PHD COURSE IN LIFE AND ENVIRONMENTAL SCIENCES

Report Form for PhD student annual evaluation (XXXVII and XXXVIII cycles)

Name of PhD student: Francesco Giannelli
Title of PhD research: Genomics of Lessepsian invaders

Name of PhD supervisor: Emiliano Trucchi, Emanuela Fanelli
Research lab name: Genomics lab

Cycle:
[ ] XXXVI
[X] XXXVII

PhD Curriculum::
[X] Marine biology and ecology
[ ] Biomolecular Sciences
[ ] Civil and environmental protection

DISVA instrumentation labs/infrastructure eventually involved in the project:
[ ] Actea Mobile Laboratory
[ ] Advanced Instrumentation lab
[ ] Aquarium
[ ] MassSpec lab
[ ] MaSBiC
[X] Simulation/informatics lab
[ ] Other. Please, indicate: ....

ABSTRACT (1000 characters, including spaces):

Since the opening of the Suez canal an increasing number of so-called Lessepsian invaders are entering the Mediterranean. Many of them are successfully colonising this basin, negatively interfering with native species. This project aims to investigate the dynamics of different types of genetic variability of marine invaders during the colonisation of new habitats, categorising the variability based on its evolutionary impact (i.e., adaptive, deleterious and neutral). Using two successful Lessepsian fish invaders, *Pterois miles* and *Siganus rivulatus* as models, we are going to investigate the genetic causes and consequences characterising their invasive processes. The results obtained so far on *P. miles* suggest that the colonisation of each new environment along the invasion route is due to a very small number of individuals (i.e., serial bottlenecks) that have been successful in surviving and reproducing in the new environmental conditions.

Part 1. Scientific case of the PhD Research (2 to 3 pages, including figures)

- BACKGROUND

In the last decades the number of invasive species has constantly increased worldwide [1] resulting in one of the most important threats for natural ecosystems and endangered species. The Mediterranean Sea is already well known for being one of the most impacted areas in the world by the direct and indirect consequences of human activities [2] with the increasing number of invasive species recorded in this sea being a major concern [2].

Many of these alien species are entering the Mediterranean from the Red Sea after the opening of the Suez canal in 1869 [3]. A concerning amount of these so-called Lessepsian invaders are successfully colonising and spreading all over the basin negatively interfering with native species [3]. This continuous spread is also favoured by the consequences of climate change that is resulting in increased sea temperatures benefiting Lessepsian species, that are well adapted to warmer environments, over native species [2].

Lessepsian invaders have been extensively studied under an ecological point of view but just a few genomics studies have been carried out so far and none of them focused on both neutral and non neutral genetic variability [4, 5, 6]. This means that the ongoing adaptation processes and the dynamics of accumulation of deleterious mutations during the invasive process to date have been largely ignored, despite the fact that these dynamics are considered fundamental to understand the basis of the invasive processes and build reliable predicting models [3].

- SCIENTIFIC AIMS

Using two Lessepsian fish species, *Pterois miles* and *Siganus rivulatus*, chosen for their different ecological and economic impacts and for their distinct trophic guilds, we are going to investigate the trajectories of different types of genetic variability (adaptive, deleterious and neutral), during the colonisation of new habitats, in order to answer the following questions:

- How did the invasion of the Mediterranean by Lessepsian species occur? And what are the dynamics of population expansion in the invasive range?
- Are there ongoing selective processes acting in the invasive range?
- Are some of the genetic variants under selection in the new environment present in the source population as possible preadaptation?

- WORKPLAN AND RESEARCH ACTIVITIES

WP 1. Data production.

Sampling and DNA extraction

During the first year of the project our aim was to obtain tissue samples (muscle or liver) of 10 individuals of both *P. miles* and *S. rivulatus* from every selected sampling location: Red Sea, Cyprus, Crete and the Ionian islands (figure 1) managing to get most of the samples we had set ourselves (table 1). The resulting tissue samples have been processed and DNA has been extracted from them. More than 80 *P .miles* and *S. rivulatus* from the native range (Red Sea) and the invasive range (Cyprus, Crete and Ionian islands) have been processed and the best samples (50 samples for now) have been sent to be sequenced in order to produce whole genome resequencing data using a short read sequencing method (DNBseq^{TM} technologies). We obtained short reads of 150 bp and an expected coverage of ~30X for both *P. miles* and *S. rivulatus*.

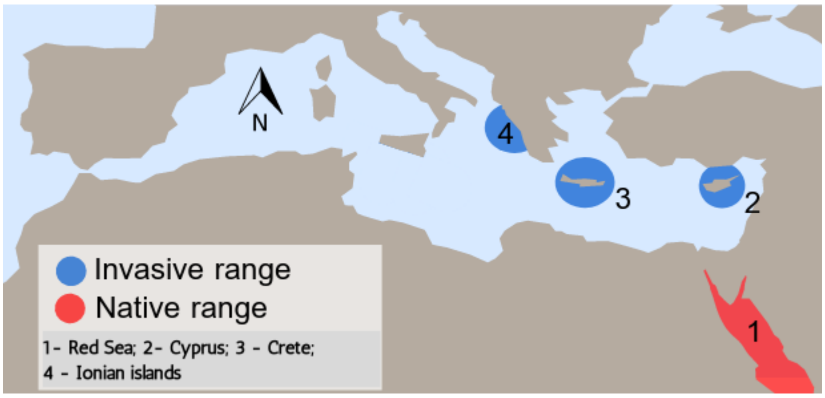

Figure 1. Sampling localities

	Pterois miles	*Siganus rivulatus*
Red Sea	3	10
Cyprus	10	10
Crete	10	10
Ionian islands	7	5

Table 1. Number of samples for both species in every sampling location

Reference genomes

For the analyses we want to carry out, a reference genome for both species is necessary. The reference genome of *P. miles* has already been sequenced and made available online on ENA [7]. While for the *S. rivulatus* no reference genome was available so we had to produce our own reference genome. For that purpose we selected one tissue sample that we collected in Cyprus in 2022 and we sent it to a sequencing service (BGI genomics) that performed the DNA extraction and long-read sequencing using PacBio Revio technology.

Both the whole genome resequencing and long-read sequencing performed excellently, giving us data of great quality for the subsequent genomic analyses and genome assembly (for *S. rivulatus).*

WP 2. Data analysis.

Genome assembly of *S. rivulatus*

The long-read sequences obtained from the PacBio Revio technology on *Siganus rivulatus* were used to produce the first reference genome for this species. We used Hifiasm [8] to perform a de novo genome assembly of our reads, obtaining from the very first steps a highly-contiguous assembly. Following that, we moved on to eliminate artificially duplicated regions using Purge Dups [9]. After, we proceeded with the scaffolding step, using LRScaf [10], in which the contigs forming our assembly were merged together in order to create longer sequences (scaffolds). Finally, we utilised TGS-GapCloser [11] to fill in the gaps within our genome, generating sequences for regions previously lacking information.

Every step was followed by three quality control analyses using BUSCO [12], merqury [13] and FASTA-tools [14] that confirmed the overall excellent quality of the genome assembly (Table 2).

Number of contigs:	41
Total size (bp):	518290999
N50 (bp):	22713224
L50:	11
N90 (bp):	17311268
L90:	21
Mean contig size (bp):	12641243.8780488
Longest contig (bp):	26743707
Third quartile (bp):	22937535
Median (bp):	17311268
First quartile (bp):	550782.5
Shortest contig (bp):	32809
Number of Ns:	0
Number of gaps (/N+/):	0
Number of other IUPACs:	0

Table 2. Results of the FASTA-tools run on the complete reference genome of *S. rivulatus*

Whole genome resequencing of *P. miles*

The whole genome short-read sequences obtained for *P. miles* underwent a first quality control, using FastQC (read length and bp quality scores) [15] and then trimmed using Trimmomatic [16] eliminating low quality regions. A second quality control, using again FastQC, has been performed afterward the trimming step that revealed a very high quality of the data produced. The sequencing reads were aligned to the reference genome using bwa-mem [17]. The resulting mapped reads underwent various processing steps for rearrangement using samtools [18] to prepare them for variant calling. Finally, Freebayes [19] was utilised for the variant calling process, resulting in an vcf file that contains a total of 26,507,408 variant sites.

With the *S. rivulatus* genome assembly in hand, we're replicating the same analyses on its sequences.

PCA

Using the newly produced vcf files of *P. miles* individuals we performed two preliminary analyses: PCA analysis [20] and an admixture analysis [21]. The results we got from PCA were totally in line with our expectations. It is clear from figure 2A that the whole Mediterranean has been colonised by the descendents of just a few individuals from the Red Sea (RS). Figure 2B shows the results we obtained from the Mediterranean individuals alone and it shows how the genetic diversity of the populations is inversely proportional to the distance from Suez. A possible interpretation for this pattern is that the colonisation of Crete (CR) was accomplished by the descendants of only a few individuals from Cyprus (CY), and similarly, the Ionian islands (KF) were colonised by the descendants of only a few individuals from Crete.

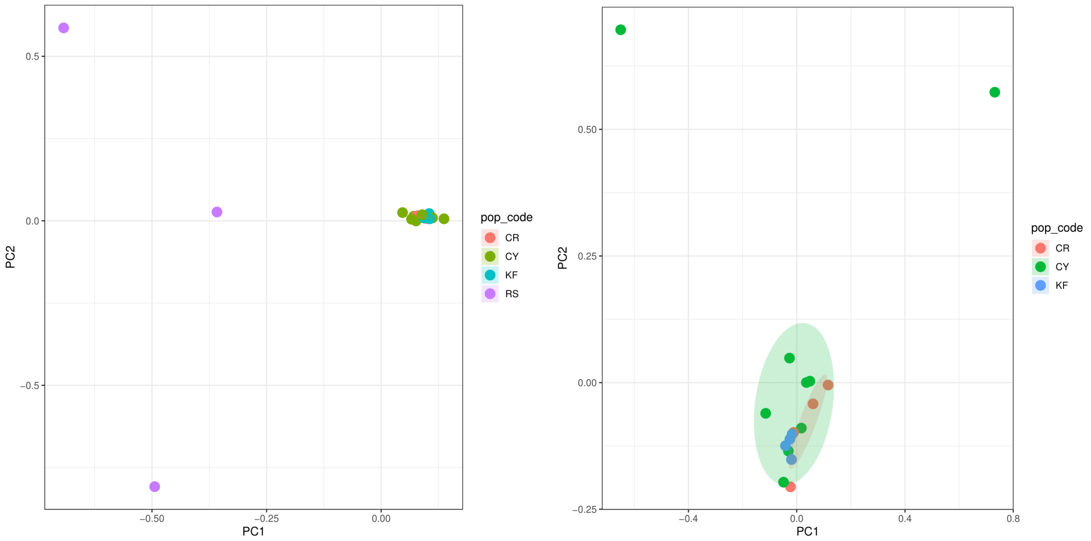

Figure 2. PCA plot showing all the individuals from all the four different populations (A), including the Red Sea (purple dots) and with the individuals from the Mediterranean alone (B)

Admixture

The admixture results (for a two-groups sample: K2) confirms the clear separation between the Red Sea and Mediterranean populations. The three individuals from Cyprus that present an important genetic component from the Red Sea must be investigated for more accurate analyses in order to properly interpret these preliminary results.

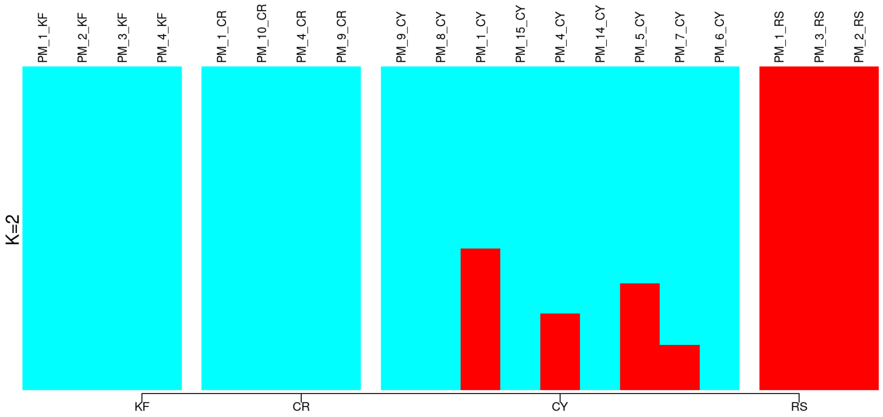

Figure 3. Admixture results for K2

Multiple genome alignment

Simultaneously with the above analyses we conducted a comprehensive whole-genome alignment using Cactus [22], aligning the *P. miles* genome with an existing multiple fish genome alignment [23]. The objective was to identify highly conserved genome regions via PhiloP [24], aiming to pinpoint mutations with a higher probability of being non-neutral. These conserved regions will be instrumental in discerning genetic variations. Their combination with RNAseq data will help determine genes with heightened expression, suggesting a significant role in the species' physiology.

Similar analyses will be replicated on the *S. rivulatus* genome, extending our comparative insights into both species.

- REFERENCES

1. Seebens, H., Blackburn, T. M., Dyer, E. E., Genovesi, P., Hulme, P. E., Jeschke, J. M., ... & Essl, F. (2017). No saturation in the accumulation of alien species worldwide. Nature communications, 8(1), 14435.
2. Templado, J. (2014). Future trends of Mediterranean biodiversity. In The Mediterranean Sea (pp. 479-498). Springer, Dordrecht.
3. Sherman, C. D. H., Lotterhos, K. E., Richardson, M. F., Tepolt, C. K., Rollins, L. A., Palumbi, S. R., & Miller, A. D. (2016). What are we missing about marine invasions? Filling in the gaps with evolutionary genomics. Marine Biology, 163(10), 1-24.
4. Chiesa, S., Azzurro, E., & Bernardi, G. (2019). The genetics and genomics of marine fish invasions: a global review. Reviews in Fish Biology and Fisheries, 29(4), 837-859.
5. Azzurro, E., Nourigat, M., Cohn, F., Souissi, J. B., & Bernardi, G. (2021). Right Out of The Gate: The Genomics of Lessepsian Invaders in The Vicinity of The Suez Canal. 10.21203/rs.3.rs-632020/v1.
6. Bernardi, G., Azzurro, E., Golani, D., & Miller, M. R. (2016). Genomic signatures of rapid adaptive evolution in the bluespotted cornetfish, a Mediterranean Lessepsian invader. Molecular ecology, 25(14), 3384-3396.
7. Christos V. Kitsoulis, Vasileios Papadogiannis, Jon B. Kristoffersen, Elisavet Kaitetzidou, Aspasia Sterioti, Costas S. Tsigenopoulos, & Tereza Manousaki. (2022). A high-quality reference genome assembly for the devil firefish, Pterois miles. Zenodo. https://doi.org/10.5281/zenodo.6380502
8. Cheng, H., Jarvis, E. D., Fedrigo, O., Koepfli, K. P., Urban, L., Gemmell, N. J., & Li, H. (2022). Haplotype-resolved assembly of diploid genomes without parental data. Nature Biotechnology, 40(9), 1332-1335.
9. Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics, 36(9), 2896-2898.
10. Qin, M., Wu, S., Li, A., Zhao, F., Feng, H., Ding, L., & Ruan, J. (2019). LRScaf: improving draft genomes using long noisy reads. BMC genomics, 20(1), 1-12.
11. Xu, M., Guo, L., Gu, S., Wang, O., Zhang, R., Peters, B. A., ... & Zhang, Y. (2020). TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. GigaScience, 9(9), giaa094.
12. Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness. Gene prediction: methods and protocols, 227-245.
13. Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome biology, 21(1), 1-27.
14. Available online at: https://github.com/CFSAN-Biostatistics/fastatools
15. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
16. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120.
17. Vaseemuddin Md, Sanchit Misra, Heng Li, Srinivas Aluru. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. IEEE Parallel and Distributed Processing Symposium (IPDPS), 2019.
18. Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., ... & Li, H. (2021). Twelve years of SAMtools and BCFtools. Gigascience, 10(2), giab008.
19. Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907.
20. Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science, 2(11), 559-572.
21. Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome research, 19(9), 1655-1664.
22. Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., ... & Paten, B. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature, 587(7833), 246-251.
23. Martin, F. J., Amode, M. R., Aneja, A., Austine-Orimoloye, O., Azov, A. G., Barnes, I., ... & Flicek, P. (2023). Ensembl 2023. Nucleic acids research, 51(D1), D933-D941.
24. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome research, 20(1), 110-121.

Part 2. PhD student information on the overall year activity (courses/seminars/schools, mobility periods, participation to conferences)

List of attended courses/seminars/schools

1. EMBO Biodiversity Informatics, 19-20 May, HCMR, Greece
2. Landscape-Seascape genomics 29 October -04 November - Tjärnö Marine Laboratory, University of Gothenburg, Sweden

List of periods spent abroad

1. EMBO Biodiversity Informatics, 19-20 May, HCMR, Greece
2. Landscape-Seascape genomics 29 October -04 November - Tjärnö Marine Laboratory, University of Gothenburg, Sweden

List of conferences/workshops attended and of contributions eventually presented

3. EMBO Biodiversity Informatics, 19-20 May, HCMR, Greece
4. Landscape-Seascape genomics 29 October -04 November - Tjärnö Marine Laboratory, University of Gothenburg, Sweden
5. SMBE 2023 - 23-27 July 2023 - Ferrara. - Participation and poster presentation
6. Organisation of SLiM workshop - 27-31 May 2024 - Università Politecnica delle Marche, Italy

Part 3. PhD student information on publications

List of publications on international journals

J1. Maroso, F., Padovani, G., Muñoz Mora, V. H., Giannelli, F., Trucchi, E., & Bertorelle, G. Fitness consequences and ancestry loss in the Apennine brown bear after a simulated genetic rescue intervention. Conservation Biology, e14133. (published)

J2. Trucchi, E., Massa, P., Giannelli, F., Fernandes, F. A., Ancona, L., Stenseth, N. C., ... & Le Bohec, C. (2023). Gene expression is the main driver of purifying selection in large penguin populations. bioRxiv, 2023-08. (submitted)

J3. Study of the relationship between mito-nuclear discordance and sex biased dispersal in natural populations, using a forward simulation method (SLiM). (In preparation)

List of publications on conference proceedings

List of other publications (books, book chapters, patents)

13/11/23

Student signature

Supervisor signature