
PGD du projet "EMERGEN : surveillance génomique et recherche sur la COVID-19 et les autres maladies infectieuses émergentes"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "Science Europe: structured template" fourni par Science Europe.

Renseignements sur le plan

Plan title	PGD du projet "EMERGEN : surveillance génomique et recherche sur la COVID-19 et les autres maladies infectieuses émergentes"
Version	First version
Plan purpose/scope	<ol style="list-style-type: none">1. Define the research products and their conditions under which they will be accessed2. Define the data fluxes from their production to the publication and conservation3. Describe the metadata model, and link it with international standards

Fields of science and technology (from OECD classification) Health sciences

Language fra

Creation date 2022-01-19

Last modification date 2023-02-02

Identifiant type DOI

License Creative Commons Attribution Share Alike 4.0 International

Associated documents (publications, reports, patents, experimental plan...), website

- EMERGEN-DB Web site : <https://emergen-db.france-bioinformatique.fr/>
- Project flow chart : https://resana.numerique.gouv.fr/public/document/afficherOnlyOffice?slug=74683&id_information=1553686
- Zenodo community : <https://zenodo.org/communities/emergen/>
- Metadata referential : <https://doi.org/10.5281/zenodo.7145749>

Renseignements sur le projet

Project title EMERGEN : surveillance génomique et recherche sur la COVID-19 et les autres maladies infectieuses émergentes

Acronym EMERGEN

Abstract EMERGEN est le projet français de surveillance génomique COVID-19, initié en janvier 2021 par le Ministère de la Santé et des Solidarités (MSS) et le Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI). Il est co-piloté par Santé publique France et l'ANRS|MIE.

Funding

- Ministère de la Santé et des Solidarités (MSS) :
- Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI) :

Start date 2021-01-01

End date 2022-12-31

Partners

- Santé publique France ()
- Agence Nationale de Recherche sur le SIDA et les Maladies Infectieuses Emergentes ()
- Centre National de la Recherche Scientifique ()
- Institut Français de Bioinformatique (201321719F)
- institut Pasteur ()

Produits de recherche :

1. le premier produit ne peut pas être effacé
2. Metadata collected by sampling laboratories (personal or technical metadata) (Dataset)
3. Technical metadata characterising the sequencing process (platform, technology, primers, ...)
4. Specification sheet for all the metadata circulating in EMERGEN data flow (Model)
5. Data resulting from the sequencing of SARS-CoV-2 genome, including raw reads, aligned reads and consensus genomes (Dataset)
6. Annotation of mutations and variants resulting from the analysis of viral sequences
7. System-level workflows to handle the data flows on the cluster facility of the EMERGEN-Bioinfo platform
8. Galaxy workflows
9. Bioinformatics platform to handle all the non-sensitive data produced by EMERGEN
10. Bioinformatics platform certified for Health Data Storage and treatment (HDS) used to handle the sensitive data produced by EMERGEN and ensure pairing with health data from other sources
11. Code of EMERGEN-DB, the database for sequences and metadata of viral genomes (Software)

Contributeurs

Name	Affiliation	Rôles
Chiapello Hélène	INRAE	

Name	Affiliation	Rôles
de Lamotte Frederic - https://orcid.org/0000-0003-4234-1172	INRAE	<ul style="list-style-type: none"> • Data contact (Sequencing metadata, EMERGEN-bioinfo, Airflow workflows, Variant metadata, EMERGEN-HDS)
Denecker Thomas - https://orcid.org/0000-0003-1421-7641	IFB - Institut Français de Bioinformatique	<ul style="list-style-type: none"> • Data contact (EMERGEN-DB software) • Data producers or collector (Metadata referential) • Person in charge of long term data preservation (Viral sequence data)
Le Corguillé Gildas - https://orcid.org/0000-0003-1742-9711	IFB - Institut Français de Bioinformatique	<ul style="list-style-type: none"> • Data contact (rien) • Data producers or collector (Galaxy Workflows)
Lieby Paulette	IFB Institut Français de Bioinformatique	
Médigue Claudine - https://orcid.org/0000-0002-3905-1054	CNRS	
Naouar Naira - https://orcid.org/0000-0003-2161-8603	Sorbonne Université	<ul style="list-style-type: none"> • Data contact (Galaxy Workflows) • Data producers or collector (Galaxy Workflows)
OUAHIOUNE Hamid	IFB - Institut Français de Bioinformatique	
Salgado David - https://orcid.org/0000-0002-5905-3591	IFB - Institut Français de Bioinformatique	<ul style="list-style-type: none"> • Data documentation manager (Sampling metadata) • Data producers or collector (Metadata referential)
van Helden Jacques - https://orcid.org/0000-0002-8799-8584	IFB - Institut Français de Bioinformatique	<ul style="list-style-type: none"> • DMP manager • Data contact (Metadata referential, Sampling metadata, Viral sequence data) • Data documentation manager (Sampling metadata) • Data producers or collector (Metadata referential) • Person in charge of long term data preservation (Viral sequence data) • Project coordinator

Budget

Type de coût (Titre)	Montant	Etape du cycle de vie
Personnel cost (Workflow developer salary)	80000 EUR	Data collection/production associated costs - Galaxy Workflows

Droits d'auteur :

Le(s) créateur(s) de ce plan accepte(nt) que tout ou partie de texte de ce plan soit réutilisé et personnalisé si nécessaire pour un autre plan. Vous n'avez pas besoin de citer le(s) créateur(s) en tant que source. L'utilisation de toute partie de texte de ce plan n'implique pas que le(s) créateur(s) soutien(nen)t ou aient une quelconque relation avec votre projet ou votre soumission.

PGD du projet "EMERGEN : surveillance génomique et recherche sur la COVID-19 et les autres maladies infectieuses émergentes"

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

le premier produit ne peut pas être effacé

1.1 Research output description

Name	le premier produit ne peut pas être effacé
Description	<p>EMERGEN-Bioinfo is a numerical platform including the hardware and software components used to handle viral genomic sequences produced by the EMERGEN consortium.</p> <p>It includes</p> <ol style="list-style-type: none">1. A cluster facility to ensure the compute and storage tasks2. A Galaxy server (covid19.usegalaxy.fr) enabling to analyse the NGS sequences with the workflows3. The EMERGEN-DB database (emergen-db.france-bioinformatique.fr)
Workpackage	WP3
Keywords (free-text)	compute & storage; cluster
May contain personal data?	No
May contain sensible data?	No
May take ethical issues into account?	Yes

1.2 Will existing data be reused?

1.3 How new data will be collected or produced?

Metadata collected by sampling laboratories (personal or technical metadata)

1.1 Research output description

Name	Metadata collected by sampling laboratories (personal or technical metadata)
Type	Dataset
Workpackage	WP1
Keywords (free-text)	
Language	fra
Issued Date	2022-12-31
May contain sensible data?	Yes
May take ethical issues into account?	Yes

1.2 Will existing data be reused?

1.3 How new data will be collected or produced?

Title of the method	Sampling metadata collection
Description	The identifying data is collected by the national health system. Additional metadata is collected by the sequencing labs by asking patients to fill in a specific form.
Data Nature	Observation

Technical metadata characterising the sequencing process (platform, technology, primers, ...)

1.1 Research output description

Name	Technical metadata characterising the sequencing process (platform, technology, primers, ...)
-------------	---

1.2 Will existing data be reused?

Question sans réponse.

1.3 How new data will be collected or produced?

Question sans réponse.

Specification sheet for all the metadata circulating in EMERGEN data flow

1.1 Research output description

Name	Specification sheet for all the metadata circulating in EMERGEN data flow
Description	A spreadsheet providing a description of all the fields handled at each step of the EMERGEN data flow.
Type	Model
Workpackage	WP3
Keywords (free-text)	data model; metadata specification
Language	fra
Identifier type	DOI
May contain personal data?	No
May contain sensible data?	No
May take ethical issues into account?	Yes

1.2 Will existing data be reused?

Justification	<p>The EMERGEN metadata referential is built by the EMERGEN partners for the specific needs of the project, but the information gathered is partly driven by the requirements of the international repositories where the data will be published (GISAID and ENA).</p> <p>The EMERGEN metadata referential thus includes a mapping from EMERGEN to GISAID and ENA, which relies on the respective data models of these international repositories.</p>
----------------------	--

1.3 How new data will be collected or produced?

Title of the method	Expert curation
Description	<p>The metadata referential is created by the following partners of the project :</p> <ol style="list-style-type: none">1. IFB - Institut Français de Bioinformatique2. SpF - Santé publique France3. CNR - Centre National de Référence des virus respiratoires
Data Nature	Model
Related references	<ul style="list-style-type: none">• Spreadsheet capture :

Data resulting from the sequencing of SARS-CoV-2 genome, including raw reads, aligned reads and consensus genomes

1.1 Research output description

Name	Data resulting from the sequencing of SARS-CoV-2 genome, including raw reads, aligned reads and consensus genomes
Description	Data resulting from SARS-CoV-2 genome sequencing of patient samples. This includes different data types and formats obtained during the analysis workflow : <ul style="list-style-type: none"> • raw reads (fastq.gz format) • mapped reads (bam format) • consensus genomes (fasta format) • variant call results (vcf files) • variant nomenclature assigned by nextclade and pangolin softwares
Type	Dataset
Workpackage	WP2
Keywords (free-text)	
Language	eng
Identifier type	local identifier

1.2 Will existing data be reused?

Justification The SARS-CoV-2 reference genome is used for the read mapping step of the analysis workflows.

1.3 How new data will be collected or produced?

Title of the method Amplicon-based Next Generation Sequencing

Description Viral genome is amplified by PCR and the amplicons are sequenced by Next Generation Sequencing (either short read or long read technologies depending on the cases).

Data Nature Experimental Data

Annotation of mutations and variants resulting from the analysis of viral sequences

1.1 Research output description

Name Annotation of mutations and variants resulting from the analysis of viral sequences

1.2 Will existing data be reused?

Question sans réponse.

1.3 How new data will be collected or produced?

Question sans réponse.

System-level workflows to handle the data flows on the cluster facility of the EMERGEN-Bioinfo platform

1.1 Research output description

Name System-level workflows to handle the data flows on the cluster facility of the EMERGEN-Bioinfo platform

1.2 Will existing data be reused?

Question sans réponse.

1.3 How new data will be collected or produced?

Question sans réponse.

Galaxy workflows

1.1 Research output description

Name Galaxy workflows

Description The Galaxy software environment is used to run the analysis of NGS data produced by the sequencing labs of the EMERGEN consortium.

Workpackage WP3

Keywords (free-text) galaxy, airflow, workflow

Language eng

Identifier type DOI

May contain personal data? No

May contain sensible data? No

May take ethical issues into account? Unknown

1.2 Will existing data be reused?

Justification

Part of these workflows have been developed by third parties from the European ELIXIR bioinformatics infrastructure , and imported from the [COVID19 Workflowhub](#) to be deployed the French COVID-19 Galaxy server ([covid19.usegalaxy.fr](#)). These workflows have been adapted to the specific configurations of the different sequence-producing labs. Other custom workflows will be developed and deposited on the [COVID19 Workflowhub](#).

1.3 How new data will be collected or produced?**Description**

New workflows will be developed and deposited on the [COVID19 Workflowhub](#). Some of the imported workflows may be adapted to face the specific configurations of some sequencing laboratories, and the adapted versions may be deposited on the [COVID19 Workflowhub](#) if relevant.

Costs

- Personnel cost : 80 000 EUR

Bioinformatics platform to handle all the non-sensitive data produced by EMERGEN**1.1 Research output description****Name**

Bioinformatics platform to handle all the non-sensitive data produced by EMERGEN

1.2 Will existing data be reused?

Question sans réponse.

1.3 How new data will be collected or produced?

Question sans réponse.

Bioinformatics platform certified for Health Data Storage and treatment (HDS) used to handle the sensitive data produced by EMERGEN and ensure pairing with health data from other sources**1.1 Research output description**

Name Bioinformatics platform certified for Health Data Storage and treatment (HDS) used to handle the sensitive data produced by EMERGEN and ensure pairing with health data from other sources

1.2 Will existing data be reused?

Question sans réponse.

1.3 How new data will be collected or produced?

Question sans réponse.

Code of EMERGEN-DB, the database for sequences and metadata of viral genomes

1.1 Research output description

Name Code of EMERGEN-DB, the database for sequences and metadata of viral genomes

Description The EMERGEN-DB database enables to collect, handle, access and visualise the viral genomic sequences and associated metadata generated by EMERGEN, the French consortium for COVID-19 genomic surveillance and research.
The software is implemented in the Django framework (python) with an underlying PostgreSQL database. The database can be accessed via a graphical user-friendly interface (GUI) and an application programmatic interface (API, REST).

Type Software

Workpackage WP3

Keywords (free-text)

Language eng

Identifier type url

May contain personal data? no

May contain sensible data? no

May take ethical issues into account? yes

1.2 Will existing data be reused?

1.3 How new data will be collected or produced?

Question sans réponse.

2. Documentation and data quality

le premier produit ne peut pas être effacé

2.1 What metadata and documentation (for example way of organising data) will accompany the data?

Question sans réponse.

2.2 What methods will be used to ensure their scientific quality?

Question sans réponse.

Metadata collected by sampling laboratories (personal or technical metadata)

2.1 What metadata and documentation (for example way of organising data) will accompany the data?

Description

The contents, format and destination of each field of the sampling metadata are described in the EMERGEN metadata referential (<https://doi.org/10.5281/zenodo.7145748>).

Related references

- EMERGEN metadata referential : <https://doi.org/10.5281/zenodo.7145748>

2.2 What methods will be used to ensure their scientific quality?

Technical metadata characterising the sequencing process (platform, technology, primers, ...)

2.1 What metadata and documentation (for example way of organising data) will accompany the data?

2.2 What methods will be used to ensure their scientific quality?

Question sans réponse.

Specification sheet for all the metadata circulating in EMERGEN data flow

2.1 What metadata and documentation (for example way of organising data) will accompany the data?

2.2 What methods will be used to ensure their scientific quality?

Question sans réponse.

Data resulting from the sequencing of SARS-CoV-2 genome, including raw reads, aligned reads and consensus genomes

2.1 What metadata and documentation (for example way of organising data) will accompany the data?

Description

The sequencing metadata is produced by the sequencing laboratories and transferred to EMERGEN-DB via a structured template file (xlsx format).

The sequencing metadata fields are described in the EMERGEN metadata referential

List of sequencing metadata fields :

1. Sequencing stream
2. Sample number from the sequencing lab
3. Name of the sequencing laboratory
4. Geographical FINESS of the sequencing laboratory
5. Date of receipt of the sample
6. Sequencing result date
7. Sequencing technology
8. Sequencing length
9. Sequence header in the fasta file
10. Sequencing platform
11. Other sequencing platform
12. Sequencing instrument
13. Other sequencing instrument
14. Library selection
15. Library source
16. Library strategy
17. Library layout
18. Amplification kit
19. Nominal length
20. Nominal sdev
21. Library construction protocol
22. Assembly method
23. Read length
24. Average coverage

Related references

- EMERGEN-DB metadata referential : referentiel_metadonnees
- EMERGEN-DB metadata collection template : fichier_trame

Metadata/data standards

- Genome Metadata : <https://rdamsc.bath.ac.uk/msc/m19>

Metadata language code

eng

2.2 What methods will be used to ensure their scientific quality?**Annotation of mutations and variants resulting from the analysis of viral sequences****2.1 What metadata and documentation (for example way of organising data) will accompany the data?**

Question sans réponse.

2.2 What methods will be used to ensure their scientific quality?

Question sans réponse.

System-level workflows to handle the data flows on the cluster facility of the EMERGEN-Bioinfo platform

2.1 What metadata and documentation (for example way of organising data) will accompany the data?

Question sans réponse.

2.2 What methods will be used to ensure their scientific quality?

Question sans réponse.

Galaxy workflows

2.1 What metadata and documentation (for example way of organising data) will accompany the data?

2.2 What methods will be used to ensure their scientific quality?

Bioinformatics platform to handle all the non-sensitive data produced by EMERGEN

2.1 What metadata and documentation (for example way of organising data) will accompany the data?

Question sans réponse.

2.2 What methods will be used to ensure their scientific quality?

Question sans réponse.

Bioinformatics platform certified for Health Data Storage and treatment (HDS) used to handle the sensitive data produced by EMERGEN and ensure pairing with health data from other sources

2.1 What metadata and documentation (for example way of organising data) will accompany the data?

Question sans réponse.

2.2 What methods will be used to ensure their scientific quality?

Question sans réponse.

Code of EMERGEN-DB, the database for sequences and metadata of viral genomes

2.1 What metadata and documentation (for example way of organising data) will accompany the data?

Question sans réponse.

2.2 What methods will be used to ensure their scientific quality?

Question sans réponse.

3. Legal and ethical requirements, codes of conduct

le premier produit ne peut pas être effacé

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

3.2 How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is

applicable?

Question sans réponse.

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

Question sans réponse.

Metadata collected by sampling laboratories (personal or technical metadata)

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

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

Question sans réponse.

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

Question sans réponse.

Technical metadata characterising the sequencing process (platform, technology, primers, ...)

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

Question sans réponse.

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

Question sans réponse.

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

Question sans réponse.

Specification sheet for all the metadata circulating in EMERGEN data flow

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

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

Description

The metadata referential will be published in open access on the Zenodo repository.

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

Question sans réponse.

Data resulting from the sequencing of SARS-CoV-2 genome, including raw reads, aligned reads and consensus genomes

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

Description

The viral sequences are not personal, they do not present any risk of exposing either the identity of the patients or health conditions.

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

Description

An important issue for sequencing laboratories is the exploitation of the sequence data for scientific publications. On the other hand, since this data is produced from a public funding, the 2016 French law "République numérique" makes it mandatory to open this data.

3.3 What ethical issues and codes of conduct are there, and how will they be taken into account?**Description**

An ethical issue is to share viral sequences without delay in order to enable international organisations to follow the evolution of the virus, detect new variants, raise international alerts etc.

The data is shared without delay on GISAID, but this provides a restricted access requiring an authorised access and under a protective license.

The data should also be opened via its submission to the European Nucleotide Access, which leaves open the possibility to delay the opening (embargo period).

Annotation of mutations and variants resulting from the analysis of viral sequences**3.1 If personal data are processed, how will compliance with legislation on personal data and on security be ensured?**

Question sans réponse.

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

Question sans réponse.

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

Question sans réponse.

System-level workflows to handle the data flows on the cluster facility of the EMERGEN-Bioinfo platform**3.1 If personal data are processed, how will compliance with legislation on personal data and on security be ensured?**

Question sans réponse.

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

Question sans réponse.

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

Question sans réponse.

Galaxy workflows

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

Question sans réponse.

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

Question sans réponse.

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

Question sans réponse.

Bioinformatics platform to handle all the non-sensitive data produced by EMERGEN

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

Question sans réponse.

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

Question sans réponse.

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

Question sans réponse.

Bioinformatics platform certified for Health Data Storage and treatment (HDS) used to handle the sensitive data produced by EMERGEN and ensure pairing with health data from other sources

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

Question sans réponse.

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

Question sans réponse.

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

Question sans réponse.

Code of EMERGEN-DB, the database for sequences and metadata of viral genomes

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

Question sans réponse.

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

Question sans réponse.

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

Question sans réponse.

4. Data processing and analysis

le premier produit ne peut pas être effacé

4.1 How and with what resources will the data be processed / analyzed?

Metadata collected by sampling laboratories (personal or technical metadata)

4.1 How and with what resources will the data be processed / analyzed?

Question sans réponse.

Technical metadata characterising the sequencing process (platform, technology, primers, ...)

4.1 How and with what resources will the data be processed / analyzed?

Question sans réponse.

Specification sheet for all the metadata circulating in EMERGEN data flow

4.1 How and with what resources will the data be processed / analyzed?

Question sans réponse.

Data resulting from the sequencing of SARS-CoV-2 genome, including raw reads, aligned reads and consensus genomes

4.1 How and with what resources will the data be processed / analyzed?

Question sans réponse.

Annotation of mutations and variants resulting from the analysis of viral sequences

4.1 How and with what resources will the data be processed / analyzed?

Question sans réponse.

System-level workflows to handle the data flows on the cluster facility of the EMERGEN-Bioinfo platform

4.1 How and with what resources will the data be processed / analyzed?

Question sans réponse.

Galaxy workflows

4.1 How and with what resources will the data be processed / analyzed?

Question sans réponse.

Bioinformatics platform to handle all the non-sensitive data produced by EMERGEN

4.1 How and with what resources will the data be processed / analyzed?

Question sans réponse.

Bioinformatics platform certified for Health Data Storage and treatment (HDS) used to handle the sensitive data produced by EMERGEN and ensure pairing with health data from other sources

4.1 How and with what resources will the data be processed / analyzed?

Question sans réponse.

Code of EMERGEN-DB, the database for sequences and metadata of viral genomes

4.1 How and with what resources will the data be processed / analyzed?

Question sans réponse.

5. Storage and backup during the research process

le premier produit ne peut pas être effacé

5.1 How will data be stored and backed up during the research?

Metadata collected by sampling laboratories (personal or technical metadata)

5.1 How will data be stored and backed up during the research?

Question sans réponse.

Technical metadata characterising the sequencing process (platform, technology, primers, ...)

5.1 How will data be stored and backed up during the research?

Question sans réponse.

Specification sheet for all the metadata circulating in EMERGEN data flow

5.1 How will data be stored and backed up during the research?

Question sans réponse.

Data resulting from the sequencing of SARS-CoV-2 genome, including raw reads, aligned reads and consensus genomes

5.1 How will data be stored and backed up during the research?

Question sans réponse.

Annotation of mutations and variants resulting from the analysis of viral sequences

5.1 How will data be stored and backed up during the research?

Question sans réponse.

System-level workflows to handle the data flows on the cluster facility of the EMERGEN-Bioinfo platform

5.1 How will data be stored and backed up during the research?

Question sans réponse.

Galaxy workflows

5.1 How will data be stored and backed up during the research?

Question sans réponse.

Bioinformatics platform to handle all the non-sensitive data produced by EMERGEN

5.1 How will data be stored and backed up during the research?

Question sans réponse.

Bioinformatics platform certified for Health Data Storage and treatment (HDS) used to handle the sensitive data produced by EMERGEN and ensure pairing with health data

from other sources

5.1 How will data be stored and backed up during the research?

Question sans réponse.

Code of EMERGEN-DB, the database for sequences and metadata of viral genomes

5.1 How will data be stored and backed up during the research?

Question sans réponse.

6. Data sharing and long-term preservation

le premier produit ne peut pas être effacé

6.1 How will data be shared?

Question sans réponse.

6.2 How will data be long-term preserved? Which data?

Question sans réponse.

Metadata collected by sampling laboratories (personal or technical metadata)

6.1 How will data be shared?

Question sans réponse.

6.2 How will data be long-term preserved? Which data?

Question sans réponse.

Technical metadata characterising the sequencing process (platform, technology, primers, ...)

6.1 How will data be shared?

Question sans réponse.

6.2 How will data be long-term preserved? Which data?

Question sans réponse.

Specification sheet for all the metadata circulating in EMERGEN data flow

6.1 How will data be shared?

Question sans réponse.

6.2 How will data be long-term preserved? Which data?

Question sans réponse.

Data resulting from the sequencing of SARS-CoV-2 genome, including raw reads,

aligned reads and consensus genomes

6.1 How will data be shared?

Modalities of sharing	<p>The data will be reused with different modalities in different contexts.</p> <ul style="list-style-type: none">• consensus genomes (fasta) are published on GISAID and ENA international repositories• raw sequences will be published at EBI-ENA (www.ebi.ac.uk/ena)• the full sequence data will be available to researchers on the EMERGEN-bioinfo digital platform operated by the Institut Français de Bioinformatique, upon request to the data access committee <p>GISAID ensures a fast sharing of the consensus genomes for the international community of virologists, under protective license. It does not include the raw sequences.</p> <p>ENA provides open access to both the raw sequences and consensus genomes. The French law specifies that all data produced by public funding must be open.</p> <p>The EMERGEN-bioinfo digital platform provides access to the data in a computing environment adapted to perform advanced analyses. It also enables to access, under a protective license, personal metadata that cannot be disclosed on public repositories.</p>
Reusability	reuse for genomic surveillance and for future research
Data repository/catalogs	<ul style="list-style-type: none">• European Nucleotide Archive (EBI-ENA) : https://www.ebi.ac.uk/ena/• GISAID : https://www.gisaid.org/• IFB core cluster : core.cluster.france-bioinformatique.fr• EMERGEN-DB : https://emergen-db.france-bioinformatique.fr/

6.2 How will data be long-term preserved? Which data?

Justification	The long-term preservation will be ensured by the European Nucleotide Archive (EBI-ENA).
Estimated volume of data	2
Unit	PB
Start date	2021-01-01
End date	2022-12-31
Archive	:

Annotation of mutations and variants resulting from the analysis of viral sequences

6.1 How will data be shared?

Question sans réponse.

6.2 How will data be long-term preserved? Which data?

Question sans réponse.

System-level workflows to handle the data flows on the cluster facility of the EMERGEN-Bioinfo platform

6.1 How will data be shared?

Question sans réponse.

6.2 How will data be long-term preserved? Which data?

Question sans réponse.

Galaxy workflows

6.1 How will data be shared?

Question sans réponse.

6.2 How will data be long-term preserved? Which data?

Question sans réponse.

Bioinformatics platform to handle all the non-sensitive data produced by EMERGEN

6.1 How will data be shared?

Question sans réponse.

6.2 How will data be long-term preserved? Which data?

Question sans réponse.

Bioinformatics platform certified for Health Data Storage and treatment (HDS) used to handle the sensitive data produced by EMERGEN and ensure pairing with health data from other sources

6.1 How will data be shared?

Question sans réponse.

6.2 How will data be long-term preserved? Which data?

Question sans réponse.

Code of EMERGEN-DB, the database for sequences and metadata of viral genomes

6.1 How will data be shared?

6.2 How will data be long-term preserved? Which data?

Archive

: