結腸直腸癌之 DNA 甲基化生物標記分析

專題編號：114-1-CSIE-S007
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：111590009　陳世昂
111590013　呂念庭
111590054　謝永宏
111590056　王逸婕

一、摘要

本研究以結腸直腸癌（Colorectal Cancer, CRC)之 DNA 甲基化資料為研究主題，旨在探討不同組織間的甲基化差異與潛在生物標記（biomarker)。研究使用 GEO 公開資料集 GSE199057（Illumina EPIC 850K）及 GSE193535（450K)，代表不同族群與晶片平台樣本。透過 R 套件 ChAMP 進行品質控制（QC)、BMIQ 正規化、差異甲基化探針分析（DMP）及差異甲基化區域分析（DMR），並以 Python 繪製火山圖呈現甲基化變化趨勢。未來將進一步應用機器學習（Machine Learning）模型驗證這些候選基因區域的分類與預測能力。

關鍵詞：DNA 甲基化、結腸直腸癌、生物標記、ChAMP、DMP、DMR

二、緣由與目的

根據世界衛生組織統計，大腸癌（Colorectal Cancer, CRC）為全球第三大常見癌症，其死亡率高且早期症狀不明顯[1]。目前臨床上主要透過糞便潛血檢查與大腸鏡檢測進行篩檢，但存在侵入性高與準確度受限等問題，因此尋找更穩定且非侵入式的分子標記（Biomarker）成為重要研究方向。

DNA 甲基化是表觀遺傳調控的一種形式，能影響基因表現與腫瘤形成，其異常甲基化已被證實與多種癌症發生有關[2]。過去研究顯示，大腸癌組織中部分基因（如 SFRP1、SEPT9）出現高甲基化現象，並被認為可作為血液或糞便檢測中的非侵入式早期診斷標記 [3]。然而，不同族群及晶片平台的甲基化模式存在差異，使得生物標記的跨資料集穩定性仍待驗證。

本研究以 GEO 資料庫中兩組資料為研究對象：GSE199057（Illumina EPIC 850K，非裔美國人及白人樣本）與 GSE193535（450K，東南亞族群樣本），透過 R 套件 ChAMP 進行資料正規化（BMIQ)、差異甲基化探針（DMP）與差異甲基化區域（DMR）分析，並以 Python 進行可視化與統計檢定。

三、研究報告內容

本研究以 GEO 公開資料庫中兩組結腸直腸癌（Colorectal Cancer, CRC)DNA 甲基化資料為研究對象，分別為 GSE199057（Illumina MethylationEPIC 850K）與 GSE193535（Illumina 450K)。前者樣本分為 Healthy(72)、Adjacent(80)及 Tumor(77)，後者為成對的 Tumor(54)與 Adjacent（54)。研究目標在於建立標準化之甲基化分析流程，並比較不同組別間的甲基化變化以篩選潛在生物標記（biomarkers)。

分析流程主要以 R 套件 ChAMP 為核心，輔以 Python 進行後處理與可視化，步驟如下：

1.資料前處理與品質控制（QC）

以 champ.load() 讀取 .idat 檔後，透過 champ.QC() 產生品質控制報表，包括 β 值分布密度圖、箱形圖與主成分分析（PCA)。偵測率（detection P-value）大於 0.01 之探針被排除，並移除低品質樣本以

確保後續分析穩定性。

2.正規化與資料清理（Normalization）
採用 BMIQ 方法修正 Infinium I/II 探針設計差異，產生全樣本的 β 值矩陣（all_beta_normalized）。經統計檢查後，樣本間分布均勻，無明顯批次效應。

3.差異甲基化探針分析（DMP Analysis）
以 champ.DMP() 比較組間甲基化差異，並輸出 p 值與平均 β 值。隨後將結果匯入 Python，進行離群值處理與群組平均計算，求得 Δβ（delta beta），以量化甲基化變化幅度。火山圖（Volcano Plot）則以 Δβ 為橫軸、−log10(adj.P.Val) 為縱軸，用於呈現顯著探針分布。

4.差異甲基化區域分析（DMR Analysis）
為捕捉區域性甲基化變化，使用 champ.DMR() 搭配 Bumphunter 演算法分析相鄰探針群。分析分別對 Healthy vs Tumor（HvT）與 Adjacent vs Tumor（AvT）兩組別執行，產生 DMR_HvT 與 DMR_AvT 結果。後續以 Python 繪製 DMR 火山圖，觀察整體甲基化趨勢。

5.可視化與報表產出
使用 Matplotlib、Pandas 與 NumPy 套件繪製火山圖及 QC 視覺化圖（密度圖、箱形圖、PCA），並輸出統一報表以便後續比較。

目前已完成 GSE199057 的完整分析流程建立，包含 QC、Normalization、DMP、DMR 與初步可視化。後續將以 GSE193535 資料集進行交叉驗證，並嘗試利用機器學習模型（如 XGBoost 或 Random Forest）進行特徵選取與預測分析。

預期成果為找出在 Tumor、Adjacent、Healthy 三組樣本中皆具有穩定甲基化差異的 CpG 位點或區域，作為潛在的結腸直腸癌早期生物標記，並驗證其跨人種與跨晶片平台的重現性。

參考文獻

[1] H. Sung, J. Ferlay, R. Siegel, et al., “Global Cancer Statistics 2020,” *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] P. A. Jones, “Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and Beyond,” *Nature Reviews Genetics*, vol. 13, pp. 484–492, 2012.

[3] W. N. Kisiel, et al., “DNA methylation biomarkers in stool and blood: a systematic review,”
Clinical Epigenetics, vol. 6, no. 1, pp. 1–13, 2015. (PMCID: PMC4286876)