MISE EN PLACE DE GENEULIKE : UN ESPACE DE DÉPÔT POUR DES LISTES D'ENTITÉS BIOLOGIQUES

Clément LANCIEN

Rapport de stage de Master 1 Bio-Informatique et Génomique de l'Université de Rennes 1.
Stage effectué au sein de l'Inserm U1085-Irset

Encadrants : Frédéric CHALMEL et Thomas DARDE

Membres de jury : Christian DELAMARCHE et Fouzia MOUSSOUNI

SOUTENUE LE
27 JUILLET 2017

UMR CNRS 6290
Institut de Génétique et Développement de Rennes (IGDR)

Annabelle MONNIER
Equipe Génomique Fonctionnelle Intégrée et Biomarqueurs

UNIVERSITÉ DE RENNES 1

ENGAGEMENT DE NON PLAGIAT

Je, soussigné(e) ……..... étudiant(e) en…….. déclare être pleinement informé que le plagiat de documents ou d'une partie de document publiés sur toute forme de support, y comprit l'internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Date :

Signature :

Document à compléter de manière manuscrite et à insérer obligatoirement en première page du rapport de stage.

Laboratoire de Génomique Médicale
BMT-HC - CHU Pontchaillou

2 rue Henri le Guilloux
35033 Rennes Cedex
FRANCE

Annabelle MONNIER
annabelle.monnier@univ-rennes1.fr
TÉL. 33 (0)2 99 28 92 54

Remerciements

Je tiens tout d'abord à remercier le Docteur Bernard JÉGOU et le Docteur Nathalie DEJUCQ-RAINSFORD de m'avoir accepté au sein de l'Inserm U1085-Irset et plus particulièrement au sein de l'équipe 8.

Je tiens vivement à remercier mes maîtres de stage, le Docteur Frédéric CHALMEL et Thomas DARDE, pour leur disponibilité, même le week-end, ainsi que leur aide pour la réalisation de ce mémoire. Je remercie également Thomas de prendre le temps de m'expliquer certaines notions informatiques malgré sa rédaction de thèse, et pour la formation GoDocker.

Je remercie Angélique BRUNOT pour son aide précieuse sur les marges des zones de texte de ma présentation au sein de l'Irset.

Enfin je tiens à remercie toute l'équipe 8 pour leur accueil, leur bonne humeur et pour la sortie *EscapeGame*.

Sommaire

I. Introduction

I.1 Contexte Scientifique

À la fin des années 1990, l'émergence des puces à ADN puis des technologies de séquençages à ultra-débit à partir du milieu des années 2000 a fait émerger de nouvelles questions relatives à l'accumulation des grands volumes de données en biologie (Barrett and Edgar 2006; Reimand et al. 2016). Cette évolution quasi-exponentielle (Barrett et al. 2011) a conduit plusieurs grandes institutions, telles que la *European Bioinformatics Institute* (EBI, http://www.ebi.ac.uk) et le *National Center for Biotechnology Information* (NCBI, https://www.ncbi.nlm.nih.gov) à développer de nouveaux outils de stockage, aussi appelés espaces de dépôts (ou *repositories* en anglais), qui permettent de centraliser et organiser les données « omiques », telles que transcript-, proté-, gén-, métabol-omiques. Ainsi, ces *repositories* ont pour objectifs de stocker de manière pérenne les informations générées par les chercheurs et de les rendre accessible à la communauté scientifique. Avant de soumettre une publication, il est généralement imposé aux auteurs de déposer les données brutes ou prétraitées (normalisées par exemple) au sein de ces espaces de dépôts.

Contrairement aux banques et *repositories* « spécialisés » qui ont pour vocation de centraliser les données d'un même sujet biologique, tels que PubChem (Kim et al. 2016) pour les substances chimiques, OpenfMRI (Poldrack et al. 2013) en neuroscience ou encore BioModels Database (Chelliah, Laibe, and Le Novère 2013) pour les modèles biologiques, certains espaces de dépôts sont qualifiés de « généralistes ». Ceux-ci ont pour fonction le stockage de données brutes sans aucune restriction sur la thématique étudiée. Plusieurs *repositories* sont ainsi disponibles selon les différentes disciplines « omiques ». Parmi eux, GEO (*Gene Expression Omnibus*) (Edgar, Domrachev, and Lash 2002) et ArrayExpress (Parkinson et al. 2005) font figure de précurseurs et de références dans le domaine. Alors que GEO est développé et maintenu par le NCBI, ArrayExpress dépend de l'EBI. Ces deux *repositories* permettent aux scientifiques de déposer leurs données brutes de transcriptomique et génomique générées issues d'expériences de puces à ADN et de séquençage à haut-débit (RNA-seq ou ChIP-seq par exemple). Ils fournissent également un certain nombre d'outils d'analyse bio-informatique permettant aux utilisateurs, par exemple, de consulter le profil d'expression de gènes d'intérêts. A l'instar de ces deux exemples, PRIDE (*PRoteomics Identifications Database*) est un des espaces de dépôts de référence pour la protéomique et notamment pour le stockage des données de spectrométrie de masse (Vizcaíno et

al. 2016). Créée en 2004 par l'EBI, PRIDE permet de stocker des séquences protéiques et peptidiques ainsi que des modifications post-traductionnelles.

Bien que développés initialement pour les biologistes, ces espaces de dépôts présentent néanmoins une restriction majeure : l'exploitation des données qu'ils contiennent nécessitent des compétences bio-informatiques et statistiques avancées. Ce qui constitue un frein pour la plupart des chercheurs non-génomistes. De plus, dans la majorité des cas, ceux-ci ne souhaitent pas ré-analyser les données brutes disponible mais désirent « simplement » comparer leurs propres résultats issus d'analyses « omiques » avec ceux publiés par leurs confrères. Ces résultats se présentent habituellement sous la forme de listes d'entités biologiques (gènes, transcrits, protéines, sondes nucléotidiques, …) générées par les différentes étapes de filtration des données décrites dans les publications. Ainsi, ces listes peuvent représenter par exemple, des listes de sondes nucléotidiques d'une puce à ADN sur- ou sous-exprimées, ou encore une liste de protéines différentielles identifiées par spectrométrie de masse (LC-MS/MS par exemple) entre deux conditions expérimentales à comparer. Bien que ces listes d'entités biologiques constituent le cœur même des résultats présentés dans les publications scientifiques, celles-ci sont la plupart du temps disponibles dans les publications scientifiques sous la forme de tableaux souvent perdus dans des fichiers supplémentaires dans des formats (PDF par exemple) difficilement exploitables à grande échelle. Certaines banques de données, telles que MSigDB (*Molecular Signature DataBase*, http://software.broadinstitute.org/gsea/msigdb) développé par le Broad Institut (Subramanian et al. 2005), permettent d'héberger des listes de gènes annotés (ou *gene sets* en anglais). Néanmoins, ces banques ne constituent pas des espaces de dépôts en ce sens qu'elles ne permettent pas aux utilisateurs de déposer eux-mêmes leurs listes d'entités biologiques. Ainsi, contrairement aux données brutes, aucun *repository* n'est dédié au stockage de ces informations précieuses décrivant les différentes étapes du processus d'analyse et des listes d'entités biologiques résultantes. C'est face à ce constat que mon encadrant de stage, Frédéric CHALMEL, a initié le projet GeneULike.

I.2 Contexte du laboratoire

Les recherches du groupe dirigé par Frédéric CHALMEL au sein de l'équipe 8 de l'unité Inserm U1085-Irset portent sur l'étude du développement et des fonctions testiculaires chez l'homme et comment l'environnement, au sens large, peut impacter ces processus et donc la fertilité masculine. Une des spécificités de ce groupe est d'utiliser des technologies à haut et ultra-haut débit pour leurs analyses. Leurs recherches impliquent donc de fortes composantes bio-informatiques et

génomiques qui ont conduit au développement de plusieurs outils bio-informatiques permettant d'analyser (Chalmel and Primig 2008), d'interpréter (Britto et al. 2012; Sallou et al. 2016) et de visualiser (Darde et al. 2015) des données « omiques ».

Dans ce contexte, un des objectifs de la thèse de Thomas Darde était de mettre en place, en collaboration avec la plateforme de bio-informatique GenOuest, un nouvel espace de dépôt multi-espèces, appelé *the TOXicological sIgNatures database* (TOXsIgN, http://toxsign.genouest.org), dédié aux signatures toxicologiques. Ces dernières se définissent comme la description des effets physiologiques, cellulaires, moléculaires, génomiques sur les individus ou leurs descendants, après exposition à des facteurs environnementaux uniques ou combinés, comme les produits chimiques et les agents physiques ou biologiques. Au cours du développement de ce *repository*, Frédéric CHALMEL a proposé de généraliser ce concept de dépôt de listes de gènes à l'ensemble des thématiques des sciences de la vie. Le projet GeneULike était né.

La première fonction de GeneULike est d'être un espace de dépôt multi-espèces et multi-technologiques destiné à héberger des listes d'entités biologiques accessible à la communauté scientifique *via* un serveur web. Ces listes sont organisées selon une architecture à quatre niveaux (Projets > Études > Filtrations > Listes). Chaque niveau, associé à un identifiant unique, est annoté (description, organisme, informations, éléments méthodologiques et expérimentaux, liens PubMed vers l'article, …) ce qui permet de décrire précisément comment ces listes ont été obtenues.

Le deuxième objectif de GeneULike est d'adosser cet espace de dépôt à un espace de travail qui permettra de mettre à disposition un ensemble d'outils (comparaison d'une liste à celles déjà présentent dans GeneULike) destinés aux utilisateurs.

I.3 Objectifs du stage

L'objectif principal de mon stage est d'initier la mise en place de l'espace de dépôt, GeneULike, en m'inspirant fortement de l'architecture de TOXsIgN (c.f. section précédente) Cet objectif se subdivise en trois sous-parties :

1. La création de la base de données destinée à héberger les listes d'entités biologiques.
2. Développer l'interface web qui permettra à un utilisateur de déposer des projets et des listes associées mais également à la communauté scientifique de consulter ces informations.
3. La conception d'un espace de travail qui aura pour vocation d'héberger différents outils de recherche et de comparaison. Un outil permettra également de convertir toutes les entités biologiques (transcrits, protéines, sondes nucléotidiques, …) en identifiants uniques

(EntrezGene et HomoloGene IDs) qui permettront de franchir à la fois les barrières technologiques et des espèces.

Cet espace de travail nécessitera la mise en place d'un système de soumission de tâches propres à chaque utilisateur.

Face à l'ampleur du travail à effectuer, la durée de mon stage a été prolongée de deux mois. Ce rapport de stage présente l'état d'avancement de mon travail au cours des trois premiers mois.

II. Matériels et méthodes

II.1 Environnement de travail

II.1.1 Système de contrôle de version

Un système de contrôle de version (VCS) ou outil de gestion de projet permet de suivre les modifications itératives réalisées sur un projet informatique ou non. La possibilité de revenir à une version antérieure spécifique d'un projet ou d'un fichier facilite le développement d'outil seul ou à plusieurs (Blischak, Davenport, and Wilson 2016). L'utilisation d'outils de gestion de version telle que GitHub ou SourceForge est recommandée dans le cadre du développement d'outils scientifiques (Prlić and Procter 2012).. Le développement de GeneULike a été réalisé avec l'outil de gestion de projet GitHub.

II.1.2 Python

Le langage de programmation Python a été choisi pour écrire la plupart des scripts nécessaires pour le développement de l'outil de conversion. La version 2.7 de Python a été sélectionnée afin de conserver une cohérence avec le travail déjà réalisé par Thomas Darde sur le projet TOXsIgN. Python est un langage de programmation de haut niveau (Bassi 2007). Il est devenu populaire au sein de la communauté scientifique, principalement par sa syntaxe (Bakker 2014), sa programmation orienté objet (Ekmekci, McAnany, and Mura 2016), son nombre élevé de librairies (ou *packages* en anglais) (Perkel 2015) qui sont utiles pour de nombreux domaines biologiques (Hart et al. 2015). D'autres scripts (R et Tcl/Tk) développés par mon laboratoire sont également utilisés et appelés par mes fonctions Python.

II.2 Environnement web

II.2.1 Modèle MVC

La mise en place d'une application web nécessite de mettre en relation différents composants afin de pouvoir gérer des données, les représenter de manière visuelle et d'exécuter les actions

effectuées par les utilisateurs sur un site web. Ce modèle relationnel est appelé architecture MVC, pour Modèle, Vue, Contrôleur (Figure 1). Le modèle contient les informations à afficher sur une page web. La vue est la représentation visuelle de la page. Le contrôleur supervise l'exécution des actions effectuées par les utilisateurs et correspond à un intermédiaire entre le modèle et la vue. Cette architecture définit un canevas qui assure la structuration et la stabilité de l'application web.

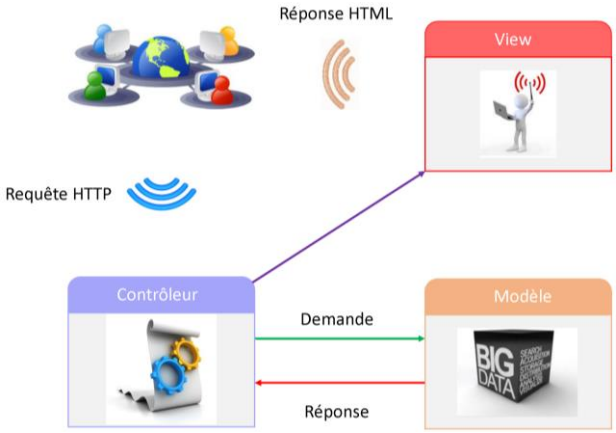


Figure 1 | Schéma de l'architecture MVC. Cette architecture est composée de trois entités distinctes, chacune ayant son propre rôle à remplir au sein d'une application web.

II.2.2 Vue : Interface Web

L'interface graphique web a été réalisée par les langages de programmations HTML5 et CSS3, pour respectivement *HyperText Markup Language* et *Cascading Style Sheets*.

II.2.3 Contrôleurs : AngularJS et Pyramid

Le développement de GeneULike a nécessité l'utilisation de deux frameworks : AngularJS (https://angularjs.org) et Pyramid (http://docs.pylonsproject.org/projects/pyramid/en/latest/#). AngularJS est un framework JavaScript libre et *open source* développé en 2009 par Google. L'un des concepts d'AngularJS est celui de *Data Binding*. Il s'agit ici du lien que fait AngularJS entre le code JavaScript (utilisé pour les fonctions et le traitement de certaines information) et le code HTML (représentation de la page). Il est ainsi possible pour chaque page HTML de créer un modèle propre et d'afficher des variables en temps réel sans prétraitement préalable en JavaScript.

L'objectif de Pyramid est de combiner la simplicité d'utilisation et la facilité de lecture offertes par python afin de mettre en place une architecture web robustes et aisément maintenable. Pyramid communique avec la banque de données et gère les requêtes effectuées depuis le site afin de fournir aux utilisateurs du site les informations demandées. Ce framework offre la possibilité d'utiliser toute la bibliothèque d'outils offerte par Python ainsi que les outils ou scripts développés de manière indépendante.

II.2.4 Modèle : Base de données

Les données ont été stockées et gérées à l'aide du système de gestion de base de données (SGBD) MongoDB. MongoDB est un système de gestion de base de données orientées documents et évolutif de la mouvance NoSQL (*Not Only SQL*). Ceci signifie que les données stockées respectent un système clé-valeur où pour chaque information une valeur est associée, (« identifiant » = « adresse mail » par exemple).

Le côté évolutif de MongoDB, ou « scalabilité », se traduit par le fait que MongoDB est capable de s'adapter en fonction des ressources informatiques mais également en fonction de la façon dont la base est sollicitée. L'utilisation de documents au format JavaScript Object Notation (JSON) encodé en objet au format binaire BSON sans schéma prédéfini permet de manipuler les données de manière souple en les ajoutant et en les retirant de la base à tout moment « à la volée ».

Les données prennent la forme de documents enregistrés eux-mêmes dans des collections, chacune contenant un ou plusieurs documents. Les collections sont comparables aux tables et les documents aux enregistrements des bases de données relationnelles. Contrairement aux bases de données relationnelles, les clés d'un enregistrement sont libres et peuvent être différentes d'un enregistrement à un autre au sein d'une même collection. La seule clé commune et obligatoire est la clé principale « _id ». Cette dernière est utilisée pour identifier de manière unique les documents au sein d'une collection. Par ailleurs, si MongoDB ne permet pas d'effectuer des requêtes très complexes, il permet en revanche de programmer des requêtes spécifiques.

II.3 Indexation de données au sein de GeneULike

La comparaison de listes de gènes appartenant à des espèces ou des technologies différentes (inter –technologies/ -espèces) dans GeneULike reposent sur la centralisation d'un ensemble hétérogènes d'entités biologiques autour d'identifiants uniques : les identifiants de gènes Entrez Gene IDs et les identifiants d'orthologies, les HomoloGene IDs. Entrez Gene est une base de données maintenue par le NCBI centrée autour de gènes (Maglott et al. 2011). Chaque gène est associé à un identifiant unique ainsi qu'à de nombreuses informations comme par exemple le nom, les synonymes, l'espèce ou encore les produits de gène (transcrits, protéines). Les EntrezGene IDs étant liés à un grand nombre de banques de données permet de s'affranchir de la barrière technologique – la majorité des entités biologiques pouvant être converties en EntrezGene IDs. Ces banques incluent la *Reference Sequence Database* (RefSeq) (O'Leary et al. 2016), Ensembl (Aken et al. 2017), UniProt (The UniProt Consortium 2008) et les sondes nucléotidiques annotées par

GEO (Clough and Barrett 2016). Enfin, la conversion des EntrezGene en HomoloGene IDs (NCBI Resource Coordinators 2016), banque d'homologie développée par le NCBI, permet de s'affranchir de la barrière inter-espèce.

III. Résultats

Pour illustrer les fonctionnalités de GeneULike et les résultats obtenus durant mon stage, je m'appuierai sur un article de mon laboratoire d'accueil publié en 2007 (Chalmel et al. 2007). Dans celui-ci, Frédéric CHALMEL et ses collègues ont analysé et comparé le transcriptome de plusieurs types de cellules testiculaires (les cellules de Sertoli, les cellules germinales mitotiques : les spermatogonies, les cellules germinales méiotiques : les spermatocytes, les cellules germinales en différenciation : les spermatides) chez l'homme, la souris et le rat. L'objectif était d'identifier les gènes possédant un profil d'expression conservé chez les mammifères. Au cours de mon stage, j'ai déposé cette étude publiée dans GeneULike pour m'en servir d'exemple. Plus spécifiquement, j'ai extrait cinq listes de sondes nucléotidiques murines correspondant : aux gènes différentiellement exprimés entre les différents types cellulaires (liste composée de 9283 sondes nommée DET, pour *Differentially Expressed in Testis)*, parmi lesquels ceux possédant un profil d'expression somatiques (2134 sondes, DET-SO), mitotiques (3423 sondes, DET-MI), méiotiques (2317 sondes, DET-ME) et post-méiotiques (1589 sondes, DET-PM)

III.1 Architecture d'un projet dans GeneULike

GeneULike est un espace de dépôt organisé en 4 niveaux (Figure 2) : Projets > Études > Stratégies > Listes. Chaque niveau est défini par un identifiant unique facilitant son accessibilité, réutilisation et citation dans les articles. Ainsi, le projet dresse le contexte général dans lequel se sont déroulées la ou les études qui lui sont associées. Chaque étude décrie l'expérimentation effectuée par le biais de différents champs et description tel que l'organisme, l'organe, la technologie ou encore le protocole expérimental utilisé. A chacune de ces études est associée à une ou plusieurs stratégies de filtration aboutissant à l'obtention de listes d'entités biologiques. Cette stratégie décrit, par exemple, la normalisation des données brutes ou encore, les étapes de filtrations statistiques et de classification ayant permis d'obtenir des listes d'entités biologiques. Enfin, les listes possèdent toutes un nom, une description ainsi que d'autres informations permettant par exemple de connaître le type d'identifiants dans ces listes (protéines, transcrits, sondes, etc). D'autre part, l'utilisateur à la possibilité de décrire une hiérarchie entre les listes, certaines n'étant que des sous-ensembles d'une autre.

Figure 2 | Représentation de la structure d'un fichier au format JSON d'une étude de l'expression différentielle des gènes impliqués dans le tube séminifère au cours de la semaine 8/9 chez *Mus Musculus* (Chalmel et al. 2007) au sein de GeneULike. Ce format contient un objet entouré par des accolades (« {} »). Un projet possède plusieurs clés ou attributs, ainsi que d'un tableau d'objets qui décrivent les études. Chaque étude possède, en plus de ces clés, une liste d'objets qui décrivent les stratégies. Ces derniers possèdent également une liste d'objets qui décrivent les listes.

Au cours du dépôt d'un nouveau projet, l'ensemble des informations saisies par l'utilisateur est associé à un statut dit « privé ». Ainsi, seul le créateur de la liste est capable d'accéder et de modifier ces données. Pour faire, évoluer ce statut en « public », l'utilisateur doit, à partir de son espace personnel, effectuer une demande de changement de statut auprès des administrateurs du site. Le projet passe alors de « privé » à « en attente de validation ». Ce statut fige toutes les informations relatives au projet et celles-ci sont accessibles par toutes les personnes possédant l'identifiant. Néanmoins, à la différence du statut « public », le projet n'est pas indexé dans le moteur de recherche de GeneULike, limitant ainsi sa diffusion. Pour pouvoir effectuer le changement privé-public, un certain nombre de critère est requis parmi lesquels figure en première place la disponibilité d'un identifiant PubMed assurant la viabilité scientifique du projet par sa publication dans une revue scientifique à comité de lecture. Une fois « public », un projet n'est plus modifiable et est accessible par tous *via* son identifiant ou le moteur de recherche.

III.2 Interface Web

III.2.1 Connection

La soumission d'un projet dans GeneULike requiert un compte utilisateur. Ce dernier doit donc s'identifier *via* une page d'inscription demandant de renseigner différents champs comme le nom, le prénom, l'adresse mail ou le laboratoire. Une fois la demande effectuée, l'utilisateur reçoit un

mail l'invitant à valider son inscription. Le nouvel utilisateur peut alors accéder à sa page personnelle, modifier ses informations, mot de passe ou bien déposer des projets dans GeneULike. L'utilisation de témoins de connections (ou *cookies* en anglais) assurent aux utilisateurs de conserver toutes les actions effectuées sur le site avant leur connexion. Ainsi, l'historique des outils utilisés ou des recherches restent accessibles avec ou sans connexion.

III.2.2 Soumission et rapport d'erreur

La procédure de dépôt d'un projet est réalisée par l'intermédiaire d'un fichier Excel structuré selon l'organisation des données et représentant un seul projet. Cette manière d'insérer les données dans GeneULike s'inspire de la méthode de soumission de GEO et permet aux utilisateurs le dépôt simple et rapide d'un projet complexe incluant de nombreuses listes sans risque de déconnexion. Ce classeur est organisé en quatre feuilles, une par niveau, récapitulant toutes les informations nécessaires. Si tous les champs ne sont pas indispensables pour le dépôt d'un projet, certaines informations sont essentielles. Il est ainsi impératif d'indiquer un nom et une description pour chaque liste ainsi que le type d'identifiant afin de permettre leur conversion en Entrez Gene IDs (c.f. dernière section des résultats).

	GeneULike Template	- A project is the global description of your studies - Each project is defined by one unique title (one by line) - Fill the fields		
ProjectID	Title	Description	PubMedID(s) (comma separated)	Contributors (comma separated)
Projet0	The conserved transcriptome in human and rodent male gametogenesis	Cross-species expression profiling analysis of the human, mouse, and rat male meiotic transcriptional program, using enriched germ cell populations, whole gonad, and high-density oligonucleotide microarrays (GeneChips)		Chamel F., Rolland AD., Niederhausser-Wiederkehr C., Chung SS., Demougin P., Gattiker A., MooreJ., Patard JJ., Wolgemuth DJ., Jégou B., Primig M.

Figure 3 | | Exemple d'un projet (Chalmel et al. 2007) avant soumission en renseignant le fichier Excel. Un projet est définit par un identifiant, un titre, une description. Dans cet exemple, l'identifiant PubMed a été volontairement omis.

Avant de mettre en ligne (ou *uploader* en anglais) les informations contenues dans le classeur Excel sur le serveur GeneULike (Figure 3), le fichier est tout d'abord parcouru afin de détecter des erreurs telles que le non remplissage de champs obligatoires. Le résultat de cette étape s'affiche sous la forme d'un rapport d'erreur (Figure 4). Ce dernier liste pour chacun des quatre niveaux (Projets > Études > Stratégies > Listes) trois types d'erreurs. Une erreur de type « Critique » signifie la présence d'une erreur empêchant l'indexation des données (absence d'un champs requis) obligeant l'utilisateur à effectuer les modifications nécessaires pour que le projet puisse être correctement *uploadé*. Un « Avertissement » signifie que le projet peut être *uploadé* mais que les informations renseignées ne permettront pas à l'utilisateur de changer le statut d'un projet en « public ». Enfin,

une erreur de type « Info » signifie que certains champs facultatifs sont manquants mais ceux-ci n'entravent pas la soumission du projet dans le système.

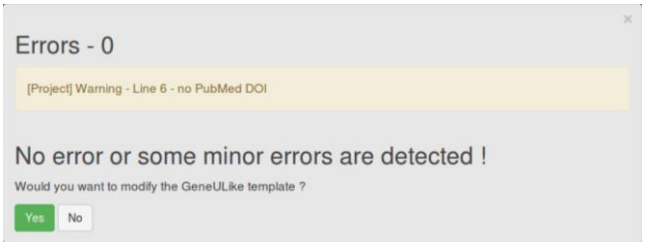
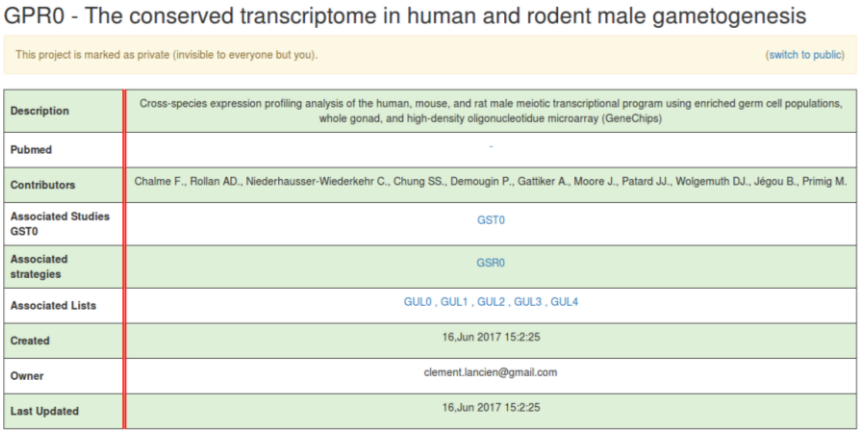


Figure 4 | Rapport d'erreur suite à la mise en ligne d'un projet *via* le fichier Excel (Figure 3). Ce rapport signale à l'utilisateur un avertissement. L'absence de la référence PubMed du projet signifie que le créateur ne pourra faire évoluer le changement de statut « privé à « public ».

III.2.3 Consultation de listes déposées

Une fois le fichier Excel déposé, l'utilisateur peut consulter les informations du projet *via* l'interface web de GeneULike (Figure 5). Depuis sa page personnel, l'utilisateur a accès à tous les projets/études/stratégies/listes dont il est le propriétaire. L'accès à la page d'un de ces objets est effectué à l'aide d'une adresse web uniformisée du type http://geneulike.genouest.org/#/browse?dataset=XXX, où XXX est l'identifiant unique de l'objet à afficher.


GPR0 - The conserved transcriptome in human and rodent male gametogenesis

This project is marked as private (invisible to everyone but you). (switch to public)

Description	Cross-species expression profiling analysis of the human, mouse, and rat male meiotic transcriptional program using enriched germ cell populations, whole gonad, and high-density oligonucleotidue microarray (GeneChips)
Pubmed	-
Contributors	Chalme F., Rollan AD., Niederhausser-Wiederkehr C., Chung SS., Demougin P., Gattiker A., Moore J., Patard JJ., Wolgemuth DJ., Jégou B., Primig M.
Associated Studies GST0	GST0
Associated strategies	GSR0
Associated Lists	GUL0 , GUL1 , GUL2 , GUL3 , GUL4
Created	16,Jun 2017 15:2:25
Owner	clement.lancien@gmail.com
Last Updated	16,Jun 2017 15:2:25

Figure 5 | Exemple d'affichage d'un projet (Chalmel et al. 2007) en statut « «privé » à partir de l'espace de travail du créateur de projet.

III.3 L'espace de travail de GeneULike

L'espace de travail de GeneULike s'inspire de celui développé dans TOXsIgN. Il permet aux utilisateurs, connectés ou non, de lancer différentes tâches (*jobs*) sur les listes indexées dans la base de données. Pour cela, un outil de soumission de tâches a été mis en place. Ce dernier gère le fonctionnement du *job,* ses résultats et les potentielles erreurs rencontrées. Les informations relatives à une tâche sont accessibles par le biais d'une page web dédiée. Sur cette dernière, le statut, l'outil utilisé, les paramètres et un lien de visualisation des résultats sont affichés.

Pour ma part, durant mon stage j'ai pu travailler sur la mise en place de l'outil de conversion d'identifiants. Celui-ci a pour objectif, quand cela est possible, la conversion des identifiants présentes dans une liste en Entrez Gene et HomoloGene IDs. Il est néanmoins possible que

certaines entités biologiques ne puissent être converties en EntrezGene et/ou HomoloGene IDs si celles-ci n'appartiennent à aucune des 17 banques intégrées.

Dans le cadre de la conversion des entités biologiques (identifiants de protéines, transcrits, sondes nucléotidiques, …) plusieurs scripts Python ont été développés afin de les associer avec des Entrez Gene et des HomoloGene IDs. Les identifiants des banques de données les plus utilisées ont ainsi été indexées dans GeneULike. Ces scripts permettent de se connecter automatiquement aux serveurs des différentes banques de données, telle que le NCBI ou UniProt, puis de télécharger les informations sous la forme de fichiers compressés. A partir de ces fichiers, les couples d'identifiants Entrez Gene IDs/identifiants de la base de données sont extraits et stockés dans les fichiers tabulés de type « Entrez_GeneTobase_de_donnees.txt » (« Entrez_Gene ToHomoloGene.txt » ou « Entrez_GeneToUniProt.txt » par exemple). Enfin, chaque fichier tabulé est ensuite transformé au format JSON avant d'être indexé dans GeneULike. Chaque indexation génère une collection nommée en fonction de la base de données d'où proviennent les couples d'identifiants (Figure 6).

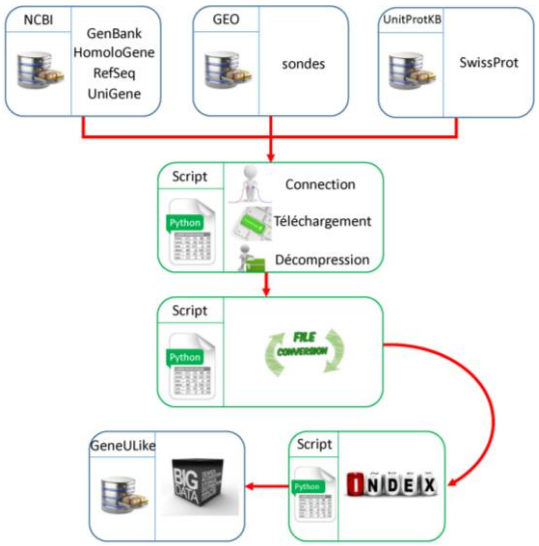


Figure 6 | Scripts python nécessaire pour indexer différentes bases de données. Un premier script se connecte aux différents serveurs FTP (*File Tansfert Protocol*) des différentes banques de données afin de télécharger plusieurs bases de données sous un format compressé. Une étape de décompression est suivie par une étape d'extraction des couples Entre Gene IDs et les identifiants des bases de données associés à ces Entrez Gene IDs. Enfin, ces fichiers sont convertis au format JSON avant d'être indexés dans la base GeneULike.

Actuellement, 17 types d'identifiants sont indexés dans le système et peuvent être convertis en Entrez Gene IDs. Ces informations représentent un volume de total de 27,7GB (Tableau 1). L'utilisation de la base de données GEO (ftp://ftp.ncbi.nlm.nih.gov/geo/), permet également de créer des associations entre des identifiants de sondes nucléotidiques de 1358 technologies de puces à ADN (152 espèces différentes) avec les Entrez Gene IDs. Ainsi, les données issues de ces technologies de puces peuvent être aisément déposées dans GeneULike.

La flexibilité des scripts développés ainsi que celle du système de gestion des données permettent également d'indexer de nouvelles bases de données rapidement et facilement (Figure 6).

Tableau 1 | Tableau des banques de données présentes dans la base GeneULike. Il montre l'étendu des données présentes dans GeneULike. Au total, 17 banques de données sont indexées dans la base GeneULike avec un total de 151 millions d'identifiants associés à 9,7 millions d'EntrezGene IDs

Nom de la banque	Banques de données	Type de données	N° identifiants indexés	Taille collection (MB)
Ensembl	Ensembl_gene	Gène	571656	104
	Ensembl_protein	Protéine	655926	107
	Ensembl_transcript	Transcrit	588973	105
HomoloGene	HomoloGene	Orthologie	275237	46
GEO	GPL	Sondes	19228568	4542
GenBank	GenBank_protein	Protéine	18510487	3497
	GenBank_transcript	Transcrit	2418535	414
NCBI	GI_protein	Protéine	38652386	6551
	GI_transcript	Transcrit	21286945	3588
	UniGene	Gène	592208	105
RefSeq	RefSeq_protein	Protéine	20141901	3550
	RefSeq_transcript	Transcrit	18868410	3509
UniProt	SwissProt	Protéine	461735	101
	trEmbl	Protéine	8742128	1448
Vega	Vega_gene	Gène	16656	3
	Vega_protein	Protéine	16629	3
	Vega_transcript	Transcrit	17837	3

Au final, environ 9,7 millions d'Entrez Gene IDs sont associés à 151 millions d'identifiants de banques de données indexées dans GeneULike. Sur ces 9,7 millions d'Entrez Gene IDs, 2,83% peuvent être converti en HomoloGene IDs, cette banque n'incluant qu'une vingtaine d'espèces différentes (NCBI Resource Coordinators 2016). C'est grâce à cet énorme jeu de données que les conversions inter-espèces/-technologies/-omiques sont possibles.

Lors du dépôt d'un projet, les entités biologiques de chaque liste sont extraites. Chaque entité biologique va subir une conversion en EntrezGene puis en HomoloGene ID. Cette conversion entraine pour chaque liste la création d'un fichier tabulé qui contient pour chaque ligne : l'identifiant de départ ; l'EntrezGene ID ; l'HomoloGene ID ; le nom du gène (*Gene Symbol*) ; l'espèce (*Taxon ID*) et la description du gène.

Dans le cas de mes listes d'intérêts extraites de l'étude publiée par mon laboratoire, le résultat de la conversion en EntrezGene et HomoloGene IDs a été regroupé dans le Tableau 2.

Tableau 2 | Tableau des résultats obtenues après conversion des identifiants issues de l'étude sur l'expression des gènes dans le tubes séminifères de la semaine 8/9 chez *Mus Musculus*

Nom puce	N° IDs de sondes	N° Entrez Gene IDs	N° HomoloGene IDs
DET	9283	7372	6611
DET-SO	2134	1634	1570
DET-MI	3243	2535	2456
DET-ME	2317	1905	1584
DET-PM	1589	1708	1394

IV. Discussion

En l'état actuel, GeneULike permet la soumission et la consultation de listes d'entités biologiques sur l'interface web. Cette étape de soumission est réalisée par l'intermédiaire d'un fichier Excel pour sa simplicité d'utilisation même sans connexion internet (hors-ligne). Néanmoins, ce choix n'est pas sans conséquence puisque ce format nous a imposé certaines restrictions. En effet, le fichier Excel ne permet pas de réaliser des sélections multiples dans les différentes cellules à remplir ce qui posera des problèmes lorsque GeneULike intégrera des vocabulaires contrôlés (ontologies) pour décrire certaines informations.

Après chaque soumission, chaque projet est associé automatiquement à un statut « privé ». Parmi tous les critères indispensables, le plus important permettant de faire évoluer un projet vers un statut « public » est de l'associer à une publication scientifique (PubMed ID), seul gage de qualité assurant que l'analyse a été évaluée par des scientifiques du domaine. Un des objectifs de GeneULike étant également de fournir des outils de comparaison, les projets « publics » soumis devraient contribuer à augmenter le nombre de citations de leurs articles associés. Ainsi, ce système permet aux anciennes études de qualité mais tombées en « désuétudes technologiques », d'être remises sur le devant de la scène à égalité avec les projets et listes issus de technologies plus récentes. Les scientifiques ont donc un intérêt majeur à soumettre leurs anciens travaux de recherche. Toutefois, GeneULike offre également la possibilité aux utilisateurs de soumettre des listes et de les maintenir en statut « privé » tout en permettant l'utilisation des outils présents dans l'espace de travail. GeneULike constitue alors un véritable support pour l'analyse de données « omiques ».

Grâce à l'indexation de 17 banques de données différentes et 1358 technologies au sein de GeneULike, l'utilisateur est peu limité par le type de données à déposer et peut centraliser dans un même endroit des listes de sondes ADN, de protéines, de transcrits. GeneULike assure une

convergence automatique de ces identifiants, lorsque cela est possible, en identifiants uniques : EntrezGene et HomoloGene IDs. La conversion en EntrezGene IDs permet de franchir la barrière technologique et la conversion en HomoloGene IDs permet de franchir la barrière des espèces. Toutefois, cette dernière conversion n'est possible que pour une vingtaine d'espèce (NCBI Resource Coordinators 2016), et restreint donc les comparaisons. Il est déjà prévu d'inclure d'autres banques d'orthologies/homologies, telles que GeneTree (Vilella et al. 2009) et OMA (Altenhoff et al. 2015) afin d'améliorer considérablement les comparaisons de listes appartenant à des espèce différentes – OMA couvrant à elle seule plus de 2000 espèces. Ceci nécessitera le développement de scripts dédiés afin d'associer les identifiants de ces nouvelles banques vers des EntrezGene IDs et ainsi vers les 17 autres types d'identifiants.

V. Conclusion et perspectives

GeneULike est un espace dépôt de listes d'entités biologiques issues d'analyses publiées. Ces listes sont déposées, par l'intermédiaire d'un fichier Excel, et organisées sous la forme d'un projet associé à une publication. Actuellement GeneULike est fonctionnel mais nécessite l'ajout à court, moyen et long terme de fonctionnalités supplémentaires pour être un espace de dépôt et de travail complet.

A court terme, l'utilisation de vocabulaires contrôlés et hiérarchisés (ontologies) pour compléter les champs du fichier de soumission Excel permettra de structurer les informations au sein de GeneULike et ainsi de considérablement améliorer la recherche dans la banque par des requêtes spécifiques. Il est déjà envisagé d'intégrer des ontologies de tissus telles que la BRENDA Tissue Ontology (BTO) (Gremse et al. 2011), une ontologie pour décrire les tissus ainsi que les lignées cellulaires, types cellulaires et cultures cellulaires - Biomedical Investigations (OBI) (Bandrowski et al. 2016), ou, plus communément, la Gene Ontologie (GO) (Gene Ontology Consortium 2015) pour décrire le(s) processus biologique(s) étudié(s).

GeneULike ambitionne également de proposer un espace de travail dont l'interface s'inspire de celle développée dans le cadre du projet Galaxy (Blankenberg et al. 2010). Sans être aussi performant, l'objectif est de rendre toutes les informations présentent dans GeneULike utilisables pour effectuer des comparaisons, créer de nouvelle listes par le croisement de listes déjà existantes ou encore de proposer des outils de conversion d'identifiants.

Avant la publication d'un article consacré à GeneULike, il est indispensable d'intégrer une grande quantité de projets et donc de listes d'entités biologiques dans la base GeneULike. En effet, si la

base ne possède pas suffisamment de données, les utilisateurs ne trouveront pas d'attraits à déposer leurs listes dans GeneULike. Une stratégie actuellement envisagée serait de couvrir un domaine biologique particulier. Ce dernier pourrait concerner le développement du testicule et ses fonctions dans plusieurs espèces, thématiques phares de mon laboratoire d'accueil (Chalmel et al. 2007; Rolland et al. 2009; Chalmel et al. 2012; Rolland et al. 2013). Une fois l'article de GeneULike accepté, il serait également indispensable de faire connaître l'existence de notre outil aux éditeurs de journaux scientifiques afin que ces derniers incitent les scientifiques à déposer leurs analyses dans GeneULike au même titre que leurs données brutes dans GEO ou ArrayExpress.

Une autre amélioration de GeneULike sur le court et moyen terme porte sur le dépôt de projets issus de données de séquençage (RNA-seq) et notamment de listes composées de transcrits assemblés. Ces technologies étant de plus en plus utilisées, il est indispensable que le *repository* soit compatible avec ces données. Néanmoins, plusieurs obstacles techniques devront être surmontés notamment concernant les différentes versions de génome disponibles et les différences de qualités des génomes. L'adaptation du dépôt pour ce type de projets est actuellement à l'étude dans le laboratoire. Une possibilité serait de fournir un outil permettant d'annoter automatiquement les transcrits assemblés et ainsi de les associer à des EntrezGene IDs pour plusieurs espèces et versions de génome.

Enfin, à plus long terme, la mise en place d'outils communautaires permettra de repousser les fonctionnalités de GeneULike en permettant notamment de travailler à plusieurs sur un même projet déposé.

Pour conclure, GeneULike est une réponse aux besoins des scientifiques de centraliser afin mieux comparer les listes d'entités biologiques publiés. Cependant, cet outil sera un succès à la seule et unique condition que les scientifiques « s'emparent » de ce dernier en faisant l'effort d'y déposer leurs données agrandissant par la même occasion l'intérêt d'autre scientifiques pour le dépôt. Le succès de GeneULike repose également sur les éditeurs qui doivent imposer, au même titre que pour les données brutes, le dépôt des données issues des étapes de l'analyse d'une publication.

Bibliographie

Aken, Bronwen L., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Friederike Bernsdorff, Jyothish Bhai, Konstantinos Billis, et al. 2017. "Ensembl 2017." *Nucleic Acids Research* 45 (D1): D635–42. doi:10.1093/nar/gkw1104.

Altenhoff, Adrian M., Nives Škunca, Natasha Glover, Clément-Marie Train, Anna Sueki, Ivana Piližota, Kevin Gori, et al. 2015. "The OMA Orthology Database in 2015: Function Predictions, Better Plant Support, Synteny View and Other Improvements." *Nucleic Acids Research* 43 (Database issue): D240–49. doi:10.1093/nar/gku1158.

Bakker, Mark. 2014. "Python Scripting: The Return to Programming." *Ground Water* 52 (6): 821–22. doi:10.1111/gwat.12269.

Bandrowski, Anita, Ryan Brinkman, Mathias Brochhausen, Matthew H. Brush, Bill Bug, Marcus C. Chibucos, Kevin Clancy, et al. 2016. "The Ontology for Biomedical Investigations." *PLoS ONE* 11 (4). doi:10.1371/journal.pone.0154556.

Barrett, Tanya, and Ron Edgar. 2006. "Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis." *Methods in Enzymology* 411: 352–69. doi:10.1016/S0076-6879(06)11019-8.

Barrett, Tanya, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, et al. 2011. "NCBI GEO: Archive for Functional Genomics Data Sets--10 Years on." *Nucleic Acids Research* 39 (Database issue): D1005-1010. doi:10.1093/nar/gkq1184.

Bassi, Sebastian. 2007. "A Primer on Python for Life Science Researchers." *PLoS Computational Biology* 3 (11). doi:10.1371/journal.pcbi.0030199.

Blankenberg, Daniel, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. 2010. "Galaxy, a Web-Based Genome Analysis Tool for Experimentalists." *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* 0 19 (January): Unit-19.1021. doi:10.1002/0471142727.mb1910s89.

Blischak, John D., Emily R. Davenport, and Greg Wilson. 2016. "A Quick Introduction to Version Control with Git and GitHub." *PLoS Computational Biology* 12 (1): e1004668. doi:10.1371/journal.pcbi.1004668.

Britto, Ramona, Olivier Sallou, Olivier Collin, Grégoire Michaux, Michael Primig, and Frédéric Chalmel. 2012. "GPSy: A Cross-Species Gene Prioritization System for Conserved Biological Processes—application in Male Gamete Development." *Nucleic Acids Research* 40 (Web Server issue): W458–65. doi:10.1093/nar/gks380.

Chalmel, Frédéric, Aurélie Lardenois, Bertrand Evrard, Romain Mathieu, Caroline Feig, Philippe Demougin, Alexandre Gattiker, et al. 2012. "Global Human Tissue Profiling and Protein Network Analysis Reveals Distinct Levels of Transcriptional Germline-Specificity and Identifies Target Genes for Male Infertility." *Human Reproduction (Oxford, England)* 27 (11): 3233–48. doi:10.1093/humrep/des301.

Chalmel, Frédéric, and Michael Primig. 2008. "The Annotation, Mapping, Expression and Network (AMEN) Suite of Tools for Molecular Systems Biology." *BMC Bioinformatics* 9 (February): 86. doi:10.1186/1471-2105-9-86.

Chalmel, Frédéric, Antoine D. Rolland, Christa Niederhauser-Wiederkehr, Sanny SW Chung, Philippe Demougin, Alexandre Gattiker, James Moore, et al. 2007. "The Conserved Transcriptome in Human and Rodent Male Gametogenesis." *Proceedings of the National Academy of Sciences* 104 (20): 8346–8351.

Chelliah, Vijayalakshmi, Camille Laibe, and Nicolas Le Novère. 2013. "BioModels Database: A Repository of Mathematical Models of Biological Processes." *Methods in Molecular Biology (Clifton, N.J.)* 1021: 189–99. doi:10.1007/978-1-62703-450-0_10.

Clough, Emily, and Tanya Barrett. 2016. "The Gene Expression Omnibus Database." *Methods in Molecular Biology (Clifton, N.J.)* 1418: 93–110. doi:10.1007/978-1-4939-3578-9_5.

Darde, Thomas A., Olivier Sallou, Emmanuelle Becker, Bertrand Evrard, Cyril Monjeaud, Yvan Le Bras, Bernard Jégou, Olivier Collin, Antoine D. Rolland, and Frédéric Chalmel. 2015. "The ReproGenomics Viewer: An Integrative Cross-Species Toolbox for the Reproductive Science Community." *Nucleic Acids Research* 43 (Web Server issue): W109–16. doi:10.1093/nar/gkv345.

Edgar, Ron, Michael Domrachev, and Alex E. Lash. 2002. "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository." *Nucleic Acids Research* 30 (1): 207–10.

Ekmekci, Berk, Charles E. McAnany, and Cameron Mura. 2016. “An Introduction to Programming for Bioscientists: A Python-Based Primer.” *PLoS Computational Biology* 12 (6): e1004867. doi:10.1371/journal.pcbi.1004867.

Gene Ontology Consortium. 2015. “Gene Ontology Consortium: Going Forward.” *Nucleic Acids Research* 43 (Database issue): D1049–56. doi:10.1093/nar/gku1179.

Gremse, Marion, Antje Chang, Ida Schomburg, Andreas Grote, Maurice Scheer, Christian Ebeling, and Dietmar Schomburg. 2011. “The BRENDA Tissue Ontology (BTO): The First All-Integrating Ontology of All Organisms for Enzyme Sources.” *Nucleic Acids Research* 39 (Database issue): D507–13. doi:10.1093/nar/gkq968.

Hart, Reece K., Rudolph Rico, Emily Hare, John Garcia, Jody Westbrook, and Vincent A. Fusaro. 2015. “A Python Package for Parsing, Validating, Mapping and Formatting Sequence Variants Using HGVS Nomenclature.” *Bioinformatics (Oxford, England)* 31 (2): 268–70. doi:10.1093/bioinformatics/btu630.

Kim, Sunghwan, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, et al. 2016. “PubChem Substance and Compound Databases.” *Nucleic Acids Research* 44 (Database issue): D1202–13. doi:10.1093/nar/gkv951.

Maglott, Donna, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. 2011. “Entrez Gene: Gene-Centered Information at NCBI.” *Nucleic Acids Research* 39 (Database issue): D52-57. doi:10.1093/nar/gkq1237.

NCBI Resource Coordinators. 2016. “Database Resources of the National Center for Biotechnology Information.” *Nucleic Acids Research* 44 (Database issue): D7–19. doi:10.1093/nar/gkv1290.

O’Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. “Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation.” *Nucleic Acids Research* 44 (D1): D733-745. doi:10.1093/nar/gkv1189.

Parkinson, H., U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, et al. 2005. “ArrayExpress--a Public Repository for Microarray Gene Expression Data at the EBI.” *Nucleic Acids Research* 33 (Database issue): D553-555. doi:10.1093/nar/gki056.

Perkel, Jeffrey M. 2015. “Programming: Pick up Python.” *Nature* 518 (7537): 125–26. doi:10.1038/518125a.

Poldrack, Russell A., Deanna M. Barch, Jason P. Mitchell, Tor D. Wager, Anthony D. Wagner, Joseph T. Devlin, Chad Cumba, Oluwasanmi Koyejo, and Michael P. Milham. 2013. “Toward Open Sharing of Task-Based fMRI Data: The OpenfMRI Project.” *Frontiers in Neuroinformatics* 7 (July). doi:10.3389/fninf.2013.00012.

Prlić, Andreas, and James B. Procter. 2012. “Ten Simple Rules for the Open Development of Scientific Software.” *PLoS Computational Biology* 8 (12): e1002802. doi:10.1371/journal.pcbi.1002802.

Reimand, Jüri, Tambet Arak, Priit Adler, Liis Kolberg, Sulev Reisberg, Hedi Peterson, and Jaak Vilo. 2016. “g:Profiler-a Web Server for Functional Interpretation of Gene Lists (2016 Update).” *Nucleic Acids Research* 44 (W1): W83-89. doi:10.1093/nar/gkw199.

Rolland, Antoine D., Aurélie Lardenois, Anne-Sophie Goupil, Jean-Jacques Lareyre, Rémi Houlgatte, Frédéric Chalmel, and Florence Le Gac. 2013. “Profiling of Androgen Response in Rainbow Trout Pubertal Testis: Relevance to Male Gonad Development and Spermatogenesis.” *PLoS ONE* 8 (1). doi:10.1371/journal.pone.0053302.

Rolland, Antoine D, Jean-Jacques Lareyre, Anne-Sophie Goupil, Jérôme Montfort, Marie-Jo Ricordel, Diane Esquerré, Karine Hugot, Rémi Houlgatte, Fréderic Chalmel, and Florence Le Gac. 2009. “Expression Profiling of Rainbow Trout Testis Development Identifies Evolutionary Conserved Genes Involved in Spermatogenesis.” *BMC Genomics* 10 (November): 546. doi:10.1186/1471-2164-10-546.

Rother, Kristian, Wojciech Potrzebowski, Tomasz Puton, Magdalena Rother, Ewa Wywial, and Janusz M. Bujnicki. 2012. “A Toolbox for Developing Bioinformatics Software.” *Briefings in Bioinformatics* 13 (2): 244–57. doi:10.1093/bib/bbr035.

Sallou, Olivier, Paula D. Duek, Thomas A. Darde, Olivier Collin, Lydie Lane, and Frédéric Chalmel. 2016. “PepPSy: A Web Server to Prioritize Gene Products in Experimental and Biocuration Workflows.” *Database: The Journal of Biological Databases and Curation* 2016 (May). doi:10.1093/database/baw070.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A

Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50. doi:10.1073/pnas.0506580102.

The UniProt Consortium. 2008. "The Universal Protein Resource (UniProt)." *Nucleic Acids Research* 36 (Database issue): D190–95. doi:10.1093/nar/gkm895.

Vilella, Albert J., Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney. 2009. "EnsemblCompara GeneTrees: Complete, Duplication-Aware Phylogenetic Trees in Vertebrates." *Genome Research* 19 (2): 327–35. doi:10.1101/gr.073585.107.

Vizcaíno, Juan Antonio, Attila Csordas, Noemi del-Toro, José A. Dianes, Johannes Griss, Ilias Lavidas, Gerhard Mayer, et al. 2016. "2016 Update of the PRIDE Database and Its Related Tools." *Nucleic Acids Research* 44 (D1): D447-456. doi:10.1093/nar/gkv1145.

Annexe 1 : Structure d'accueil

L'Institut de Recherche en Santé, Environnement et Travail (IRSET) a pour mission d'étudier les facteurs environnementaux influençant la santé humaine. Au sein de ce dispositif, mon équipe d'accueil étudie les conséquences des expositions virales et chimiques sur le tractus uro-génital et la fertilité en général. Elle a pour objectifs de répondre aux préoccupations majeures concernant l'appareil uro-génital telles que la dissémination de virus par voie sexuelle, les conséquences des infections du tractus urogénital sur la fertilité et le développement de cancers, l'augmentation des anomalies du tractus uro-génital masculin (cancer testiculaire, hypospadias et cryptorchidie) pouvant conduire à l'infertilité ou l'effet potentiellement délétère des traitements anti-cancéreux sur la fertilité. Pour cela, les recherches sont orientées selon 4 axes : la différenciation gonadique et physiologique des gonades, l'exposition du tractus uro-génital aux virus, aux médicaments et aux produits chimiques environnementaux. C'est sur ce dernier axe, l'étude de l'impact des composés environnementaux sur le système reproductif, que portent les recherches de mon groupe.

Annexe 2 : Bilan personnel du stage

J'ai eu la chance de trouver un stage dans un champ disciplinaire qui m'intéresse particulièrement. Ayant comme projet à long terme de travailler dans le traitement des données « omiques », cette expérience m'a avant tout permis d'acquérir de solide base dans ce domaine.

Au cours de cette expérience, j'ai appris à gérer un nombre important de données et à extraire des informations à partir de ces données avant de les insérer automatiquement dans une base de données.

A travers le projet GeneULike qui m'a séduit par simplicité et son ambition, j'ai réalisé les difficultés techniques qui s'imposent au cours du développement d'outils bio-informatiques. En effet, la création de tels outils nécessite de se mettre à la place des utilisateurs (biologistes) afin d'identifier les besoins.

Pour rendre compatible des problèmes biologiques à l'informatique, telles que la comparaison de différentes entités biologiques, j'ai appris à utiliser plusieurs langages de programmation ou framework, en plus d'une base de donnée.

Enfin, j'ai mis un certain temps avant de me rendre compte de toutes les possibilités que pouvaient offrir ce dépôt.

Résumé

Aujourd'hui de nombreux espaces de dépôts sont mis à la disposition de la communauté scientifique. Ces derniers ont pour vocation l'hébergement des données brutes générées lors des expérimentations. Cependant il n'existe aucun espace pour le dépôt des listes d'entités biologiques résultant d'une stratégie d'analyse effectuée dans les publications. C'est dans ce contexte que s'inscrit GeneULike. GeneULike est un espace de dépôt web pour les listes issues de stratégie d'analyse. Grace à son outil de conversion d'entités biologiques afin de les centraliser autour des identifiants de gènes Entrez Gene et des identifiants d'orthologies HomoloGene, il est possible de déposer des listes d'espèces et de technologies différentes. Le dépôt de listes est réalisé par l'intermédiaire d'une application web. Cette dernière permet la consultation des listes mais met à disposition un espace de travail ayant pour vocation d'héberger différents outils, tels que des outils de recherche ou de comparaison d'une liste avec celles présentes dans la base de GeneULike. La pertinence de ce dépôt a été mise en évidence via le dépôt et la visualisation de cinq listes issues d'une étude sur la spermatogénèse.
Mots clés : Banque de données, Analyse de données, Gène, Conversion, Application Web

Abstract Development of GeneULike : a repository dedicated for biological entities lists

Currently, many repositories are available to the scientific community. The latter are intended to host the raw data generated during the experiments. However, there is no warehouse to deposit lists of biological entities resulting from an analysis strategy carried out in scientist's publication. Inthis context, the GeneULike project was born. GeneULike is a web repository dedicated to biological entities lists extracted from the analysis described in publications. One of the unique features of GeneULike relies on its ability to archive heterogeneous data from multiple species and multiple technologies. This feature rest on the implementation of a conversion process which centralize all biological entities in EntrezGene and HomoloGene identifiers. The submission of lists carried out via a web application allowing the visualization and the reutilization of information. Finally, GeneULike encompass a workspace providing a unique space to compare and analyze hosted data. The relevance of this repository has been highlighted by the submission and reading of five lists which came from a study on spermatogenesis.
Key words : Database, Data Analysis, Gene, Conversion, Web Application