Research Networking Programmes

Short Visit Grant ☐ or Exchange Visit Grant ☒

(please tick the relevant box)

Scientific Report

The scientific report (WORD or PDF file – maximum of eight A4 pages) should be submitted online within one month of the event. It will be published on the ESF website.

***Proposal Title**:* Genotyping by sequencing: new tool for population genomics of native and aquacultured Mediterranean mussel, Mytilus galloprovincialis

Application Reference N°: 4598

1) Purpose of the visit

Main purpose of my visit to Patrik Nosil lab at University of Sheffield (United Kingdom), was to implement genotyping by sequencing approach as suitable and versatile NGS tool into population genomics of aquacultured Mediterranean mussel Mytilus galloprovincialis, which hereafter becomes new species in the field of livestock/aquaculture genomics.

2) Description of the work carried out during the visit

During my two weeks visit (February 9th to February 23th) the Patrik Nosil lab on the University of Sheffield, we have used 200 milion reads generated by RAD-seq (Restriction site associated DNA sequencing) methodology on Illumina HiSeq sequencing platform to create pseudoreference Mytilus galloprovincialis consensus genome sequence, we have mapped barcoded RADseq Illumina sequences of native and aquacultured mussel individuals collected at the aquaculture sites, called SNPs and performed conversion of genotype likelihoods to genotype probabilities.
0. Preparation of samples (previously conducted): DNA isolation and RAD tag sequencing library preparation
DNA isolation protocols were optimized in order to obtain DNA of the highest quality. Due to the high susceptibility of mussel DNA toward degradation, various extraction kits and protocols were optimized, using different mussel tissues and tissue conservation methods. On the end, DNA extracted from the muscle tissue preserved in 96% EtOH

using Sigma GenElute extraction kit, with homogenization step performed in liquid nitrogen and omitting vortexing was used.
Library preparation: In short, genomic DNA was digested with restriction enzymes (EcoRI and MseI). Custom made Illumina barcodes and adaptor sequences were ligated to the digested fragments. Only EcoRI restriction site was barcoded. These fragments were amplified by PCR using custom made Illumina primers. Amplified fragments ware polled into 30 uL of Illumina sequencing library and sent for sequencing. Sequencing was performed at NCGR (National Centre for Genomic Resources, USA) were fragments were size selected to the 300-500 bp, and sequenced in one lane of the Illumina NGS platform (HiSeq 2000 with V3 reagents).
Before starting the actual sequencing data analyses I received extensive few days bioinformatics training enabling me to work more efficiently in Unix environment and especially to successfully handle and process large amount of NGS data.
1. Quality control of raw data
Sequences were downloaded and checked for the raw data quality using FastQC tools. In total app. 200 million reads were generated out of one Illumina lane, and average quality of sequences was high.
2. Demultiplexing the data
Adaptor sequences and protector base were trimmed using custom written Perl scripts. Sequences that were too short, sequences without the barcodes or with irregular barcodes were discarded (3.1% sequences). 205 653 103 sequences were analyzed, 199 212 204 sequences contained barcodes and 6 440 865 sequences didn't contain barcodes.
104 762 sequences had a 3' MseI adapter sequence and 34 sequences were too short after removing the 3' adapter (<10 + 6 bp). 199 212 204 sequences were retained for further analyses.
3. Quality control
FastQC was run on the trimmed data. Average quality of reads was high (38), as was the overall quality, and GC content was 36%.
4. Splitting the reads
Custom made Perl scripts were used to identify and remove individual identifier (barcode) sequences from fastq files. All the sequences were assigned to the individuals and splited into separate bz2 files for each individual. Average number of reads per was 703930 bp per individual.
5. De novo assembly
Rainbow tools (Chang et al. 2012, Bioinformatics) were used to create pseudoreference Mytilus genome using Illumina RAD sequencing reads. Scripts were modified for one end sequencing protocols. We have arbitrary chose first 10 million reads and these were assembled into the contigs, yielding 459178 Mytilus contigs of length between 84-86 bp. Minimum overlap of sequences in contigs was 45 bp, and maximum 4 mismatches were allowed. Then, we made consensus seuqence by concating these contigs and used it to create first pseudoreference Mytilus genome by inclusion of arrays containing 30 N in-between two contigs. Total length of the pseudoreference was 52 805 760 bp. Afterwards, a second de novo assembly was performed using all 200 mil Mytilus reads generating pseudoreference genome of total length 472 757 920 bp.
6. Mapping the reads to the pseudoreference
Bowtie2 was used in combination with samtools to map all fastq sequence files to the pseudoreference Mytilus consensus genome assembled from 10 million reads, and sorted all the sequences in bam.bai files. Alignment sequences rate was around 95% for most of the individuals. This resulted in 199 212 204 sequences mapped and assembled onto the

Mytilus pseudoreference genome, with the average coverage depth of 438x per genomic region across all individuals.
7. SNPs calling
Samtools mpileup and bcftools incorporated in custom made Perl scripts were used to identify variable sites, and call single nucleotide polymorphisms (SNPs). We performed basic filtering, excluded all the insertions and deletion, all SNPs with the depth exceeding 10000, with the quality score below 20, and all the SNPs that were scored in less of 40% individuals. We have retained only biallelic SNPs. Final set of 180439 out of 343174 variable sites (SNPs) were retained after filtering.
8. Caculating genotype likelihoods and genotype probabilities
Custom made Perl scripts were used to convert phred-scaled genotype likelihoods to genotype probabilities using prior probability of each genotype for a given population allele frequencies.

3) Description of the main results obtained

The main purpose of this scientific exchange is fulfilled, as customized genomic tools have been developed for Mediterranean mussel, Mytilus galloprovincialis. The implementation of RAD-seq methodology and computational analytical tools in the populations genomics of aquacultured Mytilus galloprovincialis was really successful, yielding high quality genotype data and adding completely new species in the livestock genomics field.Although I am still processing data and did not reach the final step of having all population genomic results in hand, my two weeks visit to Patrik Nosil lab at the University of Sheffield already resulted in producing high quality genome wide genotyping data for Mytilus galloprovincialis. Currently, I am applying all the methods learned during my two weeks visit to address the questions of genomic differentiation between aquacultured and native mussels. Except introducing new species in the livestock genomic arena, to the best of our knowledge this is first application of GBS approach in the livestock genomics. We have shown that this methodology can be tailored to non model species to generate huge amount of genome wide SNP data at just small fraction of cost of the whole genome sequencing. In terms of genotyping by sequencing data analyses and especially in terms of acquiring new knowledge this visit to Patrik Nosil lab exceeded all my expectations.

4) Future collaboration with host institution (if applicable)

This collaboration is still ongoing, while research on aquacultured vs native mussels is active in terms of population differentiation data analyses.
Therefore, I am still learning many new methods in population genomics analyses within the frames of this collaboration.
Currently, I am also applying for the position of postdoctoral researcher on the UKF funded project led by dr. Nosil.

5) 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)*

Due to the high quality genoting data obtained, this collaboration is on good way to result in high quality scientific publication on the genomic differentiation between aquacultured and native population of Mediterranean mussel, Mytilus galloprovincialis. Naturally, ESF will be acknowledged for this travel grant in the resulting publication.

6) Other comments (if any)

Due to the availability of the host laboratory, I had to postpone the trip for 9 days form the dates proposed in the project application, and final dates of my visit was Feb 9th till Feb 23th.

I am truly greatful both to the ESF grant and host laboratory of dr. Patrik Nosil for the opportunity to gain valuable new analytical skills, and overall large amount of new knowlegde in the fast developing and exciting research field of population genomics.

My travel and life expences that exceeded the amount limits of the ESF travel grant specific categories were covered from the UKF funded project led by dr. Patrik Nosil.