Cahier des charges

Le rôle de la méthylation des protéines dans le cancer du sein

Ekaterina Flin

13/10/2023

Attention !

Les figures dans ce document sont données à titre d'exemple pour illustrer une représentation possible des résultats. Ces figures peuvent venir d'autres projets, ce ne sont pas de « vrais » résultats attendus.

Table des matières

Contexte du projet

Dans une cellule, une protéine synthétisée à partir de l'ARN messager peut subir des modifications chimiques appelées modifications post-traductionnelles (post-translational modifications, PTM). Les PTM peuvent représenter, par exemple, l'addition ou la suppression d'un groupe méthyle (CH3) sur un substrat. Ces modifications et les signalisations qui en découlent jouent un rôle essentiel dans l'activité des protéines, la régulation d'expression des gènes et le bon fonctionnement d'une cellule. La dérégulation de ces signalisations par méthylation anormale de certaines protéines, en particulier des protéines d'histones ou des facteurs de transcription, peuvent conduire à la formation du cancer.

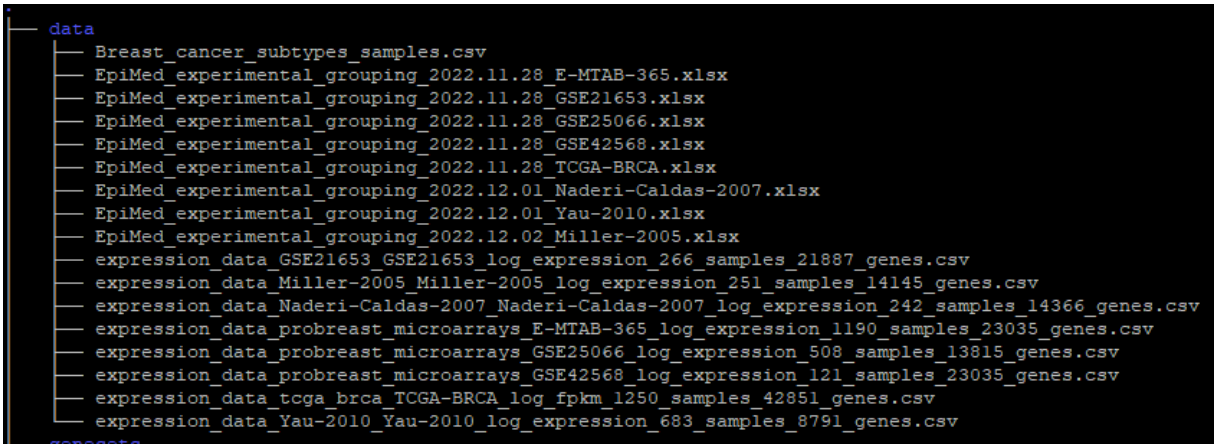
Objectif

L'objectif du projet consiste à étudier le rôle des enzymes qui régulent la méthylation des lysines dans le cancer du sein afin d'identifier de nouvelles cibles thérapeutiques en utilisant des approches dites épigénétiques.

Données

Les données sont disponibles à l'adresse suivante (voir le dossier « data ») :

http://epimed.univ-grenoble-alpes.fr/downloads/uga_imag_ssd/



Les données contiennent plusieurs datasets du cancer du sein.

Dataset	**Technologie**
TCGA-BRCA	RNA-seq
GSE25066	Microarrays
GSE21653	Microarrays
GSE42568	Microarrays
Yau-2010 (PMID 16936776)	Microarrays
E-MTAB-365	Microarrays
Miller-2005 (alias GSE3494, GSE4922)	Microarrays

Naderi-Caldas-2007 (PMID 20946665)	Microarrays

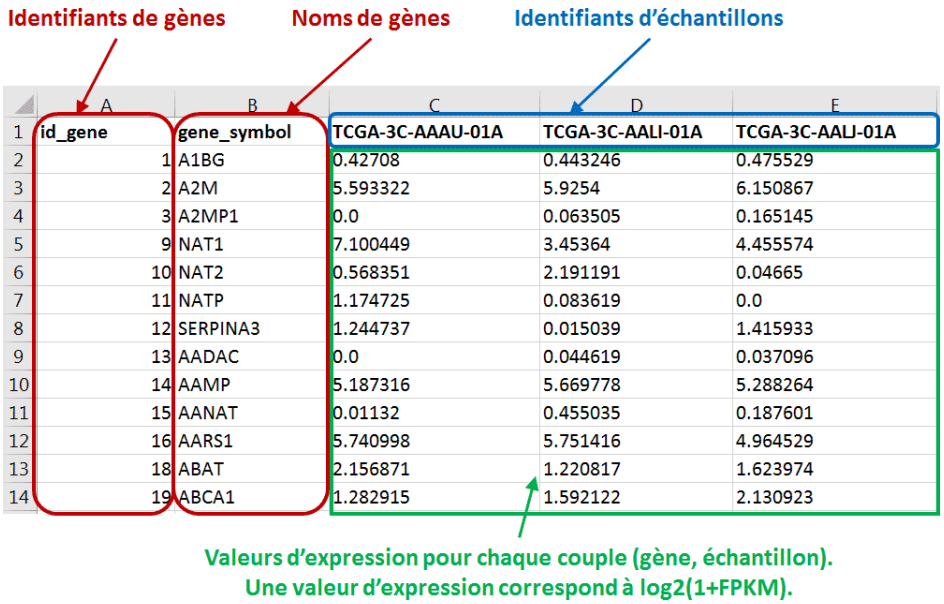
Pour chaque dataset, deux types de fichier sont disponibles :

- Les niveaux d'expression de gènes
  (fichier nommé `expression_data*.csv`, séparateur « ; »)
- Les annotations biocliniques des échantillons
  (fichier nommé `EpiMed_experimental_grouping*.xlsx`)

Données d'expression de gènes

(`expression_data*.csv`)

Les données transcriptomiques sont disponibles sous la forme d'une matrice de valeurs d'expression de gènes, normalisées et log-transformées pour chaque paire gène-échantillon.



	A	B	C	D	E
1	id_gene	gene_symbol	TCGA-3C-AAAU-01A	TCGA-3C-AALI-01A	TCGA-3C-AALJ-01A
2	1	A1BG	0.42708	0.443246	0.475529
3	2	A2M	5.593322	5.9254	6.150867
4	3	A2MP1	0.0	0.063505	0.165145
5	9	NAT1	7.100449	3.45364	4.455574
6	10	NAT2	0.568351	2.191191	0.04665
7	11	NATP	1.174725	0.083619	0.0
8	12	SERPINA3	1.244737	0.015039	1.415933
9	13	AADAC	0.0	0.044619	0.037096
10	14	AAMP	5.187316	5.669778	5.288264
11	15	AANAT	0.01132	0.455035	0.187601
12	16	AARS1	5.740998	5.751416	4.964529
13	18	ABAT	2.156871	1.220817	1.623974
14	19	ABCA1	1.282915	1.592122	2.130923

En fonction du dataset, les valeurs peuvent correspondre aux données normalisées FPKM (pour les données RNA-seq) ou à d'autres valeurs d'expression normalisées (pour les données Microarrays). Dans les deux cas, les données sont directement exploitables pour faire des analyses statistiques.

Annotations biocliniques

(`EpiMed_experimental_grouping*.xlsx`)

Les annotations biocliniques dans les fichiers Excel sont présentées avec deux onglets : « standard exp_group » et « original parameters ». Les données de survie sont disponibles dans les colonnes suivantes dans l'onglet « standard exp_group ».

Colonne	**Signification**
os_month	La durée de la survie globale OS (overall survival) en mois.
os_censor	La censure de la survie globale.

	0 – donnée censurée, c'est-à-dire, l'évènement (décès) n'a pas été observé pendant la durée de suivi indiqué dans « os_months ». 1 – l'évènement a été observé après la durée indiqué dans « os_months ».
dfs_month	La durée de la survie sans rechute DFS (disease-free survival) en mois.
dfs_censor	La censure de la survie sans rechute.

Il existe plusieurs sous-types du cancer du sein. Les échantillons sont annotés selon les sous-types. Les annotations des sous-types sont disponibles dans le fichier CSV `Breast_cancer_subtypes_samples.csv` (séparateur « ; »).

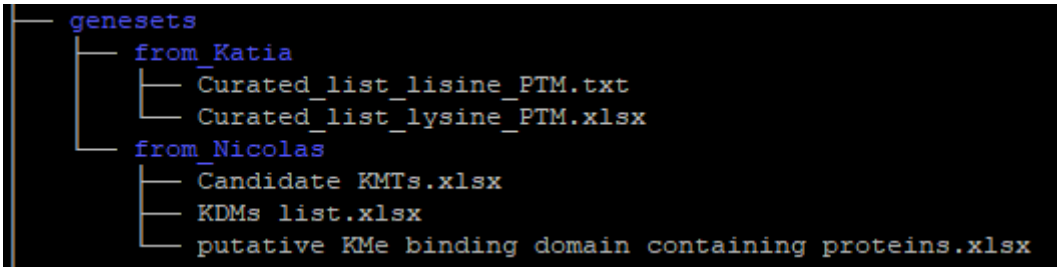
Les sous-types à considérer dans ce projet.

Nom du groupe	**Sous-types du groupe**
Cancer-normal	Non-tumour, All-tumours
Sous-type moléculaire	Luminal-A, Luminal-B, HER2-enriched, Basal-like
Stade du cancer	Stage-I, Stage-II, Stage-III, Stage-IV Est-ce qu'on regroupe Stage-III et Stage-IV en Stage-III-IV car seulement 18 échantillons sont disponibles dans Stage-IV ?
TNM	N0, N1, N2, N3, M1

Les gènes d'intérêt

Voir le dossier « genesets »

http://epimed.univ-grenoble-alpes.fr/downloads/uga_imag_ssd/



« from Nicolas » : fichiers originaux avec différentes listes des enzymes de méthylation

« from Katia » : la liste simplifiée, contient uniquement les enzymes reconnus dans la base NCBI, à voir avec Nicolas pour une mise à jour éventuelle

Méthodologie

Analyse d'expression de gènes

L'objectif de cette partie consiste à identifier les gènes d'intérêt qui sont significativement sur-exprimés dans un sous-type du cancer du sein versus les autres sous-types dans chaque groupe.

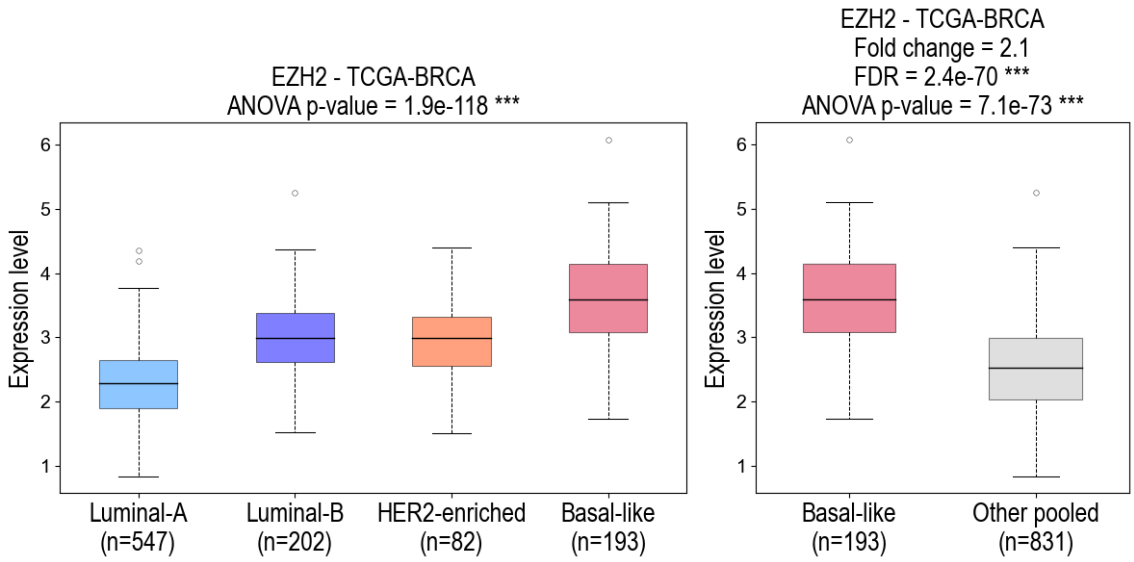
Par exemple, dans le groupe « Cancer-normal », on compare les niveaux d'expression de chaque gène d'intérêt dans les sous-types « Non-tumour » et « All-tumours ». Par le T-test (pairwise ANOVA) ou le

U-test (Mann-Whitney), on identifie les gènes avec des niveaux d'expression significativement plus élevés dans les cancers « All-tumours » comparé aux échantillons non tumoraux « Non-tumours ».

Dans les groupes avec plusieurs sous-types, on compare à tour de rôles l'expression dans un sous-type versus les autres. Par exemple, dans les sous-types moléculaires, pour chaque gène on compare d'abord « Luminal-A » versus les autres : « Luminal-B », « HER2-enriched » et « Basal-like » ensemble. Ensuite, on compare « Luminal-B » versus « Luminal-A », « HER2-enriched » et « Basal-like » ensemble, en ainsi de suite.

Exemple

Le gène EZH2 est surexprimé dans le sous-type moléculaire Basal-like.



On réalise cette analyse tout d'abord dans le dataset TCGA-BRCA où tous les sous-types sont disponibles. On la répète ensuite dans les autres datasets quand c'est possible, pour les sous-types moléculaires en particulier.

On calcule la p-value et le fold change entre les moyennes. Puisque les données sont déjà en format log2, pour estimer le fold change on peut calculer la différence des valeurs moyennes d'expression.

log2(A/B) = log2(A) — log2(B)

log2(fold change) = mean expression (« All-tumours ») — mean expression (« Non-tumour »)

fold_change = 2 ** (log2(fold_change)))

Les p-valeurs obtenues peuvent être corrigées pour des tests multiples selon l'algorithme de Benjamini-Horchberg.

Le résultat significatif correspond à la p-value ajustée (FDR) < 0.05 et le fold change > 1.5 (surexpression).

Quels gènes sont surexprimés dans quels sous-types ?

Est-ce que les résultats sont confirmés dans plusieurs datasets ?

Pour quels gènes le résultat est confirmé dans le plus grand nombre de datasets ?

Exemple de livrable

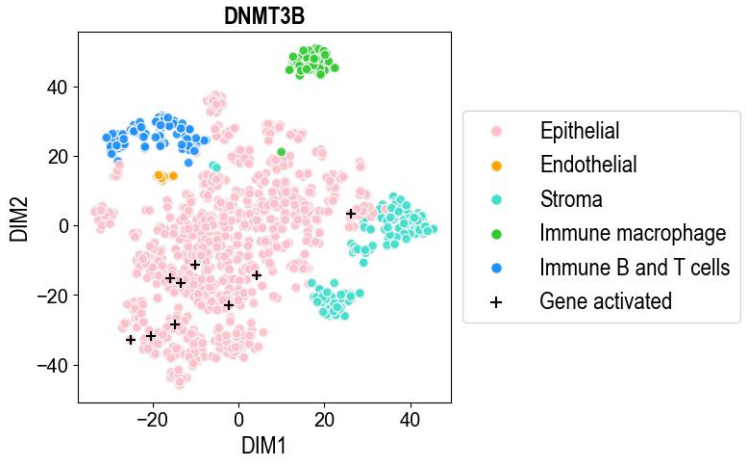
Un tableau de p-valeurs, FDR et fold change obtenus pour chaque gène dans chaque sous-type et chaque dataset.

feature	group_name	dataset_name	log2fc	fc	pval	fdr
ANTKMT	Luminal-B	GSE21653	0.607088704	1.523182389	7.51986E-09	7.32389E-08
ANTKMT	Luminal-B	E-MTAB-365	0.602809133	1.518670758	1.7062E-16	3.71412E-15
ASH2L	Luminal-B	GSE21653	0.649387343	1.568501972	8.04587E-09	7.7873E-08
ATG16L1	Luminal-A	GSE21653	0.585639854	1.500704423	5.54234E-08	4.76242E-07
BAZ2B	Luminal-A	Yau-2010	0.587090112	1.502213753	1.90775E-12	2.77182E-11
BMT2	Basal-like	GSE42568	0.646363915	1.565218335	1.10129E-05	6.35575E-05
BOP1	Basal-like	TCGA-BRCA	1.149144568	2.217823517	9.05193E-57	1.65171E-54
BOP1	Basal-like	GSE21653	0.862002629	1.817559545	3.04722E-12	4.36101E-11
BOP1	Basal-like	GSE42568	0.842374524	1.792998802	3.85897E-05	0.000201694
BOP1	Basal-like	GSE25066	0.73784087	1.667678131	4.01372E-22	1.40684E-20

[Optionnel] Analyse de données « single cell »

Pour les gènes les plus intéressants, sur-exprimés dans le sous-type « Basal-like » uniquement, regarder dans quels types de cellules ils sont exprimés : cellules tumorales, cellules immunitaires etc. Pour cela, utiliser le dataset supplémentaire « single cell » avec 6 échantillons du cancer « Basal-like ».

Exemple pour le gène DNMT3B, projection t-SNE (la figure vient d'un autre projet).



Analyse de survie

L'objectif de cette partie est d'identifier les gènes qui ont un impact sur la survie des patients dans toute la cohorte ou dans les sous-types.

En analyse de survie on utilise habituellement deux modèles : le test statistique du logrank et le modèle à risque proportionnel de Cox. Le test du logrank évalue si deux populations de patients ont une probabilité de survie similaire (hypothèse H0) ou différente (hypothèse H1). Le modèle de Cox ressemble à une régression linéaire, adaptée à des données de survie. Il permet de savoir si une variable explicative est corrélée à la probabilité de survie des patients.

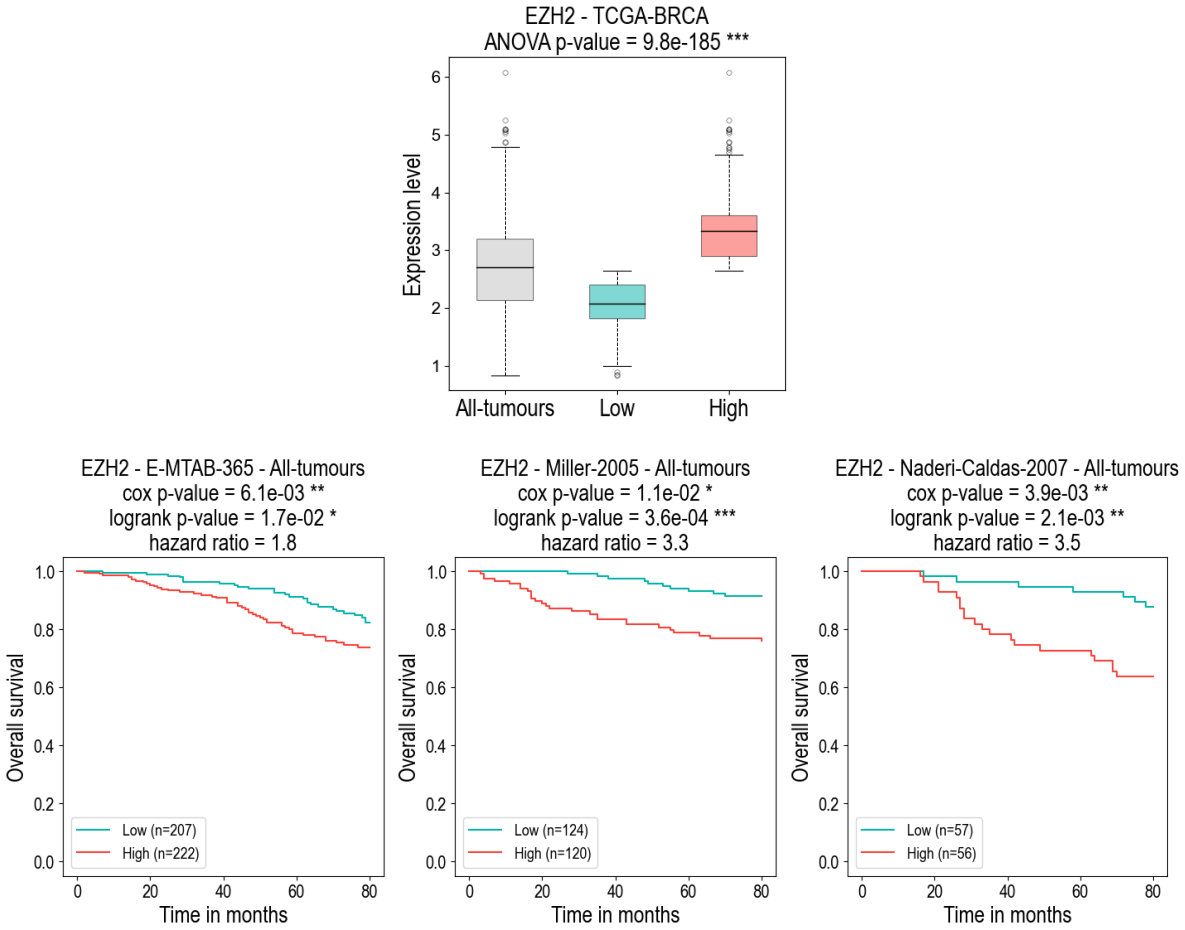
Pour le modèle de Cox, on estime l'impact de la valeur numérique du niveau d'expression du gène sur la probabilité de survie. Est-ce que l'augmentation du niveau d'expression est associée à l'augmentation du risque de décès. On calcule la p-valeur associée et le hasard ratio du modèle.

Si besoin, vous pouvez suivre des cours enregistrés (vidéos courtes de 10-20 min), consacrées à l'analyse de survie :
https://www.youtube.com/playlist?list=PLng25Lwn1Xd1yqQZRBgsOHbRoSn0kJd1h

Pour chaque gène séparément, on va appliquer les deux modèles de la façon suivante. Pour le test du logrank, on sépare les patients en deux groupes selon la valeur d'expression du gène : « low » si le niveau d'expression est inférieur à la médiane ou « high » si le niveau d'expression est supérieur à la médiane. On compare la probabilité de survie entre les groupes « low » et « high » en calculant la p-valeur associée.

Exemple

Le niveau d'expression du gène EZH2 est significativement associé à la survie des patients. Le résultat est confirmé dans plusieurs datasets.



On fera le calcul des p-valeurs de logrank et de Cox ainsi que du hasard ratio :

- Pour chaque gène
- Dans chaque dataset
- Dans chaque sous-type (sauf « Non-tumour »)
- Pour la survie globale (OS) et survie sans rechute (DFS). Plusieurs datasets n'ont pas de données sur la survie globale, uniquement sur la survie sans rechute.

En tant que critère de significativité, on peut considérer :

- Soit les deux p-valeurs de cox et logrank < 0.05 à la fois
- Soit l'une ou l'autre des p-valeurs < 0.05

Le choix sera fait en fonction des résultats obtenus. Si aucun gène ne correspond au premier critère, on utilise le second.

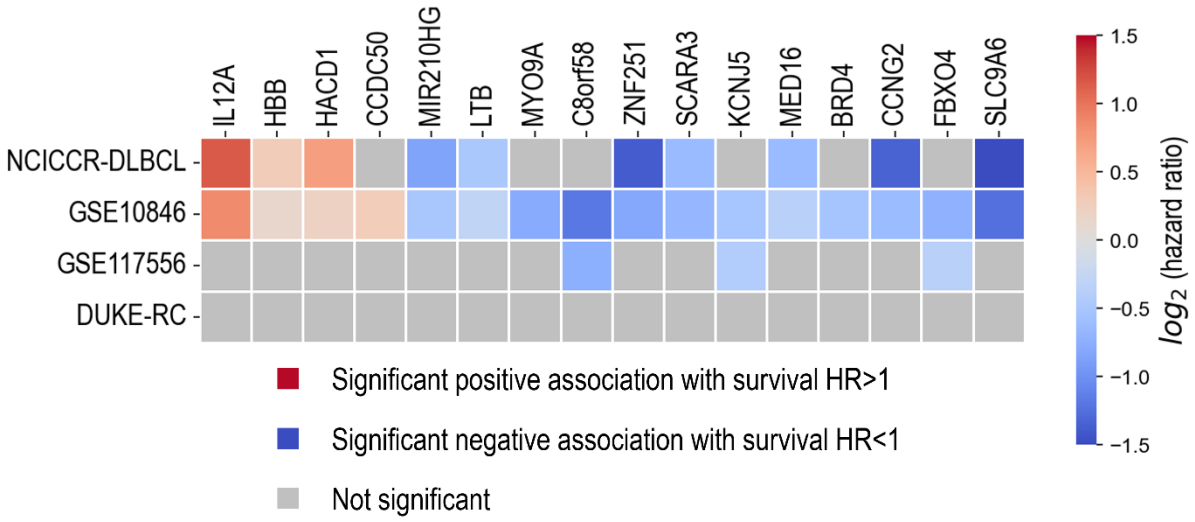
Quels sont les gènes dont le niveau d'expression est associé à la survie de façon robuste, c'est-à-dire, avec validation dans plusieurs datasets ?

Remarque

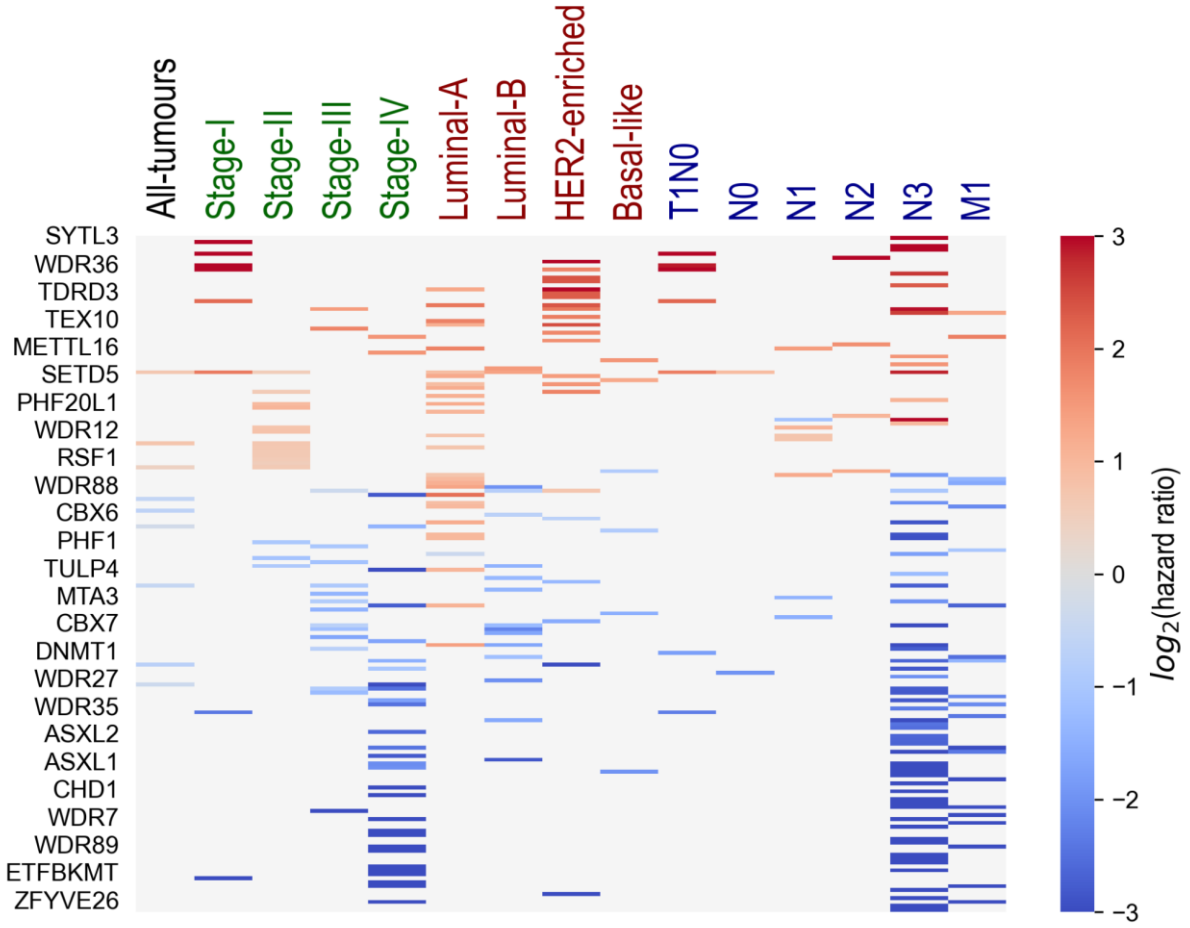
Je ne calcule pas toujours le FDR pour l'analyse de survie car la distribution des p-valeurs ajustées souvent n'est pas homogène. J'utilise plutôt le critère de validation dans plusieurs datasets pour sélectionner les candidats les plus stables.

Exemple de présentation

Pour présenter les résultats, vous pouvez sélectionner uniquement les gènes candidats les plus intéressants qui valident dans N datasets à la fois et puis afficher le hasard ratio correspondant en couleur selon cet exemple : l'axe « x » liste les datasets et l'axe « y » montre les gènes les plus intéressants (la figure vient d'un autre projet).



Ou alors, sur le même modèle, présenter une matrice similaire mais cette fois « les gènes versus les sous-types », en sélectionnant uniquement les résultats significatifs validés dans au moins N datasets.



Analyse d'enrichissement GSEA

Pour cette étape, il faut mieux qu'on fasse une réunion dédiée ensemble pour que je vous montre plus en détail comment faire le calcul. Cette partie est légèrement plus complexe que les autres.

Le Gene Set Enrichment Analysis (GSEA) est une méthode d'analyse de données d'expression de gènes largement utilisée dans la littérature en biologie. Elle vise à comparer un vecteur de données numériques, indexé par l'ensemble des gènes du génome, à un certain ensemble de gènes (gene set), connu pour être associé à une signature moléculaire d'un phénomène biologique spécifique.

La méthode GSEA a été initialement décrite par Mootha et al. (2003) et ensuite révisée et généralisée par Subramanian et al. (2005). Elle repose sur un test d'enrichissement dérivé du test pondéré de Kolmogorov-Smirnov. Dans un premier temps, les différents gènes sont ordonnés sur la base d'un score qui représente la différence entre deux conditions comparées. En pratique, le score correspond souvent soit au fold change, calculé lors de l'analyse différentielle entre deux conditions, soit à la p-valeur obtenue dans cette analyse pour chaque gène. Chaque gène obtient ainsi un rang selon le score attribué. L'objectif de la méthode est d'évaluer si le gene set est distribué au hasard au sein de la liste ordonnée de gènes ou s'il tend à être accumulé au début ou à la fin de la liste (Charmpi, 2015).

La procédure GSEA consiste, pour chaque gene set, à calculer un score d'enrichissement ES, équivalent à la statistique de Kolmogorov-Smirnov normalisée, en parcourant la liste de gènes, de la statistique la

plus élevée à la statistique la plus faible. La valeur finale utilisée pour caractériser statistiquement le groupe de gènes correspond à la valeur maximale atteinte par ES au cours de ce parcours. L'hypothèse nulle qui est testée est « le groupe de gènes n'est pas associé aux conditions comparées », et l'attribution de la p-valeur du score d'enrichissement est calculée en effectuant des permutations d'étiquettes d'échantillons (Berger, 2009). La correction pour tests multiples repose sur une normalisation de la valeur ES de chaque groupe de gènes, tenant compte de la taille du gene set (NES = score d'enrichissement normalisé). Le FDR est ensuite calculé en comparant la distribution observée et la distribution nulle de la statistique normalisée (NES).

L'analyse GSEA a été développé par Broad Institute.

https://www.gsea-msigdb.org/gsea/index.jsp

Le logiciel qui effectue l'analyse GSEA est disponible en versions Java, Python et R. La version R n'est plus maintenue, il y a des problèmes d'installation. La versions Python (package gseapy) ou Java (utilisable en batch) sont recommandées pour automatiser l'analyse.

Pour installer le package gseapy : https://pypi.org/project/gseapy/

Nous allons utiliser uniquement la fonction « prerank » qui effectue l'analyse d'enrichissement.

Documentation : https://gseapy.readthedocs.io/en/latest/gseapy_example.html#3.-Prerank-example

Remarque

La p-valeur minimale dans l'analyse GSEA dépend du nombre de permutations choisi dans les options. Le nombre de permutations par défaut est 1000. Quand l'analyse GSEA indique la p-valeur = 0.000000, cela signifie en réalité que la p-valeur est inférieure à 0.001 (=1/1000). On peut poser p-val=0.001 si p-val<0.001 et FDR=0.001 si FDR<0.001 dans ces cas.

Le seuil de significativité conseillé par l'analyse GSEA : FDR<0.25.

Détail du calcul

En utilisant les données TCGA-BRCA uniquement, pour chaque échantillon tumoral, pour chaque gène, calculer la différence entre le niveau d'expression du gène dans l'échantillon tumoral et le niveau moyen d'expression du gène dans les échantillons non-tumoraux (« Non-tumours »).

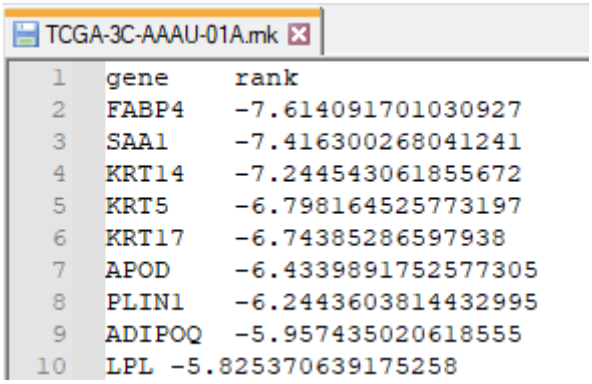
Attention, on compare bien le niveau d'expression dans chaque échantillon tumoral séparément, échantillon par échantillon, avec la valeur moyennée du niveau d'expression dans les échantillons non-tumoraux.

dif = expression in tumour — mean(normal expression)

Pour chaque échantillon, ordonner la liste de tous gènes selon la valeur dif, de la plus petite (négative) à la plus grande (positive).

Sauvegarder le résultat dans une fichier ASCII avec l'extension « .rnk » pour une utilisation future dans l'analyse GSEA. La première colonne contient le nom du gène, la deuxième colonne contient la valeur dif ordonnée. Les colonnes sont séparées par une tabulation « \t ».

Exemple pour l'échantillon TCGA-3C-AAAU-01A.rnk :



TCGA-3C-AAAU-01A.mk

	gene	rank
1		
2	FABP4	-7.614091701030927
3	SAA1	-7.416300268041241
4	KRT14	-7.244543061855672
5	KRT5	-6.798164525773197
6	KRT17	-6.74385286597938
7	APOD	-6.4339891752577305
8	PLIN1	-6.2443603814432995
9	ADIPOQ	-5.957435020618555
10	LPL	-5.825370639175258

On fait l'analyse d'enrichissement GSEA pour chaque fichier « .rnk » et pour plusieurs liste de gènes (gene sets), composées de gènes d'intérêt.

Générer un clustering hiérarchique en fonction des valeurs NES obtenues. L'axe « x » : les échantillons du cancer du sein, l'axe « y » : les gene sets.

Bibliographie

Berger, F., 2009. Développement critique de méthodes d'analyse de l'expression différentielle de gènes et de groupes de gènes, mesurée sur damiers à ADN. Thèse de doctorat soutenue aux Facultés Universitaires Notre-Dame de la Paix, Université de Namur, Belgique.

Charmpi, K., 2015. Méthodes statistiques pour la fouille de données dans les bases de données de génomique (Gene Set Enrichment Analysis). Thèse de doctorat soutenue à l'Ecole Doctorale MSTII, Université Grenoble Alpes, France.

Mootha, V., Lindgren, C., Eriksson, KF. et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34, 267–273 (2003). https://doi.org/10.1038/ng1180

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS, 102(43):15545–15550, 2005.