乳癌患者特定基因突變之甲基化生物標記分析

專題編號：113-CSIE-S009
執行期限：112 年第 1 學期至 113 年第 1 學期
指導教授：白敦文
專題參與人員：110590005 蕭耕宏
110590014 張匯吾
110590033 陳庠蓁
110590450 歐佳昀

一、摘要

近年來罹患乳癌人數逐年提升已明顯對女性健康產生極大影響，經世界衛生組織與世界癌症研究基金會報告，乳癌已成為全球罹癌人數第一名。故本專題使用乳癌樣本，探討對乳癌患者提供具早期偵測效能之甲基化生物標記分析。透過 GDC(Genomic Data Commons) 和 GEO (Gene Expression Omnibus)提供乳癌患者 DNA 甲基化晶片資料集(包含組織及血液樣本)，綜合患者所罹患之共病資料分析。針對 DNA 甲基化位點具高度差異表現候選位點挑選重要生物標記。透過基因本體論(Gene Ontology, GO)之基因功能註釋分群，結合 RFE(Recursive Feature Elimination) 和 Boruta 挑選出八個重要的 DNA 甲基化生物標記，並依基因功能聚類分為三個生物標記組合。針對組織樣本進行預測之平均 F1-score 可達 0.95~0.96，血液液態樣本則達 0.78~ 0.81。此結果將做為後續與醫療研究團隊進行臨床實驗的重要參考依據。

關鍵詞：乳癌、DNA 甲基化、生物標記、機器學習、液態活檢

二、緣由與目的

世界衛生組織與世界癌症研究基金會等醫療機構指出，罹患乳癌人數已成為全球罹癌人數第一位[1]。傳統乳癌檢測方式需進行腫瘤切片獲取樣本進行檢驗。該程序具侵入性，可能對患者造成不適和手術風險，不利於早期診斷。得益於基因科學與大數據應用快速發展，透過採集血液樣本能提供早期探索[2]，並有效檢測是否已有罹患乳癌症狀的最佳甲基化生物標記組合。本研究以乳癌為目標癌症，藉由蒐集腫瘤組織和血液樣本之甲基化基因晶片公開資料集，加上乳癌患者的共病資訊與疾病基因之關聯性，進行 DNA 甲基化候選生物標記分析，探索檢測乳癌的最佳甲基化生物標記組合。提供女性在一般健檢過程即可用血液樣本進行非侵入性之早期乳癌檢測機制，而非在腫瘤形成後才能發現乳癌。

三、研究報告內容

本研究將液態檢體樣本(GSE243529)和 GDC 腫瘤組織檢體資料集，以 8:2 比例分為訓練和驗證進行具有高度差異性之甲基化位點分析。以層次聚類分群法結合共識矩陣，對基因註釋進行功能分群。確定基因群組後，使用兩種方法挑選。第一種方法是以 Boruta [3] 結合 RFECV (recursive feature elimination with cross-validation)觀察特徵選擇及性能指標變化；第二種方法同時使用 RFE[4]搭配機器學習模型性能指標，探索乳癌特徵集的最佳篩選模型，挑選具準確性與適應性的甲基化生物標記組合。系統流程圖如圖 1 所示。

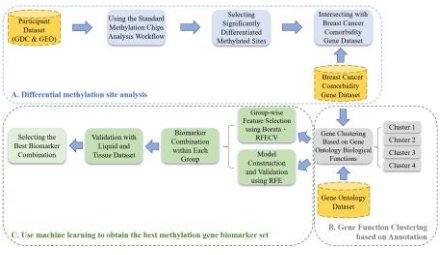

圖 1 研究流程圖

(一) 差異性甲基化分析

為提高後續挑選具正確性之生物標記，以 ChAMP (Chip Analysis Methylation Pipeline)套件計算 Δβ 並篩選甲基化差異較大的基因位點，再和乳癌共病之疾病基因資料集交集分析。為增加所選生物標記用於未來生物實驗的可行性，僅挑選位在基因啟動子區域內(TSS1500/TSS200)的位點。

(二) 基因功能註釋及標籤分群

因挑選相同基因功能群的生物標記，無法全面了解基因生物學功能之間的關係。故以基因本體論進行基因註釋，計算 GO 註釋(BP、CC、MF)基因功能之相似度矩陣，以進行層次聚類分群。接續計算同群中相異基因出現次數轉換為共識矩陣，並以 Silhouette 指標確定最佳聚類數量。

(三) 機器學習篩選特徵組合

綜合以下兩種篩選方法結合分群結果，評估各群代表基因組合及顯著差異的高甲基化位點，篩選表現最佳的候選生物標記組合。最終在液態和組織資料集進行驗證。根據 F1-score、sensitivity、specificity、precision 和 accuracy 等指標，同時使用 OncoScore 工具評估與癌症的潛在關聯性。

1.使用 Boruta 和 RFECV 選擇生物標記

2.多模型評估 RFE 生物標記選擇

四、研究結果

根據 GSE243529 訓練集 (80%) 之 |Δβ|>0.01 及 GDC 訓練集(80%)|Δβ| > 0.15 為門檻值，篩選 60 個位於 TSS1500/200 候選生物標記。再根據基因功能註釋分群，以 Silhouette 指標確定最佳聚類數量為四群。接續說明使用以下兩種方法篩選甲基化生物標記組合及使用 3 份液態和 1 份組織資料集預測表現結果。

(一) 使用 Boruta 和 RFECV 選擇生物標記

使用 Boruta 和 RFECV 進行挑選，分別挑選 33 和 36 個候選生物標記。經結果交集可篩選出 32 個。其中位於啟動子區域及具顯著高甲基化的位點共 11 個。再根據分群驗證後挑選一組最佳生物標記組合。

(二) 多模型評估 RFE 生物標記選擇

預留 25、30 或 35 個生物標記進行 RFE 篩選，以四種機器學習模型 eXtreme Gradient Boosting、Random Forest、SVM 和 Logistic Regression 比較分析。在不同數量下，DT 模型驗證 35 個生物標記預測表現最佳。故將四種模型各挑選的 35 個生物標記交集，剩下 24 個位點。為避免 DT 模型限制產生偏差，挑選前 5 名表現較佳模型所共同挑選的位點，總共 21 個。綜合兩種結果聯集獲得 26 個位點，其中在啟動子區域及具顯著高甲基化共 11 個，分群驗證後挑選兩組最佳生物標記組合。

綜合結果挑選出三組生物標記組合，其在組織樣本平均預測準確度為 0.91~0.93，液態樣本則為 0.71~0.76。其中 CMTM5、PDCD1LG2、MIR124-3、NEFM、PTF1A、CX3CL1 和 PCYT2 之 OncoScore 數值皆超過 21，表示與癌症具顯著關聯，且皆有文獻佐證在乳癌中具潛在影響。

五、參考文獻

[1] https://www.wcrf.org/cancer-trends/worldwide-cancer-data/

[2] J. Liu et al., “Genome-wide cell-free DNA methylation analyses improve accuracy of non-invasive diagnostic imaging for early-stage breast cancer,” Molecular Cancer, vol. 20, no. 1, Feb. 2021

[3] Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta–a system for feature selection. Fundamenta Informaticae, 101(4), 271-285.

[4] A. Thalor, H. Kumar Joon, G. Singh, S. Roy, and D. Gupta, “Machine learning assisted analysis of breast cancer gene expression profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer,” Computational and Structural Biotechnology Journal, vol. 20, pp. 1618–1631, Jan. 2022