Analyse d'expression différentielle pour un séquençage de l'ARN avec edgeR

-

12 décembre 2025

Table des matières

1 Résumé

L'outil bioinformatique **edgeR**, un package de R/Bioconductor, a été créé spécialement pour l'analyse de données de comptages de séquences et fragments omiques. Afin de démontrer son fonctionnement, nous avons testé **edgeR** pour identifier des gènes exprimés différentiellement entre des tissus carcinomes épidermoïdes buccaux et des tissus normaux provenant de trois patients (1). Après le traitement de nos données, incluant la mise à jour des annotations, le filtrage des gènes redondants et la normalisation TMM, des modèles linéaires généralisés binomiaux négatifs ont été appliqués. L'utilisation de ces modèles a rendu possible la comparaison des échantillons des tissus cancéreux et normaux, en tenant compte des différences initiales entre les patients.

L'analyse a révélé 1267 gènes présentant des différences significatives entre les deux groupes étudiés. Les gènes régulés à la hausse (321 gènes) dans les cellules tumorales semblent être liés au développement tissulaire, tandis que les gènes régulés à la baisse (946 gènes) semblent associés à la structure musculaire. Ces résultats démontrent que edgeR gère efficacement les données de comptages issues du séquençage de l'ARN et extrait de manière fiable des différences biologiquement significatives entre les tissus cancéreux et normaux.

2 Introduction

L'arrivée du séquençage à haut débit, communément connu sous le nom de Next Generation Sequencing, a révolutionné les études génomiques et de biologie moléculaires. Il permet le séquençage rapide de millions de fragments d'ADN ou d'ARN simultanément, offrant des avancées significatives dans les domaines de l'analyse génomique et de transcriptomique. Une application spécifique du NGS est le séquençage d'ARN (RNA-seq) à haut débit, utilisée pour le profilage de l'expression des gènes et des transcrits (2). Avec la possibilité d'avoir un si grand nombre de fragments et de segments à analyser, il faut trouver une manière simple et efficace afin d'offrir un sens à ces données. Un tel outil qui permet cette analyse est le package **edgeR**.

edgeR, dont l'acronyme signifie Empirical analysis of Digital Gene Expression in R, est un outil puissant qui analyse les données issues du NGS. Il a originalement été conçu par MD Robinson, DJ McCarthy et GK Smyth en 2008 (3). Il est un projet collaboratif dans la communauté Bioconductor où plusieurs autres chercheurs contribuent encore à son évolution et son bon fonctionnement (Yunshen Chen, Aaron Lun,...). Il se spécialise dans les analyses différentielles sur des données discrètes omiques qui produisent des comptes de fragments d'ARN ou d'ADN. Il est utilisé notamment pour la détection de gènes ou de traits qui ont changé de niveau d'abondance entre des types de cellules, des groupes expérimentaux et pour trouver la cause génomique de maladies (4).

edgeR a été le premier outil à utiliser des modèles linéaires généralisés binomiaux négatifs (BN) afin de modéliser les comptages de lecture dans la recherche du génome. En effet, puisque ces derniers ont souvent une variance beaucoup plus grande que la moyenne, il faut un paramètre de dispersion qui permet la relation *variance > moyenne*. La régression BN est donc un meilleur modèle statistique pour analyser des données de comptages qui représentent une surdispersion. De plus, **edgeR** tient compte de deux variations dans les expérimentations de séquençage de l'ARN. La première est la variation biologique entre les différents groupes ou cellules échantillonnés qui suit une loi de Gamma. La deuxième, la variation technique, incluant les erreurs de mesure et de séquençage, qui suit une loi de Poisson. Ensemble, les comptages de lecture suivent une loi de Gamma-Poisson, c'est-à-dire une loi binomiale négative (5).

Dans cette revue d'outil, les capacités de modélisation linéaires généralisés de **edgeR** seront démontrés avec une analyse différentielle d'expréssion de gènes entre des transcrits séquencés de cellules cancéreuses et de cellules normales issus du séquençage de l'ARN. L'utilisation du package ainsi que ses différentes fonctions et modèles statistiques pour la préparation de données et l'analyse différentielle seront présentés.

3 Méthodologie

3.1 Introduction

Pour démontrer les capacités de **edgeR** et de son utilisation de modèles linéaires généralisés pour l'analyse de données, deux types de tissus seront étudiés chez trois patients: les tissus carcinomes épidermoïdes buccaux et les tissus normaux qui leur correspondent. Ces données proviennent d'une expérience de séquençage d'ARN par paires entreprie par Tuch et al. en 2010.(1)
Le but de l'analyse est de trouver les gènes qui sont différentiellement exprimés entre les cellules cancéreuses et les cellules normales.

3.2 Téléchargement des données

Nous commonçons par charger les *librairies* pertinentes.

```
library(org.Hs.eg.db)
library(limma)
library(edgeR)
```

Les données sont téléchargées du **DOI** (*the Digital Object Identifier*). Avant de lire les données, nous avons supprimé les colonnes qui n'étaient pas pertinentes pour notre étude. Nous avons également renommé certaines colonnes (`idRefSeq` à `RefSeqID`, `nameOfGene` à `Symbol`, `numberOfExons` à `NbrOfExons`).

```
rawdata <- read.delim("TableS1.txt", check.names = FALSE, stringsAsFactors = FALSE)
```

RefSeqID	Symbol	NbrOfExons	8N	8T	33N	33T	51N	51T
NM_182502	TMPRSS11B	10	2592	3	7805	321	3372	9
NM_003280	TNNC1	6	1684	0	1787	7	4894	559
NM_152381	XIRP2	10	9915	15	10396	48	23309	7181

```
df <- summary(rawdata)
```

RefSeqID	Symbol	NbrOfExons	8N	8T	33N	33T	51N	51T
Length:15668	Length:15668	Min. : 1.00	Min. : 2.0	Min. : 0.0	Min. : 1	Min. : 0	Min. : 5	Min. : 0
Class :character	Class :character	1st Qu.: 6.00	1st Qu.: 119.0	1st Qu.: 88.0	1st Qu.: 143	1st Qu.: 171	1st Qu.: 317	1st Qu.: 223
Mode :character	Mode :character	Median : 10.00	Median : 256.0	Median : 219.0	Median : 291	Median : 379	Median : 679	Median : 494
NA	NA	Mean : 12.66	Mean : 771.6	Mean : 646.2	Mean : 1270	Mean : 1186	Mean : 2162	Mean : 1394
NA	NA	3rd Qu.: 16.00	3rd Qu.: 562.0	3rd Qu.: 503.0	3rd Qu.: 622	3rd Qu.: 828	3rd Qu.: 1479	3rd Qu.: 1100
NA	NA	Max. :312.00	Max. :393801.0	Max. :330105.0	Max. :581364	Max. :365430	Max. :1675945	Max. :633871

3.3 Création d'un objet *DGEList*

Nous stockerons nos données dans l'objet `DGEList` pour faciliter leur manipulation. Cet objet, conçu spécifiquement pour edgeR, peut être manipulé comme une liste. `DGEList` est une structure de données qui comprend toutes les composantes nécessaires dans un seul objet.

- une matrice `counts` contenant le nombre de chaque gène dans chaque échantillon,
- un data-frame `samples` contenant des informations sur les échantillons et les librairies,
- et un data-frame `genes` contenant des annotations sur les gènes (6).

```
y <- DGEList(counts = rawdata[,4:9], genes = rawdata[,1:3])
```

3.4 Annotation

L'étude (1) a été réalisée en 2010 ; par conséquent, certains des `RefSeqIDs` ne sont plus utilisés, car il y a eu des changements dans la configuration du génome de référence humain (*GRH*) depuis (4). Nous filtrons les `RefSeq IDs` qui ne sont plus disponibles dans l'annotation actuelle du **NCBI** (fournie par le package du *GRH* `org.Hs.eg.db`). Notez que pour une étude plus poussée, il est conseillé de retrouver l'ancien *GRH* afin de ne pas perdre de données importantes.

```
idfound <- y$genes$RefSeqID %in% mappedRkeys(org.Hs.egREFSEQ)
y <- y[idfound,]
dim(y) # 15534 sur 15668 RefSeqIDs sont retenus
```

```
## [1] 15534     6
```

Entrez Gene IDs sont les identifiants numériques uniques attribués aux gènes par le NCBI (7). Pour chaque gène, `Entrez Gene` donne des détails tels que le nom du gène et sa localisation chromosomique. Nous ajoutons les *Entrez Gene IDs* à l'annotation.

```
egREFSEQ <- toTable(org.Hs.egREFSEQ)
df <- head(egREFSEQ, 3)
```

gene_id	accession
1	NM_130786
1	NP_570602
2	NM_000014

```
m <- match(y$genes$RefSeqID, egREFSEQ$accession)
y$genes$EntrezGene <- egREFSEQ$gene_id[m]
```

Ensuite, nous mettons à jour `symbol` en utilisant les *Entrez Gene IDs*.

```
egSYMBOL <- toTable(org.Hs.egSYMBOL)
df <- head(egSYMBOL, 3)
```

gene_id	symbol
1	A1BG
2	A2M
3	A2MP1

```
m <- match(y$genes$EntrezGene, egSYMBOL$gene_id)
y$genes$Symbol <- egSYMBOL$symbol[m]
df <- head(y$genes, 5)
```

RefSeqID	Symbol	NbrOfExons	EntrezGene
NM_182502	TMPRSS11B	10	132724
NM_003280	TNNC1	6	7134
NM_152381	XIRP2	10	129446
NM_022438	MAL	3	4118
NM_001100112	MYH2	40	4620

3.5 Filtration

Plusieurs transcrits *RefSeq* correspondent au même gène, car ils peuvent représenter différents variants de transcrits ou des isoformes de protéines. Nous ne gardons qu'un seul transcrit *RefSeq* (celui avec le `count` le plus élevé) pour chaque gène. Les additionner ensemble augmenterait les biais de l'analyse puisqu'il donnerait plus d'importance au gène, alors que ce n'est pas le cas (8).

```
o <- order(rowSums(y$counts))
y <- y[o,]
d <- duplicated(y$genes$Symbol)
y <- y[!d,]
nrow(y)
```

```
## [1] 10510
```

Normalement, il faudrait filtrer les gènes faiblement exprimés (enlever ceux qui ont moins de 1 cpm par exemple). Par contre, tous nos transcrits se retrouvent au moins 50 fois dans au moins un des 6 cas, donc cela n'est pas nécessaire.

Ensuite, nous recalculons la taille du `library`.

```
y$samples$lib.size <- colSums(y$counts)
```

Et nous utilisons les *Entrez Gene IDs* comme les noms de lignes.

```
rownames(y$counts) <- rownames(y$genes) <- y$genes$EntrezGene
y$genes$EntrezGene <- NULL
```

3.6 Normalisation

La normalisation permet d'éliminer les effets techniques présents dans les données afin de minimiser l'impact des biais techniques sur les résultats (9). De même, nous pouvons nous assurer que les différences entre les échantillons sont dû à la variation biologique. Nous utilisons la fonction `normLibSizes()` qui utilise la normalisation TMM (*trimmed mean of M-values*) pour tenir compte des différences de composition entre les librairies. Cette méthode de normalisation permet la comparaison entre les échantillons qui ont des couvertures différentes. Elle modifie la taille des librairies, et non les données de comptages. Pour se faire, elle émet l'hypothèse *la plupart des gènes ne sont pas exprimés différentiellement*. Ensuite, elle choisit un échantillon de référence, et calcule le facteur TMM pour chacun des autres échantillons. Un facteur $TMM < 1$ indique qu'un petit nombre de gènes hautement exprimés contribuent largement à la taille de la librairie, donnant moins de valeur aux autres gènes du même groupe. Pour contrer cette partialité, elle diminuera la taille de la librairie. Puis un facteur $TMM > 1$ augmente la taille (4).

```
y <- normLibSizes(y)
samples<-y$samples
```

	group	lib.size	norm.factors
8N	1	7397598	1.1542497
8T	1	7124442	1.0619357
33N	1	15260793	0.6556112
33T	1	13651143	0.9484143
51N	1	19318441	1.0892960
51T	1	14382783	1.2045134

4 Résultats et discussions

4.1 Exploration des données

Nous voulons voir s'il existe des valeurs aberrantes ou d'autres corrélations dans notre échantillon. La fonction `plotMDS` produit un plot dans lequel les distances entre les échantillons représentent approximativement les différences d'expression (10).
Sur le plot (**figure 1**), la dimension 1 sépare les échantillons tumoraux des échantillons normaux et la dimension 2 correspond au numéro du patient. On peut observer que les échantillons non tumoraux sont proches les uns des autres, tandis que les échantillons tumoraux sont plus espacés. Donc, les échantillons tumoraux sont plus hétérogènes que les échantillons normaux. Même si les échantillons ne sont pas nombreux, il est évident que les tissus tumoraux et normaux ne se mélangent pas. Il ne semble pas y avoir de donnée aberrante.

```
par(mar = c(4, 4, 0, 1))
plotMDS(y, cex = 0.7, cex.lab = 0.7, cex.axis = 0.7)
```

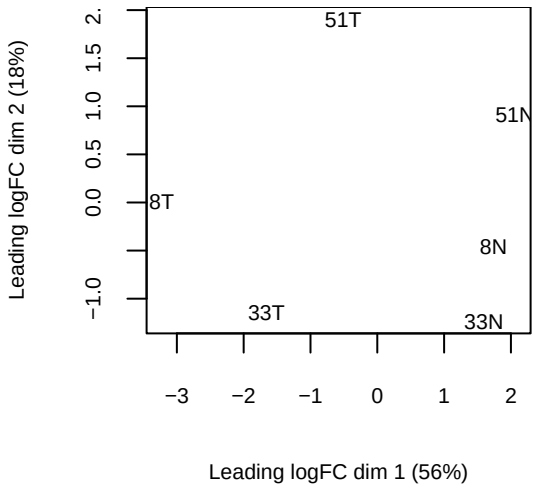

Figure 1: plotMDS(y)

4.2 Matrice de design

Ensuite, nous voulons tester l'expression différentielle entre les tissus tumoraux et les tissus normaux chez les patients. Ici, nous avons un modèle linéaire additif. Le facteur de blocage est le `Patient`, comme ça nous pouvons isoler les effets dûs au facteur `Tissue`. Les différences sont dûs aux tissus, et non aux patients. Pour se faire, il faut d'abord créer une matrice de design selon le design expérimental.

```
Patient <- factor(c(8, 8, 33, 33, 51, 51))
Tissue <- factor(c("N", "T", "N", "T", "N", "T"))
dfdm<-data.frame(Sample=colnames(y),Patient,Tissue)
```

Sample	Patient	Tissue
8N	8	N
8T	8	T
33N	33	N
33T	33	T
51N	51	N
51T	51	T

```
design <- model.matrix(~Patient+Tissue)
rownames(design) <- colnames(y)
```

	(Intercept)	Patient33	Patient51	TissueT
8N	1	0	0	0
8T	1	0	0	1
33N	1	1	0	0
33T	1	1	0	1
51N	1	0	1	0
51T	1	0	1	1

4.3 Dispersion BN des données

Pour comprendre la variabilité biologique, nous estimons la dispersion binomiale négative des données.
edgeR utilise la méthode qCML (*quantile-adjusted conditional maximum likelihood)* pour estimer les dispersions dans les expériences à un seul facteur. La fonction `estimateDisp(y)` calcule en une seule étape les dispersions communes et dispersions tagwise, et fournit des estimations de dispersion fiables, notamment pour les études RNA-seq sur des petits échantillons.

```
y <- estimateDisp(y, design, robust = TRUE)
y$common.dispersion
```

```
## [1] 0.1613756
```

La racine carrée de la dispersion commune, `0.1613`, nous donne le BCV (coefficient de variation biologique), `0,402`. Un BCV entre 0,2 et 0,4 indique les gènes hautement exprimés différentiellement (11).

```
sqrt(y$common.dispersion) # BCV
```

```
## [1] 0.4017158
```

Les estimations de dispersion peuvent être visualisées sur un BCV plot. Le plot montre, sous la ligne rouge, que les gènes hautement exprimés différentiellement ont un petit BCV.

```
par(mar = c(4, 4, 0, 1))
plotBCV(y)
```

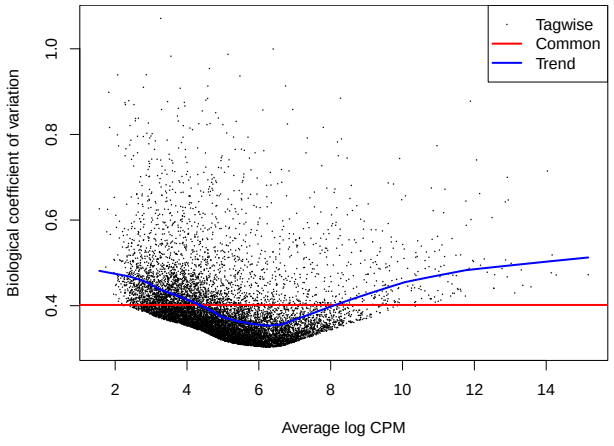

Figure 2: plotBCV(y)

4.4 Expression différentielle des gènes

Nous utilison la fonction `glmFit` pour adapter un modèle généralisé linéaire binomial négatif.

```
fit <- glmFit(y, design)
```

Ensuite, la fonction `glmLRT` est utilisée pour effectuer des tests de rapport de vraisemblance pour identifier les gènes significatifs. Il fait un test pour l'effet *tissus tumoreux vs normaux*.

```
lrt <- glmLRT(fit)
```

Nous utilisons la fonction `topTags` pour voir les résultats les plus significatifs, ceux dont l'effet testé précedemment est plus grand. Un haut `LR` (*Likelihood Ratio*) et une petite valeur-p indique que l'hypothèse nulle (aucun gène DE entre les deux types de tissus) peut être rejetée. Un petit `FDR` (*False Discovery Rate*) indique qu'il y a de très petites chances d'obtenir de faux positifs. En effet, tous les gènes qui se retrouvent dans le tableau `topTags` ont un `FDR` pratiquement nul. Donc, notre rejet de l'hypothèse nulle n'est pas un faux positif [(4).

```
tt <- topTags(lrt, n = 10)$table
```

	RefSeqID	Symbol	NbrOfExons	logFC	logCPM	LR	PValue	FDR
5737	NM_000959	PTGFR	3	-5.20	4.82	101.0	7.94e-24	8.35e-20
5744	NM_198966	PTHLH	4	3.88	5.82	84.7	3.44e-20	1.81e-16
10351	NM_007168	ABCA8	38	-4.00	5.02	78.7	7.13e-19	1.98e-15
1288	NM_001847	COL4A6	45	3.71	5.71	78.6	7.53e-19	1.98e-15
5837	NM_005609	PYGM	20	-5.49	6.08	75.2	4.34e-18	9.12e-15
487	NM_173201	ATP2A1	22	-4.62	6.04	74.0	7.73e-18	1.35e-14
27179	NM_014440	IL36A	4	-6.18	5.49	72.5	1.69e-17	2.53e-14
196374	NM_173352	KRT78	9	-4.26	7.70	69.4	8.11e-17	1.00e-13
6387	NM_199168	CXCL12	3	-3.72	5.86	69.3	8.60e-17	1.00e-13
83699	NM_031469	SH3BGRL2	4	-3.95	5.62	68.6	1.24e-16	1.30e-13

Ici nous avons une vision plus claire des `CPM` des gènes exprimés différentiellement entre les deux types de tissus chez les trois patients. Il est évident qu'il y a des changements consistants entre les tissus tumoraux et normaux.

```
o <- order(lrt$table$PValue)
cpm<-cpm(y)[o[1:10],]
```

	8N	8T	33N	33T	51N	51T
5737	53.286956	0.9252284	28.28544	0.926860	82.448258	2.5975143
5744	5.387253	78.1157149	10.59455	133.854037	5.940076	104.0737391
10351	56.449039	3.3043873	41.77849	2.239912	83.636273	6.3494794
1288	11.945647	137.0659837	6.19681	96.470682	4.514458	57.2607594
5837	163.842749	2.9078608	126.63482	1.235813	103.167244	5.9454216
487	114.537676	3.3043873	155.12015	4.016394	107.634181	9.2933289
27179	42.980907	1.3217549	182.30616	3.475725	38.111529	0.0577225
196374	399.125153	21.9411314	615.48318	50.436634	153.206446	4.7332483
6387	63.827233	3.0400363	68.46476	6.179067	188.514259	17.9517099
83699	103.177600	5.4191951	124.03615	5.715637	50.894573	5.6568089

La fonction `decideTests()` permet de voir le nombre total de gènes différentiellement exprimés avec un taux de `FDR` à 5%. En effet, environ 10% des gènes entre les deux types de tissus sont DE.

```
t1 <-summary(decideTests(lrt))
```

	TissueT
Down	938
NotSig	9255
Up	317

Le log fold change mesure le changement de l'expression des gènes entre deux conditions (tumeur vs normal). Il est le ratio logarithmique en base 2 du taux d'expression d'un gène donné (condition tumorale/condition normale). Par exemple, Un `LFC` de 1 indique une augmentation de l'expression de 2 fois alors qu'un `LFC` de -1 indique une diminution de l'expression de moitié. Ici, les gènes qui sont déterminés par leur `CPM` ayant un `LFC` entre -1 et 1 ne présentent pas assez de changements pour être significatifs. Ils sont représentés par des points noirs. Les gènes ayant un `LFC` > 1 sont significatifs et sont régulés à la hausse (points rouges). Les gènes ayant un `LFC` < 1 sont significatifs et sont régulés à la baisse (points bleus) (4).

```
par(mar = c(4, 4, 1.5, 1))
plotMD(lrt, cex = 0.7, cex.lab = 0.7, cex.axis = 0.7, cex.main = 0.8)
abline(h=c(-1, 1), col="blue")
```

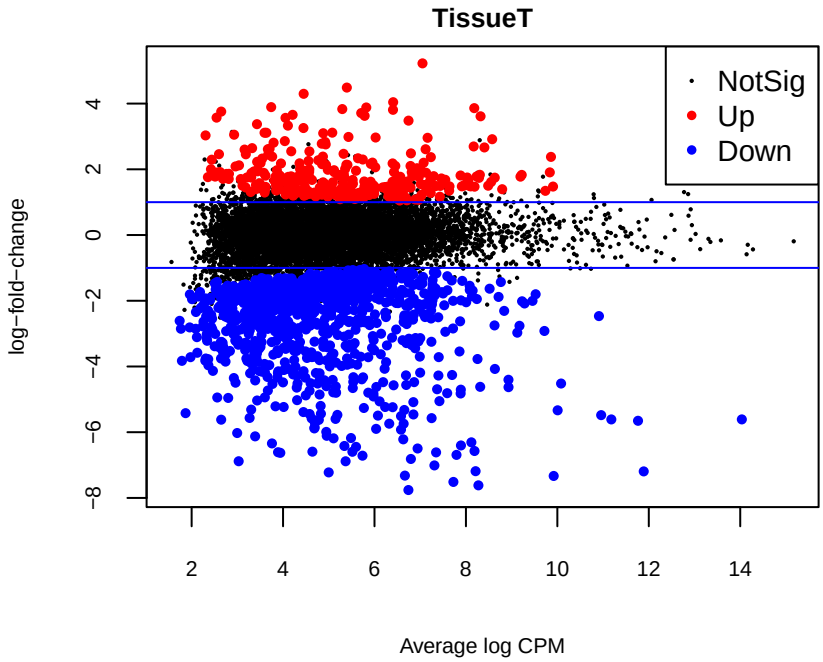

Figure 3: plotMD(lrt)

4.5 Analyse d'ontologie des gènes

Nous réalisons une analyse d'ontologie génique axée sur l'ontologie des processus biologiques (BP). Les gènes regulé à la hausse dans les tumeurs ont tendance à être associés au développement tissulaire.

```
go <- goana(lrt)
upreg <- topGO(go, ont="BP", sort="Up", n=10, truncate=30)
```

	Term	Ont	N	Up	Down	P.Up	P.Down
GO:0009888	tissue development	BP	1255	83	193	1.328042e-12	1.445920e-15
GO:0022008	neurogenesis	BP	1009	69	115	4.063003e-11	2.900881e-03
GO:0042127	regulation of cell populati...	BP	1346	83	164	5.423760e-11	9.652552e-06
GO:0008283	cell population proliferati...	BP	1921	105	218	9.108440e-11	3.584045e-05
GO:0007155	cell adhesion	BP	924	64	161	1.445713e-10	5.310000e-18
GO:0060429	epithelium development	BP	755	55	93	5.878335e-10	7.019916e-04
GO:0007399	nervous system development	BP	1456	84	156	1.172757e-09	6.523234e-03
GO:0048513	animal organ development	BP	1808	97	268	1.930662e-09	7.000000e-20
GO:0007275	multicellular organism deve...	BP	2740	129	324	5.919168e-09	1.021135e-09
GO:0030154	cell differentiation	BP	2485	120	311	6.500597e-09	2.529899e-12

Les gènes regulé à la basse dans les tumeurs ont tendance à être associés au développement musculaire.

```
go <- goana(lrt)
downreg <- topGO(go, ont="BP", sort="Down", n=10, truncate=30)
```

	Term	Ont	N	Up	Down	P.Up	P.Down
GO:0003012	muscle system process	BP	255	6	95	7.863968e-01	2.222900e-36
GO:0006936	muscle contraction	BP	196	5	79	7.094527e-01	3.817275e-33
GO:0003008	system process	BP	884	43	185	1.143903e-03	8.127874e-31
GO:0055001	muscle cell development	BP	141	2	61	9.295484e-01	8.154249e-28
GO:0042692	muscle cell differentiation	BP	278	6	83	8.491404e-01	3.348482e-24
GO:0061061	muscle structure developmen...	BP	461	11	111	8.278790e-01	1.970052e-23
GO:0051146	striated muscle cell differ...	BP	204	5	68	7.417525e-01	5.398719e-23
GO:0032501	multicellular organismal pr...	BP	3937	158	485	3.408172e-06	9.794296e-21
GO:0030239	myofibril assembly	BP	55	0	32	1.000000e+00	4.015039e-20
GO:0048513	animal organ development	BP	1808	97	268	1.930662e-09	7.158655e-20

5 Conclusion

Pour conclure, nous avons démontré que l'outil **edgeR** est important pour l'analyse de données omiques issus du séquençage à haut débit, surtout ceux issus du séquençage de l'ARN. L'expression différentielle entre les tissus carcinomes épidermoïdes buccaux et les tissus normaux a été établi, ainsi qu'une analyse d'ontologie des gènes. Bien que notre analyse soit basé sur uniquement un facteur, l'approche de modèle linéaire généralisé de l'outil permet des analyses beaucoup plus complexes, comme les expérimentations à facteurs multiples et les analyses de series temporelles par

exemple. De plus, **edgeR** a été le premier à utiliser la loi binomiale négative pour représenter les données de comptages, ce qui lui permet de faire des bonnes analyses, autant pour les petits jeux de données comme le nôtre que les gros (5). Facile et simple d'approche, cet outil nous a permit de mieux comprendre les différences entre deux tissus et d'en tirer des conclusions significatives. **edgeR** est et continuera à être un outil indispensable dans la recherche génomique et de biologie moléculaire.

6 Références

1. Tuch BB, Laborde RR, Xu X, Gu J, Chung CB, Monighetti CK, et al. Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations [Internet]. 2010. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009317#abstract0
2. Ian C. Nova PhD. RNA-seq (RNA sequencing) [Internet]. 2025. Available from: https://www.genome.gov/genetics-glossary/RNA-seq
3. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. [Internet]. 2009. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC2796818/
4. Chen Y, McCarthy D, Baldoni P, Ritchie M, Robinson M, Smyth G. edgeR: Differential analysis of sequence read count data. User's guide [Internet]. 2025. Available from: https://www.bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf
5. Chen Y, Chen L, Lun ATL, Baldoni PL, Smyth GK. edgeR v4: Powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets [Internet]. 2025. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11754124/
6. DGEList: DGEList constructor [Internet]. 2025. Available from: https://www.rdocumentation.org/packages/edgeR/versions/3.14.0/topics/DGEList
7. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: Gene-centered information at NCBI [Internet]. 2007. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC1761442/
8. Goldfarb T PS Kodali VK. NCBI RefSeq: Reference sequence standards through 25 years of curation and annotation [Internet]. 2025. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11701664/
9. Singh V, Kirtipal N, Song B, Lee S. Normalization of RNA-seq data using adaptive trimmed mean with multi-reference [Internet]. 2024. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11107385/
10. plotMDS.DGEList: Multidimensional scaling plot of distances between digital gene expression profiles [Internet]. 2025. Available from: https://www.rdocumentation.org/packages/edgeR/versions/3.14.0/topics/plotMDS.DGEList
11. Marco JDG, Fernández-Calle P, Ricós C. Models to estimate biological variation components and interpretation of serial results: Strengths and limitations [Internet]. 2020. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10270238/