PGD du projet "REtroViral Emergence @farm"

Jocelyn Turpin, Caroline Leroux

PGD du projet "REtroViral Emergence @farm"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - DMP template (english)" fourni par Agence nationale de la recherche (ANR).

Plan Details

Plan title	PGD du projet "REtroViral Emergence @farm"
Version	Mid term version
Fields of science and technology (from OECD classification)	Biological sciences (Natural sciences)
Language	eng
Creation date	2023-03-31
Last modification date	2024-09-30
Identifier	DMP du projet "REtroViral Emergence @farm"
Identifier type	Local identifier
License	**Name**: Creative Commons Attribution Share Alike 4.0 International **URL**: http://spdx.org/licenses/CC-BY-SA-4.0.json
Associated documents (publications,reports, patents, experimental plan...), website	• Particular sequence characteristics induce bias in the detection of polymorphic transposable element insertions : https://doi.org/10.1101/2024.09.25.614865 • A genome-wide study of ruminants reveals two endogenous retrovirus families still active in goats : https://doi.org/10.1101/2024.06.21.600049 • Association between genetic clades and cancer prevalence suggested by French-wide study of oncogenic small ruminant beta-retroviruses diversity : https://doi.org/10.1101/2024.07.18.604097

Project Details

Project title REtroViral Emergence @farm

Acronym REVEatFarm

Abstract

Small ruminant oncogenic beta-retroviruses induce lung and nasal cancers in sheep and goat. Mostly sporadic, severe outbreaks of cancer with a high mortality rate and transmission also occur. JSRV induced lung adenocarcinoma was recently requalified as emergent by the OIE. Despite their economic impact worldwide, these diseases are neglected with few to no regulation and epidemiological surveillance. The control of these infectious diseases, with neither vaccine nor treatment available, can only rely on effective measures to limit the diffusion of these deadly viruses. The objectives of this project are to characterize the persistence and transmission of this airborne transmitted virus and to determine if we could identify a viral factor that may explain the switch between sporadic vs epidemic cancers. To answer these questions, we will rely on our biobank of cancers from JSRV and ENTV-2 infected animals and prospective samples via our network of veterinarians, breeders and technical partners. We will investigate the presence of retroviral genomes in environmental samples in the farm but also quantify the virus shed by symptomatic and asymptomatic animals. To improve and facilitate the surveillance of beta-retroviruses, we will test if wastewater or stools could be used as an alert tool of JSRV and ENTV circulation. The second part of the project will question the nature of the retroviruses associated with high prevalence of cancers. To identify viral variants or molecular signatures associated with high or low pathogenicity phenotypes, we will sequence and compare viruses circulating in flocks with sporadic or epidemic cancers. Finally, we will determine if the oncogenic switch might be explained by a change in the viral expansion in vivo. All together this project will give us the scientific basis to prevent the emergence of oncogenic beta-retroviruses.

Funding

- Agence Nationale de la Recherche : ANR-22-CE35-0002
- French National Research Agency : ANR-22-CE35-0002

Start date 2022-10-01

End date 2026-09-30

Research outputs :

1. Sequencing data: Raw sequencing data and their derivatives (Dataset)

Contributors

Name	Affiliation	Roles
Caroline Leroux - https://orcid.org/0000-0002-7923-3127	UMR754 "Infections Virales et Pathologie Comparée" - 199517675N	• Personne contact pour les données
Turpin Jocelyn - https://orcid.org/0000-0001-5177-2471	UMR754 "Infections Virales et Pathologie Comparée" - 199517675N	• DMP manager
TURPIN Jocelyn - https://orcid.org/0000-0001-5177-2471	Université Claude Bernard Lyon 1 - 199517675N	• Project coordinator

PGD du projet "REtroViral Emergence @farm"

1. Data description and collection or re-use of existing data

1a. How will new data be collected or produced and/or how will existing data be re-used?

2 main sets of data will be produced :

- We will develop an approach of Ligation-Mediated PCR (LM PCR), a protocol to identify and quantify the integration sites of the retrovirus in the genome of their host. The principle is to amplify virus-host junctions using primers targeting the viral LTR on one side, and the linker on the other side, paired with next generation sequencing (short reads - Illumina) thus enabling to amplify the junction independently of its location on the genome.
- We will sequence the full genome of JSRV and ENTV combining third-generation long-read sequencing technologies (Oxford Nanopore technology or ONT) and a step of enrichment in exogenous genomes using the CRISPR cas 9 methodology or near-full-length proviral PCR amplification.

Over the years, our team has constituted a large biobank of tumors from JSRV, ENTV-1 and ENTV-2 induced cancers. The sequencing experiments described above will be done on DNA extracted from those samples.

No existing data produced by the team will be re-used. Only sequences available on the public databanks (NCBI, UCSC, Ensembl, NGDC...) will be used for the design of molecular tools (LM-PCR and CRISPR) for the enrichment of viral sequences in the library preparation and the analysis of the data.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

ONT:

- Raw sequencing data : fast5 (hierarchical data format 5 (HDF5))
- Basecalled sequencing data : FASTQ (text file)
- Consensus proviral sequences : FASTA (text file)

Illumina

- Raw sequencing data : FASTQ (text file)
- List of viral integration sites in the genomes : text file

Illumina & ONT

- Mapped sequence reads : BAM (binary files)
- Source Code : text file

Volume: 1TB to 10TB - the size can not be estimated more precisely and will depend on the depth of sequencing requested for the proviral sequencing.

---

2. Documentation and data quality

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

The metadata of each sequencing run is standardized and recorded in the team's dedicated database, stored on the laboratory server. For each run the following information are recorded:

- Date
- Name (scientist)
- Type of sequencing

- Sequencer
- Flow-cells' type
- Flowcell lot
- Sequenced samples
- Barcodes
- Library preparation kit
- Library preparation kit lot
- Number of reads
- Number of raw reads

Illumina:

- Cluster density
- Reads passing filter
- Cluster PF
- % > Q30
- PhyX control detected

ONT:

- Mean read length
- Number of starting pores
- QSCORE

2b. What data quality control measures will be used?

Quality control of the sequencing data

FastQC is one of the most common tools for quality control of sequencing data including Illumina and ONT. It will be used to quality-check the sequencing data.

Multiple factors will be notably checked:

- Per Base sequence quality
- Per Sequence Quality scores
- Overrepresented sequences
- Sequence length

Quality control for the preservation of the raw data:

The datasets will be hashed using the program md5sum to create a summary for every folder that assures the quality and integrity of the data each time it needs to be copied or used.

---

3. Storage and backup during the research process

3a. How will data and metadata be stored and backed up during the research?

Immediately after the acquisition the raw sequencing data will be placed on the host laboratory's local server which is managed by Caroline Leroux one of the coordinators of this project and of this DMP.

In parallel, the raw and derived data will be stored on a network-attached storage (NAS) device. This NAS uses RAID to automatically duplicate all files onto 2 independent hard drives.

In addition, the raw data will be stored on an external hard-drive.

Biobank: the samples are anonymised at reception and metadata (clinical data, age, breed, sex, geographic origin) and recorded in the team's dedicated database, stored on the laboratory server.

3b. How will data security and protection of sensitive data be taken care during the research

The raw data is stored in 3 different locations: server, NAS, and external hard drive. The server and NAS are also backed up automatically. If one location fails it can either be recovered from its back up or from the other location.

Access to the NAS and the server is restricted (login access and password) to the two coordinators of this DMP and the PhD student recruited for this project.

---

4. Legal and ethical requirements, code of conduct

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

No personal data will be processed.

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

All the data will be made publicly available (public databases, e.g. European Nucleotide Archive (ENA)).

All research articles will be published in peer-reviewed journals and be open-access.

In line with the French law (n°2016-1321 pour une République numérique du 7 octobre 2016) all data will be made publicly available on open access.

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

Samples from the biobank are collected at necropsy and hence no ethics committee is requested.

The project will be conducted following the INRAE charter of scientific integrity and research ethics.

---

5. Data sharing and long-term preservation

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

The data will be made openly available without restriction at the time of publication.

The raw sequencing datasets and consensus proviral sequences will be uploaded to public databases.

Integration sites' list (text file) will be available with the associated research article.

All the scripts and pipelines, will be available on a git repository.

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

Deposit on public repositories will guarantee sustainability and long-term access to the data.

5c. What methods or software tools are needed to access and use data?

All data correspond to text files, no specific software is required to access them.
All the software to analyse the data are open source.

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

Unique and persistent identifiers will provided once uploaded on the public database (accession number) and by scientific journal/preprint depository (doi).

---

6. Data management responsibilities and resources

6a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

The coordinator of the ANR - REVE@FARM project; Jocelyn Turpin Chargé de Recherche INRAE & the Head of the team Caroline Leroux, Directeur de recherche INRAE. Caroline Leroux is also the INRAE GDPR correspondent for the UMR754 and hence follow the updates and recommendation of INRAE.

6b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

No financial resources are specifically dedicated

At every step of the project time, a meeting is organised every two months with all the participants of the project to discuss the results, the strategies but also data management and ensure that data will be FAIR (Findable, Accessible, Interoperable, Re-usable) for the upcoming steps.