GENLOADICS

Plan de gestion de données créé à l'aide de DMP OPIDoR

Créateur du PGD : Eric Lombaert

Affiliation du créateur principal : INRAE - Institut national de recherche pour l'agriculture l'alimentation et l'environnement

Modèle du PGD : INRAE - General project template

Dernière modification du PGD : 08/10/2020

Financeur : ANR (JCJC project)

Résumé du projet :

Biological invasions are a major component of global change. Yet it is still not understood why some introduced populations are invasive and others are not. Among the most interesting hypotheses is the purging (i.e. elimination) of deleterious mutations during the introduction process. Deleterious mutations constitute what is called the genetic load because they are responsible for a decrease in the fitness of individuals by accumulating in the genome over time. After the introduction of a small number of individuals into the future invaded area, high levels of inbreeding and genetic drift lead to the exposure of deleterious mutations to natural selection. This can have two consequences: the decrease of the average fitness of the population and the purging of deleterious mutations. The assumption we make is that the populations that will actually become invasive are those that have purged part of their deleterious alleles. Indeed, they will have a major evolutionary advantage because they will be less subject to inbreeding depression and will thus better tolerate the low densities encountered during the establishment phase and during any secondary introductions. On the other hand, theory suggests that the genetic load that has not been purged could then become fixed on the fronts of geographic expansion by genetic drift (we use the expression "expansion load"), and thus compromise the success of the invasion in the longer term.

The measurement of life history traits that is typically considered to test such hypotheses is extremely difficult to implement. It requires working on living populations, which is difficult, or even impossible for some species. Today, high-throughput sequencing methods theoretically make it possible to compare the frequencies of deleterious or potentially deleterious mutations between populations on any species. We propose to test the purge hypothesis on a large taxonomic scale - 10 non-model insect species. To do this, we will use an exon capture protocol developed in our laboratory and suited to non-model species, and then we will quantify and compare the genetic load of native, invasive and spatially expanding populations.

The objective of this project will be both a technical and scientific first: (i) application of a protocol to capture exomes and compare genetic loads generalizable to non-model species, (ii) testing the hypotheses of deleterious allele purging and expansion load on key populations of a large number of invasive species. This project will provide an almost definitive answer to the purge hypothesis in invasion biology and will constitute a very important technical opening

for studying the evolution of the genetic load in a variety of organisms of interest (e.g. biological control agents, domestic animals, threatened species).

Chercheur Principal : Eric Lombaert

Identifiant ORCID : 0000-0003-0949-6690

Contact pour les Données : Eric Lombaert

Produits de recherche :

1. Genotypic data : Genotypic data for invasion routes inferences (Jeu de données)
2. Insect samples : Samples (tissues and/or DNA) of invasive insects (Collection)
3. Transcriptome : De novo transcriptome of several invasive insects (Jeu de données)
4. Exome capture probes : Exome capture probes for various invasive insect species (Jeu de données)
5. DNA Sequences : Raw sequences of various invasive insect exomes (Jeu de données)
6. Exome genotypic data : Exome capture genotypic data of various invasive insects (Jeu de données)
7. IEBs software : Software for the prediction of intron/exon boundaries (Logiciel)
8. Genetic load : Quantification of the genetic load of various invasive insect species (Jeu de données)
9. GENLOADICS website : GENLOADICS website : website dedicated to invasive insects with genetic data (Ressource interactive)

GENLOADICS

Information concerning the management plan

Genotypic data : Genotypic data for invasion routes inferences

Various authors according to the bibliographical results (see comment).

Those genotypic data will be gathered from various already published studies related to the population genetics of different invasive insect species. The published studies, the authors and the link to the associated data will be listed on a Github site dedicated to the GENLOADICS project.

Various affiliation (see comment)

Various dates (see comment)

Question sans réponse.

Insect samples : Samples (tissues and/or DNA) of invasive insects

Question sans réponse.

Université Côte d’Azur, INRAE, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France

21 september 2020

Question sans réponse.

Transcriptome : De novo transcriptome of several invasive insects

Question sans réponse.

Université Côte d’Azur, INRAE, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France

21 september 2020

Question sans réponse.

Exome capture probes : Exome capture probes for various invasive insect species

Nimblegen (Roche company)

Roche Company

21 September 2020

Question sans réponse.

DNA Sequences : Raw sequences of various invasive insect exomes

Question sans réponse.

Université Côte d’Azur, INRAE, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France

21 september 2020

Question sans réponse.

Exome genotypic data : Exome capture genotypic data of various invasive insects

Question sans réponse.

Université Côte d’Azur, INRAE, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France

21 september 2020

Question sans réponse.

IEBs software : Software for the prediction of intron/exon boundaries

Emeline Deleury (emeline.deleury@inrae.fr)

Université Côte d’Azur, INRAE, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France

21 september 2020

v1.1

Genetic load : Quantification of the genetic load of various invasive insect species

Question sans réponse.

Université Côte d’Azur, INRAE, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France

21 september 2020

Question sans réponse.

GENLOADICS website : GENLOADICS website : website dedicated to invasive insects with genetic data

Emeline Deleury (emeline.deleury@inrae.fr)

Université Côte d’Azur, INRAE, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France

21 september 2020

Question sans réponse.

Information on the research project

ANR AAPG 2019

Agence Nationale de la Recherche (ANR)

ANR JCJC

Question sans réponse.

GENLOADICS

GENLOADICS - Evolution of the genetic load during biological invasions

INRAE Centre de Recherche PACA, Sophia-Antipolis, France

Question sans réponse.

INRAE, CNRS, Université Côte d’Azur, ISA, Sophia-Antipolis, France

From 2020-01-01 to 2023-12-31
48 months

Brief presentation of project data

Genotypic data : Genotypic data for invasion routes inferences

Existing population genotypic data of different kind (microsatellites, sequences, SNP) and of various invasive insect species to be used to infer invasion routes and estimate bottleneck severity by the use of approximate Bayesian computation (project's Task 1).

Origin of data: reused data. All data will be obtain from previous studies and are thus already available in various forms (depending on the authors choices).

Some associated publications: Sherpa et al., 2019, Javal et al., 2019, Andraca-Gomeiz et al., 2020, Jacquet et al., 2015, Bras et al., in prep., Miller et al., 2005, Lombaert et al., 2018, Fraimout et al., 2017, Lombaert et al., 2011, Lesieur et al., 2019, Kerdelhue et al., 2014, Auger-Rozenberg et al., 2012, Ryan et al., 2019, Correa et al., 2019, Papura et al., 2012, Boissin et al., 2012, Guillemaud et al., 2015, Arca et al., 2015).

Insect samples : Samples (tissues and/or DNA) of invasive insects

Insect samples (tissues and/or DNA) will be obtained for every species (around 10) and population (between 3 and 5 per species) chosen in Task 1. These samples correspond to the result of the project's Task 2.

Origin of data: mix of reused (i.e. already sampled) and new data (depending on the species). The Nagoya protocol will be followed. Tissues will be preserved in alcohol. Tissues and DNA will be stored at -25°C in freezer located at INRAE Sophia-Antipolis. An Excel (*.xlsx) spreadshit database will contained all information about samples: species, number of individuals, GPS coordinates, sampling dates, name of the collector(s), etc. This file will be shared between partners of the project. Information about population samples will be listed on the GENLOADICS website.

Transcriptome : De novo transcriptome of several invasive insects

RNAseq will be done for the species for which no transcriptome nor genome is available (around 5 species among the 10). A *de novo* transcriptome will be construct (project's Task 3).

Origin of data: new data. The transcriptome will be construct by an external platform. The raw data as well as the final results (assembled transcriptomes) will be made available on public archives/sequence databases (e.g. EMBL-TSA , BIPAA BioInformatics Platform for Agroecosystem Arthropods).

Exome capture probes : Exome capture probes for various invasive insect species

Exome capture probes will be designed to cover around 5 Mb of the exome of every target species. *De novo* transcriptome, or annotated genome when available, will be used for the probe design (project's Task 3).

Origin of data: new data. We will perform the first step of the bioinformatic probe design. The final design step and the production will performed by the company Nimblegen (Roche). Consequently, we will not have a full access to the sequence of the final probes. The list of targeted transcripts and the positions of the probe blocks on these targets will be available from the GENLOADICS website.

Associated publication: https://doi.org/10.1101/583534

DNA Sequences : Raw sequences of various invasive insect exomes

Exome DNA (small read) sequences will be obtained for every species and population chosen in Task 1, and sampled in Task 2.

Origin of data: new data. Before sequencing, we will capture the fragmented DNA with exome capture probes (project's Task 3). the captured DNA will be sequenced by an external platform with an Illumina HiSeq3000 sequencer in 150 pair-end mode (project's Task 4). The obtained raw sequences will be made available on an open

plateform (e.g. ENA/NCBI, SRA). The accession numbers and links to the data in these databases will be listed on GENLOADICS website.

Exome genotypic data : Exome capture genotypic data of various invasive insects

A large amount of SNPs specific to the exome will be generated for each population and species (project's Task 4).

Origin of data: new data. On the basis of the DNA sequences captured, we will performed a SNP calling analysis to identify SNPs accross the exome of the various populations for every species. The obtained genotypes will be made fully available on an open online plateform (e.g. NCBI, genbank, Zenodo). Bioinformatic pipeline will be made available on Github. The links to the SNP list will be listed on GENLOADICS website.

IEBs software : Software for the prediction of intron/exon boundaries

We will develop a software allowing the prediction of Intron-Exon Boundaries (IEB) on *de novo* transcript sequences using result of mapping genomic reads directly onto coding sequences (CDS).

Origin of data: mix of reused and new data. A preliminary version of the software is freely available at https://github.com/edeleury/IEB-finder

Associated publication: https://doi.org/10.1101/583534

Genetic load : Quantification of the genetic load of various invasive insect species

The SNPs identified for each species will allow for the characterization and quantification of the genetic load (project's Task 4).

Origin of data: new data. The SNPs will be classified into synonymous and non-synonymous variants. Close related outgroup species will be used to polarize ancestral and derived states. We will then classify non-synonymous mutations into functional classes (“conservative missense”, “radical missense” and “nonsense”) on the basis of changes in the biochemical properties of amino acids. Finaly, we will use a number of method to compare the genetic load between native and invasive populations. All datasets and R scripts for data analyses will be made avalaible at the time of publication of the results using open archives such as Zenodo. The links to the annotated SNPs will be listed on GENLOADICS website.

GENLOADICS website : GENLOADICS website : website dedicated to invasive insects with genetic data

This website will centralize the different information available and produced for each invasive insect species studied in the project.

Origin of data: new data. For each invasive insect species, a page will list the available data (e.g. bibliography related to its invasion, links to available genetic data, link to its transcriptome or genome) and, if the species is used in the GENLOADICS project, the data and main results produced (invasion routes, insect samples, targeted transcripts and associated capture probes, links to population genomic data produced, list of SNPs found and annotated). During the project, this site will materialize the progress of the study for each species. At the end of the project, it will centralize the links to the different data and results. A home page will point to the different invasive insects species studied in the project.

Intellectual property rights

Previous data will remain the property of the providing partner.

Data generated and results obtained by our laboratory will belong to us.

The final capture probe sequences/designs will be the property of Nimblegen company (Roche).

All biological material will stay the property of the partner who have collected it (we will only use a small amount of the tissue or DNA, in agreement with the considered partner).

The use of genetic resources will be done in conformity with the Nagoya protocol. Formalities will be dealt by ourselves.

Confidentiality

We will not produce any confidential data in this project.

Only the capture probe designs (to which we will not have access) will be confidential, but the Nimblegen company will deal with this aspect.

Question sans réponse.

Question sans réponse.

Access and sharing of data at the end of the project

Yes, there is a legal obligation to share data produced by INRAE (CADA law).

All data will be shared in online public databases, with the exception of the capture probe sequences (property of the Nimblegen company).

De novo transcriptome is essential to study the biology of a non-model species (e.g. phylogeny, search for members of specific families such as P450 or carbohydrate Active Enzymes).

Exome sequences (i.e. captured DNA with probes designed from *de novo* transcripts) obtained on non-model species can be used for the extension of gene models and definition of intron/exon structure. For species with draft genome, genomic reads can be used to improve the genome assembly of the genome at the location of the genes.

The SNPs obtained can be used to address various questions in evolutionary biology (e.g. GWAS analyses).

The data will be in text or tabulated text files (free and open formats) and do not require special tools.

Large data files will be in the usual standard formats (e.g. FASTQ for raw read data, FASTA for transcriptome data, GFF for genome sequence data, CVF for SNPs data) so as to be readable by the bioinformatics tools used for their analysis.

Raw read data and transcriptome assembly will be made available via ENA or NCBI with an identifier and accession numbers referenced in published articles.
Documentation, final datasets (i.e. "genetic load" datasets) and code will be made available via Zenodo.org or Github.com or https://data.inrae.fr/ with assigned DOI.

Data will be share publicly.
A Creative Common licence will be chosen.

At the latest, at the end of the project (end of 2023), probably earlier for most of the datas.

No time limitation.

Yes, permanent identifier (DOI or accession number) will identify the data.

Question sans réponse.

Description and organisation of data

Genotypic data : Genotypic data for invasion routes inferences

Bibliography and contact with authors when necessary.

If it exists, study reference associated with the data.

All information about samples (species, taxonomy id, number of individuals, development stage, GPS coordinates, sampling dates, author names, ...) and genetic marker data (type like microsatellites or SNP, their number, accession number in database).

During the project, we will fill in predefined excel files to fill in the metadata associated with each data type.
A unique identifier will be associated to each dataset of the same type and will allow to link the data presented in different tables in order to preserve their link of origin.

We will separate raw data from processed data, deliverables and processing programs. The files will be organized by species and then by step of the project. The files will be named including the identifiers of the biological material used. The names of the output files will be named with the name of the input file, the version of the software used and the main parameters used.

Question sans réponse.

Insect samples : Samples (tissues and/or DNA) of invasive insects

Tissues (or DNA) will be obtained by partners or ourselves. They will be stored in Ethanol at -25°C in a freezer located at INRAE Sophia-Antipolis.

Question sans réponse.

All information about samples: species, taxonomy id, number of individuals, development stage, GPS coordinates, sampling dates, name of the collector(s), etc.

During the project, we will fill in predefined excel files to fill in the metadata associated with each data type.
A unique identifier will be associated to each dataset of the same type and will allow to link the data presented in different tables in order to preserve their link of origin.

All population sample will receive a specific name: [Species]-[population]-[sampling year]-[unique population sample letter]

Question sans réponse.

Transcriptome : De novo transcriptome of several invasive insects

RNA will be extract from fresh individuals using a commercial extraction kit. The transcriptome will be construct by an external platform.

Document describing how the transcriptome was obtained: description of the biological material used to extract the RNA, description of the protocol to obtain total RNAs, desciption of how the sequencing and assembly were done.

Species, strain, taxonomic id, stages.

During the project, we will fill in predefined excel files to fill in the metadata associated with each data type.
A unique identifier will be associated to each dataset of the same type and will allow to link the data presented in different tables in order to preserve their link of origin.

We will separate raw data from processed data, deliverables and processing programs. The files will be organized by species and then by step of the project. The files will be named including the identifiers of the biological material used. The names of the output files will be named with the name of the input file, the version of the software used and the main parameters used.

Question sans réponse.

Exome capture probes : Exome capture probes for various invasive insect species

For target selection, we will apply the protocol described here: https://doi.org/10.1101/583534
The probe design will be performed by the company Nimblegen.

Document describing the different filters used to select targets or link to the paper that refers to the dataset.

Species, taxonomic id, probe source (from a transcriptome or from genome), target exome size.

During the project, we will fill in predefined excel files to fill in the metadata associated with each data type.
A unique identifier will be associated to each dataset of the same type and will allow to link the data presented in different tables in order to preserve their link of origin.

We will separate raw data from processed data, deliverables and processing programs. The files will be organized by species and then by step of the project. The files will be named including the identifiers of the biological material used. The names of the output files will be named with the name of the input file, the version of the software used and the main parameters used.

Question sans réponse.

DNA Sequences : Raw sequences of various invasive insect exomes

For each population of each species, ~50 individuals will be pooled in a single vial for DNA extraction purpose. The first step consists in extracting DNA and performing a mechanic fragmentation of the DNA by sonication. The next step aims at building the DNA libraries with Illumina adapters and barcode indexes. For each population, the libraries will be prepared independently on 3 aliquots of the original DNA, which will allow for technical replicates. Finally, barcoded amplified DNA will be pooled in equimolar concentration prior to hybridization to have 1µg in total per capture reaction. After capture, DNA enriched in target fragments is amplified by PCR. Finally, the captured DNA will be sequenced with an Illumina HiSeq3000 sequencer in 150 pair-end mode.

Document describing how the exome capture and sequencing was carried out, or link to the paper that refers to the dataset.

The same as insect samples (insect sample id) and extraction number.

During the project, we will fill in predefined excel files to fill in the metadata associated with each data type.
A unique identifier will be associated to each dataset of the same type and will allow to link the data presented in different tables in order to preserve their link of origin.

We will separate raw data from processed data, deliverables and processing programs. The files will be organized by species and then by step of the project. The files will be named including the identifiers of the biological material used. The names of the output files will be named with the name of the input file, the version of the software used and the main parameters used.

The "Phred quality score" will be measured for every sequences using FastQC software.

Exome genotypic data : Exome capture genotypic data of various invasive insects

For each library, cleaned and trimmed reads will be mapped directly on the coding target sequences using BOWTIE2 (or on the genome if available). Reads with multiple alignments or that do not map properly with their mate will be discarded. Separately on each library, variants sites will then be stringently called with the SAMtools mpileup and varscan2 softwares. All variant positions will then be genotyped in all libraries. Final list of SNPs will then be selected according to a combination of (i) minimal sequencing coverage of the variant and (ii) detections of the variant in more than one sample (considering all aliquots of all population samples). A maximum-coverage filter based on the mean sequencing depth will also be applied to avoid possible replicated areas that could remain. Genotyping biases at few exon ends related to direct mapping on transcripts will be discarded using our predictor of intron-exon boundaries.

Description of how the dataset was obtained, or link to the paper that refers to the dataset.

Species, insect sample id and extraction number, replicat number, sequencer name, read length, paired read, length of insertion between pairs.

During the project, we will fill in predefined excel files to fill in the metadata associated with each data type.
A unique identifier will be associated to each dataset of the same type and will allow to link the data presented in different tables in order to preserve their link of origin.

We will separate raw data from processed data, deliverables and processing programs. The files will be organized by species and then by step of the project. The files will be named including the identifiers of the biological material used. The names of the output files will be named with the name of the input file, the version of the software used and the main parameters used.

Question sans réponse.

IEBs software : Software for the prediction of intron/exon boundaries

A preliminary version of the software IEB-finder is already available here: https://github.com/edeleury/IEB-finder
We will improve its use and performance by testing it on most of the data produce on the various species of the

project. Species with already available genome will allow for the quantification of false-negative and false-positive.

A readme file that explains how to use the predictor and a dataset to test the tool.
A README.md is available at https://github.com/edeleury/IEB-finder

Version.

Question sans réponse.

A strict control of the versions will be made (e.g. current version: v1.1).

Question sans réponse.

Genetic load : Quantification of the genetic load of various invasive insect species

As all targets will be CDS, we will be able to classify variants into synonymous and non-synonymous changes with SnpEff. Close related outgroup species will be used to polarize ancestral and derived states. We will then classify non-synonymous mutations into functional classes (“conservative missense”, “radical missense” and “nonsense”) on the basis of changes in the biochemical properties of amino acids. To compare the genetic load between populations, we will use several statistics such as the mean derived allele frequency within functional class, the relative number of derived deleterious alleles that are frequent in one population and not another, or the ratio of nonsynonymous to synonymous polymorphic sites adjusted for the frequencies of the derived allele. Synonymous SNPs will be used as a proxy for neutral sites. We will also use the properties of site frequency spectra (SFS) to infer the distribution of fitness effects (DFE) of non-synonymous mutations with Fitaai.
R scripts for data analyses will be produced and used.

Description of how the dataset was obtained, or link to the paper that refers to the dataset. Definition of each column of the tabulated file so that the data can be reused.

Species and the list of populations from which the dataset will be obtained.

During the project, we will fill in predefined excel files to fill in the metadata associated with each data type.
A unique identifier will be associated to each dataset of the same type and will allow to link the data presented in different tables in order to preserve their link of origin.

We will separate raw data from processed data, deliverables and processing programs. The files will be organized by species and then by step of the project. The files will be named including the identifiers of the biological material used. The names of the output files will be named with the name of the input file, the version of the software used and the main parameters used.
A strict control of the version of the R scripts will be made.

Question sans réponse.

GENLOADICS website : GENLOADICS website : website dedicated to invasive insects with genetic data

We will create a typical page (Rmarkdown) per species on which we will archive the available information and data (links to archives or sequence databases). The pages will be completed at the end of each task when results or data have been generated (versioning). A central page will give access to the different pages of the studied species.

Question sans réponse.

Version.

Question sans réponse.

A strict control of the versions will be made.

Question sans réponse.

Data storage and backup during the project

A copy of raw data files will be stored in a datacenter with a copy of the updated metadata file.
Scripts and lighter data files (e.g. excel spreadshit) will be copied from local computers to a shared and versioned workspace (e.g. Github).
Any file (raw data) or script required to produce a valuable data file will be saved on 2 geographically distinct media.
The intermediate files generated may be saved on local disks for the duration of the project.
Biological samples will be stored in a freezer at -25°C.

Question sans réponse.

Including the intermediate files generated during the project, the amount of data is estimated at 20 To.

Large datasets (sequences) will be stored in a datacenter and on public databases.
Small datasets (e.g. *.xlsx files) will be stored on a computer, on an external hard drive, on a cloud (pcloud) and on Github.

Mainly in France (INRAE).
And on international database servers.

Yes

Question sans réponse.

At first, only the researchers working on the project will have rights access to the data (via Access Control Policy). Latter, data will be shared publicly.

Identification statement and password.

Log files containing the parameters and versions of the software used. Access to data via user accounts with identifiers. Data accessible with identifiers (login/password). When executing scripts/software, error checking. Integrate validation steps to test file integrity (use the md5sum tool to check the integrity of data after transfer to a remote media, test the consistency of the data produced via tests built into the script code).

Data archiving and conservation after the end of the project

All biological samples (i.e. remaining insect tissues), raw data (e.g. DNA sequences, RNA sequences) and final data (e.g. *de novo* transcriptome, list of annotated SNPs, IEB positions on transcript sequence, scripts, DNA samples) will be conserved. Intermediate files will be removed shortly after using (e.g. the **.bam* files after the SNP calling procedure).

The archive platform will be (depending on the data): ENA/NCBI, SRA, Zenodo, Github,https://data.inrae.fr/, BioRxiv.

No time-limitation.

Question sans réponse.

The final amount of data (perennial in time) should be around 3 To

Question sans réponse.