

對比歐亞慢性腎臟病患者的基因甲基化分析

專題編號：114-2-CSIE-S001

執行期限：114 年第 1 學期至 114 年第 2 學期

指導教授：白敦文

專題參與人員：112590453 詹東儒

112590034 高翎竣

112590018 劉玉凱

112590017 胡育鈞

一、摘要

此專題旨在結合機器學習與富集分析技術，發掘跨族群慢性腎臟病之 DNA 甲基化通用生物標記。本研究利用 GEO 公開數據集，經 ChAMP 套件進行正規化與嚴謹之資料前處理後，採用「彈性網路自適應交叉驗證模型」(ElasticNet)進行特徵降維。為確立跨國共通之生物標記，本研究依序進行通路富集分析，並取兩國數據之通路、基因與 CpG 位點交集。最終鎖定共通 CpG 位點作為生物標記，並透過隨機森林與邏輯回歸模型進行跨世代一致性分析，評估模型在 CKD 分類與預測的準確性與穩健性。分析結果顯示這些特徵雖源於不同人種，卻共同指向特定的致病生物路徑。**關鍵詞：**DNA methylation、Chronic Kidney Disease、ElasticNet、Differential Methylation Position、Rank Normalization。

二、緣由與目的

2024 年台灣全民健康保險醫療費用前二十大疾病中，「急慢性腎臟疾病」長年高居健保支出的第一名，緊隨其後的糖尿病及高血壓性疾病與其高度相關[1]。為避免健保財務負擔過度沉重，從源頭減緩 CKD 惡化，以確保醫療福利制度的永續發展。本專案使用國際開放之基因體資料集 GEO，結合美國與日本兩大獨立跨國群體 DNA 甲基化資料集。首先依據群體之年齡、性別及種族等特徵進行資料分層與基線校正，並標記不同疾病特徵患者的甲基化數據。接著，透過跨國數據的交叉比對與外部驗證，篩選不受人種與環境限制的「一致性 CpG 甲基化」，特徵建立一套完整的機

器學習系統。」隨後，本研究運用機器學習演算法進行特徵篩選，提取最具代表性的重要生物標記，以建立高準確性的 CKD 早期預測模型，並進一步透過富集分析對系統生物學進行差異性解讀，尋找不同人種在 CKD 中共同的問題基因。本研究期望結合機器學習與富集分析發掘跨族群慢性腎臟病之 DNA 甲基化通用生物標記，落實疾病早期診斷並降低健保支出的長遠目標。

三、研究範圍

(一) 對象範圍

只針對「第二型糖尿病合併慢性腎臟病患者」與「健康對照組」，不探討其他類型的腎臟病。

(二) 數據範圍

只使用美國與日本的 GEO 公開數據庫資料，範圍只在 DNA 甲基化(CpG 位點)的微陣列表現。

(三) 技術範圍

使用純資訊科學與統計方法(如 ElasticNet、Random Forest、Rank Normalization 等)進行運算與預測，不涉及醫院後續實體的臨床抽血檢驗或濕實驗室操作。

(四) 驗證範圍

驗證跨越美、日兩類不同族群人種的生物標記一致性與預測力。

四、使用技術方法及架構流程

(一) 資料來源

此專題使用來自國際開放基因體資料庫 (GEO)之兩組獨立跨國甲基化微陣列數據，包含美國群體[2]與日本群體[3]的近端

腎小管上皮細胞樣本。資料依據臨床診斷分為慢性腎臟病實驗組與健康對照組。

(二) 資料前處理與批次效應校正

透過 Series Matrix File 自行製作 Sample sheet，並使用 R 語言與 ChAMP 套件處理 EPIC 晶片，對各別樣本做 QC 處理。針對晶片數據跨實驗室的尺度偏移與批次效應，以及避免嚴重的 data leakage，本研究並非使用生物資訊常使用的 ComBat 標準化手法，改採「Per-sample Rank Normalization」。將各樣本內的所有 CpG 位點 β 值轉換為相對排序之方式處理。

(三) 跨國一致性特徵交集分析

為萃取不受人種、地區及實驗室限制的通用致病標記，本研究首先對美、日兩國資料分別進行獨立的差異甲基化位點篩選。首先過濾出具備臨床代表性的變異探針，隨後將兩國入選的顯著位點進行交集，並比對其甲基化變動的方向性，強制篩選不同跨國群體所呈現同步超甲基化或低甲基化的核心特徵。透過此一致性篩選，排除單一族群特有的遺傳與環境噪音。

(四) ElasticNet 特徵精煉

在模型自帶的五折交叉驗證框架下，特徵標準化的參數計算 (fit) 嚴格限定於各迴圈的訓練折疊，測試折疊僅進行轉換應用。處理後的標準化特徵隨即送入 ElasticNet [1]。演算法融合 LASSO(L1)正規化的稀疏剔除特質與 Ridge(L2)正規化處理高度共線性特徵的群體效應，保留具生物學群體代表性的 CpG 位點之外，篩選出分類貢獻度最高的核心關鍵特徵。

$$Loss = MSE + \lambda(\rho L_1 + (1 - \rho)L_2) \quad (1)$$

(五) 富集分析

將關鍵特徵位點進行基因註釋，與國際生物通路資料庫進行交叉比對，識別並篩選出跨國群體共同顯著富集的臨床致病通路與關鍵調控基因，再透過 ElasticNet 篩選的 CpG 分析出關鍵 CpG，即本研究主要探討的生物標記，完成從底層表觀基因組學到高層次分子致病機制的轉譯醫學推導。

(六) 機器學習與跨國泛化能力驗證

根據篩選的核心一致性 CpG 特徵作為輸入，建構邏輯迴歸與隨機森林預測模型。評估階段採取「跨世代一致性分析 (Cross-Cohort Concordance Analysis)」機制：模型僅在單一國家世代數據進行訓練，並直接對另一國的獨立世代數據進行效能評估。

五、工具說明

此專題主要使用 Python 完成機器學習，R 語言做資料預處理，並使用 Google Colab 為機器學習運算平台。

六、實際結果與結論

本專案成功建構跨國甲基化機器學習分析管線。透過「樣本內秩歸一化」有效克服批次效應，並結合 ElasticNet 與富集分析，從數萬個探針中精煉出 31 個具高度跨國一致性的核心致病 CpG 位點。於分析中，模型展現極高的預測準確度(AUC 皆> 0.75)。此外，Boxplot 機率分佈分析顯示，健康者與 CKD 患者的預測得分呈現兩極化，排除隨機假陽性風險，證實此組通用生物標記具備高純淨度與強大的臨床預測效力。可惜泛化能力並無法取得第三個國家的資料具體驗證，須待新資料集產生後再加入研究。

七、參考文獻

- [1] 衛生福利部中央健康保險署, "2024 年國人全民健康保險就醫疾病資訊,"
- [2] T. Marumo *et al.*, "Aberrant proximal tubule DNA methylation underlies phenotypic changes related to kidney dysfunction in patients with diabetes," *American Journal of Physiology-Renal Physiology*, vol.327, no.3, pp.F315-F328, Sep.2024. doi:10.1152/ajprenal.00124.2024.
- [3] A. Bansal *et al.*, "Integrative Omics Analyses Reveal Epigenetic Memory in Diabetic Renal Cells Regulating Genes Associated With Kidney Dysfunction," *Diabetes*, vol. 69, no. 11, pp. 2490-2502, Aug. 2020. doi: 10.2337/db20-0382.