SeqOccIn

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

Créateurs du PGD : Cécile Donnadieu, Gérald Salin, celine.noirot@inra.fr

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

Modèle du PGD : Horizon 2020 FAIR DMP (anglais)

Dernière modification du PGD : 12/03/2021

Financeur : FEDER - Région Occitanie

Numéro de subvention : Programme opérationnel FEDER-FSE Midi-Pyrénées et Garonne 2014-2020, Occitanie

Résumé du projet :

En complément des technologies de séquençage de courts fragments (150 nucléotides) à très grande échelle (plusieurs milliards de fragments simultanément), la disponibilité de technologies permettant le séquençage de longs fragments (plusieurs dizaines voire centaines de milliers de nucléotides) a ouvert de nouveaux horizons dans le domaine de l'analyse des génomes. Ces technologies longs fragments permettent de traverser les régions répétées pour un assemblage de grande qualité des génomes complexes. Elles permettent aussi le séquençage de molécules natives non amplifiées donnant aussi accès à certaines marques épigénétiques. Dans ce contexte très évolutif, SeqOccIn vise à acquérir une expertise majeure sur l'utilisation et la complémentarité optimale des différentes technologies de séquençage haut débit en fonction des objectifs de l'utilisateur, objectifs qui peuvent consister en l'acquisition d'une information succincte à faible coût ou dans d'autres cas nécessiter une information de très haute qualité. SeqOccIn vise ainsi à acquérir l'expertise pour combiner au mieux ces technologies de séquençage courts et longs fragments dans 3 axes complémentaires : 1. l'analyse de la variabilité structurale des génomes, 2. l'analyse des marques épigénétiques, 3. l'analyse des méta-génomes. Outre les plateformes GeT et Bioinfo de Genotoul, SeqOccIn regroupe 25 partenaires de la recherche publique et privée travaillant sur les espèces animales, végétales et microbiennes.

Chercheur Principal : Cécile Donnadieu

Identifiant ORCID : 0000-0002-5041-6843

Contact pour les Données : Cécile Donnadieu

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.

SeqOccIn

1. Data summary

The SeqOccIn project (Séquençage Occitanie Innovation) aims to acquire an expertise on long fragment sequencing technologies in three study areas:

- genome polymorphism (source of genetic diversity),
- epigenome (mark partially reversible carried by the DNA modifying the genes expression),
- the metagenome (particularly bacterial communities present in a complex environment such as the intestine, feces, a food, etc.).

This DMP describes how the raw data, generated by this project, will be managed.
A large number of genomic sequences will be produced and stored into nG6 system (ng6.toulouse.inra.fr). Those data are formatted in fastq file and fast5, depending on the sequencer.

- origin of the data ? outputs of sequencer from biological samples provided by biologists.
- types and formats?
  - Illumina (MiSeq, HiSeq, NovaSeq) : fastq.gz
  - ONT (GridIon, PromethIon) : fast5 + fastq.gz
  - PacBio (SequelII) : bam
- the expected size:

Estimated amount of produced data (To):

Year	**Axis1**	**Axis2**	**Axis3**	**Total per year**
2019	49	11	6	65
2020	63	46	0	109
2021	355	54	33	442
Total	467	110	38	616

- to whom will it be useful ? all members of the project and later for the scientific community.

We identify several kind of persons:
* core team: platforms
* scientific teams: members of the projects out of the platforms
* scientific community

This DMP mainly describes raw data management. But other kind of data are produced. Here is a summary of how we intend to handle all the data.

	Sample data (bio)	Biological protocol	Raw data (bioinfo)	Processed data (bioinfo)	Bioinformatic guidelines (bioinfo)	Source code (bioinfo)
	sample description and quality of sample	definition of new protocol		The raw data are processed and this will generate new data.	knowledge of best practice to process data	software and workflow development and scripts
	Generated data	Generated data	Generated data	Generated data	Generated data	Generated data
Origin	Experimentation	Analysis	Experimentation	Analysis	Analysis	Code
Type of data	Text / Graphics	Text	Dataset / fastq,fast5	Dataset / bam,vcf,...	Text / Guideline	Software/workflow
During the project						
Findable	N/A	N/A	SRA metadata / Ng6 search	each axis defines where data are stored and findable	search keyword into forge mia	search keyword into forge mia
Accessible	Sharepoint	Lab book	Ng6 / SeqOccIn storage space	Ng6 / SeqOccIn storage space	Forge mia	Forge mia
Accessible to who ?	core team	core team	core team and after validation scientific team	core team and after validation scientific team	core team	core team
Interoperable	yes: SRA/FAANG sheet	N/A	Standard format (fastq, fast5, xsl/xml SRA)	Standard format (vcf, bam ...)	N/A	?
Re-usable	N/A	Yes	Yes	N/A	N/A	Licence open for reuse
End of project						
Findable	N/A	publication / get website ?	SRA metadata / Ng6 search	each axis define where data are stored and findable	search keyword into forge mia	search keyword into forge mia
Accessible	This is rawdata metadata, they will be accessible in ENA.	where is published	Public in ENA	Can be public depending on publication requirements	Forge mia	Forge mia
Interoperable	N/A	N/A	Yes	Standard format (vcf, bam ...)	N/A	?
Accessible to who ?	core team	Scientific community	Scientific community	core team+ scientific team	Scientific community	Scientific community
Re-usable	N/A	Yes	Yes	Yes: Public and standard format	Command lines reusable	Licence open for reuse

2. FAIR data

Each project is named with the following nomenclature : axis, step, team and species. Eg : A1-P1-PlaGe-Bovin. A project can hold different sequencing runs. For each run, the metadata files have to be filled :

- The BioSample file describe the biological source materials used in experimental assays.https://submit.ncbi.nlm.nih.gov/biosample/template/
- The SRA biosample file : ftp://ftp-trace.ncbi.nlm.nih.gov/sra/metadata_table/SRA_metadata_acc.xlsx

Those templates will be stored into the SeqOccIn Sharepoint site and will be used for data submission into a public repository.
Some metadata such as sequencer, species, number of reads are directly stored into nG6 web interface (ng6.toulouse.inra.fr) and accessible to the members of the project. At the end of the seqoccin project, those data will be available into public repository.
DOI: yes in SRA at the end of the project.

Fastq files will be deposited in an international archive such as SRA or ENA, at most one year after the end of the project (december 2022).
As soon as the raw data is deposited in a public repository (SRA or ENA), data will be searchable with over all metadata information and publically accessible.
During the project the data are only available for members of the project in nG6 system.

All those data are in standard format such as fastq, fast5, bam or vcf, by definition is re-usable by any bioinformatics software.

Question sans réponse.

3. Allocation of resources

A storage facility was bought for the project by SeqOccIn funding.
Storing raw data into the public Archive (eg SRA or ENA) is free. If this services became paying, list of data to keep will be redefined.

4. Data security

Raw data will be stored into a replicated storage facility.
During the project, raw data will be made accessible to concerned project members through nG6 web interface (ssl, personal credentials).
On the server where data are accessible for the computing, data are protected thanks to unix file rights. Each user or file owner is responsible of right he defines.

5. Ethical aspects

Question sans réponse.

6. Other

Question sans réponse.