DNA methylation analysis for-ovarian cancer biomarker discovery

專題編號：113-CSIE-S011
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： 110820014 張瑀珊
110820043 陳思瑾
110820048 許家睿

一、摘要

本研究專題使用美國 GEO 和 TCGA 資料進行卵巢癌患者的 DNA 甲基化分析，透過 ChAMP 套件進行資料過濾和正規化程序，篩選 263 個候選基因的甲基化位點。接著使用 GontoSim 和 K-means 演算法依基因的功能分成四群。並透過找最佳切點值對基因進行初步的排名。為了改善訓練效果及預測模型建構，採用 SMOTE 技術擴增資料集。再來使用 Random forest 和 SVM 方法進行訓練，並從多個模型挑選效果最佳的組合。最後透過與液態樣本比對，並針對最終挑選出的三個基因 NCOR2、PIP4K2A 及 GRM8 進行相關論文的驗證，發現與卵巢癌高度相關的 NCOR2 在我們的研究中也有不錯的表現。

二、緣由與目的

癌症的早期檢測對於提高治療成功率和患者生存率至關重要。然而，目前大多數癌症的檢測方法主要依賴於採集腫瘤組織樣本進行病理學檢測，這不僅對患者侵入性大，且往往在癌症已進展至中後期時才能確定診斷。因此，我們希望能透過 DNA 甲基化分析，尋找與卵巢癌相關的特定基因，評估這些基因的 DNA 甲基化表現，以及探索是否可以使用血液或子宮頸抹片樣本進行卵巢癌檢測，從而達到早期檢測和低侵入性的目的。

三、架構流程

(一) 資料蒐集和初步篩選

從 GEO 和 TCGA 資料中心下載卵巢癌患者的 IDAT 數據檔，並使用 R 語言的 ChAMP 套件進行初步甲基化表現分析。我們使用到其中的 Quality Control 和 Normalization 來對載入的資料進行過濾和正規化，最後利用套件取得具高度差異性的甲基化位點。這一系列流程將所有 121 位受測者（其中有卵巢癌的患者為 114 位，正常受測者有 7 位）最初的 376,411 個位點篩選後剩下 123,797 個位點。接著我們將從這些位點挑選主要生物標記，透過設立門檻值、刪除重複基因，以及與共病相關的次要生物標記進行交集後，選定 263 個基因位點，針對這些候選基因位點進行深入研究分析。

(二) 基因功能分群

為了避免最後挑選到的基因之間相似度太高，使結果過於偏向某一個方向，我們先進行基因功能分群，透過 DAVID（一個免費且公開的生物學平台）整理每個基因的 Gene Ontology 功能標註，首先分成 CC、BP、MF 三個功能分支，接著使用 GontoSim（Gene Ontology based Similarity）計算基因之間的功能相似度，最後將三個分支加權平均取得最終的基因相似度並轉成相似度矩陣。接著將相似度矩陣轉為距離矩陣，再進行基因功能分群分析，最後使用 K-means 將 263 個基因位點分成四個群組。

(三) 找最佳切點值

我們使用四種方法探索最佳切點值，分別使用 Youden Index (J)、The Closest to (0,1) Criteria (ER) 、 Concordance Probability Method (CZ) 和 Index of Union (IU)。分別用四種方法找出最佳切點值和正確率分析，再使用驗證資料集（共 20 位受測者資料，10

個 tumor，10 個 normal 且與原資料集交集後得到 256 個位點）計算每個位點的正確率，並對結果進行初步的排名。

(四) SMOTE 增加資料集

在進行下一個 Random forest 機器學習核心技術時，我們發現因為現在原資料集的 normal 數量太少（共有 121 位受測者，其中 tumor 為 114 位，normal 為 7 位），導致訓練的結果不佳，因次我們決定先使用 SMOTE 少數類過採樣技術（synthetic minority oversampling technique）來增加 normal 的資料集至 30 人，因此後面的資料是使用全體 144 位的甲基化資料，其中 tumor 為 114 位，normal 為 30 位。

(五) Random forest

接續擴增後的資料集進行 Random forest 預測模型建構，以擴增後的資料集（共 144 位資料，其中 tumor 為 114 位，normal 為 30 位）訓練，驗證資料集（共 20 位受測者資料，10 個 tumor，10 個 normal）驗證，設定子決策樹 500 顆，特徵值設定為 16，預測的結果就比原資料集的表現更為優異，在訓練模型的表現為 accuracy:0.75, sensitivity:1, specificity: 0.5, F1-score: 0.8。並且在 R 語言的 Random forest 套件中可以標註每個基因的個別重要性。

(六) SVM 支持向量機

我們首先進行特徵選擇，從各群中挑選出 beta difference 較大的基因，希望這樣可以更有效地區分正常與癌症樣本。初次訓練結果表明，僅使用這些特徵的分類效果不佳。因此，我們調整策略，改為從每個基因群中挑選先前 Random forest 重要度及切點值排名中前五名的基因，並進行隨機排列組合，以找出最佳的基因組合。

經過多次嘗試後，我們使用非線性核函數，進一步提升預測模型的辨識能力。最終以擴增後的資料集（共 144 位資料，其中 tumor 為 114 位，normal 為 30 位）訓練，驗證資料集（共 20 位受測者資料，10 個 tumor，10 個 normal）作為驗證， 得到前幾組優秀模型組合，分別是由: C1QTNF4, GRM8, NCOR2, PIP4K2A 以及 C1QTNF4, GRASP, IRX1, CEP55 所建構的預測模型，系統表現 Accuracy: 1, Sensitivity: 1, F1-score: 1; 而使用 KCNB1, GRASP, NCOR2, SMG6 組合得到的預測模型表現為 Accuracy: 0.95, Sensitivity: 1, F1-score: 0.95 的預測結果。其中有重複抓取的基因，分別是 NCOR2、GRASP、C1QTNF4。

(七) 液態樣本

我們從 GEO 蒐集卵巢癌患者的血液樣本，總共得到 normal 9 位，tumor 11 位。未篩選前有 866,837 個位點，經過同步驟一的篩選後，取得 220 個基因。接著與原組織樣本的 263 個基因交集後得到 20 個基因。其中有三個基因與我們先前挑選出的高度相關基因重疊，並且在液態樣本中處於高甲基化狀態，分別是 NOCR2、PIP4K2A、GRM8。這三個基因加上 C1QTNF4 的組合更是在 SVM 中得到的模型表現為 Accuracy: 1，且 NOCR2 在 SVM 中是有重複抓取到的。

四、結論

我們針對最後挑選出的三個基因尋找論文驗證，分別是 NCOR2、PIP4K2A 及 GRM8。其中值得一提的是 NCOR2 有三篇較新的論文，發布於 2023 至 2020 年間，皆表示 NCOR2 對於卵巢癌有一定的相關，顯示其作為潛在生物標記的價值，以及本研究的可靠性，更是探索這些基因在血液檢測中的可行性，為未來的研究和臨床應用提供了新的方向。

五、參考文獻

[1] “Nuclear receptor co-repressor NCOR2 and its relation to GPER with prognostic impact in ovarian cancer,” Journal of Cancer Research and Clinical Oncology, vol. 149, no. 11,pp.8719-8728,2023.DOI:10.1007/s00432-023-04708-z.

[2] “Multi-omics analysis of the Indian ovarian-cancer-cohort-revealed-histotype-specific mutation and gene expression patterns,” Frontiers in Genetics, vol. 14, article no. 1102114,Apr.6,2023.DOI:10.3389/fgene.2023.1102114.

[3] “The clinicopathological and genetic features of ovarian diffuse large B-cell lymphoma,” Pathology, vol. 52, no. 2, pp. 206-212,Feb.2020.DOI:10.1016/j.pathol.2019.09.014.