慢性腎臟病患者之 DNA 甲基化生物標記分析

專題編號：114-1-CSIE-S006
執行期限：113年第1學期至114年第1學期
指導教授：白敦文
專題參與人員： 111590021 周柏諺
111590023 林浚偉
111590051 林聖傑
111590053 歐安崙

一、摘要

本研究專題將透過 GEO 公開資料集取得 慢性腎臟病的甲基化表現數據並透過 ChAMP 套件進行甲基化數據正規化，藉由 Lasso 回歸 演算法進行基因的特徵篩選對潛在病患的患病進行風險預測，綜合受測者的共病資料分 析與篩選出的基因交叉比較分析，正確選取在不同的共病狀態都能具有穩定表現且較為顯著甲基化表現變化的基因，最後再透過與生物學家經常使用的基因本體論(Gene Ontology)與 Hierarchical Clustering 算法對基因功能分群，對每群分別挑選具有高差異表現的基因進行相關論文的驗證。

關鍵詞:**Chronic Kidney Disease**、**DNA Methylation**、**Lasso**、**Comorbidity**、**Gene Ontology**

二、緣由與目的

2023年台灣全民健康保險醫療費用前二十大疾病中，急性腎衰竭及慢性腎臟疾病為健保支出的第一名疾病，緊隨其後的第三及 第四名的糖尿病及高血壓性疾病同時也是罹患慢性腎臟病的高危險族群，為避免健保破產，需減少慢性疾病支出，確保福利制度永續[1]。

本計畫將使用國際開放之基因體資料集 GEO，依資料集之不同疾病註解及不同個人年齡、性別及種族等特徵進行完整甲基化基因之資料分群，並逐步辨識並標註不同疾病特徵患者的甲基化數據，進行挑選最具穩定表現且顯著功能的基因群。之後再透過取得對應甲基化位點的資料做正規化以及去離群值等動作，接續預期使用不同機器學習技術將個別基因群組的代表基因建立具高準確性的預測模型。

綜上所論，本計畫會先由現有數據探索 DNA 甲基化實驗的最佳基因，使用該基因進行資料前處理，最後再使用不同的機器學習方式進行 DNA 甲基化特徵的差異性分析，進行不同共病疾病資料的預測模型建構，本研究將主要使用慢性腎臟病的資料進行個人化的疾病預測模型建構。

三、架構流程

(1)資料來源

本研究使用來自 GEO 資料庫提供之糖尿病腎臟病甲基化數據，包含180個糖尿病腎臟病實驗樣本，可分為87個無腎臟病無惡化組(尿液白蛋白較少、eGFR 斜率較平緩)及93個腎臟病惡化組(尿液白蛋白較多、eGFR 斜率較陡)[2]。

(2)資料前處理

ChAMP 套件處理 EPIC 晶片數據，過濾低品質或冗餘 CpG 位點。BMIQ 正規化校正探針誤差，生成標準化 β 值矩陣。四分位距法檢測並移除每個 CpG 位點的離群值，降低雜訊，確保數據集中且具代表性，為差異分析奠定基礎。

(3)共病交集分析

共病是指在現有特定的一個疾病同時患有兩種或更多種以上的疾病，而慢性腎臟病目前常見的共病有高血壓及糖尿病。

慢性腎臟病共病疾病基因與資料前處 理後產生的基因列表進行交集分析以確認後續的候選基因群與慢性腎臟病具有關聯性。

(4)甲基化風險分數建立

計算控制組各 CpG 位點均值，實驗

組β值減去均值得差異值(Δβ)，再次去除離群值後，計算差異平均值(Δβ)，依|Δβ|排名，選出候選位點作為MRS(Methylation Risk Score)基礎，之後透過 Lasso 回歸分析進行 基因的特徵篩選以及決定甲基化位點的內部權重，MRS 的算式如(1)所示，符號 w 及 β 分別代表內部權重與甲基化程度，MRS 分 數愈高可能代表患者的腎病惡化風險愈高。

$$MRS = \sum_{i=1}^{n} w_i \times \beta_i \qquad (1)$$

(5)功能分群

為確認挑選的基因功能之間是否存在相關性，我們使用 Gene Ontology(GO)輔助我們進行基因之間的相似計算，GO是目前對基因功能描述有著廣泛使用的資料庫，針對每個基因創立有證據支持的 GO 標註項目來描述基因的生物學作用。GO 標註是以三個大分類進行定義，分別是分子功能 (Molecular Function, MF)、生物過程(Biological Process, BP)與細胞組成(Cellular Component, CC)，每個大分類之下又會存在更具體的標註(GO term)，我們使用的 GOSemSim 套件會對基因之間的 GO 標註進行相似計算產生相似矩陣，再將相似矩陣轉換為距離矩陣並透過 Hierarchical Clustering 演算法將基因分成三個群組。

(6)機器學習應用於甲基化預測模型建構

機器學習模型利用 MRS 位點的甲基化 數據與功能分群資訊，構建預測模型。決策樹或隨機森林適合處理高維甲基化數據，預測疾病進展風險。通過交叉驗證評估模型準確性與穩健性，驗證候選位點的生物標記潛力，為臨床診斷提供支持。

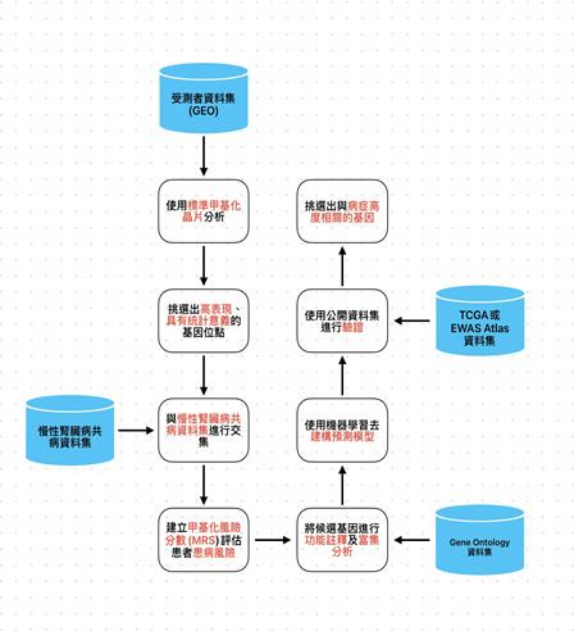

圖 1.研究流程圖

四、 預期成果

本專題計畫旨在增進慢性腎臟病檢測，結合血液 cell-free DNA 甲基化分析，提供精準生物標記，輔助早期診斷、控制病情惡化、改善 eGFR 及尿液白蛋白檢測的偽陰/偽陽問題。透過 GEO、TCGA 及 EWAS Atlas 資料庫更新預測模型，確保長期穩定性。選用穩定且可測的過甲基化位點(hypermethylation)，開發標靶藥物減緩洗腎/換腎風險。最終建立個人化風險評估，改善患者預後與生活品質。

五、 參考文獻

[1] 衛生福利部中央健康保險署(2023)。2023年國人全民健康保險就醫疾病資訊

[2]Hye Youn Sung, Sangjun Lee, Miyeun Han, Woo Ju An, Hyunjin Ryu, Eunjeong Kang, Yong Seek Park, Seung Eun Lee, Curie Ahn, Kook-Hwan Oh, Sue K. Park & Jung-Hyuck Ahn.“Epigenome-wide association study of diabetic chronic kidney disease progression in the Korean population: the KNOW-CKD study”. Published: 20 May 2023. doi:10.1038/s41598-023-35485-x