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

-

12 décembre 2025

Table des matières

1 Résumé

Le package edgeR a été utilisé pour identifier des gènes exprimés différemment sur les tissus carcinomes épidermoïdes buccaux et les 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 de tissu cancéreux et normal, en tenant compte des différences initiales entre les patients.

L'analyse a révélé 1267 gènes présentant des différences significatives, les gènes régulés à la hausse (321 gènes) dans les tumeurs étant liés au développement tissulaire, tandis que les gènes régulés à la baisse (946 gènes) sont associés à la structure musculaire. Ces résultats démontrent que edgeR gère efficacement les données de *count RNA-seq* 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. 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. 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.

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.

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. (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 *libraries* 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 : 646.1	Mean : 1270	Mean : 1187	Mean : 2162	Mean : 1393
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 un structure de données qui comprend toutes les composantes nécessaires (une matrice *counts* contenant les integer *counts*, un data-frame *samples* contenant des informations sur les échantillons et les librairies, et un data-frame *genes* contenant des annotations sur les gènes) dans un seul objet (2).

```
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. Nous filtrons les *RefSeq IDs* qui ne sont plus disponibles dans l'annotation actuelle du **NCBI** (fournie par le package `org.Hs.eg.db`).

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

```
## [1] 15533     6
```

Entrez Gene IDs sont les identifiants numériques uniques attribués aux gènes par le NCBI (3). 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
9	NAT1

```
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. Nous ne gardons qu'un seul transcrit *RefSeq* (celui avec le *count* le plus élevé) pour chaque gène.

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

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

Tous les transcrits se retrouvent au moins 50 fois dans au moins un des 6 cas, donc ce n'est pas nécessaire de filtrer les gènes qui sont faiblement exprimés.

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 (4,5). Nous utilisons la fonction `normLibSizes()` qui utilise la normalisation *TMM* (*trimmed mean of M-values*) pour tenir compte des différences de composition entre les *libraries*.

```
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 (6).
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.

```
par(mar = c(4, 4, 0.5, 1))
plotMDS(y)
```

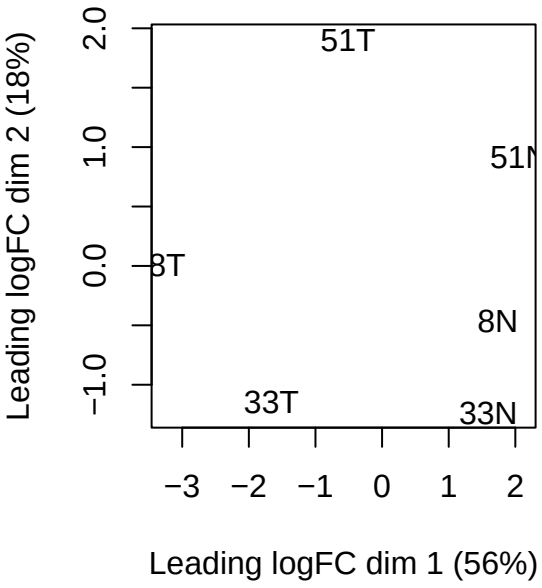

Figure 1: plotMDS(y)

4.2 Matrice de design

Ensuite, nous voulons tester l'expression différentielle entre les tissus tumoraux et les tissus normaux chez les patients.

```
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 binôme négatif 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 de petits échantillons.

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

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

La racine carrée du disperssino 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 (7).

```
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)
```

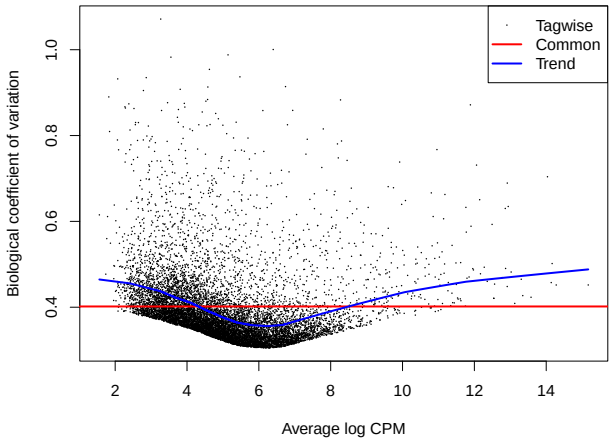

Figure 2: plotBCV(y)

4.4 Expression differential des gènes

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

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

Ensuite, la fonction `glmLRT` est utilisée pour effectuer des tests de rapport de vraisemblance pour identifier les gènes significatifs.

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

Nous utilisons la fonction `topTags` pour voir les résultats.

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

	RefSeqID	Symbol	NbrOfExons	logFC	logCPM	LR	PValue	FDR
5737	NM_000959	PTGFR	3	-5.20	4.82	100.0	1.21e-23	1.27e-19
5744	NM_198966	PTHLH	4	3.88	5.82	84.3	4.26e-20	2.24e-16
1288	NM_001847	COL4A6	45	3.71	5.71	78.3	9.00e-19	3.15e-15
10351	NM_007168	ABCA8	38	-4.00	5.02	77.5	1.31e-18	3.43e-15
5837	NM_005609	PYGM	20	-5.50	6.08	74.7	5.37e-18	1.13e-14
487	NM_173201	ATP2A1	22	-4.62	6.04	73.6	9.66e-18	1.69e-14
27179	NM_014440	IL36A	4	-6.18	5.49	72.2	1.98e-17	2.97e-14
196374	NM_173352	KRT78	9	-4.26	7.70	70.8	3.86e-17	5.07e-14
6387	NM_199168	CXCL12	3	-3.72	5.86	68.9	1.03e-16	1.20e-13
83699	NM_031469	SH3BGRL2	4	-3.95	5.62	68.4	1.36e-16	1.43e-13

```
colnames(design)
```

```
## [1] "(Intercept)" "Patient33"   "Patient51"   "TissueT"
```

```
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
1288	11.945647	137.0659837	6.19681	96.470682	4.514458	57.2607594
10351	56.449039	3.3043873	41.77849	2.239912	83.636273	6.3494794
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

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

	TissueT
Down	946
NotSig	9243
Up	321

```
par(mar = c(4, 4, 1.5, 1))
plotMD(lrt)
abline(h=c(-1, 1), col="blue")
```

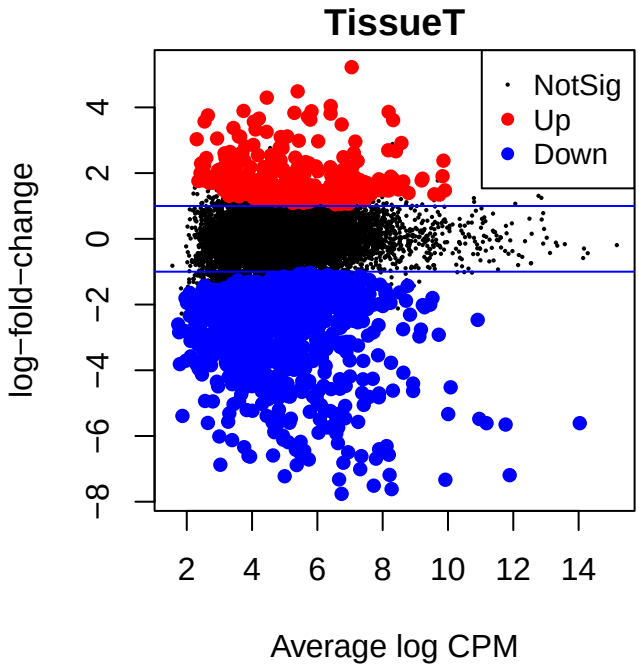

Figure 3: plotMD(lrt)

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:0022008	neurogenesis	BP	1083	74	120	1.277061e-11	7.917927e-03
GO:0009888	tissue development	BP	1294	82	199	3.631218e-11	1.150780e-15
GO:0007399	nervous system development	BP	1556	92	162	7.160014e-11	2.115949e-02
GO:0007155	cell adhesion	BP	945	63	160	1.687833e-09	2.547900e-16
GO:0048513	animal organ development	BP	1850	99	267	2.914578e-09	1.360000e-17
GO:0048731	system development	BP	2433	120	309	4.077632e-09	1.454654e-12
GO:0060429	epithelium development	BP	771	54	96	5.630044e-09	5.315945e-04
GO:0007275	multicellular organism deve...	BP	2857	134	336	7.638940e-09	2.317471e-09
GO:0048699	generation of neurons	BP	913	60	103	7.739626e-09	8.254978e-03
GO:0030154	cell differentiation	BP	2603	125	324	8.483777e-09	4.414787e-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	278	6	103	8.572576e-01	7.700000e-39
GO:0006936	muscle contraction	BP	208	5	84	7.664698e-01	5.304960e-35
GO:0003008	system process	BP	963	44	198	4.263559e-03	1.740569e-31
GO:0055001	muscle cell development	BP	152	3	67	8.480195e-01	7.743519e-31
GO:0055002	striated muscle cell develo...	BP	129	3	59	7.592912e-01	2.467449e-28
GO:0042692	muscle cell differentiation	BP	297	7	89	8.077281e-01	9.232699e-26
GO:0061061	muscle structure developmen...	BP	485	12	119	8.128848e-01	1.505117e-25
GO:0051146	striated muscle cell differ...	BP	218	6	73	6.595613e-01	1.581137e-24
GO:0030239	myofibril assembly	BP	60	0	36	1.000000e+00	5.292179e-23
GO:0032501	multicellular organismal pr...	BP	3975	163	497	1.102136e-06	3.932672e-22

5 Conclusion

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. DGEList: DGEList constructor [Internet]. 2025. Available from: https://www.rdocumentation.org/packages/edgeR/versions/3.14.0/topics/DGEList

3. 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/

4. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data [Internet]. 2010. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC2864565/

5. 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/#:~:text=Abstract,characteristic%20curve%20and%20differential%20expression.

6. 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

7. Gu Q. SOME KEY FACTORS FOR NUMBER OF SIGNIFICANT DE GENES [Internet]. 2015. Available from: https://bioinformatics.cvr.ac.uk/some-key-factors-for-number-of-significant-de-genes/#:~:text=An%20important%20factor%20that%20influences,conditions%20than%20within%2C%20as%20expected.

8. 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

9. Ian C. Nova PhD. RNA-seq (RNA sequencing) [Internet]. 2025. Available from: https://www.genome.gov/genetics-glossary/RNA-seq

10. 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/

11. Dunning M, Pereira B, Rueda O, Santiago ID, Samarajiwa S. Differential expression analysis using edgeR [Internet]. 2015. Available from: https://bioinformatics-core-shared-training.github.io/cruk-bioinf-sschool/Day3/Supplementary-RNAseq-practical.pdf