DMP du projet "GENLOADICS"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "INRAE - Project template" fourni par INRAE - Institut national de recherche pour l'agriculture l'alimentation et l'environnement.

Plan Details

Plan title	DMP du projet "GENLOADICS"
Deliverable	DMP_GL_v2
Version	Version intermédiaire
Plan purpose/scope	Mise à jour de la version initiale.
Language	fra
Creation date	2020-09-21
Last modification date	2020-10-08

Project Details

Project title	GENLOADICS
Acronym	GENLOADICS
Abstract	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).

Funding

- ANR (JCJC project) :

Start date 2020-01-01

End date 2024-06-30

Research outputs :

1. Genotypic data for invasion routes inferences (Jeu de données)
2. Samples (tissues and/or DNA) of invasive insects (Collection)
3. De novo transcriptome of several invasive insects (Jeu de données)
4. Assembled genome of several invasive insects (Jeu de données)
5. Raw sequences of various invasive insect (whole genome sequencing) (Jeu de données)
6. genotypic data (SNPs) of various invasive insects (Jeu de données)
7. Quantification of the genetic load of various invasive insect species (Jeu de données)
8. GENLOADICS website : website dedicated to invasive insects with genetic data (Resource interactive)

Contributors

Name	Affiliation	Roles
Eric Lombaert - https://orcid.org/0000-0003-0949-6690		• Coordinateur du projet • Personne contact pour les données (Genome, Transcriptome, Genotypic data, Insect samples, DNA Sequences, genotypic data, Genetic load, GENLOADICS website) • Responsable du plan de gestion de données

DMP du projet "GENLOADICS"

Information concerning the management plan

Genotypic data for invasion routes inferences

Author of the DMP (if different from the principal investigator/researcher): name, email

Various authors according to the bibliographical results (see comment).
Those genotypic data were 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.

Affiliation of the author of the DMP

Various affiliation (see comment)

Date of creation of DMP

Various dates (see comment)

Current version: (n°, date)

Question sans réponse.

Samples (tissues and/or DNA) of invasive insects

Author of the DMP (if different from the principal investigator/researcher): name, email

Question sans réponse.

Affiliation of the author of the DMP

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

Date of creation of DMP

21 september 2020

Current version: (n°, date)

Question sans réponse.

De novo transcriptome of several invasive insects

Author of the DMP (if different from the principal investigator/researcher): name, email

Question sans réponse.

Affiliation of the author of the DMP

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

Date of creation of DMP

21 september 2020

Current version: (n°, date)

Question sans réponse.

Assembled genome of several invasive insects

Author of the DMP (if different from the principal investigator/researcher): name, email

Question sans réponse.

Affiliation of the author of the DMP

Question sans réponse.

Date of creation of DMP

14 December 2021

Current version: (n°, date)

Question sans réponse.

Raw sequences of various invasive insect (whole genome sequencing)

Author of the DMP (if different from the principal investigator/researcher): name, email

Question sans réponse.

Affiliation of the author of the DMP

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

---

Date of creation of DMP

21 september 2020

Current version: (n°, date)

Question sans réponse.

genotypic data (SNPs) of various invasive insects

Author of the DMP (if different from the principal investigator/researcher): name, email

Question sans réponse.

Affiliation of the author of the DMP

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

Date of creation of DMP

21 september 2020

Current version: (n°, date)

Question sans réponse.

Quantification of the genetic load of various invasive insect species

Author of the DMP (if different from the principal investigator/researcher): name, email

Question sans réponse.

Affiliation of the author of the DMP

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

Date of creation of DMP

21 september 2020

Current version: (n°, date)

Question sans réponse.

GENLOADICS website : website dedicated to invasive insects with genetic data

Author of the DMP (if different from the principal investigator/researcher): name, email

Emeline Deleury (emeline.deleury@inrae.fr)

Affiliation of the author of the DMP

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

Date of creation of DMP

21 september 2020

Current version: (n°, date)

Question sans réponse.

Information on the research project

Identifier of the call for proposal

ANR AAPG 2019

Project funder(s)

Agence Nationale de la Recherche (ANR)

Name of research programme

ANR JCJC

Reference of funding agreement

Question sans réponse.

Project acronym

GENLOADICS

Name of research project

GENLOADICS - Evolution of the genetic load during biological invasions

---

Project leader institution, coordinator & beneficiary (name, country)

INRAE Centre de Recherche PACA, Sophia-Antipolis, France

Other partners (name, country, role of each partner other than the project leader institution)

Question sans réponse.

Unit to which project leader belongs

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

Project dates and duration

From 2020-01-01 to 2024-06-30
54 months

Brief presentation of project data

Genotypic data for invasion routes inferences

Brief presentation of project data :

- **Type, scope, scale**
- **Origin (new data collection; data conversion/transformation; data sharing/exchange; data purchase)**
- **Associated publications**

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 were obtained 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).

Samples (tissues and/or DNA) of invasive insects

Brief presentation of project data :

- **Type, scope, scale**
- **Origin (new data collection; data conversion/transformation; data sharing/exchange; data purchase)**
- **Associated publications**

Insect samples (tissues and/or DNA) are obtained for every species (12) 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 is followed. Tissues are preserved in alcohol. Tissues and DNA are stored at -25°C in freezer located at INRAE Sophia-Antipolis. An Excel (*.xlsx) spreadshit database contains all information about samples: species, number of individuals, GPS coordinates, sampling dates, name of the collector(s), etc. This file is shared between partners of the project

through a dropbox account. Information about population samples will be listed on the GENLOADICS website.

De novo transcriptome of several invasive insects

Brief presentation of project data :

- **Type, scope, scale**
- **Origin (new data collection; data conversion/transformation; data sharing/exchange; data purchase)**
- **Associated publications**

RNAseq will be done for the species for which no transcriptome is available (around 3 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).

Assembled genome of several invasive insects

Brief presentation of project data :

- **Type, scope, scale**
- **Origin (new data collection; data conversion/transformation; data sharing/exchange; data purchase)**
- **Associated publications**

Genomic DNA sequencing will be done for the species for which no assembled genome is available (around 3 species among the 12). A genome will be assembled.

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

Raw sequences of various invasive insect (whole genome sequencing)

Brief presentation of project data :

- **Type, scope, scale**
- **Origin (new data collection; data conversion/transformation; data sharing/exchange; data purchase)**
- **Associated publications**

Full-genome 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. The 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.

genotypic data (SNPs) of various invasive insects

Brief presentation of project data :

- **Type, scope, scale**
- **Origin (new data collection; data conversion/transformation; data sharing/exchange; data purchase)**
- **Associated publications**

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

Origin of data: new data. On the basis of the DNA sequences (whole genome sequencing), we will performed a SNP calling analysis to identify SNPs accross the genome, with a focus on 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.

Quantification of the genetic load of various invasive insect species

Brief presentation of project data :

- **Type, scope, scale**
- **Origin (new data collection; data conversion/transformation; data sharing/exchange; data purchase)**
- **Associated publications**

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 : website dedicated to invasive insects with genetic data

Brief presentation of project data :

- **Type, scope, scale**
- **Origin (new data collection; data conversion/transformation; data sharing/exchange; data purchase)**
- **Associated publications**

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 and genome) and, if the species is used in the GENLOADICS project, the data and main results produced (invasion routes, insect samples, 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

Who owns the rights on data and other information created during the project?

Previous data will remain the property of the providing partner.

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

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).

Will material protected by specific rights be used during the project? In this case, who will deal with the formalities required, obtain the authorisations for use and possible dissemination?

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

Confidentiality

Identification of the confidential data sets

We do not produce any confidential data in this project.

What are the measures taken and the norms that must be met to guarantee this confidentiality?

If applicable, how will data confidentiality be guaranteed when the data will be shared or made available for second level analysis?

Access and sharing of data at the end of the project

Is there an obligation to share data (or on the contrary a prohibition or restriction?

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

What data will be shared at the end of the project? If all the data are not available in the same way, or at the same time, please specify

All data will be shared in online public databases.

What are the potential reuses for these data?

De novo transcriptomes and assembled genomes are 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).

Sequences (i.e. whole genome re-sequencing) obtained on non-model species can be used for the extension of gene models and definition of intron/exon structure. They can also 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).

Does reading the data require specific software or tool? If so, which one?

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.

How will the data be shared?

Raw read data, and transcriptome and genome 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.

With whom? With what licence?

Data will be share publicly.

A Creative Common licence will be chosen.

As from when?

At the latest, at the end of the project (mid-2024), probably earlier for most of the datas.

For how long?

No time limitation.

Will the data be identified by a permanent identifier (DOI or other)?

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

Which organisation will be responsible for requesting the identifier in the case of multi-partner projects?

Description and organisation of data

Genotypic data for invasion routes inferences

What methods and tools are used to acquire and process data? Specify the different formats in which the data will be available in the different phases of research

Bibliography and contact with authors when necessary.

Documentation associated with the data

If it exists, study reference associated with the data.

What types of metadata will be produced to accompany the data? What standards or taxonomies will be used to describe 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).

How will the metadata be produced?

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

How will the data files be managed and organised during the project: control of versions, conventions for naming files, organisation of files...

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

What is the quality control procedure of the data?
Enclose the quality insurance plan if possible

Question sans réponse.

Samples (tissues and/or DNA) of invasive insects

What methods and tools are used to acquire and process data? Specify the different formats in which the data will be available in the different phases of research

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

Documentation associated with the data

An excel file, saved in a cloud (dropbox), contains all associated information.

What types of metadata will be produced to accompany the data? What standards or taxonomies will be used to describe the data?

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

How will the metadata be produced?

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

How will the data files be managed and organised during the project: control of versions, conventions for naming files, organisation of files...

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

What is the quality control procedure of the data?
Enclose the quality insurance plan if possible

Question sans réponse.

De novo transcriptome of several invasive insects

What methods and tools are used to acquire and process data? Specify the different formats in which the data will be available in the different phases of research

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

Documentation associated with the data

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.

What types of metadata will be produced to accompany the data? What standards or taxonomies will be used to describe the data?

Species, strain, taxonomic id, stages.

How will the metadata be produced?

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.

How will the data files be managed and organised during the project: control of versions, conventions for naming files, organisation of files...

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.

What is the quality control procedure of the data?
Enclose the quality insurance plan if possible

Question sans réponse.

Assembled genome of several invasive insects

What methods and tools are used to acquire and process data? Specify the different formats in which the

data will be available in the different phases of research

Genomic DNA will be extract from fresh individuals using a commercial extraction kit. The Genome will be construct by an external platform.

Documentation associated with the data

Document describing how the genome was obtained: description of the biological material used to extract the genomic DNA, description of the protocol to obtain total DNAs, desciption of how the sequencing and assembly were done.

What types of metadata will be produced to accompany the data? What standards or taxonomies will be used to describe the data?

Question sans réponse.

How will the metadata be produced?

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.

How will the data files be managed and organised during the project: control of versions, conventions for naming files, organisation of files...

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.

What is the quality control procedure of the data?
Enclose the quality insurance plan if possible

Question sans réponse.

Raw sequences of various invasive insect (whole genome sequencing)

What methods and tools are used to acquire and process data? Specify the different formats in which the data will be available in the different phases of research

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. Finally, the DNA will be sequenced with an Illumina Novaseq6000 sequencer in 150 pair-end mode.

Documentation associated with the data

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

What types of metadata will be produced to accompany the data? What standards or taxonomies will be used to describe the data?

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

How will the metadata be produced?

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.

How will the data files be managed and organised during the project: control of versions, conventions for naming files, organisation of files...

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.

What is the quality control procedure of the data?
Enclose the quality insurance plan if possible

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

genotypic data (SNPs) of various invasive insects

What methods and tools are used to acquire and process data? Specify the different formats in which the data will be available in the different phases of research

For each library, cleaned and trimmed reads will be mapped directly on the corresponding genome using BOWTIE2. 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.

Documentation associated with the data

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

What types of metadata will be produced to accompany the data? What standards or taxonomies will be used to describe the data?

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

How will the metadata be produced?

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.

How will the data files be managed and organised during the project: control of versions, conventions for naming files, organisation of files...

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.

What is the quality control procedure of the data?
Enclose the quality insurance plan if possible

Question sans réponse.

Quantification of the genetic load of various invasive insect species

What methods and tools are used to acquire and process data? Specify the different formats in which the data will be available in the different phases of research

On the exome section of the genome, 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 Fitðaði.

R scripts for data analyses will be produced and used.

Documentation associated with the data

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.

What types of metadata will be produced to accompany the data? What standards or taxonomies will be used to describe the data?

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

How will the metadata be produced?

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.

How will the data files be managed and organised during the project: control of versions, conventions for naming files, organisation of files...

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.

What is the quality control procedure of the data?
Enclose the quality insurance plan if possible

Question sans réponse.

GENLOADICS website : website dedicated to invasive insects with genetic data

What methods and tools are used to acquire and process data? Specify the different formats in which the data will be available in the different phases of research

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.

Documentation associated with the data

Question sans réponse.

What types of metadata will be produced to accompany the data? What standards or taxonomies will be used to describe the data?

Version.

How will the metadata be produced?

Question sans réponse.

How will the data files be managed and organised during the project: control of versions, conventions for naming files, organisation of files...

A strict control of the versions will be made.

What is the quality control procedure of the data?
Enclose the quality insurance plan if possible

Question sans réponse.

Data storage and backup during the project

Storage: what media will be used for data 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 are stored in a freezer at -25°C.

Storage: What types of flows will be used by the data during the project?

Manual network flow (mail, sftp)

Storage: What is the estimated amount of data?

Including the intermediate files generated during the project, the amount of data is estimated at 50 To.

Storage: Where will the data be stored, on what type of host?

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.

Storage: Where will the data be located geographically?

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

Security: Does the entity physically hosting the data have a security policy for its information system?

Yes

Security - Confidentiality: will the data de exchanged or shared with third parties?

Security - Confidentiality: how are rights of access to data determined during the research project?

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.

Security - Confidentiality: how will all the project partner researchers have access to data during the project?

Identification statement and password.

Security – Integrity – Traceability: what measures of protection will be taken to monitor data production and analysis during the project?

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

What data will be conserved in the medium and long term and what data will be destroyed?

All biological samples (i.e. remaining insect tissues, DNA and RNA), raw data (e.g. DNA sequences, RNA sequences) and final data (e.g. *de novo* transcriptome, assempled genome, 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).

On what permanent archive platform will the data that are to be conserved long-term be archived?
What procedures will be set up for long-term conservation?

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

What is the duration of data conservation?

No time-limitation.

Who will be responsible for long-term conservation?
Name an individual contact

Question sans réponse.

What will be the volume of these data?

The final amount of data (perennial in time) should be around 5 To

What funding guarantees will cover the costs of long-term conservation?

Recurent INRAE funding