Développement d'une application R Shiny pour la visualisation de résultats d'analyse de données biomarqueurs

Anne-Victoire Lagroy de Croutte de Saint Martin

AGROCAMPUS OUEST

☐ **CFR Angers** ☑ **CFR Rennes**

Année universitaire : 2020-2021 Spécialité : …Agronomie… Spécialisation (et option éventuelle) : …Statistiques et sciences de données…	**Mémoire de fin d'études** ☒ d'ingénieur d'AGROCAMPUS OUEST (École nationale supérieure des sciences agronomiques, agroalimentaires, horticoles et du paysage), école interne de L'institut Agro (Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement) ☐ de master d'AGROCAMPUS OUEST (École nationale supérieure des sciences agronomiques, agroalimentaires, horticoles et du paysage), école interne de L'institut Agro (Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement) ☐ de Montpellier SupAgro (étudiant arrivé en M2) ☐ d'un autre établissement (étudiant arrivé en M2)

Développement d'une application R Shiny pour la visualisation de résultats d'analyse de données biomarqueurs

Par : Anne-Victoire LAGROY de CROUTTE de SAINT MARTIN

Soutenu à Rennes le 08/09/2021

Devant le jury composé de :

Enseignant référent : Mathieu EMILY

Maître de stage : Caroline PACCARD et Rémi BRAZEILLES

Autres membres du jury (Nom, Qualité) : Marie-Pierre ETIENNE – enseignante

Les analyses et les conclusions de ce travail d'étudiant n'engagent que la responsabilité de son auteur et non celle d'AGROCAMPUS OUEST

Remerciements

Je tiens à remercier tout particulièrement **Rémi Brazeilles**, qui m'a encadrée tout au long de ce stage. Il a été patient et très intéressant, m'aidant dès que j'en avais besoin, et me conseillant dans mon travail. De plus, même avec la naissance de son fils il est resté très disponible et m'a consacrée beaucoup de temps, tout cela dans un cadre détendu et agréable.

J'aimerais également remercier **Caroline Paccard** d'avoir supervisé mon travail, d'avoir été toujours très disponible et à l'écoute, et de m'avoir fortement aidée dans ma recherche de thèse et VIE et dans les démarches administratives, ainsi que **Jonas Mandel** qui m'a encadrée en début de stage.

Je voudrais de même remercier toute l'équipe *Biomarker Statistics* pour les échanges, que ce soit à propos du stage, de leurs missions, ou des conseils pour l'avenir. C'est la participation de tous qui a fait de mon stage une expérience si enrichissante. Leur aide a également été précieuse dans la finalisation de l'onglet d'analyse de corrélations. Merci spécialement à **Lucie Brolon**, **Carole Ishimwe** et **Emilie Gérard** pour ces temps sur site et cette bonne ambiance de travail.

Merci aussi à **Mathieu Emily** pour son suivi et ses conseils tout au long du stage, ainsi que pour son soutien dans mes candidatures pour la suite.

Finalement, je voudrais remercier l'équipe de Sanofi Pasteur avec qui nous avons collaboré et qui m'a donné de précieux conseils pour le développement de l'onglet d'analyse de pathways.

Table des matières

Glossaire

ADNc : Acide Désoxyribo-Nuléique complémentaire.

ARNm : Acide Ribo-Nucléique messager. Sert principalement d'intermédiaire entre l'ADN et les protéines.

Barplot : Diagramme en barres.

Biomarqueur : Un biomarqueur est un élément du corps ou de ses produits, pouvant aller d'un pouls à une expression génétique, qui peut avoir des conséquences sur une maladie. Sa mesure peut alors permettre d'anticiper un résultat ou une évolution dans une maladie. Un biomarqueur omique concerne des données génomiques (ADN), transcriptomiques (ARN), protéomiques (protéines) ou métabolomiques (métabolites).

CPM : Counts Per Million – comptes par million. C'est une méthode de normalisation pour les données RNAseq.

Dotplot : Diagramme en points.

Enrichment map : L'enrichment map est un graphique de type réseau, permettant d'illustrer les résultats d'analyses de pathways en visualisant aussi les liens entre les pathways (gènes en commun).

Etude clinique : Lors d'une étude clinique, un ou plusieurs produits (éventuellement en action simultanée) sont testés sur des sujets humains sains et d'autres atteints d'un problème de santé. Une étude se découpe en plusieurs phases. La première se fait sur une dizaine de sujets, et a pour but de déterminer le mode et la fréquence d'administration du traitement. Lors de la deuxième, l'innocuité (non-toxicité) du traitement est toujours testée, et son efficacité commence à être testée. Cela se fait sur plusieurs centaines de sujets. Finalement dans la phase III, le traitement est testé à plus grande échelle, sur plus de 1000 sujets. Elle est suivie de l'Autorisation de Mise sur le Marché (AMM) et éventuellement d'une phase IV de surveillance des effets à long terme.

Fold change : décrit le changement d'une quantité (ici expression de gène ou de protéine) entre 2 conditions, par exemple entre 2 points de temps. Le fold change de A à B est B/A.

GSEA : Gene Set Enrichment Analysis – analyse d'enrichissement d'ensembles de gènes. La GSEA est une méthode d'analyse de pathways utilisant une liste ordonnée de gènes.

GO : Gene Ontology – ontologie de gènes. C'est une bibliothèque de pathways libre d'accès.

KEGG : Kyoto Encyclopedia of Genes and Genomes – encyclopédie de Kyoto des gènes et génomes. C'est une bibliothèque de pathways libre d'accès.

MSigDb : Molecular Signature Database – base de données de signatures moléculaires. C'est une bibliothèque de pathways libre d'accès.

ORA : Over-Representation Analysis – analyse de sur-représentation. L'ORA est une autre méthode d'analyse de pathways utilisant un sous-ensemble de gènes d'intérêts considérés comme différentiellement exprimés, donc significatifs.

Pathway de gènes : Un pathway de gènes correspond à une voie biologique, c'est-à-dire de groupes de gènes rassemblés pour des raisons fonctionnelles, telles que l'implication dans les voies de création d'énergie, ou caractéristiques, telles que la position sur un chromosome par exemple.

qPCR : quantitative Polymerase Chain Reaction – réaction en chaîne par polymérase quantitative. C'est une méthode de quantification de l'ADN.

RNAseq : séquençage d'Acide Ribo-Nucléique. La RNAseq est le séquençage de l'ARN, c'est-à-dire une méthode pour expliciter la séquence (l'enchaînement de nucléotides) d'une portion d'ARN cible.

Liste des figures

Introduction

Sanofi, entreprise pharmaceutique fondée en 1973 après de nombreuses fusions, est un des leaders français mais également mondiaux dans son secteur. Avec ses plus de 100,000 employés, Sanofi est implantée dans 90 pays, et ses solutions de santé sont disponibles dans 170 pays. Cette entreprise possède deux pôles principaux : production, et recherche & développement (R&D). Il existe 69 sites de production et 21 sites de R&D (Sanofi, 2020). La R&D s'organise en aires thérapeutiques avec 4 axes principaux de recherche : (i) oncologie et immuno-oncologie (cancer), (ii) immuno-inflammation (ex : eczéma, asthme), (iii) maladies rares (hématologiques (ex : hémophilie) ou neurologiques (ex : sclérose en plaque, Parkinson)), (iv) vaccins (Sanofi Pasteur) (Sanofi, 2021). Au sein de la R&D, le département de *Biostatistics & Programming* est formé des équipes de statistiques et de programmation, dont fait partie l'équipe *Biomarker Statistics*. Cette équipe se charge des analyses de biomarqueurs omiques de grande dimension, sur toutes les phases cliniques d'une étude, c'est-à-dire de la phase I à la phase IV (Paccard, 2018).

L'organisation mondiale de la santé a défini en 2001 un biomarqueur comme « toute substance, structure ou processus qui peut être mesuré dans le corps ou ses produits et influencer ou prédire l'incidence d'un résultat ou d'une maladie ». Un biomarqueur peut donc être chimique, physique ou biologique. Cela peut aller d'un pouls à une expression génétique. L'intérêt d'un biomarqueur est d'être objectif et quantifiable, c'est donc un indicateur fiable dont la mesure est reproductible. Dans le cadre d'une analyse statistique de biomarqueurs, le but est d'élucider la relation entre eux et un ou plusieurs résultat(s) clinique(s) (Strimbu & Tavel, 2010). Dans l'équipe *Biomarker Statistics*, les biomarqueurs sont de type omique. Cela correspond à des données biomarqueurs de grande dimension sur l'acide désoxyribonucléique (ADN, données génomiques), l'acide ribonucléique messager (ARNm, données transcriptomiques), les protéines (données protéomiques), et la biochimie/les molécules (données métabolomiques) (Institut Frédéric Joliot, 2020).

La visualisation des résultats d'analyse de ces données est une étape clé de la communication avec les cliniciens et de l'interprétation des résultats. Elle peut être améliorée en ajoutant de l'interactivité aux résultats, ce qui est notamment possible à travers une application web. Il existe un package du logiciel R, l'outil principalement utilisé pour cette visualisation et communication à travers l'interface de RStudio, qui permet de coder de telles applications ; c'est le package *shiny* (RStudio, 2020). Se pose alors la question de l'utilisation de cet outil pour servir au mieux à cette étape clé de la communication : la visualisation.

Comment visualiser idéalement les résultats d'analyses de données biomarqueurs à travers une application R shiny ?

Pour répondre à cette problématique, je commencerai par une partie de bibliographie et contexte qui permettra de présenter les principales analyses faites par l'équipe *Biomarker Statistics* sur des données biomarqueurs puis de décrire plus en détails une analyse précise : l'analyse de pathways, ou analyse de voies biologiques. Par la suite, je présenterai la structure des données de l'étude utilisée pendant le stage, ainsi que les outils utilisés (R et l'application BIOexplorer) dans une partie matériel et méthodes. Ensuite, une partie résultats et interprétation me permettra de présenter le développement d'un module de l'application sur l'analyse de pathways et d'en justifier les choix. Finalement, je discuterai des challenges rencontrés dans le développement de l'application ainsi que des projets pour la suite.

I. Bibliographie et contexte

A. L'analyse statistique des données biomarqueurs

Durant mon stage j'ai travaillé principalement sur un type de biomarqueurs omiques : des **données transcriptomiques** générées par séquençage d'ARN (RNAseq). Afin de bien comprendre comment ces données sont générées, j'ai eu l'opportunité de visiter les laboratoires qui génèrent ces données de séquençage à partir des échantillons (de sang ou de tumeur par exemple), avec les explications du chercheur en charge des expériences.

i) Les données biomarqueurs RNAseq

La technologie de séquençage de l'ARN (RNAseq) est une méthode de séquençage haut débit permettant d'avoir une idée précise des **niveaux d'expression des gènes**.
Cela nécessite tout d'abord la création d'une **bibliothèque RNAseq**, qui constituera le matériel de mesure. L'ARN est extrait et isolé des cellules, puis converti en ADN complémentaire (ADNc) car l'ADN est plus stable que l'ARN. Ces ADNc sont fragmentés et des amorces sont ajoutées à chaque extrémité des brins. Les amorces sont de courtes séquences d'ADN connues qui serviront par la suite à la fixation des fragments sur la puce. Les échantillons sont ensuite amplifiés par Réaction en Chaîne par Polymérase (PCR) quantitative (qPCR), une duplication exponentielle et maîtrisée. Ces manipulations donnent la bibliothèque RNAseq (Montgomery & Kukurba, 2015) (Ozsolak & Milos, 2011).
La technologie Illumina permet d'extraire les séquences de chaque fragment. Les amorces permettent la liaison des brins sur une puce puis un séquençage par synthèse a lieu. Les nucléotides servant à la synthèse sont fluorescents et l'enchaînement de couleurs donne la séquence. Sont alors obtenues les séquences de chaque fragment qu'il s'agira ensuite de localiser sur le génome, c'est-à-dire de reconnaître les gènes auxquels ils appartiennent. Les échantillons sont alignés sur les séquences correspondantes de gènes, puis assemblés et les transcrits complets sont comptés. Est alors obtenu un comptage pour chaque gène. En faisant l'analyse sur les échantillons des différents patients, une matrice des **données de comptage** est obtenue, les gènes en lignes et les échantillons (patients) en colonnes (Montgomery & Kukurba, 2015) (Ozsolak & Milos, 2011).
Les données nécessitent un traitement important notamment car, étant des données de comptage, leur distribution ne se rapproche pas d'une loi normale mais d'une loi de Poisson ou d'une loi binomiale négative (Anjum, et al., 2016). Après un filtrage sur les gènes en utilisant un seuil d'expression minimale, les données doivent être **normalisées**. En effet, la normalisation permet de corriger les données par rapport à des sources de variabilité telles que la longueur des fragments de la bibliothèque RNAseq, ou la profondeur de séquençage. De nombreuses méthodes de normalisation existent et sont utilisées selon les données. Le travail est souvent effectué sur les données normalisées en Counts Per Million (CPM) et transformées au logarithme (Law, Chen, Shi, & Smyth, 2014) (Montgomery & Kukurba, 2015).
Un **contrôle qualité** est ensuite effectué pour s'assurer qu'il n'existe pas de biais technique tel que le jour de l'expérience ou le lot de réactif. Si un tel effet est découvert, il sera alors pris en compte dans le modèle en tant que covariable pour ne pas le confondre avec un effet recherché. Ainsi, par exemple à travers une Analyse en Composantes Principales (ACP), il est possible d'observer si une variable influe sur les données alors que l'effet n'est pas intéressant à prendre en compte, comme un effet de la plaque de mesure (Paulson, et al., 2017).

ii) Les analyses biomarqueurs

Des biomarqueurs sont mesurés sur les patients lors des études cliniques : au début de l'étude clinique, c'est-à-dire avant le traitement (baseline), puis après le traitement à différents points

de mesure au long de l'étude. En fonction de la pathologie, les biomarqueurs peuvent être mesurés par exemple dans le sang, dans une tumeur ou dans le liquide céphalo-rachidien. Selon la phase clinique, les patients seront traités par différentes doses du ou des produit(s) ou sous placebo. Trois analyses principales sont effectuées sur les biomarqueurs omiques, tels que les données RNAseq décrites précédemment, par l'équipe *Biomarker Statistics*.
L'équipe cherche d'abord à savoir si chaque biomarqueur est **régulé par le traitement au cours du temps**, c'est-à-dire s'il existe un effet traitement sur le changement dans l'expression du biomarqueur au cours du temps. Le modèle est alors longitudinal et le biomarqueur est appelé pharmacodynamique (PD) si l'effet est significatif. Cette analyse peut servir à confirmer l'action du traitement, à trouver sa dose optimale, ou à développer des thérapies combinées (associations de traitements) par exemple (Rabbee, 2020).
L'équipe regarde également si chaque biomarqueur est **prédictif de la réponse au traitement**, c'est-à-dire s'il existe un effet de l'expression initiale du biomarqueur chez les patients traités sur le critère d'évaluation de l'impact du traitement. Ce sont des modèles univariés car il y en a un par biomarqueur. Un exemple de biomarqueur prédictif binaire serait un gène muté chez certains patients et un traitement qui serait efficace seulement chez les patients avec ce gène non muté (Rabbee, 2020) (Blangero, 2019).
Finalement, il s'agira d'étudier si les biomarqueurs sont **pronostiques de l'évolution de la maladie**. Le but est ici de savoir s'il est possible de déduire de l'expression d'un biomarqueur la progression de la maladie, indépendamment du traitement. Ce sont de nouveau des modèles univariés (Rabbee, 2020) (Jiao, Li, Liu, Chen, & Liu, 2019).

Un grand nombre d'autres analyses peuvent être faites sur les biomarqueurs, comme des analyses de **corrélations**. Par exemple, entre l'expression initiale des biomarqueurs et les variables cliniques initiales. Cela peut permettre d'évaluer si certains biomarqueurs sont reliés à ces variables cliniques, information qui peut servir à l'interprétation ou pour de futurs projets (Vlassenko, et al., 2016). L'équipe effectue aussi des analyses non-supervisées telles que des clustering hiérarchiques ou k-means, et des analyses multivariées supervisées permettant de trouver des combinaisons de biomarqueurs pouvant prédire l'impact du traitement, telles que des régressions LASSO ou des algorithmes de Random Forest.

B. L'analyse de pathways

Les trois analyses majeures présentées précédemment de biomarqueurs régulés par le traitement, prédictifs de la réponse au traitement ou pronostiques de l'évolution de la maladie donnent des valeurs statistiques affectées à chaque gène, notamment une pvalue et une valeur de magnitude d'un effet. Il est alors possible d'interpréter les résultats en faisant ressortir des **gènes d'intérêt**, considérés comme régulés, ou ayant un effet prédictif ou pronostique, en les filtrant par rapport à ces valeurs (pvalue et magnitude d'effet). Mais une liste de gènes n'est pas très facile à interpréter, car cela demande de s'intéresser à chaque gène individuellement et d'en connaître les propriétés. C'est pour cela qu'a été développée **l'analyse de pathways**. Cette analyse permet en effet de faire ressortir des pathways d'intérêt plutôt qu'une liste de gènes, c'est-à-dire des groupes de gènes se rapprochant les uns des autres par leurs fonction ou par des caractéristiques physiques, ce qui réduit la complexité de l'analyse et donc de l'interprétation (Khatri, Sirota, & Butte, 2012).
De plus, l'analyse de pathways permet d'obtenir un résultat plus stable que l'analyse gène par gène, soumise à moins de variabilité du fait de son agrégation à un plus haut niveau (Abatangelo, et al., 2009).
Je présenterai par la suite deux des méthodes d'analyse de pathways les plus utilisées : **l'Over-Representation Analysis** (ORA) et **la Gene Set Enrichment Analysis** (GSEA).

i) Over-Representation Analysis (ORA)

L'analyse de pathways la plus classique et ancienne est l'**Over-Representation Analysis** (ORA) : l'analyse de sur-représentation. Dans le cadre de cette analyse il s'agit de sélectionner un sous-ensemble de **gènes d'intérêts**, le plus souvent selon des seuils de significativité (pvalue) et de magnitude, puis d'étudier si, dans ce sous-ensemble, des **pathways sont sur-représentés**. Pour effectuer ce test, deux étapes sont nécessaires :

- Compter le nombre de gènes du pathway parmi les gènes d'intérêts, que l'on nomme A,
- Compter le nombre de gènes du pathway parmi les gènes de référence (gènes d'arrière-plan, c'est-à-dire tous les gènes étudiés), noté B (Khatri, Sirota, & Butte, 2012).

L'hypothèse suivante est alors faite : un pathway est considéré comme enrichi si la proportion A/B est supérieure à la proportion attendue en cas de distribution aléatoire. Le niveau de confiance de différence est évalué par des tests statistiques, le plus souvent selon la distribution **hypergéométrique**, mais parfois également avec un test du chi2, loi binomiale ou test exact de Fisher (Garcia-Campos, Espinal-Enriquez, & Hernandez-Lemus, 2015).
Comme il y a souvent de nombreux pathways, une **correction de tests multiples** est ensuite effectuée, pour conserver la précision globale voulue malgré le grand nombre de tests. Une des méthodes pouvant être utilisée est l'ajustement de Bonferroni, qui est une méthode très conservatrice, donc sévère, qui calcule la pvalue ajustée seulement à partir de la pvalue d'origine et du nombre total de pvalues calculées. Mais cette méthode est très restrictive et peut donc mener à une forte augmentation du taux de faux négatifs. Ici les faux négatifs correspondent à des pathways dits non-enrichis alors qu'en réalité ils auraient pu être considérés comme enrichis. C'est pourquoi une autre méthode est plus utilisée dans cette situation, c'est l'ajustement de **Benjamini-Hochberg**, qui contrôle le taux de faux positifs (False Discovery Rate – FDR). En effet, cette méthode est décrite comme peu conservative, c'est-à-dire peu sévère, et permet donc de garder un plus large spectre de résultats, ici de pathways. Cette procédure utilise, pour chaque calcul de pvalue ajustée, la pvalue d'origine, le nombre de tests, donc de pvalues calculées au total, et le rang de cette pvalue d'origine lorsque les pvalues sont rangées dans l'ordre croissant. Elle prend donc en compte à la fois le nombre de tests et le taux de pvalues non-significatives (Boyle, et al., 2004) (Benjamini & Hochberg, 1995).

Bien que très utilisée, l'ORA reste **limitée**, notamment parce que les gènes sont tous traités de la même manière et considérés comme indépendants, ainsi que les pathways. Or il existe de nombreuses interactions entre les gènes, et entre ces pathways. De plus, l'ORA nécessite le choix de seuils pour la sélection des gènes d'intérêt puisqu'il faut choisir une pvalue maximale et une valeur de magnitude minimale. Fixer un seuil est une forte contrainte, et cela peut par exemple écarter des gènes très proches des limites qu'il aurait été bon de prendre en compte (Khatri, Sirota, & Butte, 2012).

Cette méthode, ORA, sera présentée dans la suite du rapport. En parallèle, j'ai également eu l'occasion d'étudier une autre méthode d'analyse de pathways, la GSEA, que je détaille dans la partie suivante.

ii) Gene Set Enrichment Analysis (GSEA)

La **Gene Set Enrichment Analysis** (GSEA), analyse d'enrichissement d'ensemble de gènes, est une autre méthode d'analyse de pathways. Cette méthode, au lieu de sélectionner un sous-ensemble de gènes d'intérêt, va travailler sur l'ensemble des gènes et de leurs statistiques. Elle repose sur **une liste de gènes ordonnés** (L) selon une valeur, qui peut être la pvalue, le fold change, ou tout autre critère statistique. Ensuite, pour chaque pathway, le but est d'évaluer si les gènes de ce pathway, notés S, sont répartis aléatoirement dans la liste ordonnée L, ou plutôt au début et/ou à la fin (aux extrêmes). Afin d'évaluer cela, plusieurs étapes sont nécessaires :

- La première étape est le calcul d'un **score d'enrichissement** (ES) qui déterminera si le pathway est très représenté dans les extrêmes de la liste L. En effet, pour l'obtenir, il s'agit de parcourir L dans l'ordre en augmentant au fur et à mesure le score lorsqu'un gène de S est rencontré et en le diminuant pour les gènes n'appartenant pas à S. De plus, le coefficient ajouté ou enlevé est pondéré en fonction du rang, ce qui permet de donner une plus grande importance aux extrêmes de la liste. La méthode est une variation de la statistique de Kolmogorov-Smirnov. Le ES est alors le maximum atteint par le score calculé précédemment (cf Figure 1).
- Ensuite, ce ES est normalisé par rapport à la taille du pathway correspondant pour donner la même importance aux pathways quelle que soit leur taille. Est alors obtenu un **score d'enrichissement normalisé** (NES).
- Pour finir, la significativité de ce score est calculée par permutations, et la pvalue est corrigée pour tests multiples par le False Discovery Rate (FDR).

(Subramanian, et al., 2005) (Reimand, et al., 2019) (Abatangelo, et al., 2009) (Jiao, Li, Liu, Chen, & Liu, 2019)

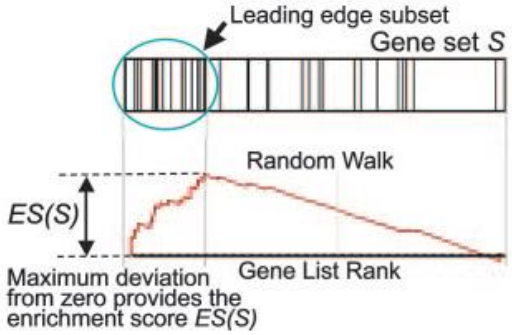


Figure 1: Calcul du score d'enrichissement

La barre horizontale représente la liste ordonnée de gènes (L), et les traits verticaux les gènes du pathway (S). Le graphique du dessous représente le calcul du score d'enrichissement le long du parcours de la liste L. (Subramanian, et al., 2005)

Cette méthode a tout de même ses **limites**, notamment car, comme l'ORA, elle considère les pathways de façon indépendante, alors que ces derniers peuvent se recouper, ce qui peut influer sur les résultats (Khatri, Sirota, & Butte, 2012).

iii) Les bases de données de pathways

Pour les deux méthodes d'analyse de pathways présentées précédemment, un grand nombre de bibliothèques de pathways sont disponibles. Les deux bibliothèques les plus connues et utilisées sont **Gene Ontology** (GO) et **Kyoto Encyclopedia of Genes and Genomes** (KEGG).
Le projet GO a commencé en 1998 et la base de données est constituée de 3 groupes : fonction moléculaire (MF), processus biologiques (BP), et composants cellulaires (CC). MF contient des pathways d'activités au niveau moléculaire, par exemple des activités catalytiques. Les pathways de BP sont des processus à plus haut niveau, comme l'apoptose (mort d'une cellule). Finalement, CC comprend les pathways décrivant des localisations, au niveau interne à la cellule ou au niveau de complexes macromoléculaires (ex : membrane interne du noyau) (Gene Ontology Consortium, 2004).
Le projet de KEGG a été lancé en 1995. Le but était de développer les descriptions fonctionnelles par gènes, c'est-à-dire de lier l'information génétique à l'information fonctionnelle. Cela est fait en 3 bases de données : PATHWAY la collection de pathways, donc à un ordre fonctionnel supérieur, GENES la collection de gènes, et LIGAND la collection de composants chimiques de la cellule. Chacun des pathways peut être visualisé sous forme de carte, et ce sont principalement des voies métaboliques (Kanehisa & Goto, 2000).

Il existe depuis 2005 une base de données très générale appelée **Molecular Signature Database** (MSigDb), qui contient GO et KEGG, ainsi que d'autres collections. MSigDb est adaptée à l'analyse de pathways car elle a été créée exclusivement pour cela. Elle a une grande

diversité et regroupe le plus grand nombre de pathways (Liberzon, et al., 2011). Elle en contient actuellement plus de 32000 répartis en 8 collections, la plupart comprenant des sous-collections. Les collections portent sur la position des gènes sur les chromosomes, les cibles potentielles de régulation par les facteurs de transcription ou micro-ARN, des collections concernant les cancers, la collection GO, et d'autres éléments intéressants (Broad Institut, 2021). C'est majoritairement cette base de données qui sera utilisée dans la suite du rapport.

Bien qu'une base de données telle que MSigDb soit très complète, il est également possible d'avoir sa propre bibliothèque de pathways, ou simplement un ou deux **pathways personnalisés** sur lesquelles travaille une équipe. Par exemple chez Sanofi, certaines équipes ont pu développer des bibliothèques personnelles concernant exclusivement leurs analyses ou thématique(s) d'intérêt, et veulent avoir les résultats des analyses d'enrichissement sur leur propre bibliothèque.

Lors de la manipulation des données d'expression à analyser, ainsi que des données extraites des bibliothèques de pathways, il est important de prêter attention à **l'identifiant de gènes** utilisé. En effet, les gènes n'ont pas un identifiant universel, il en existe plusieurs, qui ne correspondent pas toujours de façon exacte, ce qui peut causer des pertes de gènes lors des transitions entre les identifiants. Les 3 identifiants utilisés dans ce rapport et dans mes analyses sont « ensembl », « symbol » et « entrezid ». Toutes les bases de données évoquées précédemment contiennent ces 3 identifiants. Les « ensembl » sont des identifiants assez longs, et stables, c'est-à-dire qu'ils ne changent pas avec l'évolution des connaissances (ex : ENSG00000000003) (Ruffier, et al., 2017). Les « symbol » sont beaucoup plus courts (ex : TSPAN6), et sont développés par le Comité « HUGO » de nomenclature du gène (HGNC) (Povey, et al., 2001). Finalement, les « entrezid » sont numériques (ex : 7105), c'est une nomenclature développée par le Centre Américain pour l'information de la Biotechnologie (NCBI) (Maglott, Ostell, Pruitt, & Tatusova, 2011).

Suite à cette présentation des données biomarqueurs, des types d'analyses majoritairement effectuées, et de l'analyse de pathways de gènes, voyons maintenant plus précisément les outils et données utilisés durant mon stage.

II. Matériel & Méthodes

A. Outil utilisé

i) BIOexplorer, une application Rshiny pour visualiser les résultats d'analyses de données biomarqueurs

Comme nous avons pu voir dans la première partie, les analyses de données biomarqueurs omiques donnent des résultats de très grande dimension, car pour chaque biomarqueur (ici gène) l'objectif est de chercher s'il est régulé par le traitement, prédictif de la réponse au traitement et/ou pronostique de l'évolution de la maladie. Cela représente donc beaucoup de graphiques et tableaux de résultats. Par exemple, si 3 analyses sont faites sur 15000 gènes, environ 45000 graphiques peuvent être produits. La communication aux équipes cliniques se fait traditionnellement via des rapports ou présentations power point mais ce format donne des fichiers très longs, non exhaustifs, et manquant de flexibilité et d'interactivité. Pour pallier à ce problème, une application web R shiny, nommée BIOexplorer pour **Biomarkers Interactive Outputs explorer**, a été développée par l'équipe pour la visualisation et communication des résultats. Grâce au package *shiny*, l'application est déployée et peut être partagée via une adresse web (url). Une fois déployée l'application est donc accessible par tous les utilisateurs dont l'accès est autorisé et disposant du lien sécurisé.

Afin de faciliter le codage en équipe et la manipulation du code complexe et long, l'application est structurée en modules, c'est-à-dire en codes indépendants (un par onglet) pouvant être appelés dans le code principal (global). Cela permet de diviser le long code en plus courtes parties et de minimiser le risque que plusieurs membres de l'équipe modifient simultanément le même fichier R (RStudio, 2020). De plus, les mises à jour sont effectuées régulièrement sur un compte git.
BIOexplorer est divisée en 3 onglets principaux :

- **Résultats cliniques** : décrit les données cliniques, c'est-à-dire concernant les patients, avec 3 sous-onglets : un décrivant les différentes populations présentes dans l'étude clinique en question, un décrivant les variables démographiques de cette même étude, et un décrivant les critères primaires d'évaluation de l'impact du traitement (endpoints) mesurés dans l'étude clinique en question ;
- **Résultats omiques** : présente les analyses effectuées sur les données omiques. L'utilisateur peut d'abord y trouver un sous-onglet concernant l'identification de biomarqueurs régulés par le traitement (pharmacodynamiques), puis un autre concernant l'identification de biomarqueurs prédictifs de la réponse au traitement et/ou pronostiques de la progression de la maladie ;
- **Informations & paramètres** : donne à l'utilisateur des informations sur l'équipe *Biomarker statistics*, sur la session R utilisée ainsi que les versions des packages dans 2 sous-onglets, et lui donne la possibilité de personnaliser la couleur de chaque bras de traitement et chaque visite dans un 3^{e} sous-onglet.

C'est au niveau de l'onglet « Résultats omiques » que j'ai travaillé durant mon stage, en ajoutant deux nouveaux types d'analyses et donc 2 sous-onglets : un sous-onglet « Analyse de corrélations » et un sous-onglet « Analyse de pathways ». L'onglet « Analyse de corrélations », qui présente les résultats d'analyse de corrélations ou associations entre les variables cliniques et les biomarqueurs avant traitement, notamment sous forme d'une heatmap interactive, ne sera pas détaillé dans ce rapport. Je vais cependant présenter plus en détails l'onglet « Analyses de pathways ». Celui-ci comprend 2 parties car 2 types d'analyses, présentées dans la partie bibliographie & contexte : Over-Representation Analysis (ORA) et Gene Set Enrichment Analysis (GSEA). Principalement l'onglet « ORA » a été développé durant mon stage. C'est donc cet onglet qui sera détaillé dans la suite du rapport.

ii) Les packages et fonctions R de l'ORA

Le package R utilisé pour les analyses de pathways, notamment pour l'ORA, est la version 3.14.3 de ***clusterProfiler***. Ce package permet d'effectuer les différents types d'analyses de pathways voulus avec un large choix de bibliothèques de pathways. Il existe des fonctions spécialisées par bibliothèque, mais pour avoir un code plus léger et flexible, j'ai principalement utilisé la fonction *enricher*, qui permet de faire une ORA de manière générale. Cette fonction prend en entrée :

- Le vecteur des gènes d'intérêt (gene),
- Le vecteur des gènes de référence (universe),
- Un tableau de 2 colonnes : une comprenant les noms de pathways et une autre les noms de gènes, chaque ligne correspondant à un couple pathway-gène (TERM2GENE), dont la création est détaillée ci-dessous,
- Eventuellement un tableau de correspondance entre les identifiants des pathways et leurs noms complets (TERM2NAME),
- Une méthode d'ajustement de pvalue (principalement Benjamini Hochberg utilisé) (pAdjustMethod),
- Des limites de tailles de pathways à analyser (minGSSize, maxGSSize),
- Des seuils de pvalue et qvalue pour sélectionner les pathways considérées comme significatives (pvalueCutoff, qvalueCutoff).

Ainsi, au niveau de TERM2GENE il est possible de personnaliser les pathways utilisés. C'est principalement ici que se situe la différence avec les fonctions spécifiques comme *enrichGO* ou *enrichKEGG*, qui définissent automatiquement la source de pathways GO ou KEGG, respectivement. TERM2GENE peut également contenir une liste de pathways personnels, comme évoquée précédemment.

Lors de l'utilisation de la base de données MSigDb, j'utilise le package ***msigdbr*** permettant la navigation entre les différentes collections et l'extraction des données nécessaires à TERM2GENE. Ce package contient la fonction *msigdbr* qui permet d'accéder à la catégorie et sous-catégorie voulue, et de choisir l'identifiant de gènes. Il est alors possible d'obtenir directement un tableau de 2 colonnes correspondant à TERM2GENE.

La sortie de la fonction *enricher* est un objet de type enrichResult, qui comprend plusieurs éléments. Ces éléments reprennent les paramètres d'entrée de la fonction, ainsi que les résultats de l'analyse. L'élément le plus utile est 'result', un tableau de 8 colonnes (9 si un TERM2NAME est précisé) qui donne les résultats de l'ORA, une ligne par pathway analysé (cf Figure 9, page 9). Les 8 colonnes sont :

- *Description* : l'identifiant du pathway,
- *GeneRatio* : le ratio de gènes du pathway parmi les gènes d'intérêt par rapport au nombre de gènes d'intérêt,
- *BgRatio* : le ratio de gènes du pathway par rapport au nombre de gènes donnés dans l'univers (gènes de référence),
- *Count* : le nombre de gènes du pathway dans les gènes d'intérêt,
- *pvalue* : la pvalue du test d'enrichissement pour chaque pathway,
- *p.adjust* : la pvalue ajustée du test d'enrichissement (test multiple),
- *qvalue* : la qvalue du test d'enrichissement,
- *geneID* : la liste des gènes du pathway appartenant aux gènes d'intérêt.

Finalement, c'est le package ***enrichplot*** qui est utilisé pour tracer les graphiques des résultats car il est adapté aux résultats de *clusterProfiler*. En effet, les fonctions du package *enrichplot* prennent en entrée un objet de type enrichResult et tracent les graphiques selon les bons paramètres, comme nous le verrons dans la partie résultats du rapport (Yu, Wang, Han, & He, 2012).

B. Présentation des données précises pour cette étude et ce module

Au cours du stage une seule étude a été utilisée comme base de travail. Pour une question de confidentialité, la pathologie et les traitements testés ne peuvent être détaillés. Je présenterai donc seulement ses caractéristiques et la structure des données.
Pour cette étude, nous travaillons au niveau clinique, c'est-à-dire sur des données humaines, en phase II. L'étude est randomisée en double-aveugle : ni les patients ni les cliniciens ne connaissent la nature des traitements attribués. Elle comporte **4 bras de traitements** différents, notés « Treatment A », B, C et D. La randomisation est de 1:1, ce qui signifie que le nombre de patients est identique dans chaque groupe de traitement. De plus, les données omiques sont collectées à **3 points de temps**, notés T1, T2 et T3. Finalement, **5 critères primaires** sont mesurés pour évaluer l'effet des traitements, notés « Endpoint 1 », 2, etc... Les patients sont caractérisés par **11 variables cliniques** telles que l'âge, le sexe, l'indice de masse corporelle (IMC), ou la fréquence à laquelle le patient fume (ou non).
Ce sont les données RNAseq de cette étude qui sont utilisées. Comme vu dans la première partie, l'analyse de pathways, se fait après une analyse différentielle ou de biomarqueurs prédictifs et/ou pronostiques. Les jeux de données analysés seront donc les **résultats des analyses de biomarqueurs régulés** (modèle longitudinal) **et prédictifs de la réponse au**

traitement/pronostiques de l'évolution de la maladie (modèle univarié). Pour l'ORA, une sélection des gènes considérés comme différentiellement exprimés est à réaliser. Cette sélection se fait sur 2 critères : un critère de significativité du test statistique et un de magnitude de l'effet. Pour la significativité, un seuil maximal de p-value pour les 2 jeux de données (longitudinal et univarié) peut être établi, alors que pour la magnitude des effets les variables sont différentes. En effet, la magnitude apparaît dans l'analyse longitudinale à travers le fold change, alors qu'elle apparaît dans l'analyse univariée à travers la valeur calculée de l'effet d'un traitement. Ces 2 variables seront donc respectivement choisies comme seuils minimaux de magnitude.

À la suite de cette présentation des outils et données, voyons les résultats et visualisations choisies intégrées à l'onglet « ORA » développé dans l'application BIOexplorer.

III. Résultats & interprétation

Dans cette partie, il s'agira de présenter l'onglet « ORA » développé dans l'application en justifiant les choix de visualisation. Cet onglet suit la structure des autres onglets de l'application, avec une partie de filtres où l'utilisateur peut intervenir (cf boîtes bleues (gauche) sur la Figure 2), et une partie résultats, divisée en plusieurs catégories : « Figures », « Tables » et « Summary », présentant les différents résultats (cf boîtes orange (droite) sur la Figure 2). La Figure 2 donne une vision globale de l'onglet dans la section « Figures » pour replacer les éléments présentés par la suite. Je vais maintenant détailler chacune de ces parties en me basant sur un exemple de paramètres choisis.

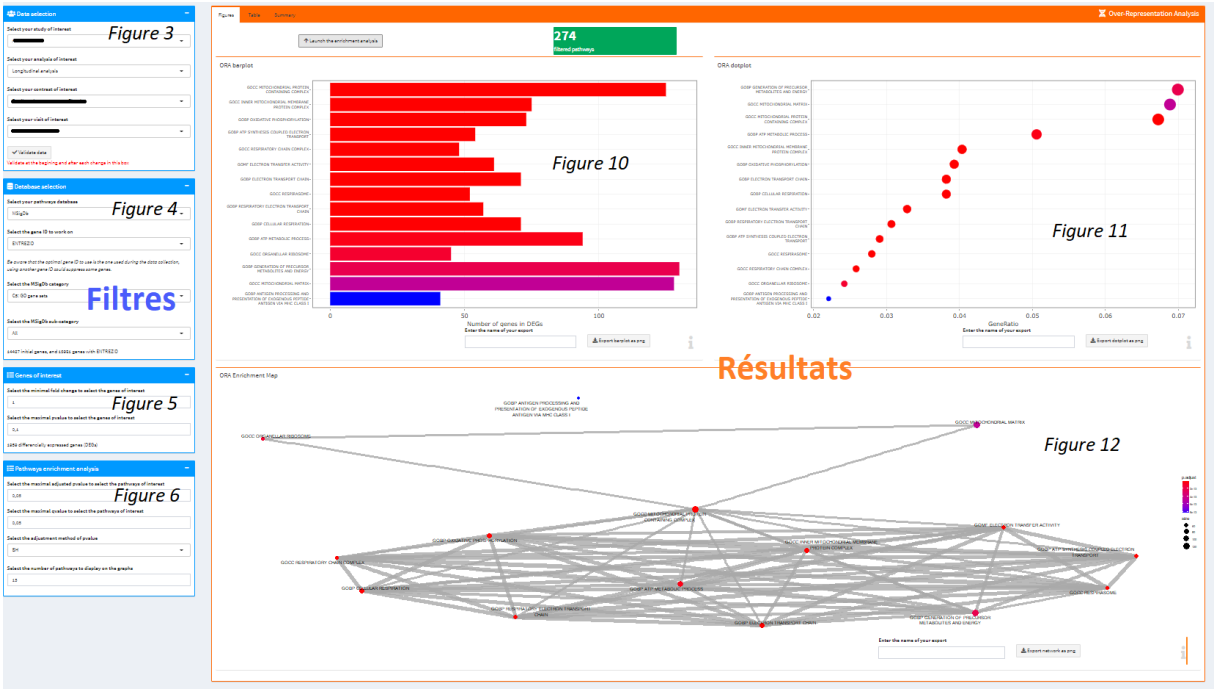


Figure 2: Capture d'écran globale de l'onglet ORA (au niveau du sous-onglet 'Figures')

A. *Une interface utilisateur-machine adaptée aux besoins des utilisateurs*

i) Une personnalisation par de nombreux filtres et paramètres

Comme précisé précédemment, le but de BIOexplorer est de fournir aux utilisateurs des résultats exhaustif et interactifs. Une grande partie de l'interactivité est permise par un grand nombre de paramètres fixés par l'utilisateur. Il s'agit ici de s'intéresser à ces paramètres, qui sont choisis dans les boîtes bleues de l'onglet.

- Le choix des données de travail

Dans la première boîte, l'utilisateur peut choisir les données sur lesquelles il souhaite travailler. En effet, comme précisé précédemment, les analyses de pathways sont faites sur les résultats des analyses de biomarqueurs régulés par le traitement (**analyse longitudinale**), ainsi que sur celle de biomarqueurs prédictifs de la réponse au traitement et/ou pronostiques de l'évolution de la maladie (**analyse univariée**). Pour l'analyse longitudinale, plusieurs traitements et temps de mesure sont disponibles. Il faut donc choisir **un contraste précis de traitement** (ex : « Treatment B vs Treatment A ») et **de temps** (ex : « T3 vs T1 »). Pour l'analyse univariée, **l'effet** à analyser (ex : effet prédictif du traitement B : « Predictive Treatment B »), ainsi que **le critère** sur lequel travailler (ex : Endpoint_1) sont à choisir. Le choix de l'étude est pour l'instant figé car je n'ai travaillé que sur une seule étude. Il existe cependant en parallèle un projet d'intégrer plusieurs études dans une même application, d'où l'utilité future de ce filtre comme mentionné en partie IV (cf Figure 3).
Pour des questions pratiques de mise à jour des données dans l'application, l'utilisateur doit valider grâce à un bouton à chaque changement dans cette boîte (cf Figure 3).

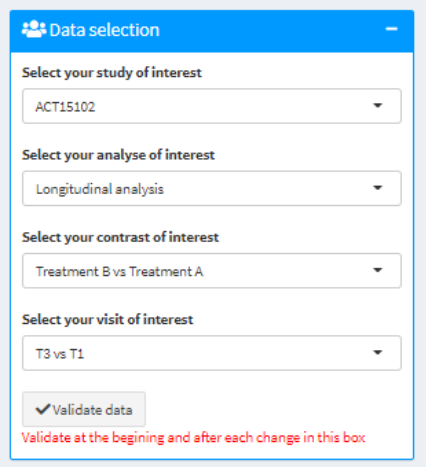
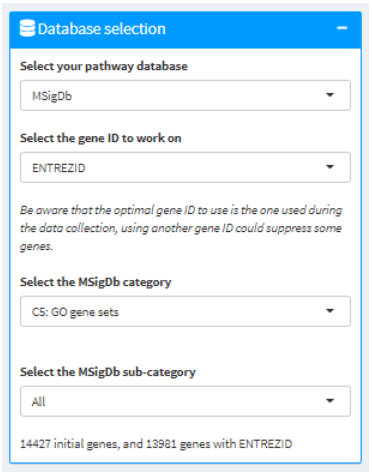


Figure 3: Capture d'écran de la boîte de filtres pour le choix des données d'étude

- Le choix de la base de données de pathways

Après avoir choisi les données à analyser, l'utilisateur doit choisir la **base de données de pathways** qu'il veut utiliser. Ici l'utilisateur a le choix entre la bibliothèque MSigDb (Figure 4a), vue en partie bibliographie, et un fichier de pathway(s) personnel (Figure 4b).
Lorsqu'il choisit de travailler avec MSigDb, l'étape suivante est le choix de la collection puis l'éventuelle sous collection, comme présenté dans la partie de bibliographie. La possibilité d'ajouter son propre fichier de pathways vient du fait que les laboratoires peuvent avoir des pathways personnels ou s'intéresser à un groupe de gènes précis. Cette option leur permet donc de tester ces ensembles de gènes en les chargeant sous forme de fichier Excel grâce à un fichier exemple structuré à télécharger (« template »). La structure est donc standardisée et adaptée au traitement qui suivra pour lancer l'analyse.
De plus, l'utilisateur doit choisir l'identifiant de gènes avec lequel il souhaite travailler : ensembl, symbol ou entrezid. Ces identifiants sont le plus universels possibles, mais il n'existe pas de correspondance exacte entre les catégories. Dans notre cas, les identifiants ensembl et symbol sont disponibles dans les données brutes, donc il se peut qu'il y ait une perte de gènes dans le cas d'une traduction en entrezid, ce qui est précisé à l'utilisateur dans la boîte.
Dans mon exemple je choisis de travailler avec la base données GO (Gene Ontology) qui est la cinquième collection de MSigDb, et d'utiliser les entrezid (cf Figure 4a).

a)



b)

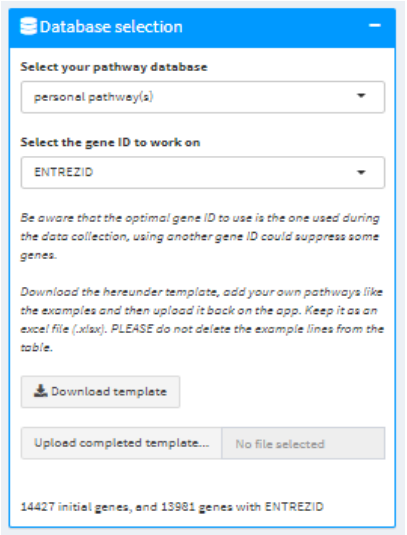


Figure 4: Capture d'écran du choix de bibliothèque de pathways a) pour MSigDb, b) pour une bibliothèque personnelle

- Les filtres pour la sélection des gènes d'intérêt

Comme expliqué dans la bibliographie, l'ORA nécessite une sélection de gènes d'intérêt d'après les résultats des analyses précédentes. Dans notre cas ce seront les gènes considérés comme régulés par le traitement (analyse longitudinale), avec un certain niveau de confiance. Comme expliqué en matériel et méthodes, la sélection se fait en considérant un seuil maximal de **pvalue** pour les deux types d'analyses, et un seuil minimal de **fold-change** pour l'analyse longitudinale, et d'**effet** pour l'analyse univariée. Dans notre exemple nous avons donc un seuil de pvalue maximale de 0.1, et de fold-change minimal de 1 (cf Figure 5).

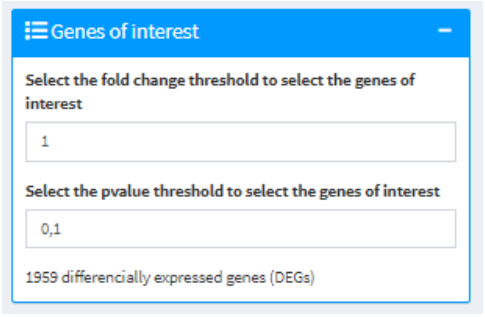


Figure 5: Capture d'écran de la boîte de filtres pour la sélection des gènes différemment exprimés

- Les paramètres de l'analyse d'enrichissement

La dernière boîte de filtres concerne les **paramètres de la fonction d'analyse d'enrichissement** du package *clusterProfiler* (*enricher*). L'utilisateur choisit une pvalue ajustée et une qvalue maximales pour filtrer les pathways considérés comme significativement enrichis. Elles sont par défaut de 0.05. Il choisit également la méthode d'ajustement de la pvalue, ici Benjamini-Hochberg par défaut. Finalement, il peut choisir le nombre de pathways affichés sur les graphiques, présentés dans la partie III.C (cf Figure 6).

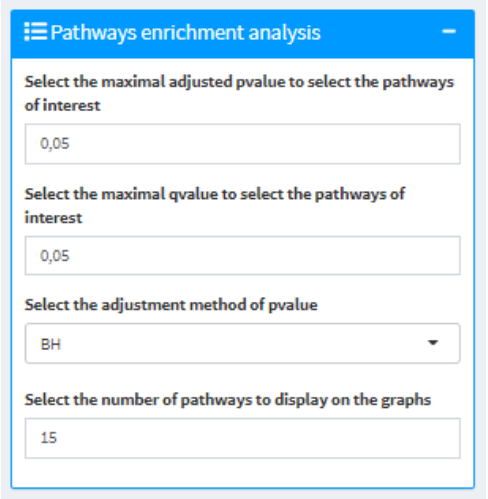


Figure 6: Capture d'écran de la boîte de filtres pour les paramètres de la fonction enricher

ii) Des graphiques interactifs pour faciliter la manipulation

Les graphiques, tracés avec la fonction *enrichplot* de *clusterProfiler*, sont des graphiques ggplot. Ainsi, grâce à la fonction *ggplotly* du package *plotly*, ces graphiques peuvent être

transformés en graphiques *plotly*, c'est-à-dire des **graphiques interactifs**. En effet, il est alors possible de zoomer et naviguer sur les graphiques. Cela peut notamment être utile lorsque l'utilisateur veut afficher un grand nombre de pathways. Le graphique est alors parfois saturé, donc zoomer puis se déplacer le long des axes peut en faciliter la lecture.
De plus, *plotly* permet d'afficher des **étiquettes** sur les données, que ce soient des points ou des barres. Ces étiquettes s'affichent lorsque la souris de l'utilisateur passe sur un élément du graphique. La légende peut alors être enlevée et ajoutée aux informations sur ces étiquettes dynamiques, ainsi que d'autres éventuelles informations (cf Figure 7a).
Finalement, l'utilisateur a la possibilité de **télécharger chaque graphique sous forme de png**. Cela peut notamment servir pour incorporer un graphique à un document, une fois les paramètres voulus choisis. C'est justement cette option que j'utilise dans ce rapport, les graphiques présentés par la suite ayant été exportés sur l'application. Dans le cas d'un export, le graphique statique est généré via le package *ggplot* de base. N'ayant plus d'interactivité, les légendes sont présentes sur le graphique car les informations ne sont plus disponibles sur les étiquettes. De plus, un titre et une description sont ajoutés avec les paramètres fixés, pour pouvoir interpréter les graphiques en connaissance de cause, même une fois exportés et sans les filtres sous les yeux (cf Figure 7b).

a)

b)

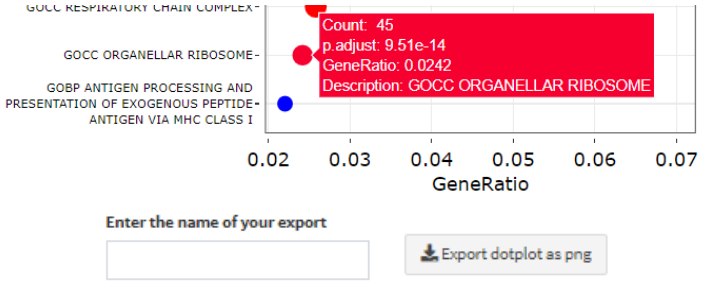


Figure 7: Captures d'écran a) de l'affichage d'une étiquette plotly ; b) de l'outils de téléchargement pour le diagramme en points

iii) Des valeurs informatives mises à jour en fonction des analyses

Finalement, la dernière spécificité de l'interface est d'afficher des **valeurs informatives** pour aider l'utilisateur. En effet, trois valeurs sont à sa disposition afin qu'il puisse se rendre compte de l'état de ses données. Elles s'affichent sous différentes formes : 2 sont sous forme de **texte** (cf Figure 8a&b) et une sous forme de **valuebox** (cf Figure 8c), un objet *shiny* qui permet d'afficher de façon esthétique une valeur.
Le premier texte affichant des valeurs se trouve en bas de la boîte de choix de bibliothèque de pathways (cf Figure 8a), et donne le nombre de gènes initial du jeu de données choisi, ainsi que le nombre de gènes une fois la traduction entre les identifiants de gènes faite. Ceci est utile lorsque l'utilisateur choisit de travailler avec les entrezid, car ils ne font pas partie du jeu de données initial. Il existe donc une légère perte de gènes au moment de la traduction. Il est important que l'utilisateur en soit conscient et puisse se rendre compte de la quantité de gènes perdus. Cette valeur se met à jour automatiquement lorsque les données et/ou l'identifiant de gènes sont modifiés.
Le deuxième texte se trouve en bas de la boîte de sélection des gènes d'intérêt (cf Figure 8b). Il permet à l'utilisateur de se rendre compte de la taille de sous-échantillon de gènes d'intérêt qu'il sélectionne. En effet, chercher un enrichissement dans un ensemble de 50 gènes d'intérêt lorsqu'il y en a 14000 au total risque d'être très peu fructueux. La Figure 8b montre qu'ici, avec un fold change minimal de 1 et une pvalue maximale de 0.1, 1959 gènes sont considérés comme régulés par le traitement, et donc sélectionnés comme gènes d'intérêt, sur les 13981 totaux.

Finalement, la dernière valeur est celle s'affichant dans une valuebox (Figure 8c). Elle correspond au nombre de pathways filtrés d'après les seuils maximaux de pvalue ajustée et de qvalue que l'utilisateur fixe dans la dernière boîte de filtres. Cette valeur est très informative car sur les graphiques l'utilisateur choisit le nombre maximal de pathways qu'il souhaite afficher. Or, ce nombre peut être plus faible que le nombre de pathways filtrés. Ainsi, les graphiques afficheront les pathways les plus significatifs, 15 par défaut, mais l'utilisateur peut se rendre compte s'il existe d'autres pathways correspondant à ses critères. Dans l'exemple, comme mentionné par la Figure 8c, 274 pathways apparaissent filtrés. Si l'utilisateur garde l'affichage de 15 pathways il y en aura donc 259 respectant les seuils mais non visualisés sur les graphiques. Il est important qu'il en soit conscient et puisse les chercher dans les tables si besoin.

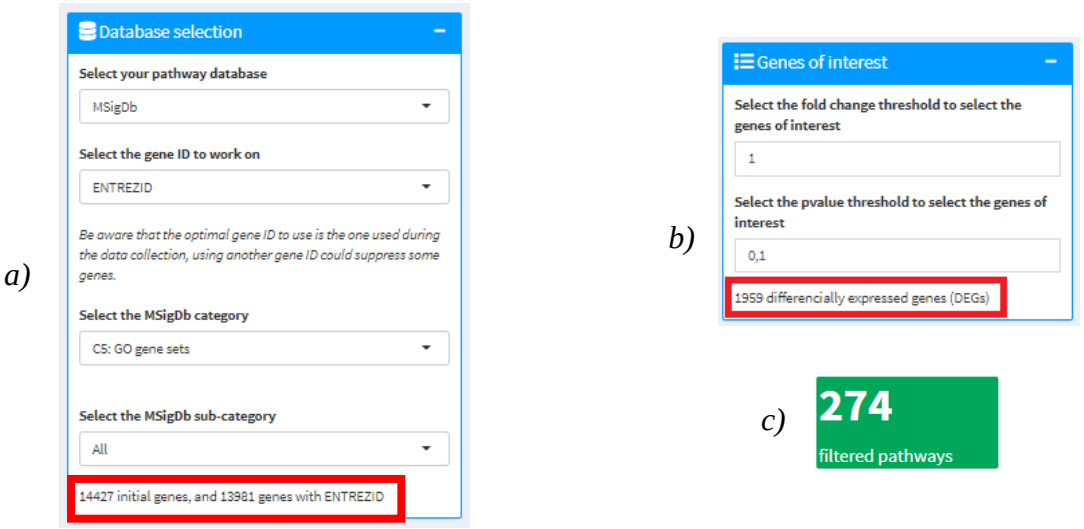


Figure 8: Captures d'écran des valeurs informatives sous forme a) de texte pour les gènes avec identifiants ; b) de texte pour les gènes d'intérêt ; c) de valuebox pour les pathways filtrés

Je vais maintenant passer à la partie résultats de l'onglet, qui présente les différents résultats des analyses, sous plusieurs formes. Comme précisé plus haut, cette partie est divisée en 3 sections : « Figures », « Tables », et « Summary ».

B. Les données sous forme de tableau

Je vais d'abord m'intéresser à la section « Tables », car il sera plus facile par la suite de comprendre les figures en visualisant la structure des données de sortie de l'ORA. En effet, l'élément principal de la section tables est le tableau de résultats obtenu en utilisant la fonction *enricher*.

Il est important de mettre à disposition les résultats **sous forme de tableau** dans l'application pour plusieurs raisons. Tout d'abord, l'utilisateur peut le **télécharger**, sous forme de CSV, Excel ou PDF, grâce au bouton 'Download' en haut à gauche de la Figure 9. Disposer des résultats exportés en tableau permet à l'utilisateur de les manipuler, et d'éventuellement les réutiliser à d'autres fins. De plus, avoir les données permet parfois de mieux comprendre les graphiques ou de disposer de plus d'informations. Surcharger un graphique en voulant mettre trop d'informations peut en effet le rendre illisible. Par exemple, le graphique peut nous permettre de remarquer un certain pathway d'intérêt, et grâce au tableau il est ensuite possible de chercher la liste des gènes de ce pathway faisant partie des gènes d'intérêt, dans la colonne geneID présentée dans la partie Matériel&Méthodes.

Finalement, le tableau est disponible grâce au package R ***datatable***. Cela permet d'afficher un tableau tout en disposant de nombreuses **options interactives**. Tout d'abord, une option de recherche est disponible pour chaque colonne. Pour les colonnes de caractères, la recherche se

fait sur une chaîne de caractères. Pour les colonnes numériques, la recherche se fait sous forme de filtre, dont le fonctionnement est expliqué dans l'onglet d'aide sous le tableau. Il existe également un filtre général si l'utilisateur souhaite faire une recherche non spécifique à une colonne. De plus, l'utilisateur a la possibilité de trier la colonne par ordre alphabétique ou l'inverse pour les colonnes caractères, et en ordre croissant ou décroissant pour les colonnes numériques. Grâce au bouton « Column visibility », l'utilisateur peut choisir les colonnes qu'il affiche dans le tableau, pour une question de visibilité, si seulement certaines données l'intéressent. Finalement, le nombre de lignes du tableau, qui correspond donc au nombre de lignes filtrées, est visible en bas à gauche de la Figure 9 comme le nombre d'« entries ». Ainsi, l'utilisateur peut chercher seulement les pathways comprenant un certain gène (faisant partie du sous-ensemble de gènes d'intérêt) par exemple et en connaître le nombre.

Column visibility | Download | Search:

Description	GeneRatio	BgRatio	Count	pvalue	p.adjust	qvalue	geneID
All	All	All	All	All	All	All	All
GOCC MITOCHONDRIAL PROTEIN CON...	125/1859	255/13058	125	8.43e-41	7.17e-37	6.82e-37	4706/7384/84545/4710/3028/6599...
GOCC INNER MITOCHONDRIAL MEMBR...	75/1859	135/13058	75	1.79e-29	7.61e-26	7.24e-26	4706/7384/4710/4708/4713/92609...
GOBP OXIDATIVE PHOSPHORYLATION	73/1859	135/13058	73	1.03e-27	2.92e-24	2.78e-24	4706/7384/4710/4708/4713/215/5...
GOBP ATP SYNTHESIS COUPLED ELE...	54/1859	94/13058	54	1.62e-22	3.44e-19	3.27e-19	4706/7384/4710/4708/4713/51142...
GOCC RESPIRATORY CHAIN COMPLEX	48/1859	79/13058	48	1.27e-21	2.16e-18	2.05e-18	4706/7384/4710/4708/4713/55101...
GOMF ELECTRON TRANSFER ACTIVIT...	61/1859	120/13058	61	1.63e-21	2.31e-18	2.19e-18	4706/7384/1535/8574/4710/4708/...
GOBP ELECTRON TRANSPORT CHAIN	71/1859	156/13058	71	3.05e-21	3.70e-18	3.52e-18	4706/1534/7384/1535/8574/4710/...
GOCC RESPIRASOME	52/1859	93/13058	52	5.37e-21	5.40e-18	5.13e-18	4706/7384/4710/4708/4713/55101...
GOBP RESPIRATORY ELECTRON TRAN...	57/1859	109/13058	57	5.71e-21	5.40e-18	5.13e-18	4706/7384/4710/4708/4713/2109/...
GOBP CELLULAR RESPIRATION	71/1859	172/13058	71	2.50e-18	2.13e-15	2.02e-15	4706/7384/4190/23596/4710/3421...

Showing 1 to 10 of 8,507 entries | Previous 1 2 3 4 5 … 851 Next

Figure 9: Capture d'écran de la table de données résultat de la fonction enricher, dans la section tables

C. Des graphiques adaptés à la transmission de l'information

Intéressons-nous maintenant à la section figures. Ici se trouvent les trois principaux graphiques de l'analyse de pathways. Un diagramme en barres (barplot), un diagramme en points (dotplot), et une carte d'enrichissement (enrichment map) peuvent y être observés. Mais pourquoi ont-ils été choisis, qu'apportent-ils à l'interprétation ?

i) Le diagramme en barres : « barplot »

La première figure dans l'ordre de lecture est le **diagramme en barres**. Sur ce graphique, les pathways les plus significatives peuvent être observées, selon le nombre maximal fixé par l'utilisateur (4^{e} boîte de filtres). Une barre représente un pathway, la longueur de la barre indique le nombre de gènes du pathway présents dans la sélection de gènes d'intérêt, et la couleur indique la valeur de la pvalue ajustée du test d'enrichissement de ce pathway. Les pathways sont affichés sur le graphique dans l'ordre de pvalue ajustée. Ce graphique est très utilisé dans ce genre d'analyses car il présente les deux informations principales. En effet, l'échelle de couleur met en avant la pvalue ajustée du test d'enrichissement, donc la force de cet enrichissement. Plus la barre est rouge, plus la pvalue ajustée est faible et donc l'enrichissement est important. Il est important de remarquer que cette échelle est seulement réalisée sur les pathways affichés, ainsi, un pathway ayant une barre bleue a un plus faible score

d'enrichissement que les rouges mais seulement parmi les pathways sélectionnés, donc déjà les plus enrichis. De plus, grâce au nombre de gènes du pathway parmi les gènes d'intérêt, l'utilisateur se rend compte d'à quel point ce pathway est représenté dans ces gènes d'intérêt. Prenons l'exemple du pathway 'GOBP generation of precursors metabolites and energy' dans le cas de l'analyse longitudinale, encadré en rouge sur la Figure 10. Les gènes appartenant à ce pathway sont impliqués dans la formation de métabolites précurseurs (c'est-à-dire de substances dont de l'énergie est dérivée), et dans les processus impliqués dans la libération d'énergie à partir de ces précurseurs (The Jackson Laboratory, 2021).
Si une grande partie des gènes d'intérêt de l'analyse longitudinale font partie de ce pathway, cela signifie qu'un certain nombre de gènes impliqués dans ces fonctions sont régulés par le traitement B par rapport au traitement A, au T3 par rapport au T1 pour notre exemple. Donc au T3, le traitement B a une action sur la libération d'énergie. Il est plus facile d'interpréter cette remarque plutôt que d'avoir une liste entière de gènes régulés par le traitement de manière individuelle.
Ainsi, ce diagramme en barres présente les principales informations de manière très intuitive, ce qui est souvent recherché par les cliniciens.

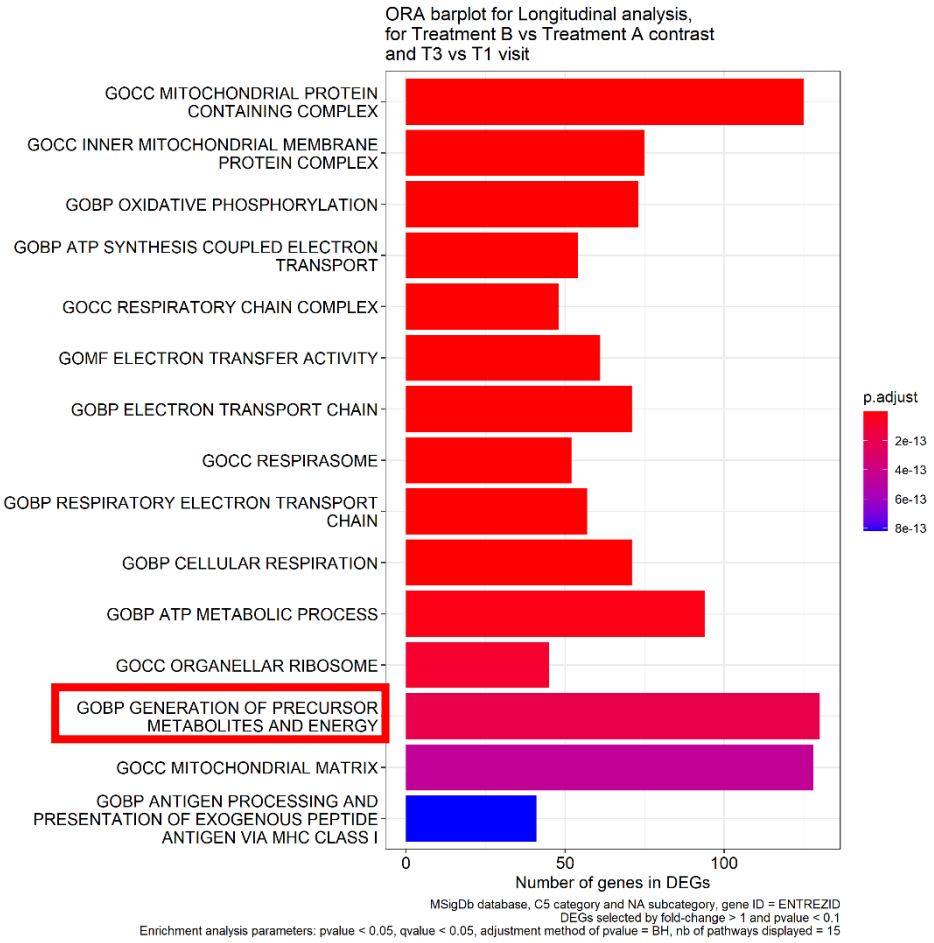


Figure 10: Diagramme en barres du test d'enrichissement sur des résultats d'analyse longitudinale, issu de l'application BIOexplorer (15 pathways affichés)

ii) Le diagramme en points : « dotplot »

La deuxième figure de l'onglet, le **diagramme en points**, se trouve au côté du diagramme en barres et lui ressemble fortement (cf Figure 11). En effet, l'utilisateur peut y retrouver la pvalue ajustée et le nombre de gènes du pathway dans les gènes d'intérêt (variable « Count »). La pvalue ajustée est présentée selon la même échelle de couleur. Cependant, ce graphique apporte une information supplémentaire : la variable « GeneRatio », qui est le ratio du nombre de gènes

du pathway parmi les gènes d'intérêt par rapport au nombre total de gènes d'intérêt. Le fait d'avoir des points et non des barres permet en effet de jouer sur un autre paramètre du graphique : la taille de ces points. C'est le nombre de gènes du pathway parmi les gènes d'intérêt qui va être affiché grâce à cette échelle de taille (« Count »). Les pathways ayant le plus de gènes parmi les gènes d'intérêt seront représentés par de plus gros points. Cette information était portée par l'axe des abscisses sur le diagramme en barres, alors que dans le diagramme en points c'est cette fois l'information du ratio de gènes qui est en abscisse. Finalement, c'est également selon ce ratio de gènes que les pathways sont ordonnés, et non plus selon les pvalue ajustées.
Il est intéressant de disposer à la fois du ratio de gènes et du nombre de gènes du pathway parmi les gènes d'intérêt, car ces deux informations permettent à l'utilisateur de se rendre compte à la fois des quantités relatives (ratio de gènes) et absolues (nombre de gènes) des gènes du pathway dans les gènes d'intérêt. Ce diagramme nous présente donc plus d'informations, de manière toujours aussi intuitive.

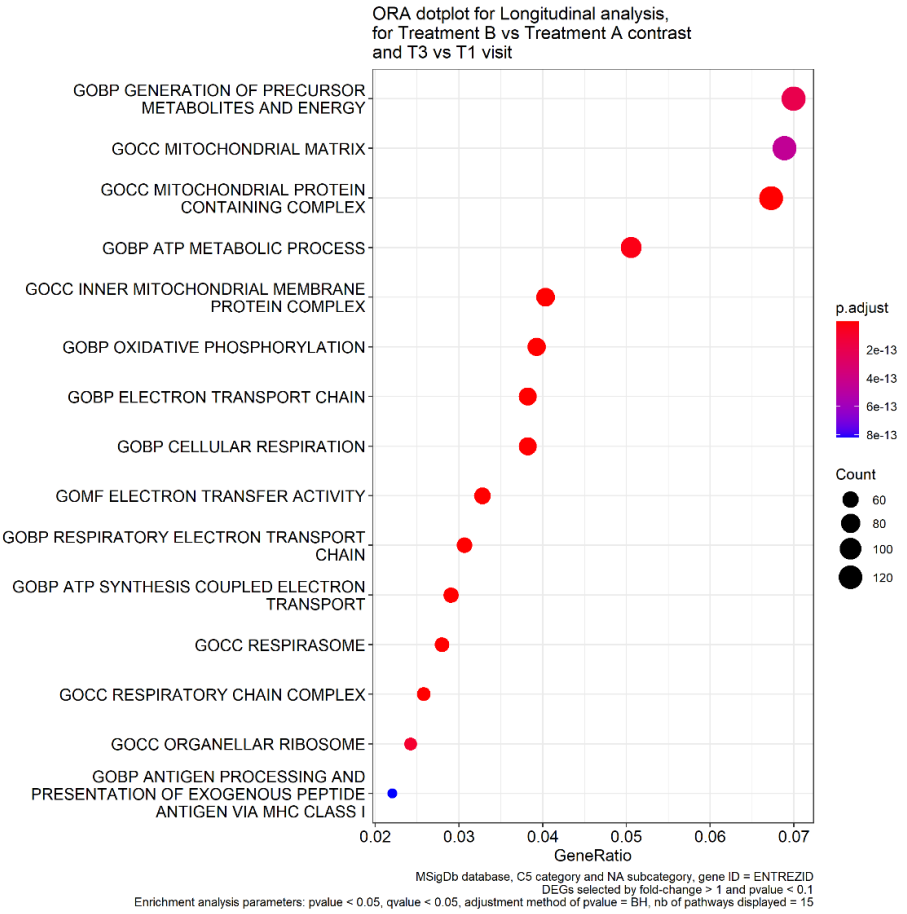


Figure 11: Diagramme en points du test d'enrichissement sur des résultats d'analyse longitudinale, issu de l'application BIOexplorer (15 pathways affichés)

iii) Carte d'enrichissement : « enrichment map »

La **carte d'enrichissement** est la troisième et dernière figure de l'onglet (cf Figure 12). C'est un graphique de type réseau, qui permet de faire des liens entre les pathways et d'éventuellement constituer des clusters lorsque certains de ces pathways sont liés entre eux. Les liens entre pathways sont faits par **recoupements de gènes**, c'est-à-dire qu'un lien apparaît entre deux pathways lorsqu'ils possèdent un ou plusieurs gène(s) en commun. Cette clusterisation de pathways permet de mettre en évidence un niveau encore plus général de fonctionnalité biologique. Dans l'exemple sur la Figure 12, beaucoup de pathways sont liés, car possédant des gènes en commun, et sont impliqués dans le processus de création d'énergie.

De plus, les mêmes échelles de couleur et de taille que pour le dotplot sont utilisées, représentant donc respectivement la pvalue ajustée et le nombre de gènes du pathway parmi les gènes d'intérêt. Conserver les mêmes échelles permet de garder une continuité entre les graphiques et d'en faciliter la lecture et la compréhension.
Ainsi, cette carte d'enrichissement met moins en évidence une hiérarchisation des pathways, mais plutôt une aide à l'interprétation en faisant ressortir des modules fonctionnels encore plus globaux que les pathways.

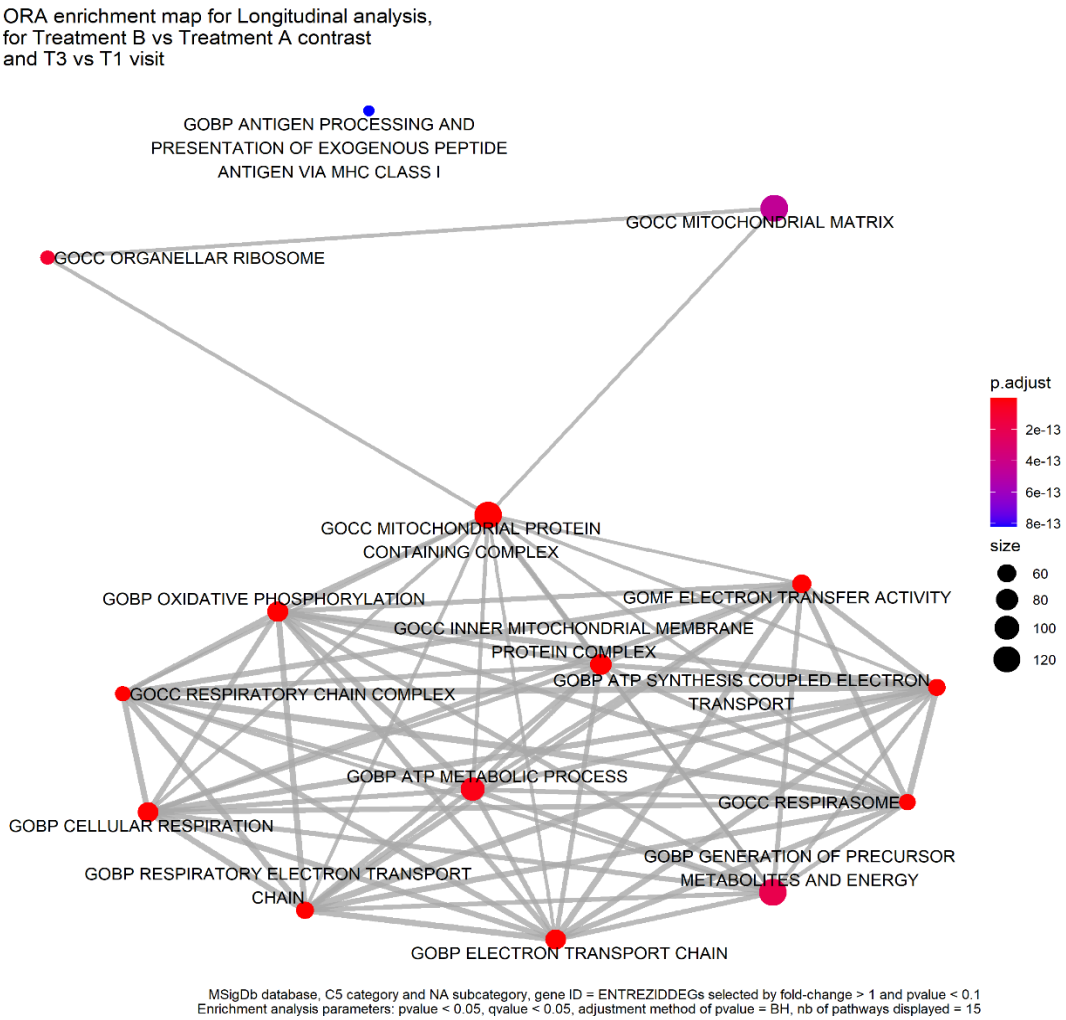


Figure 12: Carte d'enrichissement des pathways sur des résultats d'analyse longitudinale, issu de l'application BIOexplorer (15 pathways affichés)

Ces trois graphiques sont donc complémentaires et permettent une bonne compréhension et interprétation des résultats d'ORA, dans des conditions choisies par l'utilisateur (ici le contraste « Treatment B vs Treatment A », et le temps « T3 vs T1 »). Voyons maintenant comment visualiser les résultats résumant toutes les analyses d'enrichissement sous forme d'un seul et même graphique.

D. Un résumé permettant de comparer les résultats selon les visites et contrastes de traitement, ou selon les effets (prédictifs / pronostiques) et critères primaires : « panel plot »

Il est parfois utile d'avoir une **vision globale** de tous les résultats d'enrichissement (tous les contrastes de temps et de traitements (modèle longitudinal), ou les effets et critères primaires

(modèle univarié)), et de pouvoir les comparer, par exemple pour étudier **l'évolution** entre deux points de temps d'un même pathway.
La fonction *compareCluster* du package *clusterProfiler* nous permet de faire cela. Elle prend en entrée le tableau complet de résultats de l'analyse longitudinale (resp. univariée), ainsi que les variables selon lesquelles grouper les données, ici contraste de traitement et temps (resp. effet et critère). Les résultats sont affichés sous la forme d'un dotplot pour garder les informations présentées plus haut, et la fonction *facet_grid* de *ggplot2* est utilisée pour séparer en plusieurs sous-graphiques selon une des variables, ici le contraste de traitement (cf Figure 13) (resp. l'effet). Chaque sous-graphique présente ainsi les analyses d'enrichissement pour un contraste de traitement, et pour les différents temps. Des comparaisons peuvent donc être faites. Ici, très peu de pathways sont régulés au T2 vs T1 par rapport au T3 vs T1 et ce quelle que soit la comparaison entre bras de traitement. De plus, certains pathways, comme 'GOCC mitochondrial protein containing complex', regroupant des gènes codant pour des protéines mitochondriales, sont régulées significativement pour les 3 traitements.

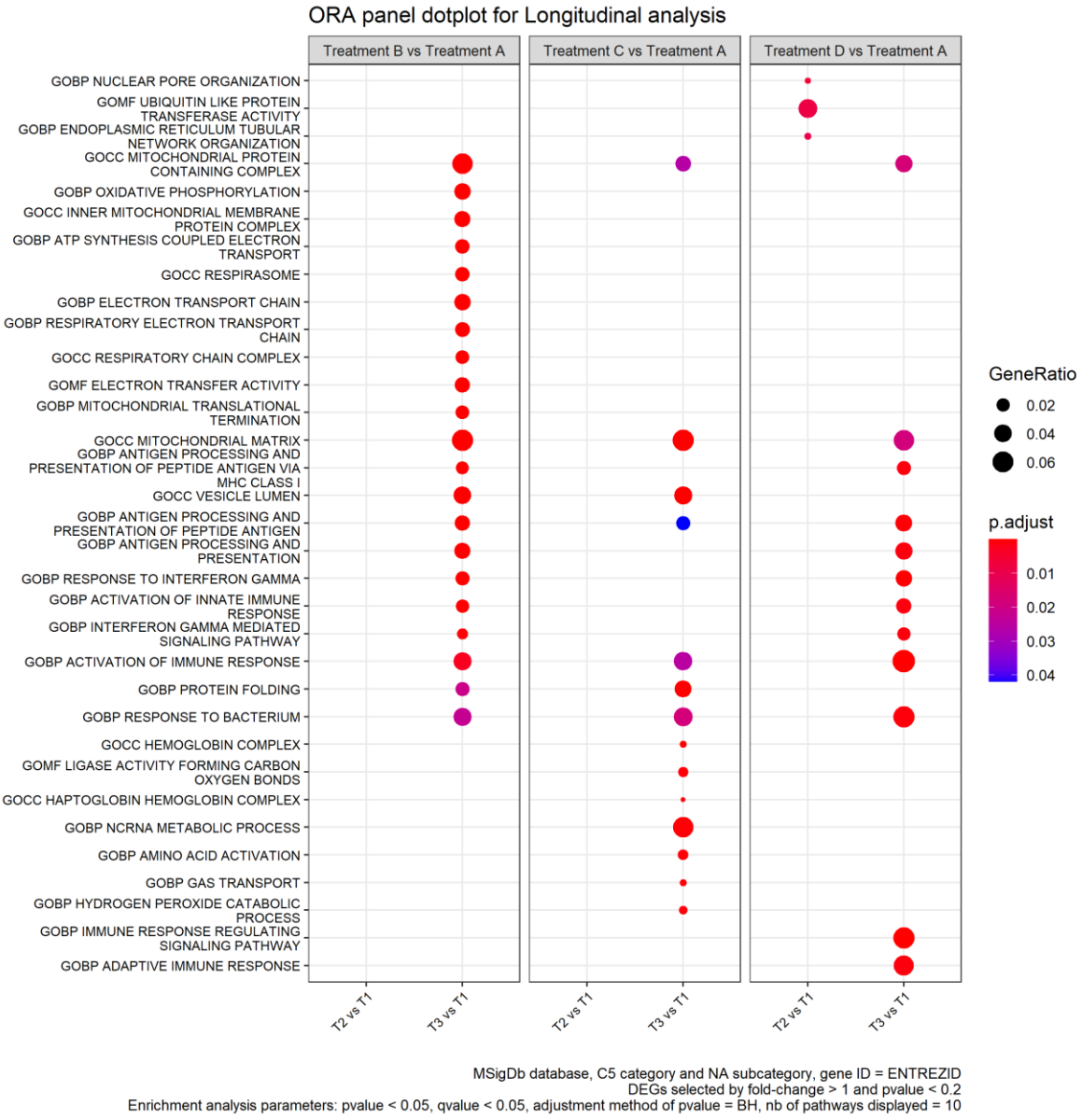


Figure 13: Panel plot des résultats d'analyse longitudinale, issu de l'application BIOexplorer (10 pathways affichés)

Cette vision globale peut être très utile, notamment lorsqu'il existe plusieurs bras de traitement à comparer.

E. Une aide à la compréhension et prise en main

Le but de l'application est d'être utilisée par l'équipe *Biomarker statistics,* mais surtout par les cliniciens par la suite. Or, les cliniciens n'ont pas forcément la même formation en statistiques. Comme l'application laisse une certaine liberté à l'utilisateur, il est nécessaire d'ajouter une **aide à la compréhension**, malgré le travail sur l'intuitivité de l'application.

Cette aide à la compréhension se fait sous différente forme. Tout d'abord il existe une **aide directe sur l'application**. Elle consiste principalement en des messages d'explication des graphiques qui s'affichent en passant la souris sur les icônes d'information à côté des boutons d'export (infobox). L'utilisateur y trouve alors un petit texte résumant l'intérêt du graphique, une aide à l'interprétation, et quelques précisions sur les légendes ou les filtres influant sur le graphique (cf Figure 14).

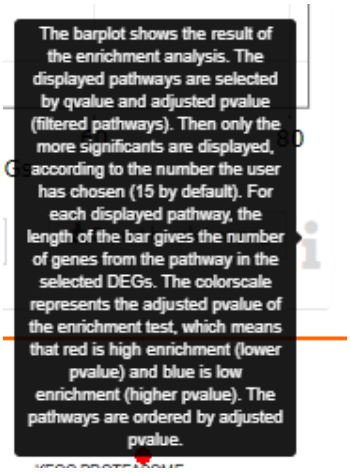
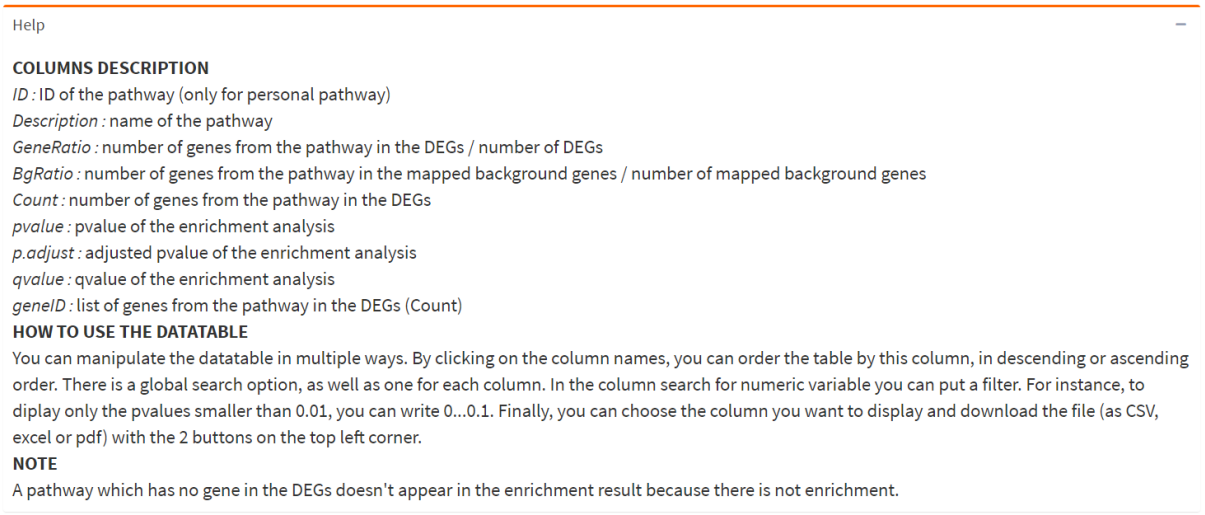


Figure 14: Capture d'écran de l'infobox pour le diagramme en barres de l'ORA

Parfois, le format de l'infobox n'est pas compatible avec la quantité d'explications ou l'importance à donner à l'explication, donc certaines aides s'affichent directement dans les pages de l'application sous forme de boîte de texte, comme pour l'aide du tableau de résultats d'ORA (cf Figure 15).


Help

COLUMNS DESCRIPTION

ID : ID of the pathway (only for personal pathway)
Description : name of the pathway
GeneRatio : number of genes from the pathway in the DEGs / number of DEGs
BgRatio : number of genes from the pathway in the mapped background genes / number of mapped background genes
Count : number of genes from the pathway in the DEGs
pvalue : pvalue of the enrichment analysis
p.adjust : adjusted pvalue of the enrichment analysis
qvalue : qvalue of the enrichment analysis
geneID : list of genes from the pathway in the DEGs (Count)

HOW TO USE THE DATATABLE

You can manipulate the datatable in multiple ways. By clicking on the column names, you can order the table by this column, in descending or ascending order. There is a global search option, as well as one for each column. In the column search for numeric variable you can put a filter. For instance, to diplay only the pvalues smaller than 0.01, you can write 0...0.1. Finally, you can choose the column you want to display and download the file (as CSV, excel or pdf) with the 2 buttons on the top left corner.

NOTE

A pathway which has no gene in the DEGs doesn't appear in the enrichment result because there is not enrichment.

Figure 15: Capture d'écran de l'aide pour le tableau de résultats d'ORA

En plus de l'aide présente dans l'application, un **guide utilisateur** est rédigé en anglais. Ce guide présente en détails la structure de chaque onglet, avec une explication précise de chaque filtre, chaque bouton, ainsi que de chaque graphique. De plus, il spécifie des points de vigilance, notamment dus aux chronologies de mise à jour des données selon les boutons et onglets.

Doté de ce guide utilisateur, ainsi que des aides fournies dans l'application, l'utilisateur peut naviguer, analyser et interpréter sans problème sur BIOexplorer.

L'utilisateur dispose donc, dans l'application BIOexplorer, d'un onglet finalisé pour l'analyse de pathways de type ORA. Cet onglet lui permet d'avoir la main sur un grand nombre de paramètres de l'analyse, et d'en visualiser les résultats de manière interactive sous forme de graphiques et de tables de données.

IV. Discussion & challenges

L'application BIOexplorer est, comme décrit dans ce rapport, **toujours en cours de développement**, car des éléments sont constamment à améliorer (efficacité du code, retour des utilisateurs) et à ajouter (nouvelles analyses, nouveaux graphiques), notamment avec des demandes qui évoluent en termes de résultats et d'analyses, ainsi qu'en termes de données. Pour l'instant, l'application est développée sur des données de type RNAseq, présentées en 1ère partie, ainsi qu'Olink (quantification de protéines) et mRNA (quantification d'ARN messager), et ce sur 7 études différentes. Par la suite, il est prévu d'adapter l'application par exemple à des données de 'single cell', c'est-à-dire au niveau de la cellule unique, ou à des données 'whole genome sequencing', c'est-à-dire séquençage complet du génome.
Le projet le plus important à court terme est de **généraliser l'application**. En effet, il n'existe en réalité pas UNE application BIOexplorer mais DES applications BIOexplorer, car pour chaque projet où elle est utilisée, une nouvelle application est développée. L'automatisation des modules des différents onglets, qui a été finalisée en début de stage, permet d'avoir très peu d'éléments à changer lors du développement d'une nouvelle application sur un nouveau projet, car il ne faut modifier que les 3 codes de base (ui, server et global). Ces modifications peuvent être par exemple le nom de l'application, les noms et nombres de bras de traitements, etc. Cependant, le but de la généralisation serait d'avoir une seule application commune à toutes les études.
De plus, il est prévu d'ajouter de **nouveaux modules** à l'application, comme un onglet d'analyse par biomarqueur, permettant de sélectionner un biomarqueur et de regrouper tous les résultats le concernant. Ou également d'ajouter des analyses de corrélations supplémentaires par rapport à celle développée pendant le stage, c'est-à-dire entre les biomarqueurs au temps initial et les variables cliniques au temps initial. Il serait possible d'ajouter les corrélations des biomarqueurs entre eux, ou des biomarqueurs avec les critères primaires.

L'application étant développée en entreprise, la demande des cliniciens et la **communication** sont des éléments très importants à prendre en compte lors du développement. En effet, les graphiques doivent être complets mais pas trop complexes, car ils peuvent poser des problèmes de compréhension s'ils donnent trop d'informations ou ne sont pas assez intuitifs. De plus, l'entreprise est une mine d'informations car beaucoup d'équipes travaillent dans différents domaines ou sur différentes analyses, et peuvent apporter leur expertise précieuse en plus de la bibliographie. La **collaboration** est importante. Par exemple, au cours du développement de l'onglet « analyse de pathways », je l'ai présenté à une équipe de Sanofi Pasteur (filière travaillant sur le développement de vaccins) qui m'a ensuite donné des conseils pour l'améliorer, en fonction de leur expérience.

Conclusion

En conclusion, nous avons pu voir à travers l'analyse de l'onglet d'over-representation analysis (ORA) que l'application *Rshiny* BIOexplorer est un bon moyen de visualiser des résultats d'analyses biomarqueurs. Elle permet en effet de les afficher de manière exhaustive et interactive, et de créer une réelle interface utilisateur-machine où l'utilisateur a de nombreuses possibilités, et peut naviguer de manière intuitive. Pour être sûr de la bonne compréhension de l'utilisateur, des aides sont disponibles interactivement sur l'application, ainsi que dans des documents annexes. De plus, les résultats des analyses peuvent se présenter sous plusieurs formes, que ce soit graphiques ou tableaux, et ces 2 formes peuvent être interactives. Finalement, il reste encore du travail à faire sur cette application, notamment sur sa généralisation, mais elle est déjà bien utilisée et appréciée au sein de l'équipe *Biomarker Statistics*, ainsi que par les autres utilisateurs plus orientés biologie.

Références bibliographiques

Abatangelo, L., Maglietta, R., Distaso, A., D'Addabbo, A., Creanza, T. M., Mukherjee, S., & Ancona, N. (2009). Comparative study of gene set enrichment methods. *BMC Bioinformatics, 10*, 275-286. doi:10.1186/1471-2105-10-275 (consulté le 11 juin 2021)

Anjum, A., Jaggi, S., Varghese, E., Lall, S., Bhowmik, A., & Rai, A. (2016, April). Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach. *Journal of Computational Biology, 23*(4) 239-247. doi:10.1089/cmb.2015.0205 (consulté le 5 avril 2021)

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300. doi: 10.1111/j.2517-6161.1995.tb02031.x (consulté le 7 juillet 2021)

Blangero, Y. (2019). Méthodologie de l'évaluation des biomarqueurs prédictifs quantitatifs et de la détermination d'un seuil pour leur utilisation en médecine personnalisée. *Thèse de doctorat en biostatistiques : Université Claude Bernard Lyon 1*, 122 p. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-02381703/document (consulté le 19 mars 2021)

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., & Sherlock, G. (2004). GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics, 20*(18), 3710-3715. doi:10.1093/bioinformatics/bth456 (consulté le 6 juillet 2021)

Garcia-Campos, M. A., Espinal-Enriquez, J., & Hernandez-Lemus, E. (2015, December). Pathway Analysis : State of the Art. *Frontiers in Physiology, 6*, 383. doi:10.3389/fphys.2015.00383 (consulté le 11 juin 2021)

Gene Ontology Consortium. (2004, January). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research, 32*(Issue suppl_1), D258-D261. doi:10.1093/nar/gkh036 (consulté le 14 juin 2021)

Jiao, Y., Li, Y., Liu, S., Chen, Q., & Liu, Y. (2019, May). ITGA3 serves as a diagnostic and prognostic biomarker for pancreatic cancer. *OncoTargets and Therapy, 12*, 4141-4152. doi:10.2147/OTT.S201675 (consulté le 19 mars 2021)

Kanehisa, M., & Goto, S. (2000, January). KEGG: Kyoto Encyclopedia og Genes and Genomes. *Nucleic Acids Research, 28*(Issue 1), 27-30. doi:10.1093/nar/28.1.27 (consulté le 14 juin 2021)

Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. (T. C. Christos A. Ouzounis, Éd.) *PLoS Computational Biology, 8*(2), e1002375. doi:10.1371/journal.pcbi.1002375 (consulté le 11 juin 2021)

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014, February). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology, 15*(2), R29. doi:10.1186/gb-2014-15-2-r29 (consulté le 5 avril 2021)

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., & Mesirov, J. P. (2011, June). Molecular signatures database (MSigDB) 3.0. *Bioinformatics, 27*(Issue 12), 1739-1740. doi:10.1093/bioinformatics/btr260 (consulté le 13 juillet 2021)

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2011, January). Entrez Gene: gene-centered information at NCBI . *Nucleic Acids Research, 39*(Issue suppl_1), D52-D57. doi:10.1093/nar/gkq1237 (consulté le 16 juillet 2021)

Montgomery, S. B., & Kukurba, K. R. (2015, April). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols, 2015*(11), 951-969. doi:10.1101/pdb.top084970 (consulté le 5 avril 2021)

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics, 12*(2), 87-98. doi:10.1101/pdb.top084970 (consulté le 5 avril 2021)

Paccard, C. (2018). Présentation de l'équipe Biomarker Statistics. Sanofi (document interne). 44 p. (consulté le 1er mars 2021)

Paulson, J. N., Chen, C.-Y., Lopes-Ramos, C. M., Kuijjer, M. L., Platig, J., Sonawane, A. R., . . . Quackenbush, J. (2017, October). Tissue-aware RNA-Seq processing and normalization for heterogeneous and sparse data. *BMC Bioinformatics, 18*(1), 437. doi:10.1186/s12859-017-1847-x (consulté le 5 avril 2021)

Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., & Wain, H. (2001, October). The HUGO Gene Nomenclature Committee (HGNC). *Human Genetics, 109*(6), 678-680. doi:10.1007/s00439-001-0615-0 (consulté le 16 juillet 2021)

Rabbee, N. (2020). *Biomarker Analysis in Clinical Trials with R.* Chapman and Hall/CRC Biostatistics Series. 197 p. (consulté le 19 mars 2021)

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., . . . Bader, G. D. (2019, January). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols, 14*(2), 482-517. doi:10.1038/s41596-018-0103-9 (consulté le 11 juin 2021)

Ruffier, M., Kähäri, A., Komorowska, M., Keenan, S., Laird, M., Longden, I., . . . Flicek, P. (2017, January). Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation. *The Journal of Biological Databases and Curation, 2017*(1), bax020. doi:10.1093/database/bax020 (consulté le 16 juillet 2021)

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS, 5*(Issue 6), p 463-466. doi:10.1097/COH.0b013e32833ed177 (consulté le 4 mars 2021)

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005, October). 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), 15545-15550. doi:10.1073/pnas.0506580102 (consulté le 11 juin 2021)

Vlassenko, A. G., McCue, L., Jasielec, M. S., Su, Y., Gordon, B. A., Xiong, C., . . . Fagan, A. M. (2016, September). Imaging and cerebrospinal fluid biomarkers in early preclinical alzheimer disease. *Annals of neurology, 80*(Issue 3), 379-387. doi:10.1002/ana.24719 (consulté le 9 avril 2021)

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012, May). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology, 16*(5), 284-287. doi:10.1089/omi.2011.0118 (consulté le 14 juin 2021)

Références sitographiques

Broad Institut. (2021). *MSigDB Collections.* [en ligne]. Disponible sur : http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp (consulté le 13 juillet 2021)

Institut Frédéric Joliot. (2020). *Les omiques*. [en ligne]. Disponible sur : https://joliot.cea.fr/drf/joliot/Pages/Institut/Enjeux/Methodo-Techno/omiques.aspx (consulté le 4 mars 2021)

RStudio. (2020). *Shiny*. [en ligne]. Disponible sur : https://shiny.rstudio.com/ (consulté le 2 mars 2021)

Sanofi. (2020). *Nous connaître*. [en ligne]. Disponible sur : https://www.sanofi.com/fr/nous-connaitre (consulté le 20 juillet 2021)

Sanofi. (2021). *Science & Innovation - Recherche & Développement*. [en ligne]. Disponible sur : https://www.sanofi.com/fr/science-et-innovation/recherche-et-developpement (consulté le 20 juillet 2021)

The Jackson Laboratory. (2021). Mouse Genome Informatics. *Gene Ontology Browser*. [en ligne]. Disponible sur : http://www.informatics.jax.org/vocab/gene_ontology/GO:0006091 (consulté le 10 août 2021)

l'institut Agro agriculture • alimentation • environnement | AGRO CAMPUS	Diplôme : Ingénieur Spécialité : Agronomie Spécialisation / option : Statistiques et sciences des données Enseignant référent : Mathieu Emily

Auteur(s) : Anne-Victoire Lagroy de Croutte de Saint Martin Date de naissance* : 18/01/1998	Organisme d'accueil : Sanofi R&D Adresse : 1 avenue Pierre Brossolette 91380 CHILLY-MAZARIN FRANCE Maître de stage : Caroline Paccard
Nb pages : 20 Annexe(s) : 0	
Année de soutenance : 2021	

Titre français : Développement d'une application R Shiny pour la visualisation de résultats d'analyse de données biomarqueurs

Titre anglais : Development of an R Shiny application for the visualization of the results of biomarker data analyses

Résumé (1600 caractères maximum) :

La génération de données biomarqueurs omiques se fait de plus en plus rapidement et facilement, ce qui entraîne une multiplication de leur quantité et de leur taille. Les analyses principales les concernant lors des études cliniques sont de déterminer s'ils sont régulés par le traitement au cours du temps, prédictifs de la réponse au traitement, pronostiques de l'évolution de la maladie. Les résultats de ces analyses représentent une importante quantité de graphiques et données, ce qui induit un questionnement sur la manière de les visualiser et communiquer. C'est pourquoi une application R Shiny, Biomarker Interactive Output explorer (BIOexplorer), est en cours de développement chez Sanofi. Cette application permettait déjà la visualisation des données cliniques et des résultats des analyses omiques citées précédemment. Le but du stage a été de développer un onglet d'analyse de corrélations entre les variables cliniques et les biomarqueurs avant traitement ; ainsi qu'un onglet d'analyse de pathways de gènes qui permet une interprétation plus globale et fonctionnelle des résultats des analyses préliminaires (biomarqueurs régulés/prédictifs/pronostiques). La méthode développée dans l'application et présentée dans le rapport est l'Over-Representation Analysis (ORA), qui se base sur la sélection d'une liste de gènes d'intérêt. Grâce à de nombreux filtres, des graphiques interactifs (barplot/dotplot/carte d'enrichissement) et un tableau interactif, l'onglet « ORA » est une interface machine-utilisateur idéale pour la visualisation et communication des résultats d'ORA.

Abstract (1600 caractères maximum) :

Omic biomarker data are more and more easy and quick to collect, and therefore very numerous and weighty. The main questions for these biomarkers in clinical trials are to study if they are regulated by treatment over time, predictive of response to treatment, and/or prognostic of disease progression. The results of these analyses represent a large amount of graphs and data, which leads to the issue: how to visualize and communicate them? As an answer to this issue, an R Shiny application, Biomarker Interactive Output explorer (BIOexplorer), is being developed at Sanofi. This application already allowed the visualization of clinical data and results of the three omics analyses mentioned above. The goal of the internship was mainly to develop two new sections presenting two new types of analyses. The first one is a correlation analysis between biomarkers at baseline and clinical variables at baseline. The second one is a gene pathway analysis, that allows a more global and functional interpretation of the results of the preliminary analyses (regulated/predictive/prognostic biomarkers). The method developed in the application and presented in the report is Over-Representation Analysis (ORA), which is based on the selection of a list of differentially expressed genes. Thanks to numerous filters, interactive graphs (barplot, dotplot, enrichment map) and an interactive table, the "ORA" tab is an ideal machine-user interface for the visualization and communication of ORA results.

Mots-clés : biomarqueur, application RShiny, visualisation, pathway, ORA

Key Words: biomarker, RShiny application, visualization, pathway, ORA

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*