探索適用於 DNA 甲基化生物標記實驗的內部控制基因

專題編號：112-CSIE-S022
執行期限：111 年第 1 學期至 112 年第 1 學期
指導教授：白敦文
專題參與人員： 109590009 楊明哲
109590010 林敬翔
109590044 蕭詠之

一、摘要

有鑑於國內適用於 DNA 甲基化實驗所使用內控基因的選擇探討仍付之闕如，因此本專題計畫將透過臺灣人體生物資料庫中針對 DNA 甲基化實驗的數據所取得的甲基化數據，綜合受測者的共病資料分析，將罹患類似疾病的病患資料進行分群，以正確選取在不同的疾病類型都能具有穩定表現的內控基因，並透過標準甲基化晶片分析流程，探索適用於 DNA 甲基化實驗分析的最佳候選內部控制基因，本研究確定了兩個新的內控基因 (MYH14、PRHOXNB)，並與傳統的管家基因 (GAPDH、ACTB) 進行比較。藉由機器學習方式（SVM、Random Forest）對使用不同內控基因作為參考受試者進行穩定性驗證。結果顯示出新的內部控制基因，為疾病的 DNA 甲基化生物標記檢測套件提供了更好的穩定性和適用性。

關鍵詞：DNA 甲基化、內部控制基因、大數據、機器學習

二、緣由與目的

基因科學發展非常快速，尤其在 DNA 甲基化的研究議題備受關注，根據莊樹諄教授研究，科學界已明確認定，DNA 甲基化是一個控制基因表現的重要關鍵，對於未來會是一大醫療趨勢[1]，其應用可對於不同特定基因的甲基化表現分析，以這些基因位點可檢視各類疾病或癌症的重要表觀遺傳生物標記[2]，使醫療團隊可針對個人的狀態提供最佳的預防或治療相關措施。在進行 DNA 研究的生物實驗時，基因表現量僅是一個數值，而這些數值無法直接表示基因反應強弱。因此經常會在實驗過程使用管家基因做為內控參考基準，因為管家基因在生物中的表現穩定之特性，被廣泛作為實驗和研究的內部對照。然而特定管家基因在不同病狀的表現量不同，導致在不同實驗條件下會產生差異，造成所選用的管家基因無法作為客觀參考的基準，以至於對不同疾病執行 DNA 實驗的判讀時造成錯誤結果解讀，從而導致嚴重錯誤的推論[3]。

綜上所論，本專題研究將藉由基因體甲基化的實驗數據進行分析，有效挑選在不同疾病影響之下，觀察那些 DNA 甲基化位點的反應仍可保留穩定表現，這些位點正適合作為甲基化實驗過程的內控基因，也就是作為 DNA 甲基化疾病的客觀參考基準。

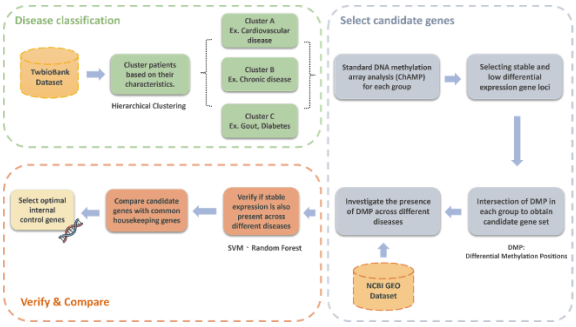
三、研究報告內容

(一)、研究方法

本研究使用台灣人體生物資料庫受測者疾病資訊進行分析。首先，使用階層式分群法，利用疾病之間所生成的距離矩陣進行共病特徵分群，分別依各資料群進行差異性甲基化位點分析，以挑選同族群樣本具有高穩定性且差異性小之甲基化基因位點，作為分析之候選內部控制基因群。

確定候選內部控制基因群之後，利用公開資料集進行實驗分析，以確認這些選取的基因在不同疾病中的表現反應是否都沒有顯著差異。同時尋找候選內部控制基因，並運用基因位點之表現差異量作為特徵，使用機器學習方法建構模型。驗證候

選內部控制基因的在預測模型中的準確率，且與生物學家常用之管家基因比較。最終選擇出最適合用於國人 DNA 甲基化實驗分析的最佳內部控制基因。



圖一、研究流程圖

(二)、研究步驟

1. 病患資料分群

本研究利用臺灣人體生物資料庫之受測者共病資料(TWBR11012_04)，計算疾病間的相似度矩陣，範圍從[-1,1]。透過將相似度矩陣轉成距離矩陣(1.1)，進行疾病的階層式分群。從初始的每種疾病作為單獨群組開始，逐步合併至預定群組數目。提供疾病相關性的分析，可以對於疾病間相互關係的深入理解，特別是尋找適用於各疾病的內控基因，更提供了有力的依據。

$$distance = 1 - correlation \qquad (1.1)$$

2. 差異性甲基化位點分析

由於差異甲基化位點在不同樣本之間具有較高的變異性，因此選擇這些位點作為候選基因可以幫助篩選出相對穩定的內部控制基因。本計畫將採用 ChAMP(Chip Analysis Methylation Pipeline) 套件配合 EPIC 晶片，找出具差異性之甲基化位點，並同時篩選組別具高穩定表現量(β 值，此為 BMIQ 正規化後的基因表現量)且差異性(Δβ)小的基因位點。

3. 探索及驗證最佳候選內控基因

挑選出具有潛力的內控基因群後，使用機器學習技術，結合公開資料集建立疾病預測模型，並驗證不同內控基因群的效能。經過甲基化晶片分析，挑選特定基因位點進行特徵差異計算，再利用 SVM 和 Random Forest 進行模型建立和預測。進一步透過 5-fold 交叉驗證確保模型的泛化能力，探討最佳的內控基因選擇。

四、研究結果

(一)、分群結果

總體而言，我們根據特徵將疾病劃分為五個緊密相關的群組。這五組分別代表：痛風與冠狀動脈相關疾病、特定女性健康問題、與胃功能相關的疾病、中樞神經和呼吸系統的問題，以及各類癌症。

(二)、差異性甲基化位點分析結果

運用台灣人體生物資料庫與公開資料集進行標準甲基化晶片分析。設定固定 β 值閾值並篩選 $\Delta\beta$ 小於 0.15 的內控候選基因，確保特徵基因位點在不同疾病皆有顯著甲基化差異。結果正確選取了 MYH14 和 PRHOXNB 作為候選內控基因，透過比較，發現常用的 ACTB 變異性較低。

(三)、內部控制基因與常用管家基因比較

使用 SVM、Random Forest 技術及 5-fold 交叉驗證，比較候選內控基因(MYH14、PRHOXNB)與常用管家基因(GAPDH、ACTB)在不同疾病資料集的穩定度。結果顯示，候選內控基因表現相近或超越實驗常用之管家基因，平均準確率與未校正差異誤差皆低於 5%，進一步發現候選內控基因在結直腸癌資料集中皆明顯優於 ACTB。

五、參考文獻

[1] 莊樹諄（2013）。就算是甲基化，也要看位置。中央研究院 基因體研究中心

[2] Alexandre How Kit & Helene Myrtue Nielsen & Jörg Tost (2012). “DNA methylation based biomarkers: Practical considerations and applications”. *Biochimie*, 94(11), 2413-2337

[3] Veronika Kloibert & Lothar Rink (2019).“Selection of an inadequate housekeeping gene leads to misinterpretation of target gene expression in zinc deficiency and zinc supplementation models”. *Journal of Trace Elements in Medicine and Biology*, 56, 192-197