Développement d'un pipeline de Gene Set Enrichment Analysis ou GSEA

Stéphanie Levon

20 juin 2016

1. Analyse du projet : méthodes, choix des outils et du jeu de données

1. Méthodes statistiques

Plusieurs méthodes statistiques existent pour la mise en évidence des enrichissements fonctionnels. La modélisation des enrichissements peut être réalisée grâce à la loi hypergéométrique, qui a les même cinditions que la loi binomiale mais qui est utilisée lorsque le nombre d'individus (ici le nombre de gènes) n'est pas négligeable devant la taille de la population.
Une autre méthode statistique très présente dans les analyses d'enrichissement est celle de Kolmogorov-Smirnov. Il s'agit d'un test d'adéquation non paramétrique. Il permet de comparer une distribution avec une autre.
Enfin, pour tester la significativité des pvalue de gènes différentiellement exprimés, un test de Fisher est couramment utilisé.

2. Choix des outils

Quelques outils interfacés vue en cours tels que DAVID semblaient pertinents pour l'analyses de données RNA-seq par GSEA. Cependant il s'avère que cet outil n'est plus à jour et est devenu obsolète. D'autre part le site de recensement d'outils OmicTools propos peu d'outil adaptés aux questions posées par le GSEA une fois l'analyse différentielle réalisée.
A côté de ces outils, de nombreux packages R sont disponibles pour analyses d'enrichissement à partir de données de séquençage. Le pipeline a donc été développé en R et intègre une sélection d'outils trouvés sur la plateforme Bioconductor qui fournit en libre accès de nombreux packages R pour l'analyses de données biologiques. Les outils choisis pour l'analyse sont les suivants : - biomaRt pour la conversion de données,
- topGO pour la récupération et la visualisation des termes GO ,
- reactomePA pour l'enrichissement des voies métaboliques *via* la base de données REACTOME,
- clusterProfiler pour l'enrichissement des voies métaboliques KEGG,
- pathview pour la visualisation des voies métaboliques KEGG.

3. Choix du jeu de données

J'ai commencé ma recherche de données avec des mots clés et termes MeSh sur Pumed. Malheureusement, bien qu'il existe de nombreux articles sur l'analyse GSEA à partir de données RNA-seq, les données sources sont souvent difficile à récupérer. J'ai alors adopté une stratégie différente en recherchant un article auxquel des membres de mon équipe d'apprentissage avait pu participer. Cet article (cf. référence ci-dessous), non seulement décrit le jeu de données qui a été utilisé pour l'étude, et le fournit sous forme de document excel.

Le jeu de données utilisé provient d'une étude portant sur les cellules souches mélanocytaires. Les mélanocytes sont responsables de la pigmentation de l'épiderme et se différencient à partir de cellule souches mélanocytaires. Le Microphtalmia-associated Transcription Factor (MITF) joue un rôle prépondérant dans la bonne différentiation de ces cellules. Une expérience de RNA-seq a été réalisée sur des cellules mélanocytaires 501 Mel d'humain. L'expression différentielle des gènes connus chez l'homme est obtenue en comparant le transcriptome de cellules contrôle avec celui de cellules 501 Mel délétées du gène MITF. Deux réplicats pour

chaque échantillons ont été séquencés. La normalisation a été réalisée avec le logiciel DESeq2 et l'annotation avec DAVID.

Koludrovic D, Laurette P, Strub T, et al. Chromatin-Remodelling Complex NURF Is Essential for Differentiation of Adult Melanocyte Stem Cells. Bickmore WA, ed. PLoS Genetics. 2015;11(10):e1005555. doi:10.1371/journal.pgen.1005555.

2. Développement de mon pipeline

1. Installation et caractéristiques des outils

Cinq packages R sont nécessaires pour le bon fonctionnement du pipeline. Chaqun de ces outils nécessite de nombreuses dépendances. La plupart du temps, l'installation de ces dépendances est gérée par R. Le tableau ci-dessous recence les outils utilisés, accompagné de leur référence, d'une description succinte ainsi que de leur rôle au sein du pipeline.

Nom de l'outil	Description général	Fonctions utilisé
biomaRt (1)	Interface vers plusieurs bases de données, permet la conversion de formats	getBM
topGO (2)	Fournit les termes GO en tenant compte de leur topologie	runTest - showSigOf
reactomePA (3)	Analyse des voies métaboliques à partir de la base de données REACTOME	enrichPathway
clusterProfiler (5) (4)	Enrichissement métaboliques des voies de la base de données KEGG	enrichKEGG
pathview (5)	Visualisation des voies métaboliques KEGG	pathview

(1) Durinck, Steffen, Paul T. Spellman, Ewan Birney, and Wolfgang Huber. "Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor Package biomaRt." Nature Protocols 4, no. 8 (2009): 1184–91. doi:10.1038/nprot.2009.97.

(2) Alexa A and Rahnenfuhrer J (2016). topGO: Enrichment Analysis for Gene Ontology. R package version 2.24.0.

(3) Yu, Guangchuang, and Qing-Yu He. "ReactomePA: An R/Bioconductor Package for Reactome Pathway Analysis and Visualization." Molecular bioSystems 12, no. 2 (February 2016): 477–79. doi:10.1039/c5mb00663e.

(4) Yu G, Wang L, Han Y and He Q (2012). "clusterProfiler: an R package for comparing biological themes among gene clusters." OMICS: A Journal of Integrative Biology, 16(5), pp. 284-287.

(5) Luo, Weijun, and Cory Brouwer. "Pathview: An R/Bioconductor Package for Pathway-Based Data Integration and Visualization." Bioinformatics 29, no. 14 (July 15, 2013): 1830–31. doi:10.1093/bioinformatics/btt285.

2. Données en entrée du pipeline

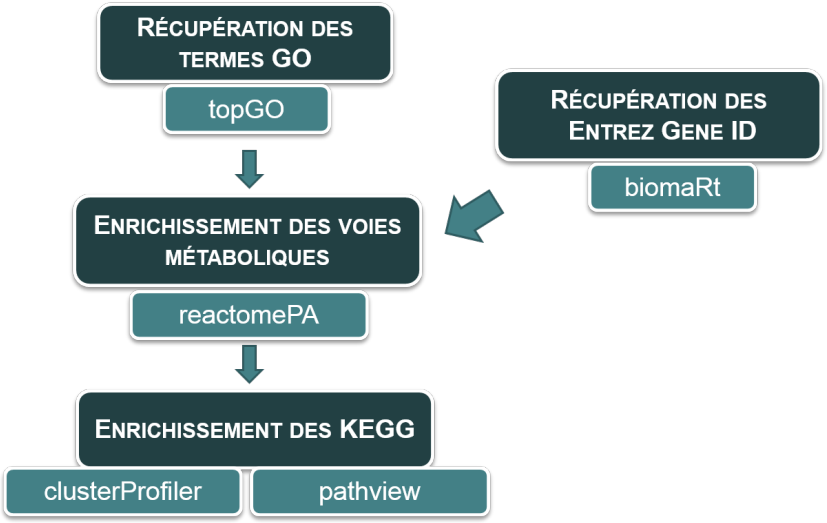
Le fichier donné en entrée est de type "csv" (Coma Separated Values). Il contient au minimum 3 colonnes énumérées ci-dessous agrémentées d'un exemple:

Ensembl gene id	gene_name	fold_change	padj
ENSG00000148677	ANKRD1	4.5018712402	1,49E-036

L'Ensembl gene id OU le nom du gène peuvent être donnés. L'outil biomaRt se chargera de récupérer l'Entrez Gene ID. Le fold change correspond à la différence d'expression entre l'échantillon test et l'échantillon control. La p-value ajustée est utilisée car de nombreux tests sont réalisés (pour chaque gène exprimé). Le modèle qui correpond le mieux à la distribution des données contient de nombreux paramètres, il est complexe, plus précis mais plus difficile à utiliser. La p-value est donc affecté (augmenté) par ce grand nombre de paramètre.

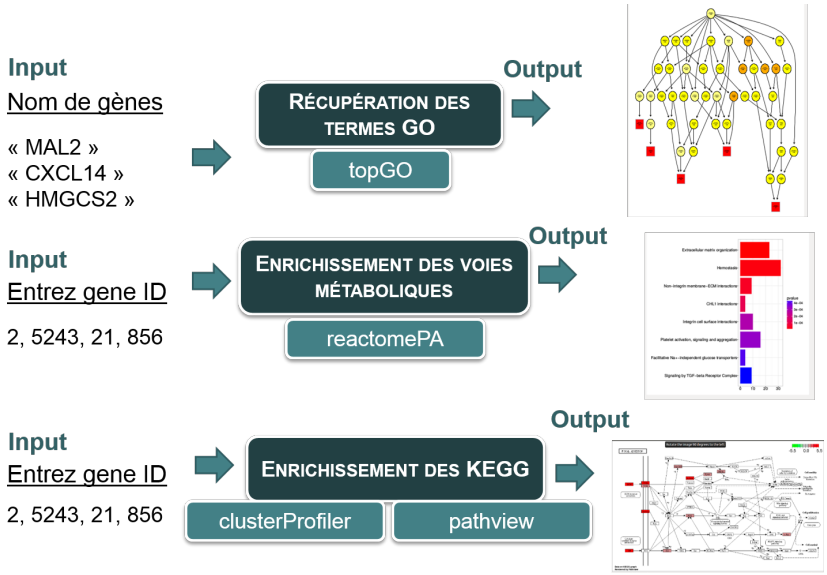
De plus, le risque α est multiplié par le nombre de test à réaliser dans une p-value, ici le risque α serait beaucoup trop élevé. C'est pourquoi on utilise la p-value ajusté qui s'affranchit de ce facteur sur le risque de première espèce α. Le jeu de données peut être récupéré *via* les données supplémentaires fournit par l'article ainsi que sur la base de données Geo (Gene Expression Omnibus) sous la référence GSE61967.

3. Schéma général vu du point de vue biologiste



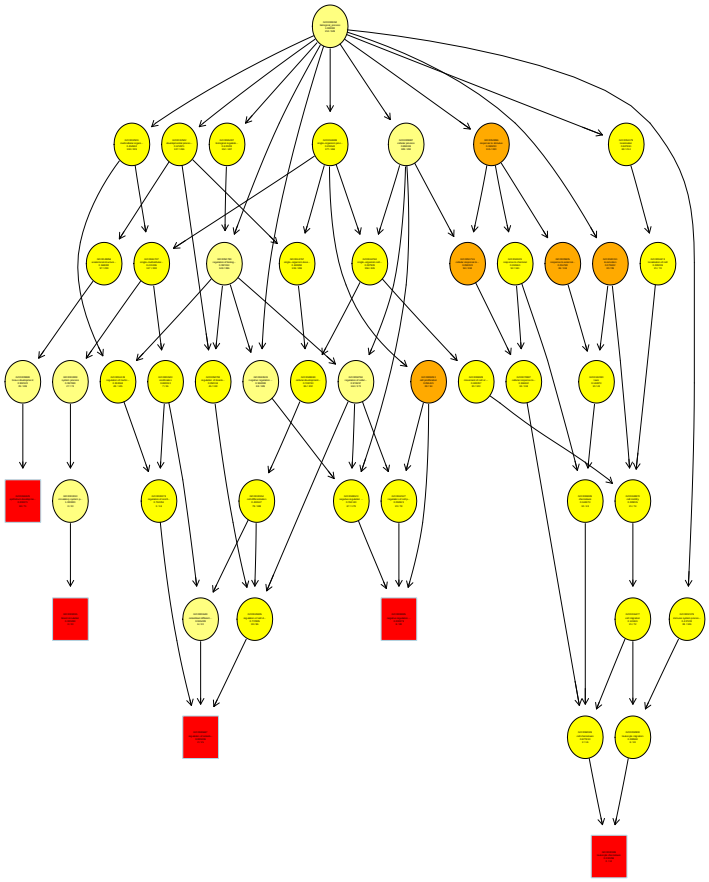
Les termes GO trouvés sont les suivants :
- développement de l'épithélium
- circulation sanguine
- différenciation des ostéoblastes
- chimiotaxie des leucocytes.

4. Schéma général vu du point de vue développeur

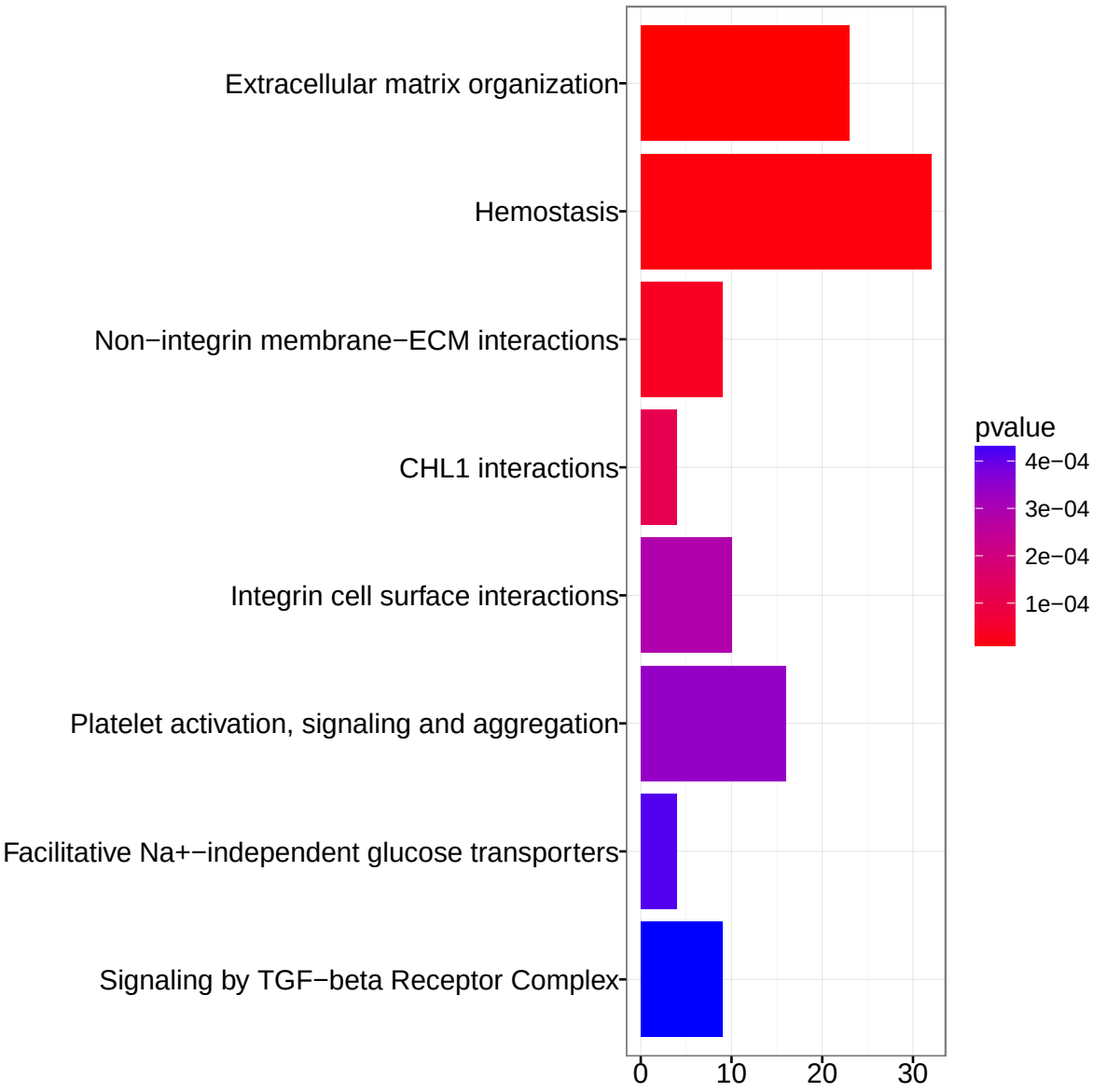


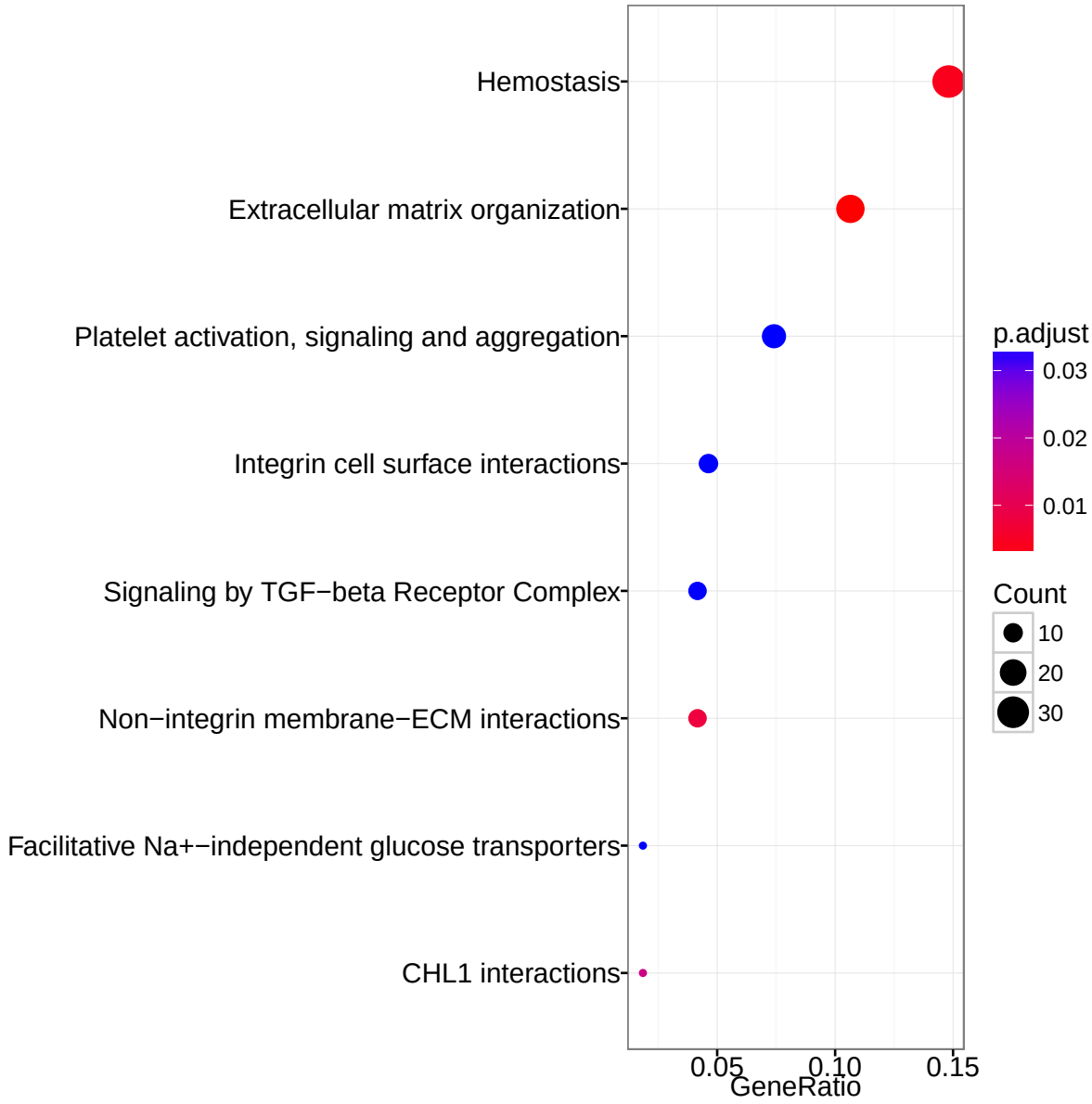
3. Résultats

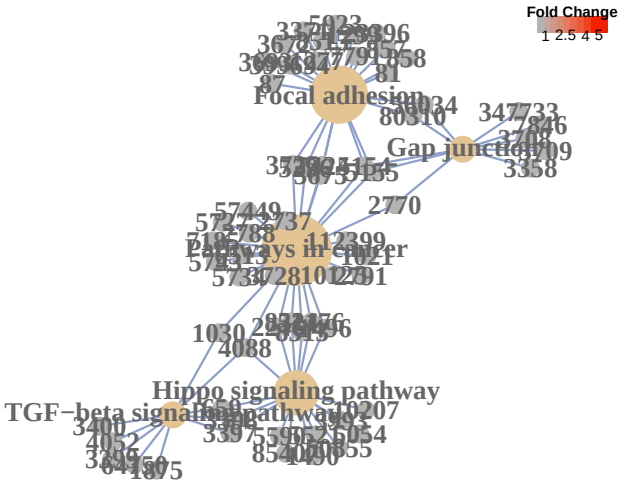
1. GSEA 1 : termes GO

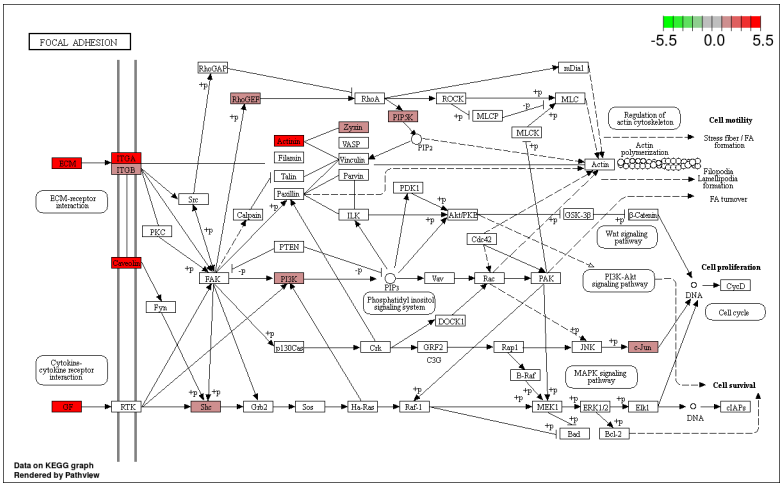


2. GSEA 2 : voie métabolique




Hemostasis
Extracellular matrix organization
Platelet activation, signaling and aggregation
Integrin cell surface interactions
Signaling by TGF−beta Receptor Complex
Non−integrin membrane−ECM interactions
Facilitative Na+−independent glucose transporters
CHL1 interactions
0.05
0.10
0.15
GeneRatio
p.adjust
0.03
0.02
0.01
Count
10
20
30





Résumé des résultats trouvés avec la base de données REACTOME et la base de données KEGG :

KEGG

pathway ID	Description	GeneRatio	BgRatio
hsa04510 hsa04510	Focal adhesion	22/228	203/7057
hsa04390 hsa04390	Hippo signaling pathway	17/228 154/7057	
hsa05200 hsa05200	Pathways in cancer	27/228 397/7057	
hsa04350 hsa04350	TGF-beta signaling pathway	10/228 84/7057	
hsa04540 hsa04540	Gap junction	10/228	88/7057
hsa04530 hsa04530	Tight junction	13/228	139/7057
hsa04360 hsa04360	Axon guidance	15/228	176/7057
hsa04810 hsa04810	Regulation of actin cytoskeleton	17/228	215/7057

pathway ID	Description	GeneRatio	BgRatio
hsa05412 hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	9/228	74/7057
hsa05231 hsa05231	Choline metabolism in cancer	10/228	101/7057

REACTOME

pathway ID	Description	GeneRatio	BgRatio
1474244	Extracellular matrix organization	23/216	249/6750
109582	Hemostasis	32/216	450/6750
3000171	Non-integrin membrane-ECM interactions	9/216	53/6750
447041	CHL1 interactions	4/216	9/6750
216083	Integrin cell surface interactions	10/216	83/6750
76002	Platelet activation, signaling and aggregation	16/216	189/6750
428790	Facilitative Na+-independent glucose transporters	4/216	12/6750
170834S	Signaling by TGF-beta Receptor Complex	9/216	72/6750

Le gène ratio correspond au nombre de gènes impliqués dans la voie métabolique par rapport au nombre de gènes différentiellement exprimés.
Le "bg ratio" pour background ratio correpond au nombre de gènes qui ont cette ontologie par rapport au nombre de gènes totaux.
Les résultats obtenus ont été comparés à la publication suivante.
Strub, T., S. Giuliano, T. Ye, C. Bonet, C. Keime, D. Kobi, S. Le Gras, et al. "Essential Role of Microphthalmia Transcription Factor for DNA Replication, Mitosis and Genomic Stability in Melanoma." Oncogene 30, no. 20 (May 19, 2011): 2319–32. doi:10.1038/onc.2010.612.

Dans cette publication, plusieurs termes se recoupent : l'adhesion focal (KEGG), l'implication dans des mécanismes de cancer (colorectal, KEGG), ainsi que la voie TGF beta (KEGG). Peu de recoupement peuvent être fait entre les voies métaboliques trouvées par REACTOME (meilleure voie mise en lumière) et par KEGG. Pour allez plus loin, il serait interessant de regarder quels gènes ont servis à l'élaboration des voies métaboliques mise en cause.

4. Discussion – conclusion

Les analysis GSEA permettent de caractériser les gènes différentiellement exprimés à l'échelle des voies métaboliques. Cependant, de nombreuses informations sont perdues au cours du traitement des données (exemple : conversion des identifiants pas toujours possible).