

Analyse multi-omique pan-cancer de la régulation de la transcription par la méthylation de l'ADN

Projet réalisé par :

Cherki Amine

Chikh Hamouda Safa

Balde Mahmoud

Quitian Jhorman

Tuteur :

M. Chuffart Florent

Sommaire

Analyse multi-omique pan-cancer de la régulation de la transcription par la méthylation de l'ADN	1
Introduction	3
I. Contexte biologique	4
II. Pattern H1F0	6
III. Scores	7
1. Corrélation	7
a. Définition du score de corrélation	9
b. Evaluation	9
c. Critique	10
2. Aire et Écart-Type	10
a. Aire - Ecart-type	10
b. Paramètre de dispersion et sélection des sondes	11
3. Score ACP_PS2	14
IV. Catalogue	16
1. ACP sur les scores	17
V. Résultats (descriptions du catalogue)	18
VI. Conclusion	25
Annexe	25

Introduction

Le cancer est une maladie caractérisée par la prolifération incontrôlée de cellules, liée à un échappement aux mécanismes de régulation qui assure le développement harmonieux de notre organisme.

Ces cellules dérégées finissent parfois par former une masse qu'on appelle tumeur maligne. Les cellules cancéreuses ont tendance à envahir les tissus voisins et à se détacher de la tumeur initiale. Elles migrent alors par les vaisseaux sanguins et les vaisseaux lymphatiques pour aller former une autre tumeur (métastase).

La recherche sur le cancer a été dominée, dès les années 80, par les avancées scientifiques démontrant l'origine génétique des processus tumoraux. Des milliers d'altérations génétiques ont ainsi été répertoriées, impliquant plus d'une centaine de gènes.

Depuis dix ans, ce modèle a évolué : les cancers sont aujourd'hui des maladies autant génétiques qu'épigénétiques. En altérant l'expression de gènes impliqués dans la régulation cellulaire, les modifications épigénétiques jouent un rôle fondamental dans l'initiation et la progression des tumeurs ; contrairement aux mutations génétiques, elles sont potentiellement réversibles. Des inhibiteurs épigénétiques sont ainsi évalués comme agents antitumoraux.

Par ailleurs, l'étude de la méthylation de l'ADN se profile comme un marqueur biologique pouvant contribuer à la classification tumorale, au diagnostic et au pronostic en pratique clinique.

I. Contexte biologique

La méthylation de l'ADN du promoteur d'un gène contrôle son niveau d'expression. Un haut niveau de méthylation du promoteur d'un gène est souvent associé à une répression de ce gène (pas de transcription). À l'inverse, un faible niveau de méthylation du promoteur d'un gène est souvent associé à une expression de ce gène (transcription).

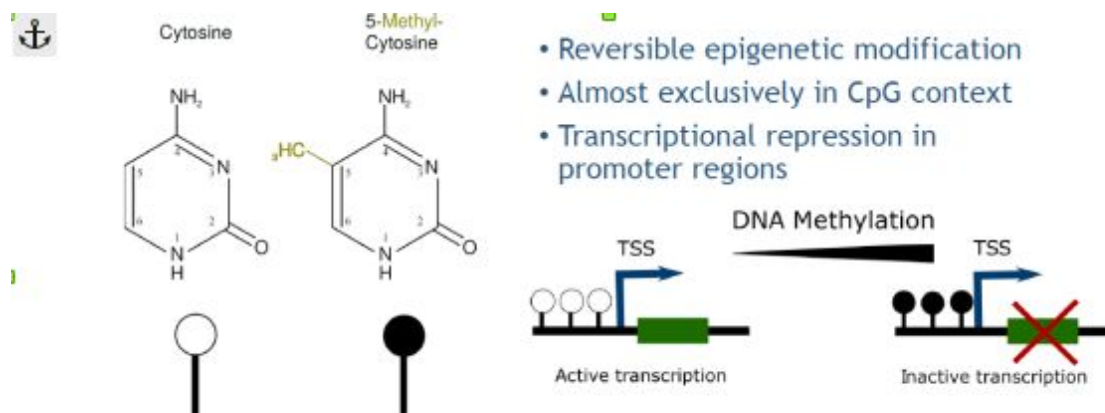


Figure 1 Méthylation décrite par Michael Scherer

Le génome humain contient autour de 25000 gènes, 2/5 de ces gènes (15000) sont dits CpG riches, c'est à dire qu'ils contiennent un îlot CpG dans leur promoteur ou encore que la séquence d'ADN située en amont de ces gènes est beaucoup plus riche en CG qu'une distribution aléatoire. Un dinucléotide CpG, parfois appelé site CpG en référence à l'anglais CpG site, est un segment d'ADN de deux nucléotides dont la séquence de bases nucléiques est CG. La notation « CpG » est une abréviation de cytosine–phosphate–guanine destinée à être clairement distinguée de la notation « CG » qui peut également désigner une paire de bases sur deux brins d'ADN distincts et non la séquence d'un brin d'ADN donné.

Dans la plupart des tissus sains, 2/3 des gènes CpG riches (16600 environ) seront constamment déméthylés, 1/3 des gènes CpG riches (8500) seront constamment méthylés sauf pour quelques tissus. On parle alors de gènes-tissus

spécifiques et c'est la méthylation de l'ADN qui va contrôler cette spécificité (contrôle de l'expression par la méthylation). Dans les faits, les promoteurs sont méthylés ou méthylés et cet état est stable dans un tissu donné.

Dans les cancers, cette stabilité est remise en cause. Par exemple cet article "**The linker histone H1.0 generates epigenetic and functional intratumor heterogeneity**" par Cristina Morales Torres et al., Science 2016, met en évidence, à l'intérieur d'une même cohorte, différents niveaux de méthylation du promoteur du gène **H1F0**.

De plus, les auteurs corrélaient négativement le niveau d'expression de ce gène et le niveau de méthylation de son promoteur. Ainsi les cancers deviennent des modèles biologiques capables de mettre en évidence des régions du génomes plastiques du point de vue de la méthylation (hotspot épigénétique). L'identification de ces régions est un enjeu majeur pour comprendre les mécanismes de l'oncogenèse (diagnostic), relier ces mécanismes à l'agressivité des cancers (pronostic) et identifier des cibles potentielles pour des futurs traitements.

L'objectif de cette étude est de produire un catalogue de hotspots épigénétiques. Cela passe par la recherche systématique de gènes dont on pourra observer une variation de l'expression en fonction du niveau de méthylation de leur promoteur. On utilisera les cohortes pan-cancer du **TCGA** (The Cancer Genome Atlas).

L'Atlas du génome du cancer (ou The Cancer Genome Atlas en anglais, abrégé en TCGA) est un projet qui fut lancé en 2005 pour cataloguer les mutations génétiques responsables du cancer en utilisant le séquençage génomique et la bio-informatique. TCGA applique des techniques d'analyse génomique à haut débit pour améliorer notre capacité à diagnostiquer, traiter et prévenir le cancer grâce à une meilleure compréhension de la base génétique du cancer.

Le projet fournit une exploration systématique des changements génétiques impliqués dans plus de 20 types de cancers chez l'humain. Le portail contient des données (information clinique, caractérisation génomique et analyse des séquences associées aux tumeurs) et des outils intégrés d'analyse.

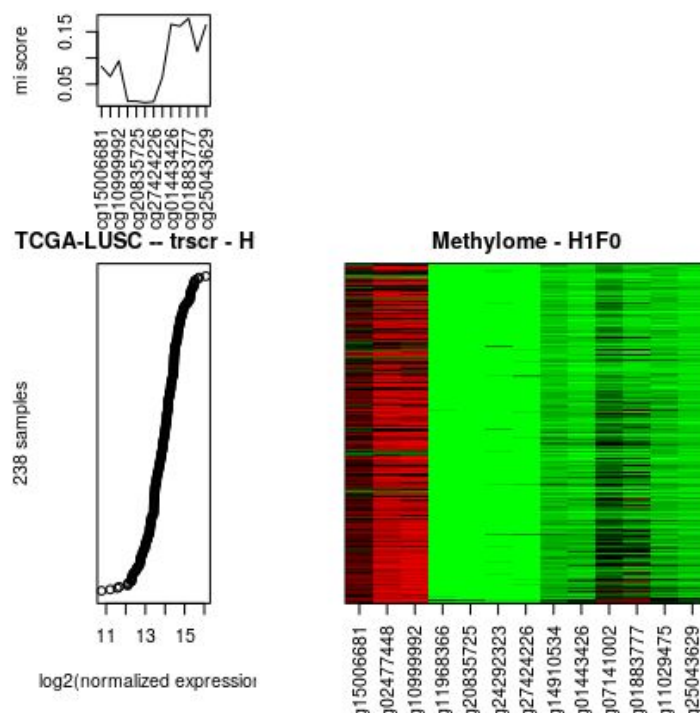
Pour ce projet nous disposons d'une base de données composée de plus de 30 types de cancers chez l'humain. Nos travaux se basent sur l'étude du gène H1F0 pour le cancer **LUSC** (Lung Squamous Cell Carcinoma - Carcinome épidermoïde pulmonaire).

Problématique :

Est-ce que le développement d'un score calibré sur le promoteur de "H1F0-LUSC" permettrait d'identifier d'autres gènes avec des promoteurs présentant une plasticité génomique ?

II. Pattern H1F0

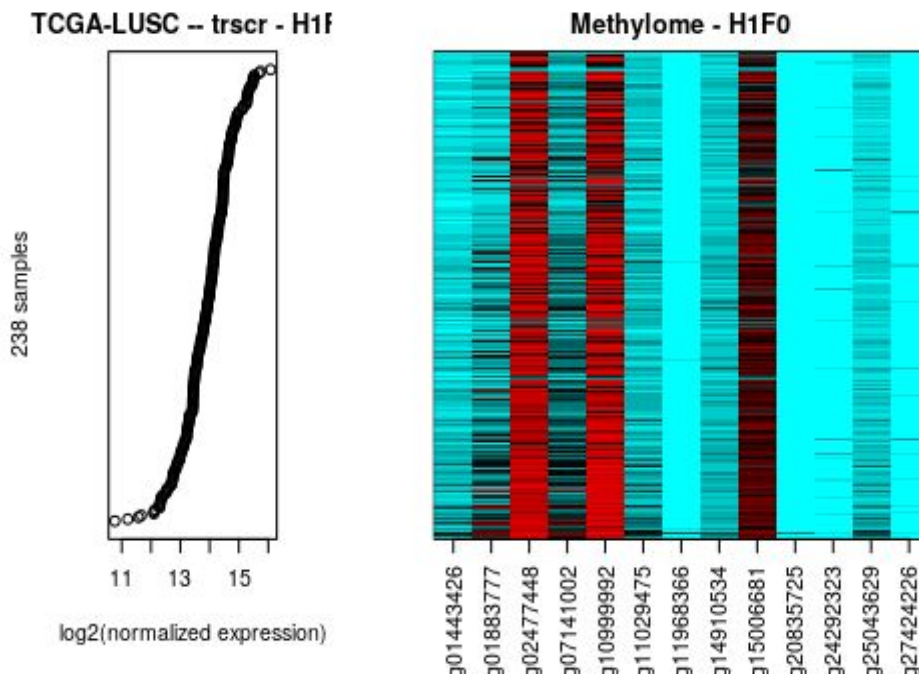
Les profils transcriptomiques et de méthylation du gène H1F0 sont le socle sur lequel se base notre travail. Nous avons utilisé une heatmap pour représenter le méthylome. Nous avons reproduit la représentation de Cristina Morales pour le gène H1F0 et TCGA-LUSC.



Le gradient des couleurs s'étale de vert à rouge ainsi une sonde méthylée sera rouge et une sonde déméthylée sera verte. Le profil transcriptomique et le méthylome se lisent en parallèle. Chaque ligne correspond à un degré de transcription et à un profil de méthylation. Le pattern intéressant est produit par les sondes suivantes : cg01443426, cg01883777, cg07141002, cg11029475,

cg14910534. Ce pattern est un hotspot épigénétique (pattern en flamme) car il témoigne de la plasticité du génome au niveau de H1F0. Ainsi pour un même cancer (Carcinome épidermoïde pulmonaire) on observe des individus avec un promoteur plus ou moins méthylé.

Dans la suite de notre étude nous présenterons la méthylation d'une autre manière. Notre gradient de couleur ira du bleu au rouge et les sondes seront triées par chromosome et par brin.



Nous allons par la suite présenter des scores calibrés sur le méthylome H1F0-LUSC.

III. Scores

1. Corrélation

Le coefficient de Pearson est un indice reflétant une relation linéaire entre deux variables continues. L'analyse conjointe du méthylome et du transcriptome pour H1F0 se présente sous la forme suivante :

- Transcriptome :

TCGA-85-A4PA-01A	10.76681
TCGA-43-A475-01A	11.20609
TCGA-96-A4JL-01A	11.57246
TCGA-77-7142-01A	11.66572
TCGA-56-5897-01A	12.09305
TCGA-56-8307-01A	12.10320
TCGA-21-5787-01A	12.14188
TCGA-98-7454-01A	12.28617
TCGA-56-A4BY-01A	12.30476
TCGA-58-8387-01A	12.30584
TCGA-39-5030-01A	12.32401
TCGA-77-A5FZ-01A	12.32925
TCGA-43-8116-01A	12.35201
TCGA-98-A53D-01A	12.37051

Le transcriptome se présente sous la forme d'un vecteur contenant 238 valeurs de transcriptions de H1F0 associées aux 238 individus retenus par l'analyse conjointe du méthylome et du transcriptome faite par la fonction "preproc_omics_data" dans la vignette "**01_creation_results.Rmd**". En effet, à partir des données omics elle sélectionne les individus pour lesquels nous avons des données de transcription et de méthylation.

- Méthylome :

	TCGA-85-A4PA-01A	TCGA-43-A475-01A	TCGA-96-A4JL-01A	TCGA-77-7142-01A	TCGA-56-5897-01A	TCGA-56-8307-01A	TCGA-21-5787-01A	TCGA-98-7454-01A	TCGA-56-A4BY-01A	TCGA-58-8387-01A	TCGA-39-5030-01A	TCGA-77-A5FZ-01A	TCGA-43-8116-01A	TCGA-98-A53D-01A
cg15006681	0.72701120	0.62836168	0.74103130	0.72675096	0.58675529	0.47516713	0.745278686	0.675110728	0.69734704	0.64114278	0.65941560	0.67896693	0.58546278	0.62209657
cg02477448	0.91450869	0.92047668	0.90314437	0.84911283	0.86790139	0.89305867	0.890438712	0.845989126	0.93939459	0.89543591	0.86668426	0.88362021	0.86858022	0.78772083
cg10999992	0.93152625	0.88743295	0.95928641	0.84641918	0.89778132	0.70160464	0.926772632	0.933256393	0.92526297	0.84984855	0.90059799	0.90348604	0.82677763	0.85778512
cg11968366	0.02612085	0.02804566	0.36016139	0.02879245	0.05043633	0.03572388	0.016486299	0.026349709	0.04598969	0.03809216	0.02251680	0.03121580	0.02937701	0.02492250
cg20835725	0.01230031	0.01140706	0.28207165	0.01091757	0.01774350	0.01310098	0.009810467	0.009476243	0.01444353	0.01230067	0.00968264	0.01828639	0.01129206	0.01386167
cg24292323	0.02681178	0.03929116	0.05015954	0.01484674	0.02603485	0.02258166	0.013980692	0.020823785	0.02056392	0.01440213	0.04045221	0.02222068	0.08089866	0.04999685
cg27424226	0.01688293	0.01883028	0.08674068	0.01437490	0.01580933	0.02114133	0.016915972	0.027303560	0.01876718	0.01577036	0.01774489	0.02173935	0.02486876	0.05921704
cg14910534	0.17692466	0.25123390	0.39776099	0.10213962	0.22175558	0.10755797	0.137684775	0.172698297	0.15008002	0.19338136	0.21355720	0.17586750	0.18773678	0.19351710
cg01443426	0.17686863	0.42106936	0.49469830	0.12640910	0.21213698	0.07738247	0.227502603	0.175024634	0.36575312	0.15243401	0.30839244	0.15508819	0.20491037	0.16002936
cg07141002	0.47446772	0.71210233	0.73055386	0.41509023	0.40425623	0.33777551	0.718762634	0.397157687	0.59013201	0.38831773	0.65056953	0.40372149	0.40756485	0.34302161
cg01883777	0.36493005	0.72569739	0.72828234	0.66517399	0.35005147	0.28710690	0.719431496	0.290627599	0.65746776	0.30780848	0.60443523	0.36297894	0.42684450	0.20879698
cg11029475	0.17234164	0.44246991	0.49147160	0.48210561	0.16809226	0.26470684	0.291951798	0.163398567	0.33035743	0.25975356	0.35885769	0.23240961	0.23108543	0.16498656
cg25043629	0.16237030	0.38931589	0.38813104	0.41055958	0.14743169	0.17325426	0.232040170	0.138385428	0.24393115	0.19699801	0.31502384	0.15139035	0.21691232	0.12697112

Le méthylome de H1F0-LUSC se présente sous la forme d'une matrice de dimension 13x238, les lignes correspondent aux sondes. Elles sont au nombre de 13 et les colonnes correspondent aux individus qui sont au nombre de 238.

a. Définition du score de corrélation

Nous avons établi la corrélation de Pearson entre la transcription et la moyenne des méthylations des 13 sondes pour chaque individu.

	V1
1	-0.5689399

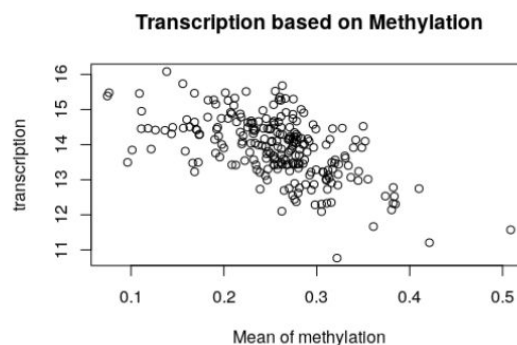
Correlation coefficient

	V1
TCGA-56-6546-01A	14.87830
TCGA-85-A50Z-01A	14.28755
TCGA-18-3409-01A	13.68268
TCGA-52-7809-01A	13.74134
TCGA-21-1080-01A	13.38907
TCGA-77-8008-01A	13.41511
TCGA-66-2780-01A	14.51737
TCGA-56-8083-01A	15.99492
TCGA-85-7843-01A	12.73268
TCGA-63-A5M9-01A	14.20519
TCGA-18-3408-01A	14.66820
TCGA-77-6845-01A	15.83549
TCGA-60-2714-01A	14.99691
TCGA-96-8169-01A	14.10411
TCGA-77-A5G8-01B	14.74105
TCGA-51-4079-11A	13.39895

Transcription values

	V1
TCGA-85-A4PA-01A	0.32177423
TCGA-43-A475-01A	0.42121033
TCGA-96-A4JL-01A	0.50873027
TCGA-77-7142-01A	0.36097637
TCGA-56-5897-01A	0.30509125
TCGA-56-8307-01A	0.26232017
TCGA-21-5787-01A	0.38054284
TCGA-98-7454-01A	0.29812321
TCGA-56-A4BY-01A	0.38457619
TCGA-58-8387-01A	0.30505275
TCGA-39-5030-01A	0.38214849
TCGA-77-A5FZ-01A	0.31084550
TCGA-43-8116-01A	0.31556241
TCGA-98-A53D-01A	0.27791718
TCGA-43-6771-11A	0.27598854
TCGA-56-6545-01A	0.38342721

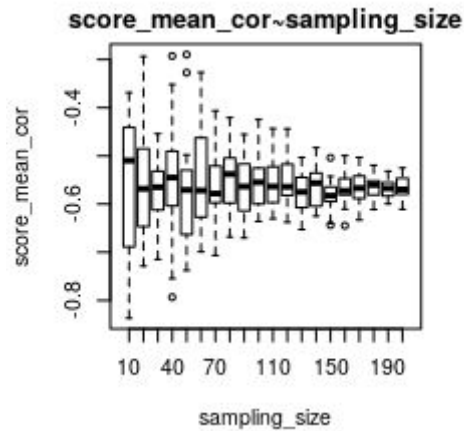
Mean of methylation



On trouve une corrélation négative de ~ -0.57 . Cela correspond à la description faite précédemment par Cristina Morales sur la relation négative entre le niveau d'expression du gène H1F0 et le niveau de méthylation de son promoteur.

b. Evaluation

Nous allons évaluer la stabilité du score par rapport aux individus en fonction du nombre d'individus. Nous avons créé 10 échantillons d'individus tirés aléatoirement allant de 10 à 200 individus. Nous avons répété la démarche 20 fois.



On voit qu'à partir de 20 individus le score moyen est assez stable.

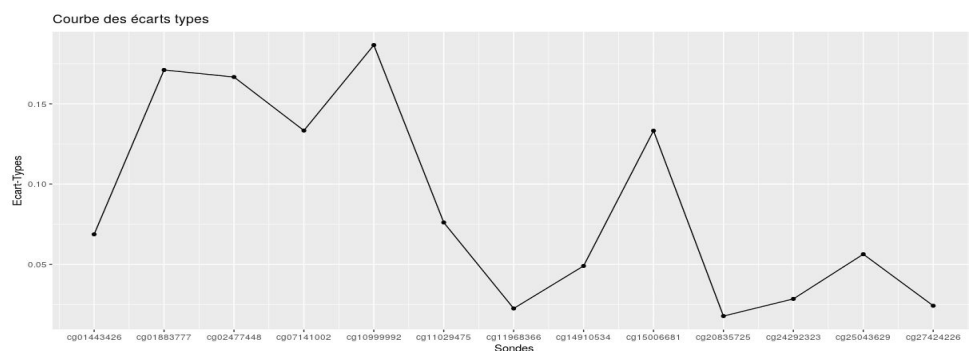
c. Critique

Les sondes non significatives viennent bruite le score, elles occasionnent une perte de puissance. Les sondes corrélées et anti-corrélées s'annulent. On voit ici la nécessité de sélectionner les sondes.

2. Aire et Écart-Type

a. Aire - Ecart-type

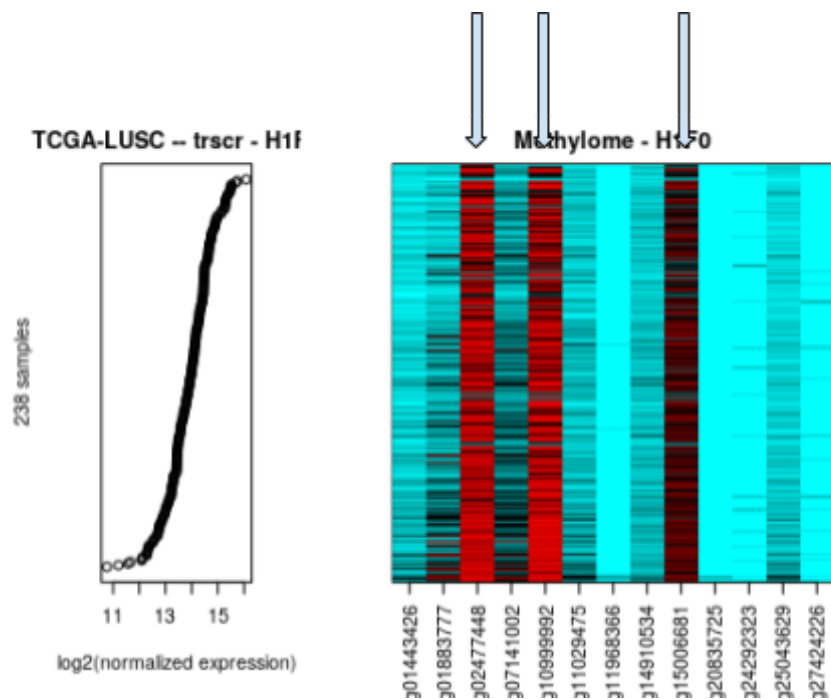
Le score "Aire - Ecart-type" repose comme son nom l'indique sur le calcul d'écart type. En effet, nous faisons la somme des écarts types de chaque sonde ce qui correspond à l'aire sous la courbe ci-dessous :



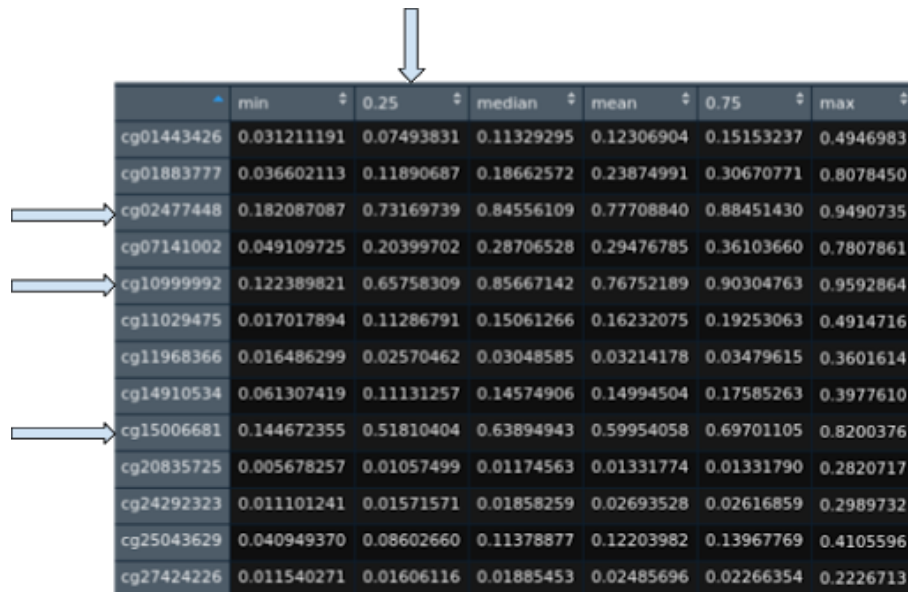
On l'obtient en faisant la somme des écarts types de méthylation de chaque sonde. Nous prenons en compte le fait que le nombre de sondes puisse influencer positivement sur ce score donc nous divisons l'aire par le nombre de sondes.

Cette opération est menée en combinant la fonction `select_probes2` (vignette: 02_catalogue) et la fonction `score`. Ce score sera stocké dans la colonne `area` du catalogue que nous présenterons par la suite.

a. Paramètre de dispersion et sélection des sondes

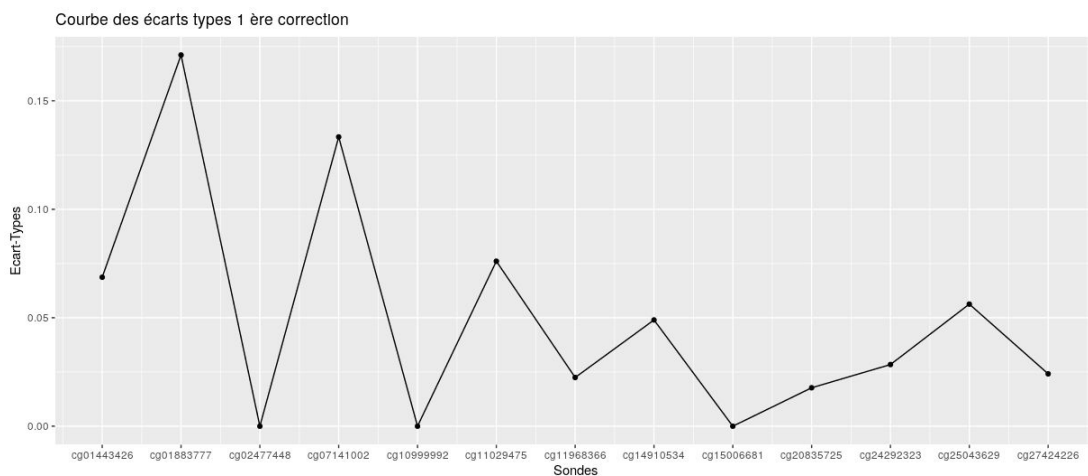


Comme nous pouvons le voir sur la heatmap ci-dessus, les sondes indiquées par une flèche sont des sondes fortement méthylées. L'objectif est de déterminer les caractéristiques de ces sondes pour pouvoir les éteindre. Pour déterminer ces caractéristiques nous avons utilisé les paramètres dispersion que sont les quartiles.

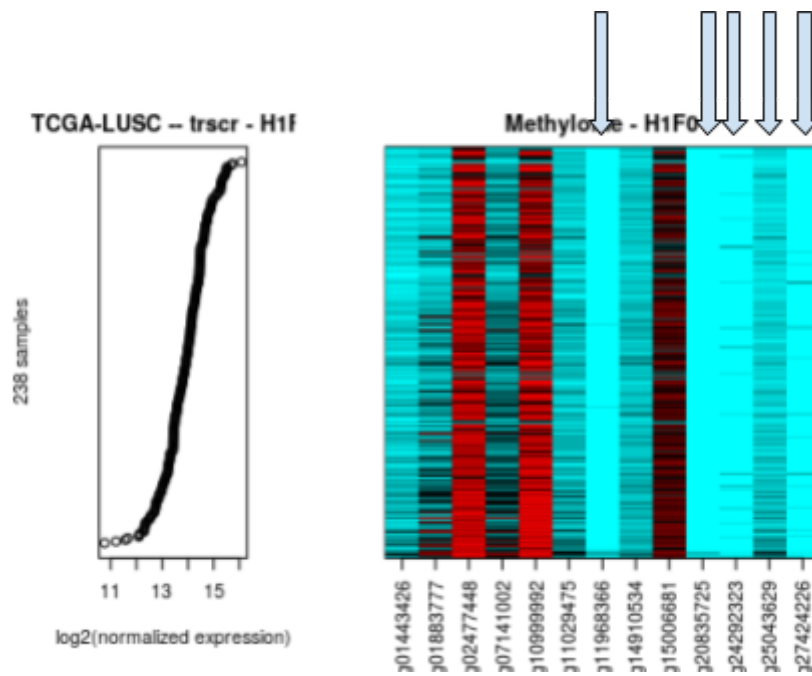


	min	0.25	median	mean	0.75	max
cg01443426	0.031211191	0.07493831	0.11329295	0.12306904	0.15153237	0.4946983
cg01883777	0.036602113	0.11890687	0.18662572	0.23874991	0.30670771	0.8078450
cg02477448	0.182087087	0.73169739	0.84556109	0.77708840	0.88451430	0.9490735
cg07141002	0.049109725	0.20399702	0.28706528	0.29476785	0.36103660	0.7807861
cg10999992	0.122389821	0.65758309	0.85667142	0.76752189	0.90304763	0.9592864
cg11029475	0.017017894	0.11286791	0.15061266	0.16232075	0.19253063	0.4914716
cg11968366	0.016486299	0.02570462	0.03048585	0.03214178	0.03479615	0.3601614
cg14910534	0.061307419	0.11131257	0.14574906	0.14994504	0.17585263	0.3977610
cg15006681	0.144672355	0.51810404	0.63894943	0.59954058	0.69701105	0.8200376
cg20835725	0.005678257	0.01057499	0.01174563	0.01331774	0.01331790	0.2820717
cg24292323	0.011101241	0.01571571	0.01858259	0.02693528	0.02616859	0.2989732
cg25043629	0.040949370	0.08602660	0.11378877	0.12203982	0.13967769	0.4105596
cg27424226	0.011540271	0.01606116	0.01885453	0.02485696	0.02266354	0.2226713

Dans un premier temps nous voulons exclure les sondes indiquées ci-dessus pour calculer l'aire sous la courbe. Pour cela nous allons discriminer les sondes avec un premier quartile supérieur à **0.4** . Ainsi nous obtenons le graphe ci dessous :

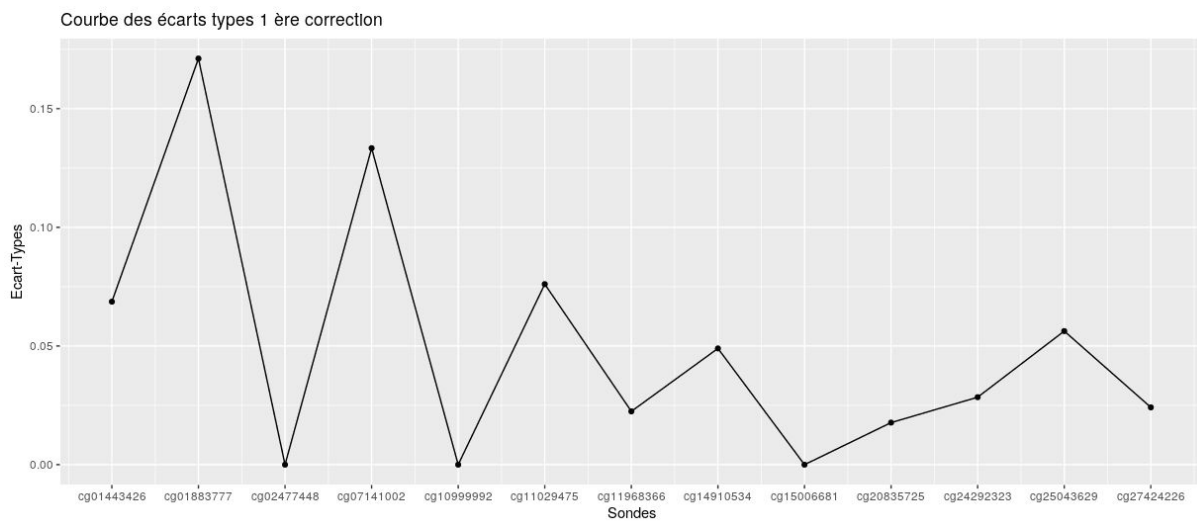


Comme nous pouvons le voir sur la heatmap ci-dessus, les sondes indiquées par une flèche sont des sondes fortement déméthylées. L'objectif est de déterminer les caractéristiques de ces sondes pour pouvoir les éteindre.



	min	0.25	median	mean	0.75	max
cg01443426	0.031211191	0.07493831	0.11329295	0.12306904	0.15153237	0.4946983
cg01883777	0.036602113	0.11890687	0.18662572	0.23874991	0.30670771	0.8078450
cg02477448	0.182087087	0.73169739	0.84556109	0.77708840	0.88451430	0.9490735
cg07141002	0.049109725	0.20399702	0.28706528	0.29476785	0.36103660	0.7807861
cg10999992	0.122389821	0.65758309	0.85667142	0.76752189	0.90304763	0.9592864
cg11029475	0.017017894	0.11286791	0.15061266	0.16232075	0.19253063	0.4914716
cg11968366	0.016486299	0.02570462	0.03048585	0.03214178	0.03479615	0.3601614
cg14910534	0.061307419	0.11131257	0.14574906	0.14994504	0.17585263	0.3977610
cg15006681	0.144672355	0.51810404	0.63894943	0.59954058	0.69701105	0.8200376
cg20835725	0.005678257	0.01057499	0.01174563	0.01331774	0.01331790	0.2820717
cg24292323	0.011101241	0.01571571	0.01858259	0.02693528	0.02616859	0.2989732
cg25043629	0.040949370	0.08602660	0.11378877	0.12203982	0.13967769	0.4105596
cg27424226	0.011540271	0.01606116	0.01885453	0.02485696	0.02266354	0.2226713

Nous voulons exclure les sondes indiquées ci-dessus pour calculer l'aire sous la courbe. Pour cela nous allons discriminer les sondes avec un troisième quartile inférieur à 0.15 . Après cette correction nous obtenons le graphe suivant :

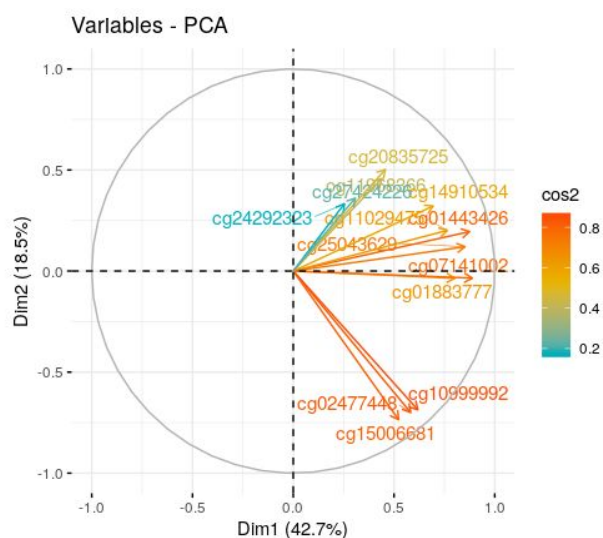


Après avoir opéré ces corrections nous retenons deux scores :

- **area** : la moyenne des écarts types de toutes les sondes multipliée par 100
- **area_selec** : la moyenne des écarts types des sondes sélectionnées après la deuxième correction multipliée par 100

3. Score ACP_PS2

Nous avons procédé à une ACP sur les sondes de H1F0-LUSC :



Nous avons projeté les sondes sur les dimensions 1 et 2. Nous avons remarqué que 3 des 5 sondes d'intérêt se trouvaient dans le quart inférieur droit. Nous avons donc paramétré une fonction `acp_probes_selec` qui compte les sondes qui se trouvent dans cette partie de notre représentation des sondes sur les dimensions 1 et 2 . Après avoir calculé cette donnée, nous faisons le rapport entre le nombre de sondes sélectionnées par l'ACP et le nombre de sondes sélectionnées dans la partie 2. Nous allons nommer ce score **ACP_PS2**.

IV. Catalogue

Après avoir conçu nos scores qui ont été testés et calibrés sur le gène H1F0 du cancer LUSC. Nous avons généralisé ces scores à tous nos gènes présents dans l'étude LUSC. Le catalogue se présente sous cette forme :

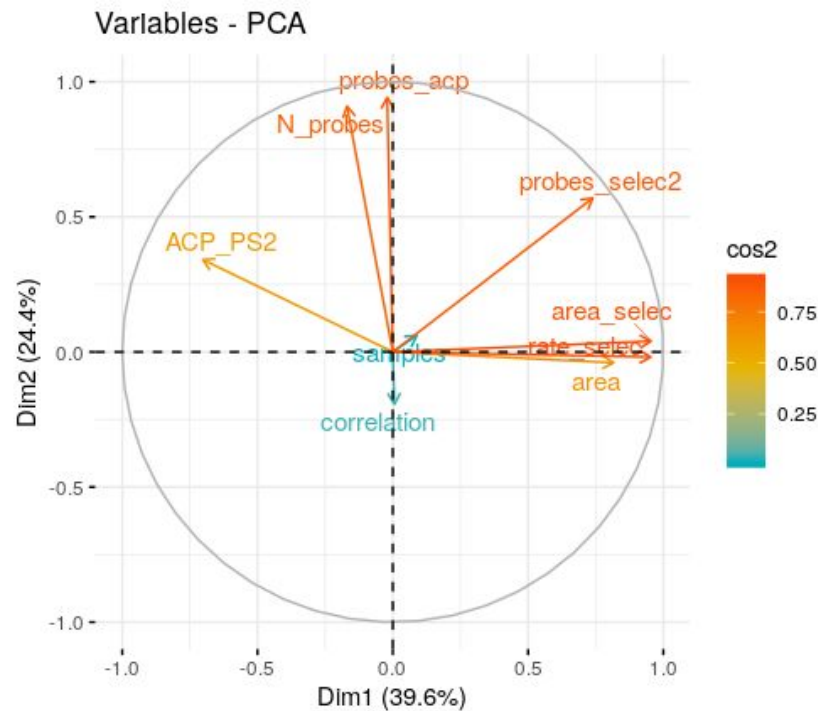
	samples	N_probes	probes_selec2	correlation	area	area_selec	rate_selec	probes_acp	ACP_PS2
WASH7P	376	3	1	0.0902350075	2.3662680	1.6532263	0.33333333	2	2.0000000
MIR6859-1	376	2	0	0.0373702891	6.3441836	0.0000000	0.00000000	0	NaN
MIR6723	264	7	4	NA	8.6209962	6.2028161	0.57142857	2	0.5000000
LINC00115	241	10	0	-0.0764616985	2.4340938	0.0000000	0.00000000	1	Inf
LINC01128	237	10	0	-0.1599160064	2.4480494	0.0000000	0.00000000	1	Inf
FAM41C	238	5	0	-0.2295182035	12.2815680	0.0000000	0.00000000	3	Inf
LOC100130417	241	17	4	0.1993838803	10.8198481	1.9023090	0.23529412	5	1.2500000
SAMD11	241	13	3	0.2798855888	4.7690677	2.1791903	0.23076923	7	2.3333333
NOC2L	241	15	2	-0.2287990138	4.4658509	0.9698697	0.13333333	5	2.5000000
KLHL17	242	17	3	-0.1248987045	5.2957223	1.4127521	0.17647059	5	1.6666667
PLEKHN1	241	14	4	-0.1303182573	5.7987749	2.3665715	0.28571429	8	2.0000000
PERM1	241	10	4	0.0429117764	12.8898698	4.3463551	0.40000000	4	1.0000000
HES4	240	19	4	0.0179958340	3.1940922	1.6533700	0.21052632	10	2.5000000
ISG15	240	10	3	-0.3608996829	8.3892624	2.3310619	0.30000000	5	1.6666667
AGRN	237	12	0	0.0179556499	1.3656830	0.0000000	0.00000000	2	Inf
RNF223	239	5	0	0.1786592114	11.4568697	0.0000000	0.00000000	3	Inf
C1orf159	240	18	4	-0.0549265322	6.7364885	2.2772629	0.22222222	7	1.7500000
LINC01342	240	9	1	-0.2041246491	9.2894774	1.0354526	0.11111111	4	4.0000000
MIR200B	240	21	3	-0.1754839586	11.9374467	2.0522150	0.14285714	9	3.0000000
MIR200A	240	20	1	-0.1889268895	11.9058717	1.0594175	0.05000000	9	9.0000000
MIR429	240	19	0	-0.2084607249	11.3207313	0.0000000	0.00000000	11	Inf

Dictionnaire des scores et des indicateurs du catalogue LUSC :

- samples : Le nombres d'individus par analyse transcriptomique.
- N_probes : Le nombres de sondes pour chaque gène.
- probes_selec2 : Le nombre de sondes sélectionnées après la correction aire_ecart type.
- correlation : corrélation entre la moyenne des méthylations des individus et la transcription.
- area : Ecart-type moyen de méthylation des sondes avant sélection
- area_selec : Ecart-type moyen de méthylation des sondes après sélection
- rate_selec : Le taux de sondes sélectionnées.
- probes_acp : Le nombre de sondes sélectionnées par l'ACP
- ACP_PS2 : probes_acp/probes_selec2

1. ACP sur les scores

Pour pouvoir appliquer des filtres efficaces sur le catalogue et discriminer les gènes qui peuvent présenter un intérêt, nous allons procéder à une ACP sur les scores. L'ACP sur les scores nous renseignera sur les scores corrélés, ainsi lorsque nous appliquerons nos filtres nous éviterons les informations redondantes.



Par exemple, nous éviterons d'utiliser comme filtre les scores `area_selec` et `area` en même temps car on voit qu'ils sont fortement corrélés. Le score `rate_selec` est le mieux représenté au vu de son `cos2`, on suppose donc qu'il aura un pouvoir discriminant plus fort que les autres scores.

V. Résultats (descriptions du catalogue)

On décrira ici le nombre de pattern trouvé, leur fréquence dans les différents types de cancer, l'association avec la survie, si cela correspond à une classe particulière de gènes, si l'association est corrélée ou anti corrélée.

1. Filtre

Les étapes que nous avons menées précédemment avaient pour unique but de produire des scores calibrés sur le pattern H1F0 car le pattern H1F0 a été décrit par Cristina Morales comme étant un gène d'intérêt.

Nous allons donc établir notre premier filtre en nous basant sur les caractéristiques de H1F0 :

```
> catalog["H1F0",]  
      samples N_probes probes_selec2 correlation      area area_selec rate_selec probes_acp ACP_PS2  
H1F0      238       13           5 -0.5689399 8.722222  3.832272  0.3846154         2      0.4
```

Le premier filtre se présente sous cette forme (vignette 03_filtre.Rmd) :

```
sub_catalog = catalog[catalog$probes_selec2 > 5 & catalog$correlation<(-0.55) & catalog$ACP_PS2>0.4 & catalog$ACP_PS2<1,]
```

Ce filtre nous permet de sélectionner 78 gènes parmi 20344, parmi ces gènes 84 % ont un pattern intéressant.

```
> rownames(sub_catalog)  
[1] "ARTN"      "FAAH"      "IL12RB2"   "GIPC2"     "SHC1"      "ZNF695"    "TRIM58"    "LOC100132215" "ITPRIPL1"  
[10] "ZNF662"    "ZNF502"    "ZNF717"    "NUDT16P1"  "SLC35G2"   "ARHGEF26"  "LXN"       "RNF212"      "EVC2"  
[19] "SH3TC1"    "UCHL1"     "MIR4458HG" "CD01"      "ALDH7A1"   "PCDH2"     "ZNF300"    "RANBP17"     "ZSCAN23"  
[28] "COL21A1"   "CGAS"      "TSPYL5"    "BRINP1"    "AKR1E2"    "SRGN"      "ACSL5"     "CAVIN3"      "ME3"  
[37] "PLEKHG6"   "PLBD1"     "PLBD1-AS1" "KRT7"      "KRT5"      "DGKA"      "EID3"      "CLDN10"      "CIDEB"  
[46] "MT1E"      "SDR42E1"   "ALOX12"    "FGF11"     "RAB34"     "SLFN12"    "HOXB4"     "CCDC68"      "ZNF69"  
[55] "EPHX3"     "ZFP82"     "ZNF382"    "ZNF285"    "ZNF415"    "ZNF677"    "ZNF542P"   "ZNF582"      "ZNF583"  
[64] "ZNF667"    "ZNF667-AS1" "ZFP28"     "ZNF671"    "ZNF418"    "ZSCAN1"    "TCF15"     "FRG1BP"     "ACTL10"  
[73] "CLIC6"     "PRAME"     "WBP2NL"    "FIRRE"     "PRKY"      "TTTY15"
```

Le but plus tard sera de faire un lien avec la survie. En effet nous essaierons de déterminer si le niveau de transcription ou de méthylation de ces gènes a une influence sur la survie des patients.

Le second filtre se présente sous cette forme (vignette 03_filtre_bis.Rmd) :

```
sub_catalog2 = catalog[catalog$probes_select > 5 & catalog$correlation<(-0.55) & catalog$rate_select > 0.38 & catalog$samples > 50,]
```

Nous avons utilisé le filtre rate_select ainsi que le nombre d'individus par étude (samples).

Ce filtre nous permet de sélectionner 109 gènes parmi 20344. Il sélectionne plus de gènes et le taux de pattern intéressant est de l'ordre de 75%. De plus les filtres partagent des similarités donc une partie des gènes sélectionnés sont similaires.

```
> rownames(sub_catalog2)
```

[1]	"ARTN"	"FAAH"	"IL12RB2"	"GIPC2"	"SHC1"	"PKP1"	"ZNF695"	"TRIM58"	"KRTCAP3"
[10]	"LOC339803"	"C2orf74"	"LOC100132215"	"ITPR1PL1"	"RARB"	"ZNF662"	"ZNF502"	"ZNF717"	"NUDT16P1"
[19]	"SLC35G2"	"ARHGEF26"	"LXN"	"RNF212"	"EVC2"	"SH3TC1"	"UCHL1"	"MIR4458HG"	"CD01"
[28]	"ALDH7A1"	"PCDHB2"	"ZNF300"	"RANBP17"	"FAM50B"	"HIST1H3G"	"ZSCAN23"	"HLA-DMA"	"COL21A1"
[37]	"CGAS"	"FAM221A"	"HOXA4"	"TSPYL5"	"GRHL2"	"NAPRT"	"BRINP1"	"TOR4A"	"AKR1E2"
[46]	"SRGN"	"ACSL5"	"IFITM1"	"CAVIN3"	"ME3"	"PLBD1"	"PLBD1-AS1"	"KRT7"	"KRT5"
[55]	"DGKA"	"EID3"	"HCAR1"	"GJB6"	"CLDN10"	"CIDEB"	"MT1E"	"SDR42E1"	"ALOX12"
[64]	"FGF11"	"SOX15"	"ALDH3A1"	"RAB34"	"SLFN12"	"HOXB2"	"HOXB4"	"CCDC68"	"ZNF69"
[73]	"ZNF844"	"EPHX3"	"BST2"	"ZFP82"	"ZNF382"	"ZNF829"	"ZNF285"	"ZNF229"	"ZNF701"
[82]	"ZNF415"	"ZNF347"	"ZNF677"	"ZNF542P"	"ZNF582"	"ZNF583"	"ZNF667"	"ZNF667-AS1"	"ZNF471"
[91]	"ZFP28"	"ZNF470"	"ZNF549"	"ZIK1"	"ZNF671"	"ZNF418"	"ZSCAN1"	"ZNF135"	"TCF15"
[100]	"FRG1BP"	"ACTL10"	"CLIC6"	"ZNF280B"	"PRAME"	"WBP2NL"	"SMC1B"	"RIBC2"	"ARMCX2"
[109]	"FIRRE"								

2. Analyse de la Survie

Nous avons utilisé la méthode de Kaplan-Meier pour analyser la survie des individus en fonction de la méthylation ou de la transcription de chaque gène qui ont été trouvés dans la partie précédente.

a. Méthode pour former les groupes

- Par transcription : Pour chaque étude du transcriptome de chaque gène nous déterminons la médiane des transcriptions. Les individus ayant une transcription du gène inférieure à la médiane font partie du groupe low_tr et les individus ayant une transcription du gène supérieure à la médiane font partie du groupe high_tr.

Nous menons ensuite une analyse de la survie avec le package "survival" et "survminer". En plus de cela nous établissons un test du log rank pour déterminer s'il y a une différence de survie entre les deux groupes.

Il teste l'hypothèse nulle d'égalité des probabilités de survie entre 2 études de survie :

Hypothèse nulle : "H0 = les probabilité de survie entre les 2 populations sont identiques"

Hypothèse alternative : "H1 = les probabilité de survie entre les 2 populations sont différentes".

- Par méthylation : Pour chaque étude du méthylome de chaque gène on sélectionne les sondes d'intérêt. Sur ces sondes d'intérêt nous établissons la moyenne des méthylations pour chaque individus. Nous déterminons ensuite la médiane des moyennes de méthylations. Un individu ayant une moyenne de méthylations inférieure à la médiane se retrouvera dans le groupe low_meth tandis qu'un individu ayant une moyenne de méthylations supérieure à la médiane se retrouve dans le groupe high_meth.

b. Résultat (vignette 05_survival_test)

Les test du log-rank appliqués aux analyses de survie se présente sous cette forme :

	By Methylation	By Transcription
EPHX3	0.001549871	0.2831416149
ALDH7A1	0.002646305	0.0086498091
GIPC2	0.004177543	0.1917241603
ZNF583	0.004522469	0.3490988295
RNF212	0.016347146	0.8172872130
CCDC68	0.019059729	0.0067254194
TRIM58	0.020189798	0.0019025153
CAVIN3	0.022623032	0.0124244153
SRGN	0.038840174	0.9599399005
ZNF677	0.053165545	0.1005533443
TSPYL5	0.065068128	0.7000385048
WBP2NL	0.073194992	0.4006243372
ZNF671	0.085382402	0.8335719539
ZSCAN1	0.086358102	0.0002397012
ZNF418	0.100954343	0.3167003847
ZNF542P	0.112592944	0.1383122549
NUDT16P1	0.118424332	0.0484129895
ZFP28	0.130515822	0.2718856322
ZFP82	0.139229331	0.4170768292
ZNF69	0.145493263	0.0622172720

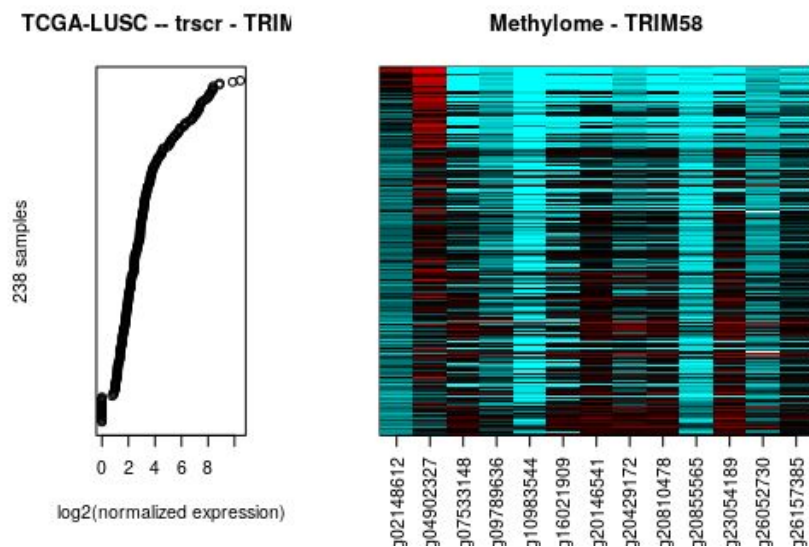
Nous déterminons la p-value de chaque test pour chaque type de classification. Nous allons garder les gènes qui ont une p-value inférieure à 5% pour au moins un type de classification. On rejettera donc pour cela l'hypothèse H0 .

Filtre 1	Filtre 2																																																																																																									
<table><tr><th></th><th>By Methylation</th><th>By Transcription</th></tr><tr><td>GIPC2</td><td>0.004177543</td><td>0.1917241603</td></tr><tr><td>TRIM58</td><td>0.020189798</td><td>0.0019025153</td></tr><tr><td>NUDT16P1</td><td>0.118424332</td><td>0.0484129895</td></tr><tr><td>SLC35G2</td><td>0.184352772</td><td>0.0387005134</td></tr><tr><td>RNF212</td><td>0.016347146</td><td>0.8172872130</td></tr><tr><td>ALDH7A1</td><td>0.002646305</td><td>0.0086498091</td></tr><tr><td>SRGN</td><td>0.038840174</td><td>0.9599399005</td></tr><tr><td>CAVIN3</td><td>0.022623032</td><td>0.0124244153</td></tr><tr><td>CLDN10</td><td>0.469781172</td><td>0.0132323231</td></tr><tr><td>CCDC68</td><td>0.019059729</td><td>0.0067254194</td></tr><tr><td>EPHX3</td><td>0.001549871</td><td>0.2831416149</td></tr><tr><td>ZNF415</td><td>0.552734087</td><td>0.0310626995</td></tr><tr><td>ZNF583</td><td>0.004522469</td><td>0.3490988295</td></tr><tr><td>ZSCAN1</td><td>0.086358102</td><td>0.0002397012</td></tr></table>		By Methylation	By Transcription	GIPC2	0.004177543	0.1917241603	TRIM58	0.020189798	0.0019025153	NUDT16P1	0.118424332	0.0484129895	SLC35G2	0.184352772	0.0387005134	RNF212	0.016347146	0.8172872130	ALDH7A1	0.002646305	0.0086498091	SRGN	0.038840174	0.9599399005	CAVIN3	0.022623032	0.0124244153	CLDN10	0.469781172	0.0132323231	CCDC68	0.019059729	0.0067254194	EPHX3	0.001549871	0.2831416149	ZNF415	0.552734087	0.0310626995	ZNF583	0.004522469	0.3490988295	ZSCAN1	0.086358102	0.0002397012	<table><tr><th></th><th>By Methylation</th><th>By Transcription</th></tr><tr><td>GIPC2</td><td>0.004177543</td><td>0.1917241603</td></tr><tr><td>TRIM58</td><td>0.020189798</td><td>0.0019025153</td></tr><tr><td>NUDT16P1</td><td>0.118424332</td><td>0.0484129895</td></tr><tr><td>SLC35G2</td><td>0.184352772</td><td>0.0387005134</td></tr><tr><td>RNF212</td><td>0.016347146</td><td>0.8172872130</td></tr><tr><td>ALDH7A1</td><td>0.002646305</td><td>0.0086498091</td></tr><tr><td>SRGN</td><td>0.038840174</td><td>0.9599399005</td></tr><tr><td>CAVIN3</td><td>0.022623032</td><td>0.0124244153</td></tr><tr><td>HCAR1</td><td>0.017903073</td><td>0.0341627769</td></tr><tr><td>CLDN10</td><td>0.469781172</td><td>0.0132323231</td></tr><tr><td>CCDC68</td><td>0.019059729</td><td>0.0067254194</td></tr><tr><td>ZNF844</td><td>0.022200798</td><td>0.1724001618</td></tr><tr><td>EPHX3</td><td>0.001549871</td><td>0.2831416149</td></tr><tr><td>ZNF415</td><td>0.552734087</td><td>0.0310626995</td></tr><tr><td>ZNF583</td><td>0.004522469</td><td>0.3490988295</td></tr><tr><td>ZSCAN1</td><td>0.086358102</td><td>0.0002397012</td></tr><tr><td>ZNF135</td><td>0.011380344</td><td>0.3304541613</td></tr><tr><td>SMC1B</td><td>0.007086642</td><td>0.0306415042</td></tr><tr><td>RIBC2</td><td>0.007086642</td><td>0.0658126171</td></tr></table>		By Methylation	By Transcription	GIPC2	0.004177543	0.1917241603	TRIM58	0.020189798	0.0019025153	NUDT16P1	0.118424332	0.0484129895	SLC35G2	0.184352772	0.0387005134	RNF212	0.016347146	0.8172872130	ALDH7A1	0.002646305	0.0086498091	SRGN	0.038840174	0.9599399005	CAVIN3	0.022623032	0.0124244153	HCAR1	0.017903073	0.0341627769	CLDN10	0.469781172	0.0132323231	CCDC68	0.019059729	0.0067254194	ZNF844	0.022200798	0.1724001618	EPHX3	0.001549871	0.2831416149	ZNF415	0.552734087	0.0310626995	ZNF583	0.004522469	0.3490988295	ZSCAN1	0.086358102	0.0002397012	ZNF135	0.011380344	0.3304541613	SMC1B	0.007086642	0.0306415042	RIBC2	0.007086642	0.0658126171
	By Methylation	By Transcription																																																																																																								
GIPC2	0.004177543	0.1917241603																																																																																																								
TRIM58	0.020189798	0.0019025153																																																																																																								
NUDT16P1	0.118424332	0.0484129895																																																																																																								
SLC35G2	0.184352772	0.0387005134																																																																																																								
RNF212	0.016347146	0.8172872130																																																																																																								
ALDH7A1	0.002646305	0.0086498091																																																																																																								
SRGN	0.038840174	0.9599399005																																																																																																								
CAVIN3	0.022623032	0.0124244153																																																																																																								
CLDN10	0.469781172	0.0132323231																																																																																																								
CCDC68	0.019059729	0.0067254194																																																																																																								
EPHX3	0.001549871	0.2831416149																																																																																																								
ZNF415	0.552734087	0.0310626995																																																																																																								
ZNF583	0.004522469	0.3490988295																																																																																																								
ZSCAN1	0.086358102	0.0002397012																																																																																																								
	By Methylation	By Transcription																																																																																																								
GIPC2	0.004177543	0.1917241603																																																																																																								
TRIM58	0.020189798	0.0019025153																																																																																																								
NUDT16P1	0.118424332	0.0484129895																																																																																																								
SLC35G2	0.184352772	0.0387005134																																																																																																								
RNF212	0.016347146	0.8172872130																																																																																																								
ALDH7A1	0.002646305	0.0086498091																																																																																																								
SRGN	0.038840174	0.9599399005																																																																																																								
CAVIN3	0.022623032	0.0124244153																																																																																																								
HCAR1	0.017903073	0.0341627769																																																																																																								
CLDN10	0.469781172	0.0132323231																																																																																																								
CCDC68	0.019059729	0.0067254194																																																																																																								
ZNF844	0.022200798	0.1724001618																																																																																																								
EPHX3	0.001549871	0.2831416149																																																																																																								
ZNF415	0.552734087	0.0310626995																																																																																																								
ZNF583	0.004522469	0.3490988295																																																																																																								
ZSCAN1	0.086358102	0.0002397012																																																																																																								
ZNF135	0.011380344	0.3304541613																																																																																																								
SMC1B	0.007086642	0.0306415042																																																																																																								
RIBC2	0.007086642	0.0658126171																																																																																																								
On a 14 gènes ayant un pattern intéressant et un lien avec la survie. Donc à peu près 18% .	Ici on a 19 gènes ayant un pattern intéressant et un lien avec la survie.																																																																																																									

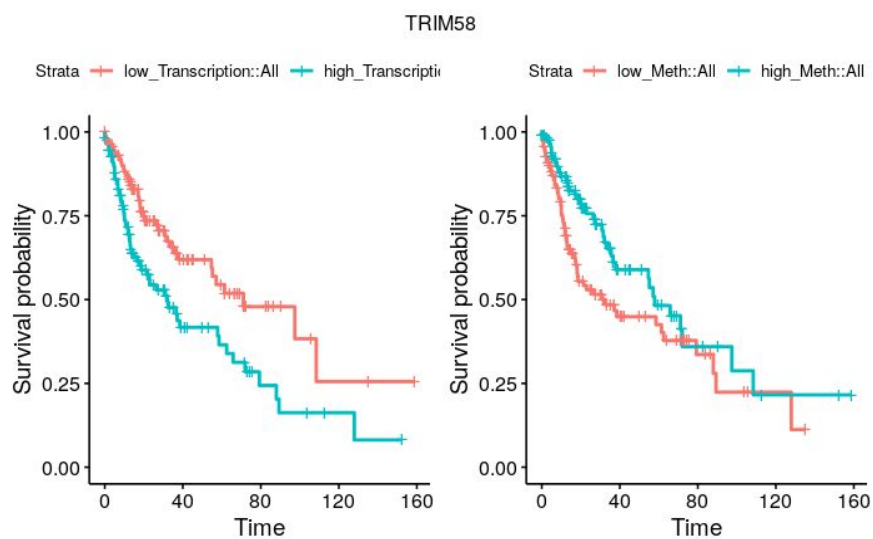
Parmi tous ces gènes, plusieurs ont déjà été identifiés par les chercheurs comme étant liés au développement de cancer. Ces gènes sont TRIM58, CAVIN3,

CLDN10, CCDC68 et SMC1B. **TRIM58** est particulièrement intéressant car il a été identifié comme étant directement lié au cancer **LUSC**. Il conforte donc en partie notre méthode de sélection.

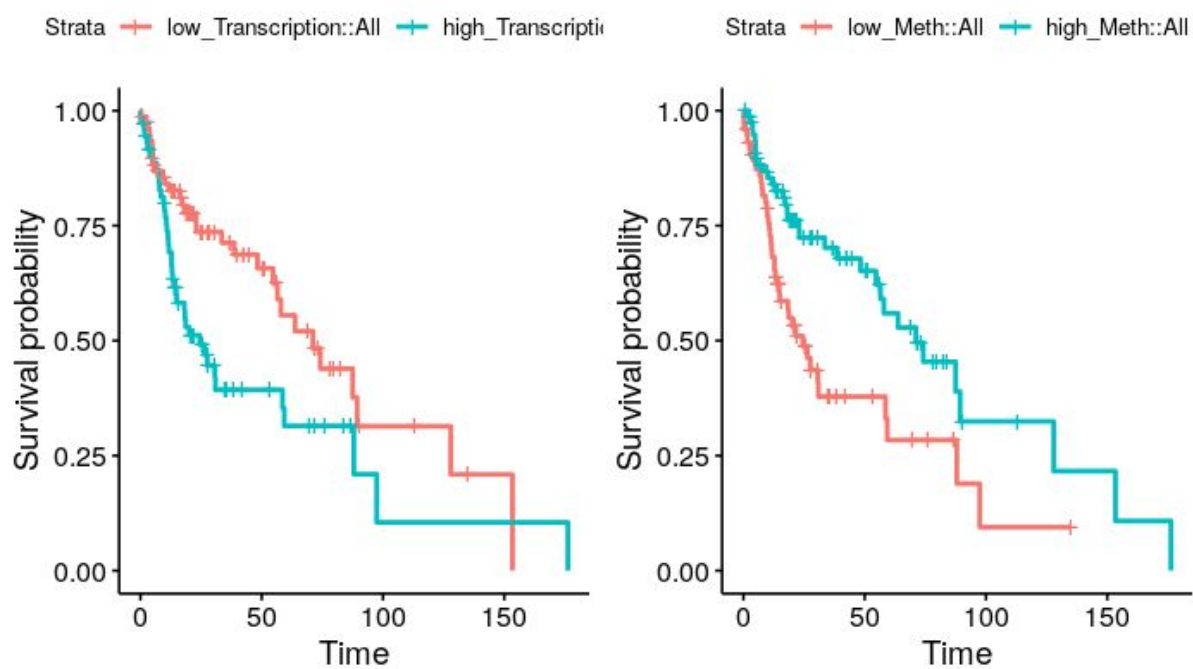
Pattern TRIM58



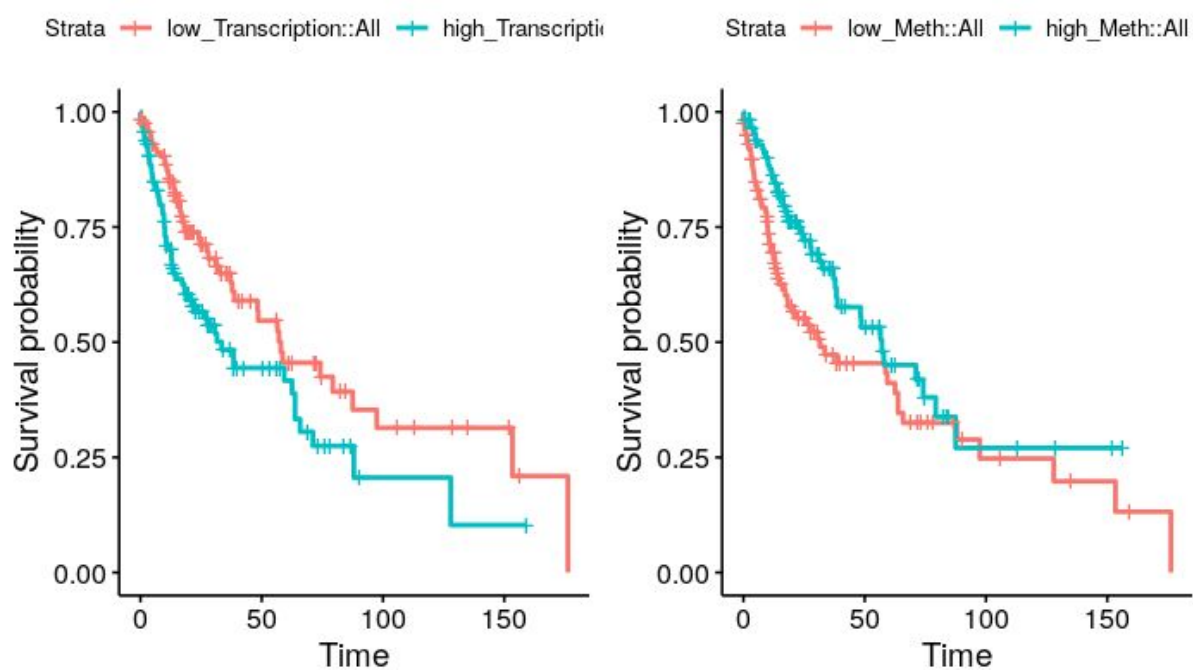
Quelques courbes de survie (06_survival_curves.Rmd)



ALDH7A1



CAVIN3



VI. Conclusion

L'analyse statistique que nous avons menée a mis en valeur plusieurs patterns intéressants. L'analyse de la survie a mis en relief des gènes qui seraient directement liés à la survie des patients. Cependant une analyse biologique est nécessaire pour étudier les implications biologiques de ces patterns. Cette méthode est généralisable à tous les cancers de la base de données TCGA. De plus la validation de nouveaux promoteurs de gène différentiellement méthylé nous permettrait d'améliorer la calibration de notre méthode.

Annexe

Pour reproduire les données de notre rapport il faut lancer les vignettes dans l'ordre suivant :

01_creation_results.Rmd : Cette vignette permet de nettoyer et de présenter les données brutes du TCGA pour pouvoir mener des études dessus. Elle nous permet de préparer les données omics de méthylation, de transcription ainsi que la liste des gènes. La vignette 1 utilise le concept de **mémoïsation** qui est une technique d'optimisation de code.

Son but est de diminuer le temps d'exécution d'un programme informatique en mémorisant les valeurs retournées par une fonction. La mémoïsation est une façon de diminuer le temps de calcul d'une fonction au prix d'une occupation mémoire plus importante.

Dans le cadre de notre analyse, nous disposons d'un grand nombre de gènes et de données omics d'où l'utilisation de la technique de mémoïsation.

Notons que sous R, toutes les données avec lesquelles on travaille sont systématiquement hébergées dans la mémoire vive de la machine.

La première vignette sert donc à charger et organiser les données en mémoire en implémentant une fonction mémoïsée ``mget_dmprocr_results (gene_symbol, tcga_project)``. Cette fonction lorsqu'elle est appelée à nouveau avec les mêmes paramètres, renvoie les valeurs stockées au lieu de réexécuter tout le code.

Ainsi pour un gène donné, les données omics associées sont accessibles avec la fonction mémoïsée créée et elles sont stockées dans ``tgca_project``.

Ceci a été fait, principalement, à l'aide de la fonction `memoise` implémentée dans le package `memoise` inclus dans R qui, à son tour, met en cache les résultats d'une fonction de sorte que lors de son appel, encore une fois avec les mêmes arguments, il retourne la valeur pré-compilée.

02_catalogue.Rmd : Nous chargeons ici les fonctions de scores que nous avons défini auparavant puis nous les appliquons à chaque gène de l'étude LUSC pour produire un catalogue des scores pour chaque gène.

03_filtre.Rmd : Le premier filtre

03_filtre_bis.Rmd : Le second filtre

04_results.Rmd : d'autres résultats mettant en avant d'autres types de filtres mais que nous n'avons pas présentés dans ce rapport.

05_survival_test.Rmd : Cette vignette crée la dataframe contenant la p-value des tests statistiques menés sur l'analyse de survie des individus

06_survival_curves.Rmd : Cette vignette présente les courbes de survie des gènes dont la p _ value obtenue avec la vignette précédente est inférieure à 5%

07_ACP.Rmd : produit les graphiques ACP de notre étude

graph_ecart_type.Rmd : produit les graphiques représentant les correction en analysant les écarts types des sondes.