肺癌患者特定基因突變之甲基化生物標記分析

專題編號：111-CSIE-S014
執行期限：110 年第 1 學期至 111 年第 1 學期
指導教授：白敦文
專題參與人員： 108820002 陳佳吟
108820022 黃士儼
108820035 林煥智

一、摘要

本研究針對不同肺癌患者族群進行全基因體 DNA 甲基化圖譜分析，使用統計模型及機器學習技術探索特定族群在特定基因(EGFR、KRAS)突變的組合，與肺癌相關的甲基化生物標記，透過甲基化生物標記的表現推斷受測者罹患肺癌的可能性與疾病預後進展，可提早啟動相關的預防機制及最佳治療建議。

關鍵詞：EGFR、KRAS、生物標記、甲基化

二、緣由與目的

世界衛生組織指出，2020 年肺癌死亡率是所有癌症中第一名[1]。在台灣肺癌死亡率也高達 19.31%，和全球肺癌死亡率同樣位居第一[2]，其居高不下的死亡率足見肺癌已成為全球必須正視的嚴重問題。

腫瘤形成早期會發生 DNA 甲基化，其對於癌症的基因表達十分重要，肺癌一般而言要有效精準預測較不易，因此需要更深入地了解肺癌腫瘤的基因表現和其前兆。目前有研究指出在 EGFR 和 KRAS 基因突變的腫瘤中有觀察到特定的甲基化圖譜，其甲基化程度與基因表現有呈現正相關及負相關[3][4]，以致影響肺癌患病率，除了在檢驗初期當作判斷是否具有肺癌的條件外，治療的過程及預後也都是很重要的指標。不同的人種，在這兩個基因的表現有明顯的差異，例如 EGFR 基因的突變比率在白種人(18.7%)和黃種人(53%)之間差異極大 [5]。

三、研究流程

使用 GEO 數據庫(Gene Expression Omnibus, GEO)所提供的肺癌資料集進行分析，包含 19 個正常細胞樣本、142 個癌症細胞樣本以及各樣本 EGFR 與 KRAS 基因的表現型。

將癌症樣本依照 EGFR 及 KRAS 基因的表現型(突變及未突變)進行分組(共 4 組)，每一組都分別加入相同未罹患癌症的細胞樣本，做為控制組。

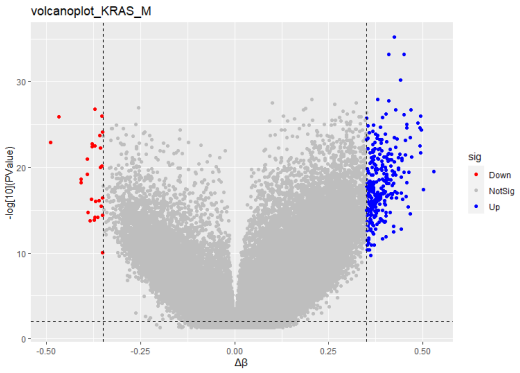
採用 ChAMP (Chip Analysis Methylation Pipeline) [6]套件進行甲基化資料分析，去除不適合進行分析的位點，再使用 BMIQ 正規化 (beta-mixture quantile normalization) [7]去除因探針設計產生的誤差。針對每個位點去除離群值(outlier)，讓同一位點的 β 值分布能夠更為集中。接著計算每一個位點控制組與實驗組的 β 值差異數值分析。

透過分析各群組之候選生物標記間之交集與差集，篩選出適合各人種檢驗肺癌之候選生物標記。

使用 SVM 技術，將具差異性的甲基化生物標記組合進行 SVM 模型建構，探索可以準確檢測肺癌的最佳基因組合及預測模型，檢視所有生物標記組合的預測正確率，以獲得最佳生物標記組合。

四、研究結果

圖(一)為 KRAS 突變組差異性甲基化位點分析的火山圖分佈(總共四組，僅舉一組結果為例)，橫軸為癌症細胞樣本與正常細胞樣本的甲基化 β 值差值，Δβ 為正值代表該基因位點是過甲基化的表現，負值則代表低甲基化的現象，|Δβ| 越大表示該位點平均差值較大，較具有分辨能力；縱軸則為 p 值，p 值可用來參考該位點統計的顯著性，數值越小顯著性越高，然而圖內 p 值為取對數後加上負號，數值越大表示原數值越小，也意味著顯著性越高。位於火山圖的左上及右上角的位點即為較佳的候選生物標記。



圖(一)KRAS 突變(145個候選位點)

將前步驟所分析出的各組候選生物標記依基因分類進行交集分析，最後會得到 50 個位點。此外，甲基化若發生於該基因啟動子(promoter)的區域，該基因的轉錄將會被直接抑制，若是該基因的原是功能具有抑制腫瘤生長的作用，而它的啟動子區域發生甲基化變異，將失去抑制腫瘤生長的作用而造成腫瘤細胞的增生，導致癌症形成[8]。因此本研究從 50 個位點中，進一步挑選 12 個位於啟動子區域，且 Δβ 為正數的過甲基化基因位點，這 12 個位點即為本研究最終篩選的甲基化生物標記。

為確認前步驟挑選的 12 個生物標記真正具有高度適用性，本研究使用 SVM 技術隨機挑選 3 個生物標記的組合進行甲基化生物標記的預測表現驗證，如此可以避免進使用單一個位點因其他疾病導致甲基化 β 值產生變化而造成預測錯誤的結果。驗證過程使用五倍交叉驗證(5-fold cross validation)及使用不同參數設定以提升預測之準確度。表二僅顯示使用 linear 為 kernel、參數 C 設為 100 所訓練的前五名的最佳表現的組合甲基化基因。

Gene1_ID	Gene2_ID	Gene3_ID
HOXD10	FOXD2	SCT
LOC149134	HOXD4	SCT
ZNF177	SERPING1	MARCHF11
HOXD4	MARCHF11	SCT
HOXD4	HOXD10	FOXD2

表(二)三位點精準度排序(前五名)

五、結論

挑選生物標記時，若直接將所有樣本一起進行差異性甲基化位點分析，以 |Δβ|>0.35 為例，可以得到 162 個具顯著差異性的位點。由於已有研究指出 EGFR 與 KRAS 突變與否與罹患肺癌的機率有關[5]，因此本研究針對此二基因的表現型將樣本分成四組再分別進行差異性甲基化位點分析，將各組結果交集後，能夠將原本 162 個候選位點縮減至 50 個位點。此作法目的為探索不論 EGFR 與 KRAS 是否突變都具有顯著差異性的位點，因此相較於前者，使用此作法取得的位點較具通用性。

六、參考文獻

[1] 2020 年最新全球癌症大數據報告 https://www.snq.org.tw/chinese/11_expertcolumn/detail.php?SID=552

[2] 癌症登記報告 https://www.hpa.gov.tw/Pages/List.aspx?nodeid=119

[3] Moore, L., Le, T. & Fan, G., "DNA Methylation and Its Basic Function" Neuropsychopharmacol, vol. 38, pp. 23-38, 2013.

[4] M. Gardiner-Garden, M. Frommer, "CpG Islands in vertebrate genomes" Journal of Molecular Biology, vol. 196, Issue 2, pp. 261-282, 1987.

[5] 張宗德，「肺腺癌驅動基因研究相關進展」，Pubmed，Chinese.doi: 10.3779/j.issn.1009-3419.2013.02.06, PMID: 23425901

[6] Wang, Z., Wu, X. & Wang, Y, "A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip" BMC Bioinformatics, 19, Article number: 115, 2018.

[7] Morris, T. J., Butcher, L. M., Feber, A., Teschendorff, A. E., Chakravarthy, A. R., Wojdacz, T. K., and Beck, S, "Champ: 450k chip analysis methylation pipeline" Bioinformatics, 30(3), pp. 428-430, 2014.

[8] Shahjehan A. Wajed, Peter W. Laird, Tom R. DeMeester, "DNA Methylation: An Alternative Pathway to Cancer, PMCID:PMC1421942"