locus being smaller than 0.7 as those are reads from repeat regions.

Depending on coverage, genotypes were called in the following 6 types: 2 monomorphic (Nx, L) and 4 polymorphic genotypes (Y, U, Nxy, P). Thereby, only the two dominant variants were considered to call bi-allelic polymorphisms and coded as alleles X and Y. Counts for a genotype had to exceed a minimal threshold (min) to be considered at all (threshold = 4). A reliable diploid genotype was called when exceeding the coverage of 15 (lim). Those thresholds were determined according to the frequency distribution of the coverage over the 2nd called variant.

The output in text format contains locus ID referring to the pseudo-reference, coverage, allele (X and/or Y), genotype call (N, L, Y, U or P) and the corresponding read sequence.

STATISTICS: The seven RAD libraries yielded in total 1.4 billion reads of which 0.86 billion passed the quality control. On average, that makes 2.93 mio reads per individual with a mean coverage of 73 per individual and RAD-tag. The pseudo-references M1 and M4 contained 73,714 and 63,477 unique loci, respectively. Overall, all results meet the expectations from earlier projects on African cichlids and European lampreys.

TROUBLE-SHOOTING: First of all did the alignments of reads against the M4 reference perform better than against the M1 reference. That means that not all individuals could be mapped against the M1 reference. Reasons for that could be the rather stringent setting within the novoalign command of -l 30. To gain a successful alignment at least 30 reads have to align to that region. From an expected mean coverage of 70 per locus that doesn't constitute a problem. But this threshold can hardly be reached when the reference was constructed in a way that homologous loci are artificially split in two or more loci by the very stringent parameter -m 1 that allows only 1 mismatch per stack. Subsequent analyses will be based on the results from the M4 reference.

The last step that has to be carried out before the FST analyses and phylogeny construction start is the calling of the actual SNPs from loci that are present in all individuals. That step is currently running and therefore no results are available.

1. **Description of the main results obtained**

 The quality of all produced libraries is good and will provide high quality SNPs for further analyses. Those analyses will be based on the M4 reference. In general, computational analyses are limited by computer resources and the capacity of such. Therefore, the work with the data is not finished but I'm constantly working on it and the next step, the SNP calling is almost finished and the first phylogeny can be constructed. That will then finally give us the first impression of our results.

1. **Future collaboration with host institution (if applicable)**

 During my stay I could interact with members of the institiute. Further collaborations are planned with Dr. Marta Barluenga exceeding the here stated aims of my research exchange. That future collaboration includes the transcriptome assembly of three species each from the different Nicaraguan crater lakes.

Further, as the analyses could not be finished within this two month stay, I will continue my work on the here described project that should finally result in a RAD-marker based phylogeny and biogeographic resolution of the here anaylsed species, a coalescent analyses that will be performed with a collaborator from the University of Sheffield, and the identification of candidate genes that might underlie that phenotypic resemblance in similar ecological niches but different phylogentic history.

1. **Projected publications / articles resulting or to result from the grant *(ESF must be acknowledged in publications resulting from the grantee’s work in relation with the grant)***

Despite no ready to publish results the data looks promising and will be similary prepared as the journal article by Mateus, Stange et al. (2013) Strong genome-wide divergence between sympatric European river and brook lampreys. Current biology : CB, 23(15), R649–R650.

1. **Other comments (if any)**