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

Institute of Epidemiology and Preventive Medicine

College of Public Health

National Taiwan University

Master Thesis

以非常態情境評估基因集合分析方法在真實基因資料下之表現研究

Statistical Evaluation for Methods of Gene-set Analysis with Multivariate Non-normal Scenarios

何奇軒

Chi-Hsuan Ho

指導教授：蕭朱杏 博士

Advisor: Chuhsing Kate Hsiao, Ph.D.

中華民國107年6月

June, 2018

**致謝**

隨著時間的過去，研究所生涯也邁入尾聲，回想起在學校這兩年來充實的時光裡，首先我要先感謝指導老師蕭朱杏教授，感謝蕭老師在研究的過程中耐心地給予指導以及協助，使我在生物資訊以及生物統計領域的專業知識以及研究方法上能夠更上一層。除此之外，老師總是在我面對問題時給予我很大的鼓勵以及支持，讓我能夠擁有更強大的自信心來面對在研究中所遇到的瓶頸，並將問題順利地解決。另外，我也想感謝盧子彬老師和曾仲瑩老師，兩位老師在所上分別開設的計算生物學原理與應用以及遺傳資料統計分析課程都使我在程式編輯能力以及統計分析方法方面有著很大的進步。而在口試期間，要特別感謝農藝所的蔡政安老師以及所上的洪弘老師能於百忙之中撥冗參加學生的學位口試，並且提供寶貴的建議，使我們的論文研究能夠更加地面面俱到。

接下來我想先感謝實驗室以及研究室的夥伴們，特別感謝書如學姊和昌憲學長在研究上對我的幫忙，同時也要感謝筱琪、奇祐、宏卿和煜鈞在研究的道路上給予我很多的意見，有你們在的實驗室總是充滿著無數笑聲以及那種渴望知識的熱情。另外，我也要感謝我的好友景榮，很高興能在碩士生涯中遇到你，不管是在修課或是在研究方面上，我們都能夠互相打氣或是彼此分享研究想法，克服困難。

最後，我要感謝父母及家人在背後的鼓勵以及支持，因為有你們的幫忙我才能夠心無旁騖地完成我的研究，順利地邁入我人生中的下一個階段，而我也必定會將這兩年來所學到的東西以及經歷的一切運用到未來的挑戰上，不辜負你們的期望。另外，特別感謝科技部的資助，使得本研究能夠順利完成！

**中文摘要**

隨著科技的進步，有越來越多的統計方法能夠幫助研究者找出致病的生物路徑或是基因，因此如何評估並且有效率地選擇這些方法來進行後續更進一步地研究便成為一個關鍵的步驟。在過去的研究中，研究者大多利用多維度(對數)常態分佈來評估這些統計方法，然而這個方式是否恰當仍具有很大的爭議性，因此，本論文的第一部分將會著重在mRNA基因表現量資料上，我們會從公開網站上蒐集各個疾病的資料集以及相關的生物路徑資訊，並且挑選四種常態性檢定(Mardia’s test, Henze-Zirkler’s test, Royston’s test, One-sample energy test)來探討這些經過正規化處理後的資料集是否符合常態性假設。而從這一部分的結果中可以得知，正規化後的真實基因資料有很高的機率並非分服從多維度常態分佈，也因此，本論文的第二個部分挑選了五種常用的基因集合分析方法(Hotelling’s T2, Global test, GlobalANCOVA, Energy test (N-statistic), GSEA (Category))來探討這些方法在非常態情境底下的表現。在這個過程中，我們利用多維度t分佈以及多維度常態混合分佈來設計一系列的多維度非常態情境，並且藉由這些分析方法在這幾種情境底下的穩健性來評估他們的好壞。而從實驗結果中可以發現，雖然大多數的統計方法在非常態情境底下的表現都不好，但Hotelling’s T2在不同情境的某些特定情況底下卻仍然擁有良好以及穩定的表現，然而這些結果都是無法從傳統多維度常態的模擬方式裡獲得的。因此，總結來說，為了要得到更可靠、更準確的評估資訊，我們建議之後的研究者在模擬階段時可以加入一些非常態情境以及其他細微的設定，並且利用各個情境下整體的穩健性來評估這些方法。最後，本論文採用雷達圖來將上述的多個資訊彙整，以提供研究者一個更清楚的視覺化方式來了解這些方法的表現。

關鍵字：基因集合分析、多維度非常態分佈、mRNA基因表現量、真實基因資料常態性、穩健性、雷達圖、模擬方法

**ABSTRACT**

As the technology improves, more and more statistical methods for gene-set analysis (GSA) are developed to find pathogenic pathways and genes. Thus, finding a suitable method to make further analysis becomes a critical procedure. In recent years, many studies use the multivariate normal distribution or multivariate lognormal distribution in simulation studies to evaluate the performance of these GSA methods. However, the normality assumption for the gene expression data has been questionable. Therefore, the first part of our study focus on the normality of mRNA gene expression data. We first collect the corresponding pathway information and the gene expression data for each cancer subtype from public website. Then, we choose four normality tests (Mardia’s test, Henze-Zirkler’s test, Royston’s test, One-sample energy test) to analyze these real data, and the results show it is very possible that the normalized gene expression data are not multivariate normally distributed. Thus, in the second part of our study, we consider five GSA methods (Hoteling’s T2, Global test, GlobalANCOVA, Energy Test, GSEA (Category)) in some multivariate non-normal scenarios (including multivariate t distribution and mixture of multivariate normal distributions) to compare the performance and the robustness of these statistical methods. The results of our experiments indicate that although the majority of these GSA methods have a very poor performance under the non-normal scenarios, surprisingly, Hoteling’s T2 still has a consistent and overwhelmingly good performance under different scenarios with some special settings. However, these results cannot be learned from the traditional multivariate normal simulation methods. Thus, to sum up, to get a more reliable and accurate information, we suggest that researchers should add some non-normal scenarios and other settings to the simulation study before using the robustness to evaluate these methods. Finally, these results are demonstrated with radar plots to visualize all the experimental outcomes more clearly.

Keywords: gene-set analysis, multivariate non-normal distribution, mRNA gene expression, the normality of real data, robustness, radar plot, simulation methods

**目錄**

[第一章、 研究背景 1](#_Toc517453820)

[第二章、 真實基因資料常態性假設之分析 5](#_Toc517453821)

[第一節 資料背景 5](#_Toc517453822)

[第二節 資料前處理以及資料篩選 7](#_Toc517453823)

[正規化： 7](#_Toc517453824)

[生物路(途)徑(pathway)的選擇： 7](#_Toc517453825)

[樣本臨床資料的整理： 8](#_Toc517453826)

[第三節 常態性檢定 9](#_Toc517453827)

[第四節 常態性檢定模擬以及其結果 10](#_Toc517453828)

[模擬設定: 11](#_Toc517453829)

[模擬結果: 12](#_Toc517453830)

[第五節 真實基因資料分析結果 14](#_Toc517453831)

[第三章、 基因集合分析方法的模擬比較 17](#_Toc517453832)

[第一節 模擬設定 17](#_Toc517453833)

[差異方式A—每個基因的邊際分佈在不同的兩種表現型間有相同的平移差異： 18](#_Toc517453834)

[差異方式B—並非每個基因在不同的兩種表現型間都有相同程度的平移差異： 21](#_Toc517453835)

[第二節 基因集合分析方法 22](#_Toc517453836)

[方法一：Hotelling’s T2 23](#_Toc517453837)

[方法二：The Global Test 23](#_Toc517453838)

[方法三：The GlobalANCOVA 24](#_Toc517453839)

[方法四：Energy Test (N-statistic) 26](#_Toc517453840)

[方法五：GSEA (Category) 26](#_Toc517453841)

[第三節 分析結果 28](#_Toc517453842)

[第四章、 討論 32](#_Toc517453843)

[第五章、 參考文獻 35](#_Toc517453844)

[附錄一、常態性檢定 96](#_Toc517453845)

[附錄二、其他附錄資料 101](#_Toc517453846)

**圖目錄**

[圖一 乳癌資料集常態性檢定分析結果(點圖) 82](#_Toc517703890)

[圖二 大腸直腸癌資料集常態性檢定分析結果(點圖) 83](#_Toc517703891)

[圖三 白血病資料集常態性檢定分析結果(點圖) 84](#_Toc517703892)

[圖四 肺癌資料集(GPL96)常態性檢定分析結果(點圖) 85](#_Toc517703893)

[圖五 肺癌資料集(GPL570)常態性檢定分析結果(點圖) 86](#_Toc517703894)

[圖六 卵巢癌資料集常態性檢定分析結果(點圖) 87](#_Toc517703895)

[圖七 GSA方法在六種不同情境底下對型一誤差率(0.05)的估計表現(雷達圖) 88](#_Toc517703896)

[圖八 GSA方法在六種不同情境底下的統計檢定力表現(雷達圖)– 89](#_Toc517703897)

[圖九 GSA方法在六種不同情境底下的統計檢定力表現(雷達圖)– 90](#_Toc517703898)

[圖十 GSA方法在六種不同情境底下的統計檢定力表現(雷達圖)– 91](#_Toc517703899)

[圖十一 GSA方法在情境四底下的統計檢定力表現(雷達圖) 92](#_Toc517703900)

[圖十二 GSA方法在情境六底下的統計檢定力表現(雷達圖) 93](#_Toc517703901)

[圖十三 GSA方法在情境三底下的統計檢定力表現(雷達圖) 94](#_Toc517703902)

[圖十四 GSA方法在情境三、四、六底下的統計檢定力表現(雷達圖) 95](#_Toc517703903)

**表目錄**

[表一1 (A) 乳癌資料集的詳細資訊 39](#_Toc517453855)

[表一2 (B) 大腸直腸癌資料集的詳細資訊 40](#_Toc517453856)

[表一3 (C) 白血病資料集的詳細資訊 41](#_Toc517453857)

[表一4 (D) 肺癌資料集的詳細資訊 42](#_Toc517453858)

[表一5 (E) 卵巢癌資料集的詳細資訊 43](#_Toc517453859)

[表二6 各個癌症和其相關的Pathway彼此之間的關係 44](#_Toc517453860)

[表三7 (A) 多維度常態性檢定對型一誤差率的估計–維度數為2, 4 52](#_Toc517453861)

[表三8 (B) 多維度常態性檢定對型一誤差率的估計–維度數為8, 9, 25, 49 53](#_Toc517453862)

[表四9 (A) 多維度常態性檢定檢定力表現–為混合分佈，維度數為2, 4, 8, 9 54](#_Toc517453863)

[表四10(B) 多維度常態性檢定檢定力表現–為混合分佈，維度數為25, 49 57](#_Toc517453864)

[表五11(A) 多維度常態性檢定檢定力表現–為t分佈，維度數為2, 4, 8, 9 58](#_Toc517453865)

[表五12(B) 多維度常態性檢定檢定力表現–為t分佈，維度數為25, 49 61](#_Toc517453866)

[表六13(A) 各個癌症資料集常態性檢定結果–單一探針，2個基因，病例組 62](#_Toc517453867)

[表六14(B) 各個癌症資料集常態性檢定結果–單一探針，2個基因，對照組 63](#_Toc517453868)

[表七15(A) 各個癌症資料集常態性檢定結果–單一探針，4個基因，病例組 64](#_Toc517453869)

[表七16(B) 各個癌症資料集常態性檢定結果–單一探針，4個基因，對照組 65](#_Toc517453870)

[表八17(A) 各個癌症資料集常態性檢定結果–單一探針(log)，4個基因，病例組 66](#_Toc517453871)

[表八18(B) 各個癌症資料集常態性檢定結果–單一探針(log)，4個基因，對照組 68](#_Toc517453872)

[表九19(A) 各個癌症資料集常態性檢定結果–2個探針，病例組 69](#_Toc517453873)

[表九20(B) 各個癌症資料集常態性檢定結果–2個探針，對照組 70](#_Toc517453874)

[表十21(A) 各個癌症資料集常態性檢定結果–4個探針，病例組 71](#_Toc517453875)

[表十22(B) 各個癌症資料集常態性檢定結果–4個探針，對照組 72](#_Toc517453876)

[表十一23(A) 各個癌症資料集常態性檢定結果–2個探針(log)，病例組 73](#_Toc517453877)

[表十一24(B) 各個癌症資料集常態性檢定結果–2個探針(log)，對照組 74](#_Toc517453878)

[表十二25GSA方法在差異表達為共變異數矩陣發生變化時的檢定力表現 75](#_Toc517453879)

[表十三26GSA方法在多維度常態分佈情境下所用的相關係數矩陣1 (非結構化) 76](#_Toc517453880)

[表十四27GSA方法在多維度常態分佈情境下所用的相關係數矩陣2 (非結構化) 78](#_Toc517453881)

[表十五28GSA方法在多維度常態分佈情境底下的表現(非結構化相關係數矩陣1) 80](#_Toc517453882)

[表十六29GSA方法在多維度常態分佈情境底下的表現(非結構化相關係數矩陣2) 81](#_Toc517453883)

# 研究背景

隨著科技的發展，在基因研究領域中，微陣列(microarray)資料已經廣泛地被應用到各種醫學以及生命科學的研究領域，科學家們可以藉由這項技術檢測基因表現量(gene expression)，並且從中找出受疾病影響的基因集合(gene set)或是基因。而這樣的方法除了讓科學家們能更精確地了解每個基因的基因表現量在正常人以及病患身上的差異以外，更省下了很多尋找致病基因的時間，加速了醫藥領域的發展。因此，如何有效地去找出那些受疾病影響的基因集合便成為一個很重要的議題，而統計學家便藉由分析整個基因集合的基因表現量在不同表現型(phenotype)組別上所形成的差異來回答這個問題，然而，在傳統的統計方法中，由於微陣列的資料型態不同於以往所處理的資料樣式，因此，有很多方法並不適用於微陣列資料，而這也就迫使統計學家們針對如此的資料型態去發展新的統計方法。然而，當我們在發展一個新的方法並且將它運用到實際資料前，必須要先知道這個新方法的優缺點，而模擬是其中一種方式；「模擬」這個步驟在統計學家發展新方法的道路上扮演著一個很重要的角色，一個適當的模擬機制除了能夠提供我們一個公平的平台去評估這些方法的好壞以外，還可以幫助統計學家們去了解並且克服新方法所遇到的問題以及彌補不足的地方。

而在現有被用來評估基因集合分析(Gene-Set Analysis, GSA)方法的模擬方式中，大致上可以分成以下三大類：首先第一大類，在絕大多數的研究中(Glazko, and Emmert-Streib 2009; Goeman, and Bühlmann 2007; Ackermann, and Strimmer 2009; Hummel, Meister, and Mansmann 2007; Kim, and Volsky 2005; Liu et al. 2007; Song, and Black 2008)，研究者使用多維度常態分佈(Multivariate Normal Distribution, MVN)來模擬整個基因集合內所有基因的基因表現量，而之所以利用多維度常態分配來進行模擬是因為比起其他複雜的分配，該分配較為簡單而且具有許多良好的性質，例如：對稱性、該分配屬於指數家族等。而在這之中，有些研究者將多維度常態分佈的相關係數矩陣設定成複合對稱架構(compound symmetry) (Kim, and Volsky 2005; Glazko, and Emmert-Streib 2009)或是對角方塊複合對稱架構(block diagonal and compound symmetry) (Goeman, and Bühlmann 2007; Liu et al. 2007)的樣式，並且利用這些架構來表達一個基因集合內部，基因與基因之間的網絡合作關係。另外一方面，他們在病例組(case group)和對照組(control group)的多維度常態分佈中分別給予期望值向量(mean vector)不一樣的設定，並且利用這樣的差別設計來呈現致病基因在兩組之間的差異表達(differentially expressed)情形。而除了上述的方式以外，有些研究者(Song, and Black 2008; Ackermann, and Strimmer 2009)則是覺得應該要將同一個基因集合底下的致病基因和其他的基因分開來看，因此他們根據一個基因集合內部可能會出現的基因種類(如：持家基因(housekeeping gene)、up/down-regulated gene等)去設計每個基因種類專屬的分配設置。然而，使用這一大類的方法來進行模擬，雖然簡單，但由於現實的基因資料往往比這些參數設置還要來的複雜許多，甚至有許多資料特性(如：偏度和峰度)是這種方法完全無法仿造的，因此利用這種方法所得到的結果未必會是正確的。

而在第二大類中，過去的研究裡，有研究者使用多維度對數常態分佈(Qiu, and Yakovlev 2006; Klebanov et al. 2007)來對資料進行模擬，然而他們只是將整體的分配換成是對數常態分配，至於分配中裡面的參數設定則和第一大類的設定方式相同，因此，此種方式的缺點和上面那一大類是差不多的。

另外，有別於前面兩大類使用分配假設的模擬方式，在第三大類中，有研究者(Nettleton, Recknor, and Reecy 2008; Heller et al. 2009; Liang, and Nettleton 2010; Benidt, and Nettleton 2015)認為，不管用甚麼樣的分配來進行模擬都無法模擬出真實基因彼此之間的網絡合作關係，而若要模擬出真實的關係，那麼最好的選擇就是直接利用實際的基因資料來進行模擬。因此，在研究中，他們大多使用接下來的這個概念來進行模擬。首先，先將原先實際的基因資料分成病例組以及對照組兩組，並且將這兩組資料當中的對照組樣本資料複製兩次當成是在模擬實驗中的病例組以及對照組資料，此時，模擬實驗中病例組以及對照組的資料是完全相同的。接著，他們再把焦點移回到原先在實際基因資料中的病例組資料，他們從這組資料去挑選出那些待檢測的基因集合它們的基因表現量，並且拿這組病人的這些基因表現量去取代模擬實驗中相對應的病例組資料，也就是說在這個時候，在模擬實驗中病例組的資料裡，那些待檢測的基因集合裡的基因表現量是來自於原先實際資料中的病例組資料，然而對於不在該基因集合裡的那些基因，他們的基因表現量則是來自於原先實際資料中的對照組資料，至於模擬實驗中對照組的資料則是原封不動地取自於原先實際資料中的對照組資料。這一類方法的好處是能夠局部地呈現出實際基因彼此之間的合作關係，但是在病例組的資料中，那些待檢測的基因集合裡的基因和基因集合外的基因彼此之間的關係卻沒有辦法被適當地呈現出來，此外，由於病例組的資料是從不同人身上所拼湊出來的，而每個人的基因可能會受到不同環境或是一些其他不明因素的干擾，因此，這樣的拼湊方式是否恰當仍然是一個很大的問題。

有鑑於上述的幾類模擬方式都各有優缺點，可能會讓我們對現有的基因集合分析方法(gene-set analysis method)產生誤解，因此，本論文的第二章會先從真實的基因資料切入，觀察實際的各種基因資料在經過正規化後是否真的符合常態分佈，並將這部分的結果運用到後續章節的模擬情境設計之中。在第三章裡，我們將五個基因集合分析方法套用到我們所建立的非常態模擬方式上進行比較，藉由評估各個檢定在不同情境(scenario)下的表現是否穩健(robust)來提供使用者對這些檢定一個更全面的觀點。最後，我們利用雷達圖(radar chart)將這些眾多複雜的結果以視覺化呈現，以利後人在比較各種方法時能夠更容易且準確地知道這些方法的好壞。另外，本研究是第一個針對真實基因資料進行常態性假設檢驗之研究，也是第一個站在多維度的角度思考基因表現量資料中可能會出現可能會出現非常態多維度的分配設定，並且利用這些想法設計跨情境模擬的研究，此外，本研究在模擬階段評估不同基因集合分析方法在各個情境間的表現穩健程度以及最後雷達圖的視覺化方式，這些都是在過往文獻鮮少甚至是沒有被提及到的概念，而這些想法除了在理論上提供研究者在研究每個基因集合分析方法時，能對這些方法有一個更全面性的認知之外，亦能在實務上提供使用者一個更有力以及更可靠的比較資訊，並且協助他們依照研究需求選擇適合的方法來進行後續更進一步地分析。­­­­­

# 真實基因資料常態性假設之分析

在現有的統計方法中，由於常態分佈具有許多良好的性質，因此，不但有很多檢定是從常態分佈的理論所發展出來的，甚至當我們在對這些檢定進行評估時也常大量地使用常態分佈來生成模擬資料。然而，這些基因的資料是否真的服從常態分佈是需要先進行了解的，如果這些基因資料確實服從多維度(對數)常態分佈的話，那麼使用這些檢定方法以及評估方式所得到的結果才具有效率以及可信度。本章節將會介紹幾個從公開網站下載的mRNA基因表現量資料的相關背景，接著再說明如何進一步地整理這些資料，並且拿這些資料來進行多維度常態檢定，藉由檢定結果來回答這個問題。

## 資料背景

在現實生活中充滿了各式各樣的基因表現量資料，而在本研究中，我們把目標鎖定在癌症(cancer)上，針對乳癌(breast cancer)、大腸直腸癌(colorectal cancer)、肺癌(lung cancer)、白血病(leukemia)以及卵巢癌(ovarian cancer)一共五個疾病、六個資料集進行研究，其中肺癌的資料集有兩個。而在這些癌症資料當中，除了卵巢癌的資料是取自於美國癌症基因體圖譜計畫(The Cancer Genome Atlas, TCGA)以外，其餘疾病的基因表現量資料皆取自於美國國家生物技術資訊中心(National Center for Biotechnology Information, NCBI)中的基因表現量資料庫(Gene Expression Omnibus, GEO)。另外，在這些資料集當中，乳癌、大腸直腸癌以及兩個肺癌資料集的資料屬性是屬於成對資料(paired data)，也就是說在這些疾病的資料裡頭，每個癌細胞組織的基因表現量資料和正常細胞組織的基因表現量資料是從同一位癌症病患身上所取得的，而這裡所指的正常細胞組織也就是癌旁正常組織(adjacent normal tissues)；因此，在這些資料集當中，病例組和控制組兩組的樣本數是相同的。另外，白血病以及卵巢癌的資料屬性不是成對資料，在這些資料集當中，病例組和控制組的每筆資料分別從不一樣的病患身上所取得，也因此兩組的樣本數不一定會相同。同時，在本研究中，我們將基因表現量的技術平台聚焦在Affymetrix GeneChip的HG-U133A(在NCBI GEO裡被編碼為GPL96)以及HG-U133 Plus 2(在NCBI GEO裡被編碼為GPL570)這兩種平台上，在我們所收集到的資料裡，僅有大腸直腸癌資料集以及一個肺癌的資料集是採用HG-U133 Plus 2平台，其餘資料集皆採用HG-U133A平台，而這兩種平台最主要的差異是差在探針(probe)數目上的不同。

另外，有關各個資料集的資訊，首先是乳癌資料，在NCBI的編號為“GSE15852”，這筆資料集主要收集的疾病亞型為浸潤性乳腺管癌Invasive Ductal Carcinoma, IDC))，樣本數為36對成對樣本，總共有72筆資料。其次是大腸直腸癌資料，在NCBI的編號為“GSE8671”，這筆資料集主要收集的疾病亞型為腺瘤(adenoma)，樣本數為32對成對樣本，總共有64筆資料。再來是兩個肺癌資料，其中在NCBI編號為“GSE7670”的這一筆資料集，它所使用的平台是HG-U133A，這筆資料集主要收集的疾病亞型為肺腺癌(lung adenocarcinoma)，樣本數為26對成對樣本，總共有52筆資料，而另一筆肺癌資料集在NCBI的編號為“GSE19804”，它所收集的疾病亞型同樣為肺腺癌，但使用的平台則是HG-U133 Plus 2，樣本數則為47對成對樣本，總共有94筆資料。接下來是白血病資料，在NCBI的編號為“GSE9476”，主要收集的疾病亞型為急性骨髓性白血病(Acute Myeloid Leukemia, AML)，由於資料屬性並非成對樣本，因此，所有的資料可以分成兩組，一組資料來源為急性骨髓性白血病患者，共有26人，以及另一組資料來源為正常人的造血細胞(hematopoietic cells)，共有38人，總共有64筆資料。最後則是卵巢癌資料，在TCGA的編號為“tcga-ov”，主要收集的疾病亞型為漿液性卵巢上皮癌(serous epithelial ovarian cancer)，這筆資料直接沿用Cheng (2015)和She (2016)所下載及整理的結果，一共282筆。這些資料可以依兩年的預後情形做為切點分成兩組，一組病患的存活時間不到兩年，共有72人，以及另一組存活時間超過兩年，共有210人。有關各個資料集詳細的資訊全部都彙整在表一（表一 (A)～(E)）裡。

## 資料前處理以及資料篩選

雖然各個疾病的基因表現量資料在NCBI的網站上都已經有提供正規化處理好的版本，但由於每位研究者在處理這些資料時所使用的正規化(normalization)技巧以及轉換方式都不盡相同，因此為了降低不同方法對我們最後實驗結果所造成的干擾，除了卵巢癌資料以外，其餘的癌症資料我們先到NCBI網站上將它們的原始數據(raw data)給下載下來，接著，再將這些原始數據拿到R語言中統一進行資料正規化，而在完成了資料正規化之後，我們便對這些處理好的資料進行更進一步的整理以利後續分析之使用，而後面的這些整理過程可以分成兩個步驟，首先是針對基因進行篩選，接著再針對所有樣本進行樣本臨床資料的整理，而詳細過程如下：

### 正規化：

在收集了各個疾病基因表現量的原始數據資料後，我們利用R語言套件包，*affy*，裡面的*ReadAffy()*函式來將基因表現量的原始數據給讀入R語言中。接著對資料進行兩個階段的正規化處理，首先第一個方法為Robust Multi-array Average (RMA)，在R語言中，我們利用*affy*套件包裡的*rma()*函式來進行此種正規化處理。接著再利用Quantile normalization方法再將這筆資料正規化了一次，在R語言中，我們利用套件包，*preprocessCore*，裡面的*normalize.quantiles()*函式來對資料進行此步驟，而原始數據資料在完成了這兩階段的正規化後，便可以進入下一個階段進行更進階地處理。

### 生物路(途)徑(pathway)的選擇：

由於本研究的最終目標是想要探討基因集合分析方法的模擬方式，因此在這裡的研究主角就是在控制組以及病例組兩組之間有著差異表達情形的那些生物路徑。我們針對本研究中的每個癌症亞型進行了文獻回顧，並且從醫學以及生物學上的角度收集與這些疾病相關的生物路徑，詳細參考資料請見表二。在收集完了生物路徑後，我們到Kyoto Encyclopedia of Genes and Genome (KEGG)網站中將每個生物路徑裡的基因成員資訊下載下來，而由於在KEGG中，有些基因會有兩個以上的名稱(gene name)，因此在接下來的研究中，我們都只擷取該欄位中的第一個名稱來做為配對使用。接著，再到Affymetrix的官網中，下載各個基因晶片所專屬的基因注釋檔案(gene annotation file)，並且將基因晶片中每一個探針位點所屬的基因資訊和經過處理過後的資料給結合在一起。最後，再拿這筆資料去和每個生物路徑裡的基因成員進行配對，而在配對過程中，由於有很多基因都擁有兩個以上的探針資料，因此我們對資料進行了兩種處理，一種是針對每個基因去挑選出變異係數(Coefficient of Variation, CV))最大的探針來代表該基因的基因表現量，另一種則是將每個基因所擁有的探針資料全部納入，在此種處理方式底下，一個基因可能會擁有數個探針的基因表現量資料，而在進行完這整個選擇的流程後，我們便可以得到每個生物路徑在這兩種處理方式底下所專屬的基因表現量資料。

### 樣本臨床資料的整理：

為了要降低樣本臨床資料來源對實驗結果的干擾，因此在本研究中，我們利用NCBI以及TCGA網站上所提供的樣本臨床資料來對樣本進行整理。整理樣本的條件為，在最後所收集的資料中，在帶有癌症的組別內，所有病患所得到的疾病必須要屬於同一個疾病的亞型，而且病患所屬的臨床階段或是腫瘤級數必須一致；而無論最後樣本的表現型組別是甚麼，每位樣本的性別、種族必須要相同；此外，每個樣本取樣細胞的來源都必須從同一種器官、位置或組織中取得。有關各個資料集詳細的樣本分類資訊，統整在表一（表一 (A)～(E)）中。雖然在收集到的臨床資料中，並不是每筆資料都有完整的臨床資訊，但只要網站上有提供，都會被利用來進行篩選。經由這樣一系列的挑選流程之後，就能對這些資料進一步地分析。

## 常態性檢定

為了要確認正規化後的基因表現量資料服從常態分佈或是對數常態分佈，我們必須得對處理完的資料進行多維度的常態檢定，或是將資料經過對數轉換後再進行多維度的常態檢定。然而，在現今的文獻研究中有著各式各樣的常態性檢定，並沒有一個固定的準則告訴使用者甚麼樣的情況底下該用甚麼樣的常態性檢定。因此，在本研究中，當我們要知道基因資料的常態性時，會同時比較多個檢定所產生出來的結果是否一致，並且拿這個結果來輔助最後的判斷以及決策。

在單維度的時候，適合度(卡方)檢定很常被使用來確認以及佐證資料分佈。而當我們在面對連續型的資料型態時，我們必須將整個資料群體拿去進行轉換，區分成數個類別(data binning)後再進行適合度檢定。然而，當我們在面對高維度資料的時候，如何選擇適當的高維度切點去區分資料會是一個很大的議題，因為不同切點的選擇將會影響到這個檢定的檢定力(power) (Baggerly 2001)。也因此，在我們的研究中，這個方法並不是一個好的選擇。而另一個在單維度很常被拿來使用的檢定是Kolmogorov-Smirnov (K-S)檢定，其原理是去找出理論分佈的累積分配函數(cumulative distribution function)以及實際資料的經驗分配函數(empirical distribution function)，並利用兩者之間最大的距離來判斷該筆資料是否服從一個特定的理論分配。雖然這個方法在過往文獻中已經有人成功地拓展到高維度(Justel, Pena, and Zamar 1997)，但其所開發出來的檢定統計量即便是在二維的情況底下仍是非常難以去計算而且複雜的。此外，有關於經驗分配函數的多維度版本該從哪一個維度方向開始進行累加也是K-S檢定的另一個困難點(Baggerly 2001)。

有鑑於將現有在單維度常用的方法直接拓展到高維度並不是想像中的那麼有效率，在本研究中，我們直接從多維度的角度來尋找檢定方法。在眾多傳統的多維度檢定裡，過去有研究指出，Henze-Zirkler’s test (Henze, and Zirkler 1990)以及Royston’s test (Royston 1992)這兩個檢定比起很多傳統的方法較能良好的控制型一誤差率(type I error rate)以及有著較佳的統計檢定力(Korkmaz, Goksuluk, and Zararsiz 2014)；此外，該研究同時也指出Mardia’s test (Mardia 1970)能夠藉由比較多維度版本的偏度以及峰度在多維度常態和實際資料彼此之間的差別，來判斷這筆資料是否服從多維度常態分佈。然而，卻也有文獻指出，偏度以及峰度在多維度的情況底下有一種以上的定義，而這些不同的定義方式會大大地影響這個檢定的表現，尤其在樣本數相對較小的情況下影響將會更為劇烈(Koizumi, Okamoto, and Seo 2009)。考量到我們所使用的基因資料其樣本數都不多，因此Mardia’s test在我們的研究中只扮演輔助的角色。另外，過去也有文獻指出Energy test (N-statistic)的單樣本檢定(one-sample test)版本能在某些情境底下察覺出樣本分配的一些細部差異，而這些差異是上面的兩個檢定所沒有辦法找出來的(Székely, and Rizzo 2005)。因此，在本研究中，我們會使用Henze-Zirkler’s test、Royston’s test、Mardia’s test以及one-sample energy test來檢測虛無假設(null hypothesis)，而在這裡的虛無假設也就是這些基因表現量資料服從常態分佈。另外，在R語言中，針對前三個檢定，我們使用套件包，*MVN*，裡面的*mvn()*函式來對我們的基因表現量資料進行這三種檢定，而針對one-sample energy test，我們使用套件包，*energy*，裡面的*mvnorm.etest()*函式來對我們的基因表現量資料進行這個檢定。有關於這四個檢定詳細的構造以及原理介紹，請見附錄一。

## 常態性檢定模擬以及其結果

由於上述的多維度常態性檢定方法其檢定統計量的構造裡都需要用到反矩陣來進行資料的標準化，因此其所適用的情況都是在資料的維度數目小於樣本數目時才有辦法正常執行，也因為這樣，我們必須將實際的樣本資料維度進行適當的切割，並且再運用一些多元常態的理論性質來幫助我們確定這些實際資料的常態性。然而，有些資料集的樣本數非常的小，這時候該如何去選擇每次檢定所使用的維度數目就變得非常重要，而這種情況也是先前開發這些檢定的作者們未顧及到的使用情境，因此我們對這些情境進行模擬，並且觀察這些檢定在各種可能會遇到的情況下的表現為何。

### 模擬設定:

為了探討這些檢定在我們感興趣的情境底下對型一誤差率的估計以及檢定力的表現，我們設定以下參數：

為每次常態檢定所檢測的基因(維度)數，有這六種設置，

為每次常態檢定所檢測的樣本數，有這六種設置，

為第位樣本的基因表現量資料，

為第個基因及第個基因彼此之間的相關係數，有這四種設置。

接著，針對型一誤差率的估計，我們設定資料服從以下常態分配：

其中，

在時，，，

在時，，，

在時，，，

至於模擬次數，我們對上面設定中的每一種組合進行1000次模擬試驗。

而針對統計檢定力表現，我們設定資料服從以下兩種類型的分配。首先，第一種類型為多維度常態混合分配：

其中，

代表混合分配中兩個多維度常態分佈的混合比例，，另外，

在時，，，

在時，，，

，

至於模擬次數，我們對上面設定中的每一種組合進行500次模擬試驗。

接著，第二種類型則為多維度t分配(Multivariate t-distribution, MVT)：

其中，

代表這個多維度t分配的自由度，有這三種設置，另外，

在時，，，，

在時，，，，

至於模擬次數，我們對上面設定中的每一種組合進行500次模擬試驗。

這些設置最主要是想要了解以下幾個問題，首先是當每次常態檢定所檢測的基因數和樣本數很靠近時，這些檢定的表現為何，以及這些檢定的表現是否會受到基因彼此間相關係數的影響，還有就是當樣本數不變時，我們是否能夠藉由降低每次常態檢定所檢測的基因數目來提升這些檢定的表現。

### 模擬結果:

首先，從型一誤差率的模擬實驗結果中表三(表三 (A) (B))，我們可以發現，這些檢定在表三 (A)大多數的情況下對型一誤差率的控制都還在可以接受的範圍內。然而，Mardia’s test和Henze-Zirkler’s test在基因數目和樣本數很靠近時(表三 (B))會出現比較不穩定的情況，這兩個檢定在此種使用情境底下計算統計量以及p值時有時候會出現問題。另外，若在每次執行常態檢定的過程中一次同時考慮多個基因(表三 (B))，那麼我們可以發現，幾乎每個檢定對型一誤差率的控制都會受到影響，而這種情況也是先前開發這些檢定的作者們未顧及到的使用情境。接著，我們也可以看到，基因之間的相關程度只有在基因的數目往樣本數靠近時才會明顯發揮影響力，在其他的情況下相關性的高低並不會大幅度地影響檢定的型一誤差控制率。這跟我們原先預期的結果有一些不同，原本我們的認知是，既然是常態性檢定，在任何情況底下都不應該受到相關性高低的影響，但從實驗結果中可以發現這個想法在某些情況底下是不會成立的。

另一方面，從檢定力的模擬實驗結果中(表四 (A) (B)、表五 (A) (B))可以發現，有些檢定在某些特殊情況底下，其拒絕率會變得異常的極端（0或是1），而發生這些狀況最主要的原因和前一部分相同，即在特殊情況下檢定受到嚴重影響，但這些數值結果的正確性大致上還是可以藉由情境的設定去區分出來。整體來說，在絕大多數的情況底下，隨著樣本數的上升，每一個檢定的拒絕率也會跟著上升。另外，當對立假設(alternative hypothesis)所設定的情境和虛無假設之間的差距越大（在混合常態模型是兩個常態分配的混合比例越相近，以及兩個分配所使用的相關係數差距越大，在多維度t分配則是自由度越小），每個檢定的拒絕率也會越高，這些結果都還蠻正常的和原先預期的差不多。值得一提的是，我們預期在其他條件不變的情況下，減少資料的維度(基因數)可以使得檢定力上升，這個假設並沒有被實驗結果給證實，每個檢定在維度產生變化時其檢定力沒有固定的趨勢。

而當每次常態檢定所檢測的基因數目很靠近樣本數的時候，相較於檢測的基因數目很少的情況下，絕大多數的檢定在這種時候的表現都不是很好，因此，當我們在進行下一個階段實際的基因資料分析時，我們每次會針對兩個或是四個基因檢查多維度常態性。另外，基因彼此之間的相關程度對檢定力的影響可以從對立假設是多維度t分配的實驗結果(表五 (A) (B))看出來，大部分的檢定只有在基因數目很靠近樣本數時才會比較容易受到相關程度的影響，而對Royston’s test而言，基因間的相關性越低，其拒絕率會越高，而這個現象和在型一誤差的實驗結果是一致的。

總結來說，由這兩個部分的實驗結果看起來，考量到各個資料集的樣本數大小不一，再加上資料檢測的效率，因此我們每次從一個生物路徑中隨機挑選二或四個基因進行常態性檢定，並且重複數次這樣的過程。

## 真實基因資料分析結果

在這一部分的分析中，主要是想要了解真實的基因資料經過正規化處理後是否真的服從多維度常態分配或者是多維度對數常態分配。而從上一個部分的模擬實驗所得到的結果中可以得知，若同時考慮所有資料集的樣本數，我們發現所有多維度常態性檢定的統計檢定力(power)在維度(基因)數目為兩個和四個時會有最好的表現。因此，在接下來的分析中，我們將會針對每個疾病資料集中的每個生物路徑去隨機選擇兩個或四個基因探針出來進行多維度常態性檢定。另外，針對多維度對數常態分配的檢測，我們則是隨機選擇四個基因(探針資料)來進行對數轉換後再進行多維度常態性檢定。所有隨機選擇和檢定的流程會重複500次，並計算每個生物路徑在這500次內的拒絕率，最後藉由比較某個生物路徑在某個檢定下的拒絕率結果和該檢定的型一誤差率來回答真實基因資料是否符合多維度常態分配。

在實際資料的各個檢定表現中，首先，我們可以發現，對於每一個基因是否有進行單一探針的篩選對最後常態性表現的影響幾乎沒甚麼差別（有進行單一探針篩選：表六 (A) (B)、表七 (A) (B)、表八 (A) (B)，沒有進行單一探針篩選：表九 (A) (B)、表十 (A) (B)、表十一 (A) (B)，有篩選和沒有篩選的結果差距大部分都小於0.1）。接著，有關於每次檢定所使用的基因數目這項議題中，我們可以發現在肺癌的兩個資料集中，同時使用4個基因(表七 (A) (B)、表十 (A) (B))比起同時使用2個基因(表六 (A) (B)、表九 (A) (B))進行常態性檢定會得到較高一點的拒絕率，至於其他資料集在這兩種設置方式底下的表現則是差不多。另外，從資料的檢定結果中也可以觀察到資料所使用的晶片平台對最後的結果並沒有特別的影響。

接著，我們針對每次都是使用4個基因來進行多維度(對數)常態檢定的結果來更進一步探討，我們將每個疾病在這一部分的結果畫成點圖(dot plot) (圖一～圖六)，在這些圖中，紅色虛線(以及藍色虛線)代表某個檢定在該疾病分組的近似樣本數(這邊採四捨五入)底下的型一誤差率，而cancer (normal)則代表使用某個疾病的病例組(對照組)資料所產生出來的結果，另外，log則代表將該筆資料取對數後所得到的檢定拒絕率，用來判斷資料是否服從對數常態分配。而從所有檢定型一誤差率的模擬結果還有實際資料的檢定結果比較來看，由於乳癌(圖一)以及大腸直腸癌(圖二)受到樣本數的影響（分別為13以及11人），再加上有關於這兩個疾病的腫瘤細胞組織(橘色和紫色)還有健康細胞組織(藍色和綠色)其常態檢定拒絕率以及取對數後的檢定拒絕率都不是說很高（0.1左右），因此我們沒有充分的證據證明這兩個疾病的基因表現量資料在經過正規化處理後並非服從於多維度(對數)常態分配。而在另一方面，白血病(圖三)、肺癌(圖四、圖五)、卵巢癌(圖六)的病例組資料結果(橘色和紫色)，因為檢定拒絕率還算高（白血病：0.4左右，肺癌：0.5、0.8左右，卵巢癌：0.5、0.6左右），再加上多維度常態分配的性質(多維度常態分佈的邊際分佈還是常態分佈)，所以我們可以說這三種疾病的基因表現量資料在經過正規化處理後可能並非服從多維度(對數)常態分配。至於白血病、肺癌以及卵巢癌在研究中所蒐集到的對照組資料，雖然比起病例組的資料結果，他們的常態檢定拒絕率在有些資料集的檢定拒絕率數值較低，但仍然可以判斷出這些基因表現量資料在經過正規化處理後可能並非服從多維度(對數)常態分配。

整體來說，雖然在大部分的資料集中，取對數後的檢定拒絕率有稍微下降，但是還是可以很明顯的說明這些正規化後的資料集可能並非服從多維度(對數)常態分佈。另外，我們也發現了，這些實際資料檢定的拒絕率結果和資料集的樣本數有著很大的關係，這個部分也和在模擬實驗中所得到的結果相同；因此對於乳癌以及大腸直腸癌資料集來說，雖然在這個實驗結果中無法確認常態分配假設，但是可以合理的懷疑這些資料事實上可能也不是服從多維度(對數)常態分佈，只是實驗數據受到樣本數干擾所致。

# 基因集合分析方法的模擬比較

從前一部份的研究證據中可以了解到，幾個癌症資料集中的基因表現量資料在經過正規化處理後可能並非服從多維度常態分佈，而這和目前在學術上大多使用多維度常態分佈來進行模擬的方式互相矛盾。現階段並沒有相關的文獻討論這些基因集合分析方法在其他種模擬狀況底下的表現，也因此我們收集了五種方法(Hotelling’s T2 (Lu et al. 2005; Schäfer, and Strimmer 2005)、The Global test (Goeman et al. 2003)、The GlobalANCOVA (Hummel, Meister, and Mansmann 2007)、Energy Test (N-statistic) (Székely, and Rizzo 2004; Baringhaus, and Franz 2004; Klebanov et al. 2007)、GSEA (Category) (Gentleman, Morgan, and Huber 2008))，並且設計了六種情境來探討這個議題，期望這些結果能讓後人在使用這些方法分析真實資料時，可以更全面地了解這些方法的實際表現。

## 模擬設定

為了了解這些基因集合分析方法在非常態分佈下的表現，我們一共設計了六種情境，在這六種不同的情境之中，包括了一般在GSA中常用來進行模擬的多維度常態分配，以及其他各種非常態的多維度分配，如：多維度t分配，以及多個由兩個多維度常態分佈所混合形成的混合分配(Mixture of Two Multivariate Normal Distributions, Mixture of Two MVN)。在資料生成方面，我們先將總人數設為70人，並且讓每個人都有30個基因。接著，再將所有人的表現型分成兩種（組），一組為控制組，另一組則為病例組，兩組表現型的人數相同，都是35人。而這兩組的基因表現量在同一種情境底下，則分別服從各自的多維度分配。在這兩組多維度分配中，我們僅針對期望值向量進行不一樣的參數設計，其餘的參數在同一種情境底下則維持相同的設計，也就是說利用兩組期望值向量間的差異，來呈現控制組以及病例組之間，基因表現量資料的差異表達情形。而在兩組的期望值向量上，我們設計了兩大種類的差異方式，各個情境及設定的詳細內容將於下方詳述之。

另外，在每一種情境以及差異方式底下，我們會重複執行1000次試驗，每次的試驗資料分別使用R語言中不同的種子(seed)去生成。在生資料的過程中，會先使用R語言套件包，*mvtnorm*，裡面的*rmvnorm()*以及*rmvt()*函式，分別來生成多維度常態分配以及多維度t分配的資料後，再視每一個情境、組合所需去加以調整或混合，最後再利用這些資料進行分析。

### 差異方式A—每個基因的邊際分佈在不同的兩種表現型間有相同的平移差異：

在此種差異方式下，我們總共設定了六種情境，並且讓每一個基因（維度）的邊際分佈在兩種表現型間都有相同的平移差異，意即在**（**差異向量，定義在下方詳述**）**內的每一個元素都是相同的，而當時的結果也就代表著各個方法在各種情境底下的型一誤差率，當為其他設定值時則為各個方法在各種情境底下的統計檢定力。

情境一、二：多維度常態分配(MVN)

在這兩種情境下，我們使用架構為複合對稱架構且相關係數分別為0.1、0.9的相關係數矩陣來模擬基因集合裡基因彼此之間的交互作用關係，並且設定控制組資料來自以下分配：

，

病例組資料則來自以下分配：

，

其中，

：控制組中第位樣本的基因表現量，，

：病例組中第位樣本的基因表現量，，

**：**多維度常態分佈的期望值向量，

**：**多維度常態分佈的相關係數矩陣，

：基因和基因彼此之間的相關係數，

（情境一：（低相關性），情境二：（高相關性）），

：兩組資料在期望值向量上的差異向量，設定為**0**, **0.05**, **0.1**, **0.3**,

**0.5**, **0.7**, **0.9**, **1**, **2**。

情境三：多維度t分配(MVT)

在此情境下，我們使用架構為複合對稱架構且相關係數為0.9的相關係數矩陣來模擬基因集合裡基因彼此之間的交互作用關係，並且設定控制組資料來自以下分配：

，

病例組資料則來自以下分配：

，

其中，

：控制組中第位樣本的基因表現量，，

：病例組中第位樣本的基因表現量，，

**：**自由度為3的多維度t分佈期望值向量，

**：**自由度為3的多維度t分佈相關係數矩陣，需要再將此矩陣乘上

才會是在多維度t分佈中常使用的形狀矩陣(shape matrix)，

：兩組資料在期望值向量上的差異向量，設定為**0**, **0.05**, **0.1**, **0.3**,

**0.5**, **0.7**, **0.9**, **1**, **2**。

情境四、五、六：多維度常態混合分佈(Mixture of MVN)

在接下來的這三個情境下都是由兩個多維度常態分配分別以0.5的比例去混合出一個新的分配，這樣設置的目的是想去探討如果資料的分配不是一個單一母體的分配，而是混合分配時，這些方法的表現又會變得如何。在這些情境底下，我們使用兩個架構為複合對稱架構且相關係數分別為0.1、0.9的相關係數矩陣來模擬基因集合裡基因彼此之間的交互作用關係，並且設定控制組資料來自以下的分配組合：

，

病例組資料則來自以下的分配組合：

，

其中，

：控制組中第位樣本的基因表現量，，

：病例組中第位樣本的基因表現量，，

**：**多維度常態分佈的期望值向量，

**：**混合分配中第一個分佈的期望值向量

（情境四、五、六：），

**：**混合分配中第二個分佈的期望值向量

（情境四：，情境五：，情境六：），

**：**多維度常態分佈的相關係數矩陣，

：兩組資料在期望值向量上的差異向量，設定為**0**, **0.05**, **0.1**, **0.3**,

**0.5**, **0.7**, **0.9**, **1**, **2**。

### 差異方式B—並非每個基因在不同的兩種表現型間都有相同程度的平移差異：

在這種差異方式下，我們採用上述差異方式中的情境三、四、六，共三種非常態的多維度分配情境，來進行此種差異方式的研究。然而，與上述情境設定不同的地方在於，在這三個情境中，內的元素並非全部相同，我們假設內的元素為，並在下方針對內的這30個元素分別給予六種不同的參數組合設計。

組合一：0.1： 0個，0.3：30個，0.5： 0個

組合二：0.1： 7個，0.3：15個，0.5： 8個

組合三：0.1：15個，0.3： 8個，0.5： 7個

組合四：0.1： 7個，0.3： 8個，0.5：15個

組合五：0.1：10個，0.3：10個，0.5：10個

組合六： 0 ：12個，0.5：18個

在這些組合設計之中，組合一也就是在前一種差異方式裡將設定為**0.3**的案例。另外，我們將組合一、二、五、六，這四種組合設計的期望值都設在0.3左右，但把變異程度設為不同，目的是要探討內30個元素的變異程度對這些基因集合分析方法的影響。同樣地，我們將組合三、四，這兩種組合設計的變異程度設成是同一個，但把期望值設為不同，目的則是要探討內30個元素的期望值對這些基因集合分析方法的影響。在另一方面，之所以會只選擇上一種差異方式中的情境三、四、六來進行這一部分的研究是因為在現實資料中，正規化後的基因表現量分佈不容易是常態分佈，也因此我們藉由這些非常態的情境來探討此種差異方式對這些分析方法產生的影響。

## 基因集合分析方法

在這五種基因集合分析方法中，我們使用下列符號表示方式：

、分別表示控制組以及病例組的人數，

表示總人數，，

為一個的向量，用來表示個人的表現型(有病，正常)，

表示一個基因集合裡的基因（維度）總數目，

為一個的大型矩陣，用來表示某個我們感興趣的基因集合中，所有人的基因表現量，

為一個的向量，用來表示某個我們感興趣的基因集合中，所有人的第個基因表現量，

為一個的向量，用來表示控制組中第位病患的基因表現量，

為一個的向量，用來表示病例組中第位病患的基因表現量，

皆為的向量，分別表示控制組位以及病例組位病患基因表現量的樣本平均向量(sample mean vector)，

、分別表示控制組以及病例組基因表現量的樣本共變異數矩陣(sample covariance matrix)。

在本研究中，由於我們採用self-contained檢定的架構，因此，虛無假設就是控制組資料和病例組資料是來自相同的分配，相對地，對立假設則是控制組資料和病例組資料是來自不同的分配，但在參數設定中，因為在兩組表現型的分配之間唯一不同的地方只有期望值向量，所以在這裡可以把目標簡化成是檢定兩組表現型的期望值向量間是否有異同。

另外，計算每個檢定統計量的p值(p-value)時，我們針對這70位樣本的表現型（有病和正常）進行1000次的隨機排列(permutation)，並將這1000次排列樣本結果分別拿去計算統計檢定量，接著，再根據這1000次排列樣本的統計檢定量得到該統計檢定量的經驗分佈(empirical distribution)，最後再利用該經驗抽樣分佈去計算該檢定統計量之p值。

### 方法一：Hotelling’s T2

在一般Hotelling’s T2檢定中，常使用以下的檢定統計量來進行檢定：

其中，。在上述的檢定統計量中需要去找出的反矩陣，然而，在基因領域中往往會因其資料特性是基因數目遠超過樣本數（）而沒有辦法順利地找到反矩陣。因此，在這裏我們改成由Schäfer和Strimmer在2005年所提出的shrinkage估計量(以及)來取代Hotelling’s T2檢定統計量中的和矩陣。在這之中，裡的各個元素為，而

，，，，，，則代表控制組中第個人的第個基因表現量，則是使用標準化後的資料所得到的結果，則代表shrinkage intensity，經由此方式所計算出來的會使得的均方誤差(Mean Square Error, MSE)達到最小，而的計算方式則與大同小異，在此不再贅述。在R軟體中，我們使用套件包，*Hotelling*，裡面的*hotelling.test()*函式來對我們生成的各種資料進行檢定。

### 方法二：The Global Test

在本研究中，Global Test把樣本的表現型當成是一個應變數(dependent variable)，把樣本中某個基因集合裡，每個基因的基因表現量分別當成是一個解釋變數(independent variable)，也因此總共會有個解釋變數。接著，Global Test再利用這些變數建立一套羅吉斯回歸模型(logistic regression model)，並藉由此模型探討樣本的表現型和某個基因集合彼此之間的關聯型。而在檢測這兩者關聯性的過程中，Global Test先將這個模型中個解釋變數的迴歸係數視為隨機效應(random effect)，並且提出以下假設：

，

其中為某一個分配，則為某個基因集合裡的個基因在迴歸式裡的迴歸係數。若這個基因集合和病人的表現型之間沒有任何關係，則，意即，反之，則。接著，Global Test將分數檢定(score test)的檢定統計量簡化，再拿簡化後的檢定統計量探討是否會等於0，而在本研究中此檢定使用的檢定統計量簡化版本如下：

在這之中，，和的定義如同前方所述，代表裡所有元素的樣本平均數，是一個的向量，在這個向量裡的所有元素都是。檢定統計量中分母的部分是因為本研究的依變項種類是二元的(binary)，所以才有這樣的架構。而在R軟體中，我們使用套件包，*globaltest*，裡面的*gt()*函式來對我們生成的各種資料進行檢定。

### 方法三：The GlobalANCOVA

在本研究中，GlobalANCOVA和Global Test一樣，都是將迴歸模型運用到基因領域，然而不同的是，GlobalANCOVA將所有人的基因表現量經過轉換後當成是一個依變項，並且將任何我們感興趣的性狀(如：表現型)以及其他會影響到基因表現量的干擾因子(confounder)當成是自變項，在這個方法中，我們感興趣的變數和那些干擾因子是站在同一個基準點上的，而這個概念和Global Test是完全不同的。然後，GlobalANCOVA便利用這些變數再使用一般的線性回歸去建模，詳細的模型如下：

其中，

，為一個的向量，在這裏面放的是所有人的基因表現量，

，為一個的大型矩陣，在這之中，

，為一個的矩陣，用來表示此回歸模型中每個基因的設計矩陣(design matrix)，

代表所有自變項的數目，而在本研究中的自變項也就是每一個樣本的表現型，因此，為一個的向量，在這之中，

，為一個的向量，代表此模型中第個基因的迴歸係數，

，為一個的向量，在這之中，

，為一個的向量，代表此模型中第個基因模型裡的隨機誤差(random error)。接著，GlobalANCOVA利用模型比較的技巧來檢定樣本表現型這個變數和基因表現量之間的關聯性，比較模型在加入樣本表現型這個變數之前和加入之後，殘差的平方和(Residual Sum of Squares, RSS)是否能夠顯著的被樣本表現型這個變數給解釋掉一些，而此方法所使用的檢定統計量如下：

其中，，在這之中，，為帽子矩陣(hat matrix)，如果在矩陣中每個基因的設計矩陣裡沒有放入樣本表現型這個變數的資訊，則最後計算出來的表示成，而此時模型中向量裡的參數個數表示成；相反地，若有加入該變項的資訊，則最後計算出來的表示成，而此時模型中向量裡的參數個數表示成。而在R軟體中，我們使用套件包，*GlobalAncova*，裡面的*GlobalAncova()*函式來對我們生成的各種資料進行檢定。

### 方法四：Energy Test (N-statistic)

在一般的Energy Test中，常使用以下的檢定統計量來進行檢定：

這個統計量被稱為Energy statistic，有時也被稱為N-statistic。在上式裡，為某一種核(kernel)函數，為一種特殊的維度映射方式。在本研究中，我們採用在高維度中，用來計算任意兩點的歐式距離(Euclidean distance)公式來當成這裡的核，也就是說

,

其中，和分別代表和中的第維度元素。相同的概念也可以推廣到 和 ，在此就不再贅述。而上述的Energy statistic在經過此種核的轉換後，就可以想像成是去探討控制組和病例組的基因表現量資料，排除兩組各自組內的隨機誤差之後的差距到底有多大。也因此，若我們想研究的基因集合和病人表現型之間是沒有關聯的，那麼最後所得到的檢定統計量其值將會趨近於0。而在R軟體中，我們使用套件包，*energy*，裡面的*eqdist.etest()*函式來對我們生成的各種資料進行這個檢定。

### 方法五：GSEA (Category)

有別於一般文獻使用Subramanian等人在2005年所提出的GSEA方法，在本研究中，由於我們使用self-contained檢定，因此，這裏採用另一種檢定方法，該方法將舊有的GSEA方法進行一些調整，使GSEA的概念可在self-contained檢定的框架下實現，其所使用的檢定統計量如下：

其中，為第個基因在病例組以及對照組之間的雙樣本t檢定(two-sample t-test)檢定統計量，而最後的檢定統計量只有針對研究者感興趣的基因集合其內部的基因成員進行加總，其餘非基因集合內的基因則全部忽略。在這裡，由於我們將全部的基因視為同一個基因集合內的成員，因此該檢定統計量加總的範圍是從到。而在R軟體中，我們使用套件包，*Category*，裡面的*gseattperm()*函式來對我們生成的各種資料進行這個檢定。

## 分析結果

從最後的表現結果（圖七～圖十三）中，我們大致上可以分成三個面向去討論，第一個面向是各個方法在各種情境底下的型一誤差率，第二個面向是當差異方式為A時，這些方法在各種情境底下的統計檢定力，以及第三個面向，當差異方式為B時，這些方法在各種差異向量組合影響之下的統計檢定力表現。

另外，有別於以往在研究中用來評估檢定分析方法所使用的統計檢定力曲線(power curve)及數值表格的呈現方式，由於本研究使用的模擬情境有很多種，再加上每一種情境都有五種檢定分析方法，因此在本研究中，我們採用了雷達圖來將眾多圖表資訊合併，並且彙整呈現出本研究想傳達的主軸。在雷達圖中，每層虛線圓圈都分別代表著某一個統計檢定力值，因此，隨著每一層虛線圓圈從裡往外擴大，統計檢定力值也會隨著上升，而在雷達圖中的每一條虛線軸線就代表著某一種情境（或是差異組合設計）。另外，在圖中有五種不同顏色的多邊形，這些多邊形分別代表著在本研究中所使用的五種分析方法，而每個多邊形的角則代表該檢定在某一種特定情境（或是差異組合設計）底下的統計檢定力表現。

在第一個面向中，我們從圖七裡可以發現，各個方法在各種不同的情境底下對型一誤差率(0.05)的估計還在可以接受的範圍內，而且這些方法彼此之間對型一誤差率的估計大致上並沒有非常顯著的差別。但若詳細地去比較，我們仍可以發現Hotelling’s T2（星號、橘線）和其他檢定比起來，其對型一誤差率有較微寬鬆(liberal)的估計。另外，若從這六種不同情境的角度下去進行比較，我們也可以發現，若撇除Hotelling’s T2（星號、橘線）不看，相較於那兩種多維度常態的情境，大部分的檢定在模擬情境為非常態時，對型一誤差率都有稍微較寬鬆的估計。

在第二個面向中，整體來說，當（圖十）時，除了Hotelling’s T2（星號、橘線）以外，其餘剩下的檢定不管是在哪一種情境底下都能有很不錯的統計檢定力表現(power > 0.8)，而當（圖八）時，所有的檢定不管是在哪一種情境底下的統計檢定力表現都非常糟糕(power < 0.25)。

另外，由於每個檢定在這六種情境底下的表現都可以在的時候較清楚地被區分出來，因此，我們選擇（圖九）的雷達圖來更進一步地探討這些數據結果。在圖九中，先從各個檢定的角度看，可以發現到，無論是在哪一種情境底下，Hotelling’s T2（星號、橘線）的統計檢定力表現始終明顯地低於其他的檢定分析方法。此外，在高相關性多維度t分配的情境中，Energy test (N-statistic)（菱形、深藍色）的統計檢定力表現明顯地高於其他檢定分析方法；而這也就代表著比起其他分析方法，Energy test藉由比較不同表現型分配之間的歐式距離能夠更有效地在此情境底下偵測出此種平移的差異方式。而從這六種不同情境的立場下比較，我們可以發現，在兩個同為多維度常態性的情境下（情境一、二），基因彼此之間的相關性()在每個檢定分析方法的表現中扮演極為重要的角色，當相關性上升，檢定分析方法的檢定力會明顯下滑。而在另一方面，在這六種情境中，所有的分析方法都是在低相關性多維度常態的情境下擁有最好的統計檢定力表現（圖九的右上區塊），但在混合分配這種較為分散型的分佈()的情境下有著最糟糕的統計檢定力表現（圖九的左下區塊）。此外，當我們從雷達圖右上區塊的情境以順時針方向觀察到左下區塊的情境，整體分配從集中逐漸演變成混合分配這種較為分散的型態，而且分散的程度越來越大，在這個趨勢底下，每個檢定分析方法的統計檢定力表現也跟著急速下滑，而這也說明，若單純使用多維度常態分配的情境來進行模擬，其所得到的評估結果並沒有辦法推廣到其他種模擬情境以及真實的基因表現量資料上面。另外，我們還可以發現到，雖然多維度t分佈和多維度常態分佈一樣是單峰分佈，但是在該分佈中一些特有的資料特性對這五種檢定分析方法的影響並沒有辦法藉由多維度常態的情境仿造出來，而這也就證明了，若單純只使用多維度常態去進行模擬，其所得到的結果有可能高估，甚至誤判了檢定表現的這一個論述。

在第三個面向中，整體來說，若不考慮這三張雷達圖（圖十一、圖十二、圖十三）中組合一的部分，也就是所有基因的邊際分佈在兩種表現型間產生相同平移的差異方式，那麼我們大致上可以看到Hotelling’s T2檢定在此種差異方式底下的表現是最好的，Energy test (N-statistic)次之，再來則是Global Test以及GlobalANCOVA，而GSEA (Category)在此種差異方式底下的表現則是最差的。而從這一部份的實驗結果中也可以發現到，在此種差異方式下，若將這三張雷達圖（圖十一、圖十二、圖十三）擺在一起觀察，除了Hotelling’s T2檢定的表現比較不會因為分配情境的改變而大幅度的產生變化之外，其餘檢定方法的表現都會因分配情境從集中型的分配（圖十三、圖十一）變成分散型的分配（圖十二）而大幅度地下降。這也可以讓我們了解到，在現今的這些基因集合分析方法中，除了少數的一些檢定之外，其餘大多數的分析方法在非常態情境下或是在真實資料下的表現都不是很理想。因此，若我們單純使用舊有的模擬設計來評估現有的基因集合分析方法，將會誤判這些基因集合分析方法的表現，進而影響我們對這些方法的認知以及選擇。

另外，若站在各個檢定的角度下去觀察，無論是在哪一種情境底下，Hotelling’s T2檢定除了在某些特定情況以外，其在針對此種差異方式均有著令人出乎意料的表現，即便使用的情境是對所有檢定皆非常不利的分散型混合分配（情境六，圖十二），其檢定統計力的表現依然可在大多數的差異組合設計中得到很好的表現，甚至遠勝於其他的基因集合分析方法，而有關於Hotelling’s T2檢定如此特別的行為，相似的情況在過去的文獻中(Glazko, and Emmert-Streib 2009; Tsai, and Chen 2009; Bubeliny 2011; Tryputsen et al. 2014)有被提到以及更進一步地研究，當維度彼此之間的相關性很高()時，如果只有一個維度的邊際分佈在兩種表現型間產生平移，則此時Hotelling’s T2檢定的統計檢定力表現會比每個維度的邊際分佈在兩種表現型間產生相同平移現象時來的好，而這也印證了Hotelling’s T2檢定真的對此種差異方式特別地靈敏。此外，在情境為混合分配（情境五、六，圖十一、圖十二）時可以觀察到，隨著內元素()的變異數變大，Hotelling’s T2檢定的表現也就跟著上升，其對Hotelling’s T2檢定表現的影響甚至超越了內元素的平均對Hotelling’s T2檢定表現的影響，而從過去的研究(Bubeliny 2011)中也發現，在維度彼此之間的相關性有一定大小的條件下，當所有維度中若有一半維度的邊際分佈在兩種表現型間產生相同程度的平移，另一半維度的邊際分佈維持不變，則此時Hotelling’s T2檢定的統計檢定力表現會達到最好，而這也與我們從實驗中所得到的猜測結果一致，因為此時正好是內元素變異數達到最大的時候，也因此，Hotelling’s T2檢定的表現會在此時達到最好。在另一篇研究(Glazko, and Emmert-Streib 2009)的真實基因資料分析中也發現，Hotelling’s T2檢定對於那種僅有少數基因成員在兩組不同表現型產生差異表達情形的生物路徑會特別靈敏，而這也和我們的實驗結果一致。

# 討論

在經過這一系列的研究之後我們可以發現，基因彼此間的相關係數、分配情境的選擇以及在控制組和病例組之間的差異方式，都會影響基因集合分析方法的表現，而後兩者在舊有的模擬設計中常被忽略。過去的模擬方法往往在使用了唯一的一種分配情境以及差異方式後，便對各個方法以及新發展的方法進行比較及評論，這樣子的做法在本研究中已經被證實是不恰當的，利用這樣舊有的模擬方式所得到的評估結果只是冰山一角。這也是本研究中之所以設計了多種分配情境來探討各種基因集合分析方法表現的原因之一。

而從最後所有分析結果的雷達圖中我們可以清楚地了解到，比起在分散型的多維度分配情境下，所有檢定在集中型的多維度分配情境中有著更好的表現，而這項結果也告訴我們，當我們使用這些基因集合分析方法進行實際的基因資料分析時，應該要先將待檢測的資料經過初步的整理，使得兩組表現型組內的資料分佈能夠有機會變的更集中。如此一來，利用這些方法所得到的結果將會更準確以及更有效率。而在前面的研究中，我們將每個癌症資料的來源盡可能地同質化，這樣子的做法也可以輔助我們整理樣本。然而，為了要讓篩選後的資料能夠維持一定的大小，我們必須要擁有大量的樣本資料以及良好的資料品質才能夠達成這項目標，而這也有賴於之後的研究者，若能夠在收集樣本時就將這些因素考慮進去，那麼後續分析的過程將會更順利。

在另一方面，從最後的結果中，我們發現了Hotelling’s T2檢定在本研究中設定的某幾種情境底下的確擁有不錯的表現。然而，不可否認的是，這些結果僅限於去探討兩組表現型的期望值向量彼此之間是否有顯著的差異。如果我們遇到的狀況是，兩組表現型的差異表達情形不是出現在期望值向量上，而是在相關係數矩陣上有不同的結構，又或者是兩組表現型的分配完全不同，那麼Hotelling’s T2檢定在此時將會毫無用武之地，而在其他的檢定中，雖然從部分檢定的檢定統計量架構中可以看出，某些檢定是有辦法處理這一類更複雜的問題的，但這些檢定目前在本研究中的表現並不是說很理想，如果要拿這些檢定到各種常態以及非常態情境底下去回答那些更進階的問題，想必最後所得到的評估結果以及穩定性將會比現在的結果更不理想。

另外，在前面的研究中，我們利用皮爾森相關係數(Pearson's correlation coefficient)並搭配複合對稱的矩陣架構來描述基因彼此之間複雜的網絡合作關係。然而在現實資料中，相關係數矩陣的架構未必會是複合對稱的矩陣架構。因此針對這個議題，我們也同時設計了兩個非結構化(unstructured)的相關係數矩陣(表十三、表十四)來做這一塊的研究，不過使用此方法所得到的結果(表十五、表十六)仍有很多無法合理解釋的現象，值得更進一步地探討。在另一方面，基因彼此之間的關係往往比皮爾森相關係數所能描述的線性關係還要來的複雜許多，而在多維度的聯立分佈中，維度彼此之間的關聯性往往扮演著非常重要的角色，因此，如何適當地描繪出基因彼此之間的工作關係就會變成是一個非常重要的課題。關於這方面的研究，Sklar在1959年時提出了「耦合」(copula) (Sklar 1959)的想法，他將多維度聯合分佈的組成切成了兩大部分，分別是各個維度的邊際分佈，以及用來描述各個維度之間複雜關係的耦合，當我們確定了各個維度的邊際分佈之後，此時只要再確認好耦合的長相，便可以決定一個唯一的多維度分配。這種概念看似是一個能夠對複雜的基因環境進行建模的一套靈活工具，但這個方法在目前基因領域的實踐上遇到了很大的問題。由於基因資料的樣本數有限，再加上我們想研究的基因數目過多的緣故，進而導致了「維度的詛咒」(the curse of dimensionality)的發生。當我們想透過收集到的資料去識別出某一個疾病所特有的耦合時，有很大的機率會在各種耦合的適合度檢定中得到不顯著或是誤判的結果，而造成這個問題的原因其中之一是因為這些有限的樣本在超高維度的空間裡會顯得非常的稀疏(sparse)，而這樣的性質會導致這類型的資料不容易達到顯著。在後來的研究(Joe 1996)中，Joe提出vine copula的想法，他將條件分佈與聯合分佈之間的關係推廣到耦合身上，使得研究者可以先決定好每兩個維度的耦合之後，在把所有的耦合給結合再一起。但這樣子的作法對超高維度的基因網絡系統來說仍然是非常地沒有效率，因此如果我們想要使用這個方面的構想去建立出模型，並且使用該模型來建造模擬資料，那麼除了盡可能地多收集一些樣本資料以外，還是得仰賴未來科技發展的幫忙才有機會達成這個目標。

在另一方面，本研究是從各種不同模型的角度出發來評估現有的基因集合分析方法。然而，在現今的研究中，有研究者直接利用真實資料評估這些方法的好壞(Wang et al. 2015; She 2016)。有別於在第一章所提到的方法，這些研究者把實際的基因表現量資料直接套用到回歸模型中，利用回歸模型來重新產生出每位樣本的表現型，並且藉由改變每個基因的迴歸係數（效果量(effect size)）來評估各種不同分析方法彼此之間的好壞，而這樣的評估方式在未來也可以納入我們的比較情境中。

總結來說，由於目前真實基因的資料量還不夠多，再加上科技發展在超高維度資料上所面臨的瓶頸，因此想知道基因表現量在各個疾病的確切分佈更是難上加難。然而，也正因為無法得到確切分佈，我們現階段在模擬方法上能做的除了直接使用真實資料來建模進行評估以外，也可以多使用不同的分配情境、差異方式以及其他種不同的設定來進行模擬，並觀察待評估的方法在各種不同的情境及設定之下，是否都還能夠擁有良好以及穩定的表現。如此一來，藉由這樣的機制所評選出來的方法才會更加地穩健，若可將評選出來的方法運用到未知分配的實際資料上，則其分析出來的結果也才會更具有效率、穩定性以及說服力。

# 參考文獻

Ackermann, M., and Strimmer, K. (2009), “A General Modular Framework for Gene Set Enrichment Analysis,” *BMC Bioinformatics,* 10, 47.

Baggerly, K. A. (2001), “Probability Binning and Testing Agreement between Multivariate Immunofluorescence Histograms: Extending the Chi-Squared Test,” *Cytometry*, 45, 37-46.

Baringhaus, L., and Franz, C. (2004), “On a New Multivariate Two-Sample Test,” *Journal of Multivariate Analysis*, 88, 190-206.

Benidt, S., and Nettleton, D. (2015), “SimSeq: A Nonparametric Approach to Simulation of RNA-Sequence Datasets,” *Bioinformatics*, 31, 2131-2140.

Bubeliny, P. (2011), “Hotelling Test for Highly Correlated Data,” *Acta Universitatis Carolinae. Mathematica et Physica*, 52, 67-75.

Cheng, S. J. (2015), “Identification of Methylation-Driven Genes with Bayesian Conditional Autoregressive Model,” *Master’s thesis*, National Taiwan University, Taiwan.

Gentleman, R., Morgan, M., and Huber, W. (2008), “Gene Set Enrichment Analysis,” in: Hahne, F., Gentleman, R., and Falcon, S., (Eds.), (2008), “Bioconductor Case Studies,” New York: Springer, 193-205.

Glazko, G. V., and Emmert-Streib, F. (2009), “Unite and Conquer: Univariate and Multivariate Approaches for Finding Differentially Expressed Gene Sets,” *Bioinformatics*, 25, 2348-2354.

Goeman, J. J., and Bühlmann, P. (2007), “Analyzing Gene Expression Data in Terms of Gene Sets: Methodological Issues,” *Bioinformatics*, 23, 980-987.

Goeman, J. J., van de Geer, S. A., de Kort, F., and van Houwelingen, H. C. (2003), “A Global Test for Groups of Genes: Testing Association with a Clinical Outcome,” *Bioinformatics*, 20, 93-99.

Heller, R., Manduchi, E., Grant, G. R., and Ewens, W. J. (2009), “A Flexible Two-Stage Procedure for Identifying Gene Sets That are Differentially Expressed,” *Bioinformatics*, 25, 1019-1025.

Henze, N., and Zirkler, B. (1990), “A Class of Invariant Consistent Tests for Multivariate Normality,” *Communications in Statistics - Theory and Methods*, 19, 3595-3617.

Hummel, M., Meister, R., and Mansmann, U. (2007), “GlobalANCOVA: Exploration and Assessment of Gene Group Effects,” *Bioinformatics*, 24, 78-85.

Joe, H. (1996), “Families of m-Variate Distributions with Given Margins and m(m-1)/2 Bivariate Dependence Parameters,” in: Rüschendorf, L., Schweizer, B., and Taylor, M. D., (Eds.) “Distributions with Fixed Marginals and Related Topics,” *Lecture Notes-Monograph Series*, 28, 120-141.

Justel, A. , Pena, D., and Zamar, R. (1997), “A Multivariate Kolmogorov-Smirnov Test of Goodness of Fit,” *Statistics & Probability Letters*, 35, 251-259.

Kim, S. Y., and Volsky, D. J. (2005), “PAGE: Parametric Analysis of Gene Set Enrichment,” *BMC Bioinformatics,* 6, 144.

Klebanov, L., Glazko, G., Salzman, P., Yakovlev, A., and Xiao, Y. (2007), “A Multivariate Extension of the Gene Set Enrichment Analysis,” *Journal of Bioinformatics and Computational Biology*, 5, 1139-1153.

Koizumi, K., Okamoto, N., and Seo, T. (2009), “On Jarque-Bera Tests for Assessing Multivariate Normality,” *Journal of Statistics: Advances in Theory and Applications*, 1, 207-220.

Korkmaz, S., Goksuluk, D., and Zararsiz, G. (2014), “MVN: An R Package for Assessing Multivariate Normality,” *The R Journal*, 6, 151-162.

Liang, K., and Nettleton, D. (2010), “A Hidden Markov Model Approach to Testing Multiple Hypotheses on a Tree-Transformed Gene Ontology Graph,” *Journal of the American Statistical Association*, 105, 1444-1454.

Liu, Q., Dinu, I., Adewale, A. J, Potter, J. D., and Yasui, Y. (2007), “Comparative Evaluation of Gene-Set Analysis Methods,” *BMC Bioinformatics,* 8, 431.

Lu, Y., Liu, P. Y., Xiao, P., and Deng, H. W. (2005), “Hotelling’s T2 Multivariate Profiling for Detecting Differential Expression in Microarrays,” *Bioinformatics*, 21, 3105-3113.

Mardia, K. V. (1970), “Measures of Multivariate Skewness and Kurtosis with Applications,” *Biometrika*, 57, 519-530.

Nettleton, D., Recknor, J., and Reecy, J. M. (2008), “Identification of Differentially Expressed Gene Categories in Microarray Studies Using Nonparametric Multivariate Analysis,” *Bioinformatics*, 24, 192-201.

Qiu, X., and Yakovlev, A. (2006), “Some Comments on Instability of FDR Estimation,” *Journal of Bioinformatics and Computational Biology*, 4, 1057-1068.

Royston, P. (1992), “Approximating the Shapiro-Wilk W-test for Non-Normality.” *Statistics and Computing*, 2, 117-119.

Schäfer, J., and Strimmer, K. (2005), “A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics,” *Statistical Applications in Genetics and Molecular Biology*, 4, Article 32.

She, C. X. (2016), “Pathway-Based Bayesian Integrative Analysis for Genetic Association Studies,” *Master’s thesis*, National Taiwan University, Taiwan.

Sklar, M. (1959), “Fonctions de Répartition à n Dimensions et Leurs Marges,” *Publications de l’Institut de Statistique de Université de Paris,* 8, 229-231.

Song, S., and Black, M. A. (2008), “Microarray-Based Gene Set Analysis: A Comparison of Current Methods,” *BMC Bioinformatics,* 9, 502.

Székely, G. J., and Rizzo, M. L. (2004), “Testing for Equal Distributions in High Dimension,” *InterStat*, November (5).

Székely, G. J., and Rizzo, M. L. (2005), “A New Test for Multivariate Normality,” *Journal of Multivariate Analysis*, 93, 58-80.

Tryputsen, V., Cabrera, J., de Bondt, A., and Amaratunga, D. (2014), “Using Fisher’s Method to Identify Enriched Gene Sets,” *Statistics in Biopharmaceutical Research*, 6:2, 154-162.

Tsai, C. A., and Chen, J. J. (2009), “Multivariate analysis of variance test for gene set analysis,” *Bioinformatics*, 25, 897-903.

Wang Z., Maity A., Hsiao C. K., Voora D., Kaddurah-Daouk R., and Tzeng J. Y. (2015), “Module-based association analysis for omics data with network structure,” *PLoS One*, 10, e0122309.

表一1(A) 乳癌資料集的詳細資訊

紅色粗體代表該組為最後被選擇來分析的樣本組別

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Diseases | | Breast Cancer | | | | | |
| Dataset Series | | GSE15852 | | | | | |
| Platform | | HG-U133A  (GPL96) | | | | | |
| Disease Subtype | | Invasive Ductal Carcinoma(ICD) | | | | | |
| Data Type | | Paired data | | | | | |
| Gene Sets Used in Study | | PI3K-Akt signaling pathway,  Wnt signaling pathway | | | | | |
| Sample Size | | 36 pairs(All: 72) | | | | | |
| Data Summary | **Cancer** | Race:  Malay | | | Race:  Chinese | | |
| Grade  1 | Grade  2 | Grade  3 | Grade  1 | Grade  2 | Grade  3 |
| 3 | **13** | 11 | 3 | 5 | 1 |
| **Normal** | Malay | | | Chinese | | |
| 27 | | | 9 | | |
| Normality Test Results | | Not sure(small sample size) | | | | | |

Grade: Tumor grade

表一2(B) 大腸直腸癌資料集的詳細資訊

紅色粗體代表該組為最後被選擇來分析的樣本組別

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Diseases | | Colorectal Cancer | | | | | | | | | | |
| Dataset Series | | GSE8671 | | | | | | | | | | |
| Platform | | HG-U133 Plus 2  (GPL570) | | | | | | | | | | |
| Disease Subtype | | Adenocarcinoma | | | | | | | | | | |
| Data Type | | Paired data | | | | | | | | | | |
| Gene Sets Used in Study | | Wnt signaling pathway,  Ras signaling pathway,  PI3K-Akt signaling pathway,  p53 signaling pathway,  TGF-beta signaling pathway | | | | | | | | | | |
| Sample Size | | 32 pairs(All: 64) | | | | | | | | | | |
| Data Summary | **Cancer** | Position:  Transversum | | Position:  Ascending  Colon | | Position:  Descending  Colon | | Position:  Rectum | | Position:  Sigmoid  colon | | |
| Size  3cm | Size  3cm | Size  3cm | Size  3cm | Size  3cm | Size  3cm | Size  3cm | Size  3cm | Size  3cm | Size  3cm |
| 1 | 0 | 3 | 1 | 4 | 0 | 4 | 3 | **11** | 5 |
| **Normal** | Transversum | | Ascending  Colon | | Descending  Colon | | Rectum | | Sigmoid  colon | | |
| 1 | | 4 | | 4 | | 7 | | 16 | | |
| Normality Test Results | | Not sure(small sample size) | | | | | | | | | | |

Position: Tumor position, Size: Tumor size, Stage: Tumor staging

表一3(C) 白血病資料集的詳細資訊

紅色粗體代表該組為最後被選擇來分析的樣本組別

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Diseases | | Leukemia | | | |
| Dataset Series | | GSE9476 | | | |
| Platform | | HG-U133A  (GPL96) | | | |
| Disease Subtype | | Acute Myeloid Leukemia(AML) | | | |
| Data Type | | Non-paired data | | | |
| Gene Sets Used in Study | | mTOR signaling pathway,  Jak-STAT signaling pathway,  PI3K-Akt signaling pathway,  Wnt signaling pathway | | | |
| Sample Size | | 64 | | | |
| Data Summary | **Cancer** | BM | | PB | |
| 7 | | **19** | |
| **Normal** | BM CD34+ | PBSC CD34+ | BM | PB |
| 8 | 10 | 10 | **10** |
| Normality Test Results | | It is not normally distributed. | | | |

表一4(D) 肺癌資料集的詳細資訊

紅色粗體代表該組為最後被選擇來分析的樣本組別

Stage: Tumor staging

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Diseases | | Lung Cancer | | |
| Dataset Series | | GSE15852 | | GSE19804 |
| Platform | | HG-U133A  (GPL96) | | HG-U133 Plus 2  (GPL570) |
| Disease Subtype | | Adenocarcinoma | | |
| Data Type | | Paired data | | |
| Gene Sets Used in Study | | MAPK signaling pathway,  Wnt signaling pathway,  Cell cycle pathway,  mTOR signaling pathway,  p53 signaling pathway,  EGFR tyrosine kinase inhibitor resistance,  Jak-STAT signaling pathway | | |
| Sample Size | | 26 pairs(All: 52) | | 47 pairs(All: 94) |
| Data Summary | **Cancer** | Male | Female | Female, early stage |
| 5 | **21** | **47** |
| **Normal** | Male | Female | Female, early stage |
| 5 | **21** | **9** |
| Normality Test Results | | It is not normally distributed. | | |

表一5(E) 卵巢癌資料集的詳細資訊

紅色粗體代表該組為最後被選擇來分析的樣本組別

Grade: Tumor grade, Position: Tumor position, Stage: Tumor staging

|  |  |  |
| --- | --- | --- |
| Diseases | | Ovarian Cancer |
| Dataset Series | | Tcga-ov |
| Platform | | HG-U133A  (GPL96) |
| Disease Subtype | | Serous Cystadenocarcinoma |
| Data Type | | Non-paired data |
| Gene Sets Used in Study | | PI3K-Akt signaling pathway,  mTOR signaling pathway,  Jak-STAT signaling pathway,  p53 signaling pathway |
| Sample Size | | 163 |
| Data Summary | **Cancer** | Survival time 2(year) |
| Grade: 3, Race: white, Position: Bilateral, Stage: III and IV |
| **47** |
| **Normal** | Survival time 2(year) |
| Grade: 3, Race: white, Position: Bilateral, Stage: III and IV |
| **116** |
| Normality Test Results | | It is not normally distributed. |

表二6各個癌症和其相關的Pathway彼此之間的關係

|  |  |  |
| --- | --- | --- |
| Disease (Subtype) | Related Pathway  (Number of Genes) | Relationship |
| Breast Cancer  (Invasive Ductal Carcinoma) | PI3K-Akt signaling  (351) | The HER2/PI3K/Akt pathway plays a causal role in conferring a broad chemoresistance in breast cancer and justifies a strategy for targeting this pathway for enhancing breast cancer chemotherapy (Park, and Kim 2007). |
| Wnt signaling  (143) | The Wnt/β-catenin signaling cascade is an important signal transduction pathway in human cancers. Overexpression of β-catenin and its downstream effector, cyclin D1, is implicated in malignant transformation and acquisition of an invasive tumor phenotype (Prasad et al. 2007). |
| Colorectal Cancer  (Adenoma) | Wnt signaling  (143) | The earliest molecular abnormality in Wnt signaling pathway is mutation in APC, and the gene known to be involved in the hereditary condition familial adenomatous polyposis coli (APC) (Noffsinger 2009). The changes of APC lead to deregulation of the WNT-APC-B-CATENIN signaling pathway and changes in expression in a range of WNT targets including C-MYC (Arends et al. 2013). |
| Ras signaling  (236) | The KRAS activating mutations is thought mostly to occur during the early stages of adenoma progression, and it tends to have an influence on locking the KRAS protein into the active guanosine triphosphate-bound state mediating excessive signaling through the RAS-RAF-MEK-ERK pathway that drives cell proliferation (Arends et al. 2013). |
| PI3K-Akt signaling  (351) | The KRAS protein, produced by KRAS activating mutations, feeds signals into the PI3K pathway with activation of AKT signaling to suppress apoptosis of tumor cell (Arends et al. 2013). |
| p53 signaling  (68) | The BAX mutations are more frequently found and can block p53-induced apoptosis as an alternative to TP53 mutation (Arends et al. 2013). Results from a large number of studies have unequivocally evidence demonstrated that mutant p53 not only plays a pivotal role in the transformation of colorectal cancer, but also contributes to the aggressiveness and invasiveness of colorectal cancer. Thus, it is not surprising that manipulation of the p53 pathway has attracted interest soon after the discovery of p53 gene (Li et al. 2015). |
| TGF-beta signaling  (84) | The most frequent and characteristic genetic changes in colorectal carcinogenesis include alterations to APC, KRAS, SMAD4, TP53. The SMAD4 gene is inactivated by either deletion (in ~60% cancers) of a large part of chromosome 18q where the SMAD4 genes resides or by SMAD4 gene mutation in 35% of cancers. The SMAD4 protein plays a role in signal transduction in the beta transforming growth factor growth inhibitory signaling pathway, so loss of such inhibitory signals allows more autonomous tumor cell growth (Arends et al. 2013). |
| Leukemia  (Acute Myeloid Leukemia) | mTOR signaling  (151) | The aberrant activity of the PI3K/AKT/mTOR pathway, modulating critical cellular functions such as proliferation, metabolism, and survival, has been well established as a major component of AML pathogenesis (Carneiro et al. 2015). |
| Jak-STAT signaling  (162) | In adult AML, gain-of-function mutations that result in constitutive activation of KIT gene have been identified in the extracellular domain and in the activation loop of the kinase domain. Mutations in the KIT gene in activation loop result in amino acid substitutions or small insertions that occur mainly at codon D816 or N822. Codon 816 mutations constitutively activate the PI3K/AKT and STAT3 signaling pathways (Scholl, Gilliland, and Fröhling 2008). |
| PI3K-Akt signaling  (351) | The protein kinase Akt/PKB, a major downstream target of PI3K, is phosphorylated and activated in the majority of primary AML cells and that constitutive activation of the PI3K/ Akt pathway is necessary for both the sustained proliferation (Kubota et al. 2004) and the survival of AML cells (Xu et al. 2003; Grandage et al. 2005; Billottet et al. 2006). |
| Wnt signaling  (143) | A critical role in adult stem cell biology has been identified for the Wnt pathway, an evolutionarily highly conserved signaling cascade critical for normal embryogenesis and hematopoiesis (Griffiths et al. 2010). |
| Lung Cancer  (Lung Adenocarcinoma) | MAPK signaling  (255) | The signal intensity of the MAPK pathway is a critical determinant not only in lung adenocarcinoma development, but also in dictating the nature of the cancer-initiating cell and ultimately the resulting tumor phenotype (Nieto et al. 2017). |
| Wnt signaling  (143) | The Wnt pathway activation, inactivation of various Wnt pathway inhibitors and overexpression of various Wnt pathway components are associated with poor prognosis in non-small cell lung cancer patients (Stewart 2013). |
| Cell cycle  (124) | Genetic alterations, such as mutation and amplification, in lung adenocarcinoma frequently occur in this pathway, and it was found by statistical method (Ding et al. 2008). |
| mTOR signaling  (151) | The deficiency of NF1 mutations has been implicated in RAS-and PI3K-dependent hyperactivation of the mTOR pathway in some patients suffering from lung adenocarcinoma in some study (Ding et al. 2008). |
| p53 signaling  (68) | Some mutations and amplifications of gene members in P53 pathway in lung cancer are strongly associated with smoking, and are common in Adenocarcinoma (Shtivelman et al. 2014). |
| EGFR tyrosine kinase inhibitor resistance  (554) | EGFR is a receptor tyrosine kinase that binds to epidermal growth factor and other growth factor ligands to become activated. EGFR is implicated in a variety of cancers and frequently mutated in Adenocarcinoma. It has been shown to be deregulated by various mechanisms in Adenocarcinoma, including overexpression, amplification, and mutation (Shtivelman et al. 2014). |
| Jak-STAT signaling  (162) | Reduced expression of negative regulators can cause increased pathway activation. In non-small cell lung cancer (NSCLC) tumour samples, expression of SOCS3 is lost due to promoter hypermethylation, an epigenetic change that reduces gene transcription. The impact of this on JAK/STAT signalling pathway activation was validated using a NSCLC cell line, where restoration of SOCS3 expression reduced constitutive STAT3 phosphorylation (Thomas et al. 2015). |
| Ovarian Cancer  (Serous Cystadenocarcinoma) | PI3K-Akt signaling,  (351)  mTOR signaling  (151) | The PI3K/Akt/mTOR pathway is frequently deregulated in ovarian cancer. Array comparative genomic hybridization studies have identified this pathway as the most frequently altered in ovarian cancer. Intrinsic activation of this pathway, via PIK3CA mutations and PTEN loss, has been shown to initiate ovarian tumors in mice, and inhibition of PI3K/mTOR in these models was found to delay tumor growth and prolong survival, thus providing critical proof of concept for the oncogenic relevance of this pathway in ovarian cancer and its potential as a therapeutic target (Cheaib, Auguste, and Leary 2015). |
| Jak-STAT signaling  (162) | JAK/STAT3 is one of the major signaling pathways that is aberrantly activated in ovarian cancer and associated with tumor progression and poor prognosis in ovarian cancer patients (Wen et al. 2014). |
| p53 signaling  (68) | Some study showed that the p53 signaling pathway can characterize a novel subtype of patients with High-grade serous ovarian cancer (Hayano et al. 2014). |

References for the relationships between the pathways and the diseases:

Arends, M. J. (2013), “Pathways of Colorectal Carcinogenesis,” *Applied Immunohistochemistry & Molecular Morphology*, 21, 97-102.

Billottet, C., Grandage, V. L., Gale, R. E., Quattropani, A., Rommel, C., Vanhaesebroeck, B., and Khwaja, A. (2006), “A Selective Inhibitor of the p110δ Isoform of PI3-Kinase Inhibits AML Cell Proliferation and Survival and Increases the Cytotoxic Effects of VP16,” *Oncogene*, 25, 6648-6659.

Carneiro, B. A., Kaplan, J. B., Altman, J. K., Giles, F. J., and Platanias, L. C. (2015), “Targeting mTOR Signaling Pathways and Related Negative Feedback Loops for the Treatment of Acute Myeloid Leukemia,” *Cancer Biology & Therapy*, 16, 648-656.

Cheaib, B., Auguste, A., and Leary, A. (2015), “The PI3K/Akt/mTOR Pathway in Ovarian Cancer: Therapeutic Opportunities and Challenges,” *Chinese Journal of Cancer*, 34, 4-16.

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D. M., Morgan, M. B., Fulton, L., Fulton, R. S., Zhang, Q., Wendl, M. C., Lawrence, M. S., Larson, D. E., Chen, K., Dooling, D. J., Sabo, A., Hawes, A. C., Shen, H., Jhangiani, S. N., Lewis, L. R., Hall, O., Zhu, Y., Mathew, T., Ren, Y., Yao, J., Scherer, S. E., Clerc, K., Metcalf, G. A., Ng, B., Milosavljevic, A., Gonzalez-Garay, M. L., Osborne, J. R., Meyer, R., Shi, X., Tang, Y., Koboldt, D. C., Lin, L., Abbott, R., Miner, T. L., Pohl, C., Fewell, G., Haipek, C., Schmidt, H., Dunford-Shore, B. H., Kraja, A., Crosby, S. D., Sawyer, C. S., Vickery, T., Sander, S., Robinson, J., Winckler, W., Baldwin, J., Chirieac, L. R., Dutt, A., Fennell, T., Hanna, M., Johnson, B. E., Onofrio, R. C., Thomas, R. K., Tonon, G., Weir, B. A., Zhao, X., Ziaugra, L., Zody, M. C., Giordano, T., Orringer, M. B., Roth, J. A., Spitz, M. R., Wistuba, I. I., Ozenberger, B., Good, P. J., Chang, A. C., Beer, D. G., Watson, M. A., Ladanyi, M., Broderick, S., Yoshizawa, A., Travis, W. D., Pao, W., Province, M. A., Weinstock, G. M., Varmus, H. E., Gabriel, S. B., Lander, E. S., Gibbs, R. A., Meyerson, M., and Wilson, R. K. (2008), “Somatic Mutations Affect Key Pathways in Lung Adenocarcinoma,” *Nature*, 455, 1069-1075.

Grandage, V. L., Gale, R. E., Linch, D. C., and Khwaja, A. (2005), “PI3-Kinase/Akt is Constitutively Active in Primary Acute Myeloid Leukaemia Cells and Regulates Survival and Chemoresistance via NF-KappaB, Mapkinase and p53 Pathways,” *Leukemia*, 19, 586-594.

Griffiths, E. A. , Gore, S. D., Hooker, C., McDevitt, M. A., Karp, J. E., Smith, B. D., Mohammad, H. P., Ye, Y., Herman, J. G., and Carraway, H. E. (2010), “Acute Myeloid Leukemia is Characterized by Wnt Pathway Inhibitor Promoter Hypermethylation,” *Leukemia & Lymphoma*, 51, 1711-1719.

Hayano, T., Yokota, Y., Hosomichi, K., Nakaoka, H., Yoshihara, K., Adachi, S., Kashima, K., Tsuda, H., Moriya, T., Tanaka, K., Enomoto, T., and Inoue, I. (2014), “Molecular Characterization of an Intact p53 Pathway Subtype in High-Grade Serous Ovarian Cancer,” *PLoS One*, 9, e114491.

Kubota, Y., Ohnishi, H., Kitanaka, A., Ishida, T., and Tanaka, T. (2004), “Constitutive Activation of PI3K is Involved in the Spontaneous Proliferation of Primary Acute Myeloid Leukemia Cells: Direct Evidence of PI3K Activation,” *Leukemia*, 18, 1438-1440.

Li, X. L., Zhou, J., Chen, Z. R., and Chng, W. J. (2015), “P53 Mutations in Colorectal Cancer-Molecular Pathogenesis and Pharmacological Reactivation,” *World Journal of Gastroenterology*, 21, 84-93.

Nieto, P., Ambrogio, C., Esteban-Burgos, L., Gómez-López, G., Blasco, M. T., Yao, Z., Marais, R., Rosen, N., Chiarle, R., Pisano, D. G., Barbacid, M., and Santamaría, D. (2017), “A Braf Kinase-Inactive Mutant Induces Lung Adenocarcinoma,” *Nature*, 548, 239-243.

Noffsinger, A. E. (2009), “Serrated Polyps and Colorectal Cancer: New Pathway to Malignancy,” *Annual Review of Pathology*, 4, 343-364.

Park, S. S., and Kim, S. W. (2007), “Activated Akt Signaling Pathway in Invasive Ductal Carcinoma of the Breast: Correlation with HER2 Overexpression,” *Oncology Reports*, 18, 139-143.

Prasad, C. P., Gupta, S. D., Rath, G., and Ralhan, R. (2007), “Wnt Signaling Pathway in Invasive Ductal Carcinoma of the Breast: Relationship between β-Catenin, Disheveled and Cyclin D1 Expression,” *Oncology*, 73, 112-117.

Scholl, C., Gilliland, D. G., and Fröhling, S. (2008), “Deregulation of Signaling Pathways in Acute Myeloid Leukemia,” *Seminars in Oncology*, 35, 336-345.

Shtivelman, E., Hensing, T., Simon, G. R., Dennis, P. A., Otterson, G. A., Bueno, R., and Salgia R. (2014), “Molecular Pathways and Therapeutic Targets in Lung Cancer,” *Oncotarget*, 5, 1392-1433.

Stewart, D. J. (2014), “Wnt Signaling Pathway in Non–Small Cell Lung Cancer,” *Journal of the National Cancer Institute*, 106, djt356.

Thomas, S. J., Snowden, J. A., Zeidler, M. P., and Danson, S. J. (2015), “The Role of JAK/STAT Signalling in the Pathogenesis, Prognosis and Treatment of Solid Tumours,” *British Journal of Cancer*, 113, 365-371.

Wen, W., Liang, W., Wu, J., Kowolik, C. M., Buettner, R., Scuto, A., Hsieh, M. Y., Hong, H., Brown, C. E., Forman, S. J., Horne, D., Morgan, R., Wakabayashi, M., Dellinger, T. H., Han, E. S., Yim, J. H., and Jove, R. (2014), “Targeting JAK1/STAT3 Signaling Suppresses Tumor Progression and Metastasis in a Peritoneal Model of Human Ovarian Cancer,” *Molecular Cancer Therapeutics*, 13, 3037-3048.

Xu, Q., Simpson, S. E., Scialla, T. J., Bagg, A., and Carroll, M. (2003), “Survival of Acute Myeloid Leukemia Cells Requires PI3-Kinase Activation,” *Blood*, 102, 972-980.

表三7(A) 多維度常態性檢定對型一誤差率的估計–維度數為2, 4

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | *Mardia* | *HZ* | *Royston* | *energy* |
| 2 | 0.1 | 10 | 0.048 | 0.028 | 0.052 | 0.047 |
| 20 | 0.029 | 0.043 | 0.072 | 0.053 |
| 30 | 0.045 | 0.049 | 0.078 | 0.053 |
| 40 | 0.045 | 0.048 | 0.064 | 0.05 |
| 50 | 0.058 | 0.055 | 0.067 | 0.056 |
| 100 | 0.065 | 0.04 | 0.057 | 0.039 |
| 0.9 | 10 | 0.048 | 0.028 | 0.058 | 0.047 |
| 20 | 0.029 | 0.043 | 0.066 | 0.053 |
| 30 | 0.045 | 0.049 | 0.066 | 0.053 |
| 40 | 0.045 | 0.048 | 0.062 | 0.05 |
| 50 | 0.058 | 0.055 | 0.055 | 0.056 |
| 100 | 0.065 | 0.04 | 0.065 | 0.039 |
| 4 | 0.1 | 10 | 0.021 | 0.031 | 0.064 | 0.054 |
| 20 | 0.017 | 0.047 | 0.066 | 0.055 |
| 30 | 0.044 | 0.037 | 0.074 | 0.05 |
| 40 | 0.06 | 0.041 | 0.073 | 0.042 |
| 50 | 0.064 | 0.033 | 0.072 | 0.034 |
| 100 | 0.083 | 0.05 | 0.064 | 0.044 |
| 0.9 | 10 | 0.021 | 0.031 | 0.064 | 0.054 |
| 20 | 0.017 | 0.047 | 0.062 | 0.055 |
| 30 | 0.044 | 0.037 | 0.062 | 0.05 |
| 40 | 0.06 | 0.041 | 0.06 | 0.042 |
| 50 | 0.064 | 0.033 | 0.056 | 0.034 |
| 100 | 0.083 | 0.05 | 0.072 | 0.044 |

表三8(B) 多維度常態性檢定對型一誤差率的估計–維度數為8, 9, 25, 49

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  | *Mardia* | *HZ* | *Royston* | *energy* |
| 8 | 0.1 | 10 | 0.005 | 0.074 | 0.051 | 0.057 |
| 30 | 0.066 | 0.054 | 0.074 | 0.048 |
| 0.9 | 10 | 0.005 | 0.074 | 0.048 | 0.057 |
| 30 | 0.066 | 0.054 | 0.067 | 0.048 |
| 9 | 0.1 | 10 | 1 | 1 | 0.043 | 0.055 |
| 30 | 0.094 | 0.073 | 0.087 | 0.047 |
| 0.9 | 10 | 1 | 1 | 0.056 | 0.131 |
| 30 | 0.094 | 0.073 | 0.061 | 0.047 |
| 25 | 0.1 | 50 | 0.901 | 0.04 | 0.108 | 0.056 |
| 100 | 0.447 | 0.028 | 0.098 | 0.047 |
| 0.9 | 50 | 0.901 | 0.04 | 0.073 | 0.056 |
| 100 | 0.447 | 0.028 | 0.072 | 0.047 |
| 49 | 0.1 | 50 | 1 | 0.004 | 0.143 | 0.043 |
| 100 | 0.998 | 0 | 0.123 | 0.044 |
| 0.9 | 50 | 1 | 0.037 | 0.091 | 0.111 |
| 100 | 0.998 | 0 | 0.066 | 0.044 |

表四9(A) 多維度常態性檢定檢定力表現–為混合分佈，維度數為2, 4, 8, 9

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | *Mardia* | *HZ* | *Royston* | *energy* |
| 0.5 | 0.5 | 0.9 | 2 | 10 | 0.106 | 0.058 | 0.046 | 0.096 |
| 30 | 0.19 | 0.176 | 0.066 | 0.208 |
| 50 | 0.208 | 0.284 | 0.068 | 0.312 |
| 100 | 0.266 | 0.622 | 0.06 | 0.636 |
| 4 | 10 | 0.152 | 0.154 | 0.052 | 0.22 |
| 30 | 0.54 | 0.63 | 0.098 | 0.698 |
| 50 | 0.61 | 0.906 | 0.088 | 0.914 |
| 100 | 0.708 | 1 | 0.082 | 1 |
| 8 | 10 | 0 | 0.168 | 0.078 | 0.144 |
| 30 | 0.966 | 0.966 | 0.096 | 0.982 |
| 50 | 0.992 | 1 | 0.102 | 1 |
| 100 | 1 | 1 | 0.114 | 1 |
| 9 | 10 | 0 | 1 | 0.094 | 0.084 |
| 30 | 0.986 | 0.976 | 0.116 | 0.988 |
| 50 | 0.996 | 1 | 0.114 | 0.998 |
| 100 | 1 | 1 | 0.098 | 1 |
| 0.5 | 2 | 10 | 0.046 | 0.038 | 0.04 | 0.054 |
| 30 | 0.066 | 0.052 | 0.062 | 0.06 |
| 50 | 0.074 | 0.038 | 0.066 | 0.052 |
| 100 | 0.096 | 0.044 | 0.06 | 0.054 |
| 4 | 10 | 0.032 | 0.034 | 0.06 | 0.048 |
| 30 | 0.108 | 0.078 | 0.076 | 0.088 |
| 50 | 0.132 | 0.058 | 0.058 | 0.086 |
| 100 | 0.148 | 0.084 | 0.056 | 0.098 |
| 8 | 10 | 0 | 0.066 | 0.064 | 0.054 |
| 30 | 0.182 | 0.076 | 0.08 | 0.108 |
| 50 | 0.274 | 0.092 | 0.094 | 0.17 |
| 100 | 0.376 | 0.098 | 0.08 | 0.278 |
| 9 | 10 | 0 | 1 | 0.052 | 0.044 |
| 30 | 0.198 | 0.086 | 0.08 | 0.126 |
| 50 | 0.262 | 0.098 | 0.076 | 0.182 |
| 100 | 0.404 | 0.106 | 0.06 | 0.298 |
| 0.1 | 0.9 | 0.9 | 2 | 10 | 0.124 | 0.09 | 0.036 | 0.114 |
| 30 | 0.326 | 0.218 | 0.068 | 0.274 |
| 50 | 0.428 | 0.248 | 0.062 | 0.336 |
| 100 | 0.6 | 0.458 | 0.062 | 0.592 |
| 4 | 10 | 0.114 | 0.082 | 0.066 | 0.132 |
| 30 | 0.726 | 0.422 | 0.074 | 0.618 |
| 50 | 0.858 | 0.61 | 0.08 | 0.79 |
| 100 | 0.972 | 0.85 | 0.086 | 0.966 |
| 8 | 10 | 0 | 0.088 | 0.042 | 0.072 |
| 30 | 0.894 | 0.426 | 0.098 | 0.808 |
| 50 | 0.976 | 0.692 | 0.102 | 0.946 |
| 100 | 1 | 0.954 | 0.096 | 0.998 |
| 9 | 10 | 0 | 1 | 0.07 | 0.14 |
| 30 | 0.866 | 0.386 | 0.094 | 0.8 |
| 50 | 0.976 | 0.696 | 0.06 | 0.956 |
| 100 | 1 | 0.956 | 0.108 | 1 |
| 0.5 | 2 | 10 | 0.036 | 0.022 | 0.036 | 0.036 |
| 30 | 0.066 | 0.046 | 0.056 | 0.046 |
| 50 | 0.058 | 0.048 | 0.072 | 0.056 |
| 100 | 0.094 | 0.04 | 0.048 | 0.048 |
| 4 | 10 | 0.028 | 0.032 | 0.052 | 0.046 |
| 30 | 0.078 | 0.058 | 0.06 | 0.082 |
| 50 | 0.09 | 0.048 | 0.07 | 0.074 |
| 100 | 0.112 | 0.07 | 0.06 | 0.074 |
| 8 | 10 | 0 | 0.074 | 0.038 | 0.064 |
| 30 | 0.114 | 0.046 | 0.11 | 0.076 |
| 50 | 0.2 | 0.058 | 0.092 | 0.11 |
| 100 | 0.27 | 0.078 | 0.076 | 0.164 |
| 9 | 10 | 0 | 1 | 0.062 | 0.076 |
| 30 | 0.134 | 0.062 | 0.078 | 0.076 |
| 50 | 0.208 | 0.072 | 0.104 | 0.126 |
| 100 | 0.302 | 0.066 | 0.09 | 0.164 |

表四10(B) 多維度常態性檢定檢定力表現–為混合分佈，維度數為25, 49

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | *Mardia* | *HZ* | *Royston* | *energy* |
| 0.5 | 0.5 | 0.9 | 25 | 50 | 1 | 1 | 0.166 | 1 |
| 100 | 1 | 1 | 0.174 | 1 |
| 49 | 50 | 0 | 0.012 | 0.218 | 0.06 |
| 100 | 1 | 1 | 0.226 | 1 |
| 0.5 | 25 | 50 | 0.634 | 0.036 | 0.096 | 0.592 |
| 100 | 0.992 | 0.164 | 0.102 | 0.96 |
| 49 | 50 | 0 | 0.006 | 0.152 | 0.03 |
| 100 | 1 | 0.012 | 0.16 | 1 |
| 0.1 | 0.9 | 0.9 | 25 | 50 | 0.966 | 0.088 | 0.102 | 0.958 |
| 100 | 1 | 0.744 | 0.096 | 1 |
| 49 | 50 | 0 | 0.036 | 0.094 | 0.09 |
| 100 | 1 | 0.03 | 0.078 | 1 |
| 0.5 | 25 | 50 | 0.25 | 0.03 | 0.126 | 0.23 |
| 100 | 0.746 | 0.04 | 0.138 | 0.534 |
| 49 | 50 | 0 | 0.006 | 0.204 | 0.04 |
| 100 | 0.802 | 0 | 0.17 | 0.748 |

表五11(A) 多維度常態性檢定檢定力表現–為t分佈，維度數為2, 4, 8, 9

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | *Mardia* | *HZ* | *Royston* | *energy* |
| 0.9 | 3 | 2 | 10 | 0.278 | 0.196 | 0.224 | 0.238 |
| 30 | 0.718 | 0.602 | 0.572 | 0.656 |
| 50 | 0.872 | 0.804 | 0.772 | 0.844 |
| 100 | 0.996 | 0.962 | 0.924 | 0.968 |
| 4 | 10 | 0.242 | 0.172 | 0.266 | 0.252 |
| 30 | 0.874 | 0.78 | 0.614 | 0.884 |
| 50 | 0.99 | 0.952 | 0.812 | 0.98 |
| 100 | 1 | 0.998 | 0.972 | 1 |
| 8 | 10 | 0.002 | 0.13 | 0.262 | 0.116 |
| 30 | 0.97 | 0.864 | 0.638 | 0.976 |
| 50 | 1 | 0.99 | 0.824 | 1 |
| 100 | 1 | 1 | 0.978 | 1 |
| 9 | 10 | 1 | 1 | 0.25 | 0.216 |
| 30 | 0.968 | 0.864 | 0.67 | 0.982 |
| 50 | 1 | 0.998 | 0.85 | 1 |
| 100 | 1 | 1 | 0.972 | 1 |
| 9 | 2 | 10 | 0.104 | 0.064 | 0.074 | 0.082 |
| 30 | 0.192 | 0.112 | 0.18 | 0.158 |
| 50 | 0.322 | 0.158 | 0.232 | 0.212 |
| 100 | 0.562 | 0.27 | 0.38 | 0.356 |
| 4 | 10 | 0.068 | 0.052 | 0.102 | 0.108 |
| 30 | 0.31 | 0.172 | 0.22 | 0.302 |
| 50 | 0.524 | 0.24 | 0.27 | 0.396 |
| 100 | 0.862 | 0.44 | 0.414 | 0.666 |
| 8 | 10 | 0.002 | 0.098 | 0.102 | 0.086 |
| 30 | 0.364 | 0.186 | 0.184 | 0.482 |
| 50 | 0.756 | 0.358 | 0.256 | 0.704 |
| 100 | 0.978 | 0.732 | 0.408 | 0.948 |
| 9 | 10 | 1 | 1 | 0.094 | 0.138 |
| 30 | 0.334 | 0.182 | 0.204 | 0.49 |
| 50 | 0.81 | 0.418 | 0.306 | 0.794 |
| 100 | 0.994 | 0.726 | 0.432 | 0.97 |
| 15 | 2 | 10 | 0.07 | 0.056 | 0.056 | 0.074 |
| 30 | 0.116 | 0.096 | 0.14 | 0.12 |
| 50 | 0.18 | 0.088 | 0.152 | 0.12 |
| 100 | 0.348 | 0.088 | 0.222 | 0.144 |
| 4 | 10 | 0.042 | 0.052 | 0.092 | 0.08 |
| 30 | 0.16 | 0.1 | 0.144 | 0.176 |
| 50 | 0.298 | 0.112 | 0.174 | 0.218 |
| 100 | 0.536 | 0.206 | 0.256 | 0.332 |
| 8 | 10 | 0.004 | 0.104 | 0.084 | 0.086 |
| 30 | 0.154 | 0.096 | 0.132 | 0.248 |
| 50 | 0.468 | 0.152 | 0.158 | 0.422 |
| 100 | 0.814 | 0.352 | 0.24 | 0.674 |
| 9 | 10 | 1 | 1 | 0.074 | 0.156 |
| 30 | 0.14 | 0.094 | 0.144 | 0.252 |
| 50 | 0.478 | 0.134 | 0.178 | 0.46 |
| 100 | 0.832 | 0.346 | 0.228 | 0.712 |
| 0 | 3 | 2 | 10 | 0.278 | 0.196 | 0.27 | 0.238 |
| 30 | 0.718 | 0.602 | 0.712 | 0.656 |
| 50 | 0.872 | 0.804 | 0.85 | 0.844 |
| 100 | 0.996 | 0.962 | 0.982 | 0.968 |
| 4 | 10 | 0.242 | 0.172 | 0.38 | 0.252 |
| 30 | 0.874 | 0.78 | 0.84 | 0.884 |
| 50 | 0.99 | 0.952 | 0.96 | 0.98 |
| 100 | 1 | 0.998 | 1 | 1 |
| 8 | 10 | 0.002 | 0.13 | 0.444 | 0.116 |
| 30 | 0.97 | 0.864 | 0.956 | 0.976 |
| 50 | 1 | 0.99 | 0.998 | 1 |
| 100 | 1 | 1 | 1 | 1 |
| 9 | 10 | 1 | 1 | 0.446 | 0.07 |
| 30 | 0.968 | 0.864 | 0.948 | 0.982 |
| 50 | 1 | 0.998 | 0.998 | 1 |
| 100 | 1 | 1 | 1 | 1 |
| 9 | 2 | 10 | 0.104 | 0.064 | 0.084 | 0.082 |
| 30 | 0.192 | 0.112 | 0.228 | 0.158 |
| 50 | 0.322 | 0.158 | 0.296 | 0.212 |
| 100 | 0.562 | 0.27 | 0.474 | 0.356 |
| 4 | 10 | 0.068 | 0.052 | 0.116 | 0.108 |
| 30 | 0.31 | 0.172 | 0.338 | 0.302 |
| 50 | 0.524 | 0.24 | 0.44 | 0.396 |
| 100 | 0.862 | 0.44 | 0.618 | 0.666 |
| 8 | 10 | 0.002 | 0.098 | 0.116 | 0.086 |
| 30 | 0.364 | 0.186 | 0.44 | 0.482 |
| 50 | 0.756 | 0.358 | 0.582 | 0.704 |
| 100 | 0.978 | 0.732 | 0.854 | 0.948 |
| 9 | 10 | 1 | 1 | 0.12 | 0.066 |
| 30 | 0.334 | 0.182 | 0.428 | 0.49 |
| 50 | 0.81 | 0.418 | 0.634 | 0.794 |
| 100 | 0.994 | 0.726 | 0.864 | 0.97 |
| 15 | 2 | 10 | 0.07 | 0.056 | 0.076 | 0.074 |
| 30 | 0.116 | 0.096 | 0.156 | 0.12 |
| 50 | 0.18 | 0.088 | 0.192 | 0.12 |
| 100 | 0.348 | 0.088 | 0.276 | 0.144 |
| 4 | 10 | 0.042 | 0.052 | 0.098 | 0.08 |
| 30 | 0.16 | 0.1 | 0.2 | 0.176 |
| 50 | 0.298 | 0.112 | 0.266 | 0.218 |
| 100 | 0.536 | 0.206 | 0.364 | 0.332 |
| 8 | 10 | 0.004 | 0.104 | 0.08 | 0.086 |
| 30 | 0.154 | 0.096 | 0.256 | 0.248 |
| 50 | 0.468 | 0.152 | 0.35 | 0.422 |
| 100 | 0.814 | 0.352 | 0.528 | 0.674 |
| 9 | 10 | 1 | 1 | 0.082 | 0.052 |
| 30 | 0.14 | 0.094 | 0.21 | 0.252 |
| 50 | 0.478 | 0.134 | 0.354 | 0.46 |
| 100 | 0.832 | 0.346 | 0.542 | 0.712 |

表五12(B) 多維度常態性檢定檢定力表現–為t分佈，維度數為25, 49

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | *Mardia* | *HZ* | *Royston* | *energy* |
| 0.9 | 3 | 25 | 50 | 1 | 0.92 | 0.864 | 1 |
| 100 | 1 | 1 | 0.99 | 1 |
| 49 | 50 | 1 | 0.048 | 0.862 | 0.166 |
| 100 | 1 | 1 | 0.992 | 1 |
| 9 | 25 | 50 | 0.686 | 0.156 | 0.268 | 0.976 |
| 100 | 1 | 0.892 | 0.41 | 1 |
| 49 | 50 | 1 | 0.044 | 0.298 | 0.1 |
| 100 | 1 | 0.328 | 0.398 | 1 |
| 15 | 25 | 50 | 0.282 | 0.068 | 0.164 | 0.818 |
| 100 | 0.996 | 0.356 | 0.232 | 0.994 |
| 49 | 50 | 1 | 0.036 | 0.19 | 0.114 |
| 100 | 0.99 | 0.064 | 0.226 | 1 |
| 0 | 3 | 25 | 50 | 1 | 0.92 | 1 | 1 |
| 100 | 1 | 1 | 1 | 1 |
| 49 | 50 | 1 | 0.012 | 1 | 0.036 |
| 100 | 1 | 1 | 1 | 1 |
| 9 | 25 | 50 | 0.686 | 0.156 | 0.904 | 0.976 |
| 100 | 1 | 0.892 | 0.984 | 1 |
| 49 | 50 | 1 | 0.002 | 0.968 | 0.038 |
| 100 | 1 | 0.328 | 0.998 | 1 |
| 15 | 25 | 50 | 0.282 | 0.068 | 0.652 | 0.818 |
| 100 | 0.996 | 0.356 | 0.848 | 0.994 |
| 49 | 50 | 1 | 0 | 0.824 | 0.048 |
| 100 | 0.99 | 0.064 | 0.956 | 1 |

表六13(A) 各個癌症資料集常態性檢定結果–單一探針，2個基因，病例組

(針對每個基因進行單一探針挑選，並在500次檢定裡，每次檢定抽取2個基因探針資料來進行分析所得到的病例組(Cancer)的檢定拒絕率)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Diseases | Platform | Pathway | Mardia | HZ | Royston | energy |
| Breast Cancer | GPL96 | PI3K-Akt | 0.146 | 0.088 | 0.134 | 0.112 |
| Wnt | 0.07 | 0.066 | 0.062 | 0.08 |
| Colorectal Cancer | GPL570 | P53 | 0.074 | 0.05 | 0.07 | 0.074 |
| PI3K-Akt | 0.112 | 0.066 | 0.102 | 0.094 |
| Ras | 0.102 | 0.064 | 0.108 | 0.094 |
| TGF- | 0.1 | 0.08 | 0.072 | 0.118 |
| Wnt | 0.096 | 0.074 | 0.104 | 0.086 |
| Leukemia | GPL96 | Jak-STAT | 0.352 | 0.266 | 0.45 | 0.316 |
| mTOR | 0.262 | 0.192 | 0.284 | 0.208 |
| PI3K-Akt | 0.322 | 0.282 | 0.364 | 0.32 |
| Wnt | 0.38 | 0.292 | 0.39 | 0.318 |
| Lung Cancer | GPL96 | Cell cycle | 0.314 | 0.464 | 0.548 | 0.49 |
| EGFR | 0.246 | 0.288 | 0.358 | 0.32 |
| Jak-STAT | 0.362 | 0.436 | 0.532 | 0.464 |
| MAPK | 0.292 | 0.394 | 0.488 | 0.428 |
| mTOR | 0.192 | 0.32 | 0.412 | 0.35 |
| P53 | 0.294 | 0.468 | 0.49 | 0.484 |
| Wnt | 0.202 | 0.384 | 0.468 | 0.418 |
| GPL570 | Cell cycle | 0.69 | 0.63 | 0.736 | 0.682 |
| EGFR | 0.55 | 0.532 | 0.582 | 0.56 |
| Jak-STAT | 0.58 | 0.534 | 0.632 | 0.574 |
| MAPK | 0.6 | 0.58 | 0.656 | 0.622 |
| mTOR | 0.582 | 0.51 | 0.63 | 0.56 |
| P53 | 0.642 | 0.58 | 0.68 | 0.616 |
| Wnt | 0.598 | 0.548 | 0.68 | 0.612 |
| Ovarian Cancer | GPL96 | mTOR | 0.578 | 0.472 | 0.612 | 0.538 |
| Jak-STAT | 0.75 | 0.62 | 0.788 | 0.69 |
| P53 | 0.594 | 0.48 | 0.634 | 0.558 |
| PI3K-Akt | 0.656 | 0.564 | 0.72 | 0.63 |

表六14(B) 各個癌症資料集常態性檢定結果–單一探針，2個基因，對照組

(針對每個基因進行單一探針挑選，並在500次檢定裡，每次檢定抽取2個基因探針資料來進行分析所得到的對照組(Normal or Adjacent Normal)的檢定拒絕率)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Diseases | Platform | Pathway | Mardia | HZ | Royston | energy |
| Breast Cancer | GPL96 | PI3K-Akt | 0.114 | 0.072 | 0.112 | 0.09 |
| Wnt | 0.152 | 0.11 | 0.134 | 0.118 |
| Colorectal Cancer | GPL570 | P53 | 0.082 | 0.082 | 0.098 | 0.096 |
| PI3K-Akt | 0.11 | 0.076 | 0.134 | 0.096 |
| Ras | 0.118 | 0.06 | 0.134 | 0.092 |
| TGF- | 0.078 | 0.036 | 0.076 | 0.046 |
| Wnt | 0.104 | 0.092 | 0.134 | 0.128 |
| Leukemia | GPL96 | Jak-STAT | 0.272 | 0.162 | 0.226 | 0.2 |
| mTOR | 0.296 | 0.202 | 0.282 | 0.238 |
| PI3K-Akt | 0.284 | 0.184 | 0.312 | 0.234 |
| Wnt | 0.24 | 0.144 | 0.254 | 0.186 |
| Lung Cancer | GPL96 | Cell cycle | 0.188 | 0.214 | 0.336 | 0.25 |
| EGFR | 0.21 | 0.214 | 0.304 | 0.25 |
| Jak-STAT | 0.154 | 0.176 | 0.278 | 0.206 |
| MAPK | 0.214 | 0.174 | 0.292 | 0.222 |
| mTOR | 0.214 | 0.188 | 0.304 | 0.234 |
| P53 | 0.192 | 0.164 | 0.294 | 0.222 |
| Wnt | 0.166 | 0.168 | 0.252 | 0.224 |
| GPL570 | Cell cycle | 0.648 | 0.642 | 0.696 | 0.678 |
| EGFR | 0.684 | 0.694 | 0.766 | 0.746 |
| Jak-STAT | 0.702 | 0.652 | 0.786 | 0.706 |
| MAPK | 0.652 | 0.63 | 0.718 | 0.668 |
| mTOR | 0.726 | 0.646 | 0.764 | 0.716 |
| P53 | 0.656 | 0.666 | 0.742 | 0.698 |
| Wnt | 0.724 | 0.694 | 0.76 | 0.734 |
| Ovarian Cancer | GPL96 | mTOR | 0.806 | 0.642 | 0.838 | 0.75 |
| Jak-STAT | 0.946 | 0.894 | 0.952 | 0.906 |
| P53 | 0.856 | 0.678 | 0.862 | 0.778 |
| PI3K-Akt | 0.926 | 0.824 | 0.926 | 0.872 |

表七15(A) 各個癌症資料集常態性檢定結果–單一探針，4個基因，病例組

(針對每個基因進行單一探針挑選，並在500次檢定裡，每次檢定抽取4個基因探針資料來進行分析所得到的病例組(Cancer)的檢定拒絕率)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Diseases | Platform | Pathway | Mardia | HZ | Royston | energy |
| Breast Cancer | GPL96 | PI3K-Akt | 0.12 | 0.094 | 0.19 | 0.118 |
| Wnt | 0.068 | 0.066 | 0.092 | 0.086 |
| Colorectal Cancer | GPL570 | P53 | 0.042 | 0.034 | 0.058 | 0.066 |
| PI3K-Akt | 0.04 | 0.066 | 0.15 | 0.086 |
| Ras | 0.07 | 0.036 | 0.11 | 0.08 |
| TGF- | 0.078 | 0.07 | 0.128 | 0.104 |
| Wnt | 0.07 | 0.068 | 0.112 | 0.102 |
| Leukemia | GPL96 | Jak-STAT | 0.424 | 0.26 | 0.62 | 0.384 |
| mTOR | 0.304 | 0.188 | 0.446 | 0.272 |
| PI3K-Akt | 0.416 | 0.29 | 0.55 | 0.348 |
| Wnt | 0.446 | 0.322 | 0.594 | 0.4 |
| Lung Cancer | GPL96 | Cell cycle | 0.38 | 0.582 | 0.752 | 0.7 |
| EGFR | 0.266 | 0.316 | 0.534 | 0.456 |
| Jak-STAT | 0.354 | 0.482 | 0.728 | 0.598 |
| MAPK | 0.348 | 0.47 | 0.692 | 0.568 |
| mTOR | 0.184 | 0.316 | 0.568 | 0.422 |
| P53 | 0.37 | 0.538 | 0.69 | 0.61 |
| Wnt | 0.194 | 0.4 | 0.7 | 0.452 |
| GPL570 | Cell cycle | 0.83 | 0.742 | 0.916 | 0.862 |
| EGFR | 0.786 | 0.628 | 0.8 | 0.784 |
| Jak-STAT | 0.778 | 0.632 | 0.88 | 0.788 |
| MAPK | 0.788 | 0.7 | 0.872 | 0.826 |
| mTOR | 0.726 | 0.638 | 0.832 | 0.764 |
| P53 | 0.792 | 0.702 | 0.894 | 0.812 |
| Wnt | 0.782 | 0.654 | 0.878 | 0.792 |
| Ovarian Cancer | GPL96 | mTOR | 0.508 | 0.388 | 0.698 | 0.536 |
| Jak-STAT | 0.648 | 0.528 | 0.826 | 0.666 |
| P53 | 0.516 | 0.472 | 0.716 | 0.606 |
| PI3K-Akt | 0.558 | 0.472 | 0.784 | 0.586 |

表七16(B) 各個癌症資料集常態性檢定結果–單一探針，4個基因，對照組

(針對每個基因進行單一探針挑選，並在500次檢定裡，每次檢定抽取4個基因探針資料來進行分析所得到的對照組(Normal or Adjacent Normal)的檢定拒絕率)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Diseases | Platform | Pathway | Mardia | HZ | Royston | energy |
| Breast Cancer | GPL96 | PI3K-Akt | 0.142 | 0.058 | 0.182 | 0.118 |
|  |  |  |  |  |
| Colorectal Cancer | GPL570 | P53 | 0.04 | 0.056 | 0.148 | 0.078 |
|  |  |  |  |  |
| Ras | 0.062 | 0.054 | 0.17 | 0.074 |
|  |  |  |  |  |
| Wnt | 0.072 | 0.054 | 0.16 | 0.1 |
| Leukemia | GPL96 |  |  |  |  |  |
| mTOR | 0.136 | 0.114 | 0.402 | 0.178 |
|  |  |  |  |  |
| Wnt | 0.152 | 0.134 | 0.298 | 0.206 |
| Lung Cancer | GPL96 | Cell cycle | 0.174 | 0.22 | 0.428 | 0.346 |
| EGFR | 0.228 | 0.262 | 0.436 | 0.358 |
| Jak-STAT | 0.172 | 0.162 | 0.394 | 0.246 |
| MAPK | 0.186 | 0.216 | 0.414 | 0.298 |
| mTOR | 0.23 | 0.208 | 0.43 | 0.33 |
| P53 | 0.182 | 0.172 | 0.45 | 0.256 |
| Wnt | 0.196 | 0.16 | 0.376 | 0.264 |
| GPL570 | Cell cycle | 0.87 | 0.838 | 0.918 | 0.892 |
| EGFR | 0.904 | 0.872 | 0.928 | 0.93 |
| Jak-STAT | 0.844 | 0.796 | 0.956 | 0.88 |
| MAPK | 0.862 | 0.818 | 0.91 | 0.888 |
| mTOR | 0.908 | 0.854 | 0.932 | 0.942 |
| P53 | 0.818 | 0.802 | 0.92 | 0.89 |
| Wnt | 0.906 | 0.842 | 0.912 | 0.908 |
| Ovarian Cancer | GPL96 | mTOR | 0.828 | 0.654 | 0.888 | 0.826 |
| Jak-STAT | 0.948 | 0.846 | 0.984 | 0.928 |
| P53 | 0.748 | 0.594 | 0.87 | 0.732 |
| PI3K-Akt | 0.932 | 0.804 | 0.966 | 0.92 |

表八17(A) 各個癌症資料集常態性檢定結果–單一探針(log)，4個基因，病例組

(針對每個基因進行單一探針挑選，接著將基因表現量取，並在500次檢定裡，每次檢定抽取4個基因探針資料來進行分析所得到的病例組(Cancer)的檢定拒絕率)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Diseases | Platform | Pathway | Mardia | HZ | Royston | energy |
| Breast Cancer | GPL96 | PI3K-Akt | 0.12 | 0.09 | 0.182 | 0.116 |
|  |  |  |  |  |
| Colorectal Cancer | GPL570 | P53 | 0.048 | 0.04 | 0.07 | 0.064 |
|  |  |  |  |  |
| Ras | 0.068 | 0.022 | 0.106 | 0.064 |
|  |  |  |  |  |
| Wnt | 0.062 | 0.056 | 0.088 | 0.088 |
| Leukemia | GPL96 |  |  |  |  |  |
| mTOR | 0.29 | 0.168 | 0.434 | 0.228 |
|  |  |  |  |  |
| Wnt | 0.416 | 0.288 | 0.564 | 0.364 |
| Lung Cancer | GPL96 | Cell cycle | 0.304 | 0.546 | 0.712 | 0.65 |
| EGFR | 0.248 | 0.32 | 0.504 | 0.464 |
| Jak-STAT | 0.286 | 0.422 | 0.658 | 0.544 |
| MAPK | 0.282 | 0.442 | 0.628 | 0.546 |
| mTOR | 0.156 | 0.29 | 0.524 | 0.396 |
| P53 | 0.298 | 0.484 | 0.678 | 0.564 |
| Wnt | 0.174 | 0.378 | 0.66 | 0.444 |
| GPL570 | Cell cycle | 0.77 | 0.678 | 0.854 | 0.796 |
| EGFR | 0.766 | 0.674 | 0.856 | 0.826 |
| Jak-STAT | 0.694 | 0.59 | 0.872 | 0.706 |
| MAPK | 0.734 | 0.67 | 0.856 | 0.782 |
| mTOR | 0.704 | 0.636 | 0.828 | 0.752 |
| P53 | 0.678 | 0.604 | 0.792 | 0.712 |
| Wnt | 0.722 | 0.652 | 0.86 | 0.768 |
| Ovarian Cancer | GPL96 | mTOR | 0.362 | 0.306 | 0.624 | 0.38 |
| Jak-STAT | 0.46 | 0.388 | 0.692 | 0.464 |
| P53 | 0.558 | 0.474 | 0.742 | 0.584 |
| PI3K-Akt | 0.448 | 0.384 | 0.696 | 0.486 |

表八18(B) 各個癌症資料集常態性檢定結果–單一探針(log)，4個基因，對照組

(針對每個基因進行單一探針挑選，接著將基因表現量取，並在500次檢定裡，每次檢定抽取4個基因探針資料來進行分析所得到的對照組(Normal or Adjacent Normal)的檢定拒絕率)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Diseases | Platform | Pathway | Mardia | HZ | Royston | energy |
| Breast Cancer | GPL96 | PI3K-Akt | 0.146 | 0.06 | 0.164 | 0.112 |
|  |  |  |  |  |
| Colorectal Cancer | GPL570 | P53 | 0.04 | 0.056 | 0.14 | 0.078 |
|  |  |  |  |  |
| Ras | 0.06 | 0.056 | 0.154 | 0.082 |
|  |  |  |  |  |
| Wnt | 0.068 | 0.05 | 0.148 | 0.09 |
| Leukemia | GPL96 |  |  |  |  |  |
| mTOR | 0.128 | 0.114 | 0.396 | 0.182 |
|  |  |  |  |  |
| Wnt | 0.152 | 0.146 | 0.312 | 0.208 |
| Lung Cancer | GPL96 | Cell cycle | 0.178 | 0.228 | 0.516 | 0.37 |
| EGFR | 0.254 | 0.268 | 0.446 | 0.384 |
| Jak-STAT | 0.162 | 0.162 | 0.374 | 0.24 |
| MAPK | 0.156 | 0.196 | 0.38 | 0.284 |
| mTOR | 0.25 | 0.216 | 0.458 | 0.346 |
| P53 | 0.142 | 0.158 | 0.426 | 0.226 |
| Wnt | 0.178 | 0.148 | 0.372 | 0.254 |
| GPL570 | Cell cycle | 0.866 | 0.818 | 0.91 | 0.876 |
| EGFR | 0.898 | 0.898 | 0.938 | 0.942 |
| Jak-STAT | 0.828 | 0.794 | 0.938 | 0.876 |
| MAPK | 0.84 | 0.826 | 0.9 | 0.91 |
| mTOR | 0.918 | 0.864 | 0.932 | 0.932 |
| P53 | 0.834 | 0.796 | 0.918 | 0.874 |
| Wnt | 0.906 | 0.848 | 0.92 | 0.908 |
| Ovarian Cancer | GPL96 | mTOR | 0.754 | 0.582 | 0.85 | 0.738 |
| Jak-STAT | 0.874 | 0.738 | 0.936 | 0.874 |
| P53 | 0.774 | 0.652 | 0.88 | 0.764 |
| PI3K-Akt | 0.866 | 0.716 | 0.944 | 0.836 |

表九19(A) 各個癌症資料集常態性檢定結果–2個探針，病例組

(不針對每個基因進行單一探針挑選，並在500次檢定裡，每次檢定抽取2個基因探針資料來進行分析所得到的病例組(Cancer)的檢定拒絕率)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Diseases | Platform | Pathway | Mardia | HZ | Royston | energy |
| Breast Cancer | GPL96 | PI3K-Akt | 0.082 | 0.066 | 0.102 | 0.076 |
|  |  |  |  |  |
| Colorectal Cancer | GPL570 | P53 | 0.084 | 0.05 | 0.06 | 0.064 |
|  |  |  |  |  |
| Ras | 0.126 | 0.066 | 0.128 | 0.094 |
|  |  |  |  |  |
| Wnt | 0.072 | 0.068 | 0.096 | 0.086 |
| Leukemia | GPL96 |  |  |  |  |  |
| mTOR | 0.262 | 0.156 | 0.272 | 0.174 |
|  |  |  |  |  |
| Wnt | 0.316 | 0.258 | 0.366 | 0.278 |
| Lung Cancer | GPL96 | Cell cycle | 0.304 | 0.42 | 0.54 | 0.45 |
| EGFR | 0.34 | 0.418 | 0.484 | 0.448 |
| Jak-STAT | 0.344 | 0.388 | 0.504 | 0.434 |
| MAPK | 0.31 | 0.4 | 0.488 | 0.444 |
| mTOR | 0.238 | 0.356 | 0.424 | 0.398 |
| P53 | 0.278 | 0.43 | 0.476 | 0.448 |
| Wnt | 0.224 | 0.382 | 0.45 | 0.378 |
| GPL570 | Cell cycle | 0.644 | 0.53 | 0.674 | 0.598 |
| EGFR | 0.532 | 0.416 | 0.556 | 0.478 |
| Jak-STAT | 0.488 | 0.414 | 0.534 | 0.484 |
| MAPK | 0.582 | 0.462 | 0.626 | 0.544 |
| mTOR | 0.53 | 0.462 | 0.59 | 0.514 |
| P53 | 0.666 | 0.53 | 0.68 | 0.596 |
| Wnt | 0.568 | 0.458 | 0.618 | 0.528 |

表九20(B) 各個癌症資料集常態性檢定結果–2個探針，對照組

(不針對每個基因進行單一探針挑選，並在500次檢定裡，每次檢定抽取2個基因探針資料來進行分析所得到的對照組(Normal or Adjacent Normal)的檢定拒絕率)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Diseases | Platform | Pathway | Mardia | HZ | Royston | energy |
| Breast Cancer | GPL96 | PI3K-Akt | 0.166 | 0.104 | 0.172 | 0.124 |
|  |  |  |  |  |
| Colorectal Cancer | GPL570 | P53 | 0.088 | 0.056 | 0.082 | 0.068 |
|  |  |  |  |  |
| Ras | 0.104 | 0.076 | 0.156 | 0.096 |
| TGF- | 0.086 | 0.04 | 0.086 | 0.058 |
| Wnt | 0.088 | 0.068 | 0.11 | 0.09 |
| Leukemia | GPL96 | Jak-STAT | 0.272 | 0.172 | 0.266 | 0.21 |
| mTOR | 0.226 | 0.14 | 0.246 | 0.17 |
| PI3K-Akt | 0.274 | 0.194 | 0.312 | 0.224 |
| Wnt | 0.246 | 0.132 | 0.214 | 0.15 |
| Lung Cancer | GPL96 | Cell cycle | 0.188 | 0.206 | 0.302 | 0.252 |
| EGFR | 0.218 | 0.178 | 0.304 | 0.224 |
| Jak-STAT | 0.174 | 0.208 | 0.328 | 0.238 |
| MAPK | 0.132 | 0.142 | 0.21 | 0.174 |
| mTOR | 0.22 | 0.214 | 0.338 | 0.268 |
| P53 | 0.204 | 0.198 | 0.324 | 0.242 |
| Wnt | 0.172 | 0.146 | 0.238 | 0.166 |
| GPL570 | Cell cycle | 0.604 | 0.552 | 0.66 | 0.61 |
| EGFR | 0.646 | 0.612 | 0.728 | 0.666 |
| Jak-STAT | 0.608 | 0.562 | 0.69 | 0.61 |
| MAPK | 0.64 | 0.578 | 0.71 | 0.626 |
| mTOR | 0.662 | 0.626 | 0.726 | 0.668 |
| P53 | 0.658 | 0.612 | 0.69 | 0.668 |
| Wnt | 0.668 | 0.612 | 0.714 | 0.666 |

表十21(A) 各個癌症資料集常態性檢定結果–4個探針，病例組

(不針對每個基因進行單一探針挑選，並在500次檢定裡，每次檢定抽取4個基因探針資料來進行分析所得到的病例組(Cancer)的檢定拒絕率)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Diseases | Platform | Pathway | Mardia | HZ | Royston | energy |
| Breast Cancer | GPL96 | PI3K-Akt | 0.082 | 0.044 | 0.128 | 0.068 |
| Wnt | 0.102 | 0.088 | 0.204 | 0.11 |
| Colorectal Cancer | GPL570 | P53 | 0.038 | 0.03 | 0.064 | 0.048 |
| PI3K-Akt | 0.066 | 0.044 | 0.12 | 0.07 |
| Ras | 0.066 | 0.048 | 0.156 | 0.102 |
| TGF- | 0.08 | 0.062 | 0.136 | 0.094 |
| Wnt | 0.058 | 0.056 | 0.09 | 0.078 |
| Leukemia | GPL96 | Jak-STAT | 0.338 | 0.208 | 0.508 | 0.284 |
| mTOR | 0.278 | 0.14 | 0.388 | 0.206 |
| PI3K-Akt | 0.384 | 0.276 | 0.566 | 0.37 |
| Wnt | 0.366 | 0.236 | 0.5 | 0.34 |
| Lung Cancer | GPL96 | Cell cycle | 0.358 | 0.524 | 0.74 | 0.638 |
| EGFR | 0.374 | 0.508 | 0.696 | 0.6 |
| Jak-STAT | 0.364 | 0.4 | 0.7 | 0.548 |
| MAPK | 0.352 | 0.446 | 0.708 | 0.57 |
| mTOR | 0.262 | 0.374 | 0.594 | 0.458 |
| P53 | 0.276 | 0.458 | 0.652 | 0.554 |
| Wnt | 0.262 | 0.448 | 0.68 | 0.528 |
| GPL570 | Cell cycle | 0.802 | 0.642 | 0.872 | 0.822 |
| EGFR | 0.678 | 0.482 | 0.788 | 0.66 |
| Jak-STAT | 0.656 | 0.512 | 0.782 | 0.666 |
| MAPK | 0.744 | 0.558 | 0.844 | 0.742 |
| mTOR | 0.688 | 0.542 | 0.806 | 0.696 |
| P53 | 0.812 | 0.63 | 0.89 | 0.754 |
| Wnt | 0.73 | 0.63 | 0.82 | 0.748 |

表十22(B) 各個癌症資料集常態性檢定結果–4個探針，對照組

(不針對每個基因進行單一探針挑選，並在500次檢定裡，每次檢定抽取4個基因探針資料來進行分析所得到的對照組(Normal or Adjacent Normal)的檢定拒絕率)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Diseases | Platform | Pathway | Mardia | HZ | Royston | energy |
| Breast Cancer | GPL96 | PI3K-Akt | 0.162 | 0.088 | 0.23 | 0.138 |
| Wnt | 0.166 | 0.094 | 0.288 | 0.152 |
| Colorectal Cancer | GPL570 | P53 | 0.066 | 0.064 | 0.096 | 0.09 |
| PI3K-Akt | 0.044 | 0.052 | 0.14 | 0.088 |
| Ras | 0.054 | 0.07 | 0.154 | 0.1 |
| TGF- | 0.038 | 0.032 | 0.082 | 0.054 |
| Wnt | 0.056 | 0.05 | 0.128 | 0.084 |
| Leukemia | GPL96 | Jak-STAT | 0.128 | 0.094 | 0.37 | 0.162 |
| mTOR | 0.086 | 0.116 | 0.344 | 0.168 |
| PI3K-Akt | 0.14 | 0.114 | 0.412 | 0.206 |
| Wnt | 0.116 | 0.078 | 0.292 | 0.156 |
| Lung Cancer | GPL96 | Cell cycle | 0.178 | 0.172 | 0.442 | 0.298 |
| EGFR | 0.22 | 0.194 | 0.394 | 0.32 |
| Jak-STAT | 0.148 | 0.178 | 0.44 | 0.244 |
| MAPK | 0.118 | 0.156 | 0.326 | 0.232 |
| mTOR | 0.218 | 0.194 | 0.43 | 0.312 |
| P53 | 0.19 | 0.196 | 0.486 | 0.326 |
| Wnt | 0.15 | 0.138 | 0.372 | 0.25 |
| GPL570 | Cell cycle | 0.804 | 0.746 | 0.87 | 0.848 |
| EGFR | 0.882 | 0.754 | 0.91 | 0.856 |
| Jak-STAT | 0.802 | 0.708 | 0.902 | 0.844 |
| MAPK | 0.82 | 0.732 | 0.896 | 0.85 |
| mTOR | 0.862 | 0.75 | 0.908 | 0.892 |
| P53 | 0.816 | 0.766 | 0.884 | 0.848 |
| Wnt | 0.866 | 0.768 | 0.896 | 0.882 |

表十一23(A) 各個癌症資料集常態性檢定結果–2個探針(log)，病例組

(不針對每個基因進行單一探針挑選，接著將基因表現量取，並在500次檢定裡，每次檢定抽取4個基因探針資料來進行分析所得到的病例組(Cancer)的檢定拒絕率)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Diseases | Platform | Pathway | Mardia | HZ | Royston | energy |
| Breast Cancer | GPL96 | PI3K-Akt | 0.08 | 0.046 | 0.128 | 0.068 |
| Wnt | 0.094 | 0.086 | 0.2 | 0.11 |
| Colorectal Cancer | GPL570 | P53 | 0.034 | 0.028 | 0.072 | 0.042 |
| PI3K-Akt | 0.068 | 0.04 | 0.104 | 0.062 |
| Ras | 0.068 | 0.046 | 0.136 | 0.092 |
| TGF- | 0.076 | 0.064 | 0.142 | 0.096 |
| Wnt | 0.058 | 0.054 | 0.092 | 0.078 |
| Leukemia | GPL96 | Jak-STAT | 0.3 | 0.14 | 0.462 | 0.242 |
| mTOR | 0.246 | 0.122 | 0.346 | 0.176 |
| PI3K-Akt | 0.35 | 0.222 | 0.516 | 0.324 |
| Wnt | 0.326 | 0.214 | 0.504 | 0.306 |
| Lung Cancer | GPL96 | Cell cycle | 0.272 | 0.486 | 0.66 | 0.604 |
| EGFR | 0.322 | 0.472 | 0.636 | 0.578 |
| Jak-STAT | 0.31 | 0.354 | 0.634 | 0.494 |
| MAPK | 0.296 | 0.394 | 0.65 | 0.512 |
| mTOR | 0.218 | 0.34 | 0.532 | 0.434 |
| P53 | 0.212 | 0.382 | 0.564 | 0.474 |
| Wnt | 0.2 | 0.434 | 0.638 | 0.514 |
| GPL570 | Cell cycle | 0.754 | 0.578 | 0.832 | 0.756 |
| EGFR | 0.668 | 0.486 | 0.79 | 0.654 |
| Jak-STAT | 0.62 | 0.492 | 0.762 | 0.638 |
| MAPK | 0.692 | 0.552 | 0.798 | 0.684 |
| mTOR | 0.688 | 0.54 | 0.778 | 0.694 |
| P53 | 0.722 | 0.586 | 0.838 | 0.696 |
| Wnt | 0.708 | 0.596 | 0.806 | 0.732 |

表十一24(B) 各個癌症資料集常態性檢定結果–2個探針(log)，對照組

(不針對每個基因進行單一探針挑選，接著將基因表現量取，並在500次檢定裡，每次檢定抽取4個基因探針資料來進行分析所得到的對照組(Normal or Adjacent Normal)的檢定拒絕率)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Diseases | Platform | Pathway | Mardia | HZ | Royston | energy |
| Breast Cancer | GPL96 | PI3K-Akt | 0.162 | 0.092 | 0.228 | 0.136 |
| Wnt | 0.164 | 0.098 | 0.292 | 0.154 |
| Colorectal Cancer | GPL570 | P53 | 0.06 | 0.06 | 0.102 | 0.092 |
| PI3K-Akt | 0.044 | 0.054 | 0.144 | 0.084 |
| Ras | 0.052 | 0.07 | 0.142 | 0.098 |
| TGF- | 0.036 | 0.03 | 0.074 | 0.056 |
| Wnt | 0.062 | 0.052 | 0.136 | 0.082 |
| Leukemia | GPL96 | Jak-STAT | 0.122 | 0.086 | 0.354 | 0.158 |
| mTOR | 0.088 | 0.106 | 0.32 | 0.17 |
| PI3K-Akt | 0.136 | 0.104 | 0.382 | 0.2 |
| Wnt | 0.116 | 0.084 | 0.288 | 0.16 |
| Lung Cancer | GPL96 | Cell cycle | 0.17 | 0.196 | 0.464 | 0.306 |
| EGFR | 0.224 | 0.186 | 0.362 | 0.322 |
| Jak-STAT | 0.144 | 0.16 | 0.41 | 0.228 |
| MAPK | 0.098 | 0.142 | 0.296 | 0.214 |
| mTOR | 0.198 | 0.188 | 0.404 | 0.306 |
| P53 | 0.16 | 0.174 | 0.436 | 0.274 |
| Wnt | 0.134 | 0.144 | 0.332 | 0.25 |
| GPL570 | Cell cycle | 0.818 | 0.754 | 0.864 | 0.854 |
| EGFR | 0.858 | 0.77 | 0.906 | 0.866 |
| Jak-STAT | 0.764 | 0.696 | 0.862 | 0.802 |
| MAPK | 0.804 | 0.736 | 0.876 | 0.836 |
| mTOR | 0.862 | 0.766 | 0.904 | 0.882 |
| P53 | 0.802 | 0.752 | 0.872 | 0.846 |
| Wnt | 0.856 | 0.774 | 0.888 | 0.872 |

表十二25 GSA方法在差異表達為共變異數矩陣發生變化時的檢定力表現

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Hotelling’s T2  (S.E.) | Hotelling’s T2  (N.S.E.) | Global test | GlobalAncova | Energy |
| 0.1 | 1 | 0.055 | 0.055 | 0.054 | 0.049 | 0.052 |
| 2 | 0.091 | 0.062 | 0.055 | 0.05 | 0.947 |
| 3 | 0.147 | 0.068 | 0.056 | 0.053 | 1 |
| 4 | 0.199 | 0.074 | 0.057 | 0.054 | 1 |
| 5 | 0.254 | 0.084 | 0.06 | 0.053 | 1 |
| 0.5 | 1 | 0.055 | 0.058 | 0.053 | 0.047 | 0.047 |
| 2 | 0.091 | 0.103 | 0.053 | 0.049 | 0.298 |
| 3 | 0.147 | 0.169 | 0.054 | 0.047 | 0.939 |
| 4 | 0.199 | 0.259 | 0.061 | 0.045 | 1 |
| 5 | 0.254 | 0.315 | 0.058 | 0.045 | 1 |
| 0.9 | 1 | 0.055 | 0.062 | 0.053 | 0.046 | 0.052 |
| 2 | 0.091 | 0.111 | 0.054 | 0.043 | 0.149 |
| 3 | 0.147 | 0.206 | 0.056 | 0.044 | 0.427 |
| 4 | 0.199 | 0.311 | 0.058 | 0.045 | 0.754 |
| 5 | 0.254 | 0.374 | 0.061 | 0.052 | 0.92 |

**S.E.: with shrinkage estimator**

**N.S.E.: without shrinkage estimator**

表十三26GSA方法在多維度常態分佈情境下所使用的相關係數矩陣1 (非結構化)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | G10 | G11 | G12 | G13 | G14 | G15 | G16 | G17 | G18 | G19 | G20 | G21 | G22 | G23 | G24 | G25 | G26 | G27 | G28 | G29 | G30 |
| G1 | **1** | **0.39** | **0.07** | **0.03** | **-0.24** | **-0.01** | **-0.04** | **-0.3** | **-0.16** | **-0.1** | **0.12** | **-0.26** | **0.15** | **0.23** | **-0.17** | **-0.09** | **0.25** | **-0.19** | **0.12** | **-0.27** | **-0.23** | **0.2** | **-0.04** | **0.03** | **0.02** | **0.15** | **-0.17** | **-0.04** | **0** | **0.2** |
| G2 | **0.39** | **1** | **-0.11** | **-0.42** | **0.07** | **-0.09** | **0** | **-0.28** | **-0.17** | **0.11** | **-0.3** | **-0.4** | **0.09** | **0.11** | **-0.35** | **0.03** | **0.17** | **-0.19** | **-0.03** | **0.28** | **-0.25** | **0.33** | **-0.17** | **-0.29** | **-0.03** | **-0.1** | **-0.06** | **-0.35** | **0.26** | **-0.17** |
| G3 | **0.07** | **-0.11** | **1** | **0.28** | **-0.47** | **0.07** | **0.01** | **0.39** | **-0.33** | **-0.35** | **0.14** | **0.36** | **-0.13** | **-0.58** | **0.36** | **-0.29** | **0.13** | **-0.14** | **0.1** | **-0.2** | **0.19** | **-0.16** | **0.23** | **0.42** | **0.2** | **-0.14** | **0.1** | **-0.11** | **-0.07** | **0.35** |
| G4 | **0.03** | **-0.42** | **0.28** | **1** | **0** | **0.22** | **0.48** | **0.27** | **-0.02** | **-0.37** | **0.25** | **0.37** | **-0.11** | **-0.16** | **0.56** | **-0.16** | **0.18** | **0.22** | **-0.16** | **-0.31** | **0.34** | **-0.31** | **-0.06** | **0.58** | **0.42** | **0.05** | **-0.05** | **0.1** | **-0.29** | **0.4** |
| G5 | **-0.24** | **0.07** | **-0.47** | **0** | **1** | **0.01** | **0.33** | **-0.01** | **0.14** | **0.08** | **-0.17** | **0.08** | **0.17** | **0.14** | **0.22** | **0.06** | **-0.05** | **0.19** | **-0.19** | **0.15** | **0.06** | **-0.43** | **-0.12** | **-0.15** | **-0.06** | **0.13** | **-0.36** | **0.17** | **0.15** | **-0.28** |
| G6 | **-0.01** | **-0.09** | **0.07** | **0.22** | **0.01** | **1** | **0.09** | **-0.03** | **-0.2** | **-0.04** | **-0.18** | **0.08** | **0.1** | **-0.21** | **0.26** | **0.16** | **0.07** | **-0.13** | **-0.1** | **0.2** | **-0.13** | **-0.25** | **-0.08** | **0.44** | **0.21** | **-0.1** | **-0.1** | **0.26** | **0.3** | **-0.06** |
| G7 | **-0.04** | **0** | **0.01** | **0.48** | **0.33** | **0.09** | **1** | **0.2** | **0.13** | **-0.29** | **-0.19** | **0.3** | **-0.07** | **-0.2** | **0.57** | **-0.15** | **-0.04** | **0.3** | **-0.24** | **-0.11** | **0.21** | **-0.45** | **-0.13** | **0.34** | **-0.12** | **-0.13** | **-0.06** | **-0.02** | **-0.01** | **0.14** |
| G8 | **-0.3** | **-0.28** | **0.39** | **0.27** | **-0.01** | **-0.03** | **0.2** | **1** | **-0.18** | **0.01** | **-0.06** | **0.62** | **-0.45** | **-0.58** | **0.44** | **-0.28** | **0.01** | **0.08** | **-0.2** | **-0.03** | **0.42** | **-0.26** | **-0.27** | **0.21** | **-0.07** | **-0.24** | **0.07** | **-0.23** | **-0.07** | **0** |
| G9 | **-0.16** | **-0.17** | **-0.33** | **-0.02** | **0.14** | **-0.2** | **0.13** | **-0.18** | **1** | **-0.09** | **0.15** | **-0.26** | **-0.16** | **0.26** | **-0.15** | **-0.08** | **-0.08** | **0.29** | **0.07** | **-0.08** | **-0.08** | **-0.04** | **-0.06** | **-0.14** | **-0.09** | **-0.06** | **0.2** | **0.06** | **-0.28** | **0.2** |
| G10 | **-0.1** | **0.11** | **-0.35** | **-0.37** | **0.08** | **-0.04** | **-0.29** | **0.01** | **-0.09** | **1** | **-0.36** | **-0.25** | **-0.03** | **0.08** | **-0.35** | **-0.01** | **0.22** | **0.17** | **0.18** | **0.51** | **-0.04** | **0.22** | **-0.17** | **-0.37** | **-0.08** | **-0.3** | **0.13** | **-0.02** | **0.09** | **-0.11** |
| G11 | **0.12** | **-0.3** | **0.14** | **0.25** | **-0.17** | **-0.18** | **-0.19** | **-0.06** | **0.15** | **-0.36** | **1** | **0.03** | **0.23** | **0.21** | **-0.06** | **0.02** | **-0.16** | **-0.13** | **0.25** | **-0.52** | **-0.04** | **-0.01** | **0.02** | **0.18** | **0.44** | **0.39** | **0.14** | **0.23** | **-0.22** | **0.36** |
| G12 | **-0.26** | **-0.4** | **0.36** | **0.37** | **0.08** | **0.08** | **0.3** | **0.62** | **-0.26** | **-0.25** | **0.03** | **1** | **-0.1** | **-0.32** | **0.53** | **-0.45** | **-0.03** | **0.19** | **-0.17** | **-0.25** | **0.2** | **-0.38** | **-0.06** | **0.41** | **0.02** | **0.1** | **-0.31** | **0.02** | **-0.04** | **0** |
| G13 | **0.15** | **0.09** | **-0.13** | **-0.11** | **0.17** | **0.1** | **-0.07** | **-0.45** | **-0.16** | **-0.03** | **0.23** | **-0.1** | **1** | **0.15** | **0.03** | **-0.03** | **0.13** | **-0.16** | **0.14** | **-0.28** | **-0.24** | **-0.03** | **-0.05** | **-0.1** | **0.29** | **0.32** | **-0.11** | **0.05** | **0.19** | **-0.04** |
| G14 | **0.23** | **0.11** | **-0.58** | **-0.16** | **0.14** | **-0.21** | **-0.2** | **-0.58** | **0.26** | **0.08** | **0.21** | **-0.32** | **0.15** | **1** | **-0.38** | **0.11** | **-0.07** | **0.13** | **0.17** | **-0.16** | **-0.39** | **0.39** | **-0.14** | **-0.26** | **0.03** | **0.41** | **-0.2** | **0.19** | **-0.12** | **-0.13** |
| G15 | **-0.17** | **-0.35** | **0.36** | **0.56** | **0.22** | **0.26** | **0.57** | **0.44** | **-0.15** | **-0.35** | **-0.06** | **0.53** | **0.03** | **-0.38** | **1** | **-0.33** | **0.08** | **0.06** | **-0.12** | **-0.27** | **0.39** | **-0.43** | **-0.15** | **0.37** | **0.02** | **-0.16** | **-0.19** | **0.09** | **0.17** | **0.25** |
| G16 | **-0.09** | **0.03** | **-0.29** | **-0.16** | **0.06** | **0.16** | **-0.15** | **-0.28** | **-0.08** | **-0.01** | **0.02** | **-0.45** | **-0.03** | **0.11** | **-0.33** | **1** | **-0.27** | **-0.26** | **0.2** | **0.34** | **-0.17** | **0.05** | **0.22** | **-0.09** | **0.04** | **0.14** | **0.18** | **0.26** | **0.05** | **-0.21** |
| G17 | **0.25** | **0.17** | **0.13** | **0.18** | **-0.05** | **0.07** | **-0.04** | **0.01** | **-0.08** | **0.22** | **-0.16** | **-0.03** | **0.13** | **-0.07** | **0.08** | **-0.27** | **1** | **0.02** | **-0.01** | **0.26** | **-0.06** | **0.09** | **-0.07** | **0.01** | **0.4** | **-0.19** | **-0.19** | **-0.23** | **0.16** | **0.26** |
| G18 | **-0.19** | **-0.19** | **-0.14** | **0.22** | **0.19** | **-0.13** | **0.3** | **0.08** | **0.29** | **0.17** | **-0.13** | **0.19** | **-0.16** | **0.13** | **0.06** | **-0.26** | **0.02** | **1** | **-0.18** | **0.13** | **-0.07** | **-0.35** | **0.04** | **0.13** | **-0.11** | **-0.17** | **-0.04** | **0.06** | **0.07** | **-0.08** |
| G19 | **0.12** | **-0.03** | **0.1** | **-0.16** | **-0.19** | **-0.1** | **-0.24** | **-0.2** | **0.07** | **0.18** | **0.25** | **-0.17** | **0.14** | **0.17** | **-0.12** | **0.2** | **-0.01** | **-0.18** | **1** | **-0.19** | **-0.05** | **0.3** | **0.14** | **-0.06** | **0.07** | **0** | **0.46** | **0.04** | **-0.05** | **0.44** |
| G20 | **-0.27** | **0.28** | **-0.2** | **-0.31** | **0.15** | **0.2** | **-0.11** | **-0.03** | **-0.08** | **0.51** | **-0.52** | **-0.25** | **-0.28** | **-0.16** | **-0.27** | **0.34** | **0.26** | **0.13** | **-0.19** | **1** | **-0.17** | **0.03** | **0.16** | **-0.13** | **-0.05** | **-0.31** | **-0.01** | **0.16** | **0.29** | **-0.39** |
| G21 | **-0.23** | **-0.25** | **0.19** | **0.34** | **0.06** | **-0.13** | **0.21** | **0.42** | **-0.08** | **-0.04** | **-0.04** | **0.2** | **-0.24** | **-0.39** | **0.39** | **-0.17** | **-0.06** | **-0.07** | **-0.05** | **-0.17** | **1** | **-0.01** | **-0.25** | **-0.04** | **-0.12** | **-0.39** | **0.29** | **-0.23** | **-0.28** | **0.35** |
| G22 | **0.2** | **0.33** | **-0.16** | **-0.31** | **-0.43** | **-0.25** | **-0.45** | **-0.26** | **-0.04** | **0.22** | **-0.01** | **-0.38** | **-0.03** | **0.39** | **-0.43** | **0.05** | **0.09** | **-0.35** | **0.3** | **0.03** | **-0.01** | **1** | **-0.17** | **-0.35** | **-0.06** | **0.05** | **0.28** | **-0.28** | **-0.18** | **0.03** |
| G23 | **-0.04** | **-0.17** | **0.23** | **-0.06** | **-0.12** | **-0.08** | **-0.13** | **-0.27** | **-0.06** | **-0.17** | **0.02** | **-0.06** | **-0.05** | **-0.14** | **-0.15** | **0.22** | **-0.07** | **0.04** | **0.14** | **0.16** | **-0.25** | **-0.17** | **1** | **0.11** | **-0.13** | **0.3** | **0.12** | **0.29** | **-0.11** | **-0.02** |
| G24 | **0.03** | **-0.29** | **0.42** | **0.58** | **-0.15** | **0.44** | **0.34** | **0.21** | **-0.14** | **-0.37** | **0.18** | **0.41** | **-0.1** | **-0.26** | **0.37** | **-0.09** | **0.01** | **0.13** | **-0.06** | **-0.13** | **-0.04** | **-0.35** | **0.11** | **1** | **0.41** | **-0.04** | **-0.11** | **0.32** | **-0.12** | **0.02** |
| G25 | **0.02** | **-0.03** | **0.2** | **0.42** | **-0.06** | **0.21** | **-0.12** | **-0.07** | **-0.09** | **-0.08** | **0.44** | **0.02** | **0.29** | **0.03** | **0.02** | **0.04** | **0.4** | **-0.11** | **0.07** | **-0.05** | **-0.12** | **-0.06** | **-0.13** | **0.41** | **1** | **0.1** | **-0.05** | **0.01** | **-0.14** | **0.24** |
| G26 | **0.15** | **-0.1** | **-0.14** | **0.05** | **0.13** | **-0.1** | **-0.13** | **-0.24** | **-0.06** | **-0.3** | **0.39** | **0.1** | **0.32** | **0.41** | **-0.16** | **0.14** | **-0.19** | **-0.17** | **0** | **-0.31** | **-0.39** | **0.05** | **0.3** | **-0.04** | **0.1** | **1** | **-0.14** | **0.26** | **-0.35** | **-0.14** |
| G27 | **-0.17** | **-0.06** | **0.1** | **-0.05** | **-0.36** | **-0.1** | **-0.06** | **0.07** | **0.2** | **0.13** | **0.14** | **-0.31** | **-0.11** | **-0.2** | **-0.19** | **0.18** | **-0.19** | **-0.04** | **0.46** | **-0.01** | **0.29** | **0.28** | **0.12** | **-0.11** | **-0.05** | **-0.14** | **1** | **-0.27** | **-0.28** | **0.31** |
| G28 | **-0.04** | **-0.35** | **-0.11** | **0.1** | **0.17** | **0.26** | **-0.02** | **-0.23** | **0.06** | **-0.02** | **0.23** | **0.02** | **0.05** | **0.19** | **0.09** | **0.26** | **-0.23** | **0.06** | **0.04** | **0.16** | **-0.23** | **-0.28** | **0.29** | **0.32** | **0.01** | **0.26** | **-0.27** | **1** | **-0.07** | **-0.19** |
| G29 | **0** | **0.26** | **-0.07** | **-0.29** | **0.15** | **0.3** | **-0.01** | **-0.07** | **-0.28** | **0.09** | **-0.22** | **-0.04** | **0.19** | **-0.12** | **0.17** | **0.05** | **0.16** | **0.07** | **-0.05** | **0.29** | **-0.28** | **-0.18** | **-0.11** | **-0.12** | **-0.14** | **-0.35** | **-0.28** | **-0.07** | **1** | **-0.18** |
| G30 | **0.2** | **-0.17** | **0.35** | **0.4** | **-0.28** | **-0.06** | **0.14** | **0** | **0.2** | **-0.11** | **0.36** | **0** | **-0.04** | **-0.13** | **0.25** | **-0.21** | **0.26** | **-0.08** | **0.44** | **-0.39** | **0.35** | **0.03** | **-0.02** | **0.02** | **0.24** | **-0.14** | **0.31** | **-0.19** | **-0.18** | **1** |

表十四27GSA方法在多維度常態分佈情境下所使用的相關係數矩陣2 (非結構化)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | G1 | G2 | G3 | G4 | G5 | G6 | G7 | G8 | G9 | G10 | G11 | G12 | G13 | G14 | G15 | G16 | G17 | G18 | G19 | G20 | G21 | G22 | G23 | G24 | G25 | G26 | G27 | G28 | G29 | G30 |
| G1 | **1** | **0.11** | **0.03** | **0.23** | **0.32** | **-0.06** | **0.18** | **0.17** | **-0.24** | **0.15** | **0.26** | **0.25** | **0.03** | **-0.19** | **-0.18** | **-0.14** | **0.12** | **0.27** | **-0.5** | **-0.45** | **-0.01** | **0.29** | **-0.32** | **-0.24** | **-0.17** | **-0.24** | **0.2** | **-0.12** | **0.12** | **0.01** |
| G2 | **0.11** | **1** | **0.09** | **0.1** | **-0.23** | **-0.29** | **0.06** | **0.26** | **-0.12** | **-0.03** | **-0.03** | **0.05** | **-0.17** | **0.22** | **-0.24** | **-0.33** | **0.2** | **0.32** | **-0.19** | **-0.34** | **0.41** | **-0.11** | **0.31** | **0.56** | **0.12** | **0.15** | **0.1** | **0.12** | **-0.34** | **0.56** |
| G3 | **0.03** | **0.09** | **1** | **-0.23** | **-0.26** | **0** | **-0.12** | **-0.24** | **0.17** | **0.12** | **-0.02** | **0.06** | **0.27** | **0.21** | **-0.37** | **-0.03** | **0.08** | **-0.3** | **0.14** | **-0.09** | **0.08** | **-0.26** | **-0.06** | **0.2** | **0.36** | **-0.09** | **0.32** | **-0.15** | **0.28** | **0.12** |
| G4 | **0.23** | **0.1** | **-0.23** | **1** | **0.35** | **0.12** | **-0.04** | **0.39** | **-0.28** | **-0.15** | **0.05** | **0.07** | **0.07** | **-0.18** | **-0.04** | **-0.15** | **0.17** | **0.28** | **-0.43** | **-0.12** | **0.14** | **0.15** | **-0.27** | **0.07** | **-0.35** | **0.19** | **-0.4** | **-0.2** | **-0.07** | **0.13** |
| G5 | **0.32** | **-0.23** | **-0.26** | **0.35** | **1** | **0.18** | **0.21** | **0.04** | **-0.05** | **-0.15** | **0.18** | **0.26** | **-0.01** | **-0.06** | **-0.06** | **-0.08** | **-0.01** | **0.03** | **-0.11** | **-0.04** | **-0.39** | **0.12** | **-0.15** | **-0.47** | **-0.2** | **-0.11** | **-0.17** | **-0.21** | **-0.02** | **-0.22** |
| G6 | **-0.06** | **-0.29** | **0** | **0.12** | **0.18** | **1** | **0.03** | **-0.05** | **0.23** | **-0.05** | **-0.12** | **-0.01** | **-0.21** | **-0.13** | **0.12** | **0.2** | **0.03** | **-0.24** | **0.06** | **0.49** | **-0.07** | **-0.09** | **-0.07** | **-0.45** | **-0.14** | **0.24** | **-0.39** | **0.02** | **0.27** | **-0.29** |
| G7 | **0.18** | **0.06** | **-0.12** | **-0.04** | **0.21** | **0.03** | **1** | **0.09** | **-0.13** | **0.19** | **0.43** | **0.08** | **-0.01** | **-0.09** | **-0.06** | **0.43** | **-0.12** | **-0.02** | **-0.18** | **-0.04** | **-0.25** | **0.02** | **0.02** | **-0.1** | **0.21** | **0** | **0.16** | **0.13** | **0.11** | **-0.02** |
| G8 | **0.17** | **0.26** | **-0.24** | **0.39** | **0.04** | **-0.05** | **0.09** | **1** | **-0.56** | **-0.14** | **0.18** | **-0.13** | **-0.2** | **-0.03** | **0.03** | **-0.17** | **0.18** | **0.53** | **-0.5** | **-0.08** | **0.34** | **0.12** | **0.17** | **0.23** | **-0.25** | **0.32** | **-0.22** | **-0.11** | **-0.11** | **0.2** |
| G9 | **-0.24** | **-0.12** | **0.17** | **-0.28** | **-0.05** | **0.23** | **-0.13** | **-0.56** | **1** | **0.13** | **-0.13** | **-0.31** | **-0.21** | **0.21** | **-0.24** | **0.2** | **0.08** | **-0.27** | **0.67** | **0.31** | **-0.24** | **-0.07** | **-0.07** | **0.04** | **0.2** | **-0.43** | **0.13** | **0.02** | **0.34** | **-0.11** |
| G10 | **0.15** | **-0.03** | **0.12** | **-0.15** | **-0.15** | **-0.05** | **0.19** | **-0.14** | **0.13** | **1** | **0.35** | **-0.05** | **-0.04** | **-0.01** | **-0.08** | **0.41** | **0.02** | **-0.29** | **0.28** | **0.36** | **-0.32** | **0.28** | **0.01** | **-0.19** | **0.06** | **-0.19** | **0.24** | **0.35** | **0.3** | **0** |
| G11 | **0.26** | **-0.03** | **-0.02** | **0.05** | **0.18** | **-0.12** | **0.43** | **0.18** | **-0.13** | **0.35** | **1** | **-0.07** | **0.06** | **0.27** | **-0.11** | **0.42** | **0.23** | **0.04** | **-0.13** | **-0.08** | **-0.29** | **0.41** | **-0.03** | **-0.05** | **0.11** | **-0.33** | **0.34** | **0.24** | **0.3** | **0.14** |
| G12 | **0.25** | **0.05** | **0.06** | **0.07** | **0.26** | **-0.01** | **0.08** | **-0.13** | **-0.31** | **-0.05** | **-0.07** | **1** | **0.45** | **0.04** | **-0.11** | **-0.09** | **-0.36** | **-0.27** | **-0.19** | **-0.14** | **-0.01** | **0.12** | **-0.05** | **-0.22** | **0.03** | **0.29** | **-0.04** | **-0.02** | **-0.12** | **0.13** |
| G13 | **0.03** | **-0.17** | **0.27** | **0.07** | **-0.01** | **-0.21** | **-0.01** | **-0.2** | **-0.21** | **-0.04** | **0.06** | **0.45** | **1** | **0.02** | **0.06** | **-0.12** | **-0.13** | **-0.33** | **-0.13** | **-0.13** | **0.02** | **-0.05** | **-0.31** | **0.03** | **0.13** | **0** | **-0.07** | **-0.02** | **-0.11** | **0.1** |
| G14 | **-0.19** | **0.22** | **0.21** | **-0.18** | **-0.06** | **-0.13** | **-0.09** | **-0.03** | **0.21** | **-0.01** | **0.27** | **0.04** | **0.02** | **1** | **-0.39** | **0.02** | **0.12** | **0.11** | **0.09** | **-0.01** | **-0.13** | **-0.23** | **0.19** | **0.32** | **0.3** | **-0.27** | **0.33** | **0.05** | **0.05** | **0.5** |
| G15 | **-0.18** | **-0.24** | **-0.37** | **-0.04** | **-0.06** | **0.12** | **-0.06** | **0.03** | **-0.24** | **-0.08** | **-0.11** | **-0.11** | **0.06** | **-0.39** | **1** | **0.02** | **-0.22** | **-0.11** | **0.09** | **0.14** | **-0.08** | **-0.06** | **0.33** | **-0.29** | **-0.09** | **0.14** | **-0.28** | **0.1** | **-0.13** | **-0.48** |
| G16 | **-0.14** | **-0.33** | **-0.03** | **-0.15** | **-0.08** | **0.2** | **0.43** | **-0.17** | **0.2** | **0.41** | **0.42** | **-0.09** | **-0.12** | **0.02** | **0.02** | **1** | **0.02** | **-0.27** | **0.29** | **0.31** | **-0.53** | **0.05** | **0.03** | **-0.25** | **0.03** | **-0.28** | **0.26** | **0.39** | **0.41** | **-0.29** |
| G17 | **0.12** | **0.2** | **0.08** | **0.17** | **-0.01** | **0.03** | **-0.12** | **0.18** | **0.08** | **0.02** | **0.23** | **-0.36** | **-0.13** | **0.12** | **-0.22** | **0.02** | **1** | **0.18** | **-0.16** | **-0.04** | **0.06** | **0.03** | **-0.24** | **0.32** | **-0.08** | **-0.18** | **0.08** | **-0.11** | **0.16** | **0.24** |
| G18 | **0.27** | **0.32** | **-0.3** | **0.28** | **0.03** | **-0.24** | **-0.02** | **0.53** | **-0.27** | **-0.29** | **0.04** | **-0.27** | **-0.33** | **0.11** | **-0.11** | **-0.27** | **0.18** | **1** | **-0.43** | **-0.4** | **0.29** | **0.12** | **-0.01** | **0.33** | **-0.31** | **-0.17** | **0.03** | **-0.03** | **-0.32** | **0.1** |
| G19 | **-0.5** | **-0.19** | **0.14** | **-0.43** | **-0.11** | **0.06** | **-0.18** | **-0.5** | **0.67** | **0.28** | **-0.13** | **-0.19** | **-0.13** | **0.09** | **0.09** | **0.29** | **-0.16** | **-0.43** | **1** | **0.54** | **-0.42** | **-0.06** | **0.22** | **-0.06** | **0.05** | **-0.28** | **0.15** | **0.24** | **0.31** | **-0.37** |
| G20 | **-0.45** | **-0.34** | **-0.09** | **-0.12** | **-0.04** | **0.49** | **-0.04** | **-0.08** | **0.31** | **0.36** | **-0.08** | **-0.14** | **-0.13** | **-0.01** | **0.14** | **0.31** | **-0.04** | **-0.4** | **0.54** | **1** | **-0.17** | **-0.07** | **0.02** | **-0.2** | **-0.06** | **0.17** | **-0.31** | **0.22** | **0.21** | **-0.31** |
| G21 | **-0.01** | **0.41** | **0.08** | **0.14** | **-0.39** | **-0.07** | **-0.25** | **0.34** | **-0.24** | **-0.32** | **-0.29** | **-0.01** | **0.02** | **-0.13** | **-0.08** | **-0.53** | **0.06** | **0.29** | **-0.42** | **-0.17** | **1** | **-0.14** | **-0.12** | **0.53** | **-0.2** | **0.4** | **-0.3** | **0** | **-0.34** | **0.38** |
| G22 | **0.29** | **-0.11** | **-0.26** | **0.15** | **0.12** | **-0.09** | **0.02** | **0.12** | **-0.07** | **0.28** | **0.41** | **0.12** | **-0.05** | **-0.23** | **-0.06** | **0.05** | **0.03** | **0.12** | **-0.06** | **-0.07** | **-0.14** | **1** | **-0.32** | **-0.17** | **-0.06** | **-0.14** | **0.16** | **0.1** | **0.19** | **0.03** |
| G23 | **-0.32** | **0.31** | **-0.06** | **-0.27** | **-0.15** | **-0.07** | **0.02** | **0.17** | **-0.07** | **0.01** | **-0.03** | **-0.05** | **-0.31** | **0.19** | **0.33** | **0.03** | **-0.24** | **-0.01** | **0.22** | **0.02** | **-0.12** | **-0.32** | **1** | **-0.02** | **0.2** | **0.25** | **0** | **-0.08** | **-0.18** | **0.02** |
| G24 | **-0.24** | **0.56** | **0.2** | **0.07** | **-0.47** | **-0.45** | **-0.1** | **0.23** | **0.04** | **-0.19** | **-0.05** | **-0.22** | **0.03** | **0.32** | **-0.29** | **-0.25** | **0.32** | **0.33** | **-0.06** | **-0.2** | **0.53** | **-0.17** | **-0.02** | **1** | **0.09** | **0.04** | **0.13** | **0.07** | **-0.12** | **0.5** |
| G25 | **-0.17** | **0.12** | **0.36** | **-0.35** | **-0.2** | **-0.14** | **0.21** | **-0.25** | **0.2** | **0.06** | **0.11** | **0.03** | **0.13** | **0.3** | **-0.09** | **0.03** | **-0.08** | **-0.31** | **0.05** | **-0.06** | **-0.2** | **-0.06** | **0.2** | **0.09** | **1** | **0.03** | **0.36** | **-0.32** | **0** | **0.23** |
| G26 | **-0.24** | **0.15** | **-0.09** | **0.19** | **-0.11** | **0.24** | **0** | **0.32** | **-0.43** | **-0.19** | **-0.33** | **0.29** | **0** | **-0.27** | **0.14** | **-0.28** | **-0.18** | **-0.17** | **-0.28** | **0.17** | **0.4** | **-0.14** | **0.25** | **0.04** | **0.03** | **1** | **-0.42** | **-0.26** | **-0.24** | **0.16** |
| G27 | **0.2** | **0.1** | **0.32** | **-0.4** | **-0.17** | **-0.39** | **0.16** | **-0.22** | **0.13** | **0.24** | **0.34** | **-0.04** | **-0.07** | **0.33** | **-0.28** | **0.26** | **0.08** | **0.03** | **0.15** | **-0.31** | **-0.3** | **0.16** | **0** | **0.13** | **0.36** | **-0.42** | **1** | **0.04** | **0.39** | **0.13** |
| G28 | **-0.12** | **0.12** | **-0.15** | **-0.2** | **-0.21** | **0.02** | **0.13** | **-0.11** | **0.02** | **0.35** | **0.24** | **-0.02** | **-0.02** | **0.05** | **0.1** | **0.39** | **-0.11** | **-0.03** | **0.24** | **0.22** | **0** | **0.1** | **-0.08** | **0.07** | **-0.32** | **-0.26** | **0.04** | **1** | **0.02** | **-0.13** |
| G29 | **0.12** | **-0.34** | **0.28** | **-0.07** | **-0.02** | **0.27** | **0.11** | **-0.11** | **0.34** | **0.3** | **0.3** | **-0.12** | **-0.11** | **0.05** | **-0.13** | **0.41** | **0.16** | **-0.32** | **0.31** | **0.21** | **-0.34** | **0.19** | **-0.18** | **-0.12** | **0** | **-0.24** | **0.39** | **0.02** | **1** | **-0.17** |
| G30 | **0.01** | **0.56** | **0.12** | **0.13** | **-0.22** | **-0.29** | **-0.02** | **0.2** | **-0.11** | **0** | **0.14** | **0.13** | **0.1** | **0.5** | **-0.48** | **-0.29** | **0.24** | **0.1** | **-0.37** | **-0.31** | **0.38** | **0.03** | **0.02** | **0.5** | **0.23** | **0.16** | **0.13** | **-0.13** | **-0.17** | **1** |

表十五28GSA方法在多維度常態分佈情境底下的表現(非結構化相關係數矩陣1)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Hotelling’s T2 | Global test | GlobalANCOVA | energy | GSEA(Category) |
| 0 | 0.057 | 0.05 | 0.047 | 0.055 | 0.045 |
| 0.1 | 0.212 | 0.104 | 0.1 | 0.114 | 0.643 |
| 0.2 | 0.948 | 0.393 | 0.392 | 0.431 | 0.995 |
| 0.25 | 1 | 0.685 | 0.682 | 0.735 | 1 |
| 0.3 | 1 | 0.918 | 0.926 | 0.958 | 1 |
| 0.5 | 1 | 1 | 1 | 1 | 1 |
| 0.7 | 1 | 1 | 1 | 1 | 1 |
| 0.9 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 |

表十六29GSA方法在多維度常態分佈情境底下的表現(非結構化相關係數矩陣2)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Hotelling’s T2 | Global test | GlobalANCOVA | energy | GSEA(Category) |
| 0 | 0.062 | 0.058 | 0.052 | 0.055 | 0.05 |
| 0.1 | 0.214 | 0.099 | 0.09 | 0.107 | 0.58 |
| 0.2 | 0.893 | 0.405 | 0.399 | 0.444 | 0.988 |
| 0.25 | 0.997 | 0.697 | 0.693 | 0.736 | 0.999 |
| 0.3 | 1 | 0.927 | 0.922 | 0.945 | 1 |
| 0.5 | 1 | 1 | 1 | 1 | 1 |
| 0.7 | 1 | 1 | 1 | 1 | 1 |
| 0.9 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 |

圖一 乳癌資料集常態性檢定分析結果(點圖)

圖二 大腸直腸癌資料集常態性檢定分析結果(點圖)

圖三 白血病資料集常態性檢定分析結果(點圖)

圖四 肺癌資料集(GPL96)常態性檢定分析結果(點圖)

圖五 肺癌資料集(GPL570)常態性檢定分析結果(點圖)

圖六 卵巢癌資料集常態性檢定分析結果(點圖)

圖七 GSA方法在六種不同情境底下對型一誤差率(0.05)的估計表現(雷達圖)

圖八 GSA方法在六種不同情境底下的統計檢定力表現(雷達圖)–

圖九 GSA方法在六種不同情境底下的統計檢定力表現(雷達圖)–

圖十 GSA方法在六種不同情境底下的統計檢定力表現(雷達圖)–

圖十一 GSA方法在情境四底下的統計檢定力表現(雷達圖)–針對設計六種不同的比例組合

圖十二 GSA方法在情境六底下的統計檢定力表現(雷達圖)–針對設計六種不同的比例組合

圖十三 GSA方法在情境三底下的統計檢定力表現(雷達圖)–針對設計六種不同的比例組合

圖十四 GSA方法在情境三、四、六底下的統計檢定力表現(雷達圖)–針對為組合六的情況

此圖形是將圖十一～圖十三中右上軸（組合六）的結果一起整併起來

# 附錄一、常態性檢定

在本研究中，為了要去確認真實基因資料的常態性假設，我們使用R語言套件包，*MVN*以及*energy*，裡面所提供的多維度常態性檢定來進行分析，在這些檢定中，虛無假設都是資料來自常態分佈，而我們所使用到的幾個檢定以及其詳盡的檢定統計量構造如下：

1. Mardia’s test (Korkmaz, Goksuluk, and Zararsiz 2014)

在這個檢定中，Mardia將多維度版本的偏度(skewness, )和峰度(kurtosis, )定義成以下式子：

以及 ，

其中，為每次多維度常態性假設檢驗的基因數目，，為一個的向量，代表所有的個人裡面，第個人個基因的基因表現量，則代表所有人基因表現量的樣本共變異數矩陣。接著，此檢定便將這兩個統計量和多維度常態分佈的版本分別進行比較，其使用的檢定統計量分別如下：

檢測偏度的檢定統計量：

檢測峰度的檢定統計量：

而在本研究中，只要上述的這兩個檢定統計量裡有一個顯著，那麼我們將會拒絕虛無假設，意即這筆資料並非服從多維度常態分佈。

1. Henze-Zirkler’s test (Henze, and Zirkler 1990; Korkmaz, Goksuluk, and Zararsiz 2014)

Henze-Zirkler’s檢定利用一個非負的距離函數()來測量資料分配()和多維度常態分配()彼此之間的差距，若和的分配相同，則該距離函數的數值就會是，若不相同，則數值將會越大。而此距離函式詳細的構造如下：

其中，和分別為和的傅立葉轉換式(Fourier transforms)，也就是和的特徵函數(characteristic function) （對而言是經驗特徵函數(empirical characteristic function)），可以是一個權重函數(weight function)或是核函數(kernel function)，在此檢定中使用的密度函數(density finction)來當成，也因此算是計算兩個分配的特徵函數的期望平方差距。而則為平滑參數(smoothness parameter)，由每次常態性檢定的基因數和樣本數一起決定，詳細的關係如下：

而上述的距離函式在經過一系列的轉換及簡化後會變成以下統計量：

其中，，為第個人和第個人的個基因表現量彼此之間馬式距離(Mahalanobis distance)的平方，，為第個人的個基因表現量和所有人的個基因表現量中心(centroid)彼此之間馬式距離的平方，另外，這裡的也等於Mardia’s test中的，

。

而在Henze-Zirkler’s檢定中，我們採取了Wald檢定統計量來對資料進行檢驗，詳細的檢定統計量構造如下：

其中，，。

1. Royston’s test (Korkmaz, Goksuluk, and Zararsiz 2014)

Royston’s檢定將在進行單維度常態性檢測時很常被拿來使用的Shapiro-Wilk以及Shapiro-Francia檢定拓展至高維度，其檢定統計量的構造如下：

其中，，為標準常態分配的分配函數(distribution function)，，，和是由過去文獻(Royston 1982)中使用的多項式所近似出來的結果，，是所有個彼此之間相關係數的平均的估計量(estimator)，代表每次常態性檢定的基因數，而則代表每次常態性檢定的樣本數。

1. One-sample energy test (Székely, and Rizzo 2005)

有別於一般大家常使用的energy test，在這裡的energy test為單樣本(one-sample)的版本，藉由比較資料和多維度常態分佈彼此之間的平均距離來檢測虛無假設是否成立，當距離越大，則代表這筆資料並非來自於多維度常態分佈，而其檢定統計量的構造如下：

其中，代表歐氏範數(Euclidean norm)，，代表每次常態性檢定的個樣本中，第個人的個基因表現量()經過標準化(standardized)後的結果，為所有人的個基因表現量平均，則是所有人個基因表現量的樣本共變異數矩陣，和都是代表度的標準常態分佈，而上述的檢定統計量在經過一系列的轉換以及近似後可以變成以下式子：

其中，，至於這個檢定的p-value則是由拔靴法(bootstrap)來進行計算。

參考文獻

Henze, N., and Zirkler, B. (1990), “A Class of Invariant Consistent Tests for Multivariate Normality,” *Journal Communications in Statistics - Theory and Methods*, 19, 3595-3617.

Korkmaz, S., Goksuluk, D., and Zararsiz, G. (2014), “MVN: An R Package

for Assessing Multivariate Normality,” *The R Journal*, 6, 151-162.

Royston, J. P. (1982), “An Extension of Shapiro and Wilk’s W Test for Normality to Large Samples,” *Applied Statistics*, 31, 115-124.

Székely, G. J., and Rizzo, M. L. (2005), “A New Test for Multivariate Normality,” *Journal of Multivariate Analysis*, 93, 58-80.

# 附錄二、其他附錄資料

**表一 GSA方法在多維度常態分佈情境（低相關）底下的表現**

(此表的數據結果已統整到前面的雷達圖(圖七～圖十)中)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Hotelling’s T2 | Global test | GlobalANCOVA | energy | GSEA(Category) |
| 0.1 | 0 | 0.055 | 0.054 | 0.048 | 0.052 | 0.05 |
| 0.05 | 0.085 | 0.085 | 0.086 | 0.087 | 0.088 |
| 0.1 | 0.144 | 0.175 | 0.171 | 0.173 | 0.217 |
| 0.3 | 0.746 | 0.864 | 0.85 | 0.86 | 0.925 |
| 0.5 | 0.995 | 0.999 | 0.999 | 0.999 | 1 |
| 0.7 | 1 | 1 | 1 | 1 | 1 |
| 0.9 | 1 | 1 | 1 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 |

**表二 GSA方法在多維度常態分佈情境（高相關）底下的表現**

(此表的數據結果已統整到前面的雷達圖(圖七～圖十)中)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Hotelling’s T2 | Global test | GlobalANCOVA | energy | GSEA(Category) |
| 0.9 | 0 | 0.062 | 0.053 | 0.045 | 0.052 | 0.049 |
| 0.05 | 0.064 | 0.054 | 0.051 | 0.056 | 0.055 |
| 0.1 | 0.074 | 0.074 | 0.067 | 0.072 | 0.074 |
| 0.3 | 0.134 | 0.261 | 0.246 | 0.249 | 0.258 |
| 0.5 | 0.259 | 0.572 | 0.56 | 0.55 | 0.573 |
| 0.7 | 0.461 | 0.847 | 0.836 | 0.842 | 0.841 |
| 0.9 | 0.722 | 0.975 | 0.969 | 0.962 | 0.971 |
| 1 | 0.823 | 0.991 | 0.99 | 0.991 | 0.992 |
| 2 | 1 | 1 | 1 | 1 | 1 |

**表三 GSA方法在多維度常態混合分佈情境底下的表現**

(此表的數據結果已統整到前面的雷達圖(圖七～圖十)中)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Hotelling’s T2 | Global test | GlobalANCOVA | energy | GSEA(Category) |
| 0 | 0.069 | 0.056 | 0.05 | 0.059 | 0.051 |
| 0.05 | 0.074 | 0.068 | 0.065 | 0.068 | 0.07 |
| 0.1 | 0.083 | 0.098 | 0.091 | 0.096 | 0.093 |
| 0.3 | 0.173 | 0.415 | 0.399 | 0.406 | 0.408 |
| 0.5 | 0.389 | 0.807 | 0.793 | 0.815 | 0.812 |
| 0.7 | 0.685 | 0.977 | 0.971 | 0.974 | 0.975 |
| 0.9 | 0.907 | 0.999 | 0.999 | 1 | 0.999 |
| 1 | 0.952 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 |

**表四 GSA方法在多維度常態混合分佈情境底下的表現**

(此表的數據結果已統整到前面的雷達圖(圖七～圖十)中)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Hotelling’s T2 | Global test | GlobalANCOVA | energy | GSEA(Category) |
| 0 | 0.069 | 0.061 | 0.052 | 0.064 | 0.059 |
| 0.05 | 0.072 | 0.066 | 0.062 | 0.069 | 0.065 |
| 0.1 | 0.078 | 0.086 | 0.078 | 0.086 | 0.093 |
| 0.3 | 0.122 | 0.31 | 0.291 | 0.278 | 0.31 |
| 0.5 | 0.254 | 0.64 | 0.624 | 0.609 | 0.647 |
| 0.7 | 0.463 | 0.904 | 0.894 | 0.898 | 0.901 |
| 0.9 | 0.711 | 0.988 | 0.984 | 0.987 | 0.987 |
| 1 | 0.816 | 0.995 | 0.995 | 0.995 | 0.995 |
| 2 | 1 | 1 | 1 | 1 | 1 |

**表五 GSA方法在多維度常態混合分佈情境底下的表現**

(此表的數據結果已統整到前面的雷達圖(圖七～圖十)中)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Hotelling’s T2 | Global test | GlobalANCOVA | energy | GSEA(Category) |
| 0 | 0.062 | 0.06 | 0.056 | 0.065 | 0.062 |
| 0.05 | 0.066 | 0.068 | 0.059 | 0.068 | 0.067 |
| 0.1 | 0.065 | 0.072 | 0.065 | 0.077 | 0.075 |
| 0.3 | 0.09 | 0.19 | 0.177 | 0.158 | 0.192 |
| 0.5 | 0.146 | 0.369 | 0.357 | 0.337 | 0.377 |
| 0.7 | 0.232 | 0.63 | 0.61 | 0.581 | 0.636 |
| 0.9 | 0.39 | 0.849 | 0.839 | 0.823 | 0.854 |
| 1 | 0.463 | 0.915 | 0.909 | 0.894 | 0.916 |
| 2 | 0.997 | 1 | 1 | 1 | 1 |

**表六 GSA方法在多維度t分佈情境底下的表現**

(此表的數據結果已統整到前面的雷達圖(圖七～圖十)中)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Hotelling’s T2 | Global test | GlobalANCOVA | energy | GSEA(Category) |
| 0 | 0.062 | 0.053 | 0.05 | 0.049 | 0.059 |
| 0.05 | 0.077 | 0.07 | 0.061 | 0.067 | 0.066 |
| 0.1 | 0.091 | 0.086 | 0.081 | 0.104 | 0.088 |
| 0.3 | 0.215 | 0.317 | 0.305 | 0.425 | 0.32 |
| 0.5 | 0.467 | 0.658 | 0.648 | 0.823 | 0.667 |
| 0.7 | 0.752 | 0.882 | 0.868 | 0.986 | 0.881 |
| 0.9 | 0.918 | 0.962 | 0.959 | 1 | 0.963 |
| 1 | 0.961 | 0.982 | 0.976 | 1 | 0.984 |
| 2 | 1 | 1 | 1 | 1 | 1 |

**表七 GSA方法在情境四底下的表現**

(此表的數據結果已統整到前面的雷達圖(圖十一)中)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Number of  (0.1, 0.3, 0.5) | Hotelling’s T2 | Global test | GlobalANCOVA | energy | GSEA  (Category) |
| (15, 7, 8) | 0.951 | 0.423 | 0.403 | 0.477 | 0.315 |
| (15, 8, 7) | 0.922 | 0.397 | 0.391 | 0.442 | 0.3 |
| (7, 15, 8) | 0.853 | 0.492 | 0.478 | 0.534 | 0.426 |
| (8, 15, 7) | 0.852 | 0.469 | 0.456 | 0.499 | 0.399 |
| (7, 8, 15) | 0.953 | 0.624 | 0.61 | 0.668 | 0.521 |
| (8, 7, 15) | 0.949 | 0.624 | 0.611 | 0.669 | 0.513 |
| (10, 10, 10) | 0.946 | 0.512 | 0.499 | 0.563 | 0.411 |
| 0.5: 18, 0: 12 | 1 | 0.655 | 0.645 | 0.776 | 0.414 |
| 3: 3, 0: 27 | 1 | 1 | 1 | 1 | 0.452 |
| 0.3: 30 | 0.173 | 0.415 | 0.399 | 0.406 | 0.408 |

**表八 GSA方法在情境六底下的表現**

(此表的數據結果已統整到前面的雷達圖(圖十二)中)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Number of  (0.1, 0.3, 0.5) | Hotelling’s T2 | Global test | GlobalANCOVA | energy | GSEA  (Category) |
| (15, 7, 8) | 0.866 | 0.178 | 0.17 | 0.163 | 0.162 |
| (15, 8, 7) | 0.832 | 0.169 | 0.162 | 0.157 | 0.156 |
| (7, 15, 8) | 0.737 | 0.215 | 0.198 | 0.179 | 0.192 |
| (8, 15, 7) | 0.741 | 0.202 | 0.189 | 0.171 | 0.189 |
| (7, 8, 15) | 0.871 | 0.253 | 0.243 | 0.226 | 0.232 |
| (8, 7, 15) | 0.888 | 0.258 | 0.24 | 0.228 | 0.228 |
| (10, 10, 10) | 0.859 | 0.213 | 0.203 | 0.184 | 0.191 |
| 0.5: 18, 0: 12 | 0.997 | 0.241 | 0.229 | 0.239 | 0.192 |
| 3: 3, 0: 27 | 1 | 1 | 1 | 1 | 0.202 |
| 0.3: 30 | 0.09 | 0.19 | 0.173 | 0.158 | 0.192 |

**表九 GSA方法在情境六底下的表現**

(此表的數據結果已統整到前面的雷達圖(圖十三)中)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Number of  (0.1, 0.3, 0.5) | Hotelling’s T2 | Global test | GlobalANCOVA | energy | GSEA  (Category) |
| (15, 7, 8) | 0.994 | 0.308 | 0.298 | 0.571 | 0.249 |
| (15, 8, 7) | 0.993 | 0.289 | 0.284 | 0.535 | 0.242 |
| (7, 15, 8) | 0.992 | 0.384 | 0.365 | 0.607 | 0.333 |
| (8, 15, 7) | 0.991 | 0.363 | 0.348 | 0.578 | 0.31 |
| (7, 8, 15) | 0.994 | 0.48 | 0.464 | 0.746 | 0.422 |
| (8, 7, 15) | 0.994 | 0.473 | 0.467 | 0.748 | 0.414 |
| (10, 10, 10) | 0.994 | 0.39 | 0.375 | 0.652 | 0.319 |
| 0.5: 18, 0: 12 | 0.995 | 0.487 | 0.476 | 0.93 | 0.323 |
| 3: 3, 0: 27 | 1 | 0.998 | 0.998 | 1 | 0.348 |
| 0.3: 30 | 0.215 | 0.317 | 0.306 | 0.425 | 0.32 |

**程式碼(R code)**:

library(Hotelling)

library(globaltest)

library(Category)

library(GlobalAncova)

library(MVN)

library(mvtnorm)

library(energy)

#Random sample from multivariate normal distribution (MVN) with

#customized parameters settings

mvnormal = function(n, probenumber, rho, MEAN = rep(0, probenumber),

rhostructure = "CS", SIGMA = NULL){

#n: Sample size. (樣本數)

#probenumber: The number of dimensions(genes). (維度，基因數目)

#rho: Correlation between any two dimensions(genes).

#(任意兩個維度，基因之間的相關係數)

#rhostructure: Structure of correlation matrix.

#(關係係數矩陣的結構樣式), e.g.:"CS": Compound Symmertry,

#"UN": Unstructured Correlation Matrix type

#SIGMA: the Correlation Matrix in "UN"

#Program start:

#Create correlation matrix

if(rhostructure == "CS"){

SIGMA = matrix(rho, probenumber, probenumber)

diag(SIGMA) = 1

}else if(rhostructure == "UN"){

SIGMA = SIGMA

}

#Generate observation from MVN from package "mvtnorm".

obs = rmvnorm(n, mean = MEAN, sigma = SIGMA)

return(obs)

}

#Random sample from multivariate t distribution(MVT) with customized

#parameters settings

mvt = function(n, probenumber, rho, rhostructure = "CS", DF){

#n: Sample size. (樣本數)

#probenumber: The number of dimensions(genes). (維度，基因數目)

#rho: Correlation between any two dimensions(genes).

#(任意兩個維度，基因之間的相關係數)

#rhostructure: Structure of correlation matrix.

#(關係係數矩陣的結構樣式), e.g.:"CS": Compound Symmertry,

#"UN": Unstructured

#DF: Degrees of freedom of multivariate t distribution, DF > 2.

#(多為t分配的自由度,DF 需要大於 2)

#Program start:

#Create correlation matrtix

if(rhostructure == "CS"){

rhomatrix = matrix(rho, probenumber, probenumber)

diag(rhomatrix) = 1

}else if(rhostructure == "UN"){

rhovector = c(0.1, 0.3, 0.5, 0.7, 0.9)

rhomatrix = matrix(c(sample(rhovector,

size = probenumber\*probenumber,

replace = TRUE)), probenumber,

probenumber)

diag(rhomatrix) = 1

}

#Transform correlation matrix into scale matrix.

#(將相關係數矩陣轉換成多維度t分配中的scale matrix)

if(DF > 2){

SIGMA <<- rhomatrix\*((DF - 2)/DF)

}else{

cat("error: df must > 2 \n")

}

#Generate observation from MVT from package "mvtnorm".

obs = rmvt(n, sigma = SIGMA, df = DF)

return(obs)

}

#Use the concept of probability integral transformation(PIT) to

#create various scenarios: Mixture of MVN, MVT(df=3),

# MVN(high correlation: 0.9,

# low correlation: 0.1)

generateobs = function(n, probenumber,

alternative = c("normal", "normal"),

alternumber = 2,

proportion = rep(1/alternumber, alternumber),

MEAN = list(rep(0, probenumber),

rep(0, probenumber)),

rho, rhostructure = rep("CS", alternumber),

DF = Inf, SIGMA = NULL){

#n: Sample size of the (mixture) distribution.

# (從(混合)分配生出來的樣本數)

#probenumber: The number of dimensions(genes) for the

# (mixture) distribution.

# ((混合)分配的維度(變數、基因)數目)

#alternative: The distribution types vector used for creating the

# (mixture) distribution.

# (創造(混合)分配所需的分配類型)

#alternumber: The number of components in the mixture distribution.

# (混合分配所要用到的分配個數)

#proportion: The proportion for each distribution used for creating

# mixture distribution.

# (各個分配在混合分配中所要的混合比例)

#MEAN: The mean vectors for all distributions used for creating

# mixture distribution.

# (混合分配中每個分配的期望值向量)

#rho: Correlation between any two dimensions(genes) for all

# distributions used for creating mixture distribution.

# (混合分配中每個分配裡任意兩個維度(變數、基因)之間的相關係數)

#rhostructure: Structure of correlation matrix for all distributions

# used for creating mixture distribution.

# (混合分配中的每個分配所需要用到的關係係數

# 矩陣的結構樣式),

# e.g.:"CS": Compound Symmertry, "UN": Unstructured

#DF: Degrees of freedom of multivariate t distribution

# (if it is needed).

# (多維t分配的自由度(如果需要用到的話), DF 需要大於 2)

# DF > 2

#SIGMA: The unstructured correlation matrix ("SIGMA" is needed when

# the rhostructure is set to "UN")

#Program start:

#seed: All data seed

#obs: All generated observation.

#j: The value of indicator function:

# 1 = Distribution 1 of "alternative" vector,

# 2 = Distribution 2 of "alternative" vector.

#p: The cumulative probability of the proportion for

# each distribution in the mixture distribution.

#per: Each generated observation.

seed = runif(n)

obs = c()

for(i in seed){

j = 1

p = proportion[1]

while(TRUE){

if(j > alternumber){

break

}else if(i <= p){

if(alternative[j] == "normal"){

per = mvnormal(n = 1, probenumber = probenumber,

MEAN = MEAN[[j]],

rho = rho[j],

rhostructure = rhostructure[j],

SIGMA = SIGMA[[j]])

}else if(alternative[j] == "t"){

per = mvt(n = 1, probenumber = probenumber, rho = rho[j],

DF = DF, rhostructure = rhostructure[j])

}else{

per = NA

}

obs = rbind(obs, per)

break

}else{

j = j + 1

p = p + proportion[j]

}

}

}

return(obs)

}

######################################################################

#All Test

Test = function(Control, Case, all.sample.size, probenumber,

AA.test, BB.test, CC.test, DD.test, EE.test,

sig = 0.05){

ALL = rbind(Control, Case)

colnames(ALL) = paste0("V",(1:ncol(ALL)))

rownames(ALL) = paste0("N",(1:nrow(ALL)))

m = all.sample.size

print("Hotelling")

resultH\_Pseudo\_Perm = hotelling.test(Control, Case, shrinkage = TRUE,

perm = TRUE, B = 1000,

progBar = FALSE)

print("Global")

resultG\_Perm = gt(rep(c(TRUE, FALSE), c(nrow(Control),nrow(Case))),

ALL ,permutations = 1000)

print("GlobalAncova")

resultGlobalAncova = GlobalAncova(xx = t(ALL),

group = rep(c(0, 1), c(m/2, m/2)),

covars=NULL, perm=1000)

print("N-statistic")

resultN\_Perm = eqdist.etest(ALL, sizes=c(m/2, m/2) , R = 1000)

print("GSEA(Category)")

gseadata = ExpressionSet(t(ALL))

response = as.factor(rep(c("A", "B"), c(m/2, m/2)))

signset = matrix(1, nrow = 1, ncol = probenumber)

colnames(signset) = rownames(gseadata)

rownames(signset) = "geneset"

resultGSEA\_Category = gseattperm(gseadata, response, signset, 1000)

if(resultH\_Pseudo\_Perm$pval < sig){

AA.test = AA.test + 1

}

if(resultG\_Perm@result[1] < sig){

BB.test = BB.test + 1

}

if(resultGlobalAncova$test.result[2,] < sig){

CC.test = CC.test + 1

}

if(resultN\_Perm$p.value < sig){

DD.test = DD.test + 1

}

if((resultGSEA\_Category[1]<sig/2)||(resultGSEA\_Category[2]<sig/2)){

EE.test = EE.test + 1

}

return(c(AA.test, BB.test, CC.test, DD.test, EE.test))

}

######################################################################

#Multivariate Normal distribution (rho=0.1 and 0.9)

#(scenario 1 & 2, elements in the delta vector are the same)

simulationresult = c()

rholist = list()

rholist[[1]] = c(0.1)

rholist[[2]] = c(0.9)

proportionlist = list()

proportionlist[[1]] = c(1)

times = 1000

for(prop in proportionlist){

print(prop)

for(rho in rholist){

print(paste0("rho = ", rho))

for(probenumber in 30){

print(paste0("probenumber = ", probenumber))

for(m in c(70)){

print(paste0("sample = ", m))

for (difference in c(0, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 1, 2)){

print(difference)

#GSA Tools start

names = c("countH\_Pseudo\_Perm", "countG\_Perm",

"countGlobalAncova", "countN\_Perm",

"countGSEA\_Category")

for(i in 1: length(names)) assign(names[i], 0)

for(s in 1: times){

print(s)

set.seed(s)

obs1 = generateobs(n = m/2, probenumber = probenumber,

alternative = "normal",

alternumber = 1,

MEAN = list(rep(0, probenumber)),

proportion = prop, rho = rho)

set.seed(s + 2\*times)

obs2 = generateobs(n = m/2, probenumber = probenumber,

alternative = "normal",

alternumber = 1,

MEAN = list(rep(0 + difference,

probenumber)),

proportion = prop, rho = rho)

result = Test(Control = obs1, Case = obs2,

all.sample.size = m,

probenumber = probenumber,

AA.test = countH\_Pseudo\_Perm,

BB.test = countG\_Perm,

CC.test = countGlobalAncova,

DD.test = countN\_Perm,

EE.test = countGSEA\_Category)

for(i in 1: length(names)) assign(names[i], result[i])

}

count = matrix(c(countH\_Pseudo\_Perm, countG\_Perm,

countGlobalAncova, countN\_Perm,

countGSEA\_Category), 1, length(names))

#GSA Tools end

colnames(count) = names

rownames(count) = paste0("composition:",

paste0(as.character(prop),

collapse = "", sep = " "),

",rho:",

paste0(as.character(rho),

collapse = "", sep = " "),

",probes:", probenumber,

",size:", m,

",difference:", difference)

simulationresult = rbind(simulationresult, (count/times))

}

}

}

}

}

######################################################################

#Mixture of Normal with different mean setting (0 and 1)

#(scenario 5, elements in the delta vector are the same)

#The code for scenario 3, 4 and 6 is omitted.

simulationresult = c()

rholist = list()

rholist[[1]] = c(0.1, 0.9)

proportionlist = list()

proportionlist[[1]] = c(0.5, 0.5)

times = 1000

for(prop in proportionlist){

print(prop)

for(rho in rholist){

print(rho)

for(probenumber in c(30)){

print(probenumber)

for(m in c(70)){

print(m)

for (difference in c(0, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 1, 2)){

print(difference)

#GSA Tools start

names = c("countH\_Pseudo\_Perm", "countG\_Perm",

"countGlobalAncova", "countN\_Perm",

"countGSEA\_Category")

for(i in 1: length(names)) assign(names[i], 0)

for(s in 1: times){

print(s)

set.seed(s)

obs1 = generateobs(n = m/2, probenumber = probenumber,

MEAN = list(rep(0, probenumber),

rep(1, probenumber)),

proportion = prop, rho = rho)

set.seed(s + 2\*times)

obs2 = generateobs(n = m/2, probenumber = probenumber,

MEAN = list(rep(0+difference,probenumber),

rep(1+difference,probenumber)),

proportion = prop, rho = rho)

result = Test(Control = obs1, Case = obs2,

all.sample.size = m,

probenumber = probenumber,

AA.test = countH\_Pseudo\_Perm,

BB.test = countG\_Perm,

CC.test = countGlobalAncova,

DD.test = countN\_Perm,

EE.test = countGSEA\_Category)

for(i in 1: length(names)) assign(names[i], result[i])

}

count = matrix(c(countH\_Pseudo\_Perm, countG\_Perm,

countGlobalAncova, countN\_Perm,

countGSEA\_Category), 1, length(names))

#GSA Tools end

colnames(count) = names

rownames(count) = paste0("composition:",

paste0(as.character(prop),

collapse = "", sep = " "),

",rho:",

paste0(as.character(rho),

collapse = "", sep = " "),

",probes:", probenumber, ",size:",

m, ",difference:", difference)

simulationresult = rbind(simulationresult, (count/times))

}

}

}

}

}

######################################################################

#Multivariate t-distribution df = 3

#(scenario 3, elements in the delta vector are different, Difference: B)

simulationresult = c()

rholist = list()

rholist[[1]] = c(0.9)

proportionlist = list()

proportionlist[[1]] = c(1)

DF = 3

set.seed(2)

difference1 = sample(c(rep(0.1, 15), rep(0.3, 7), rep(0.5, 8)))

set.seed(2)

difference2 = sample(c(rep(0.1, 15), rep(0.3, 8), rep(0.5, 7)))

set.seed(2)

difference3 = sample(c(rep(0.1, 7), rep(0.3, 15), rep(0.5, 8)))

set.seed(2)

difference4 = sample(c(rep(0.1, 8), rep(0.3, 15), rep(0.5, 7)))

set.seed(2)

difference5 = sample(c(rep(0.1, 7), rep(0.3, 8), rep(0.5, 15)))

set.seed(2)

difference6 = sample(c(rep(0.1, 8), rep(0.3, 7), rep(0.5, 15)))

set.seed(2)

difference7 = sample(c(rep(0.1, 10), rep(0.3, 10), rep(0.5, 10)))

set.seed(2)

difference8 = sample(c(rep(0, 12), rep(0.5, 18)))

set.seed(2)

difference9 = sample(c(rep(0, 27), rep(3, 3)))

difference10 = sample(c(rep(0, 0), rep(0.3, 30)))

differencelist = list()

differencelist[[1]] = difference1

differencelist[[2]] = difference2

differencelist[[3]] = difference3

differencelist[[4]] = difference4

differencelist[[5]] = difference5

differencelist[[6]] = difference6

differencelist[[7]] = difference7

differencelist[[8]] = difference8

differencelist[[9]] = difference9

differencelist[[10]] = difference10

lapply(X = differencelist, FUN = table)

times = 1000

for(prop in proportionlist){

print(prop)

for(rho in rholist){

print(rho)

for(probenumber in c(30)){

print(probenumber)

for(m in c(70)){

print(m)

for (difference in differencelist) {

print(table(difference))

#GSA Tools start

names = c("countH\_Pseudo\_Perm", "countG\_Perm",

"countGlobalAncova", "countN\_Perm",

"countGSEA\_Category")

for(i in 1: length(names)) assign(names[i], 0)

for(s in 1: times){

print(s)

set.seed(s)

obs1 = generateobs(n = m/2, probenumber = probenumber,

alternative = "t",

alternumber = 1, proportion = prop,

rho = rho, DF = DF)

set.seed(s + 2\*times)

obs2 = t(difference+t(generateobs(n = m/2, probenumber=probenumber,

alternative = "t",

alternumber = 1,

proportion = prop,

rho = rho, DF = DF)))

result = Test(Control = obs1, Case = obs2,

all.sample.size = m,

probenumber = probenumber,

AA.test = countH\_Pseudo\_Perm,

BB.test = countG\_Perm,

CC.test = countGlobalAncova,

DD.test = countN\_Perm,

EE.test = countGSEA\_Category)

for(i in 1: length(names)) assign(names[i], result[i])

}

count = matrix(c(countH\_Pseudo\_Perm, countG\_Perm,

countGlobalAncova, countN\_Perm,

countGSEA\_Category), 1, length(names))

#GSA Tools end

colnames(count) = names

rownames(count)=paste0("composition:",

paste0(as.character(prop),

collapse = "", sep = " "),

",rho:",

paste0(as.character(rho),

collapse = "", sep = " "),

",probes:",

probenumber, ",size:", m,

paste0(",difference(",

paste(names(table(difference)),

collapse = " "), ")"),

paste0(as.character((table(difference))),

collapse = "", sep = " "))

simulationresult = rbind(simulationresult, (count/times))

}

}

}

}

}

######################################################################

#Unstructured Correlation Matrix Generation

seed = 1691

#AAA: Correlation Matrix 1, seed: 1691,

#BBB: Correlation Matrix 2, seed: 1718

set.seed(seed)

rhovector = c(-0.9, -0.8, -0.7, -0.6, -0.5, -0.4, -0.3, -0.2, -0.1, 0,

0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9)

probenumber = 30

allrho = sample(rhovector, size = choose(probenumber, 2),

replace = TRUE)

SIGMA = matrix(0, probenumber, probenumber)

ct = 1

for (i in 1:nrow(SIGMA)) {

for (j in (i+1) : ncol(SIGMA)) {

if(i+1 > ncol(SIGMA)){

next

}else{

SIGMA[i, j] = allrho[ct]

ct = ct + 1

}

}

}

SIGMA = SIGMA + t(SIGMA)

diag(SIGMA) = 1

SIGMA = SIGMA%\*%SIGMA

SIGMA = cov2cor(SIGMA)

SIGMA = apply(SIGMA, 1, function(x) round(x, 2))

AAA = SIGMA

#BBB = SIGMA

#######################################################################Run the GSA under MVN Scenario with UN correlation matrix

simulationresult = c()

rholist = list()

rholist[[1]] = NA

proportionlist = list()

proportionlist[[1]] = NA

SIGMA = list()

SIGMA[[1]] = AAA

SIGMA[[2]] = BBB

times = 1000

for(prop in proportionlist){

print(prop)

for(rho in rholist){

print(rho)

for(probenumber in c(30)){

print(probenumber)

for(m in c(70)){

print(m)

for (difference in c(0, 0.1, 0.2, 0.25, 0.3, 0.5, 0.7, 0.9, 1, 2)){

print(difference)

#GSA Tools start

names = c("countH\_Pseudo\_Perm", "countG\_Perm",

"countGlobalAncova", "countN\_Perm",

"countGSEA\_Category")

for(i in 1: length(names)) assign(names[i], 0)

for(s in 1: times){

print(s)

set.seed(s)

obs1 = generateobs(n = m/2, probenumber = probenumber,

proportion = 1,

MEAN = list(rep(0, probenumber)),

rho = rho, alternumber = 1,

rhostructure = rep("UN", 1),

SIGMA = SIGMA)

set.seed(s + 2\*times)

obs2 = generateobs(n = m/2, probenumber = probenumber,

proportion = 1,

MEAN = list(rep(0 + difference, probenumber)),

rho = rho, alternumber = 1,

rhostructure = rep("UN", 1),

SIGMA = SIGMA)

result = Test(Control = obs1, Case = obs2,

all.sample.size = m,

probenumber = probenumber,

AA.test = countH\_Pseudo\_Perm,

BB.test = countG\_Perm,

CC.test = countGlobalAncova,

DD.test = countN\_Perm,

EE.test = countGSEA\_Category)

for(i in 1: length(names)) assign(names[i], result[i])

}

count = matrix(c(countH\_Pseudo\_Perm, countG\_Perm,

countGlobalAncova, countN\_Perm,

countGSEA\_Category), 1, length(names))

#GSA Tools end

colnames(count) = names

rownames(count) = paste0("composition:",

paste0(as.character(prop),

collapse = "", sep = " "),

",rho:",

paste0(as.character(rho),

collapse = "", sep = " "),

",probes:",

probenumber, ",size:", m,

",difference:", difference)

simulationresult = rbind(simulationresult, (count/times))

}

}

}

}

}