Master 2 internship (February – June/July 2020 ; dates can be flexible)

Company or University + lab: *Ifremer La Tremblade (Quantitative genetic team) – Brest (Bioinformatic team)*

Address: Ifremer Centre Brest 1625 Route de Sainte-Anne, 29280 Plouzané

Supervisor (to be contacted for applying):
- Last name: Jean-Baptiste
- First name: Lamy
- Position: Quantitative genetic and genomic
- Email: Jean.Baptiste.Lamy@ifremer.fr

- Last name: Patrick
- First name: Durand
- Position: Head of the Bioinformatic team at Ifremer
- Email: Patrick.Guido.Durand@ifremer.fr

Internship title:
Assembly and *de Novo* Annotation of *Crassostrea gigas* long-reads assembly

Keywords: Assembling polymorphic species, assembly trio, functional annotation, gene prediction, transcriptomic validation, mollusk, long-reads, pseudo-diploid assembly..

Internship description (½ page à 1 page) :

Since 2005, Ifremer has been working for the sequencing of C. gigas with the international scientific community involved in (Hedgecock et al., 2005). This has been translated into 3 successive responses to Genoscope's "DNA project" calls in 2005, 2006 and 2008. The project submitted in 2008 authorized the sequencing of BAC sequence clones of an inbred line provided by P. Gaffney (Univ. Delaware). The goal was to contribute to genome assembly in collaboration with the Sino-American consortium "Project Oyster Genome". Ifremer's co-financing came too late for the data to be included in the first version of the genome published in 2012 (Zhang et al., 2012).

This first public version of the genome was found to be highly fragmented despite the combination of fosmid approaches and whole genome sequencing on second-generation sequencers (Zhang et al., 2012). Several teams have also independently highlighted systematic assembly errors on the longest contigs (Hedgecock et al., 2015) using genetic maps and BAC-ends data obtained at Genoscope (Gagnaire et al. 2018). Faced with the growing need of the community and the "poor quality" of the available genome, a French consortium composed of Ifremer (Brest, La Tremblade and Montpellier), CNRS (ISEM), INRA (GenoToul), the University of Perpignan and the University of Caen decided at the end of 2015 to reassemble the oyster genome using third-generation sequencers (PacBio) in the start-up phase at the time. This first assembly test showed an assembly quality comparable to that available and with a correction of the errors of previous assemblies. However, several obstacles have emerged. In fact, the assemblers do not know how to manage the important heterozygosity of the oyster as well as the presence of structural variants in the heterozygous state and in large quantities. Assemblers can’t merge the alleles in the same contig, resulting

in assemblies with sizes twice that expected. Also, finding genetic material with the lowest possible heterozygosity had become essential. For this reason, during the years 2017 and 2018, the teams worked to develop highly consanguineous lines.

In the meantime, new third-generation sequencing technologies are emerging (Oxford Nanopore technology, hereinafter called ONT and 10X genomics) and provide lower cost access to 5 times more data with a 10 times lower error rate in 2017 The contribution of the 10X technology is decisive, because the sequenced individual turns out to be a spontaneous self-triploid which adds complexity during assembly. At the end of 2018, the characterization in flow cytometry of highly consanguineous individuals generated by our teams shows that they are diploid. In addition, a collaboration in progress with Pasteur-Paris allows the addition of a state-of-the-art technique (3c) to help during scaffolding, a limiting step for the moment (Flot, Marie-Nelly, and Koszul 2015, Marie-Nelly et al. 2014, Marbouty et al., 2017).

At the time of writing this proposal, we have a **polished haploid assembly with numerous contigs at the diploid states**. We will make another round of sequencing on ONT technology and 10X on trios of individuals (father-Mother and progeny). Our aims is to assemble both haplotypes independently thank to newer assembling methods (Koren et al, 2018).

The technical and scientific objectives of this internship consist in implementing on either the actual assembly or the next one:

- **Assemble ONT (10X genomics) datasets on trio of oyster (sequencing planned in January)**
- **Annotate and Mask transposable elements. LPBA have already developed a solution to find and annotate transposable elements. (https://github.com/JBerthelier/PiRATE). The actual question is to know if the software will support the size of an eukaryote genome.**
- **Annotate genes with high quality experimental evidence. Large bunch of work is already present in GigaTON database (http://gigaton.sigenae.org/). We aim to use methods based on similarity (with external evidences) and with ad initio methods (training methods as augustus)**
- **Make a functional annotation from the gene prediction.**

Such steps should be well documented and computationally repeatable and reproducible. The aim is to implement (and document) a robust pipeline ready to be used on other models without all the debugging steps and the format problems.

We are searching a motivated and highly skilled person on bioinformatics with a good understanding the biological process.

Salary or allowance: >550 €/per month

The candidate will work at Brest on the bioinformatics services, but he/she have on a weekly basis a working meeting (through webconference) with teams in France namely Ifremer La Tremblade and GenoToul (INRA Toulouse).