國立臺灣大學公共衛生學院流行病學與預防醫學研究所

碩士論文

Graduate Institute of Epidemiology and Preventive Medicine
College of Public Health
National Taiwan University

Master Thesis

# 以 RNA 序列資料偵測差異表現基因的統計方法比較

# Comparisons of Statistical Methods for Detecting Differentially Expressed Genes with RNA-seq Data

呂泓廷

Hung-Ting Lu

指導教授：林菀俞 博士

Advisor: Wan-Yu Lin, Ph.D.

中華民國 103 年 7 月

July, 2014

# 誌　謝

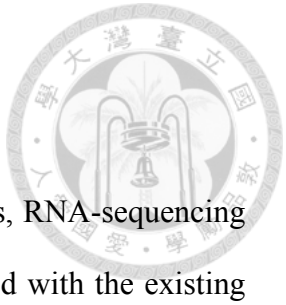
歷經兩年的碩士生涯淬鍊，此論文能夠完成，首先我特別要感謝的是我的指導教授，林菀俞老師，我非常感謝老師在我身上耗費相當多心思與精力對我的碩士論文不遺餘力地耐心指導我，從老師身上我看到了老師對學術研究的認真與堅持，與做學問嚴謹的態度，這些都是我必須學習的地方。在論文的修訂上，我特別感謝口試委員李文宗老師、楊欣洲老師、郭柏秀老師(依照姓氏筆畫排序)提供了寶貴與專業的意見，使得本論文內容能夠更趨全面與完整。

此外，我還要感謝我的朋友，所有幫助過我的同學，你們的陪伴使我的碩士生涯不孤單。最後，我要感謝我的家人，你們的支持是我完成碩士學位最大的動力。

# 中 文 摘 要

由於次世代定序(next-generation sequencing, NGS)的發展，核醣核酸定序(RNA-seq)實驗對基因體研究將出現改革性的影響。與既有的微陣列實驗相比，RNA-seq 實驗常提供了更準確的訊號。隨著定序成本的下降，RNA-seq 被認為將取代微陣列實驗。數個分析 RNA-seq 資料的統計方法已被發展，如 edgeR、DESeq2、baySeq、TSPM、NOISeq、SAMseq、Limma、EBSeq 以及 PoissonSeq 等，這九個統計方法在 RNA-seq 資料分析中常被使用。然而，這些方法之前置正規化處理、序列讀數的統計分布假設、偵測差異表現基因的統計方法以及錯誤發現率控制法皆有不同。如何選擇一個較具檢定力的方法仍是待解決的問題。在本論文中，我們針對此九個方法提供一系統性的整理與比較。除了蒙地卡羅(Monte-Carlo)模擬外，亦呈現兩則實際資料分析。根據我們的模擬結果，一般言之，Limma 方法和 baySeq 方法的表現最好，其中 Limma 方法的計算時間較短，為所有方法中較具分析 RNA-seq 資料潛力的方法。

關鍵詞：次數資料；差異表現基因；蒙地卡羅模擬；次世代定序；核醣核酸定序。



# 英 文 摘 要

As the development of next-generation sequencing technologies, RNA-sequencing (RNA-seq) experiment is revolutionizing genomic studies. Compared with the existing microarray technology, RNA-seq usually provides more precise signals. RNA-seq experiment is considered to replace microarray technology when the sequencing cost decreases. Several statistical methods for RNA-seq data analysis have been developed, such as edgeR, DESeq2, baySeq, TSPM, NOISeq, SAMseq, Limma, EBSeq, and PoissonSeq, etc. These nine statistical methods are popular in current RNA-seq data analyses. However, the normalization strategies, the distribution assumptions for reads counts, the statistical methods to detect differentially expressed genes, and the false discovery rate control for the nine methods are different. How to choose a more powerful method is still an open question. In this thesis, we provide a systematic summary and comparison of these nine methods. In addition to Monte-Carlo simulation studies, two real data analyses were also performed. Based on our simulation results, generally Limma and baySeq had the best performance among the nine methods we compared. Moreover, Limma was more computationally feasible than baySeq. Among these nine methods, Limma has more potential to analyze RNA-seq data.

Key words: Count data; Differentially expressed genes; Monte-Carlo simulation; Next-generation sequencing; RNA Sequencing.

# 目　錄

# 圖　目　錄

# 表　目　錄

# 第一章　前言

在生物體上，核醣核酸 (Ribonucleic acid; RNA)經由去氧核醣核酸 (Deoxyribonucleic acid; DNA)轉錄，附帶訊息的核醣核酸經過修飾後，形成信使核醣核酸 (messenger RNA; mRNA)。生物體在不同環境或是不同表現型 (phenotype) 下，信使核醣核酸存在的序列及基因表現量也會改變，進而產生不同的基因表現量。因此，分析基因體中所有被轉錄出來的信使核醣核酸序列及表現量，可探討生物間不同生理特性或是表現型的差異基因，並提供有用的資訊。

過去十幾年來，微陣列實驗 (microarray) 為探討大量基因表現差異的主要技術工具，但微陣列實驗在使用上有許多的限制，例如：微陣列實驗需經由已知的探針 (probe) 並用雜交 (hybridization) 的方式來判斷核醣核酸表現量，假使探針配對處的序列與基因發生變異，即便基因具有表現，但也會因為序列無法雜交而無法偵測到基因表現。且微陣列實驗還需仰賴已知全基因體序列作為雜交的平台。此外，在分析基因表現量時，微陣列實驗常藉由螢光顏色作為判斷依據，在分析過程中，如果基因表現量極低或極高下將無法準確偵測。近幾年來，由於新技術次世代定序 (next-generation sequencing, NGS) 的出現，發展出核醣核酸定序分析 (RNA-seq)，它可以對 RNA 剪接 (splicing) 和異構物 (isoform) 進行分析，且偵測到的差異表現量基因比微陣列實驗來的多。不僅如此，RNA-seq 克服了微陣列實驗的限制，即便是未完全解序的生物體，仍能偵測其基因表現量，RNA-seq 也逐漸取代了微陣列實驗，其應用範圍日漸廣泛，包括基因體學、生物醫學、環境科學、農業研究，…等。

RNA-seq 實驗可以由 Illumina's Genome Analyzer, Helicos BioSciences HeliScope, Applied Biosystems SOLiD, Pacific Biosciences SMRT 或是 Roche's 454 Life Sciences sequencing systems 來進行。圖一簡示 RNA-seq 實驗流程。RNA-seq 實驗是利用序列的讀數 (reads) 來判斷基因表現量，此表現量屬於離散型的個數資料，然而微陣列分析的方法是適用於連續型資料，所以過去分析微陣列實驗的統計

方法無法直接套用至 RNA-seq 實驗資料上,所以近年來許多學者開始發展 RNA-seq 實驗差異性檢定統計方法，例如 Robinson 和 Smyth (2010)所提出的 edgeR 方法[1-3]、Anders 和 Huber (2010)所提出的 DESeq 方法[4] (同研究團隊 Love、Huber 和 Anders 等人隨後於 2014 年提出 DESeq2 方法[5]，為 DESeq 的更新方法，本文以 DESeq2 為研究對象)、Hardcastle 和 Kelly (2010)所提出的 baySeq 方法[6]、Auer 和 Doerge (2011)所提出的 TSPM (two-stage Poisson model)方法[7]、Tarazona 等人 (2011) 所提出的 NOISeq 方法[8]、Li 和 Tibshirani (2013) 所提出的 SAMseq 方法[9]、Smyth (2004) 提出的 Limma 方法[10]、Leng 等人 (2013) 所提出的 EBSeq 方法[11]以及 Li 等人(2011)所提出的 PoissonSeq 方法[12]等，這些統計方法對序列讀數的分布假設不同，且各方法設定的模式、序列深度(sequencing depth)之處理、差異性檢定及錯誤發現率(false discovery rate, 簡稱 FDR)之校正也有不同。

本研究目的為比較多個差異性檢定統計方法在不同狀況的 RNA-seq 實驗表現之差異，我們透過不同的統計設定模擬 RNA-seq 實驗資料，並利用多個差異性檢定統計方法分析模擬的資料，比較這些差異性檢定統計方法在模擬資料上的表現，以探討差異性檢定統計方法之選擇。

本論文將於第二章介紹各方法、如序列讀數的分布假設、序列深度處理以及如何控制錯誤發現率等議題。在第三章模擬研究與探討中，將設計各種統計模擬情境來比較各方法在不同情境下的表現。第四章提供兩例實際資料分析。第五章除總結本研究外，將提出對研究者進行 RNA-seq 實驗分析時，選擇適當統計方法之建議。

# 第二章　方法

以下我們整理九個方法之 2.1 資料前置正規化處理、2.2 機率分布、2.3 統計方法以及 2.4 錯誤發現率控制，分別敘述於下四小節，並整理如表一。

## 2.1 資料前置正規化處理

在資料前置處理中，需先將序列作正規化 (normalization)，正規化的方法可以區分為以下幾類：

(1) 總數正規法 (total-count normalization)：使用總基因讀數來估計序列深度 (sequencing depth)。令 $Y_{jg}$ 為第 $j$ 個樣本在第 $g$ 個基因上的讀數，若共有 101 個基因，$d_j$ 代表第 $j$ 個樣本的序列深度，一個很直接的估計為 $\hat{d}_j = \sum_{g=1}^{101} Y_{jg}$，若有兩個樣本之序列深度分別為 $\hat{d}_1$ 與 $\hat{d}_2$，且 $\hat{d}_1 = 1.25\hat{d}_2$，則第二個樣本需乘以 1.25 倍，方可與第一個樣本的資料作比較，此稱為總數正規法。

(2) 總數正規法改良版 (revised total-count normalization)：相較於傳統的總數正規法使用全部的基因讀數和來估計序列深度，PoissonSeq 只使用不具顯著差異的基因來估計序列深度，Li 等人(2012)[12]提出一迭代方式找出不具顯著差異的基因，除此之外，亦可使用持家基因(house-keeping gene)估計序列深度，故稱為「總數正規法改良版」。PoissonSeq 和 SAMseq 皆採用此正規法。之所以如此改良的原因可由下例看出：若 $Y_{1g} = 100,\ Y_{2g} = 80$ (針對 $g = 1,2,\cdots,100$)，$Y_{1g} = 0,\ Y_{2g} = 2000$ (針對 $g = 101$)，若由總數正規法，$\hat{d}_1 = \sum_{g=1}^{101} Y_{1g} = 10000 = \hat{d}_2 = \sum_{g=1}^{101} Y_{2g}$，此意為第 1、2 個樣本在各個基因上的讀數是直接可相比的(directly comparable)，如此一來，則所有 101 個基因都將被判斷為有顯著差異表現。事實上，很可能只有第 101 個基因是有顯著差異表現

的，若只由不具顯著差異的基因(前 100 個基因)來估計序列深度，將得到 $\hat{d}_1 = 1.25\hat{d}_2$，則第二樣本需乘以 1.25 倍，方可與第一個樣本資料作比較，此即「總數正規法改良版」([9,12])。

(3) 七十五分位正規法 ($75^{th}$ percentile of nonzero count distribution)：去除所有讀數為 0 的基因後，以第七十五分位的基因讀數作為正規化係數。Bullard 等人 (2010)[13] 發現七十五分位正規法比一般的總數正規法來的穩健，其表現比許多不同的序列正規化方法來的好。baySeq 方法和 TSPM 方法皆採用七十五分位正規法。

(4) 中位數比值正規法 (median count ratio normalization)：有鑑於總數正規法易受少數高度顯著差異表現基因的影響，Anders 與 Huber (2010)提出中位數比值正規法[4]。令共有 $n$ 個樣本，第 $j$ 個樣本的序列深度估計為 $\hat{d}_j = \underset{g}{median} \frac{Y_{jg}}{\left(\prod_{j=1}^{n} Y_{jg}\right)^{\frac{1}{n}}}$，其中分母為 $n$ 個樣本在第 $g$ 個基因上的讀數幾何平均值 (Anders and Huber 2010 式 5)。DESeq2 和 EBSeq 皆採用此正規法。

(5) M 值截尾平均數法 (the trimmed mean of M values normalization, TMM)： 由 Robinson 與 Oshlack (2010)[14] 提出，亦為一有別於總數正規法的穩健正規法，M 值截尾平均數法亦使用較不具差異的基因來估計序列深度。Kvam 等人 (2012)[15]指出 M 值截尾平均數法和七十五分位正規法表現相差不大。edgeR 和 Limma 採用 M 值截尾平均數法來作正規化。

(6) RPKM (reads per kilobase per million mapped reads) 正規法：由 Mortazavi 等人 (2008)[16]提出，在 RNA-seq 實驗中，長度愈長的基因，被定序到的讀數一般會愈多，RPKM 將基因的長度作為正規化參數的考量。NOISeq 方法採用 RPKM 正規法。

## 2.2 機率分布

(1) 卜瓦松分布(Poisson distribution)：由於 RNA-seq 實驗之產出為離散型的個數資料，在統計上吾人常採用卜瓦松分布來描述離散型資料。TSPM 方法及 PoissonSeq 方法採用此分布假設。

(2) 負二項分布(negative binomial distribution)：在 RNA-seq 實驗中的生物樣本(biological replicate)常面臨過度離散(overdispersion)的問題，卜瓦松分布設定為變異數與期望值相等，故無法處理，許多學者便訴諸負二項分布。edgeR 方法、DESeq2 方法、baySeq 方法與 EBSeq 方法皆採用此分布假設。

(3) 無分布假設：母數的方法，不論是假設讀數呈卜瓦松分布或負二項分布，其結果常易受到離群值(outlier)的影響。若干研究者便提出無分布假設的統計方法，如：SAMseq 方法及 NOISeq 方法。

(4) 其它：Limma 方法原為微陣列分析統計方法，故只適合分析連續型資料，由於 RNA-seq 實驗為個數資料，一般作法將個數資料取對數(以 2 為底)，使資料轉化為連續型資料，再使用原先的 Limma 方法分析此轉換後的數據，並估計平均數與變異數之間的關係以決定每個觀察值的權重，此方法在文獻上稱為 voom (variance modeling at the observational level)轉換。

## 2.3 統計方法

(1) 條件概似函數(conditional likelihood)：代表方法為 edgeR。

edgeR 方法的提出本是為了分析 SAGE(serial analysis of gene expression)資料[2,3]，其可被視為小規模的 RNA-seq 資料，待後來的 RNA-seq 資料出現之後，edgeR 遂成為第一個可用以分析 RNA-seq 資料的方法[1]。edgeR 假設讀數呈負二項分布，負二項分布有兩個參數，除期望值之外，另一為過度離散參數(overdispersion parameter，文獻中常以$\phi$來表示)。在估計過度離散參數時，期望值成為干擾參數

(nuisance parameter)，因期望值的充分統計量(sufficient statistic)為基因的組內總讀數，給定基因的組內總讀數可建構條件概似函數，條件概似函數中便不再有期望值這個干擾參數，過度離散參數遂得以順利估計。最後，使用類似費雪精確檢定(Fisher's exact test)的方式來檢定各基因是否有差異表現。因採用精確檢定之故，edgeR 亦可運用於樣本數小的情況下[3]。

edgeR 因估計一個共同的過度離散參數，當所有基因的過度離散參數相去不遠時，edgeR 的方法表現良好；但針對實際資料，所有基因有相同的過度離散參數是不太可能的，此時 edgeR 的表現將會受到影響。

(2) 局部迴歸(local regression)：代表方法為 DESeq2。

Anders 與 Huber 等人於 2010 年提出 DESeq 方法[4]，該團隊於 2014 年提出 DESeq2 方法[5]，DESeq2 為 DESeq 的更新版本，故以下討論將針對 DESeq2。DESeq2 與 edgeR 相同，假設讀數呈負二項分布，不同於 edgeR 假設所有基因的過度離散參數相同，DESeq2 允許基因有不同的離散參數，將表現量相似的基因綜合起來，以局部迴歸(local regression)的方式來估計變異數與期望值之間的關係。當基因有不同的離散參數時，DESeq2 的表現應優於 edgeR。最後，使用華德檢定(Wald's test)來檢定各基因是否有差異表現。DESeq2 的 R 套件允許使用者選擇華德檢定或概似比檢定(likelihood-ratio test)，因華德檢定不需配適小模式(reduced model，即在虛無假設下的模式)，成為 DESeq2 套件的預設選擇，故本文之後的模擬研究亦採用此檢定。

(3) 對數線性模式(log-linear model)：代表方法為 PoissonSeq 與 TSPM。

PoissonSeq 將所有基因分為十組，讀數和接近的基因歸為一組，使用冪次轉換(power transformation)將次數資料的過度離散情形去除，使得變異數約等於期望值，轉換後的資料遂呈卜瓦松分布，接著採用對數線性模式下的分數檢定(score test)來檢定基因的差異表現[12]。若資料過度離散參數的變動很大，無法以十組來完美描述，PoissonSeq 的表現將會受到影響。

另一由對數線性模式為出發點的方法為 TSPM，TSPM 採用二階段處理方式，第一階段先使用分數檢定判別個別基因是否有過度離散現象，將基因分為兩群，一群有過度離散現象，另一群則無。第二階段則使用概似比檢定判別基因是否有顯著差異表現，針對有過度離散現象的基因，大樣本之下其概似比檢定呈 $F$ 分布；而無過度離散現象的基因，大樣本之下其概似比檢定呈一般的卡方分布。

(4) 經驗貝氏法(empirical Bayesian method)：代表方法為 baySeq、EBSeq 與 Limma。

baySeq 法假設讀數呈負二項分布，以此建構概似函數(likelihood)，先驗分布(prior distribution)採擬概似函數(quasi-likelihood function)，先驗分布之參數則使用拔靴法(bootstrapping)由資料來估計，再結合先驗分布與概似函數來導出後驗分布(posterior distribution)。擬概似函數是一種半母數的方法(semi-parametric method)[17]，Hardcastle 與 Kelly 發現擬概似函數比起母數方法(整個分布皆被指定)表現更佳，因其具備強韌性(robustness)，故採擬概似函數法來決定先驗分布[6]。由於先驗分布之參數使用拔靴法由資料來估計，baySeq 的計算耗時較久。

EBSeq 方法也假設讀數呈負二項分布，參數之先驗分布則假設為貝塔分布(Beta distribution)，先驗分布內的參數則從資料估計而來，再結合先驗分布與概似函數來導出後驗分布[11]。EBSeq 方法很類似 baySeq 法，不同的是，先驗分布假設為貝塔分布，是有母數的假設。在先驗分布接近貝塔分布時，EBSeq 表現較佳；而在先驗分布偏離貝塔分布時，半母數設定的 baySeq 表現較佳。

Limma 方法為微陣列實驗分析之方法，故資料與先驗分布皆假設為常態分布，先驗分布中的參數則由經驗貝氏法估計得來。在此設定之下，後驗分布亦為常態分布，故使用 $T$ 檢定統計量來檢定基因是否存在顯著差異[10]。由經驗貝氏法借力使力之特性[18]，可增加檢定的效率(efficiency)，提高統計檢定力，即便是樣本數小的情況下仍能有不錯的表現。故 Limma 被廣泛地運用在微陣列資料分析上，後有研究者運用其於 RNA-seq 資料上，惟須先將 RNA-seq 資料的個數型態經

由 voom 轉換，以接近常態分布，方能使用 Limma 套件加以分析。

(5) 無母數統計方法：代表方法為 SAMseq 與 NOISeq。

SAMseq 由 Li 與 Tibshirani 提出[9]，此團隊亦曾提出 PoissonSeq 法[12]，差別在於 SAMseq 是無母數方法，利用魏克森排序統計量(Wilcoxon rank statistic)，並由樣本重抽法(resampling)來檢定基因是否存在顯著差異。根據 Li 與 Tibshirani [9]，SAMseq 適用於樣本數≥10 的情況。在母數分布假設不合時，SAMseq 方法的表現常比母數方法好。但在樣本數很小的情況下，如：兩組資料、每組各兩個樣本，魏克森排序統計量的統計檢定力將會很低。

NOISeq 從實際數據中的組內資料來建立讀數相除和相減之聯立分布，稱為「噪音分布」(noise distribution)，組間的讀數相除和相減便與噪音分布比較，以檢定各基因上的讀數是否達顯著差異[8]。此為一個資料適性的無母數方法(data adaptive non-parametric method)。不同於 SAMseq，NOISeq 將資料獨特性考量在統計量(讀數相除和相減)的抽樣分布裡。

## 2.4 錯誤發現率控制

RNA-seq 可高通量地分析生物體所有基因的表現量，一次對成千上萬的基因作檢定，需作多重檢定校正。各方法軟體套件所產出的結果可區分為以下幾種：

(1) BH 方法(Benjamini-Hochberg procedure, 簡稱 BH)：如下表，FDR 被定義為 $E\left(\frac{V}{R} \mid R>0\right)\times \Pr(R>0)$，為所有被判為顯著的檢定中犯錯的比率。Benjamini 與 Hochberg (1995)[19]提出一控制 FDR 的方法，為最典型控制 FDR 的算則，Limma、TSPM、edgeR 和 DESeq2 皆使用 BH 法來作多重檢定校正。

| | 不拒絕<br>虛無假設 $H_0$ | 拒絕<br>虛無假設 $H_0$ | |
|---|---|---|---|
| 虛無假設 $H_0$ 為真 | $U$ | $V$ | $m_0$ |
| 對立假設 $H_1$ 為真 | $T$ | $S$ | $m_1$ |
| | $W$ | $R$ | $m$ |

(2) 排列錯誤發現率(permutation-based FDR)：此方法乃是經由隨機重排基因的讀數資料，建構出在虛無假設之下的檢定統計量分布，此稱排列分布(permutation distribution)。計算排列分布中有多少的檢定統計量大於某門檻值，此即在此門檻值之下的偽陽性個數(the number of false positives)，相關文獻可參考[20-24]。但因卜瓦松分布次數資料期望值與變異數之間的關聯性，具差異表現基因之排列分布往往具有較大的變異程度，與真正的排列分布並不相仿。PoissonSeq 與 SAMseq 之研究團隊[9,12]便提出僅使用傾向不具差異表現基因(檢定統計量值較低者)的排列分布，來計算排列 FDR，以免高估 FDR。

(3) 貝氏 FDR：EBSeq 和 baySeq 皆採用貝氏方法估計 FDR，計算出基因具有差異表現之後驗機率，即 $P\left(H_1 \mid Data\right)$。

(4) 其它：NOISeq 方法依據無母數方法所建立之產出為

$q_{NOISeq} \equiv P\left(\left|M^*\right| < \left|m^g\right|, D^* < d^g\right)$，其中 $m^g = \log_2\left(\frac{x_1^g}{x_2^g}\right)$ 與 $d^g = \left|x_1^g - x_2^g\right|$ 分別為第 $g$ 個基因組間的讀數相除和相減值，而 $M^*$ 與 $D^*$ 為噪音分布(組內資料建立的讀數相除和相減之聯立分布)隨機變數[8]。$P\left(\left|M^*\right| < \left|m^g\right|, D^* < d^g\right)$ 為噪音分布裡比觀

察到的$\left(\left|m^g\right|, d^g\right)$小的機率，概念相當於 1−P 值。

# 第三章　模擬研究與探討

## 3.1　模擬設定

吾人利用蒙地卡羅模擬(Monte Carlo Simulation)比較 edgeR、DESeq2、PoissonSeq、TSPM、baySeq、EBSeq、Limma、SAMseq 與 NOISeq 等九個方法在分析 RNA-seq 資料上的表現，由 R 統計軟體 2.15.2 版本計算，可由(http://www.R-project.org)下載。在模擬中，因 PoissonSeq 預設只考慮讀數總和(the total number of reads of a gene across all experiments)超過 5 的基因，為公平起見，我們讓所有方法皆只分析讀數總和超過 5 的基因。根據每次的模擬結果，可計算真陽性比率(true positive rate, TPR)和偽陽性比率(false positive rate, FPR)。如表二所示，考慮有 $m$ 個基因，$m$ 即為檢定總數，經由統計檢定判斷為無差異表現與有差異表現的基因分別為 $W$ 與 $R$ 個。根據真實狀態及檢定結果可得表二交叉表：其中 TP 為在真實狀況下有差異表現，且被正確地判斷為有差異表現的基因個數；FP 為在真實狀況下為無差異表現，卻被錯誤判斷為有差異表現的基因個數；FN 為在真實狀況下為有差異表現，卻被錯誤判斷為無差異表現的基因個數；TN 為在真實狀況下為無差異表現，且被正確地判斷為無差異表現的基因個數。TPR 為在所有實際為有差異表現的基因中，被正確地判斷為有差異表現之比率，定義如下：

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad 。$$

FPR 為在所有實際為無差異表現的基因中，被錯誤地判斷為有差異表現之比率，定義如下：

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \quad 。$$

根據檢定結果，各方法可得某一度量衡，藉以判斷基因是否具差異表現，此度量衡可能是 FDR、排列 FDR、貝氏 FDR 或其它(如上一章所提，NOISeq 產出為

$P\left(\left|M^*\right|<\left|m^g\right|, D^*<d^g\right)$)。一個好的統計方法與套件，其產出度量衡在區分真陽性與真陰性上應具備良好的能力，吾人將使用模擬研究與接收者操作特徵曲線(receiver operating characteristic curve，ROC curve)來衡量之。

首先，統一各度量衡的方向，將 NOISeq 的產出調整為 $1-P\left(\left|M^*\right|<\left|m^g\right|, D^*<d^g\right)$，此值愈小愈傾向於拒絕虛無假設，與其它方法的度量衡：FDR、排列 FDR、貝氏 FDR 的方向一致。舉例來說，若我們設定度量衡門檻值為 0.05，度量衡小於此門檻值的基因遂被宣稱為具有顯著差異表現，再根據檢定結果和真實情況計作比對，計算出一組 TPR 與 FPR，TPR 與 FPR 皆介於 0 到 1 之間。隨著度量衡門檻值的變動(自 0 至 1，每隔 0.0005 設一門檻值)，可計算不同門檻值之下所對應的 TPR 與 FPR，再將 TPR 視為座標上縱軸，FPR 作為橫軸，繪出 ROC 曲線。從 ROC 曲線中，若控制各方法橫軸 FPR 不變的情況下，TPR 愈高則代表其表現愈好，以此可評估各方法的表現。

另外，在比較不同的統計方法時，吾人計算 ROC 曲線下方的面積(area under the ROC curve, 簡稱 AUC)，作為評估各方法優劣的指標。由於多數研究者並無法忍受過大的 FPR，吾人有興趣的部分通常是在 ROC 曲線圖中 FPR 小的部分(此處吾人呈現 FPR≤0.1 的部分)，所以我們改而計算此部分的 ROC 曲線下方面積，此稱為部分 AUC(partial area under the curve of ROC, 簡稱 pAUC)[25-27]，利用數值積分的方式計算出 pAUC，由於是在 0.1x1 的長方格裡求面積，pAUC 範圍介在 0 到 0.1 之間，若 pAUC 愈大，則代表該方法的判別正確率愈高。

## 3.2 模擬設計一

我們共設定了三個模擬情境，第一個模擬的設定參考 Auer 與 Doerge (2011)文章中的模擬設定[7]，Auer 與 Doerge(2011)[7]稱此設定最接近真實的 RNA-seq 資料。我們考慮一萬個基因，假設一半基因來自卜瓦松分布，另一半基因來自不同

變異程度的過度變異卜瓦松分布(over-dispersed Poisson with different degrees of overdispersion)。在這一萬個基因裡，兩千個基因存在差異表現，由於 RNA-seq 資料樣本數通常相當少，所以每一個治療組別中各組樣本數分別設為 2,4,6,8,10 個，並分為控制組和對照組，所以總樣本數分別為 4,8,12,16,20。總共模擬一百次，之後呈現的 ROC 曲線與 pAUC 皆是一百次模擬的平均結果。假設原始的讀數資料經由正規化後服從下列的卜瓦松分布，令 $i=1,2,\ j=1,2,...,n_i,\ g=1,...,10000,$

$$Y_{ijg} \sim \text{Poi}\left(\lambda_{ig}\nu_{ijg}\right). \tag{1}$$

首先由柏拉圖(Pareto)分布產生亂數，並取其指數(exponential)，作為 $\lambda_{ig}$ 值。

當 $g=1,2,...,8000,$ 指定 $\lambda_{1g}=\lambda_{2g}$ ，

$$\lambda_{1g}=\lambda_{2g} \sim \exp\left(\text{Pareto}\left(\text{location}=3,\text{shape}=7\right)\right),$$

當 $g=8001,8002,...,10000,$ $\lambda_{1g}$ 與 $\lambda_{2g}$ 則分別抽出，

$$\lambda_{1g},\lambda_{2g} \overset{\text{iid}}{\sim} \exp\left(\text{Pareto}\left(\text{location}=3,\text{shape}=7\right)\right),$$

即代表前 8000 個基因於組間無差異表現，而第 8001 至第 10000 個基因則在不同組間存在差異表現。此外，為模擬 RNA-seq 資料之過度離散的現象，針對每一個基因，我們產生伯努利隨機變數，

$$Z_g \sim \text{Bernoulli}\left(p\right), g=1,...,10000, p\in\left[0,1\right].$$

若 $Z_g=0$，則設模型(1)中之 $\nu_{ijg}=1$，表示第 g 個基因在兩組中的讀數皆來自平均數為 $\lambda_{ig}$ 的卜瓦松分布。若 $Z_g=1$，則該基因的 $\nu_{ijg}$ 值由伽瑪(Gamma)分布來產生：

$$\nu_{ijg} \sim \text{Gamma}\left(\frac{\lambda_{ig}}{\phi_g-1},\frac{\phi_g-1}{\lambda_{ig}}\right),$$

$$\phi_g \overset{\text{iid}}{\sim} \exp\left(\text{Pareto}\left(\text{location}=1,\text{shape}=3\right)\right)-1.$$

因此隨機值之影響，每一組中第 g 個基因讀數皆來自平均數為$\lambda_{ig}\nu_{ijg}$的卜瓦松分布，使得該基因的各筆資料之平均數有變異，進而產生過度變異情形。我們的模擬考慮 p=0.5，代表有 50%的基因有過度變異的情形。

圖二為根據一百次模擬之 ROC 圖，而各方法的 pAUC 列在表三，在樣本數為二的情況下，可發現 Limma 方法和 baySeq 方法的表現最好，其次是 DESeq2 方法，由於 Limma 與 baySeq 採經驗貝氏法，即使是樣本數小的情況下仍能有不錯的表現。另外六個方法表現較差，SAMseq 表現較差的原因可能來自於樣本數很小情況下，魏克森排序統計量的統計檢定力很低(樣本數很小時排序只能有幾種情形)。NOISeq 之產出度量衡為$P\left(\left|M^*\right|<\left|m^g\right|,D^*<d^g\right)$，即噪音分布裡比觀察到的$\left(\left|m^g\right|,d^g\right)$小的機率，此度量衡無法在效應小(small-size effect)時提供好的檢定力，故 NOISeq 之表現差，Xu 等人亦發現效應小時 NOISeq 的表現不好[28]。

當樣本數為四、六、八、十的情況下，其中表現較好的方法為 Limma 方法、baySeq 方法、TSPM 方法(隨樣本數增加而變好)、DESeq2 方法和 SAMseq 方法，另外四方法(edgeR, NOISeq, EBSeq, PoissonSeq)表現較差。值得注意的是在樣本數二表現較差的 SAMseq 方法其表現大幅改善，在樣本數達四或以上時晉升為優良的方法。在任何樣本數情況下，DESeq2 的表現皆優於 edgeR，可能的原因為 DESeq2 假設基因有不同的離散參數，與模擬資料來自不同變異程度的卜瓦松分布較符合。edgeR 對所有基因估計一個共同的離散參數，故表現不佳。PoissonSeq 將所有基因分為十組，讀數和(total reads)接近的基因歸為一組，以冪次轉換來處理過度離散情形，可能由於此資料過度離散參數無法以十組來完美描述，故 PoissonSeq 的表現亦受到影響，但其仍比只估計一個共同的離散參數的 edgeR 來得好一些(特別是在 FPR 小的那一端)。

此外，我們亦模擬當 p=0.2 與 0.8 時，代表有 20%或 80%的基因有過度變異的情形，圖三與圖四分別為其根據一百次模擬之 ROC 圖，而各方法的 pAUC 分別列

在表四與表五。比較圖二至圖四(表三至表五)，可發現各方法的相對表現是一致的。另外，我們亦發現，整體而言，當基因有過度變異的比例愈低時，各方法的表現愈佳。舉兩組內樣本數各為二的情形來說，Limma 與 baySeq 的 pAUC 為 0.046 (當有 20%的基因有過度變異時), 0.041 (當有 50%的基因有過度變異時), 與 0.036 (當有 80%的基因有過度變異時)。隨著過度變異基因的比例上升，pAUC 呈現下降的現象。但不同的過度變異基因比例並不至於影響各方法的相對表現。

## 3.3 模擬設計二

第二個模擬的設定由真實的玉米資料而來。Li 等人(2010)[29]從玉米樹上挑選生長中的葉子，並利用 Illumina 定序科技從葉上四個代表部位和由雷射微切除設備（laser capture microdissection, LCM）所得的維管束鞘與葉肉細胞樣本當中獲得基因表現量，分別得到兩個生物樣本。我們的模擬將依據上述 LCM 所得樣本之統計量進行設定。利用電腦模擬經由以下步驟產生資料。考慮一萬個基因，其中 6,666 個基因存在差異表現，每一個治療組別中樣本數分別為 2,4,6,8,10 個，分為控制組和對照組，總共模擬一百次。假設原始的讀數資料經由正規化後服從下列的負二項分布，令 $i = 1,2,\ j = 1,2,...,n_i,\ g = 1,...,10000,$

$$Y_{ijg} \sim NB\left(\text{mean} = \lambda_g \text{Exp}\left((-1)^i \delta_g\right), \text{dispersion} = \phi_g\right). \quad (2)$$

首先由真實玉米葉子資料中[29]，採用實證估計的方式，利用最大概似估計法估計伽瑪(Gamma)分布中參數，得到 $\alpha = 0.28, \beta = 666$，並由此伽瑪分布產生亂數，作為 $\lambda_g$ 值。

$$\lambda_g \overset{iid}{\sim} \text{Gamma}\left(\alpha = 0.28, \beta = 666\right).$$
$$\text{E}\left(\lambda_g\right) = \alpha\beta = 0.28 \times 666 = 186.67.$$

另一方面，為估計「折半對數倍數變化」(half log fold-change)，$\delta_g$ 值，依據 Kvam 等人(2012)[15]所作之模擬設定，利用期望最大化演算法 (expectation-maximization,

EM)配適三成分的混合常態分布(three-component normal mixture distribution)，並由此混和常態分布產生亂數，作為$\delta_g$值。

當$g = 1, 2, ..., 3333,$

$$\delta_g \overset{iid}{\sim} \text{Normal}\left(\mu = 0.96, \sigma = 0.725\right),$$

當$g = 3334, 3335, ..., 6666,$

$$\delta_g \overset{iid}{\sim} \text{Normal}\left(\mu = -0.96, \sigma = 0.725\right),$$

當$g = 6667, 6668, ..., 10000$時，所有$\delta_g$值皆為 0。亦即，前 6,666 個基因在不同組間存在差異表現，而第 6,667 個至第 10,000 個基因在組間不存在差異表現。此外，為模擬 RNA-seq 資料過度離散的現象，參照 Hardcastle 與 Kelly(2010)[6]對過度離散參數的設定，針對每一個基因，皆由伽瑪分布產生亂數作為$\phi_g$值。

$$\phi_g \overset{iid}{\sim} \text{Gamma}\left(\alpha = 0.85, \beta = 0.5\right).$$

$$\text{E}\left(\phi_g\right) = \alpha\beta = 0.85 \times 0.5 = 0.425.$$

圖五為根據一百次模擬之 ROC 圖，而各方法的 pAUC 列在表六，在兩組樣本數各為二的情況下，與模擬設計一的結果相同，仍以 Limma 方法和 baySeq 方法的表現最好，由於 Limma 與 baySeq 採經驗貝氏法，即使是樣本數小的情況下仍能有不錯的表現。SAMseq, NOISeq, EBSeq, 與 TSPM 的表現最差，SAMseq 表現差的原因可能來自於樣本數很小情況下，魏克森排序統計量的統計檢定力很低(樣本數很小時排序只能有幾種情形)；NOISeq 表現差的原因可能來自於其產出度量衡 $P\left(\left|M^*\right| < \left|m^g\right|, D^* < d^g\right)$ 無法在效應小(small-size effect)時提供好的檢定力[28]；EBSeq 表現差可能是因其貝塔先驗分布參數估計不佳(樣本數太小)；TSPM 表現差，因其第一階段判別個別基因是否有過度離散現象乃使用分數檢定，第二階段則使用概似比檢定判別基因是否有顯著差異表現，此兩階段的檢定皆仰賴大樣本

理論，在每組樣本數為二時大樣本理論恐較難成立，將影響 TSPM 於小樣本時的表現。

隨著樣本數的增加，Limma 與 baySeq 一直維持著最佳表現，TSPM、EBSeq 與 SAMseq 的表現大幅改善，NOISeq 亦有些許改善。大略說來，DESeq2 的表現優於 edgeR，可能的原因為 DESeq2 假設基因有不同的離散參數，與資料實際情形較符合。edgeR 對所有基因估計一個共同的離散參數，故表現稍差。PoissonSeq 將所有基因分為十組，讀數和(total reads)接近的基因歸為一組，以冪次轉換來處理過度離散情形，由於此資料過度離散參數無法以十組來完美描述(見下一節的論述與表八的整理)，故 PoissonSeq 的表現亦受到影響，但其仍比只估計一個共同的離散參數的 edgeR 來得好一些(特別是在 FPR 小的那一端)。值得注意的是，組內樣本數為二時，edgeR 比 PoissonSeq 好；組內樣本數為四時，edgeR 比 PoissonSeq 稍好一點，此乃因 PoissonSeq 採用對數線性模式下的分數檢定(score test)來檢定基因的差異表現，此仰賴大樣本理論，在組內樣本數為二或四時，大樣本理論的結果將稍受影響。

## 3.4 模擬設計三

第三個模擬的設定亦是根據 Li 等人(2010)[29]的真實資料，由 LCM 得到玉米的維管束鞘與其葉肉細胞樣本，此資料當中共有 110,185 個基因，兩組樣本數各為二(共四個)，我們採用實證估計的方式決定模擬參數。本模擬設定與上一個模擬設定略有相同之處，皆由 Li 等人(2010)[29]的實證資料而來。讀數資料假設來自負二項分布，組內樣本數分別為 2,4,6,8,10 個，總共模擬一百次，經由以下步驟產生資料。假設原始的讀數資料經由正規化後服從下列的負二項分布：

$$Y_{ijg} \sim NB\left(\text{mean} = \lambda_g \text{Exp}\left((-1)^i \delta_g\right), \text{dispersion} = \phi_g\right), \qquad (3)$$

其中 $i = 1,2,\ j = 1,2,...,n_i,\ g = 1,2,...,110185$, 我們由 LCM 資料中計算跨組別之下各個

基因的平均數，作為$\lambda_g$的值。再利用 edgeR 方法分析此資料，獲得各基因差異表現檢定之 *P* 值(*P*-value)，若 *P* 值小於 0.01，即設定此基因具有差異表現(這樣的基因有 3,510 個)，其「折半對數倍數變化」(half log fold-change)$\delta_g$值由 edgeR 套件來估計；反之，若 *P* 值大於 0.01，即設定此基因無差異表現(有 106,675 個基因)，則$\delta_g$值即設為 0。過度離散參數$\phi_g$值亦由 edgeR 套件所估計。

圖六為根據一百次模擬之 ROC 圖，而各方法的 pAUC 列在表七，在各組樣本數皆為二的情況下，與前兩個模擬設計結果不同的是，最好的方法是 edgeR，其次為 PoissonSeq。此乃因模擬設計三基因的過度離散參數變動不大，見表八，過度離散參數的變異數是三個模擬設計裡最低的，故只估計一個共同離散參數的 edgeR 反能有最好的表現；PoissonSeq 僅以十個過度離散參數來描述資料，也能有不錯的表現。而表現最差的依序是 TSPM、EBSeq 與 SAMseq(特別是在 FPR 小時)，TSPM 表現差，因其兩階段的檢定皆仰賴大樣本理論，在每組樣本數為二時大樣本理論恐較難成立，將影響 TSPM 於小樣本時的表現；EBSeq 概念上類似 baySeq 法，但表現卻常比 baySeq 差，可能原因是其假設先驗分布為貝塔分布，其中的參數(hyper-parameter)估計在樣本數小時較不準確；在樣本數很小的情況下，魏克森排序統計量的統計檢定力很低(樣本數很小時排序只能有幾種情形)，故 SAMseq 的表現不佳，這與模擬設計一、二的結果類似。

隨著樣本數的增加，各方法皆有改善，尤其以 TSPM 與 SAMseq 的進步最顯著，這與模擬設計一、二的結果類似。EBSeq 亦改善許多，應是因為樣本數增加之後，先驗分布裡的參數(hyper-parameter)得以較準確地估計所致。

如同之前所述，Xu 等人 2013 年亦提出，NOISeq 的表現和效應大小(effect size)有關[28]，在真正的效應較小時，NOISeq 的表現相對較差。我們比對模擬設計一、二、三的結果，發現在模擬設計一所有的樣本數之下，NOISeq 皆為最差的方法；在模擬設計三時，NOISeq 的表現不錯；在模擬設計二時，NOISeq 的表現尚可(特

別是在小樣本數時，比幾個競爭方法還要好)。經由我們分析效應大小(如表八)，模擬設計一之效應果真是最低的(對數倍數變化平均值為 0.561)；模擬設計三之效應則較高(對數倍數變化平均值為 1.689)，且變異數不高(對數倍數變化變異數為 0.726)，表示各個基因的效應一致地偏高；模擬設計二之效應雖然也高(對數倍數變化平均值為 1.920)，但變異數很高(對數倍數變化變異數為 2.1025)，顯示部分基因效應偏高但部分偏低(變動很大)，NOISeq 於分析效應小的基因時比較不利，導致在模擬設計二時，NOISeq 的相對表現比不上在模擬設計三的相對表現。NOISeq 的表現和基因的效應大小約略有正相關性，雖所有方法的表現皆會與效應大小有關，但 NOISeq 對效應大小的反應較其它方法更為敏感。

若改以 DESeq2 套件來估計各基因的效應大小時，由圖七與表九發現 DESeq2 的表現最好(在 FPR 小的區域)，上述以 edgeR 套件來估計各基因的效應大小時表現最好的 edgeR 在此時表現亦頗佳，僅次於 DESeq2(在 FPR 小的區域)。這表示雖然在模擬設定三的作法會有球員兼裁判的盲點存在，以何種統計套件來估基因效應大小的確有利於該方法，但因過度離散參數的變動不大(過度離散參數的變異數 0.013 是三個模擬設計裡最低的)，故 edgeR 的表現仍然頗佳。

# 第四章 實證資料分析

我們將各方法運用到 Li 等人(2010)[29]的 RNA-seq 資料和 Marioni (2008)[30]的RNA-seq 資料，並比較各方法在真實序列資料上表現的差異。我們經由R的dexus封包(http://www.bioconductor.org/packages/devel/bioc/html/dexus.html)中的資料矩陣(名稱：countsLi)取得 Li 等人(2010)的資料，其中 countsLi 中的 SRR039509 和 SRR039510 為葉肉細胞樣本，SRR039512 和 SRR039514 為維管束鞘樣本，此資料由 LCM 設備得到，兩組樣本數各為二(兩組共四個)，作為生物複製(biological replicate)樣本，排除資料當中讀數總和不超過五之基因數目，剩下 22,745 個基因。

而 Marioni (2008)資料來自一人類身體中的肝和腎臟，各個器官上由 Illumina 次世代定序平台(NGS platform)各自重複檢測五次，作為技術複製(technical replicate)樣本，故此資料包含了五個腎臟資料和五個肝臟資料，排除資料當中讀數總和不超過五之基因數目，剩下 18,228 個基因。Marioni (2008)資料是一技術複製(technical replicate)樣本。

本論文呈現一生物複製樣本(Li 等人(2010)[29]的玉米資料)與一技術複製樣本(Marioni(2008)[30]人的肝與腎資料)，此二者主要差異在過度離散程度的大小。一般而言，生物複製樣本的過度離散程度較技術複製樣本大。經計算，Li 等人(2010)生物複製資料的過度離散參數之平均為 0.076，Marioni (2008)技術複製資料的過度離散參數平均為 0.005，確實是以技術複製樣本的過度離散程度較小。

表十為 Li 等人(2010)資料 FDR 水準控制在 5%時各方法所偵測出的顯著基因個數。由於 NOISeq 方法並未產出 FDR 值，其產出為 $q_{NOISeq} \equiv P\left(\left|M^*\right| < \left|m^g\right|, D^* < d^g\right)$，我們採用原發明者 Tarazona 等人的建議，FDR$\leq 0.05$ 約略相當於 $q_{NOISeq} \geq 0.8$ [8]，故表中 NOISeq 所偵測出的顯著基因個數是 $q_{NOISeq} \geq 0.8$ 的基因個數。PoissonSeq 方法和 TSPM 方法分別為 4,693 和 4,234，在所有方法中偵測出較多的顯著差異表現基因個數，此二方法表現較樂觀；而 baySeq 方法和

NOISeq 方法分別為 1,712 和 1,525，此二法偵測出較少的顯著差異表現基因個數，較為保守。另一方面，TSPM 方法偵測出較多異於其它方法所偵測到的基因，表示其在此資料分析裡和其它方法的差異甚大，根據我們在三個模擬設計裡觀察的結果，TSPM 在總樣本數為四時(兩組樣本數各為二)的表現頗差，故吾人對只由 TSPM 找出的 1,627 個基因需持保留態度。反之，baySeq 方法、edgeR 方法、DESeq2 方法、Limma 方法和 NOISeq 方法在所有方法中偵測出較少異於其它方法所偵測之基因，這些方法所找出顯著的基因多有其它方法加持。另外，所有方法共同偵測出的顯著差異表現基因只有 176 個，在時間、成本有限時，可優先研究這 176 個基因，因其顯著差異表現乃由眾方法所公認。

表十二為 Marioni (2008)資料的分析結果，SAMseq 方法、PoissonSeq 方法和 TSPM 方法偵測出較多的顯著差異表現基因，表現較樂觀；EBSeq 方法和 NOISeq 方法所偵測顯著差異表現基因個數則明顯低於其它方法，其表現較其它方法保守。另一方面，除 SAMseq 方法和 PoissonSeq 方法偵測出較多異於其它方法所偵測的基因，絕大多數被偵測出顯著差異表現的基因是由兩個或兩個以上的方法所認可。而由九方法共同找出的基因有 3,992 個，其顯著差異表現由眾方法所公認。

# 第五章　結論與討論

過去在探討基因差異性的主要工具為微陣列(microarray)實驗，然近年來RNA-seq 實驗逐漸取代了微陣列實驗，愈來愈多研究者開始發展 RNA-seq 資料的差異性檢定統計方法。然而，該選用哪一種 RNA-seq 資料分析方法卻莫衷一是。在此研究中，吾人希望能夠提供研究者分析 RNA-seq 資料之準則。首先，我們模擬 RNA-seq 資料，盡量使模擬設定接近真實情況，模仿實際情況來產生讀數資料，根據 Auer 與 Doerge(2011)[7]所述，模擬設計一非常接近真實的 RNA-seq 資料。除此之外，亦採用實證 RNA-seq 資料來進行模擬，如模擬設計二與三。根據模擬結果，我們比較九種常用的 RNA-seq 資料差異性檢定統計方法。

在第三章模擬研究中我們利用 ROC 曲線來探討各方法在 FPR 較小時的表現狀況，綜合各模擬狀況，在過度離散參數變動較大時(如模擬設計一、二)，Limma 方法和 baySeq 方法的表現最好。我們使用的分析工作站是 Intel Xeon E5-2690，其中央處理器時脈為 2.9 GHz，記憶體容量為 4 GB，表十四～表十六列出在三個模擬設計下各方法平均計算時間(單位：秒)。雖然 baySeq 方法的表現很好，但其計算時間過長；Limma 方法非但表現良好，其分析時間亦是所有方法中最短的，尤其是在樣本數增加時，未見其計算時間明顯增加，堪稱所有方法中最具備分析 RNA-seq 資料潛力的方法。雖然 Limma 方法在 2004 年便由 Smyth 提出，當初是為了分析微陣列的資料[10]，但經由 voom 轉換後，其表現比後來專為 RNA-seq 次數型資料發展的方法還要好，計算亦非常迅速。

在過度離散參數變動較小時(如模擬設計三，以 edgeR 來估計各基因的效應大小時)，edgeR 方法(對所有基因只估計一個過度離散參數)和 PoissonSeq 方法(對所有基因估計十個過度離散參數)的表現變好。所有方法在樣本數增加時表現皆能變好，尤其以 TSPM、SAMseq 與 EBSeq 為最。若改以 DESeq2 來估計各基因的效應大小，edgeR 仍能維持頗佳的結果(尤其是在 FPR 值較低時)，僅次於 DESeq2。

整體而言，NOISeq 的表現最差，其產出度量衡$P\left(\left|M^*\right|<\left|m^g\right|,D^*<d^g\right)$，即噪音分布裡比觀察到的$\left(\left|m^g\right|,d^g\right)$小的機率，此度量衡無法在效應小(small-size effect)時提供好的檢定力。且 NOISeq 的產出度量衡並無法與 FDR 對應(除 FDR≤0.05 約略相當於$q_{NOISeq}\geq 0.8$ [8] 之外)，此度量衡在許多 RNA-seq 分析套件中顯得相當特異，難以和其它方法比對，加上其不佳的有/無差異表現基因判斷力、不甚省時的分析時間(見表十四～表十六)，將降低其對研究者的吸引力。

RNA-seq 資料常有樣本數很小的問題，若總樣本數只有四(兩組各二)，研究者應避免使用 SAMseq、TSPM 與 EBSeq 方法，SAMseq 表現較差的原因可能來自於樣本數很小情況下，排序只能有幾種情形，魏克森排序統計量的統計檢定力便很低；TSPM 表現差，因其第一、二階段分別使用分數檢定與概似比檢定來判別基因是否有過度離散現象或顯著差異表現，此兩種檢定皆仰賴大樣本理論，在每組樣本數為二時大樣本理論恐難以成立，此將影響 TSPM 於小樣本之下的表現；EBSeq 樣本數小時表現差，應與其先驗分布參數值(hyper-parameter)估計不準確有關。

當過度離散參數變動較大時(如模擬設計一、二)，PoissonSeq 與 edgeR 的表現在眾方法中不甚好。PoissonSeq 將所有基因分為十組，讀數和(total reads)接近的基因歸為一組，以冪次轉換來處理過度離散情形，若資料的過度離散參數無法以十組來完美描述時，PoissonSeq 的表現將受到影響；edgeR 對所有基因估計一個共同的離散參數，故甚至比 PoissonSeq 的表現還差一些。但樣本數很小時(例如總樣本數為四時)，edgeR 可能會比 PoissonSeq 好，此乃因 PoissonSeq 是仰賴大樣本理論，採用對數線性模式下的分數檢定(score test)來檢定基因的差異表現，故 PoissonSeq 難以在小樣本且過度離散參數變動大時有好的結果。

在過度離散參數變動較大時(如模擬設計一、二)，DESeq2 緊追在表現最好的 Limma 和 baySeq 之後，優於 edgeR，可能的原因為 DESeq2 允許基因有不同的離散參數，與 RNA-seq 資料實際情形較符合，將表現量相似的基因綜合起來，以局

部迴歸(local regression)的方式來估計變異數與期望值之間的關係。在過度離散參數變動較小時(如模擬設計三)，DESeq2 亦有頗好的結果。此外，其計算時間亦短，是除了 Limma 之外值得考慮的方法。

從第四章實證資料分析中，我們發現在 Li 等人(2010)[29]資料上九個方法共同偵測出顯著基因相當少，僅 176 個(總共分析 22,745 個基因)；然而在分析 Marioni (2008)[30]資料時，九方法所共同偵測出的基因多達 3,992 個(總共分析 18,228 個基因)。可能原因為在 Li 等人(2010)的資料中，總樣本數只有四個，對照我們模擬總樣本數為四的結果，此時各方法的表現優劣差異頗大，吾人應避免選擇 SAMseq、TSPM 與 EBSeq 等在樣本數小時表現不佳的方法。惟有審慎評估選擇統計方法，方能有助於 RNA-seq 的資料分析與研究解讀。

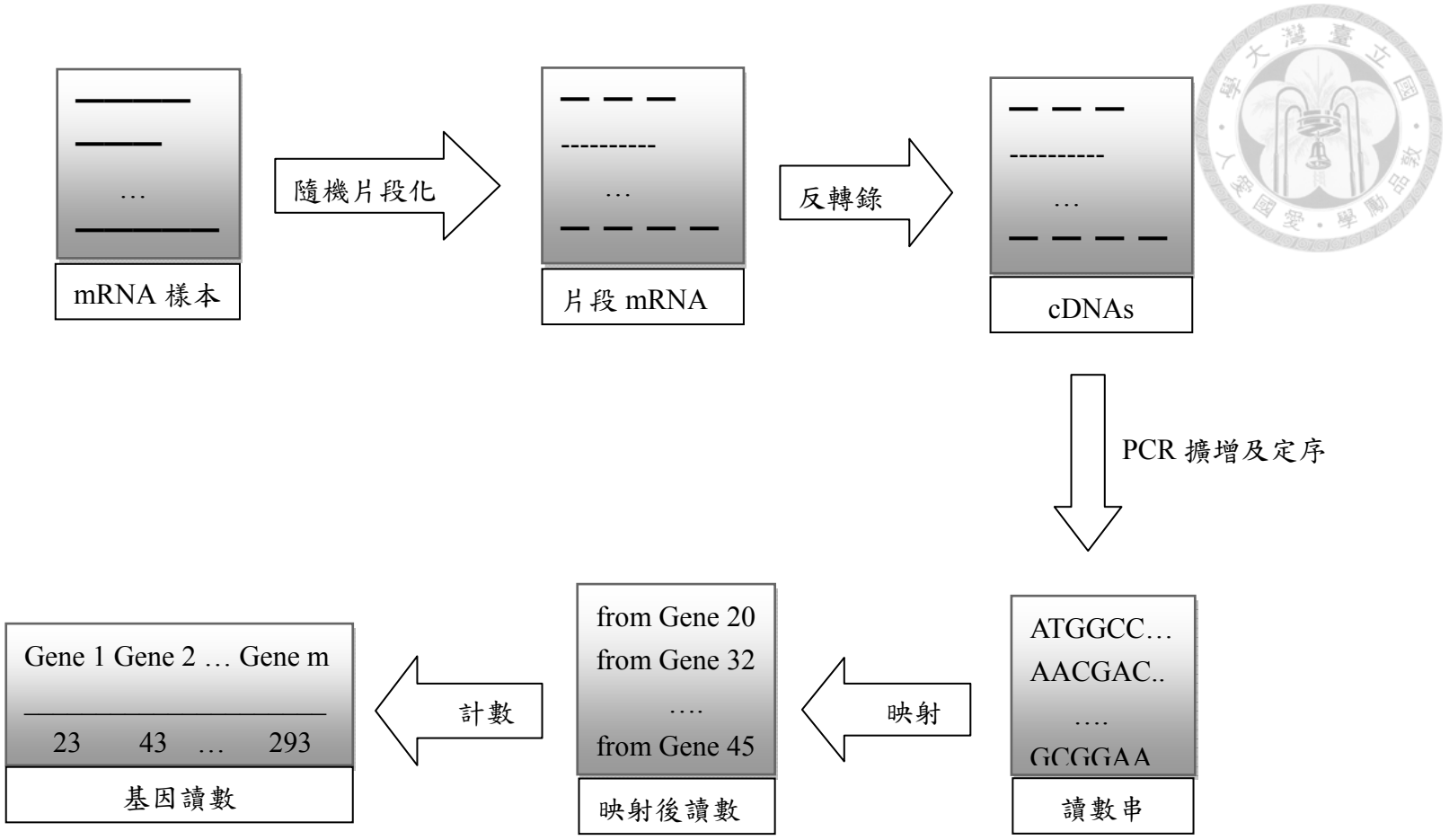
除了第三章以 ROC 曲線來探討各方法在 FPR 較小時的表現狀況，我們亦計算各方法的真實 FDR 值。表十七與表十八列出在模擬設計一，當 FDR 欲控制在 0.05 時，各方法的真實 FDR 值，分別是當過度變異基因比例為 20%與 80%時。真實 FDR 值愈接近名目值 0.05 愈好，baySeq 與 Limma 的真實 FDR 值較接近 0.05，此評比的結果與 ROC 曲線評比的結果一致(baySeq 與 Limma 較佳)。但稍嫌遺憾的是，每個方法都有不同程度的過度宣稱顯著基因(too liberal, not conservative)的趨勢，因各個方法的真實 FDR 值都略高於名目值 0.05。尤其當過度變異基因比例愈高時，過度宣稱顯著基因的趨勢愈加明顯，且此時各方法的 pAUC 呈現系統性的下降趨勢(見表三至表五)。由此可見，過度變異基因比例愈高將愈不利於 RNA-seq 資料分析，至少以現行常用的統計方法皆無法擺脫這樣的限制。

最後，本文的模擬研究有下列幾點限制(limitation)：(1)本文探討範疇為各基因彼此獨立的情況，若考慮基因間彼此相關，情況將更加複雜。(2)本文假設讀數資料為一常數(constant)，實際上，隨著實驗中隨機片段化等過程，每個樣本的讀數資料應來自於一統計分布，而吾人所觀察到的基因讀數只是該統計分布的一個實現值，更加精密複雜的 RNA-seq 統計方法可考慮每個基因的讀數皆源於一個統計

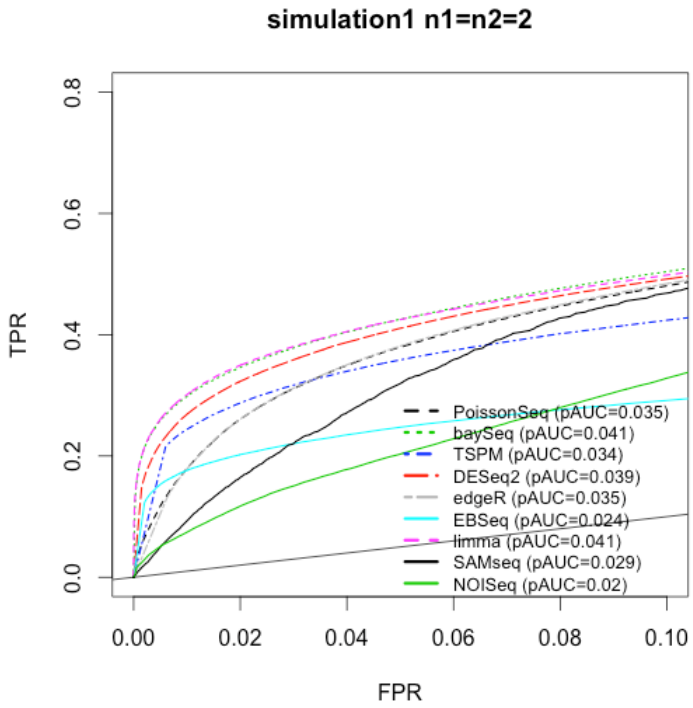
分布。(3)模擬設計三是試著以實證資料裡的效應大小來產生新的次數資料，期使該模擬能更真實地反應實際情況，然而，吾人需先選定一個統計套件來估計 Li 等人(2010)[29]玉米資料中各基因的效應大小，此作法會有球員兼裁判的盲點存在，以何種統計套件來估基因效應大小的確有利於該方法，當吾人改以 DESeq2 來估計各基因的效應大小，edgeR 的表現便退居第二了，僅次於 DESeq2(在 FPR 值低時)。edgeR 仍表現頗佳應是因為模擬設計三中過度離散參數的變動不大(過度離散參數的變異數 0.013 是三個模擬設計裡最低的)。

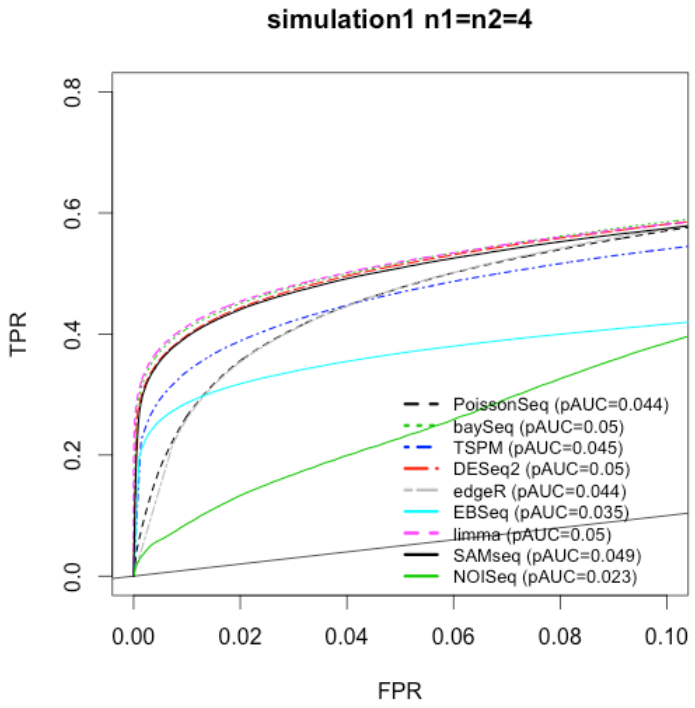
在實際資料分析時，我們建議先計算 RNA-seq 資料過度離散參數的變動程度，再決定適當的統計分析方法。若過度離散參數的變動程度較大(如模擬設計一、二)，Limma 方法和 baySeq 方法的表現最好，其中尤以 Limma 的計算更為迅速；若過度離散參數的變動程度較小(如模擬設計三)，edgeR 方法的表現頗佳。此外，不論過度離散參數的變動程度大小如何，DESeq2 則一直都有不差的表現。

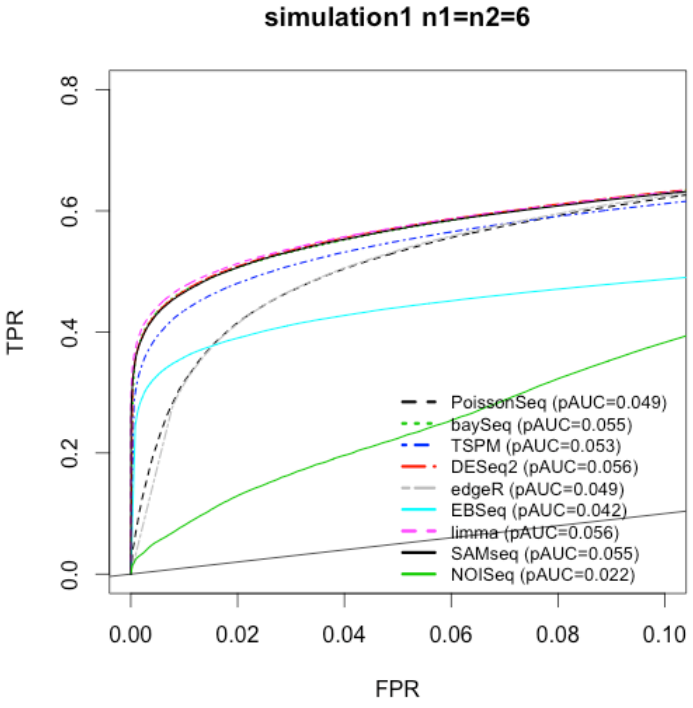
本文模擬多種情境，以 ROC 曲線、pAUC、真實 FDR 值與實際計算時間等四個面向來探討九個常用的 RNA-seq 資料分析方法，雖真實世界的資料更為多樣複雜，仍期望我們的模擬研究能提供研究者一些分析資料時的建議。

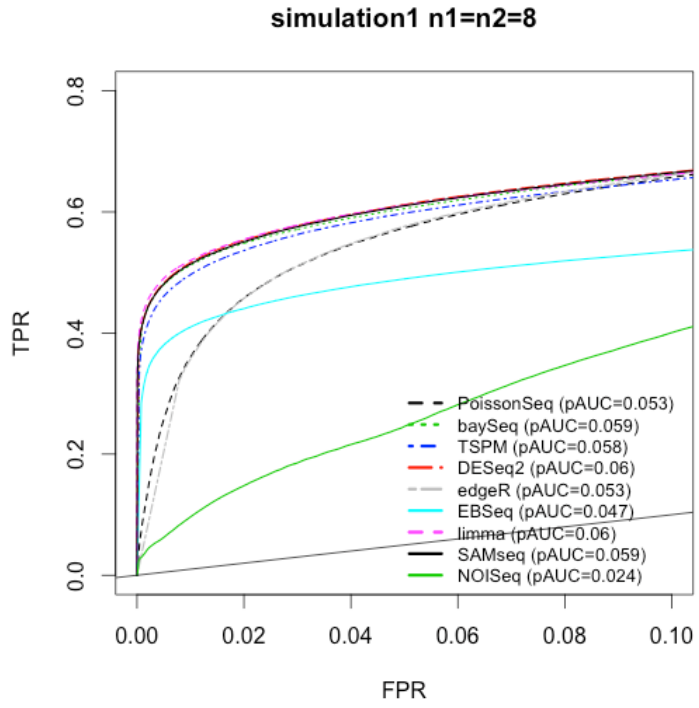


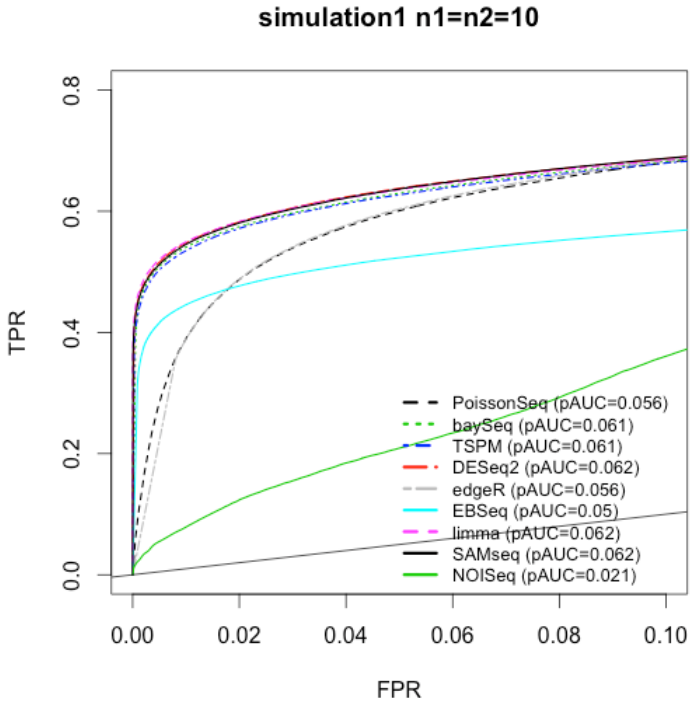

圖一　RNA-seq 實驗流程示意圖



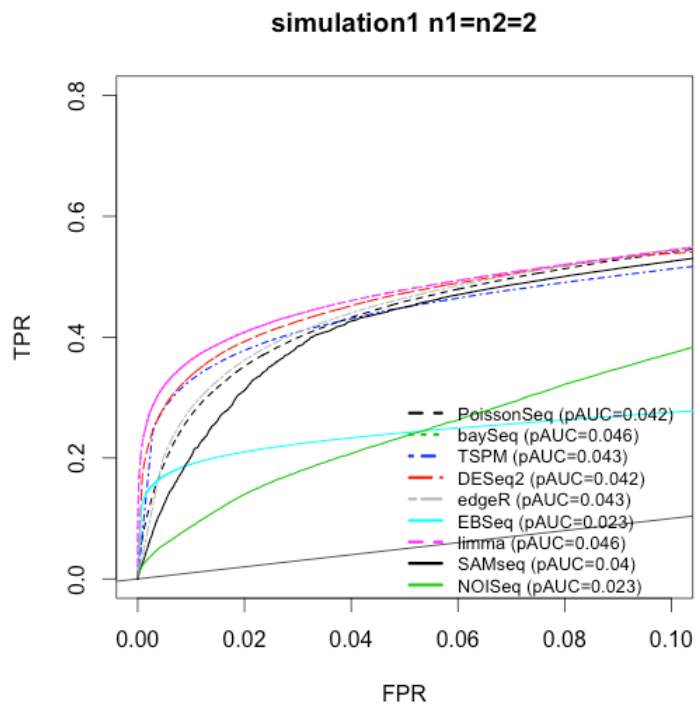



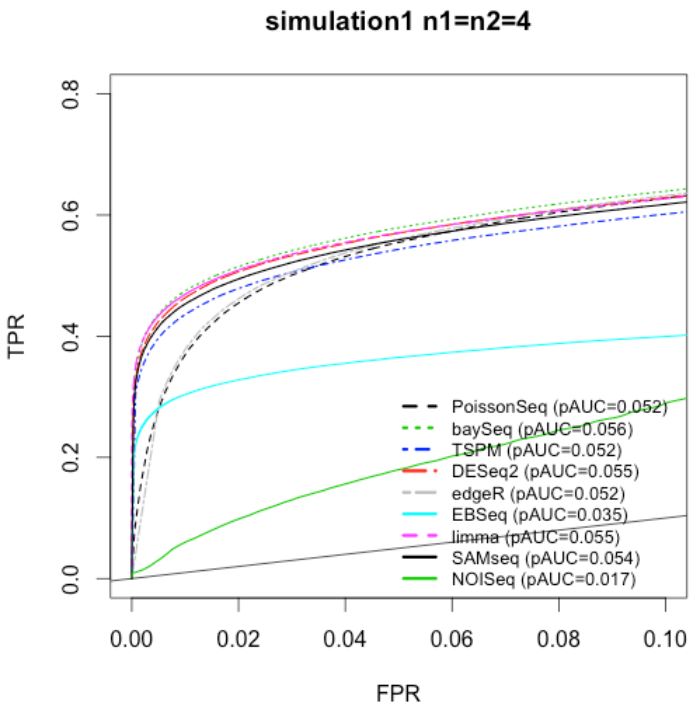



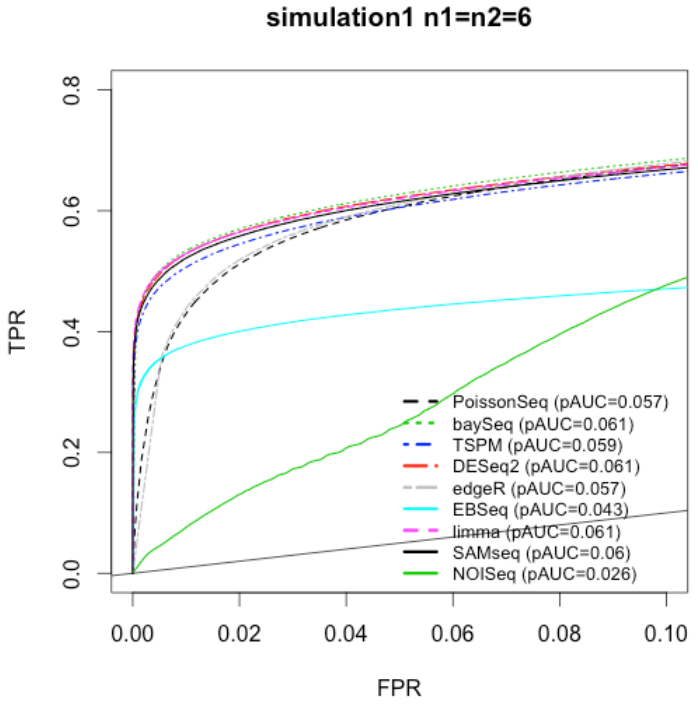



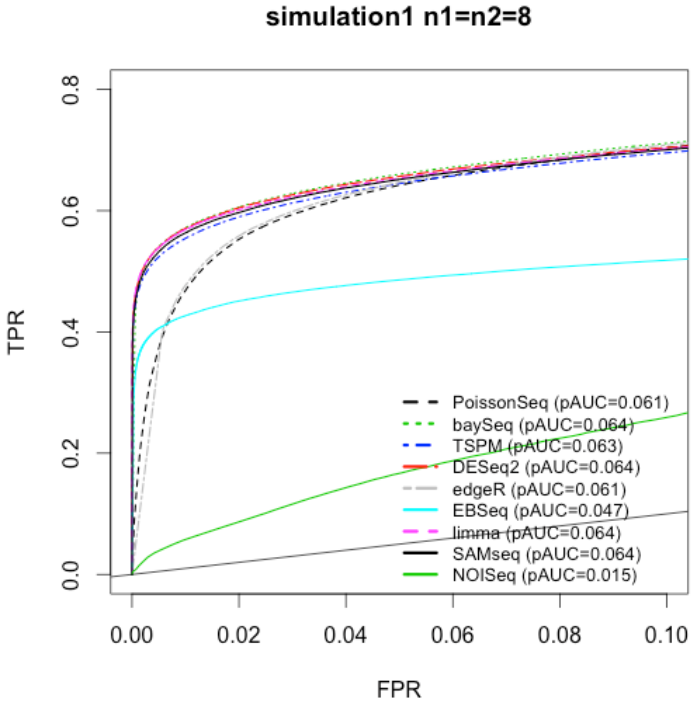



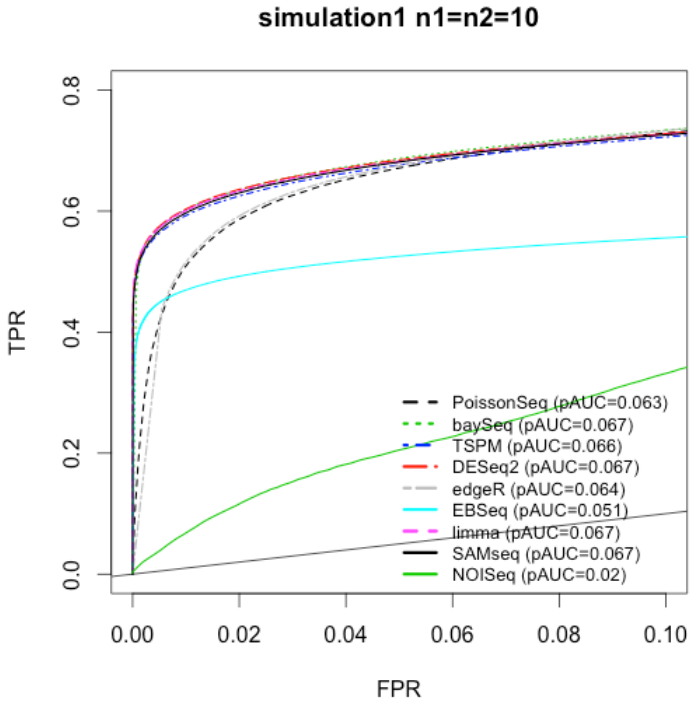

圖二:模擬設定一之各方法檢測差異表現基因 ROC 圖(過度變異基因比例為 50%)。組內樣本數依序為二、四、六、八、十。各圖中下方的灰色直線代表 TPR=FPR。



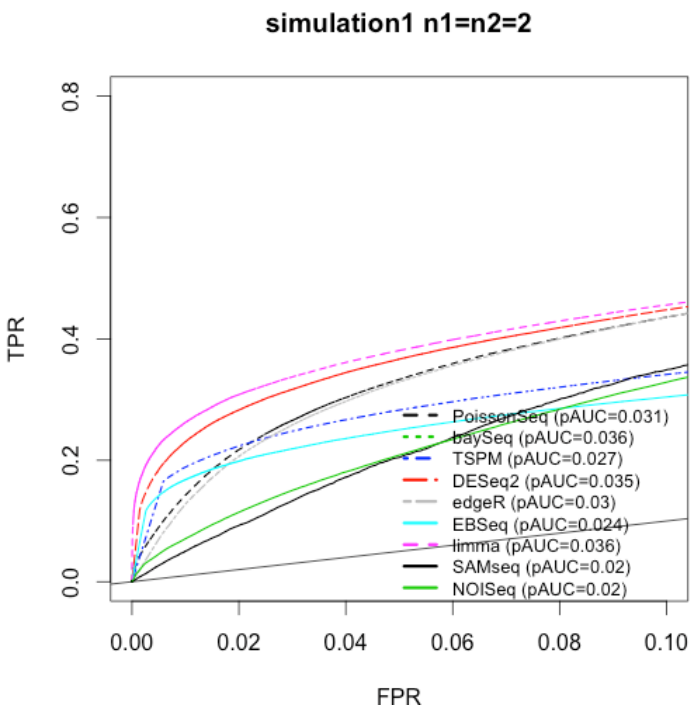



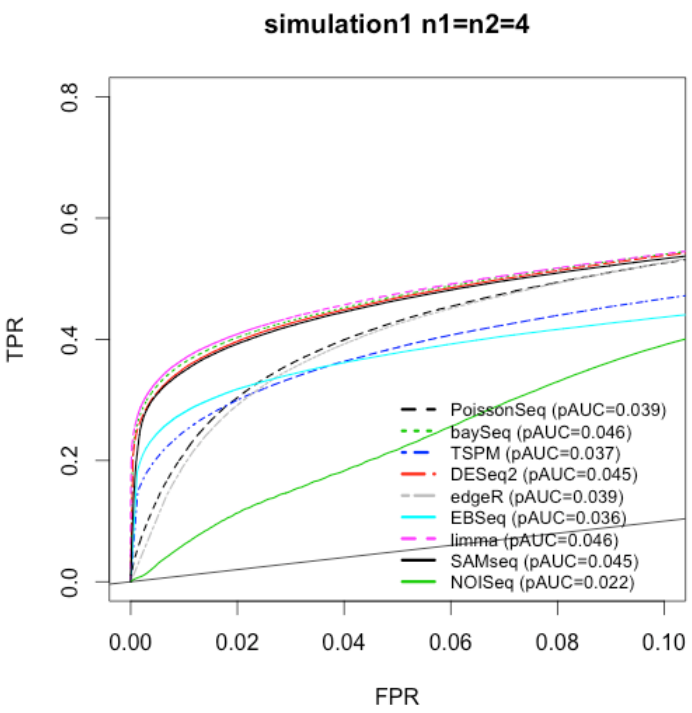



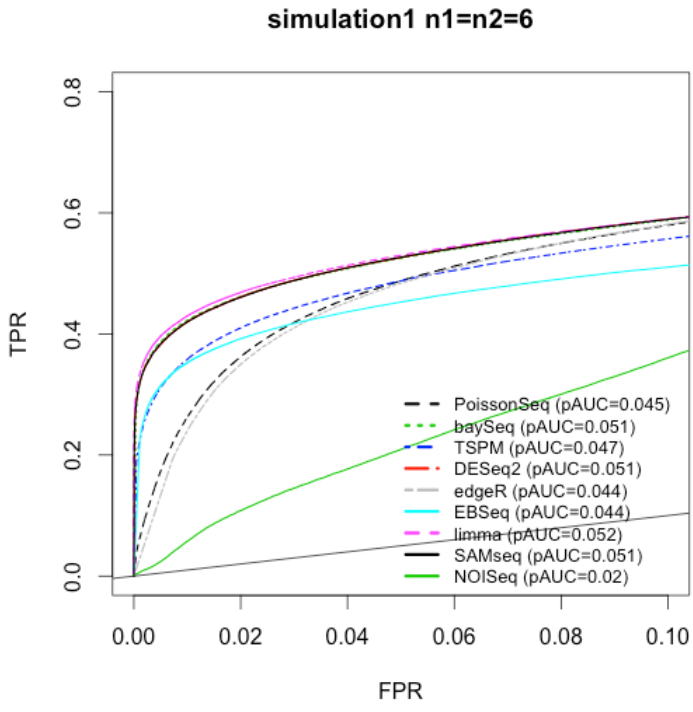



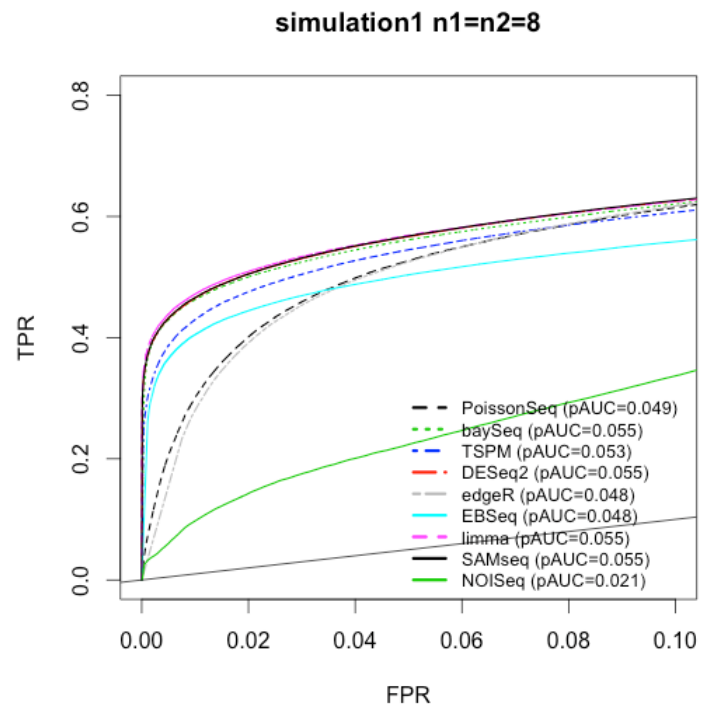



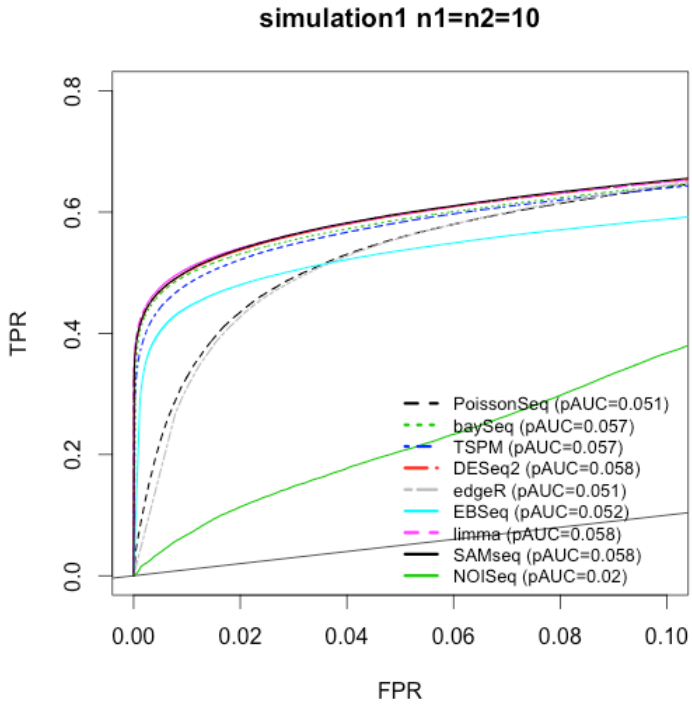

圖三：模擬設定一之各方法檢測差異表現基因 ROC 圖(過度變異基因比例為 20%)。組內樣本數依序為二、四、六、八、十。各圖中下方的灰色直線代表 TPR=FPR。



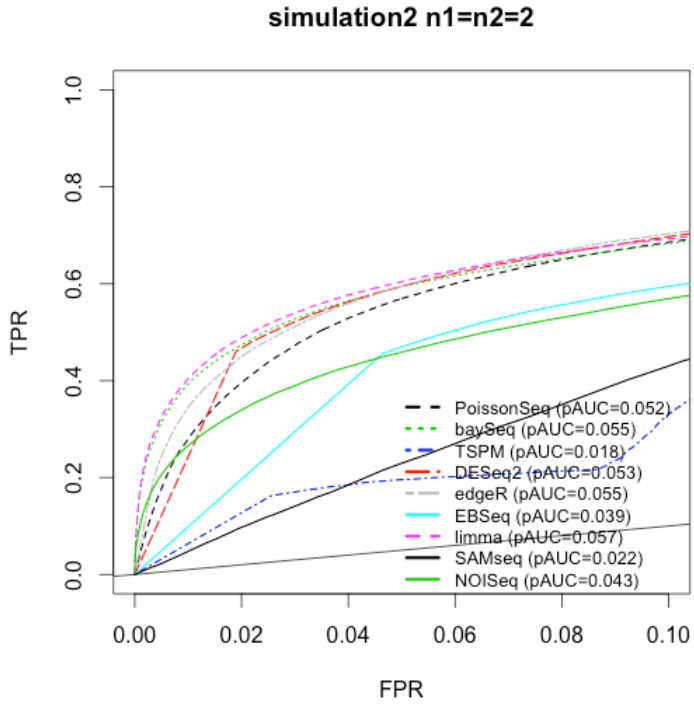



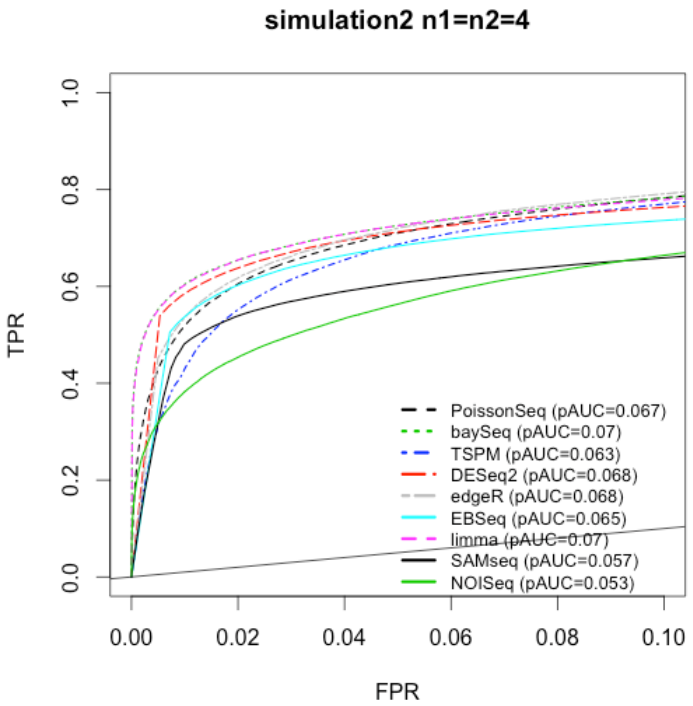



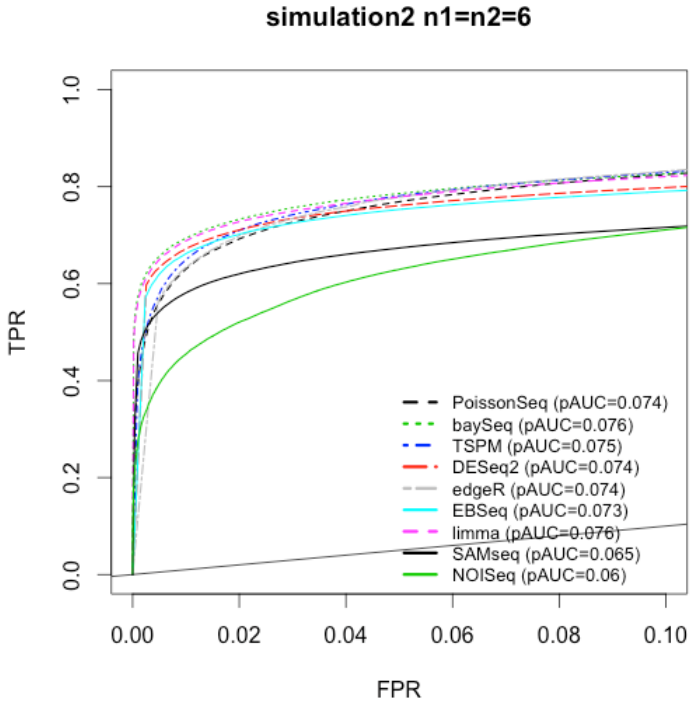



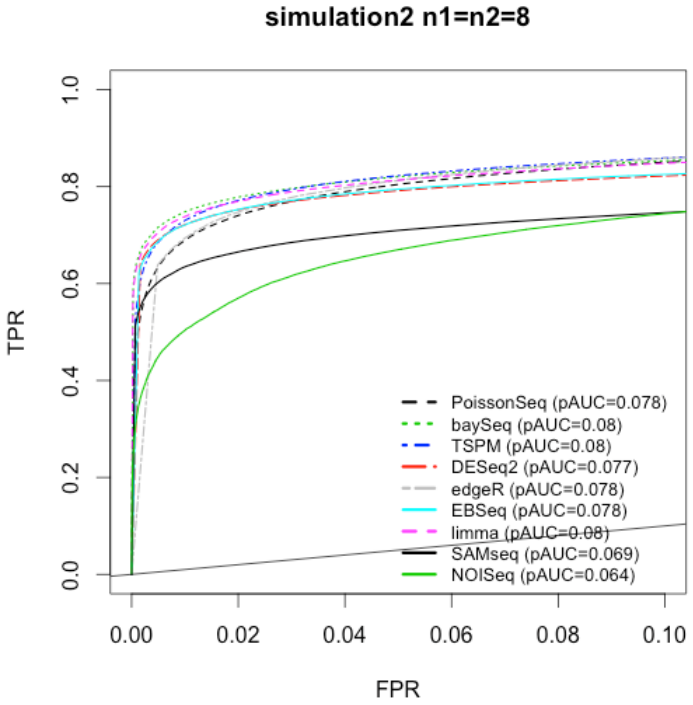



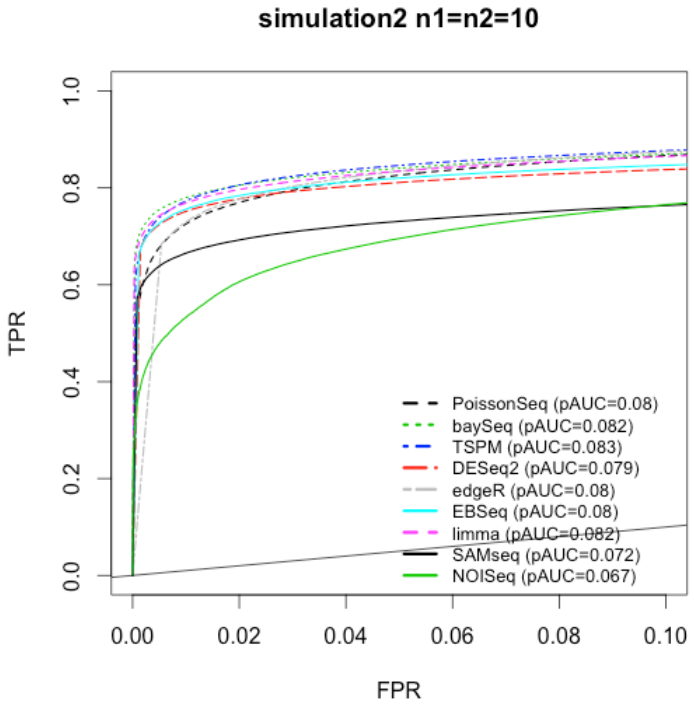

圖四：模擬設定一之各方法檢測差異表現基因 ROC 圖(過度變異基因比例為 80%)。組內樣本數依序為二、四、六、八、十。各圖中下方的灰色直線代表 TPR=FPR。



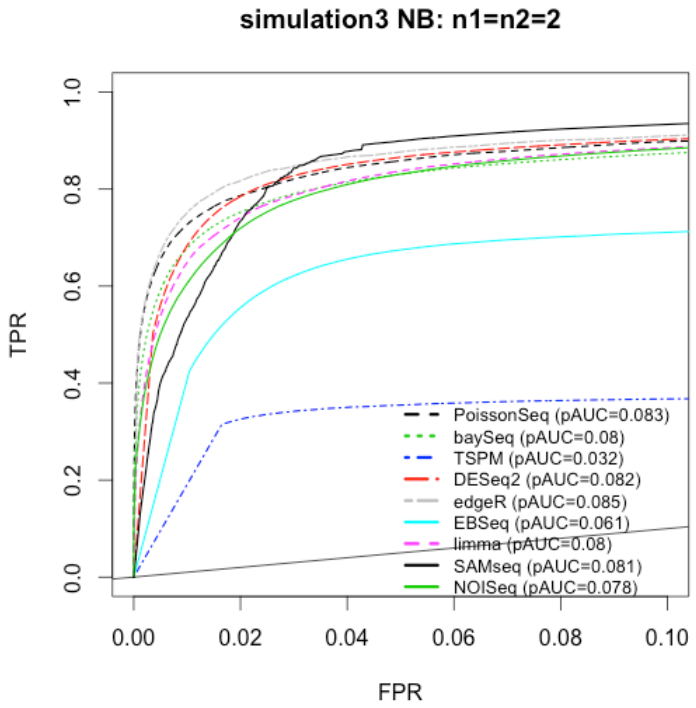



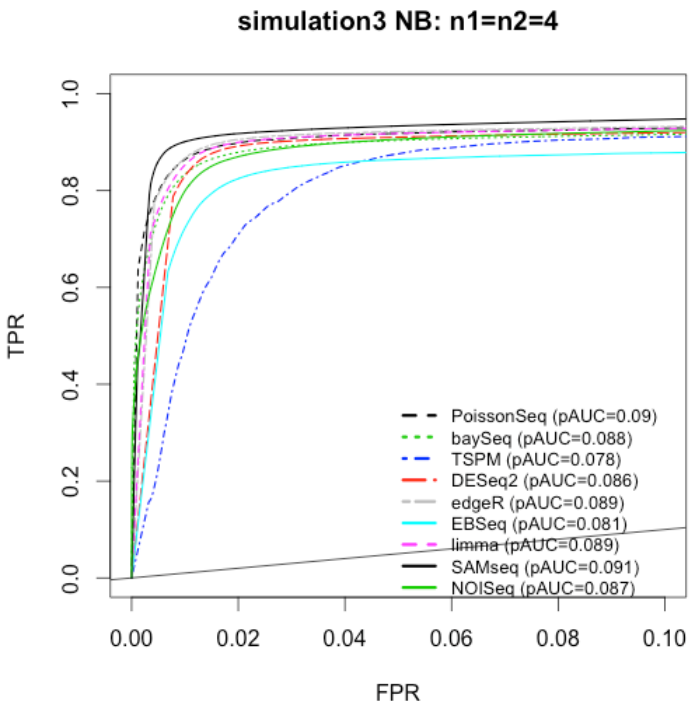



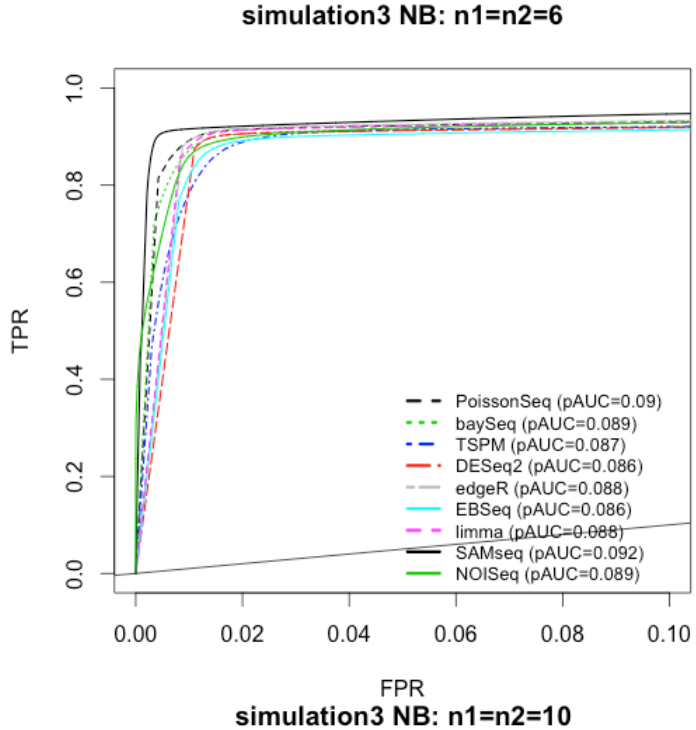



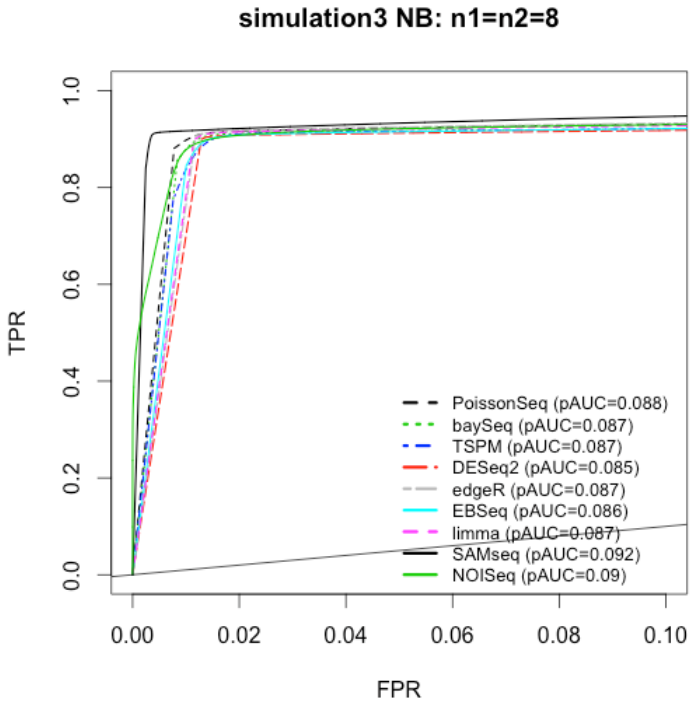



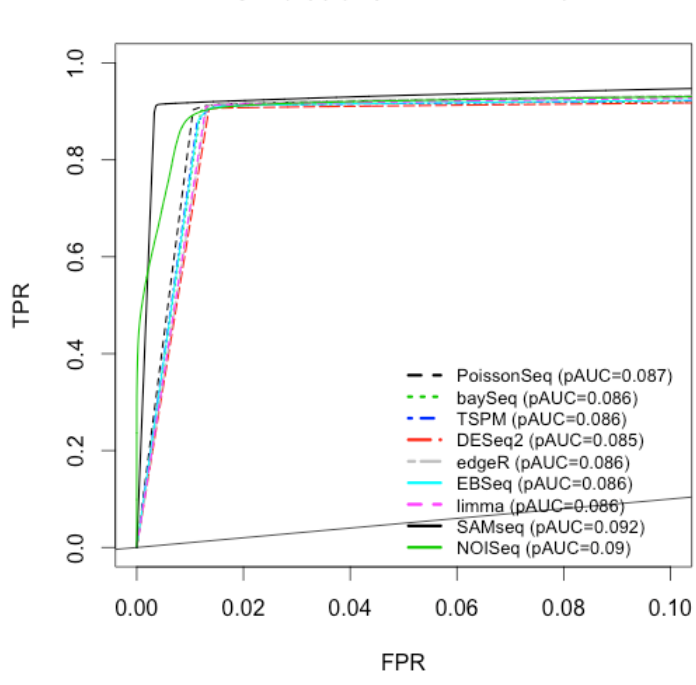

圖五：模擬設定二之各方法檢測差異表現基因 ROC 圖。組內樣本數依序為二、四、六、八、十。各圖中下方的灰色直線代表 TPR=FPR。



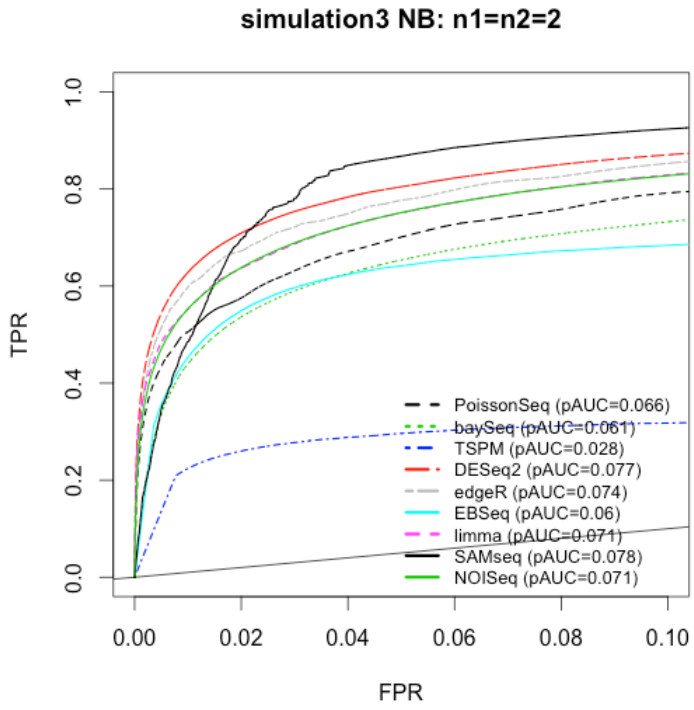



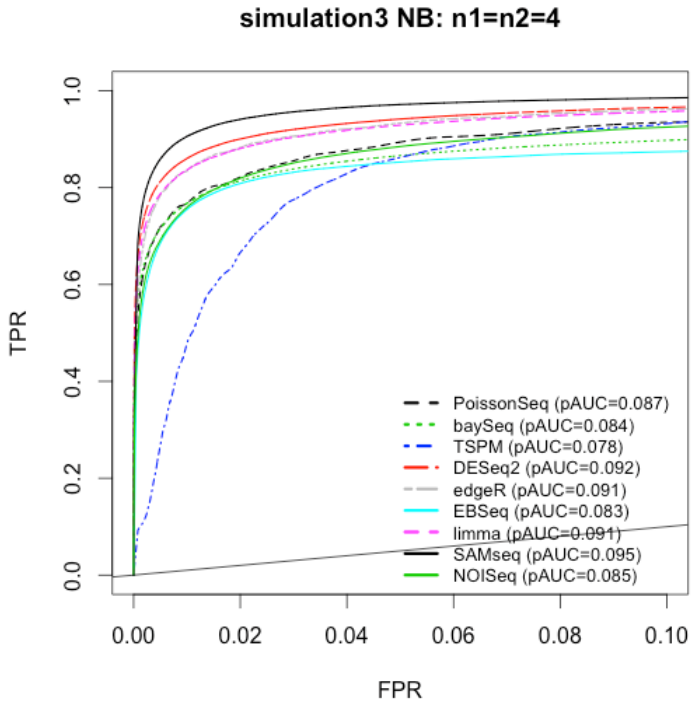



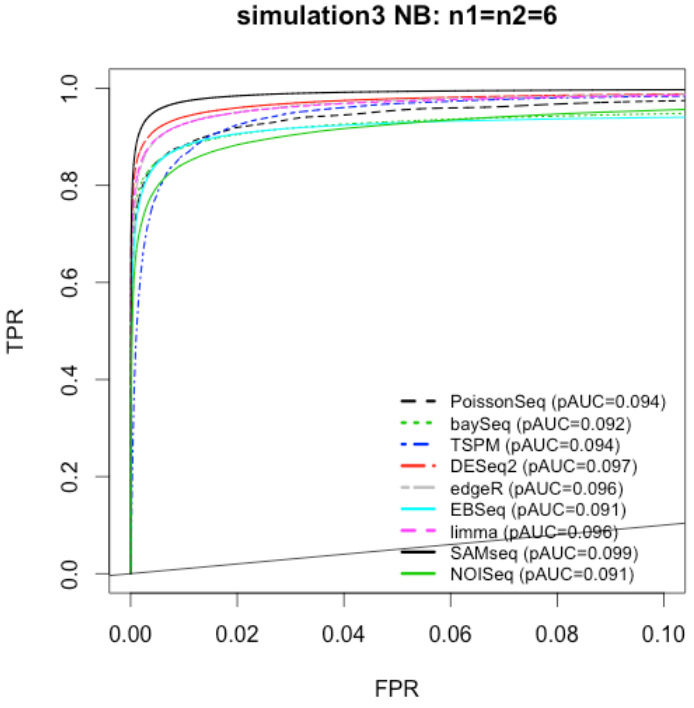



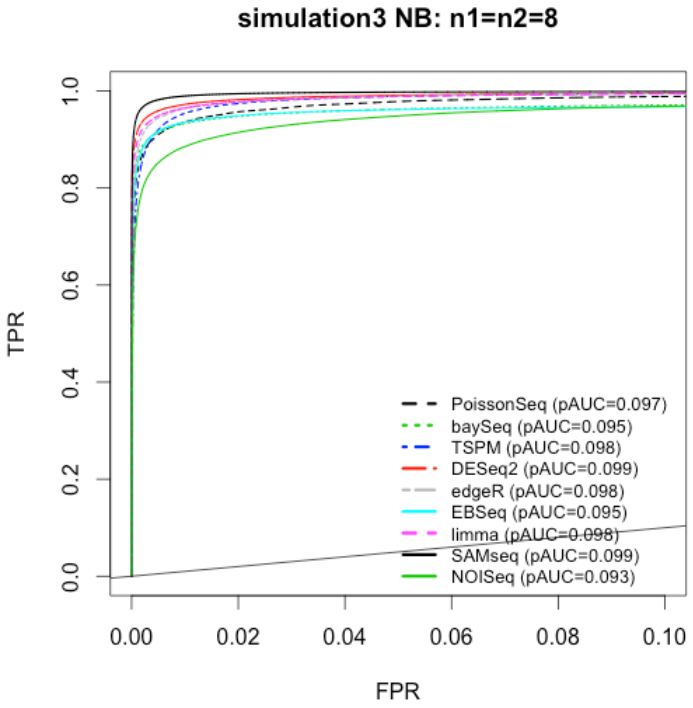



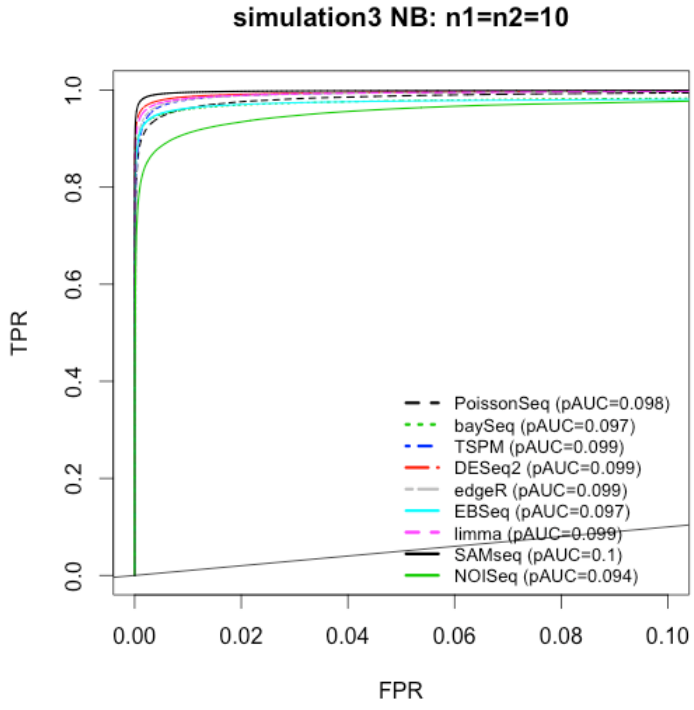

圖六：模擬設定三(當以 edgeR 來估計各基因的效應大小時)之各方法檢測差異表現基因 ROC 圖。組內樣本數依序為二、四、六、八、十。各圖中下方的灰色直線代表 TPR=FPR。




圖七：模擬設定三(當以 DESeq2 來估計各基因的效應大小時)之各方法檢測差異表現基因 ROC 圖。組內樣本數依序為二、四、六、八、十。各圖中下方的灰色直線代表 TPR=FPR。

| 方法 | 版本 | 文獻 | 正規化方式 | 統計分布假設 | 顯著性檢定 | 錯誤發現率控制 |
|---|---|---|---|---|---|---|
| edgeR | 3.4.2 | [1-3] | M 值截尾平均數法 | 負二項分布 | 精確檢定 (Exact test) | BH 校正法 |
| PoissonSeq | 1.1.2 | [12] | 總數正規法改良版 | 卜瓦松分布 | 分數檢定 (score test) | 排列 FDR (permutation-based FDR) |
| DESeq2 | 1.2.10 | [4,5] | 中位數比值正規法 | 負二項分布 | 華德檢定 (Wald's test) | BH 校正法 |
| TSPM | 作者的 R 程式 | [7] | 七十五分位正規法 | 卜瓦松分布 | 概似比檢定 (likelihood-ratio test) | BH 校正法 |
| baySeq | 1.16.0 | [6] | 七十五分位正規法 | 負二項分布 | 經驗貝氏法 | 貝氏 FDR |
| SAMseq | 2.0 | [9] | 總數正規法改良版 | 無分布假設 | 魏克森排序統計量(Wilcoxon rank statistic)，顯著性則由樣本重抽法(resampling)來決定 | 排列 FDR (permutation-based FDR) |
| EBSeq | 1.2.0 | [11] | 中位數比值正規法 | 負二項分布 | 經驗貝氏法 | 貝氏 FDR |
| Limma | 3.18.13 | [10] | M 值截尾平均數法 | voom 轉換 | 經驗貝氏法 | BH 校正法 |
| NOISeq | 2.6.0 | [8] | RPKM 正規法 | 無分布假設 | 以相除和相減的型式來比較兩組在各基因上的讀數是否達顯 | $P\left(\left\|M^*\right\| < \left\|m^g\right\|, D^* < d^g\right)$ ， |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | 著差異(讀數相除和相減之虛無假設下分布由組內資料來建構) | 噪音分布裡比觀察到的 $\left(\left\|m^g\right\|, d^g\right)$ 小的機率 |

**表一　九個方法之資料前置正規化處理、統計分布假設、顯著性檢定統計方法以及錯誤發現率控制法**

**表二　真實狀態與檢定結果交叉表**

| | 不拒絕<br>虛無假設 $H_0$ | 拒絕<br>虛無假設 $H_0$ | |
|---|---|---|---|
| 虛無假設 $H_0$ 為真 | 真陰性(TN) | 偽陽性(FP) | $m_0$ |
| 對立假設 $H_1$ 為真 | 偽陰性(FN) | 真陽性(TP) | $m_1$ |
| | $W$ | $R$ | $m$ |

**表三　模擬設計一各方法的 pAUC (當過度變異基因比例=50%時)**

| **每組樣本數** | **2** | **4** | **6** | **8** | **10** |
|---|---|---|---|---|---|
| PoissonSeq | 0.035 | 0.044 | 0.049 | 0.053 | 0.056 |
| baySeq | 0.041 | 0.050 | 0.055 | 0.059 | 0.061 |
| TSPM | 0.034 | 0.045 | 0.053 | 0.058 | 0.061 |
| edgeR | 0.035 | 0.044 | 0.049 | 0.053 | 0.056 |
| DESeq2 | 0.039 | 0.050 | 0.056 | 0.060 | 0.062 |
| SAMseq | 0.029 | 0.049 | 0.055 | 0.059 | 0.062 |
| Limma | 0.041 | 0.050 | 0.056 | 0.060 | 0.062 |
| EBSeq | 0.024 | 0.035 | 0.042 | 0.047 | 0.050 |
| NOISeq | 0.020 | 0.023 | 0.022 | 0.024 | 0.021 |

**表四　模擬設計一各方法的 pAUC (當過度變異基因比例=20%時)**

| 每組樣本數 | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| PoissonSeq | 0.042 | 0.052 | 0.057 | 0.061 | 0.063 |
| baySeq | 0.046 | 0.056 | 0.061 | 0.064 | 0.067 |
| TSPM | 0.034 | 0.045 | 0.053 | 0.058 | 0.061 |
| edgeR | 0.043 | 0.052 | 0.057 | 0.061 | 0.064 |
| DESeq2 | 0.042 | 0.055 | 0.061 | 0.064 | 0.067 |
| SAMseq | 0.040 | 0.054 | 0.060 | 0.064 | 0.067 |
| Limma | 0.046 | 0.055 | 0.061 | 0.064 | 0.062 |
| EBSeq | 0.023 | 0.035 | 0.043 | 0.047 | 0.051 |
| NOISeq | 0.023 | 0.017 | 0.026 | 0.015 | 0.020 |

**表五　模擬設計一各方法的 pAUC (當過度變異基因比例=80%時)**

| 每組樣本數 | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| PoissonSeq | 0.031 | 0.039 | 0.045 | 0.049 | 0.051 |
| baySeq | 0.036 | 0.046 | 0.051 | 0.055 | 0.057 |
| TSPM | 0.027 | 0.037 | 0.047 | 0.053 | 0.057 |
| edgeR | 0.030 | 0.039 | 0.044 | 0.048 | 0.051 |
| DESeq2 | 0.035 | 0.045 | 0.051 | 0.055 | 0.058 |
| SAMseq | 0.020 | 0.045 | 0.051 | 0.055 | 0.058 |
| Limma | 0.036 | 0.046 | 0.052 | 0.055 | 0.058 |
| EBSeq | 0.024 | 0.036 | 0.044 | 0.048 | 0.052 |
| NOISeq | 0.020 | 0.022 | 0.020 | 0.021 | 0.022 |

**表六　模擬設計二各方法的 pAUC**

| 每組樣本數 | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| PoissonSeq | 0.052 | 0.067 | 0.074 | 0.078 | 0.080 |
| baySeq | 0.055 | 0.070 | 0.076 | 0.080 | 0.082 |
| TSPM | 0.018 | 0.063 | 0.075 | 0.080 | 0.083 |
| edgeR | 0.055 | 0.068 | 0.074 | 0.078 | 0.080 |
| DESeq2 | 0.053 | 0.068 | 0.074 | 0.077 | 0.079 |
| SAMseq | 0.022 | 0.057 | 0.065 | 0.069 | 0.072 |
| Limma | 0.057 | 0.070 | 0.076 | 0.080 | 0.082 |
| EBSeq | 0.039 | 0.065 | 0.073 | 0.078 | 0.080 |
| NOISeq | 0.043 | 0.053 | 0.060 | 0.064 | 0.067 |

**表七　模擬設計三各方法的 pAUC (當以 edgeR 來估計各基因的效應大小時)**

| 每組樣本數 | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| PoissonSeq | 0.083 | 0.090 | 0.090 | 0.088 | 0.087 |
| baySeq | 0.080 | 0.088 | 0.089 | 0.087 | 0.086 |
| TSPM | 0.032 | 0.078 | 0.087 | 0.087 | 0.086 |
| edgeR | 0.085 | 0.089 | 0.088 | 0.087 | 0.086 |
| DESeq2 | 0.082 | 0.086 | 0.086 | 0.085 | 0.085 |
| SAMseq | 0.081 | 0.091 | 0.092 | 0.092 | 0.092 |
| Limma | 0.080 | 0.089 | 0.088 | 0.087 | 0.086 |
| EBSeq | 0.061 | 0.081 | 0.086 | 0.086 | 0.086 |
| NOISeq | 0.078 | 0.087 | 0.089 | 0.090 | 0.090 |

**表八　各模擬設計下效應大小(effect size)與過度離散參數之平均值與變異數**

| | 對數倍數變化絕對值 (僅針對有差異表現的基因) Absolute value of log fold-change for differentially expressed genes | | 過度離散參數(針對所有基因) Overdispersion parameters for all genes | |
|---|---|---|---|---|
| | 平均值 | 變異數 | 平均值 | 變異數 |
| 模擬設計一 | 0.561 | 0.432 | 0.066 | 0.047 |
| 模擬設計二 | 1.920 | 2.1025 | 0.425 | 0.2125 |
| 模擬設計三 | 1.689 | 0.726 | 0.076 | 0.013 |

註：模擬設計一的數據為一百次模擬資料之平均；
模擬設計二的數據為模擬設定之分布理論值；
模擬設計三的數據為實證資料之估計值。

**表九　模擬設計三各方法的 pAUC (當以 DESeq2 來估計各基因的效應大小時)**

| **每組樣本數** | **2** | **4** | **6** | **8** | **10** |
|---|---|---|---|---|---|
| PoissonSeq | 0.066 | 0.087 | 0.094 | 0.097 | 0.098 |
| baySeq | 0.061 | 0.084 | 0.092 | 0.095 | 0.097 |
| TSPM | 0.028 | 0.078 | 0.094 | 0.098 | 0.099 |
| edgeR | 0.074 | 0.091 | 0.096 | 0.098 | 0.099 |
| DESeq2 | 0.077 | 0.092 | 0.097 | 0.099 | 0.099 |
| SAMseq | 0.078 | 0.095 | 0.099 | 0.099 | 0.100 |
| Limma | 0.071 | 0.091 | 0.096 | 0.098 | 0.099 |
| EBSeq | 0.060 | 0.083 | 0.091 | 0.095 | 0.097 |
| NOISeq | 0.071 | 0.085 | 0.091 | 0.093 | 0.094 |

**表十　Li 等人(2010)資料分析結果，FDR=5%**

| | 個別方法偵測出顯著差異基因個數 | 只由該法偵測出顯著差異基因個數(其它八個方法未支持) | 所有方法共同偵測出顯著差異基因個數(所有九個方法皆支持) |
|---|---|---|---|
| PoissonSeq | 4,693 | 545 | 176 |
| baySeq | 1,712 | 0 | |
| TSPM | 4,234 | 1,627 | |
| edgeR | 2,998 | 0 | |
| DESeq2 | 3,866 | 19 | |
| SAMseq | 2,989 | 216 | |
| Limma | 2,272 | 0 | |
| EBSeq | 2,884 | 114 | |
| NOISeq | 1,525 | 0 | |

**表十一　Li 等人(2010)資料分析結果，FDR=5%**

| | PoissinSeq | baySeq | TSPM | edgeR | DESeq2 | SAMseq | Limma | EBSeq | NOISeq |
|---|---|---|---|---|---|---|---|---|---|
| PoissonSeq | **4,693** | 1,696 | 1,987 | 2,928 | 3,209 | 2,240 | 2,148 | 2,272 | 1,504 |
| baySeq | 1,696 | **1,712** | 973 | 1,673 | 1,704 | 1,093 | 1,430 | 1,121 | 952 |
| TSPM | 1,987 | 973 | **4,234** | 1,155 | 1,694 | 1,438 | 824 | 670 | 305 |
| edgeR | 2,928 | 1,673 | 1,155 | **2,998** | 2,832 | 1,669 | 2,150 | 1,982 | 1,497 |
| DESeq2 | 3,209 | 1,704 | 1,694 | 2,832 | **3,866** | 2,138 | 2,246 | 2,462 | 1,408 |
| SAMseq | 2,240 | 1,093 | 1,438 | 1,669 | 2,138 | **2,989** | 1,428 | 1,568 | 757 |
| Limma | 2,148 | 1,430 | 824 | 2,150 | 2,246 | 1,428 | **2,272** | 1,797 | 1,323 |
| EBSeq | 2,272 | 1,121 | 670 | 1,982 | 2,462 | 1,568 | 1,797 | **2,884** | 1,294 |
| NOISeq | 1,504 | 952 | 305 | 1,497 | 1,408 | 757 | 1,323 | 1,294 | **1,525** |

註：對角線記錄在 FDR 控制於 5%時，各方法偵測出顯著差異的基因個數；

非對角線記錄在 FDR 控制於 5%時，任兩個方法所共同偵測出顯著差異的基因個數。

**表十二　Marioni (2008)資料分析結果，FDR=5%**

| 方法 | 個別方法偵測出顯著差異基因個數 | 只由該法偵測出顯著差異基因個數(其它八個方法未支持) | 所有方法共同偵測出顯著差異基因個數(所有九個方法皆支持) |
|---|---|---|---|
| PoissonSeq | 13,239 | 342 | 3,992 |
| baySeq | 11,473 | 0 | |
| TSPM | 12,227 | 0 | |
| edgeR | 11,643 | 0 | |
| DESeq2 | 11,415 | 0 | |
| SAMseq | 14,104 | 706 | |
| Limma | 11,963 | 1 | |
| EBSeq | 6,319 | 0 | |
| NOISeq | 5,690 | 0 | |

**表十三　Marioni (2008)資料分析結果，FDR=5%**

| | PoissinSeq | baySeq | TSPM | edgeR | DESeq2 | SAMseq | Limma | EBSeq | NOISeq |
|---|---|---|---|---|---|---|---|---|---|
| PoissonSeq | **13,239** | 11,109 | 11,720 | 11,101 | 10,955 | 12,846 | 11,414 | 6,136 | 5,317 |
| baySeq | 11,109 | **11,473** | 11,420 | 11,210 | 11,182 | 11,458 | 11,296 | 6,312 | 5,643 |
| TSPM | 11,720 | 11,420 | **12,227** | 11,535 | 11,378 | 12,194 | 11,663 | 6,311 | 5,670 |
| edgeR | 11,101 | 11,210 | 11,535 | **11,643** | 11,291 | 11,633 | 11,445 | 6,298 | 5,685 |
| DESeq2 | 10,955 | 11,182 | 11,378 | 11,291 | **11,415** | 11,406 | 11,292 | 6,313 | 5,663 |
| SAMseq | 12,846 | 11,458 | 12,194 | 11,633 | 11,406 | **14,104** | 11,932 | 6,314 | 5,663 |
| Limma | 11,414 | 11,296 | 11,663 | 11,445 | 11,292 | 11,932 | **11,963** | 6,306 | 5,685 |
| EBSeq | 6,136 | 6,312 | 6,311 | 6,298 | 6,313 | 6,314 | 6,306 | **6,319** | 4,172 |
| NOISeq | 5,317 | 5,643 | 5,670 | 5,685 | 5,663 | 5,690 | 5,685 | 4,172 | **5,690** |

註：對角線記錄在 FDR 控制於 5%時，各方法偵測出顯著差異的基因個數；

非對角線記錄在 FDR 控制於 5%時，任兩個方法所共同偵測出顯著差異的基因個數。

表十四　模擬設計一平均每次計算時間(單位：秒)

| 每組樣本數 | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| PoissonSeq | 2.1 | 3.3 | 4.6 | 5.4 | 6.3 |
| baySeq | 2121.4 | 2549.8 | 3203.0 | 3595.0 | 3911.2 |
| TSPM | 101.4 | 109.8 | 110.6 | 141.5 | 135.0 |
| edgeR | 1.5 | 2.3 | 3.4 | 4.3 | 5.1 |
| DESeq2 | 7.5 | 8.1 | 9.5 | 11.1 | 13.4 |
| SAMseq | 4.3 | 11.6 | 14.9 | 19.0 | 21.2 |
| Limma | 1.7 | 1.6 | 1.7 | 1.7 | 1.8 |
| EBSeq | 9.4 | 20.6 | 27.1 | 36.4 | 44.8 |
| NOISeq | 5.9 | 22.3 | 57.0 | 101.1 | 111.5 |

表十五　模擬設計二平均每次計算時間(單位：秒)

| 每組樣本數 | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| PoissonSeq | 1.5 | 2.7 | 3.4 | 4.1 | 5.1 |
| baySeq | 1805.4 | 2456.6 | 2384.0 | 2563.0 | 3062.4 |
| TSPM | 91.3 | 113.3 | 103.4 | 109.1 | 123.6 |
| edgeR | 2.2 | 2.1 | 2.8 | 3.3 | 4.0 |
| DESeq2 | 6.7 | 7.1 | 8.4 | 9.5 | 12.4 |
| SAMseq | 3.6 | 10.4 | 14.4 | 18.0 | 21.3 |
| Limma | 1.3 | 1.4 | 1.6 | 1.6 | 1.8 |
| EBSeq | 13.1 | 24.4 | 33.7 | 43.7 | 50.5 |
| NOISeq | 8.0 | 30.8 | 65.3 | 116.8 | 124.3 |

表十六　模擬設計三平均每次計算時間(單位：秒)

| 每組樣本數 | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| PoissonSeq | 3.8 | 9.4 | 13.1 | 17.0 | 22.8 |
| baySeq | 2668.4 | 5580.4 | 7898.3 | 10369.4 | 11362.6 |
| TSPM | 278.5 | 338.3 | 326.8 | 337.8 | 344.9 |
| edgeR | 3.8 | 7.0 | 10.2 | 12.8 | 15.4 |
| DESeq2 | 15.1 | 20.3 | 26.2 | 30.9 | 38.1 |
| SAMseq | 18.1 | 59.6 | 82.9 | 102.3 | 126.0 |
| Limma | 7.0 | 7.8 | 8.6 | 9.3 | 9.8 |
| EBSeq | 31.6 | 70.5 | 107.0 | 152.4 | 248.2 |
| NOISeq | 29.3 | 131.8 | 337.6 | 623.1 | 693.4 |

表十七　模擬設計一各方法的真實 FDR 值(當過度變異基因比例=20%時)　(FDR 控制值=0.05)

| 每組樣本數 | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| PoissonSeq | 0.097 | 0.094 | 0.090 | 0.082 | 0.065 |
| baySeq | 0.078 | 0.072 | 0.075 | 0.062 | 0.054 |
| TSPM | 0.123 | 0.096 | 0.093 | 0.082 | 0.080 |
| edgeR | 0.099 | 0.102 | 0.084 | 0.082 | 0.068 |
| DESeq2 | 0.088 | 0.084 | 0.086 | 0.074 | 0.061 |
| SAMseq | 0.087 | 0.092 | 0.090 | 0.085 | 0.074 |
| Limma | 0.078 | 0.070 | 0.068 | 0.070 | 0.062 |
| EBSeq | 0.086 | 0.090 | 0.086 | 0.075 | 0.070 |
| NOISeq | 0.130 | 0.113 | 0.094 | 0.090 | 0.081 |

**表十八　模擬設計一各方法的真實 FDR 值 (當過度變異基因比例=80%時)**
**(FDR 控制值=0.05)**

| 每組樣本數 | 2 | 4 | 6 | 8 | 10 |
|---|---|---|---|---|---|
| PoissonSeq | 0.102 | 0.098 | 0.094 | 0.096 | 0.070 |
| baySeq | 0.080 | 0.078 | 0.079 | 0.070 | 0.060 |
| TSPM | 0.134 | 0.111 | 0.098 | 0.087 | 0.084 |
| edgeR | 0.113 | 0.099 | 0.097 | 0.092 | 0.089 |
| DESeq2 | 0.099 | 0.093 | 0.088 | 0.080 | 0.072 |
| SAMseq | 0.093 | 0.090 | 0.093 | 0.086 | 0.080 |
| Limma | 0.080 | 0.072 | 0.069 | 0.073 | 0.066 |
| EBSeq | 0.090 | 0.095 | 0.089 | 0.077 | 0.074 |
| NOISeq | 0.145 | 0.132 | 0.113 | 0.094 | 0.092 |

# 參考文獻


1. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-140.

2. Robinson MD, Smyth GK (2007) Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23: 2881-2887.

3. Robinson MD, Smyth GK (2008) Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 9: 321-332.

4. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106.

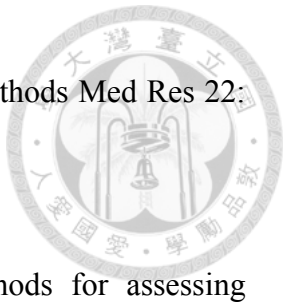
5. Love MI, Anders S, Huber W (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. bioRxiv.

6. Hardcastle TJ, Kelly KA (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11: 422.

7. Auer PL, Doerge RW (2011) A two-stage Poisson model for testing RNA-Seq data. Statistical Applications in Genetics and Molecular Biology 10: 1-26.

8. Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A (2011) Differential expression in RNA-seq: a matter of depth. Genome Res 21: 2213-2223.

9. Li J, Tibshirani R (2013) Finding consistent patterns: a nonparametric approach for



identifying differential expression in RNA-Seq data. Stat Methods Med Res 22: 519-536.

10. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3: Article 3.

11. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, et al. (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics 29: 1035-1043.

12. Li J, Witten DM, Johnstone IM, Tibshirani R (2012) Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics 13: 523-538.

13. Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11: 94.

14. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11: R25.

15. Kvam VM, Liu P, Si Y (2012) A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am J Bot 99: 248-256.

16. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and

quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621-628.

17. Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear models, and the Gauss—Newton Method. Biometrika 61: 439-447.

18. Lin WY, Lee WC (2012) Presenting the uncertainties of odds ratios using empirical-Bayes prediction intervals. PLoS One 7: e32022.

19. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc B 57: 289-300.

20. Tusher V, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. Proceedings of the National Academy of Sciences of the United States of America 98: 5116-5121.

21. Storey JD (2002) A direct approach to false discovery rates. J R Stat Soc B 64: 479-498.

22. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440-9445.

23. Xie Y, Pan W, Khodursky AB (2005) A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. Bioinformatics 21: 4280-4288.

24. Lin WY, Lee WC (2010) Incorporating prior knowledge to facilitate discoveries in a genome-wide association study on age-related macular degeneration. BMC Res

Notes 3: 26.

25. Jiang Y, Metz CE, Nishikawa RM (1996) A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 201: 745-750.

26. McClish DK (1989) Analyzing a portion of the ROC curve. Med Decis Making 9: 190-195.

27. Thompson ML, Zucchini W (1989) On the statistical analysis of ROC curves. Stat Med 8: 1277-1290.

28. Xu X, Zhang Y, Williams J, Antoniou E, McCombie WR, et al. (2013) Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets. BMC Bioinformatics 14 Suppl 9: S1.

29. Li P, Ponnala L, Gandotra N, Wang L, Si Y, et al. (2010) The developmental dynamics of the maize leaf transcriptome. Nat Genet 42: 1060-1067.

30. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18: 1509-1517.