輔仁大學統計資訊學系

第二十三屆專題研究成果報告

指導老師：游鎮瑋 博士

利用線上資料庫
即時分析肺癌治療的潛在因子

學生：陳婕昕、袁晨諭、陳子毓、林瑀倢、
楊洹綺、廖彥甄、王乙竹 撰

中華民國 112 年 11 月

摘　要

題目：利用線上資料庫即時分析肺癌治療的潛在因子

校（院）系所組別：輔仁大學管理學院統計資訊學系

學生： 陳婕昕、袁晨諭、陳子毓、林瑀倢、楊洹綺、廖彥甄、王乙竹

指導教授：游鎮瑋 博士

論文頁數：72

關鍵詞：肺癌、肺腺癌、肺鱗癌、免疫基因

論文摘要內容：

根據衛生福利部統計，肺癌為台灣癌症十大死因之首，於 111 年約有超過萬人死於肺癌，肺癌致死率極高的原因主要是發現的時間過晚，造成治療的效果有限。肺癌中肺腺癌是最為常見的肺癌，患者多半沒有任何自覺症狀；常見於老年吸菸男性的是肺鱗癌，其生存機率較高。本研究使用 Xena 線上即時資料庫蒐集資料，主要針對總共 1304 筆肺腺癌、肺鱗癌患者與正常人的免疫細胞基因進行敘述性統計、U-test、Spearman 等級相關、生存分析、分群及主成分分析，主要探討 13 個免疫細胞基因在肺腺癌、肺鱗癌患者與正常人是否有所不同，並希望能夠找出對於提早診斷與治療肺癌有益的基因。

Abstract

Title of Thesis：Real-time analysis of potential factors of lung cancer treatment using online database

Name of Department: Department of Statistics and Information Science, College of Management, Fu Jen Catholic University

Names of Students: Chieh-Hsin, Chen , Chen-Yu, Yuan , Zih-Yu, Chen , Yu-Jie,Lin , Yuan-Qi,Yang , Yan-Jhen, Liao , Yi-Chu, Wang

Advisor: Chen-Wei Yu, MT, PhD

Total Pages:72

Key Words: Lung cancer, lung adenocarcinoma, lung squamous cell carcinoma, immune genes

Abstract:

According to statistics from the Ministry of Health and Welfare, lung cancer is a leading cause of cancer mortality in Taiwan, resulting in over 10,000 deaths in 2021. The primary factor contributing to the high mortality rate associated with lung cancer is the late-stage diagnosis, resulting in limited treatment efficacy. Lung adenocarcinoma stands as the most prevalent form of lung cancer, with a noteworthy proportion of patients exhibiting no overt symptoms. Squamous cell carcinoma of the lung predominantly affects elderly male individuals who smoke, but it tends to offer a higher probability of survival. In this study, we harnessed the capabilities of the Xena online real-time database to gather data. We conducted comprehensive statistical analyses, including narrative statistics, U-tests, Spearman rank correlation, survival analysis, clustering, and principal component analysis. Our focus was exploring differences in the expression signatures of several immune genes among 1304 patients diagnosed with lung adenocarcinoma, lung squamous cell carcinoma, and healthy individuals. Our primary objective

was to discern whether 13 specific immune genes exhibited differential expression between patients with lung adenocarcinoma, those with lung squamous cell carcinoma, and individuals without cancer. We aim to identify genes that may hold promise for early lung cancer diagnosis.

謝辭

專題製作的過程充滿了許多挑戰，由衷感謝在這個過程中給予我們幫助的教授，同時也要感謝參與其中的自己，持續努力並且沒有放棄。

由於整個研究主要涵蓋了醫學領域，在這邊想要感謝我們的專題指導老師－游鎮瑋教授，不僅經常提供我們相關資料和報告，讓我們可以順利的從頭了解癌症、免疫細胞等背景，還指導我們如何在分析過程中運用這些數據，以及如何以更專業的方式解釋研究結果，這些知識和技能是我們在過程中學到的，即使沒有相關背景，在當今科技高度發展的情況下，我們仍然可以透過自學和運用我們所學到的知識來進行研究。

在專題過程中，團隊成員之間的相互協助是無比重要的，真誠地感謝我們每一位成員，大家都毫不吝嗇地提供幫助，沒有人將自己置於優先，使成員間的溝通和分工大多順暢無阻。這一年來，大家都辛苦了！

全體組員謹致於

輔仁大學統計資訊學系

中華民國 112 年 11 月

目錄

表目錄

圖目錄

第壹章 諸論

第一節 研究動機

人的一生短暫而脆弱，充滿了無數的挑戰和不確定性，其中，生老病死是每個人都必須面對的四個不可避免的階段。而當談到疾病，特別是像癌症這樣具有重大威脅者時，我們往往感到最脆弱和無助。癌症不僅對患者的生活質量產生深遠影響，還可能導致生命的突然終結。根據我國衛生福利部於 111 年公告的統計數據（圖 1-3- 2）顯示，肺癌為我國癌症死亡率之首，其重要性更是不言而喻，導致罹患肺癌的成因不勝枚舉，不論是生活習慣亦或是基因病變，都有可能增加罹癌風險。因此，本研究將深入探討免疫基因在肺癌表達上的差異，並分為四大部分進行深入研究。

第二節 研究問題與目的

肺癌的致死率一直居高不下，其主要原因在於多數患者在確診肺癌時幾乎都已是晚期，導致其治療的效果十分有限，如何盡早發現以提供更有效的治療一直是醫學界首當其衝的議題。因此，本研究將利用美國政府研發與推出的 TCGA（The Cancer Genome Atlas）與加州大學線上資料庫，藉由其統計數據進行以下研究：一、Mann－Whitney U test ：判斷不同基因分別於肺鱗癌、肺腺癌間是否有統計上的差異；二、迴歸分析：判斷不同基因在表達上是否有統計上的相關性；三、生存曲線：判斷不同基因在表達上的高低是否對生存率有統計上的顯著影響；四、PCA 分群統計：建立分群模型，觀測不同基因的表達所帶來的罹癌風險，探討及分析肺癌好發的潛在因子，藉以期望達到觀測可能罹患肺癌的高危險族群，期望達到提早偵測，提早治療的效果，以提高存活率。

第三節 研究背景介紹

肺是許多陸生動物和水生、半水生動物於呼吸系統中最關鍵的一個器官，其作用為透過吸入空氣進行氣體交換，獲得細胞呼吸必須的氧氣。然而這樣一個人類賴以維生的器官，卻免不了遭受疾病的攻擊，也就是大家常聽到的肺癌。指的是肺部內生長失去控制的，分化後的惡性細胞可能會造成，如侵入相鄰的組織和滲透到肺部以外等，而絕大多數肺癌是肺部，由病變所造成。根據衛生福利部於 6 月 12 日公布 111 年國人十大死因（圖 1-3- 1）之統計結果，癌症已連續 41 年位居國人十大死因之榜首， 其中不論男女，肺癌都是我國癌症死亡率第一名（圖 1-3- 2）。

111年我國十大死因

死因別	死亡人數 (人)	年增率 (%)	死亡率 (每十萬人口)	年增率 (%)	標準化死亡率 (每十萬人口)	年增率 (%)
所有死亡原因	208,438	13.2	893.8	13.9	443.9	9.5
1.癌症	51,927	0.5	222.7	1.2	116.0	-1.8
2.心臟疾病(高血壓性疾病除外)	23,668	8.3	101.5	9.0	47.8	4.8
3.嚴重特殊傳染性肺炎(COVID-19)	14,667	1,536.9	62.9	1,547.4	28.6	1,365.5
4.肺炎	14,320	5.7	61.4	6.4	26.3	3.8
5.腦血管疾病	12,416	1.9	53.2	2.6	25.1	-0.3
6.糖尿病	12,289	7.3	52.7	8.0	24.7	3.7
7.高血壓性疾病	8,720	10.6	37.4	11.3	16.3	7.2
8.事故傷害	6,953	2.6	29.8	3.3	20.0	0.2
9.慢性下呼吸道疾病	6,494	4.1	27.8	4.8	11.8	1.6
10.腎炎、腎病症候群及腎病變	5,813	6.3	24.9	6.9	11.3	3.5

圖 1-3- 1 111 年我國十大死因

根據衛生福利部統計結果顯示，癌症為我國十大死因之首

111年十大癌症死因

癌症別	死亡人數 (人)	年增率 (%)	死亡率 (每十萬人口)	年增率 (%)	標準化死亡率 (每十萬人口)	年增率 (%)
所有癌症死亡原因	51,927	0.5	222.7	1.2	116.0	-1.8
1.氣管、支氣管和肺癌	10,053	0.1	43.1	0.8	21.8	-1.9
2.肝和肝內膽管癌	7,781	-2.4	33.4	-1.8	17.0	-5.2
3.結腸、直腸和肛門癌	6,853	2.9	29.4	3.6	14.7	0.9
4.女性乳癌	2,834	-2.7	24.1	-2.2	13.1	-5.6
5.前列腺(攝護腺)癌	1,830	8.3	15.9	9.2	8.0	7.2
6.口腔癌	3,479	2.5	14.9	3.1	8.5	0.2
7.胰臟癌	2,769	4.1	11.9	4.8	6.1	1.4
8.胃癌	2,277	-1.4	9.8	-0.8	4.9	-3.2
9.食道癌	1,980	-2.5	8.5	-1.8	4.8	-5.0
10.卵巢癌	765	9.9	6.5	10.5	3.7	9.5

圖 1-3- 2 111 年十大癌症死因

根據衛生福利部的統計顯示，氣管、支氣管和肺癌為我國十大癌症死因之首

第四節 肺癌高危險族群簡介

細胞的癌變與 DNA 及表觀遺傳等遺傳信息的突變有關，這些變化會影響細胞的正常功能，包括細胞增殖、程序性細胞死亡（細胞凋亡）和 DNA 修復。損傷累積的越多，癌症發生的風險就越高。常見的肺癌高危險族群大致可分為以下幾點：

一吸菸

吸菸是目前為止導致肺癌的主要原因，其又可細分為常吸菸者、接觸二手菸者以及近年來發現也有可能導致罹癌的三手菸接觸者。香菸的菸霧中含有至少 73 種已知的致癌物質，包括苯並芘、NNK、1,3-丁二烯，以及釙-210 等放射性物質。肺癌是世界上癌症死亡的主要原因，2008 年，超過 160 萬人診斷出罹患肺癌，占所有癌症的 13 %，140 萬人死於肺癌，占所有癌症死亡人數的 18 %（Barta, Powell et al. 2019）。

被動吸菸，又稱吸二手菸，意為吸入其他吸菸者產生的菸草煙霧，也是導致不吸菸的人患上肺癌的原因。美國（Control and Prevention 2001）、歐洲（Jaakkola and Jaakkola 2006），和英國（Parkin 2011）的研究都證實被動吸菸者罹患肺癌風險在顯著增加，而研究亦顯示二手菸比直接吸菸更危險。美國每年有大約 3,000 人死於被動吸菸造成的肺癌（Barta）。

二肺癌家族病史

有肺癌家族病史者，研究發現，肺癌的發生與家族傾向有潛在關係，即肺癌病患的父母、子女、兄弟姊妹，得到肺癌的風險比沒有家族史的高，可能與遺傳因素或是暴露在相同致癌環境中有關。

三基因

女性雖吸菸人口較少，更有高達九成肺腺癌女性病人無吸菸習慣。近年來台大校長楊泮池院士研究發現，華人肺腺癌與高風險遺傳

因子 YAP1 變異基因密切相關，若帶有基因異常者，罹患肺腺癌風險為基因正常者的 5.9 倍，尤其易發生在女性中（臺大醫院健康電子報, 2015 "華人肺腺癌高風險遺傳基因異常"）。

四 職業

常接觸油煙的主婦或廚房工作者；職業為石棉、鈹、鈾與氡氣工人；長期暴露在某些重金屬物質，如鉻、鎘、砷等；柴油引擎廢氣；接觸化學物質的工作環境，如金屬業、冶礦業、石棉業等或暴露在放射線環境下工作等等，有較高罹患肺癌的危險性（Tobias and Hochhauser ,2009）。

五 空汙與油煙

戶外空氣汙染物會增加罹患肺癌的風險，特別是燃燒化石燃料產生的廢氣。空氣污染的影響已有國際實證對人類具致癌性，其主要組成成分「懸浮微粒（PM）」也被個別評估一併列為第一級致癌物，長期暴露在室外空污將增加肺癌的風險。懸浮微粒來源包括自然界，如火山爆發、地殼岩石崩解；及人類行為產出，如石化燃料工業與汽機車排放、移動源廢氣等燃燒行為。據估計，二氧化氮濃度每增加 10 ppb，人患肺癌風險就上升 14 %（Clapp, Jacobs et al. ,2008），1-2 %的肺癌由戶外空氣污染所致（Murray and Mason ,2016）。其他流行病學研究發現女性肺腺癌可能與烹調時，尤其是油炸、熱炒所產生的油煙有關，初步證據支持室內空氣污染會使患肺癌風險增加，包括做飯和取暖時燃燒木材、木炭、糞便或作物殘茬。暴露於室內煤煙的女性，罹患肺癌風險會提升大概一倍，許多生物質燃燒後的副產品是已知或可疑的致癌物，這一風險被證實影響全球大約 24 億人（LIM and Seow ,2012）。

六 肺部病變病史

曾得過肺結核或其他肺部慢性疾病者，如氣喘、慢性阻塞性肺病、肺纖維化、肺結核等肺部疾病，會使罹患肺癌的風險增加（衛生福利部國民健康署, 2019）。

第五節 台灣肺癌類型與五年存活率

肺癌主要分成兩大類型：非小細胞肺癌和小細胞肺癌（圖 1-5-1）。這兩種類型的癌細胞生長、分化及擴散速度都不相同，對於臨床治療的方式和化學治療的反應也有極大差別。這個區別對採取不同治療手段有非常重要意義。非小細胞肺癌是通過肺癌手術；而小細胞肺癌常常對化療和放療的反應比較好。其生存率也常隨者不同種類而確診類型而不同（圖 1-5-2），生存率最高的為肺鱗癌（22%），最低的則是小細胞癌（8%）。

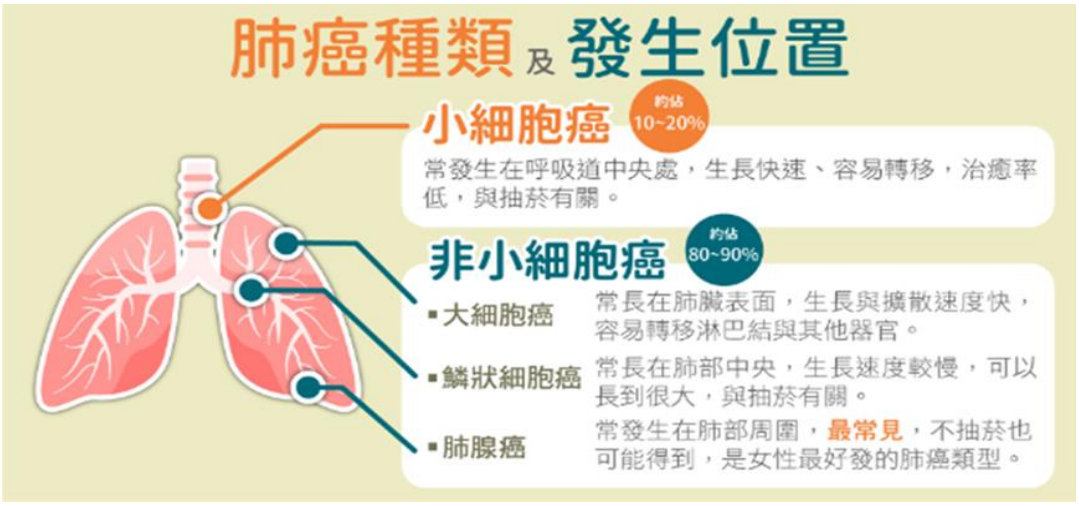

圖 1-5- 1 肺癌種類及發生位置

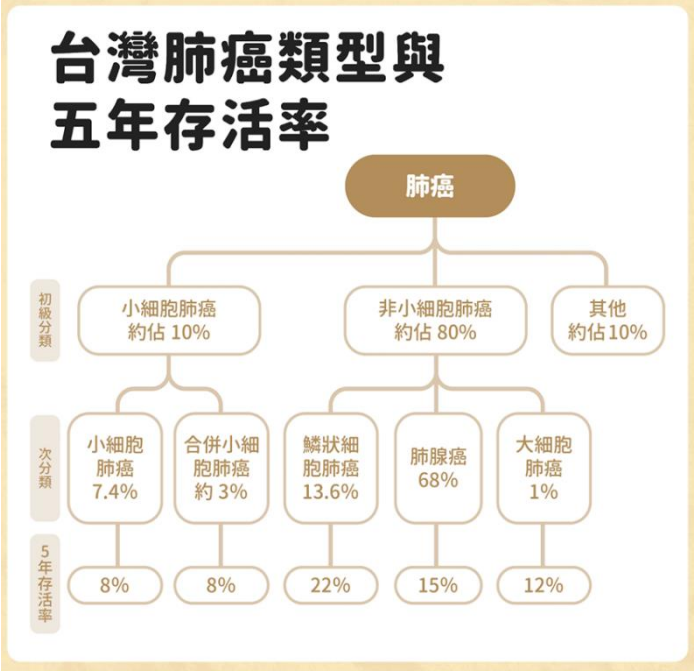

圖 1-5- 2 台灣肺癌類型與五年存活率

一 非小細胞肺癌（Non-small-cell lung cancer；NSCLC）

在台灣約佔全部肺癌百分之八十五，和小細胞肺癌比較起來，非小細胞肺癌的生長速度較慢，轉移發生也較慢，但是只有少數人在診斷出來時是屬於可以開刀治療的。其依據細胞型態可分成主要三種：肺鱗癌（Squamous-cell Carcinoma）、肺腺癌（Adenocarcinoma）、大細胞癌（Large-cell lung carcinoma）。

肺鱗癌，又稱鱗狀上皮細胞癌，常見於老年吸菸男性，腫瘤常長在肺部中央靠肺門位置，容易堵塞氣管造成肺葉萎陷，且有縱膈腔淋巴結擴散，擴散速度比其他類型肺癌慢。手術切除機會較多，5 年內生存率較高。其初期以手術治療為主，醫師會依腫瘤生長情況差異，替患者進行節狀切除術、楔狀切除術、肺葉切除術或肺切除術等，以有效切除癌細胞。若病患不適合接受手術或癌細胞已轉移者，醫師則會建議先進行化學治療，透過從靜脈注射或口服抗癌藥物，控制癌細

胞生長或減輕不適症狀，並視情況搭配放射治療，以高能量放射線殺死癌細胞；另一種方法為使用標靶治療，抑制體內腫瘤生長，達到治療鱗狀上皮細胞癌的效果。經治療後，患者 5 年存活率可達 22 %（連珮妤, 2023 "鱗狀上皮細胞肺癌是什麼？存活率、初期症狀、治療與原因."）。

肺腺癌，約佔非小細胞肺癌的 50 %，是肺癌中最常見類型，多為空汙引起，發病年齡較輕，常發生在肺部周圍，發病早期即可侵犯血管和淋巴管並引起肝、腦、骨等遠處轉移，沒有抽菸者所罹患的肺癌多為此類。此類型腫瘤通常長在肺部邊緣屬於周邊型病變，患者多半沒有任何自覺症狀，為女性最好發的肺癌類型（心血管保健諮詢網 "肺腺癌–初期症狀及第四期存活率."）。

大細胞癌，又稱為大細胞未分化癌，為高度惡性上皮腫瘤，常生長於肺臟表面，生長及擴散速度快，易轉移到淋巴結與其他器官。在臨床實踐中，肺大細胞癌常表現為周圍型、巨塊型，無癌性空洞，因癌細胞分化程度差，腫瘤生長迅速，呈侵襲性生長，突破胸膜屏障，累及鄰肺、膈肌、胸壁及縱隔大血管，且淋巴和血行轉移迅速，故病程短，即使手術切除，預後也極差。(醫學網, 2020 "肺大細胞癌：症狀、病因及如何治療.")。

二小細胞肺癌（Small cell lung cancer；SCLC）

小細胞肺癌腫瘤，又稱小細胞未分化癌，惡性程度最高，約占肺癌的 10 %～15 %。小細胞肺癌對放療和化療較敏感，這種類型的肺癌多發生在男性，通常出現在較大的支氣管上，與抽菸關係極為密切。它生長快速，惡性程度最高，約占肺癌 10 % 至 15 %，此類癌症所佔比例較少，且易由淋巴或血液循環系統蔓延至身體其它組織或器官，所以一般無法以手術切除，而對化學治療及放射治療的反應比非小細胞肺癌好。但整體而言，小細胞肺癌病患的預後比非小細胞肺癌病患差。其病灶通常位於肺部中央靠肺門位置，容易往大的支氣管

發展，使氣管堵塞造成肺葉萎陷，診斷時常已有縱膈腔淋巴結的擴散，甚至有三分之二的病人發生遠處轉移，因此後續展望並不樂觀（馬偕紀念醫院 "肺癌可分成那幾類型." 小細胞肺癌）。

第貳章 文獻探討

第一節 免疫基因

一細胞表面蛋白

(一) CD4

它們通常被稱為「CD4 細胞」、「輔助 T 細胞」或「T4 細胞」。它們被稱為「輔助細胞」是因為其中一個主要功能是將訊號送到其他免疫細胞，包括可以殺死感染細胞的 CD8 胞殺細胞。如果沒有 CD4 細胞（如 HIV 感染者、器官移植者），人體將無法對抗大量的病原菌並暴露於危險中（林政道,長庚婦產通訊--第 46 期）。

存在於 T 淋巴細胞（T 細胞）的表面。通過與 MHC-II 分子結合，促使 T 細胞受到活化，從而釋放細胞激素並啟動其他免疫細胞的應答。

當抗原呈現細胞將外來病菌分解，把抗原與 MHC-II 結合後，呈遞給輔助 T 細胞（即與輔助 T 細胞表面的 CD4 受體結合），輔助 T 細胞再接著刺激 B 細胞產生抗體，此即體液性免疫反應的基本過程。

健康人體內「CD4 輔助 T 細胞」的數量在 700-1000／μL 之間甚至以上，而膠質母細胞瘤病人卻只有 200／μL 甚至更少。同時，其他 T 細胞相比正常人也有大幅下降，如此低下的免疫功使他們更易受到各種感染，並可能導致癌症惡化（林以璿,2018）。

針對癌症而言，免疫力定義的好為應用 CD4/CD8（T suppressor）:>1.0；至於宿主可以成功達成長期癌症緩解，必須處於自體免疫（autoimmunity）情況下，宿主藉由細胞激素微環境（cytokinesmicroenvironment）來隨時做免疫微調（immunotuning）（Leahy DJ, Axel R, Hendrickson WA,1992）。

(二) CD68

主要存在於巨噬細胞、樹突細胞和其他單核細胞中，在常規染色基礎上進行免疫組織化學染色，可用來標記組織中的巨噬細胞。在病理學中，它通常用於確定腫瘤組織、發炎組織等免疫反應程度和評估疾病的病理特徵（Holness CL; et al. Blood, 1993）。

CD68 陽性細胞即腫瘤相關巨噬細胞（tumorassociated macrophages，TAMs）、E-cadherin、TGF-β1 可能參與了乳癌的進展過程，檢測其表達為乳癌的臨床病理診斷及治療提供理論依據和指導（中華內分泌外科雜誌, 2019）。

(三) CD86

存在於抗原呈現細胞（如樹突細胞、B 細胞和巨噬細胞等）的表面。在 T 細胞活化過程中的與 T 細胞受體相互作用，增強 TCR（T-cell receptor T 細胞受體）信號的信號，缺乏此信號，T 細胞將進入無反應狀態或免疫耐受，甚至引起細胞程序性死亡（昶安科技,2021）（全球醫藥新知,2018）。

CD80 和 CD86 是與 CD28 相互作用的配體，它們會與 CTLA-4 相互競爭，可刺激 T 細胞存活並阻止由 CTLA-4 引發的抑制 T 細胞功能，CD28、CTLA-4、CD86、CD80 和 VEGF-A 表現量可能對胃癌有益，因此，CD80 和 CD86 也被視為標靶治療的目標之一（Bolandi, Nadia et al.,2021）（ Zhenbiao Wu et al,2017）。

(四) CD3D

T 細胞受體（TCR）複合體的一部分。與其他 T 細胞受體相關蛋白一起組成功能完整的 T 細胞受體，當 T 細胞受體與 MHC-抗原複合物結合時，CD3D 以及其他 TCR 相關蛋白會傳遞訊號，啟動 T 細胞的免疫（Gaglia J、Kissler S,2019）。

(五) CD8A

存在於細胞毒性 T 細胞（又稱為 CD8+ T 細胞）的表面。細胞毒性 T 細胞負責對抗感染細胞、癌細胞和其他異常細胞，CD8A 助於確保細胞毒性 T 細胞對異常細胞進行適當的響應，同時避免對正常細胞造成過度的攻擊。

當病菌入侵人體，被抗原呈現細胞吞噬或潛入正常的體細胞中。其構成抗原的部分必會與細胞內的 MHC 結合而表現在細胞外。當抗原呈現細胞「移動」至最近的淋巴結，淋巴結中其中一種 CD8 序列、「不」具毒殺能力的 T 細胞，對抗原有反應時，該 T 細胞便開始分裂、並成熟為毒殺型 T 細胞。毒殺型 T 細胞在周遊人體時若遇到一個受感染的正常細胞，其表面有如前所述的「抗原與 MHC 的結合體」，則該 T 細胞便會「毒殺」受感染的細胞。

二轉錄因子蛋白

(一) FOXP3

是調節性 T 細胞（Treg 細胞）分化和功能的關鍵調節子。當 T 淋巴細胞被誘導分化為調節性 T 細胞並與受體結合時，會誘導 FOXP3 的表現，以遽增調結 T 細胞抑制發炎反應的作用。*調節性 T 細胞（Treg 細胞）：防止免疫系統對自身組織產生過度的反應，從而避免自身免疫疾病的發生（李聰亮、周宛怡、黃建仁,2016）。

(二) GATA3

GATA3 是一種轉錄因子，其對於各種組織的胚胎發育、發炎、體液免疫反應以及血管內皮的正常功能至關重要。目前的臨床和實驗室研究重點是確定直接或間接阻斷 GATA3 對發炎和過敏性疾病（如氣喘）的作用的益處。其也被認為是各種類型癌症，特別是乳癌的臨床重要標記（Naylor MJ, Ormandy CJ ,2007）。

(三) TBX21

T-box 轉錄因子 TBX21，也稱為 T-bet，現已確定 TBX21 有助於維持黏膜穩態和黏膜免疫反應。缺乏適應性免疫細胞和 TBX21 的小鼠會出現與人類潰瘍性結腸炎相似的疾病，而後將其歸因於革蘭氏陰性細菌（即螺旋桿菌）的生長（Haybar H,2019）。

三細胞膜蛋白

(一) BTN3A1

2020 年 8 月 21 日，美國 H. Lee Moffitt 癌症中心 Jose R. Conejo-Garcia 研究組在 Science 雜誌發表 BTN3A1 governs antitumor response by coordinating $\alpha\beta$ and $\gamma\delta$ T cells 的研究文章，該研究發現：嗜乳脂蛋白 BTN3A1 透過阻止 N-糖基化 CD45 從免疫突觸的分離來抑制腫瘤反應性$\alpha\beta$ T 細胞的激活。值得注意的是，CD277（BTN3A1-3）特異性抗體與 BTN3A 蛋白結合能重塑$\alpha\beta$ T 細胞的抗腫瘤效應，同時恢復$\gamma\delta$ T 淋巴細胞對表達 BTN3A1 癌細胞的細胞毒性作用，最終消除腫瘤的惡性進展。因此，靶向 BTN3A1 的治療手段不僅能協同$\alpha\beta$和$\gamma\delta$ T 細胞來殺傷已確立的腫瘤，同時對抵抗現有免疫療法的腫瘤提出一個全新的潛在治療策略（BioArt 生物藝術,2020）。

(二) MS4A1

也稱為膜表面抗原 CD20，是 B 細胞特有的膜表面抗原，本身不直接參與抗體的合成，而是用於調節 B 細胞激活、分化和增殖，且對於 B 細胞的識別、功能調節以及腫瘤治療中的應用都具有重要意義（患者病理學,2023）。

它存在於 B 細胞淋巴瘤、毛細胞白血病、B 細胞慢性淋巴球白血病和黑色素瘤癌症幹細胞。免疫組織化學可用於確定組織學組織切片中細胞上 CD20 的存在。由於 CD20 仍然存在於大多數 B 細胞腫瘤的

細胞上，而在其他類似的 T 細胞腫瘤中則不存在，因此它對於診斷 B 細胞淋巴瘤和白血病等疾病非常有用。然而，此類腫瘤中是否存在 CD20 與預後無關，兩種情況下疾病的進展大致相同，CD20 陽性細胞有時也見於何杰金氏病、骨髓瘤和胸腺瘤病例。

以淋巴癌為例：因 90% 的 B 淋巴細胞非何杰金氏淋巴瘤，其淋巴瘤細胞上會有 CD20 表面抗原，新近發展出之對抗 CD20 的單株抗體 Rituximab，便是利用其與 CD20 抗原結合，所引起的補体依賴性細胞毒殺作用（CDCC）及抗體依賴性細胞毒殺作用（ADCC）等免疫反應，來達到消滅淋巴瘤細胞的目的（Hardy R ,2008）。

(三) PDCD1

也稱為 PD-1 和 CD279，是 T 細胞和 B 細胞表面的一種蛋白質，在調節免疫系統對人類細胞的反應中發揮作用。PD-L1 是 PD1 的配體，在多種癌症中有較高的表達，因此 PD1 在癌症免疫逃脫中的作用已得到充分證實。針對 PD-1 的單株抗體可增強免疫系統，目前正在開發用於治療癌症。許多腫瘤細胞表達 PD-L1，一種免疫抑制性 PD-1 配體，抑制 PD-1 和 PD-L1 之間的相互作用可以增強體外 T 細胞反應並介導臨床前抗腫瘤活性，稱為免疫檢查點封鎖（Loftus, Peter ,2014）。

(四) CD274

也稱為 PD-L1 或 B7 同源體，是人類體內的一種蛋白質，其在多種惡性腫瘤中展現高表達力，尤其是肺癌。為了預測基因療法或全身免疫療法在阻斷 PD-1 和 PD-L1 檢查點方面的有效性，PD-L1 被用作預後標記和抗癌免疫的目標，例如對腎細胞癌患者的 196 份腫瘤樣本進行分析發現，PD-L1 的高腫瘤表達與腫瘤侵襲性增加，其死亡風險將提高 4.5 倍。

(五) CTLA4

又稱 CTLA-4，是一種蛋白質受體，可充當免疫檢查點並下調免疫反應。此基因的變異與第 1 型糖尿病、格雷夫茲病、橋本甲狀腺炎、乳糜瀉、系統性紅斑狼瘡、甲狀腺相關眼眶疾病、原發性膽汁性肝硬化和其他自體免疫疾病有關 。有 CTLA-4 突變的有症狀患者的特徵是免疫失調綜合徵，包括腸道、肺、骨髓、中樞神經系統等多個器官中廣泛的 T 細胞浸潤（Dariavach P, Mattéi MG, December 1988）。

第二節 淋巴細胞

一T 細胞

是淋巴細胞的一種，在免疫反應中扮演著重要的角色。T 是胸腺（thymus）的英文縮寫。T 細胞在骨髓被製造出來之後，在胸腺內進行「新兵訓練」分化成熟為不同亞型的效應 T 細胞，成熟後就移居於周圍淋巴組織中開始工作。其中，根據T細胞表面標誌功能的差異可將T細胞劃分為不同亞群：根據 TCR 類型差異可分為αβ T 細胞和γδ T 細胞（Janeway, Charles,2001）。

二αβ T 細胞

αβ T 細胞站 T 細胞 95%以上，是體內 T 細胞分化標記、執行 T 細胞功能主要的細胞群，代表 T 細胞的多樣性。依據αβ T 細胞基因重排類型，可分辨不同的 T 細胞克隆，一個αβ T 細胞只能識別 APC 細胞呈遞的一種抗原片段，根據αβ T 細胞基因重排的方式分型，更能反映 T 細胞克隆性。

三γδ T 細胞

受體由特定的γ和δ鏈組成，其抗原識別不同於αβ T 細胞，不需要特異性抗原呈遞細胞，它能直接與抗原相互作用，其中最明顯的功能是細胞溶解活性，也能產生大量生物活性因子如 IFN-gamma、

TNF-α、IL-2 和 IL-4。

四 B 細胞

也稱 B 淋巴球，是白血球中一種淋巴細胞的亞型。B 細胞屬於後天免疫系統的體液免疫，作用為分泌抗體。此外，B 細胞能呈現抗原（也屬於抗原呈遞細胞）並分泌細胞激素。它來源於骨髓中的造血幹細胞，在骨髓中成熟，在體液免疫中產生抗體，起到重要作用。當遇到抗原時，會分化成核比例較大的漿細胞。漿細胞的細胞質中且會出現一些顆粒，這些顆粒容易被甲基藍等天青染料所染色，同時會出現抗體，表現在細胞膜或釋放出去。另一部分 B 細胞經過抗原激活後並不成為漿細胞，而是成為記憶 B 細胞。當再次遇到相同抗原時，記憶 B 細胞能迅速作出反應，大量分化增殖（Murphy, Kenneth,2012）。

第參章 研究方法

第一節 樣本選取

XenaUCSC 為線上即時資料庫（圖 3-1-1），我們選擇的資料庫為 TCGA （The Cancer Genome Atlas），TCGA 為美國癌症基因體圖譜計畫搜集之資料庫（圖 3-1-2），其內容為搜集病人的臨床病歷紀錄和腫瘤及其相對應正常組織樣本、血液樣本等等，進行全面的基因體資料擷取和整合性的分析資料（鄒佩玲&吳昌俊,2013, 美國癌症基因體圖譜計畫 TCGA 簡介）。

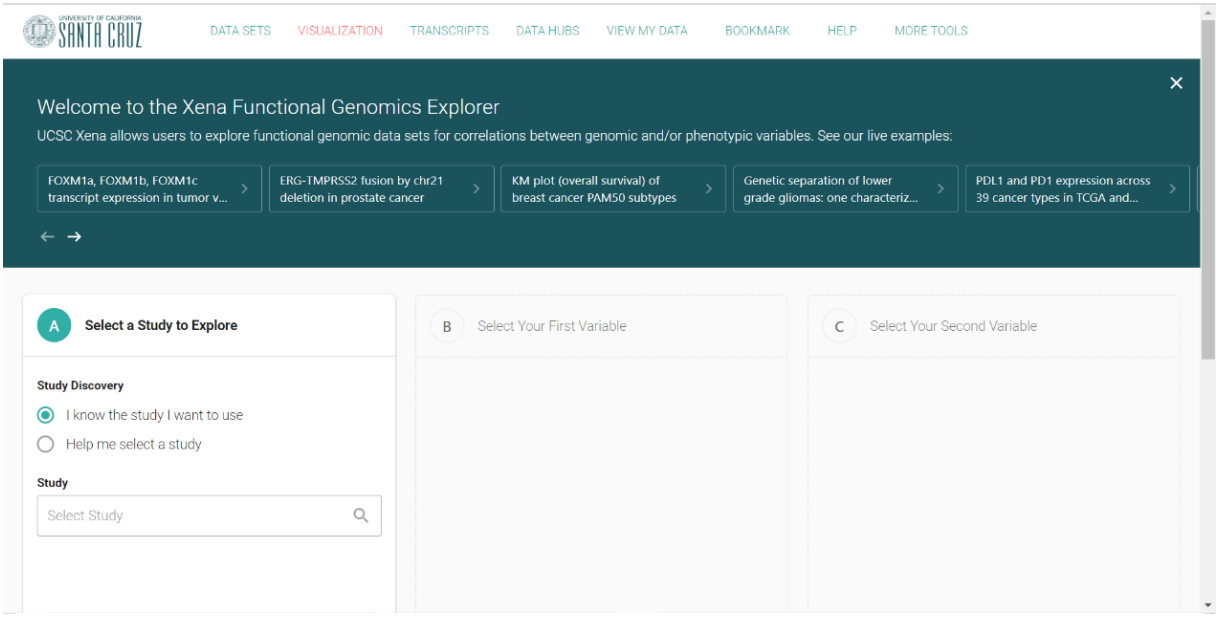

圖 3-1- 1 線上即時資料庫

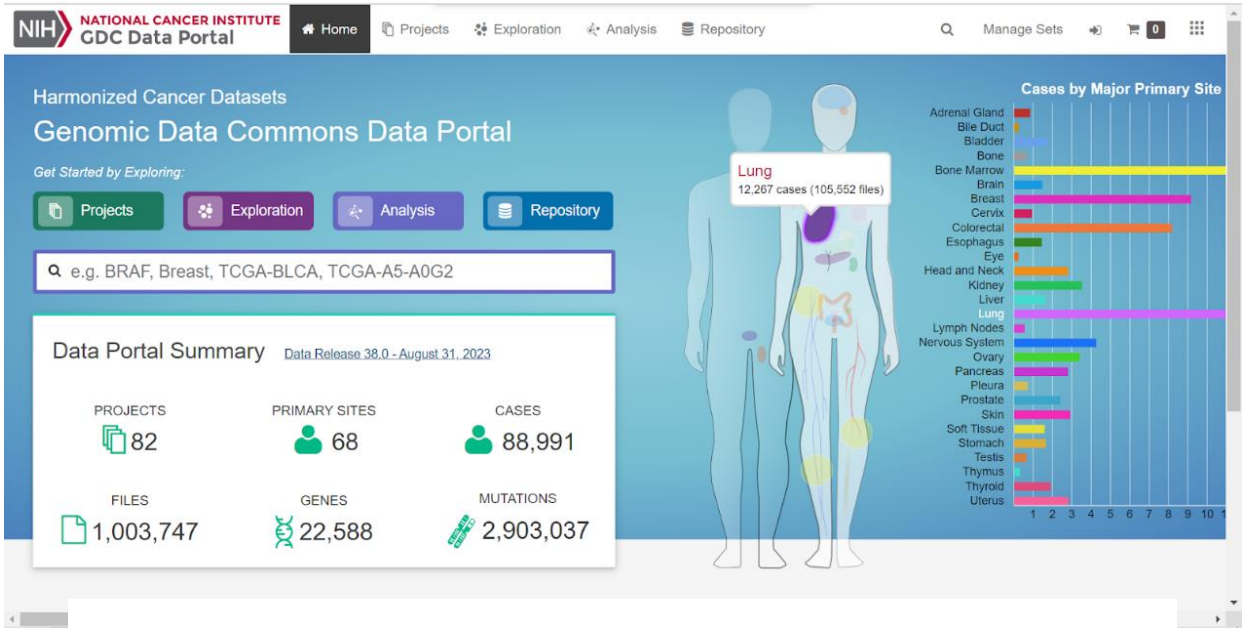

圖 3-1- 2 美國癌症基因體圖譜計畫搜集之資料庫

我們從 XenaUCSC 資料庫中找到 TCGA 並篩選出我們所需要的資料，其中有肺腺癌患者和肺鱗癌患者，再去一一選擇我們所需要的變數，像是性別、年齡、存活時間、抽菸史、目標基因...等，而我們也需要和正常人做比對，在正常人的部分我們則是使用了 GETx 的資料庫，並從中選出肺部基因的目標變數。

第二節 研究架構及流程

本研究使用 XenaUCSC 的資料庫來進行研究，下載後為 CSV 檔，並搭配 R 語言以及 Python，其中視覺化的圖表則是使用到 Excel 以及 GraphPad 來呈現。

本研究先進行敘述性統計，再使用相關的統計分析進行檢定，如卡方、相關係數、無母數檢定等等，去觀測一般人免疫細胞中的基因表達量和肺癌病患是否有差異，並更進一步的探討免疫細胞基因是否會因肺癌種類不同，基因表達量也不同，接著討論基因表達的高低與存活之間的相關性，最後使用主成分分析和分群來論證我們的結果。

第三節 檢定工具

一敘述性統計

本研究中針對 CD4, CD68, CD86, CD3D, CD8A, MS4A1, FOXP3, GATA3, TBX21, BTN3A1, PDCD1, CD274, CTLA4 等 13 個基因進行基本統計，分析各個基因表現量的基本資訊；後續針對性別、年齡、抽菸史的部分進行統計分析比較以及視覺化，初步了解肺腺癌與肺鱗癌的情況。

二曼－惠特尼 （Mann-Whitney） U 檢定

此檢定的目的為檢驗兩組獨立樣本或群體的均值是否有顯著差異，其類似於獨立樣本 t 檢定，但相較於獨立樣本 t 檢定，此檢定是將兩組樣本或群體的數值轉換為等級後再進行比較及計算，使用時間點則是在獨立樣本 t 檢定需要的基本假設交到違反時，也就是說若兩組樣本或群體不符合變異數同質性以及常態分配時，此檢驗方法會更適合被使用（H.B.Mann&D.R.Whitney,1947）。

曼－惠特尼 U 檢定的檢驗步驟如下，會有兩組獨立樣本或群集，一開始會先建立虛無假設和設定信心水準，兩組樣本數分別為 n1 和 n2，會將兩組獨立樣本或群集合在一起看並進行排序且由小到大分別給予 1~（n1+n2）的等級，接著再將兩組樣本或群集的等級分別加總為 R1 和 R2，而利用 n1、n2、R1、R2 帶入公式（1）、公式（2），可以計算出曼－惠特尼 U 檢定的檢定統計量 U、U’，最後則是利用 n1 和 n2 以及一開始設定的信心水準查詢曼－惠特尼 U 檢定臨界值表，得出兩個數值，並進行比較，去判斷是否拒絕虛無假設。

$$\mathrm{U}=n_1 n_2+\frac{n_1（n_1+1）}{2}-\mathrm{R}_1（1）$$

$$U' = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \quad (2)$$

而曼－惠特尼 U 檢定臨界值表中 n1、n2 的最大值皆只到 20，是提供給小樣本數所使用的，若兩獨立樣本數中有一超過 20，曼－惠特尼的抽樣分配則會趨於常態，因此可使用 n1、n2、U 檢定統計量以及公式（3）來計算 Z 檢定統計量，接著使用常態分配表來查看是否拒絕虛無假設。

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \quad (3)$$

本研究中，欲評估兩個隨機無母數的獨立樣本之間的差異，但檢驗過程中發現各項免疫基因的數值皆不符合常態分配，其中我們推測是由於數量太大而趨於常態但仍為偏態，因此在此使用 Mann-Whitney U test 來檢測在肺鱗癌與肺腺癌中各個蛋白質間是否具有顯著差異（Dr.Fish 漫遊社會統計,曼－惠特尼 U 檢定）。

三卡方檢定

卡方檢定約是在 1990 年由 Pearson 提出，檢定目的為檢驗兩類別資料間是否有關連性，此檢定為常態分布中所變化出來，卡方值是由標準常態分布 Z 統計檢定量的平方和變化所得到，其為公式（4）。卡方檢定使用時機通常有兩種，一種為適合度檢定（Goodness of Fit），此為檢驗同一母體中互斥的類別變數間是否具有差異或是是否依循比例，使用上可從觀察次數 （observed frequency, O） 及期望次數（expected frequency, E） 中求得卡方值，如公式（5）；另一種為獨立

性檢定，為檢驗兩項類別變數間是否有關聯性，通常虛無假設為兩類別之間互相獨立，使用上會列出 R*C 的列連表並算出個別期望值（expected frequency, E），再去計算卡方統計量，如公式（6）（Dr.Fish 漫遊社會統計,2023,卡方獨立性的假設檢定）。

$$\chi^2_{(n-1)} = \frac{\sum (x_i - \bar{x})^2}{\sigma^2} \quad (4)$$

$$\chi^2_{(n-1)} = \sum \frac{(O - E)^2}{E} \quad (5)$$

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (6)$$

在本研究中使用到卡方的獨立性檢定，目的為分別比較癌症和男女人數、各個年齡層以及三種抽菸階段是否有關聯性。

四 存活分析

存活分析為想要研究某個群體在某特定時間（starting event time point, zero time point）起經過一段時間的觀察追蹤，直到特定事件（event）的發生，即探討事件發生所需的時間（time to event）。存活函數（7）呈現的是在某特定時間點之下，個案可以活過特定時間點的機率是多少，是一個遞減的函數（Decreasing function），其中 t 為某個特定時間點，T 為觀測生存時間。

$$S(t) = P(T > t) \quad (7)$$

而用於生存時間最簡單的分布為指數分布 （exponential distribution），指數分布默認風險率 （hazard rate, λ） 不會因為時間而產生變化，在指數分布中生存方程如公式（8）。但使用指數分布的前提太過於強硬，因此有另一個更可以靈活運用的分布為 Weibull 分布，其生存方程如公式（9）。

$$S(t) = e^{-\lambda t} \quad (8)$$

$$S(t) = \exp\left(-\lambda t^{K}\right) \quad (9)$$

在存活函數的估算中有兩種方法較常使用，一種為 Life-table method，另一種為 Kaplan-Meier method，兩種差異在於前者是由固定時間設定區間，後者則是以每一個事件（event）發生時間點與設限（censoring）來設定區間。

而我們也有幾種可以檢定存活曲線的方法，分別為 log-rank test 跟 Wilcoxon test 以及 Cox Proportional Hazards Model。

Log-rank test 可以用來比較不同組別的存活曲線是否相同，Log-rank test 的優點是為無母數統計量，缺點則為只能提供 p-value，無法提供相關的處理效應（treatment effect），而在臨床上較少使用 Wilcoxon test。

若要計算處理效應（treatment effect）則可使用 Cox Proportional Hazards Model，其使用限制也為兩條存活曲線不可相交，主要假設兩組間危險函數比與時間無關、為一常數，這個常數我們稱為危險比（Hazard Ratio；HR）， H0：HR=1 versus Ha：HR≠1，危險比（Hazard Ratio；HR）之計算如公式（10）。

$$HR = \frac{\exp（\beta_0+\beta_1）}{\exp（\beta_0）}（10）$$

在本研究中用來估計生存函數的方法為 Kaplan-Meier method，並且使用 Log-rank test 對其進行檢定，最後使用 Cox Proportional Hazards Model 來計算處理效應，以上三種方法都會在後面進行更詳細的說明（吳雅琪,2011,當代醫藥法規）（王超辰,2023,醫學統計學）。

五 Log-rank test

假設有 A、B 兩群體，兩群體從觀察至死亡發生的時間共有 m 個，t_m為最長的時間長度，i=1,2,…,m，即表 4-3-1。H0: A、B 兩群體的存活曲線相同 vs. H1: A、B 兩群體的存活曲線不相同。

表 3-3- 1 假設 A、B 兩群體

組別	死亡人數	存活數量	每組人數
A	$d_{A,i}$	$n_{A,i} - d_{A,i}$	$n_{A,i}$
B	$d_{B,i}$	$n_{B,i} - d_{B,i}$	$n_{B,i}$
總和	d_i	$n_i - d_i$	n_i

其中，A、B 符合下面的統計分配、期望值以及變異數，如公式（11）、（12）、（13），而 A、B 兩群體的觀察與預期死亡人數差異互為相反數，如公式（14），並經由公式（14）即可推導出 Log-rank 統計量，如公式（15）。

$$d_{A,i} \sim Hyper(n_i, n_{A,i}, d_i)，\ d_{B,i} \sim Hyper(n_i, n_{B,i}, d_i)，$$

$$i = 1,2,\dots,m（11）$$

$$E[d_{A,i}] = e_{A,i} = \frac{n_{A,i}}{n_i} * d_i \text{，} E[d_{B,i}] = e_{B,i} = \frac{n_{B,i}}{n_i} * d_i \text{（12）}$$

$$\text{Var}[d_{A,i}] = \frac{n_{A,i}\text{（}n_i - n_{A,i}\text{）（}n_i - d_i\text{）}d_i}{r_i^2\text{（}r_i - 1\text{）}} \text{，} \text{Var}[d_{B,i}] = \frac{n_{B,i}\text{（}n_i - n_{B,i}\text{）（}n_i - d_i\text{）}d_i}{r_i^2\text{（}r_i - 1\text{）}}$$

$$\text{（13）}$$

$$\sum_{i=1}^{m}(d_{A,i} - e_{A,i}) = \sum_{i=1}^{m}\text{（}d_{A,i} - \text{（}d_i - e_{B,i}\text{））} = -\sum_{i=1}^{m}\text{（}d_{B,i} - e_{B,i}\text{）（14）}$$

$$\frac{[\sum_{i=1}^{m}\text{（}d_{A,i} - e_{A,i}\text{）}]^2}{Var\text{（}[\sum_{i=1}^{m}\text{（}d_{A,i} - e_{A,i}\text{）}]\text{）}} = \frac{[\sum_{i=1}^{m} d_{A,i} - E[\sum_{i=1}^{m} d_{A,i}]]^2}{Var\text{（}\sum_{i=1}^{m} d_{A,i}\text{）}} \sim X^2 \text{（15）}$$

使用公式（15）計算出的 Log-rank 統計量之 p-value 若是小於 α，則會拒絕$H0$，表示 A、B 群體之存活曲線整體上有顯著差異（廖佳馨,2023,失智症身心障礙者之障礙等級轉換及住院醫療成本）。

六 Kaplan-Meier method

Kaplan-Meier method 又可稱 Product-limit method，是用來估計存活曲線，以每一個事件發生時間點與設限點（censoring） 來設定區間，較適用於一般臨床相關研究。在蒐集存活資料時，有時無法確切的記錄所感興趣事件所發生的時間，此現象稱之為設限（censoring），造成設限的原因可能是在進行資料分析時所關切的事件尚未發生；也可能因受測者失去聯絡或中途離開研究。存活函數（survival function）為公式（16）。

$$S_r(t) = P_r(T \geq t) = \int_t^{\infty} f_T(u)du = 1 - F_t\text{（}t\text{）}, t \geq 0\text{（16）}$$

Kaplan and Meier （1958）利用 product limit 的無母數方法，提出存活函數$S_r(t)$之估計式為公式（17）。

$$\hat{S}_r(t) = \prod_{t_j<t}\left(\frac{n_j - d_j}{n_j}\right)\text{（17）}$$

其中n_j表示存活時間，t_j前確定存活之個數，d_j表示存活時間t_j前之「發生」之個數（蘇秀媛, 蘇志雄, 謝鑫能,2006,長期存活資料比率之統計分析）。

七 Cox Proportional Hazards Model

風險函數（hazard function），風險函數$h_T(t)$值不隨時間變動，其計算如公式（18）。

$$h_T(t) = \lim_{\varepsilon \to 0} \frac{P_r(t < T < t + \varepsilon | T > t)}{\varepsilon} = \frac{f_T\text{（}t\text{）}}{S_T\text{（}t\text{）}}, t \geq 0\text{（18）}$$

而在 Cox proportional hazard model 之下的風險函數、生存函數、生存累積函數分別如公式（19）、（20）、（21）。

$$h_T(t) = \lambda\text{（19）}$$

$$S_T(t) = exp\left(-\int_0^t h(u)du\right) = exp(-\lambda t) = e^{-\lambda t}\ （20）$$

$$f_T(t) = h_T(t) * S_T(t) = \lambda * e^{-\lambda t} = \lambda * exp\ （-\lambda \mathrm{t}）（21）$$

在公式（19）、（20）、（21）中，λ為一定值，由此可推導出 Cox proportional hazard model 為公式（22）。

$$log\frac{h_T(t)}{h_0(t)} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon\ （22）$$

八 Kruskal－Wallis 檢定

Kruskal－Wallis 檢定也為一種無母數檢定，是用來研究三個或三個以上的獨立母體是否有顯著差異，Kruskal－Wallis 檢定的使用時機為若是單因子變異數分析的假設受到嚴重違反時，例如樣本為非常態、樣本數極少或變異數非同質性時，Kruskal－Wallis 檢定會比單因子變異數分析更適合被使用。

Kruskal－Wallis 檢定的假設過程首先也是先建立虛無假設、選擇顯著水準，接著計算統計檢定量（H），和曼－惠特尼 U 檢定一樣，會先將樣本由小到大進行排序，並給予評級，在將其各群的級分加總（Ri），公式（23）為 Kruskal－Wallis 之檢定統計量之計算，其中 k 為群組組數，ni 為第 i 個群組的個數，N 為樣本總數。

$$\mathrm{H} = \frac{12}{N\ （N+1）}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3\ （\mathrm{N}+1）（23）$$

算出檢定統計量後，在 Kruskal–Wallis 檢定中，若是每個中的樣本數都大於五個，那統計檢定量會趨於卡方分配，其中自由度為（k-1），因此查找卡方分配臨界值表後再去判斷是否拒絕虛無假設（Dr.Fish 漫遊社會統計,2023,Kruskal–Wallis 檢定的假設檢定）。

九 主成分分析（Principal components analysis, PCA）

主成分分析（Karl Pearson,1901）為一降維的方法，其目的是將一群變數經由線性轉換的方式，轉換成新的變數，而主成分分析的基本假設則是希望資料可以在特徵空間找到一個投影軸，投影後可以獲得這組資料的最大變異量，其中會使用到線性代數去解出其中共變異數矩陣的特徵值（eigenvalue, λ）以及特徵向量（eigenvector, v）（黃志勝,2018,機器/統計學習:主成分分析）。

在舊變數轉換新變數的過程中，新變數稱為主成分，而其個數一定是小於或等於原始變數之個數，其中要選幾個主成分則是要看其可以解釋原變數的程度為多少，也就是說依照自己選擇主成分需要表示原資料多少百分比的變異量，來選擇需要幾個主成分，而舊變數經轉換後落在新座標軸的向量則稱為主成分計分（principal component scores）（黃文彥,2021,R 統計）。

十 分群分析

(一) K-means

K-means 為非監督式學習的演算法，若是要將一群資料分成 k 群（cluster），需透過計算資料間的距離來作為分群的依據，相近的資料會成形成一群再透過加權計算或簡單平均來找出中心點，透過多次反覆計算與更新各群中心點後，最後中心點不再改變，即可以找出最後代表該群的中心點，之後便可以透過與中心點的距離來判定與測試資料屬於哪一分群（王建凱,2018,K-means）。

(二) 階層式分群（Hierarchical Clustering）

階層式分群不需要預設分群數，產生出來的分群結果為樹狀結構圖。其中群數可由大變小（divisive hierarchical clustering），或是由小變大（agglomerative hierarchical clustering），透過群聚反覆的分裂和合併後，選取最佳的群數。

階層式分群有兩種演算法，首先為聚合法，又稱作 AGNES（Agglomerative Nesting），資料會由樹狀結構的底部開始開始逐次合併 （bottom-up），擅於處理與識別小規模群聚；另一種為分裂法，又稱作 DIANA（Divisive Analysis），資料則會由樹狀結構的頂部開始逐次分裂（top-down），擅於處理與識別大規模群聚（Hsuan,2018,Hierarchical Clustering 階層式分群）。

第四節 資料處理

本研究僅有針對年齡以及抽菸史的部分進行資料預處理，其餘性別、13 個基因等等，皆無做改變。

一 年齡

在本研究中，僅年齡的部分含有遺失值，使用十歲為一個級距將各個年齡層劃分，有研究指出肺癌患者在 50 歲以上人群的發病率會有不同趨勢（Huang, Deng et al. 2022），因此本研究最後使用 50~59 歲作為遺失值的替代。

二抽菸史

在抽菸史的部分，因為僅有收集到停止吸菸年份以及開始吸菸年分，我們根據這兩項變項，將癌症患者分為從未吸菸、曾抽菸但已戒菸以及持續抽菸三個面向。

第肆章 研究結果

本研究欲探討肺腺癌、肺鱗癌患者與正常人三者進行相關性分析與生存分析，最後針對資料進行分群，本章節分為四個部分，第一節為敘述性統計，第二節為性別、年齡、抽菸史與肺腺癌、肺鱗癌及正常人的卡方分析，第三節為各肺腺癌、肺鱗癌及正常人的獨立樣本檢定，第四節相關係數，第五節為肺腺癌、肺鱗癌患者的存活分析，第六節為分群結果。

第一節 敘述性統計

本研究樣本資料取自 Xena 資料庫，總筆數共 1304 筆，其中肺腺癌資料 514 筆，肺鱗癌資料 502 筆，正常資料 288 筆，共 16 個變數，14 個連續型變數，2 個類別變數。

一 性別

在正常人資料中，女性有 100 人，男性有 188 人，分別 34.7%與 65.3%；在肺癌患者中，女性有 407 人，男性有 609 人，分別為 40.1%與 59.9%。

二 年齡

由於無正常人資料，因此針對罹癌患者做敘述性統計，平均值為 66.8 歲，最大值為 90.1 歲，最小值為 38.5 歲，標準差為 9.3，40 歲以下罹患肺癌有 3 人，約占 0.3%；40~49 歲罹患肺癌有 47 人，約占 4.6%；50~59 歲罹患肺癌有 213 人，約占 21%；60~69 歲罹患肺癌有 351 人，約占 34.5%；70~79 歲罹患肺癌有 345 人，約占 34%；80 歲以上罹患肺癌有 57 人，約占 6%。

三 抽菸史

此因子無正常人資料，因此也針對罹癌患者做敘述性統計，從未抽菸有 273 人，約占 26.9%，曾抽菸，但已戒菸有 565 人，約占 56%，持續抽菸有 178 人，約占 17.5%。

四 其他基因因子

表 4-1-1 本研究的基因因子包含 CD4、CD3D、TBX21、FOXP3 等等共 13 個變數的敘述統計表。

表 4-1- 1 各項基因因子敘述統計

基因名稱	CD4	CD3D	TBX21	GATA3	CD8A	FOXP3	BTN3A1	CD86	MS4A1	CD68	PDCD1	CD274	CTLA4
平均數	10.91	6.91	5.19	6.66	8.07	5.83	10.08	8.30	6.53	12.12	5.76	7.13	5.99
標準差	1.24	1.37	1.84	1.59	1.45	1.35	1.08	1.09	2.31	1.13	1.44	2.28	1.4
變異數	1.53	1.88	3.39	2.53	2.10	1.81	1.71	1.18	5.35	1.27	2.06	5.19	1.95
中位數	11.10	7.00	5.02	6.78	8.14	5.84	10.09	8.43	6.72	12.17	5.89	6.64	6.11
最大值	13.74	11.34	11.21	14.32	12.19	9.65	13.60	11.43	14.63	15.10	10.02	13.33	10.78
最小值	5.96	0.47	0	1.06	2.00	1.31	6.17	4.18	0	8.11	0	1.05	0

第二節 卡方分析

本研究中性別、抽菸史兩變項均屬類別變數，而年齡屬於連續變數，本研究將年齡段人數轉換為 2*6 列聯表，分成 40 歲以下、40~49 歲、50~59 歲、60~69 歲、70~79 歲、80 歲以上共六個區段，將上述三變數進行卡方檢定。由於在正常人中並無紀錄抽菸史與年齡的資料，因此在抽菸史和年齡的類別下，虛無假設在 95%的信心水準下，抽菸史（年齡）與肺癌種類無關；而在性別變數的類別下，虛無假設則為在 95%信心水準下，性別與是否得肺癌無關。

一 性別

圖 4-2-1 為性別分類下的肺癌與正常人數圖，患有肺癌的男性 609 人，女性 407 人；沒有罹患肺癌的男性有 188 人，女性則有 100 人，利用卡方檢定得 p-value 值為 0.11 大於 0.05，表示性別與是否得肺癌沒有相關。由於性別與是否得肺癌沒有相關，因此再將肺癌細分為肺腺癌、肺鱗癌，其中患有肺腺癌的男性 237 人，女性 277 人；患有肺鱗癌的男性 372 人，女性 130 人，並進行卡方檢定得 p 值小於 2.2e-16 小於 0.05，表示性別與肺癌種類有相關，由圖 4-2-2 可知，肺腺癌患者以女性居多，肺鱗癌患者則以男性居多。

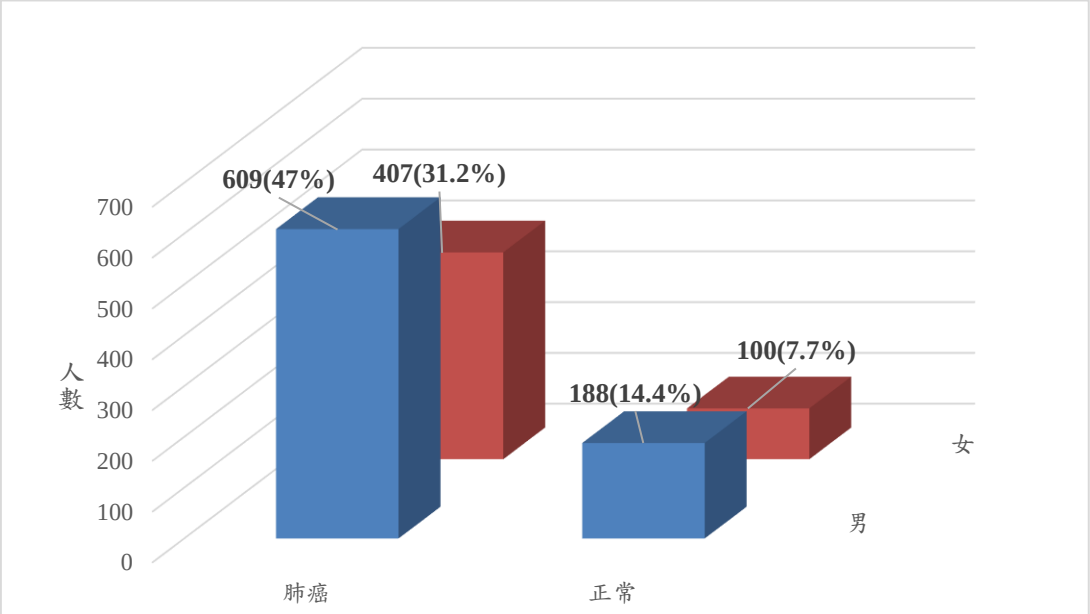

圖 4-2- 1 肺癌患者與正常人的性別人數圖

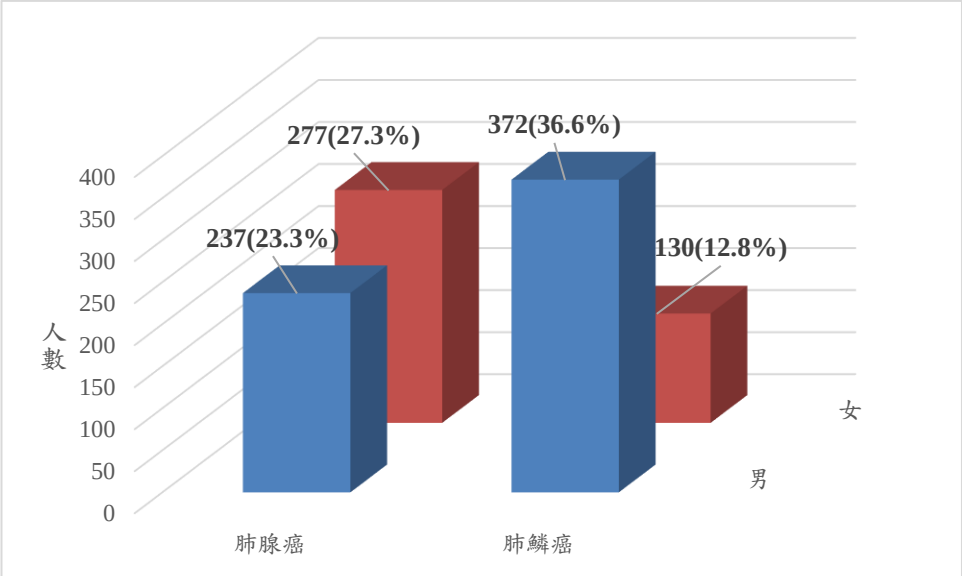

圖 4-2- 2 肺腺癌、肺鱗癌患者的性別人數圖

二 年齡

圖 4-2-3 為年齡分類下的肺腺癌、肺鱗癌人數圖，其中「40 歲以下」罹患肺腺癌只有 2 人，肺鱗癌則只有 1 人；「40~49 歲」罹患肺腺癌有 30 人，肺鱗癌則有 17 人；「50~59 歲」罹患肺腺癌有 130 人，肺鱗癌則有 83 人；「60~69 歲」罹患肺腺癌有 168 人，肺鱗癌則有 183 人；「70~79 歲」罹患肺腺癌有 154 人，肺鱗癌則有 191 人；「80 歲以上」罹患肺腺癌有 30 人，肺鱗癌則有 27 人，利用卡方檢定得 p-value 值小於 2.2e-16 小於 0.05，表示年齡與肺癌類型有關聯，由圖 4-2-3 可知，年齡介於 60~79 歲確診肺癌的比例較高。

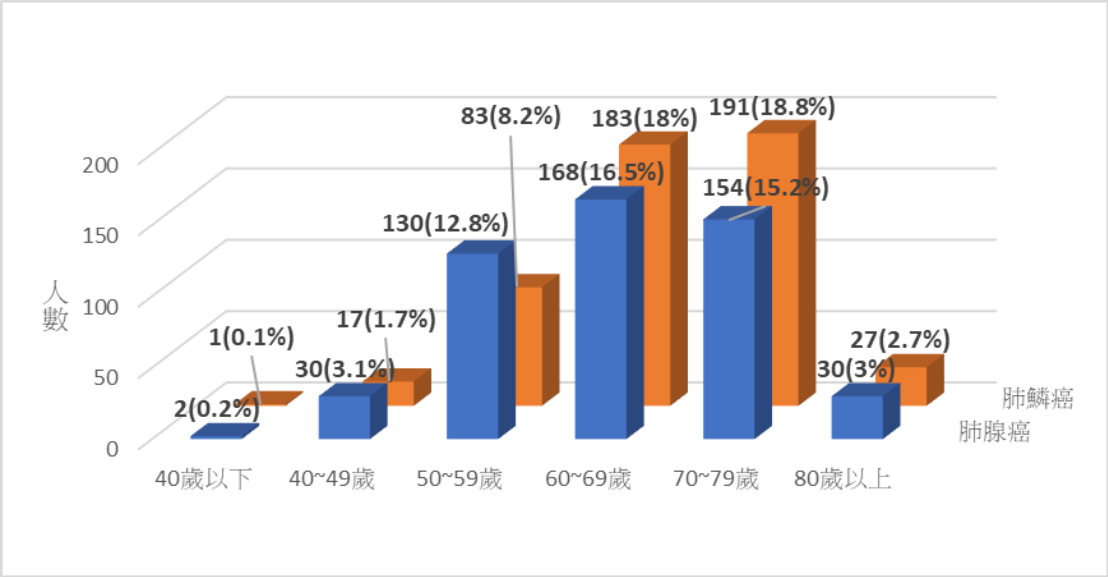

圖 4-2- 3 肺癌患者年齡人數圖

三抽菸史

圖 4-2-5 為抽菸史分類下的肺腺癌、肺鱗癌人數圖，其中「從未抽菸」且罹患肺腺癌有 174 人，肺鱗癌則有 99 人；「曾抽菸，但已戒菸」且罹患肺腺癌有 258 人，肺鱗癌則有 307 人；「持續抽菸」且罹患肺腺癌有 82 人，肺鱗癌則有 96 人，利用卡方檢定得 p 值為 0.2393 大於 0.05，表示抽菸史與肺癌類型沒有相關，由圖 4-2-4、圖 4-2-5 可知，男性有抽菸的比例最高，佔 46.1%，有抽菸史的患者罹患肺鱗癌的比例高於肺腺癌，沒有抽菸史的患者則以肺腺癌居多。

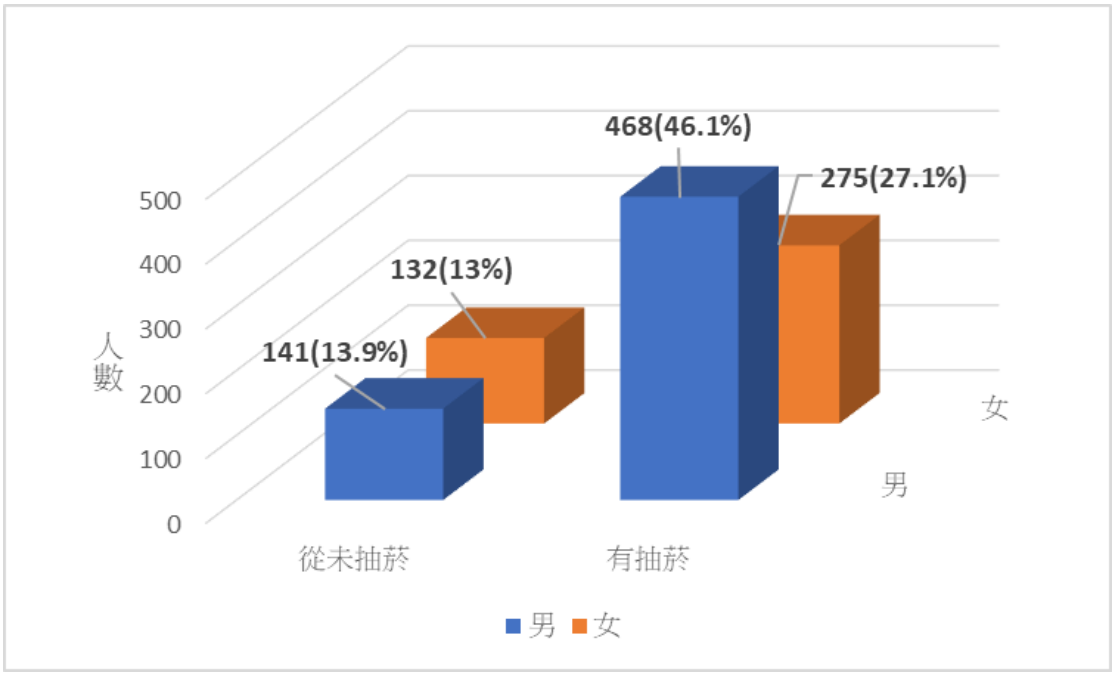

圖 4-2- 5 男女抽菸概況

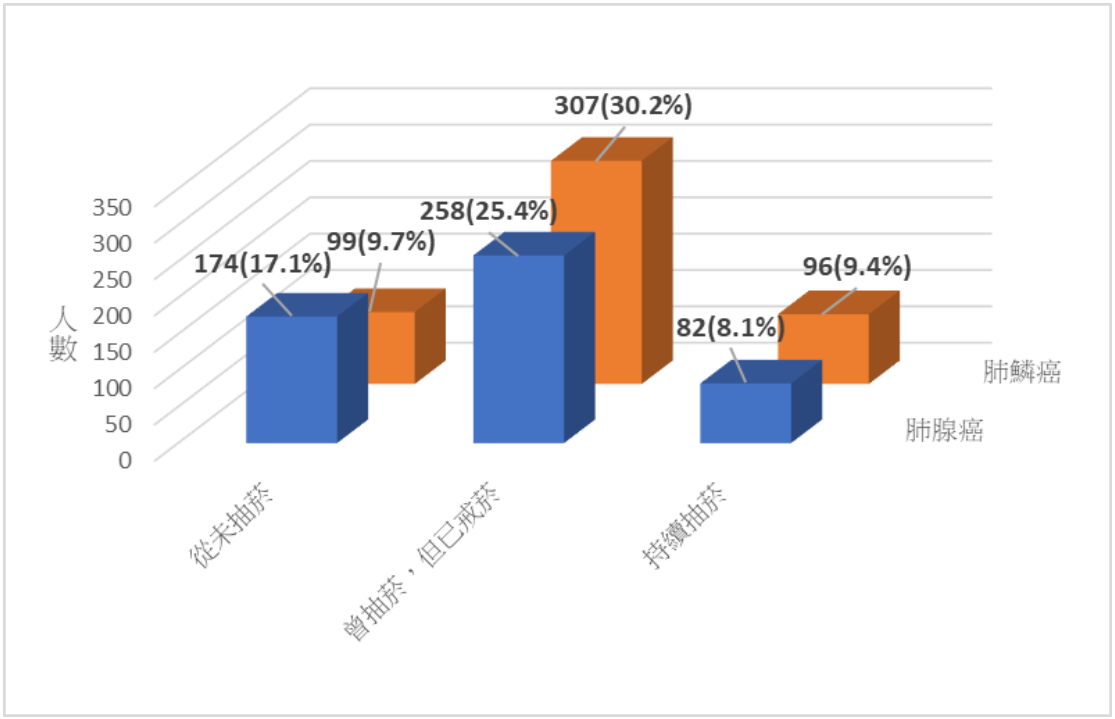

圖 4-2- 4 肺癌患者抽菸

第三節 獨立樣本檢定

本研究中針對 CD4 （圖 4-3-1）、CD3D（圖 4-3-2）、TBX21（圖 4-3-3）、GATA3（圖 4-3-4）、CD8A（圖 4-3-5）、FOXP3（圖 4-3-6）等等共 13 個基因變數先進行 Kruskal-Wallis 檢測肺腺癌、肺鱗癌患者與正常人三者之間基因表達是否有差異，H0 假設: 肺腺癌、肺鱗癌患者與正常人之間基因表達無差異，結果顯示所有 p 值皆小於 0.05，表示肺腺癌、肺鱗癌患者與正常人之間基因表達有顯著不同。因此，進一步使用 Mann-Whitney U 檢定個別檢測肺腺癌與正常人、肺鱗癌與正常人、肺腺癌與肺鱗癌的基因表達是否有顯著差異，結果顯示肺腺癌、肺鱗癌患者分別和正常人檢定的 p 值皆小於 0.05 有顯著差異；肺腺癌與肺鱗癌患者在 GATA3 及 FOXP3 的 p 值大於 0.05 沒有顯著差異，其餘基因皆有顯著差異。由圖 4-2-11 得知 FOXP3 在肺癌患者中基因表達較正常人高，其他基因則是肺癌患者表達較低；CD274（圖 4-3-12）、GATA3（圖 4-3-4）在肺鱗癌表達較高，其餘基因則是肺腺癌表達高。

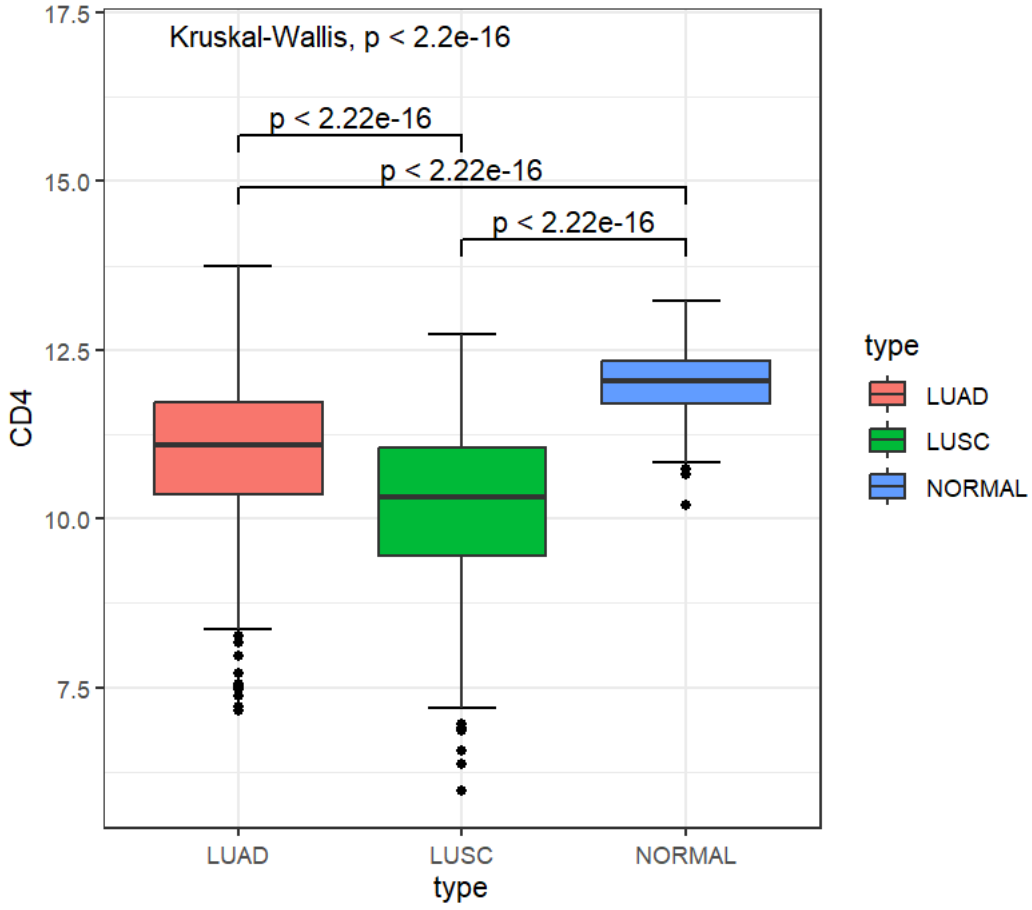

圖 4-3- 1 CD4 基因表現盒狀圖

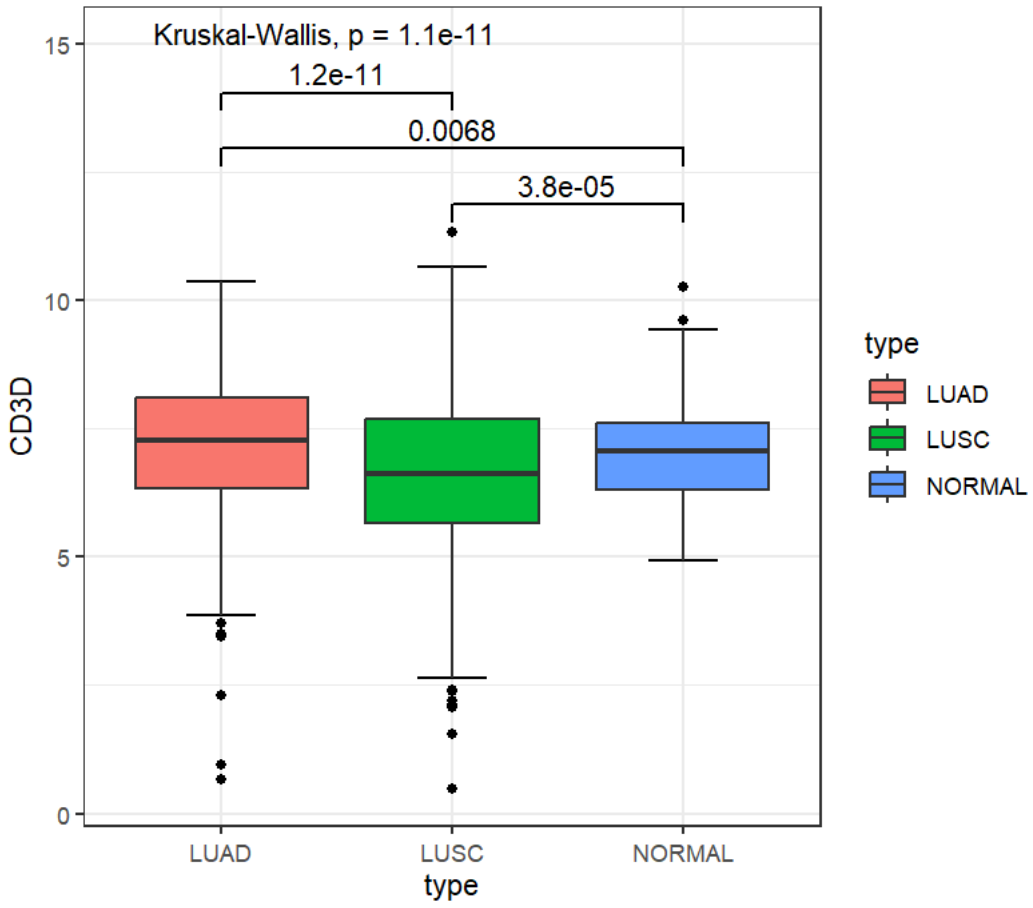

圖 4-3- 2 CD3D 基因表現盒狀圖

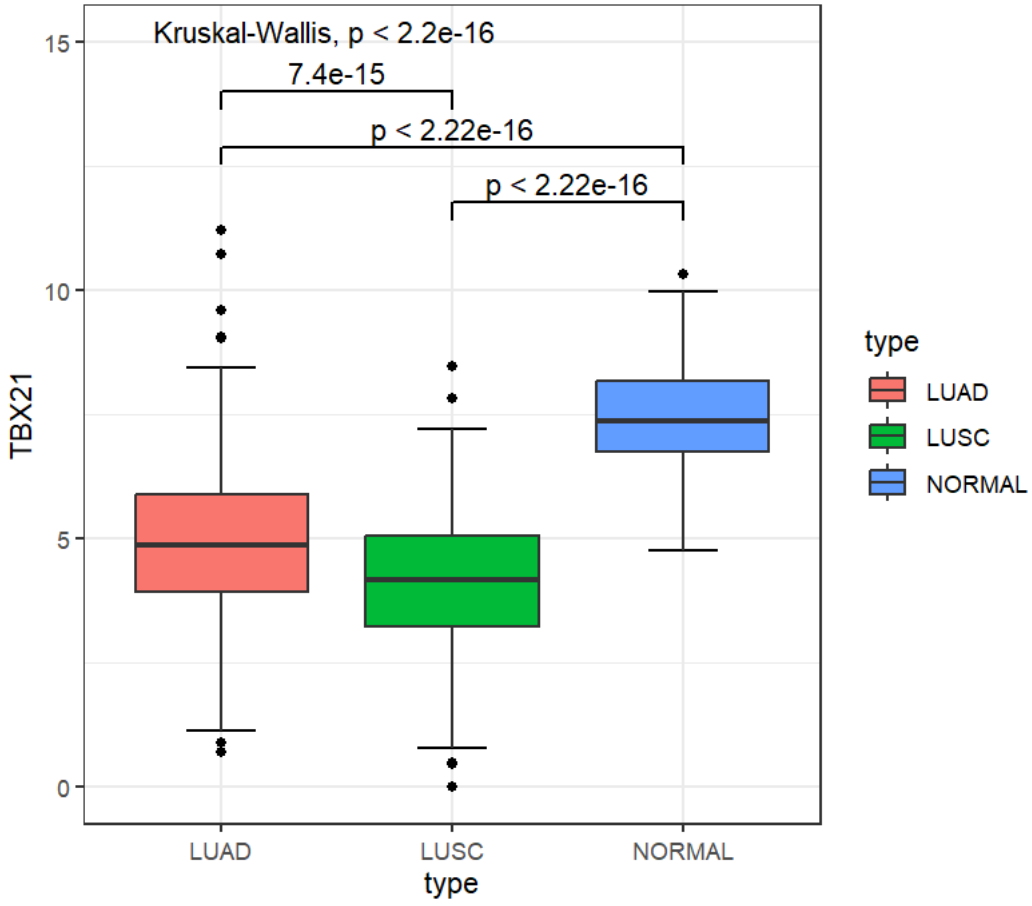

圖 4-3- 3 TBX21 基因表現盒狀圖

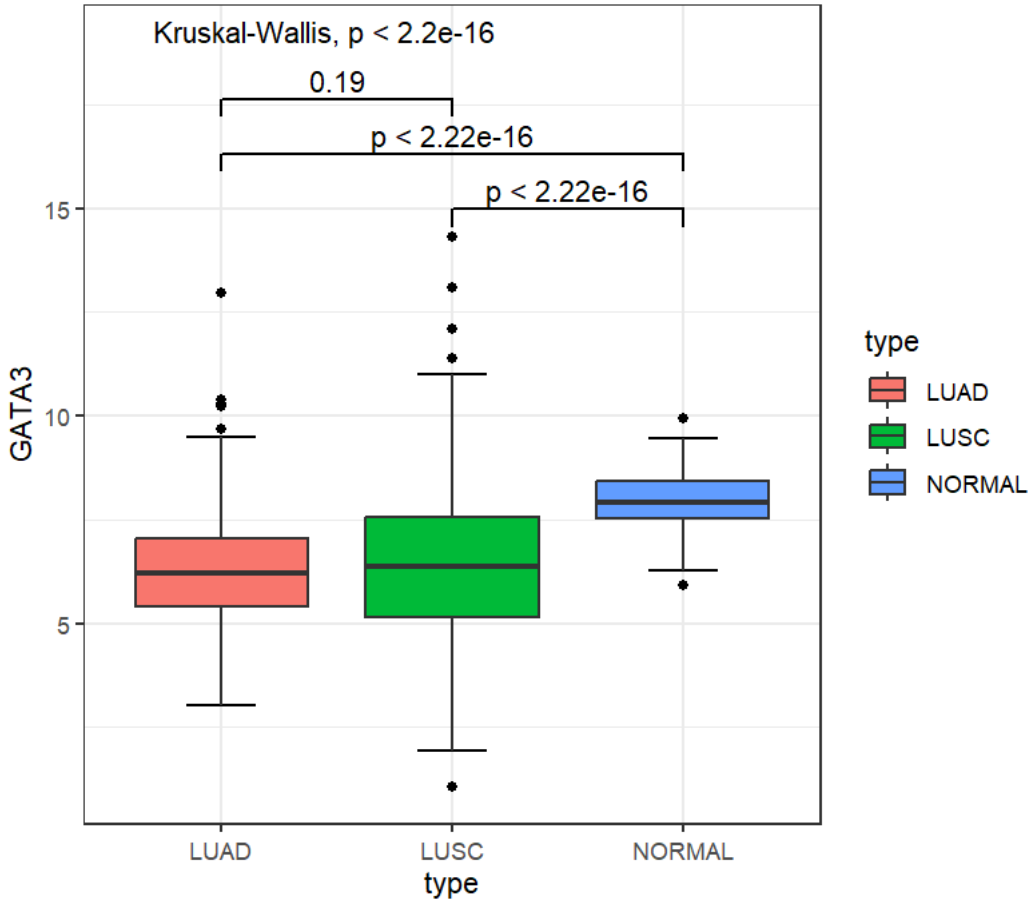

圖 4-3- 4 GATA3 基因表現盒狀圖

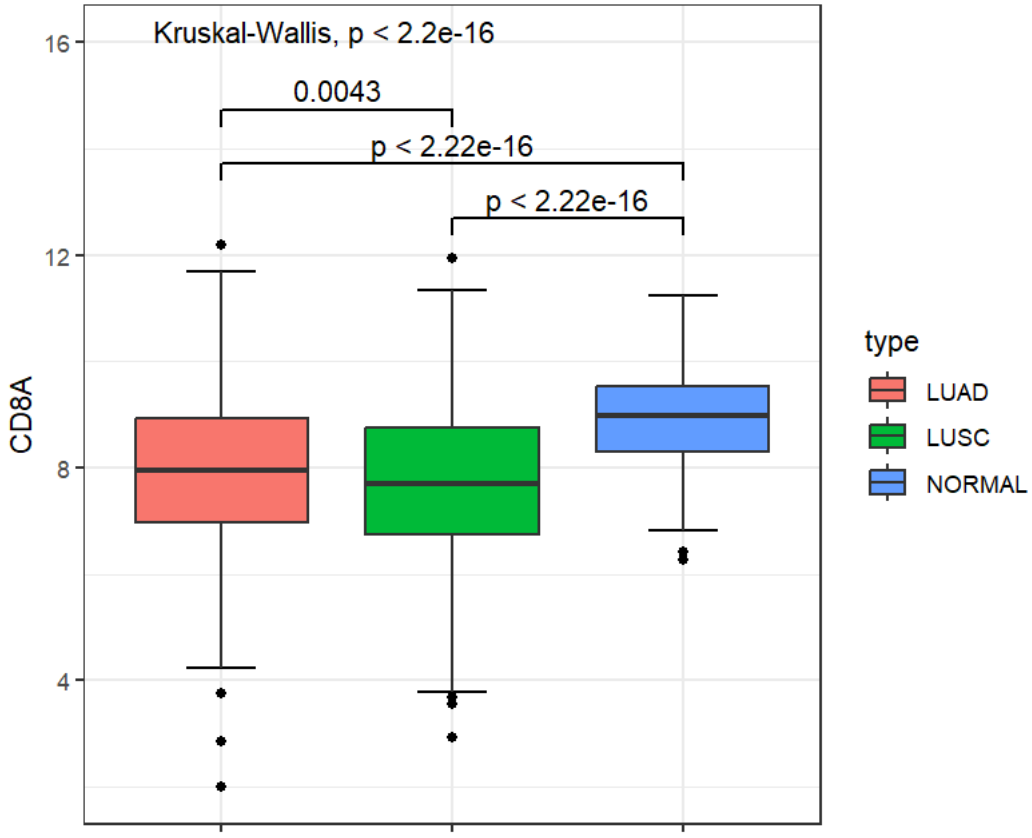

圖 4-3- 5 CD8A 基因表現盒狀圖

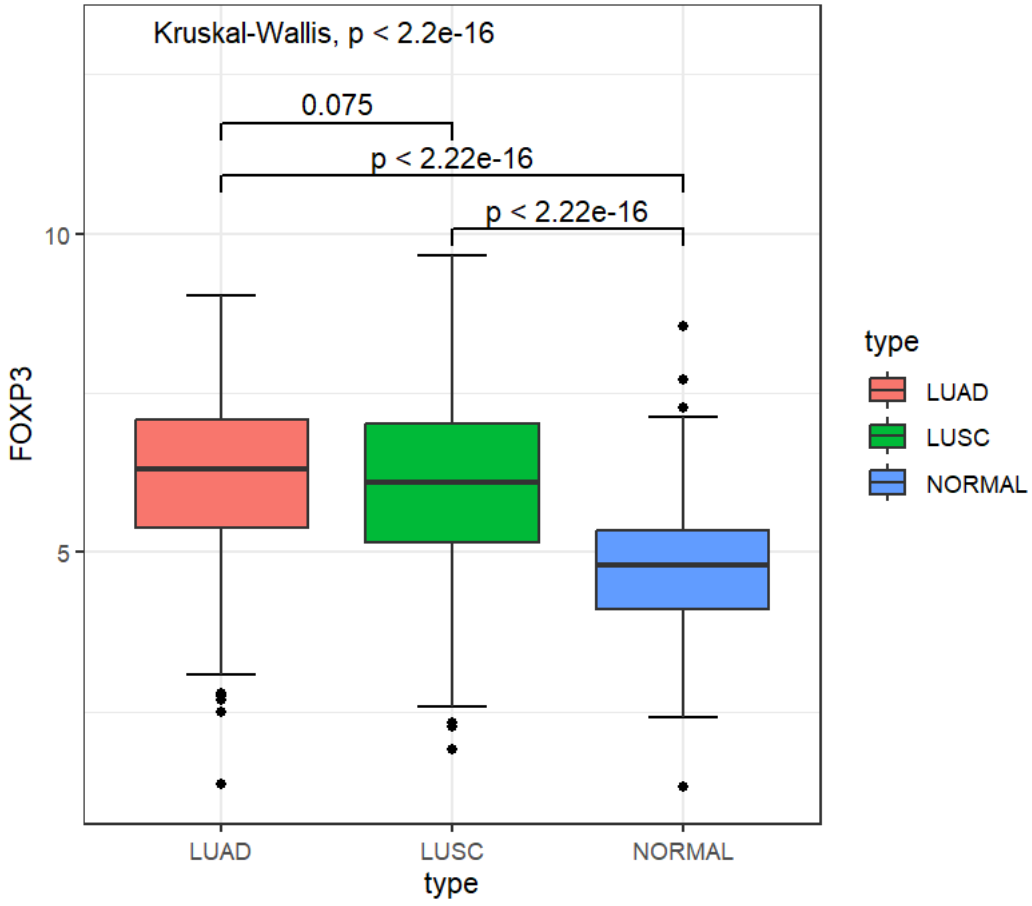

圖 4-3- 6 FOXP3 基因表現盒狀圖

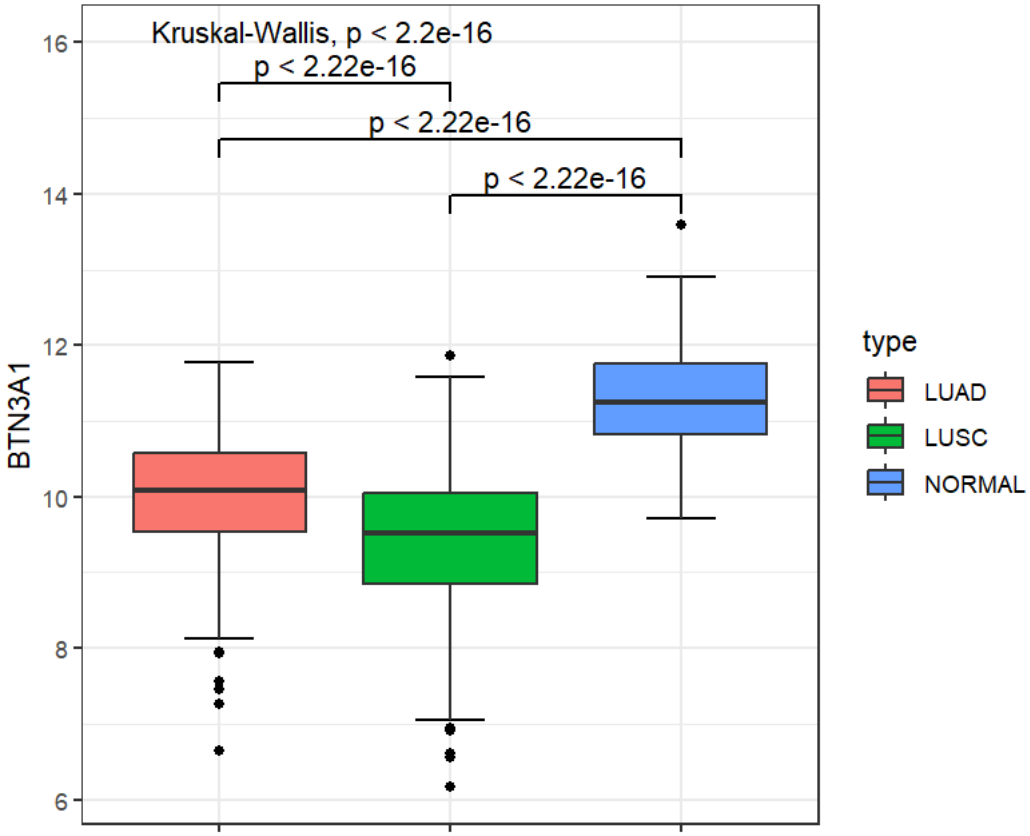

圖 4-3- 7 BTN3A1 基因表現盒狀圖

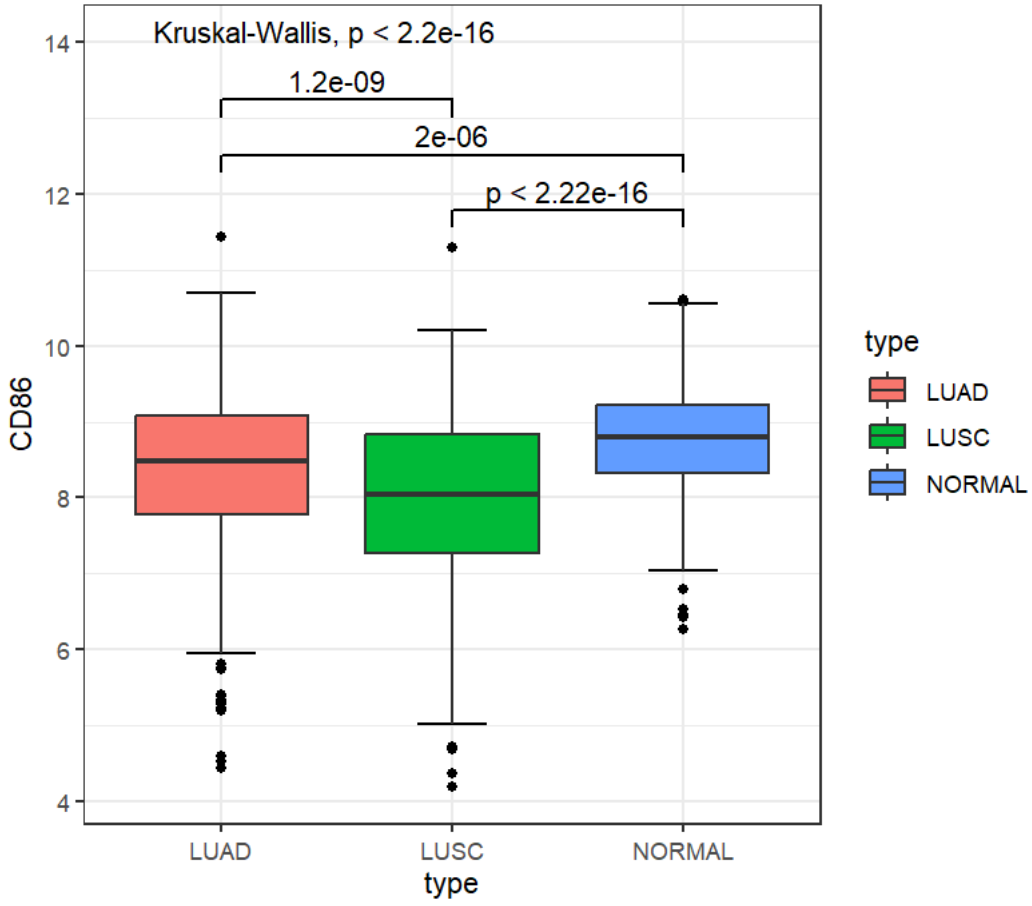

圖 4-3- 8 CD86 基因表現盒狀圖

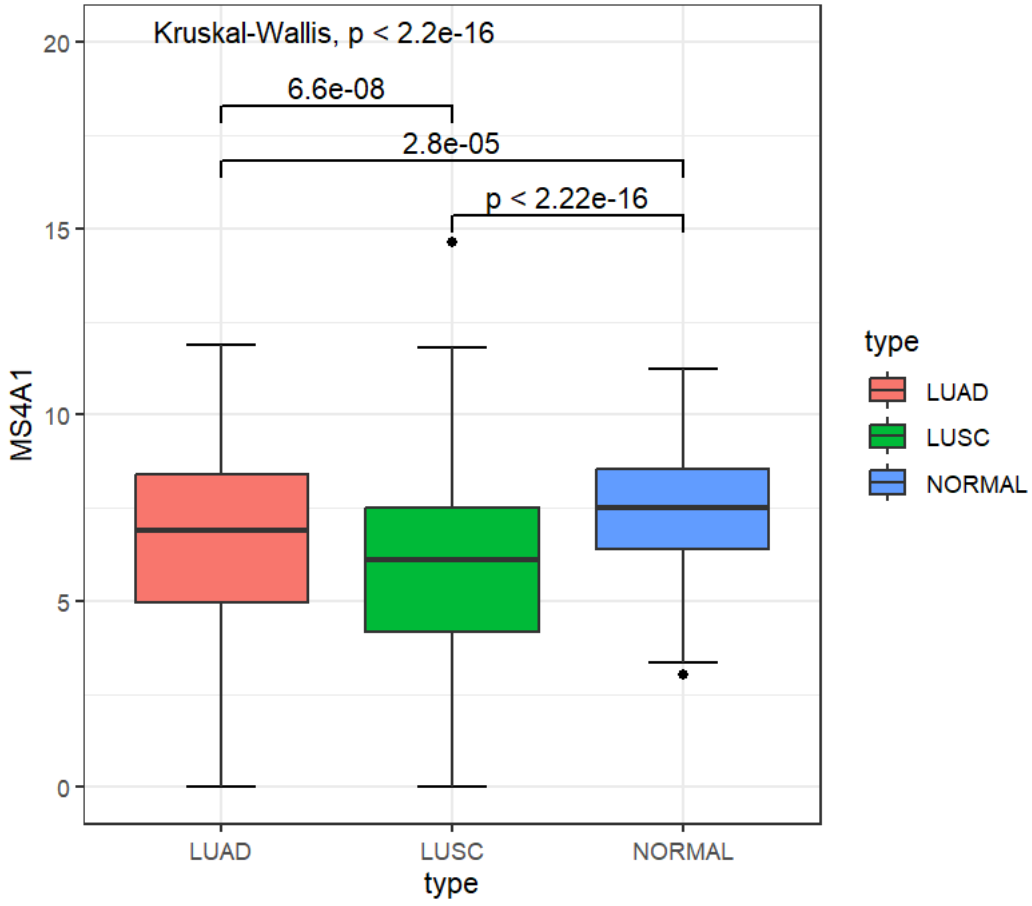

圖 4-3- 9 MS4A1 基因表現盒狀圖

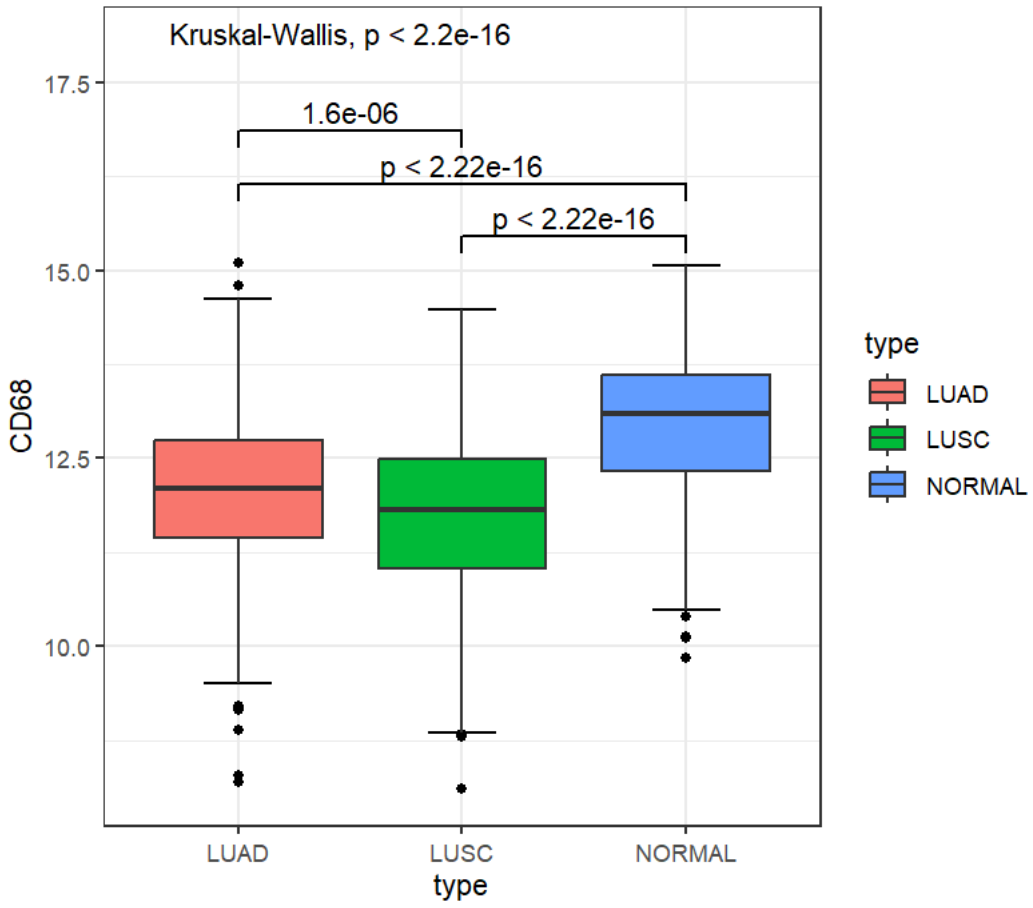

圖 4-3- 10 CD68 基因表現盒狀圖

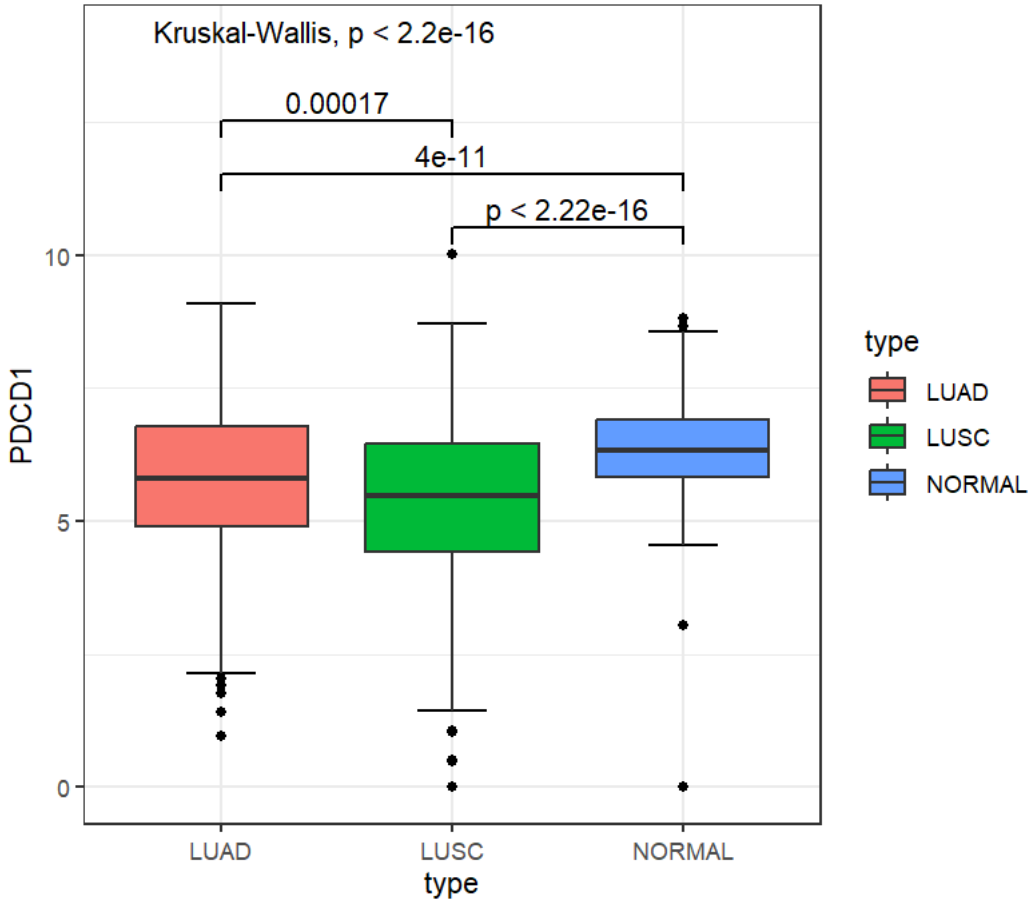

圖 4-3- 11 PDCD1 基因表現盒狀圖

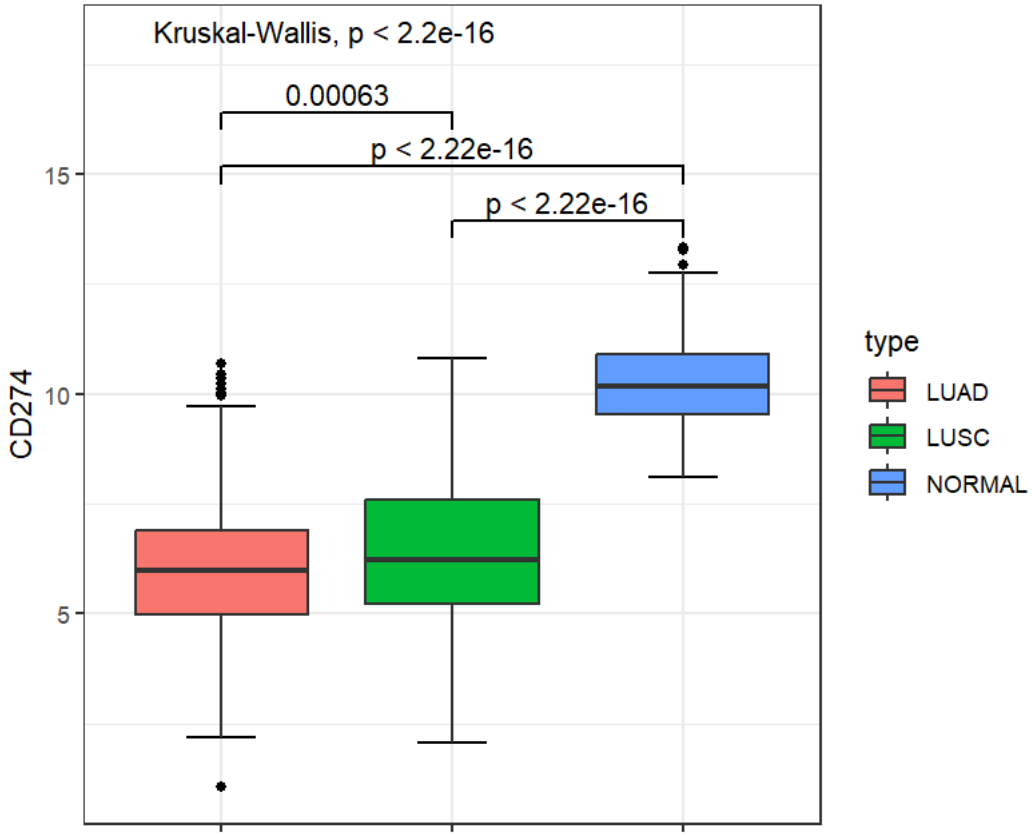

圖 4-3- 12 CD274 基因表現盒狀圖

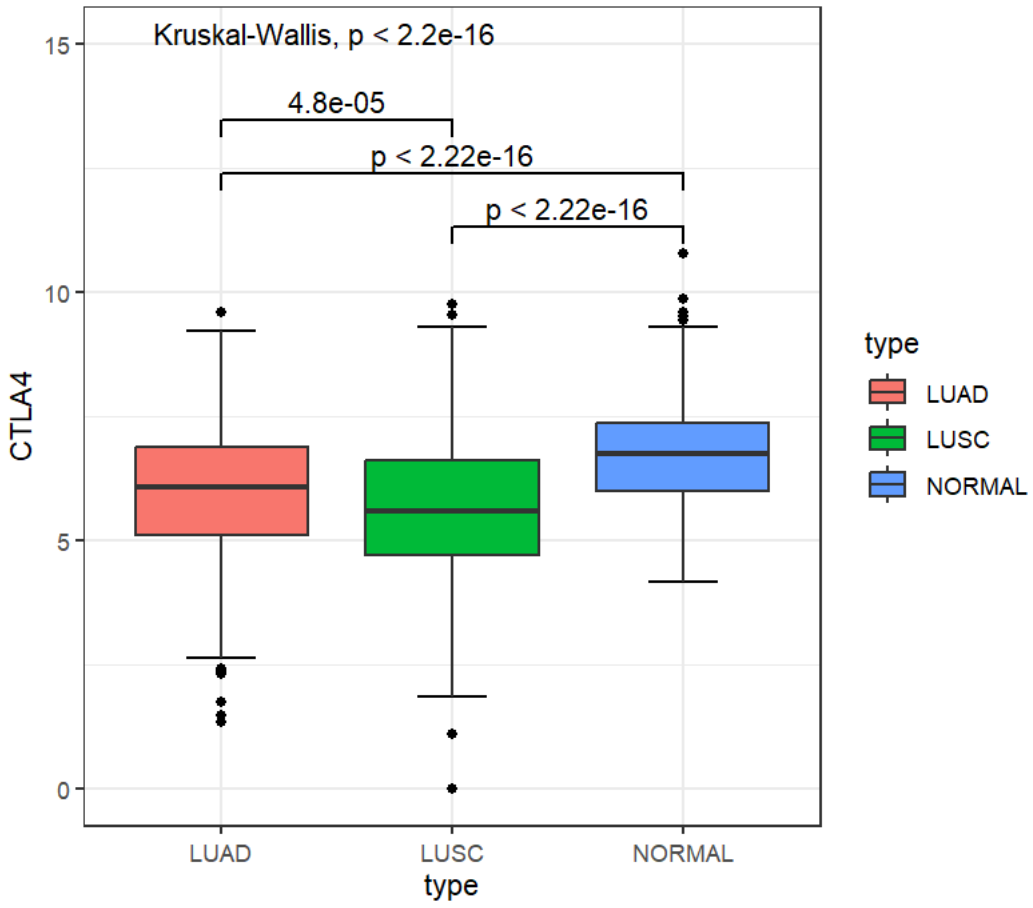

圖 4-3- 13 CTLA4 基因表現盒狀圖

第四節 相關性分析－Spearman 等級相關

本研究針對 13 個免疫細胞基因使用 Spearman 等級相關，分析兩獨立基因變數間的相關程度，虛無假設: $\rho = 0$ 兩變數之間無顯著相關。結果顯示所有 p-value 皆小於 0.05，表示兩變數之間有顯著相關，相關係數越接近 1 或是-1，代表其線性關係越明確。肺鱗癌下 GATA3 與 CD274 之間的相關係數為-0.1349，表示兩者間呈負相關。我們可從中推斷，在肺鱗癌下 GATA3 表現越高，CD274 表現可能越低。而其它變數之間都呈正相關，其之間的相關程度如 表 4-4-1、表 4-4-2，顏色愈深代表相關程度越高。

表 4-4- 1 LUAD 下免疫基因之間表達的相關程度

	CD4	CD3D	TBX21	GATA3	CD8A	FOXP3	BTN3A1	CD86	MS4A1	CD68	PDCD1	CD274	CTLA4
CD4		0.6009	0.5549	0.5115	0.5072	0.6864	0.5122	0.8649	0.4712	0.7384	0.5526	0.5456	0.5946
CD3D			0.7559	0.5189	0.8178	0.7007	0.5822	0.6001	0.6374	0.4199	0.7861	0.4826	0.7826
TBX21				0.5109	0.8109	0.5986	0.5922	0.5039	0.6140	0.4232	0.7812	0.5556	0.7146
GATA3					0.4902	0.5586	0.4674	0.4423	0.3805	0.3653	0.4989	0.3561	0.4873
CD8A						0.5697	0.5992	0.5515	0.5662	0.4013	0.7811	0.5372	0.7172
FOXP3							0.5090	0.6097	0.5105	0.4403	0.7071	0.4762	0.7572
BTN3A1								0.5216	0.4799	0.4426	0.5287	0.5345	0.5616
CD86									0.3720	0.7758	0.5338	0.6406	0.6116
MS4A1										0.2376	0.5628	0.2623	0.5912
CD68											0.4191	0.5481	0.4090
PDCD1												0.5270	0.7532
CD274													0.5245
CTLA4													

表 4-4- 2 LUSC 下免疫基因之間表達的相關程度

	CD4	CD3D	TBX21	GATA3	CD8A	FOXP3	BTN3A1	CD86	MS4A1	CD68	PDCD1	CD274	CTLA4
CD4		0.7519	0.7179	0.4220	0.6990	0.7925	0.6585	0.8703	0.5954	0.7638	0.7112	0.3609	0.7501
CD3D			0.8564	0.3318	0.9134	0.7677	0.7093	0.7561	0.6775	0.5508	0.8741	0.4040	0.8551
TBX21				0.3628	0.8359	0.7263	0.6300	0.6580	0.6287	0.5065	0.8596	0.3709	0.8050
GATA3					0.2947	0.4313	0.4522	0.3292	0.2239	0.1774	0.3890	(0.1349)	0.3827
CD8A						0.6975	0.7020	0.7047	0.5932	0.5112	0.8676	0.4204	0.8072
FOXP3							0.6195	0.7211	0.5867	0.5420	0.7477	0.3352	0.8469
BTN3A1								0.6621	0.4927	0.4299	0.6962	0.2833	0.6780
CD86									0.5409	0.7567	0.6765	0.3550	0.7489
MS4A1										0.3435	0.6407	0.2948	0.6456
CD68											0.4830	0.3828	0.5267
PDCD1												0.3521	0.8334
CD274													0.3827
CTLA4													

第五節 生存分析

本節欲觀察 CD4、CD3D、TBX21、GATA3、CD8A、FOXP3、BTN3A1、CD86、MS4A1、CD68、PDCD1、CD274、CTLA4 此十三種基因之表達含量高低，分別對肺腺癌及肺鱗癌患者的存活率曲線是否產生明顯差異，因此本研究藉由 Kaplan-Meier 方法估計存活率，並利用 Log rank test 進行檢定。

一 各基因表達高低對肺腺癌的影響

(一) 中位數分割

對所有基因以中位數分割成高表達組和基因低表達組，高於中位數者為高表達組，低於中位數者為低表達組，見表 4-5-1。

表 4-5- 1 各基因之中位數及 P-value

基因變數	CD4	CD3D	TBX21	GATA3	CD8A	FOXP3	BTN3A1	CD86	MS4A1	CD68	PDCD1	CD274	CTLA4
中位數	11.09	7.2865	4.8715	6.201	7.948	6.303	10.09	8.4795	6.8895	12.095	5.8045	5.9895	6.091
P-value	0.0074	0.0459	0.4826	0.3964	0.1265	0.6166	0.4227	0.1211	0.0006	0.9583	0.9125	0.5068	0.1239

1. 使用 Log-rank test 檢測

其中只有 CD4 和 CD3D 和 MS4A1 的 p-value 值小於 0.05，CD4 的 p-value 值為 0.0074，CD3D 的 p-value 值為 0.0459，MS4A1 的 p-value 值為 0.0006，表示在 95%的信心水準下，CD4、CD3D、MS4A1 的基因表現量高低有顯著差異。

(1) CD4 基因表達量低的存活率較低，見圖 4-5-1。

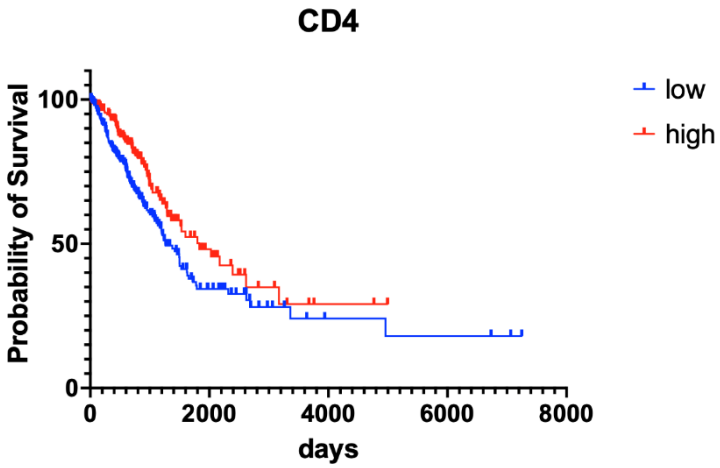

圖 4-5- 1 CD4 存活曲線圖

(2) CD3D 基因表達量低的存活率較低，見圖 4-5-2。

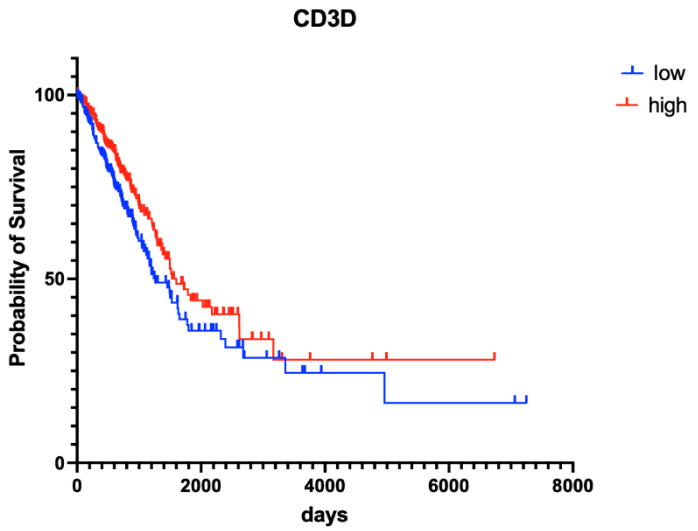

圖 4-5- 2 CD3D 存活曲線圖

(3) MS4A1 基因表達量低的存活率較低，見圖 4-5-3。

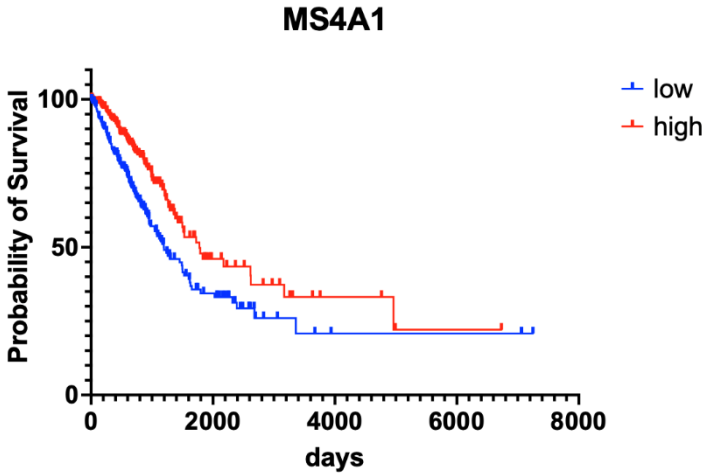

圖 4-5- 3 MS4A1 存活曲線圖

2. 危險比率：

由表 4-5-2 可知，在 CD4 的基因中，A 組（低表達組）比 B 組（高表達組）危險 1.51 倍；在 CD3D 的基因中，A 組（低表達組）比 B 組（高表達組）危險 1.345 倍；在 MS4A1 的基因中，A 組（低表達組）比 B 組（高表達組）危險 1.678 倍。

表 4-5- 2 基因表達具顯著差異之危險比

基因變數	CD4	CD3D	MS4A1
Hazard Ratio（A/B）	1.510	1.345	1.678

(二) 對不顯著的基因進行四分位數分割

對不顯著的基因進一步做四分位數分割，小於第一四分位數者為低表達組，大於第三四分位數者為高表達組，見表 4-5-3。

表 4-5- 3 各基因之第一四分位數、第三四分位數及 P-value

基因變數	TBX21	GATA3	CD8A	FOXP3	BTN3A1	CD86	CD68	PDCD1	CD274	CTLA4
1st	3.9455	5.399	6.9705	5.37525	9.5315	7.78375	11.44	4.90475	4.97425	5.122
3nd	5.8955	7.046	8.926	7.067	10.58	9.083	12.7275	6.79175	6.87775	6.891
P-value	0.1732	0.5588	0.6459	0.7248	0.3287	0.0355	0.8428	0.3920	0.8156	0.0126

1. 使用 Log-rank test 檢測

進一步將其餘不顯著的基因以四分位數分割，分成小於第一四分位數的低表達組和大於第三四分位數的高表達組，其中 CD86 和 CTLA4 的 p-value 值小於 0.05，CD86 的 p-value 值為 0.0355，CTLA4 的 p-value 值為 0.0126，表示在 95%的信心水準下，CD86、 CTLA4 的基因表現高低有顯著差異。

(1) CD86 基因表達量低的存活率較低，見圖 4-5-4。

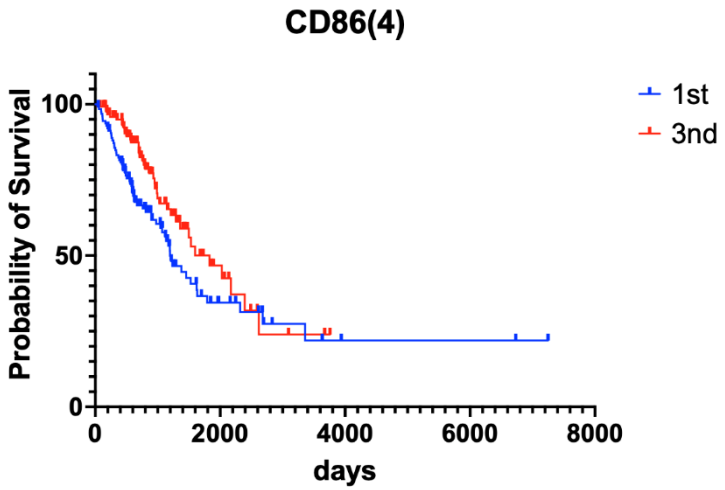

圖 4-5- 4 CD86 存活曲線圖

(2) CTLA4 基因表達量低的存活率較低，見圖 4-5-5。

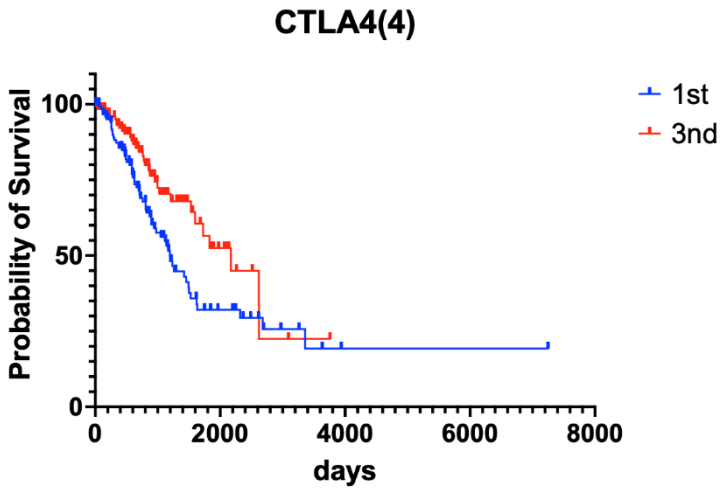

圖 4-5- 5 CTLA4 存活曲線圖

2. 危險比率：

由表 4-5-4 可知，在 CD86 的基因中，A 組（低表達組）比 B 組（高表達組）危險 1.524 倍；在 CTLA4 的基因中，A 組（低表達組）比 B 組（高表達組）危險 1.709 倍。

表 4-5- 4 基因表達具顯著差異之危險比

基因變數	CD86	CTLA4
Hazard Ratio（A/B）	1.524	1.709

二 各基因表達高低對肺鱗癌的影響

(一) 使用中位數進行分割

對所有基因以中位數分割成高表達組和基因低表達組，小於中位數者為低表達組，大於中位數者為高表達組，見表 4-5-5。再使用 Log-rank test 檢測，由表 4-5-5 可知，由於 CD4、CD3D、TBX21、GATA3、CD8A、FOXP3、BTN3A1、CD86、MS4A1、CD68、PDCD1、CD274、CTLA4 的 P-value 在 95%的信心水準下皆大於 0.05，因此各基因在肺鱗癌中的表現量高低對於患者生存時間無顯著差異。

表 4-5- 5 各基因之中位數及 P-value

基因變數	CD4	CD3D	TBX21	GATA3	CD8A	FOXP3	BTN3A1	CD86	MS4A1	CD68	PDCD1	CD274	CTLA4
中位數	10.32	6.6375	4.1735	6.3685	7.6945	6.098	9.515	8.043	6.1065	11.81	5.473	6.2175	5.595
P-value	0.1242	0.7643	0.8445	0.5397	0.5683	0.3203	0.2501	0.5057	0.9666	0.1941	0.8007	0.2884	0.5779

(二) 進行四分位數分割

對不顯著的基因進一步做四分位數分割，小於第一四分位數者為低表達組，大於第三四分位數者為高表達組，見表 4-5-6。再使用 Log-rank test 檢測，由表 4-5-6 可知，即使利用四分位數將其分群，使群間差異更大，但是 CD4、CD3D、TBX21、GATA3、CD8A、FOXP3、BTN3A1、CD86、MS4A1、CD68、PDCD1、CD274、CTLA4 的 P-value 在 95%的信心水準下仍大於 0.05，因此各基因在肺鱗癌中的表現高低對於患者生存時間無顯著差異。

表 4-5- 6 各基因之第一四分位數、第三四分位

基因變數	CD4	CD3D	TBX21	GATA3	CD8A	FOXP3	BTN3A1	CD86	MS4A1	CD68	PDCD1	CD274	CTLA4
1st	9.4485	5.65275	3.241	5.1375	6.7365	5.1455	8.8455	7.258	4.163	11.0425	4.42975	5.23625	4.71025
3rd	11.06	7.68575	5.0525	7.54775	8.7465	7.0235	10.05	8.83575	7.4995	12.4975	6.45175	7.57925	6.61825
P-value	0.2098	0.7070	0.9953	0.8349	0.3742	0.9924	0.6172	0.5428	0.7033	0.3116	0.9203	0.5137	0.3432

三 結論

綜合上述結果，CD4、CD3D、MS4A1、CD86、CTLA4 此五種基因的表達高低對於肺腺癌患者的生存天數有顯著影響，若基因表達低，患者的死亡機率較高；而 CD4、CD3D、TBX21、GATA3、CD8A、FOXP3、BTN3A1、CD86、MS4A1、CD68、PDCD1、CD274、CTLA4 此十三種基因的表達高低對肺鱗癌患者的生存時間皆無顯著影響。

第六節 主成分分析與分群分析

在本研究中，我們將主成分分析及分群分析應用於區分肺腺癌、鱗狀細胞肺癌與健康個體在免疫相關基因表達上的差異，以進一步證實上述研究的這些基因參數具有潛在的應用價值。

一 主成分分析

主成分分析（Principal Component Analysis, PCA）是一種廣泛應用於生物醫學研究的統計方法。透過 PCA，我們可以識別哪些主成分在健康和 NSCLC 個體之間具有較大的變異性。我們運用 R 程式語言進行分析，並根據累積變異量來選擇要保留多少特徵。根據圖 4-6-1，紅線為累積變異量，藍線為個別主成分解釋變異量。在本研究中我們選擇了累積解釋的變異量達到 79.4780% 的前四個主成分 （表 4-6-1），並以其分數（scores）做分群分析（Kaloyanova, 2021b）。

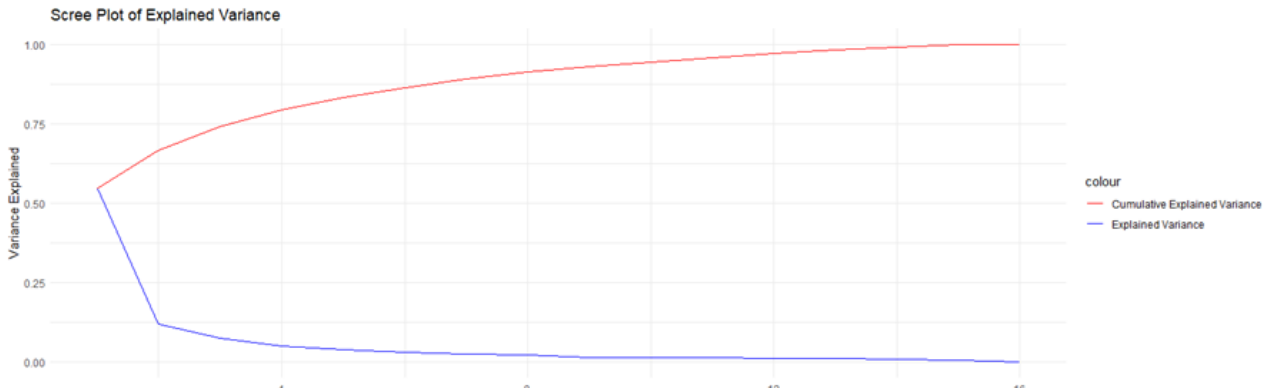

圖 4-6- 1 主成分解釋變異量

表 4-6- 1 累積解釋的變異量

	PC1	PC2	PC3	PC4
累積解釋變異量（%）	54.8626	66.8831	74.3202	79.4780

接下來我們以特徵向量來分析每個變數在主成分的解釋量，而正向負荷量表示該變數的原始資料與主成分分數呈正相關，反之亦然。如（表 4-6-2）可以了解四個主成分中負荷量較大或較小的變數，其中為 PC1 中 CD4、CD3D、TBX21、CD8A、BTN3A1、CD86、PDCD1、CTLA4 都為負向、PC2 中 FOXP3 為正向、CD274 為負向、PC3 中 CD86 為正向、CD68 為正向，最後是 PC4 中 GATA3 為負向。

而其中可以了解到 PC1 的基因類型主要為輔助 T 細胞因子，PC2 的基因類型主要為免疫系統蛋白，PC3 的基因類型主要為抗原細胞蛋白，PC4 的基因類型是轉錄因子。

表 4-6- 2 每個變數在主成分的特徵向量

	PC1	PC2	PC3	PC4
CD4	-0.3171	-0.0966	0.2676	-0.1040
CD3D	-0.3026	0.3113	-0.0978	0.1124
TBX21	-0.3104	-0.2475	-0.2105	0.1206
GATA3	-0.2015	-0.2122	-0.2478	-0.8638
CD8A	-0.3150	0.0504	-0.1959	0.1813
FOXP3	-0.1695	0.6011	0.1189	-0.2055
BTN3A1	-0.2888	-0.2613	-0.1360	-0.0465
CD86	-0.2951	0.0726	0.4607	-0.0598
MS4A1	-0.2583	0.1549	-0.2769	0.1689
CD68	-0.2413	-0.1274	0.6400	-0.0284
PDCD1	-0.3115	0.1520	-0.1750	0.0631
CD274	-0.2223	-0.4724	0.0373	0.3088
CTLA4	-0.3148	0.1392	-0.1034	0.0569
OS	-0.0316	-0.1902	-0.0076	-0.0097
GenderFEMALE	-0.0224	0.0636	0.0211	0.0383
GenderMALE	0.0224	-0.0636	-0.0211	-0.0383

二分群分析模型評估

我們運用 Python 程式語言進行 K-means 與階層式分群，並選擇了前四個主成分分數 PC1、PC2、PC3、PC4 作為分群分析的資料輸入。在本研究中，我們對 K-means 模型採用七種模型評估指標來評估，對階層式分群採用八種模型評估指標來評估，以選擇最合適研究目標的模型。

在準確率方面，所有模型都表現相似，但階層式分 3 群和 K-means 分 3 群略高（表 4-6-3）。基於中位數的 SSE（Sum of Squared Errors）在階層式分 5 群的情況下最低，這意味著數據點更接近其群集中心。Adjusted Rand Index 和 Normalized Mutual Information 在階層式分三群和 K-means 分三群的情況下相對較高，這意味著這些模型在找到實際群集方面表現得相對較好。Fowlkes-Mallows Index 和 Silhouette Method 也在階層式分三群和 K-means 分三群的情況下較

高，這些指標評估了模型及群集的質量。

表 4-6- 3 模型評估指標

	階層式			k – means		
分群數	3	4	5	3	4	5
準確率 (%)	67.3062	67.3062	67.3062	67.3062	64.9271	65.9248
Adjusted Rand Index	0.3469	0.3234	0.2729	0.3453	0.2725	0.2448
Normalized Mutual Info	0.4814	0.4441	0.3929	0.4759	0.3860	0.3525
Fowlkes-Mallows Index	0.5785	0.5483	0.4812	0.5758	0.4987	0.4594
Silhouette Method	0.3404	0.2963	0.2601	0.3485	0.3269	0.2887
Mutual Information	0.5125	0.5163	0.5169	0.5083	0.4665	0.4651
SSE	6076.4105	5104.5312	4259.4984	5861.7121	4537.3821	3952.2182
Cophenetic Coefficient	0.6148	0.6148	0.6148			

本研究中，如果每個變數都自成一組，那麼分群實際上失去了其意義。因此依據評估指標的結果，我們決定使用模型評估指標較好的階層式分群並分三群為分群分析的模型。

三 免疫相關基因及其他因素區分肺鱗癌、肺腺癌和健康個體

我們採用了階層式分群方法將前四項 PCA scores 分為三群。交叉分析表顯示肺鱗癌、肺腺癌和健康個體分別在每群的樣本數（表 4-6-4）。第一群主要由肺鱗癌（LUSC）樣本組成，佔該群總數的 59.23%。第二群主要由肺腺癌（LUAD）樣本組成，佔該群總數的 57.74%。第三群則幾乎完全由健康組織（Normal Tissue）樣本組成，佔該群總數的 98.6159% （表 4-6-5）。綜合以上結果，本模型具有 67.3062%的準確率能夠區分肺腺癌、肺鱗癌和健康組織，尤其辨識健康個體的準確性非常高。

表 4-6- 4 交叉分析表

分群數/類別	Cluster 1	Cluster 2	Cluster 3
LUAD	179	332	3
LUSC	260	241	1
Normal Tissue	0	2	285

表 4-6- 5 每群主要類別佔比

分群	主要類別	比例
Cluster 1	LUSC	59.2255 %
Cluster 2	LUAD	57.7391 %
Cluster 3	Normal Tissue	98.6159 %

四探討前四項主成份在三個分群的表達差異

在 PC1 中，Cluster 1（LUSC）及 Cluster 2（LUAD）的平均值顯著高於健康群集，其中 Cluster 1（LUSC）又高於 Cluster2（LUAD），反映了 PC1 基因的平均表達率在肺腺癌及肺鱗癌中的低活性或抑制狀態 （圖 4-6-2）。

在 PC2 中，Cluster 1（LUSC）及 Cluster 2（LUAD）的平均值也高於健康群集，其中 Cluster2（LUAD）又高於 Cluster 1（LUSC）（表 4-6-6）。推測非小細胞肺癌在 FOXP3 會有較高的表達，在 CD274 會有較低的表達。

而在 PC3 及 PC4 中肺癌患者與健康組織平均主成分成績相差較小。

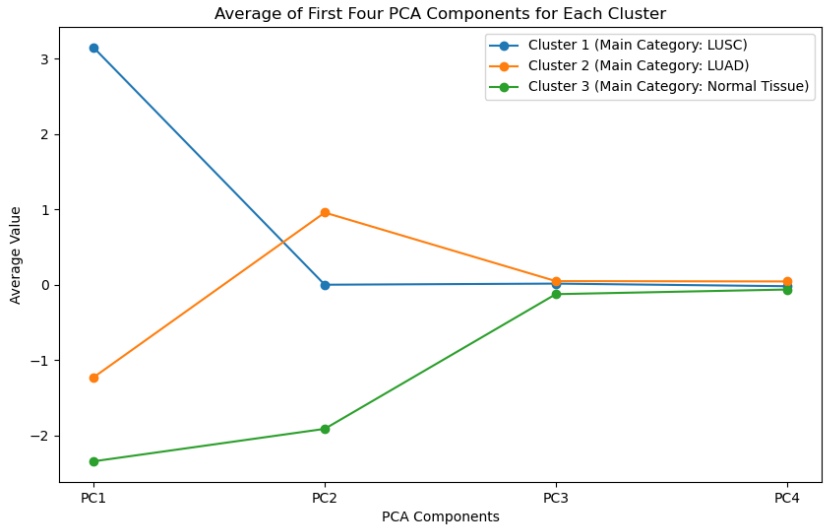

圖 4-6- 2 前四項主成分分析每群平均值

表 4-6- 6 前四項主成分分析每群平均值

	PC1	PC2	PC3	PC4
Cluster 1 (LUSC)	3.1481	0.0016	0.0164	-0.0182
Cluster 2 (LUAD)	-1.2265	0.9592	0.0496	0.0451
Cluster 3 (Normal Tissue)	-2.3417	-1.9108	-0.1236	-0.0620

五 綜合分析

我們以（圖 4-6-3）來驗證第肆章結果分析中的前四節分析結果，以表示本研究的可信度：

(一) 在性別獨立性檢定結果中，性別與是否得肺癌沒有相關，而分群分析中三個分群都以男性居多，可以推測分群分析中的性別沒有參考意義。

(二) 驗證 U-Test 分析結果：

1. FOXP3 在肺癌患者中平均基因表達較健康組織高。

2. CD4、CD3D、TBX21、GATA3、CD8A、BTN3A1、CD68、CD274、CTLA4 中健康組織平均基因表達較肺癌患者高。

3. CD4、CD3D、TBX21、CD8A、FOXP3、BTN3A1、CD86、MS4A1、CD68、PDCD1、CTLA4 平均基因表達率肺腺癌表達高於肺鱗癌。

4. 與 U-Test 結果不相符：

CD86、MS4A1、PDCD1 在健康組織平均表達率較肺癌患者低，CD274、GATA3 在肺腺癌表達較肺鱗癌高。結果的不相符可能需再進一步設計實驗分析，或是獲取更多的樣本數。

(三) 驗證 Spearman 等級相關：

相關係數越接近 1 或是-1，代表其線性關係越明確，我們可以利用 R 值來預估 PCA 的結果，同一群的變數他們的 R 值會比較接近 1 或是-1，而：

1. CTLA4 與 CD3D、TBX21、CD8A、FOXP3、PDCD1

2. PDCD1 與 CD3D、TBX21、CD8A、FOXP3

3. CD68 與 CD4、CD86

4. CD86 與 CD4

5. CD8A 與 CD3D、TBX21

6. CD3D 與 TBX21

在肺腺癌和肺鱗癌的相關性都為高度正相關，所以進一步利用 PCA 來分析，是否這些高度相關的免疫細胞基因，在 PCA 分析後也是在同一群，結果顯示這些高度相關的免疫細胞基因幾乎都在 PC1 中。

(四) Cluster 1（LUSC）及 Cluster 2（LUAD）都以死亡（OS = 0）居多，而 Cluster 3（Normal Tissue）則都存活。

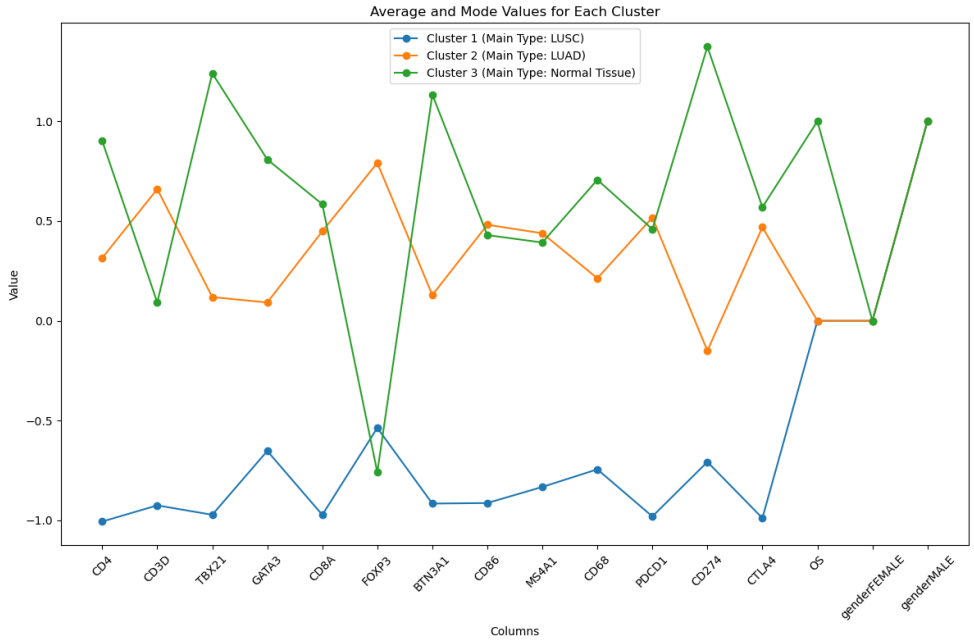

圖 4-6- 3 分群結果對應標準化原始資料

表 4-6- 7 標準化原始資料分群結果表

	Cluster 1 (LUSC)	Cluster 2 (LUAD)	Cluster 3 (Normal Tissue)
CD4	-1.0057	0.3151	0.9007
CD3D	-0.9248	0.6599	0.0919
TBX21	-0.9718	0.1188	1.2398
GATA3	-0.6521	0.0916	0.8083
CD8A	-0.9721	0.4485	0.5843
FOXP3	-0.5364	0.7911	-0.7592
BTN3A1	-0.9157	0.1292	1.1339
CD86	-0.9130	0.4812	0.4294

MS4A1	-0.8324	0.4384	0.3922
CD68	-0.7443	0.2131	0.7066
PDCD1	-0.9802	0.5177	0.4589
CD274	-0.7070	-0.1504	1.3732
CTLA4	-0.9885	0.4687	0.5689
OS	0	0	1
genderFEMALE	0	0	0
genderMALE	1	1	1

六 總結

未來檢測非小細胞肺癌的趨勢將會利用免疫基因來直接識別，本研究的模型已達到 67.3062%的準確度識別肺腺癌與肺鱗癌，而識別健康個體的準確度達到 98.6159%。因肺腺癌及肺鱗癌早期症狀並不明顯，期望能夠以免疫基因檢測可能確診肺鱗癌及肺腺癌的高危險族群，並且能達到提早偵測，提早治療的效果，以提高存活率。

持續監測患者的基因免疫表達率可以幫助調整治療方案，以達到最佳效果，而在有條件下，肺腺癌及肺鱗癌確診者可以直接利用對應的免疫抑制劑治療。根據以上綜合分析結果，肺鱗癌患者在 13 種免疫基因的表達率都偏低，現階段在臨床試驗上沒有提高病患基因表達的實驗與法規，所以目前本研究中的基因變數可能不適合肺鱗癌進行免疫治療，未來將會尋找合適治療肺鱗癌的免疫基因。對於肺腺癌患者，可以進一步分析 FOXP3 是否可以為抑制劑為肺腺癌患者做免疫治療，以增強抗癌或抗病毒的免疫反應（昶安科技, 2023）。

第伍章 結語與討論

第一節 結論

本研究使用肺腺癌、肺鱗癌患者與正常人的性別、年齡、抽菸史及免疫細胞基因之資料，經過卡方檢定後發現性別和年齡會影響肺癌種類，抽菸史不會影響肺癌種類的不同；使用 Kruskal-Wallis 和 Mann-Whitney U 檢定後得到除了 FOXP3 以外的 12 個基因都在正常人的表現較高，其中 GATA3 和 FOXP3 在肺腺癌與肺鱗癌的基因表達沒有顯著差異，肺鱗癌基因表現高於肺腺癌的為 CD274，反之則為其餘基因；生存分析的檢測結果顯示，大多基因表達高低不會影響本研究樣本的生存天數，但 CD4、CD3D、MS4A1、CD86、CTLA4 的基因表達高低會顯著影響肺腺癌患者的生存天數，只要這些基因表現越高，病人的生存機率就會越高，反之則越低，而這些基因位於免疫細胞上，表示病人若能提高自身免疫力，生存的機會就越高。

最後利用主成分分析進行降維，再使用 K-means 與階層式做分群。主成分分析結果顯示，取前四個主成分能達到近八成的解釋變異量，定義 PC1 為輔助 T 細胞因子，PC2 為免疫系統蛋白，PC3 為抗原細胞蛋白，PC4 為轉錄因子，並將主成分分析結果作為分群分析資料，在不超過變數總數且評估指數最好的情況下，選擇階層式並分三群建立最終分群模型，該模型的準確率有 67%，其中第一群多為肺鱗癌患者，第二群為肺腺癌患者，第三群為正常人。本研究可以分析出肺腺癌、肺鱗癌與正常人的群體，但是肺腺癌與肺鱗癌的準確率只有近六成，而正常人可以達到 98%的準確率，這些結果顯示，目前本研究所分析的基因變數可以順利區分出正常人與肺癌患者，如果再細分肺癌種類就需要更多基因變數分析。

現今肺癌治療仍以侵入性治療為最大宗，如手術切除、化學治療、放射治療、標靶治療，透過藥物、放射線等等方式消除癌細胞，但是癌症復發的機率高，問題無法得到根本的解決；免疫檢查點抑制

劑是免疫治療其中一種方式，藉由免疫治療靶點調控自身 T 細胞功能，增強對腫瘤細胞攻擊的效果，此外免疫抑制劑治療後產生的副作用比化療低且能顯著提高肺癌病人的整體生存率，所以免疫治療也將成為未來癌症治療之趨勢 （施穎銘, 洪淑涝, & 張純琪, 2017）。針對免疫治療部分並綜合研究結果，因為 FOXP3 在肺癌患者較正常人有顯著高表達，只要 FOXP3 基因表達越高，PC2 的結果也越高，因此 FOXP3 預計可以成為肺癌的免疫治療靶點。而 PC2 的分群結果定義為肺腺癌患者，推測肺腺癌患者適合利用 FOXP3 做免疫抑制劑治療。由於第二群的準確率只有近六成，另有四成的肺鱗癌患者在本研究的基因變數中無法更準確的分離，因此本研究目前無法確認潛在的肺鱗癌患者是否適合利用 FOXP3 進行免疫治療。

雖然免疫治療使得癌症患者存活率提高，可是仍有身體排斥、治療無效及嚴重副作用的風險，現今有三種免疫治療方法，CAR-T 細胞治療、自體免疫細胞治療及免疫檢查點抑制劑，但是免疫細胞治療在臨床上沒有對照的試驗與法規，而且三者皆需耗費大量的時間與金錢，因此醫學界仍致力於找出對人體最有效的治療方式。「預防勝於治療」比起確診後治療不如及早預防，有研究指出保持運動習慣可以提高自身免疫力降低癌症發生的風險（Bigley, Spielmann, LaVoy, & Simpson, 2013），持續運動並維持良好的生活習慣，避免接觸空氣汙染環境、定期做健康篩檢才是預防肺癌的最佳辦法。

第二節 未來展望

由於本研究只選取 13 個免疫細胞基因作為分群模型的變數，所以此模型在判別肺腺癌與肺鱗癌上準確率不高也無法找出肺腺癌免疫治療靶點，後續研究可以增加免疫細胞基因變數以提高其準確率，找出最適配的分群模型及合適治療肺鱗癌的免疫基因。

本研究正常人資料中有許多遺失值及不存在之變數，因此無法判斷除了免疫細胞基因與性別外，是否存在其他變數影響肺腺癌、肺鱗癌與正常人的不同，後續研究可以找出正常人資料，如年齡、抽菸情況等等，才能進行更全面且完整的分析。

參考資料

中文參考資料

Dr.Fish 漫遊社會統計（2023）.曼－惠特尼 U 檢定.

Dr.Fish 漫遊社會統計（2023） .卡方獨立性的假設檢定

Dr.Fish 漫遊社會統計（2023） Kruskal – Wallis 檢定的假設檢定

Hsuan（2018） "Hierarchical Clustering 階層式分群 | Clustering 資料分群 | R 統計"

王超辰（2023）醫學統計學

王建凱（2018）K-means

心血管保健諮詢網. "肺腺癌-初期症狀及第四期存活率."

李聰亮, 周宛儀, & 黃建仁. （2016）. 自體免疫疾病之致病機轉. 臺灣醫界雜誌, *59:3 期*, 19–25.

吳雅琪.（2011）."臨床試驗樣本數計算簡介."當代醫藥法規月刊 RegMed, 5.

馬偕紀念醫院."肺癌可分成那幾類型."

昶安科技 （2023）."以免疫檢查點為目標的癌症治療."

施穎銘, 洪淑涝, & 張純琪. （2017）. 非小細胞肺癌的免疫治療. 內科學誌, 28, 271-278.

連珮妤. （2023）."鱗狀上皮細胞肺癌是什麼？存活率、初期症狀、治療與原因."

黃文彥（2021）R 統計

鄒佩玲 and 吳昌俊 （2013）."美國癌症基因體圖譜計畫 TCGA（The Cancer Genome Atlas）簡介." 內科學誌 24（1）: 43-47.

臺大醫院健康電子報 （2015）."華人肺腺癌高風險遺傳基因異常."

廖佳馨（2023） "失智症身心障礙者之障礙等級轉換及住院醫療成本"

衛生福利部國民健康署（2019）."肺癌危險因子." 肺癌家族病史.

醫學網（2020）."肺大細胞癌：症狀、病因及如何治療."

英文參考資料

Barta, J. A., et al. （2019）. "Global epidemiology of lung cancer." Annals of global health 85（1）.

Bigley, A. B., Spielmann, G., LaVoy, E. C., & Simpson, R. J. （2013）. Can exercise-related improvements in immunity influence cancer prevention and prognosis in the elderly? Maturitas, 76（1）, 51-56.

Bolandi, Nadia et al. “The Positive and Negative Immunoregulatory Role of B7 Family: Promising Novel Targets in Gastric Cancer Treatment.” International journal of molecular sciences vol. 22,19 10719. 3 Oct. 2021, doi:10.3390/ijms221910719. PubMed.

Clapp, R. W., et al. （2008）. "Environmental and occupational causes of cancer: new evidence 2005-2007." Reviews on environmental health **23**（1）: 1-38.

Control, C. f. D. and Prevention （2001）. "State-specific prevalence of current cigarette smoking among adults, and policies and attitudes about secondhand smoke--United States, 2000." MMWR. Morbidity and mortality weekly report **50**（49）: 1101-1106.

Dariavach P, Mattéi MG, Golstein P, Lefranc MP （December 1988）. "Human Ig superfamily CTLA-4 gene: chromosomal localization and identity of protein sequence between murine and human CTLA-4 cytoplasmic domains". European Journal of Immunology. 18（12）: 1901–5.

Gaglia J、Kissler S（2019）”Anti-CD3 Antibody for the Prevention of Type 1 Diabetes: A Story of Perseverance”. Biochemistry.58（40）:4107-4111.

Haybar H, Rezaeeyan H, Shahjahani M, Shirzad R, Saki N （June 2019）. "T-bet transcription factor in cardiovascular disease: Attenuation or inflammation factor?". *Journal of Cellular Physiology*. **234**（6）: 7915–7922.

Hardy R （2008）. "Chapter 7: B Lymphocyte Development and Biology". In Paul W （ed.）. *Fundamental Immunology*（Book）（6th ed.）.

Huang, J., et al. （2022）. "Distribution, risk factors, and temporal trends for lung cancer incidence and mortality: a global analysis." Chest 161（4）: 1101-1111.

Jaakkola, M. and J. Jaakkola （2006）. "Impact of smoke-free workplace legislation on exposures and health: possibilities for prevention." European Respiratory Journal **28**（2）: 397-408.

Janeway, Charles. Immunobiology: the immune system in health and disease 5th ed. New York: Garland Pub. 2001

Kaloyanova, E. （2021）. "How to Combine PCA and K-means Clustering in Python?".

Kyle K. Payne *et al,*. BTN3A1 governs antitumor responses by coordinating αβ and γδ T cells.*Science***369**,942-949（2020）.

LIM, W. Y. and A. Seow （2012）. "Biomass fuels and lung cancer." Respirology **17**（1）: 20-31.

Leahy DJ, Axel R, Hendrickson WA. Crystal structure of a soluble form of the human T cell coreceptor CD8 at 2.6 A resolution. Cell. March 1992, **68**（6）: 1145–62.

Loftus, Peter （16 Nov 2014）. "New Bristol-Myers Drug Helped Skin-Cancer Patients in Trial Live Longer". Wall Street Journal. Retrieved 24 Nov 2014.

Mann, H.B. and Whitney, D.R. （1947） On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. Annals of Mathematical Statistics, 18, 50-60.

Murray, J. F. and R. J. Mason （2016）. "Murray & Nadel's textbook of respiratory medicine." （No Title）.

Murphy, Kenneth. Janeway's Immunobiology 8th. New York: Garland Science. 2012.

Naylor MJ, Ormandy CJ （2007）. "Gata-3 and mammary cell fate". *Breast Cancer Research*. **9** （2）: 302.

Parkin, D. （2011）. "2. Tobacco-attributable cancer burden in the UK in 2010." British journal of cancer **105**（2）: S6-S13.

Ryu SE, Truneh A, Sweet RW, Hendrickson WA. Structures of an HIV and MHC binding fragment from human CD4 as refined in two crystal lattices. Structure. January 1994, **2** （1）: 59–74.

Tobias, J. S. and D. Hochhauser （2009）. Cancer and its management, John Wiley & Sons.

Wu, Z., et al. （Nov 28, 2017）. "Expression and significance of CD28, CTLA-4, CD80 and CD86 in gastric cancer."