

國立臺灣大學生物資源暨農學院生物機電工程學系

碩士論文

Department of Biomechatronics Engineering

College of Bioresources and Agriculture

National Taiwan University

Master Thesis

評估 ezGeno 在分析轉錄因子結合特徵之表現並應用
於跨細胞株的比較研究

Evaluating the performance of ezGeno in analyzing
transcription factor binding profiles and applying it to a
comparative study across cell types

張淘喻

Tao-Yu Zhang

指導教授：陳倩瑜 博士

Advisor: Chien-Yu Chen, Ph.D.

中華民國 111 年 7 月

July 2022




誌謝

時光飛逝，轉眼間研究生生活即將結束，回憶兩年前，因緣際會考上台大生機所碩士班並加入倩瑜老師的 c4Lab 實驗室，這兩年的求學期間學到了很多東西，首先，非常感謝倩瑜老師這兩年來的用心指導，提供了很多的研究建議與方向，使得本論文能順利完成，經過碩士班這兩年的學習，讓我找到了自己未來的目標。

感謝實驗室的學長們文策、東祈、欽祥、昀翔與弘曄，時常給予建議與經驗分享，也感謝實驗室同學又瑋、如秀一起修課與討論，與我互相砥礪，還有學弟名翔、毓聰、伯豪，謝謝實驗室的各位，讓我在研究生涯中增添了許多歡樂與回憶。

最後，感謝家人們與宏聖一路上的支持與鼓勵，讓我可以不用擔心生活，專心完成研究。

中文摘要



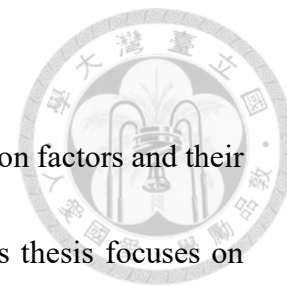
在基因表現的相關研究議題中，轉錄因子及其結合位的交互作用關係一直受到很大的關注，一直以來，轉錄因子如何辨識基因體中特定結合位置並與之結合，進而調控後續基因表現，最終影響生物行為，是生物資訊學者想了解的重要問題。本論文著重於研究轉錄因子於不同細胞株間的結合位差異，藉由收集不同細胞株中，數個轉錄因子之染色質免疫沉澱定序資料，透過深度學習工具進行分析，但由於對不同轉錄因子而言，所適合的深度網路模型並不相同，因此本研究使用自動機器學習工具 ezGeno，加速建立不同轉錄因子在不同細胞株之預測模型，並將訓練後之模型應用於尋找可能影響轉錄因子結合之變異位點。

本論文使用本實驗室與台灣人工智慧實驗室合作開發之 ezGeno，該工具先以自動機器學習的方式去挑選適合的卷積神經網路模型後，再進行轉錄因子結合位的預測。本研究主要使用 ENCODE 資料庫中的染色體免疫沉澱定序資料進行分析，為了評估 ezGeno 在學習時所需的最適正樣本數目，本研究從資料庫裡蒐集兩種資料集，第一種為隨機挑選 K562 細胞株的 10 個轉錄因子，第二種則為 2 種轉錄因子於五種細胞株，皆分別取出不同峰值數量作為正樣本，並固定測試資料正樣本數目，由實驗結果發現，當正樣本數目高於 1000 時，預測表現會趨於穩定。另一方面，針對跨細胞株之轉錄因子結合分析，本研究使用資料庫中，五種最常見的細胞株之 24 種轉錄因子，分別取出相同數量作為正樣本，進行預測準確度分析，本研究將 ezGeno 認為重要的序列片段，利用 MEME 工具進行序列特徵分析並與 JASPAR 資料庫中做對照，發現除了主要結合序列特徵以外，模型也學到一些額外的特徵。此外，分析後發現使用相同資料所建構模型具有穩定性，而不同轉錄因子或不同細胞株間模型會因結合特性差異而造成建構模型有所不同。最後，本研究將建好的預測模型應用於預測單核苷酸變異位點對轉錄因子結合所造成的影響，變異位點資料分別為胸、肝及肺組織，藉由設定不同 p-value 之閾值分析，於三種組織中，正樣本中具顯著性的變異位點數量皆多於負樣本，

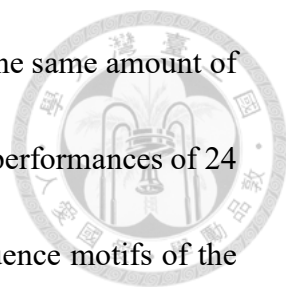
顯示本研究所建立的預測模型在未來應用於尋找可能影響轉錄因子結合之變異位點具可行性。總結，本研究利用自動機器學習工具 ezGeno 有效建立轉錄因子於不同細胞株之結合位預測模型，大幅加速深度學習在基因轉錄調控相關研究之應用。

關鍵字:轉錄因子結合位、自動機器學習、結合特異性、單核苷酸變異位點

英文摘要



In gene expression studies, the interactions between transcription factors and their binding sites have been of great interest to bioinformaticians. This thesis focuses on comparing binding behaviors of transcription factors between different cell lines by collecting chromatin immunoprecipitation sequencing data of several transcription factors in different cell lines and analyzing them with deep learning models. However, the property of different TFs requiring sophisticated network architecture tuning to achieve satisfied performance complicates the situation. For this reason, an AutoML tool, ezGeno, was used to construct models for predicting binding specificity. Finally, the prediction models were used to analyze the effect of sequence variations on transcription factor binding. This thesis uses ezGeno to automatically build deep CNN models for predicting TF binding sites. The chromatin immunoprecipitation sequencing (ChIP-seq) data is downloaded from the ENCODE database for analysis. To evaluate the performance of ezGeno, we randomly selected 10 TFs from K562 cell line as the first dataset and 2 TFs from 5 cell lines as the second, and then extracted different numbers of peaks to build and test the models, respectively. We found that using more than a certain number of positive samples is sufficient to obtain satisfied prediction performance, even though we observed that the larger number of sequences predicted the better slightly. For the study of cross-cell type comparison, we further downloaded



the ChIP-seq data of 24 TFs from five primary cell types and used the same amount of data as positive samples for prediction. We analyzed the prediction performances of 24 TFs in five primary cell types and used MEME to analyze the sequence motifs of the subsequences highlighted by ezGeno and compare them with the JASPAR database. In addition, we found that the model architectures selected by ezGeno is usually stable, while the models differed among transcription factors or cell lines due to differences in binding characteristics. Finally, the prediction model was applied to predicting the effect of single nucleotide variants on binding. The variants that affect gene expression in breast, liver and lung were used in this study. Paired sample t-test (two-tailed) was used to calculate the significance (p-values) between reference and alternative sequences. In these tissues, the number of significant variants in the positive variant list was higher than the negative one, indicating the feasibility of this analysis method. In the future, the models can be used to identify variants causing abnormal binding of transcription factors and thus affecting gene expression. In summary, this study demonstrates that ezGeno can accelerate model construction of TF binding to largely facilitate the study of transcription factor binding upon sequence variants.

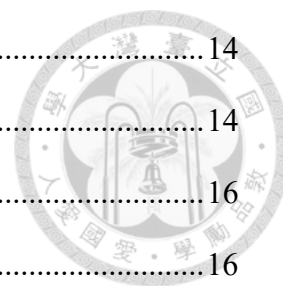
Keyword: Transcription factor binding site, Automated Machine Learning, Binding specificity, Single Nucleotide Variant

目錄



誌謝.....	i
中文摘要.....	ii
英文摘要.....	iv
目錄.....	vi
圖目錄.....	viii
表目錄.....	x
第一章 研究目的.....	1
第二章 文獻探討.....	2
2.1 分子生物學中心法則.....	2
2.2 染色質免疫沉澱定序技術.....	2
2.3 轉錄因子與轉錄因子結合位.....	2
2.4 細胞株與細胞系 (Cell strain and cell line).....	3
2.5 自動化機器學習.....	3
2.5.1 高效神經網路架構搜索.....	4
2.5.2 AutoKeras.....	4
第三章 研究方法.....	5
3.1 實驗資料.....	5
3.1.1 ENCODE 資料庫.....	5
3.1.2 各細胞株轉錄因子數目.....	5
3.1.3 ChIP-seq 實驗資料收集.....	6
3.1.4 變異位點資料.....	11
3.1.5 資料前處理.....	12
3.2 實驗流程.....	12
3.2.1 ezGeno 概述.....	13

3.2.2 ezGeno 的使用	14
3.2.3 結果分析.....	14
第四章 結果與討論.....	16
4.1 ezGeno 訓練資料的峰值數目對模型準確度的影響	16
4.2 跨細胞株間 ezGeno 預測轉錄因子結合位的表現差異	18
4.3 序列特徵分析.....	23
4.4 模型架構的挑選分析.....	24
4.5 變異對基因表達所造成的影響.....	28
第五章 結論.....	32
參考文獻.....	34
附錄 A.....	36
附錄 B.....	49
附錄 C.....	59



圖目錄

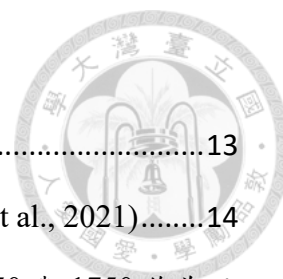


圖 3-1 實驗流程圖	13
圖 3-2 為 ezGeno 的工作流程圖，本圖改自 ezGeno 文獻(Lin et al., 2021).....	14
圖 4-1 針對 K562 中 10 種轉錄因子，分別使用 250、750、1250 與 1750 作為正 樣本的訓練資料的峰值數目對模型準確度之影響，括號內為訓練與測試資 料比例.....	16
圖 4-2 針對 2 種轉錄因子於五種細胞株中，分別使用 250、750、1250 與 1750 作為正樣本的訓練資料的峰值數目對模型準確度之影響，括號內為訓練與 測試資料比例.....	17
圖 4-3 分別使用 1750、3750、5750、7750 與 9750 作為正樣本的訓練資料的峰 值數目對模型準確度之影響(轉錄因子為 SIN3A) ，括號內為訓練與測試資 料比例.....	18
圖 4-4 分別使用 1750、3750、5750、7750、9750、19750、29750、39750 與 49750 作為正樣本的訓練資料峰值數目對模型準確度之影響，括號內為訓 練與測試資料比例.....	18
圖 4-5 五種細胞株中不同轉錄因子之峰值數量熱圖	19
圖 4-6 模型準確度之分群熱圖，固定訓練資料中正樣本數為 1750，測試資料正 樣本數為 250.....	20
圖 4-7 使用轉錄因子 RAD21 與 MAFK 於五種細胞株的前、末 1750 峰值作為 訓練正樣本的預測準確度.....	20
圖 4-8 ezGeno 針對 MCF-7 中轉錄因子 RAD21 所建構之模型.....	25
圖 4-9 ezGeno 針對 GM12878 中轉錄因子 GABPA 所建構之模型	26
圖 4-10 轉錄因子 ELF1 於五種細胞株所建模型之 UMAP 圖.....	26
圖 4-11 轉錄因子 ELK1 於五種細胞株所建模型之 UMAP 圖.....	27
圖 4-12 轉錄因子 JUND 於五種細胞株所建模型之 UMAP 圖	27

圖 4-13 隨機選取六種轉錄因子之 UMAP 圖.....	28
圖 4-14 肺組織中的變異位點以顯著性 $p\text{-value}<0.005$ 標註(第一部分).....	30
圖 4-15 肺組織中的變異位點以顯著性 $p\text{-value}<0.005$ 標註(第二部分).....	31



表目錄



表 3-1 ENCODE 資料庫中六種細胞株之轉錄因子實驗數量統計	6
表 3-2 ENCODE 資料庫中，常見六種細胞株的轉錄因子交集數量	6
表 3-3 K562 細胞株中 10 個轉錄因子的峰值數量	7
表 3-4 不同轉錄因子跨細胞株的數量統計	8
表 3-5 24 個轉錄因子跨細胞株的峰值序列數量	8
表 3-6 GTEx 資料庫中，三種組織的正、負樣本的變異位點數量	11
表 4-1 不同峰值數量間對預測表現的顯著性差異	17
表 4-2 JASPAR 資料庫中 24 種轉錄因子的資料	21
表 4-3 訓練資料正樣本為 1750 個峰值的預測模型應用於不同序列資料之結果	23
表 4-4 三種組織中，對結合預測有影響之變異數量與括號中為平均佔比	29
附表 A-1 十種轉錄因子於 K562 中，使用不同正樣本峰值數目的 ezGeno 預測 準確度.....	36
附表 A-2 兩種轉錄因子於五種細胞株中，使用不同正樣本峰值數目的 ezGeno 預測準確度.....	40
附表 A-3 轉錄因子 SIN3A 於五種細胞株中，使用不同正樣本峰值數目的 ezGeno 預測準確度	44
附表 A-4 RAD21(MCF-7)、CEBPB(HepG2)、MAFK(HepG2)和 MAFK(A549)， 使用不同正樣本峰值數目的 ezGeno 預測準確度	46
附表 B-1 二十四種轉錄因子在五種細胞株中，使用 2000 正樣本峰值數目 1750 作為訓練資料的 ezGeno 預測準確度	49
附表 B-2 兩種轉錄因子在五種細胞株中，使用 1750 末位峰值作為訓練正樣本 的 ezGeno 預測準確度	58

附表 C-1 胸腺組織中的變異位點對 MCF-7 相關模型於結合預測影響之序列數量.....	59
附表 C-2 肝組織中的變異位點對 HepG2 相關模型於結合預測影響之序列數量.....	61
附表 C-3 肺組織中的變異位點對 A549 相關模型於結合預測影響之序列數量.....	63



第一章 研究目的

隨著許多生物科學家的努力，基因表現相關研究蓬勃發展與資料量增加，進而推動相關研究往大數據分析方向前進，機器學習的分析工具如雨後春筍般一一問世，其中基因表達研究中轉錄因子及其結合位的調控關係受到很大的關注，一直以來，轉錄因子如何辨識其特定結合位與之結合進而有後續基因表現、若結合位序列中出現變異是否會對後續基因表達造成影響甚至可能致病，都是生物資訊研究的重要課題。

近年來，在深度學習技術的快速發展下，不僅於電腦視覺、語音辨識和自然語言處理已有相關應用，在生物資訊學領域也發展快速，開發出經由大量染色質數據預測轉錄因子結合位的技術，例如 DeepBind (Alipanahi et al., 2015)，但過於簡單的模型架構限制了 DeepBind 預測的準確度，而要如何改良模型架構是一問題，且每種轉錄因子適合的訓練模型又各不相同，而自動機器學習工具 ezGeno (Lin et al., 2021)能夠解決這個問題，該工具針對不同資料的特性以高效神經網路架構搜索去建構適合的深度學習之卷積神經網路模型，進而進行轉錄因子結合位的預測。

本研究的主要目的為使用自動機器學習工具 ezGeno 針對跨細胞株間轉錄因子結合特異性分析，選擇 ENCODE 資料庫(Consortium, 2012)中資料數量相對較多的幾種細胞株進行實驗並建立預測模型，藉由分析預測準確度、模型架構之穩定度、不同轉錄因子或不同細胞株間模型差異試圖解釋結合特異性問題。隨後將建好的模型應用於表達定量性狀遺傳位點(Expression quantitative trait loci，簡稱 eQTL)資料(Avsec et al., 2021)，預測單核苷酸變異位點對轉錄因子結合所造成的影響，以所建立之預測模型分別預測參考序列和變異序列上之轉錄因子結合機率，使用成對樣本 t 檢定(雙尾)計算兩者間的顯著性 p 值，藉由設定不同 p 值之閾值去分析變異位點對後續基因表達所造成的影響。

第二章 文獻探討



本章節分為五部分，首先簡述分子生物學中心法則，第二、三部分分別簡單介紹染色質免疫沉澱技術與介紹轉錄因子與轉錄因子結合位，第四部分介紹本論文所使用的細胞株資訊，第五部分則介紹自動機器學習工具 ezGeno 相關的知識。

2.1 分子生物學中心法則

分子生物學中心法(Crick, 1970)則是由英國生物物理學家 Francis Crick 於 1957 年提出的假說，認為 DNA 會將訊息傳遞給 RNA，而 RNA 再將訊息傳遞給蛋白質。隨後，於 1970 年，Francis Crick 做了補充與修改，生物遺傳訊息的傳遞為一種 DNA 自我複製的過程，透過複製的過程藉以保留遺傳物質。

2.2 染色質免疫沉澱定序技術

染色質免疫沉澱定序(Chromatin immunoprecipitation sequencing, 簡稱 ChIP-seq) (Park, 2009)為用於分析染色質與相結合的調控因子相互作用的關係。實驗過程先加入甲醛以固定當下蛋白質與 DNA 結合情況，接著透過超聲波或核酸酶的處理將染色體片段化，再利用欲研究特定蛋白質之抗體，抓取與該蛋白質結合的基因片段，完成沉澱步驟，最後移除該蛋白質與所加之抗體，留下 DNA 片段，回貼至參考基因組即完成實驗。

2.3 轉錄因子與轉錄因子結合位

轉錄因子(Transcription factor, 簡稱 TF)是能與特異 DNA 序列結合的蛋白質，藉以調控其基因的表達。此類蛋白質一般具有不同的功能區域，如 DNA 結合結構域和效應結構域，使轉錄因子不單能夠與基因結合，也可以和其它轉錄因子形成轉錄因子複合體影響基因的轉錄。

轉錄因子結合位(Transcription factor binding site, 簡稱 TFBS)是指轉錄因子與 DNA 序列發生結合的位置，TFBS 的長度通常在 5~20bp，轉錄因子通過氫鍵和凡德瓦力與這些區域產生化學性結合，由於這些化學性相互作用的特性，使大

多數轉錄因子都能與個別 DNA 序列發生特異性結合。然而，轉錄因子結合位上的鹼基並非一定都與轉錄因子有接觸，所以產生的結合力大小有所不同。因此，轉錄因子並非只與單一 DNA 序列相結合，而是能夠與類似的序列結合。

2.4 細胞株與細胞系(Cell strain and cell line)

細胞株是由選擇法(Cell selection)或克隆形成法(Clonal Selection)從原代培養物或細胞系中取得特殊性質或標誌物的培養物，細胞株為使用單細胞分離培養或通過篩選，由單細胞增殖形成的細胞群，因此，在培養過程中其特殊性質或標誌必須保持存在作為依據。原代培養物經首次傳代成功後即為細胞系(Cell line)(McGahon et al., 1995)，由原先存在於原代培養物中的細胞世系所組成，細胞系的特點是無限的壽命、穩定的表型、高可用性與容易處理。如果不能繼續傳代培養，或傳代次數有限，稱為有限細胞系(Finite cell line)，反之則稱為連續細胞系(Continuous cell line)。對於人體的腫瘤細胞，體外培養半年以上且生長穩定，並連續傳代的則可稱為連續性株或系。

本研究所使用到的五種細胞株，第一種為 HepG2 其源自患有肝癌的 15 歲白人男性之肝臟組織中，第二種為 K562 源自患有慢性骨髓性白血病的 53 歲白人女性之骨髓組織中，第三種為人類淋巴母細胞株 GM12878 源自患有人類疱疹病毒第四型的女性之血液組織，第四、五種則為人類乳腺癌細胞株 MCF-7 與人類非小細胞肺癌細胞株 A549，分別來自患有乳癌的 69 歲白人女性之胸組織以及患有上皮癌的 58 歲白人男性之肺臟組織。

2.5 自動化機器學習

自動化機器學習(Automated Machine Learning，簡稱 AutoML)提供一系列自動化之學習流程加速機器學習相關研究發展，過程中有效降低機器學習過程中建模的困難，目前有許多企業皆開發了相關平台提供有需求者使用，像是 Google Cloud AutoML、Microsoft Azure Machine Learning 和 Amazon SageMaker Autopilot 等。



2.5.1 高效神經網路架構搜索

神經網路架構搜索(Neural Architecture Search, 簡稱 NAS) (Zoph & Le, 2016), 基於強化學習(Reinforcement Learning, 簡稱 RL)的方式設計, 目的是自動根據預測表現、硬體資源限制或其他使用者的需求, 從神經網路組件中尋找並建立出合適的模型結構。NAS 主要分三部分, 分別為搜索空間(Search Space)也就是神經網路的元件像是架構的層數與卷積核大小等、搜索策略(Search Strategy)是在給定的搜索空間中要透過何種方式得到適合的模型架構, 像是調整超參數(Hyperparameter)時常用的隨機搜索(Random search)與窮舉搜索(Grid search)、性能估計策略(Performance Estimation Strategy)則是如何評估挑選出的模型架構好壞。高效神經網路架構搜索(Efficient Neural Architecture Search, 簡稱 ENAS) (Pham et al., 2018)提高了 NAS 在訓練子網路時的計算效率, 透過模型之間共享參數藉以加快模型收斂速度。本研究 ezGeno 為了使網路結構搜索過程的效率提升, 設計出簡易版 ENAS 稱為 ezNAS, 提供較簡單之殘差連線(Residual connection), ezGeno 的完整介紹於 3.2.1。

2.5.2 AutoKeras

Keras 由 Python 編寫而成的開源神經網路庫。AutoKeras (Jin et al., 2019)基於 Keras 使用 ENAS 方法的 AutoML, 由德克薩斯農工大學(Texas A&M University)的 DATA 實驗室開發的自動機器學習開源軟體庫, 因此 AutoKeras 具有高效的特性, 可用於圖像、文本和結構化資料的分類與回歸。於 ezGeno 文獻中, 與比較同樣為 AutoML 工具的 AutoKeras, 不管是在預測準確度或是建構預測模型所耗費的時間成本上, ezGeno 皆優於 AutoKeras, 而預測準確度也優於單層架構的 DeepBind 模型。

第三章 研究方法



此章節將分成兩個部分，第一部分為介紹 ENCODE 資料庫、資料庫中目標資料的數量統計、資料篩選、變異位點的資料來源與資料前處理，第二部分說明實驗流程、介紹所使用的自動機器學習工具 ezGeno，最後使用 MEME (Bailey, Johnson, Grant, & Noble, 2015)分析工具進行序列特徵探勘與變異序列分析所用的成對 t 檢定。

3.1 實驗資料

本小節分成五個部分，第一部分先介紹 ENCODE 資料庫，第二部分則統計資料庫中 ChIP-seq 的實驗數量，第三部分說明挑選目標資料的過程，第四部分為變異位點資料介紹，最後說明分析過程中使用到的 MEME 工具與成對 t 檢定。

3.1.1 ENCODE 資料庫

ENCODE (Consortium, 2012)為公開的資料庫，稱為 DNA 元件百科全書 (Encyclopedia of DNA Elements) 是由美國國家人類基因組研究所(US National Human Genome Research Institute，簡稱 NHGRI) 在 2003 年 9 月開始的一項公共聯合研究項目，目的是希望找出人類基因組中所有功能組件，包含註解基因、RNA、染色質狀態、轉錄調節的相關區域、DNA 甲基化等。採用多種分析方法來識別功能元件，通常通過 DNA 超敏性測定、DNA 甲基化測定和與 DNA 和 RNA 相互作用的蛋白質（即修飾的組蛋白、轉錄因子、染色質調節劑和 RNA 結合蛋白）的免疫沉澱進行研究，然後進行測序。

3.1.2 各細胞株轉錄因子數目

在 ENCODE 資料庫中，轉錄因子進行染色質免疫沉澱定序技術之實驗數量共有 2643 個，針對實驗數量前六多的細胞株(HepG2、K562、GM12878、MCF-7、A549 以及 HEK293)共有 2140 個實驗數量，去除相同轉錄因子在相同細胞株的實驗後為 1614 個實驗，如表 3-1 所示。表 3-2 為針對 1614 個實驗做交集數量

分析，發現 ENCODE 資料庫中有不少轉錄因子，且於不同細胞株中有不少相關數據資料，得以進一步去研究轉錄因子於跨細胞株間的結合特異性。



表 3-1 ENCODE 資料庫中六種細胞株之轉錄因子實驗數量統計

細胞株	實驗數量	轉錄因子數量
HepG2	679	591
K562	674	477
A549	252	80
HEK293	198	195
GM12878	188	155
MCF-7	149	116
總計	2140	1614

表 3-2 ENCODE 資料庫中，常見六種細胞株的轉錄因子交集數量

	HepG2	K562	GM12878	MCF-7	A549	HEK293
HepG2	591	232	102	75	53	58
K562	232	477	120	88	51	42
GM12878	102	120	155	65	42	8
MCF-7	75	88	65	116	30	8
A549	53	51	42	30	80	7
HEK293	58	42	9	8	7	195

3.1.3 ChIP-seq 實驗資料收集

第一部分，為了有效率地使用 ezGeno 進行預測，首先針對 ezGeno 之訓練資料大小進行預測表現分析，使用 ENCODE 資料庫之 K562 細胞株下載隨機選取

的 10 個轉錄因子峰值資料(ATF2、ATF3、CBX2、CTCF、GATA1、GATA2、IRF1、JUN、RNF2 以及 SETDB1)與兩個轉錄因子：JUND 和 ELF1 於 HepG2、K562、GM12878、MCF-7 與 A549 細胞株，其各別的轉錄因子峰值數量如表 3-3、3-5 所示。



表 3-3 K562 細胞株中 10 個轉錄因子的峰值數量

TF	Peak numbers	TF	Peak numbers
ATF2	41768	GATA2	21311
ATF3	18780	IRF1	14180
CBX2	2984	JUN	9889
CTCF	49765	RNF2	4717
GATA1	14676	SETDB1	4667

第二部分為分析轉錄因子於跨細胞株間的結合特異性，我們針對 ENCODE 資料庫中擁有較多實驗數量的六種細胞株 (HepG2、K562、GM12878、MCF-7、A549 以及 HEK293) 進行轉錄因子於不同細胞株間的 ChIP-seq 實驗數量統計。ENCODE 資料庫總共有 2140 個 ChIP-seq 實驗，去除相同轉錄因子在相同細胞株的實驗後為 1614 個實驗，不同轉錄因子跨細胞株的數量統計如表 3-4 所示，由於發現只有三個轉錄因子在六種細胞株都有資料，因此最後我們選擇了跨五個不同細胞株的 24 個轉錄因子去做結合特異性分析，其各別的轉錄因子峰值數量如表 3-5 所示。

表 3-4 不同轉錄因子跨細胞株的數量統計

跨細胞株數量	轉錄因子數量
1	626
2	223
3	88
4	35
5	24
6	3



表 3-5 24 個轉錄因子跨細胞株的峰值序列數量

TF name	Cell line	Peak numbers	TF name	Cell line	Peak numbers
RAD21	HepG2	41335	SP1	HepG2	27292
	K562	42109		K562	7877
	GM12878	44515		GM12878	15592
	MCF-7	52253		MCF-7	2016
	A549	26595		A549	43342
MYC	HepG2	2220	CEBPB	HepG2	53225
	K562	31378		K562	27107
	GM12878	4950		GM12878	5174
	MCF-7	26198		MCF-7	37726
	A549	9505		A549	47928
ELK1	HepG2	9721	MAX	HepG2	14545
	K562	3793		K562	37342
	GM12878	7245		GM12878	13605

表 3-5(續表) 24 個轉錄因子跨細胞株的峰值序列數量

TF name	Cell line	Peak numbers	TF name	Cell line	Peak numbers
	MCF-7	6463		MCF-7	38332
	A549	443		A549	12694
POLR2A	HepG2	3369	RCOR1	HepG2	4631
	K562	38755		K562	6705
	GM12878	20700		GM12878	7460
	MCF-7	26526		MCF-7	8285
	A549	19580		A549	752
MAFK	HepG2	72325	ELF1	HepG2	22166
	K562	27213		K562	32683
	GM12878	7368		GM12878	27369
	MCF-7	11301		MCF-7	21039
	A549	61039		A549	11614
CREB1	HepG2	24964	ZBTB33	HepG2	2417
	K562	17621		K562	3360
	GM12878	28276		GM12878	2652
	MCF-7	42278		MCF-7	7737
	A549	18814		A549	11229
SIN3A	HepG2	21930	HDAC2	HepG2	38000
	K562	15656		K562	5646
	GM12878	13401		GM12878	1496
	MCF-7	38054		MCF-7	20946

表 3-5(續表) 24 個轉錄因子跨細胞株的峰值序列數量

TF name	Cell line	Peak numbers	TF name	Cell line	Peak numbers
SIN3A	A549	11204	HDAC2	A549	4046
YY1	HepG2	17128	RFX5	HepG2	7330
	K562	11638		K562	2631
	GM12878	37261		GM12878	5827
	HEK293	30726		MCF-7	9368
	A549	23451		A549	5106
JUND	HepG2	41541	GABPA	HepG2	8204
	K562	35379		K562	16493
	GM12878	7602		GM12878	13948
	MCF-7	14020		MCF-7	14817
	A549	21652		A549	17261
TAF1	HepG2	17279	ESRRA	HepG2	19315
	K562	20460		K562	29519
	GM12878	17473		GM12878	4821
	MCF-7	8544		MCF-7	8663
	A549	17421		A549	2795
TCF12	HepG2	4576	SREBF1	HepG2	2835
	K562	13907		K562	2848
	GM12878	25023		GM12878	3498
	MCF-7	11602		MCF-7	7832
	A549	31502		A549	2483
EP300	HepG2	34041	ZNF24	HepG2	5596

表 3-5(續表) 24 個轉錄因子跨細胞株的峰值序列數量

TF name	Cell line	Peak numbers	TF name	Cell line	Peak numbers
EP300	K562	28710	ZNF24	K562	25919
	GM12878	19537		GM12878	16596
	MCF-7	11939		MCF-7	11392
	A549	25015		HEK293	18518

3.1.4 變異位點資料

基因型-組織表達計畫(Genotype-Tissue Expression, 簡稱 GTEx)研究人類基因表達和基因調控機制, 而其中表達定量性狀遺傳位點(Expression quantitative trait locus, 簡稱 eQTL)為利用 mRNA 的表現量來找出控制性狀的基因, 分析遺傳變異與基因表達的關聯性, 主要研究皮膚、血液、肝臟與脂肪等。

研究序列中變異位點對基因表達所造成的影響, 主要使用與建構模型相關之胸腺、肝和肺組織的變異位點, 如表 3-6 所示, 其中包含正樣本與負樣本, 分別利用 ezGeno 對 MCF-7、HepG2 和 A549 細胞株所建立的模型進行序列中轉錄因子結合位預測, 以統計方法之成對樣本 t 檢定進行參考序列與變異序列間平均數差異, 藉以了解變異於該序列是否會影響基因表達。

表 3-6 GTEx 資料庫中, 三種組織的正、負樣本的變異位點數量

	Breast	Liver	Lung
Positive	1129	352	1653
Negative	1129	352	1653



3.1.5 資料前處理

由 ENCODE 資料庫中下載的目標檔案皆為 BED 格式，其儲存該轉錄因子結合位的染色體位置信息，依照峰值分數排序後，使用 bedtools (Quinlan & Hall, 2010) 中的 getfasta 指令以及利用人類參考基因組 hg38 取得峰值的序列資料 FASTA 檔，為符合 ezGeno 所規定的輸入序列長度 101bp，最後進行序列片段資料的裁切處理，使用的裁切方法為將資料中峰值中位數的位置前後取 50bp 的範圍。

第一部分為了評估訓練資料的峰值數對模型準確度的影響，取出的正樣本數量分別 250、750、1250 以及 1750，並固定測試資料之正樣本數量為 250，觀察 ezGeno 預測之準確度並將評估後之較合適的正樣本數目用於第二部分的分析。

第二部分為針對跨五種細胞株間的 24 個轉錄因子進行結合特異性的分析，使用第一部分評估的結果作為訓練資料的正樣本數目，此外，使用峰值數量大於 10,000 與 50,000 的轉錄因子進一步分析增加訓練資料之正樣本數量對 ezGeno 預測之準確度的影響，小於 10,000 的區間內訓練資料之正樣本數量以 2,000 為單位增加，大於 10,000 的區間內訓練資料之正樣本數量以 10,000 為單位增加。

第三部分由 GTEx 資料庫取得胸腺、肝及肺組織之正、負樣本 vcf 檔案，將變異位點對回人類參考基因組 hg38 並取該位置前後 50bp 範圍共 101bp 的序列資料 FASTA 檔，即取得變異位點之參考序列，替代變異位點則取得含變異位點之序列資料。

3.2 實驗流程

本小節先介紹自動機器學習工具 ezGeno 與其使用方法，接著說明如何利用 ezGeno 分析後的資料進行特徵探勘分析以及分析變異位點對轉錄因子結合率之成對 t 檢定說明，圖 3-1 為實驗流程圖，研究主要分成三大階段。

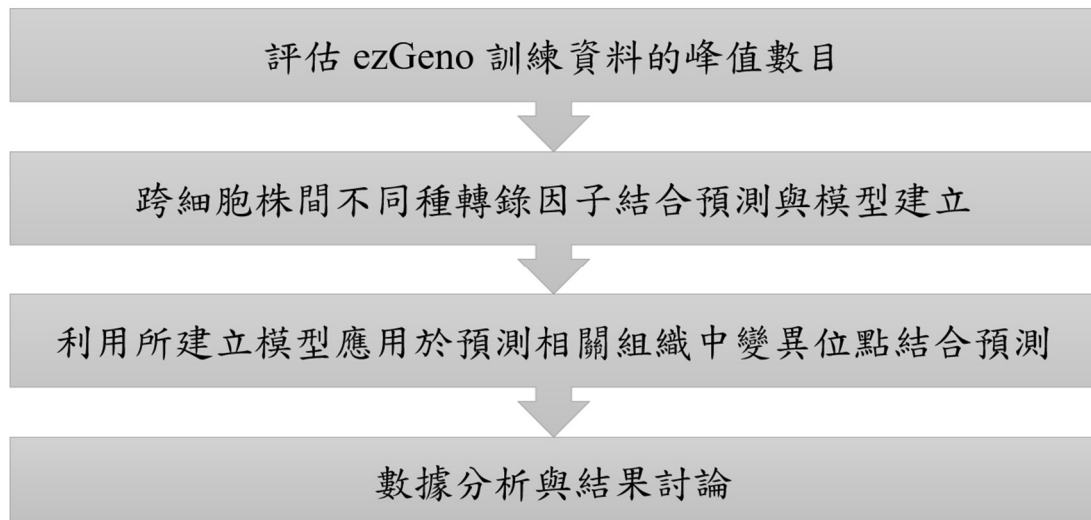


圖 3-1 實驗流程圖

3.2.1 ezGeno 概述

ezGeno (Lin et al., 2021)為本實驗室與台灣人工智慧實驗室(Taiwan AI Labs)合作開發的方法，結合 AutoML 和高效神經網路架構搜尋的技術，挑選各種參數及搭建深度學習之卷積神經網路模型，進而預測染色質免疫沉澱定序片段上轉錄因子可能的結合位置。主要使用一維卷積作為擷取序列中轉錄因子結合特徵方法，而卷積核選擇有 3、7、11、15 和 19 與是否為擴張卷積共 10 種模型單元，以及模型各層間是否相連，總計有 6000 個模型種類。

輸入的序列資料當作為正樣本，並分為 A、B 和 C 三部分，其中種類 A 為依照峰值數量排序後，前五百名偶數排名的序列，種類 B 則為前五百名基數排名的序列，最後種類 C 為去除 A、B 後其餘之序列。訓練資料集之正樣本由 A 和 C 組成，測試資料集則為種類 B 的序列，依據輸入之正樣本，ezGeno 流程如圖 3-2 所示，會自動生成以正樣本之反向序列作為負樣本，隨後進入建構最佳模型步驟以達到預測轉錄因子結合位之目的，最後呈現測試資料預測準確度並擷取出序列資料中可能之結合子序列以 FASTA 檔格式儲存。

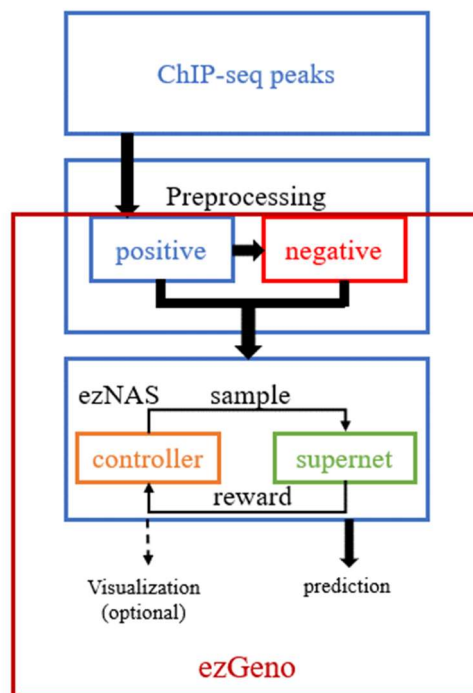


圖 3-2 為 ezGeno 的工作流程圖，本圖改自 ezGeno 文獻(Lin et al., 2021)

3.2.2 ezGeno 的使用

將處理好的序列依需求分成不同比例之訓練與測試資料，利用 ezGeno (Lin et al., 2021)自動生成負樣本的功能得到等量負樣本，接著進入建構模型的過程，最後搭建好的模型預測於測試資料與擷取序列上可能結合特徵之子序列，蒐集預測準確度評估 ezGeno 需使用的資料大小、預測模型的穩定度與分析轉錄因子於不同細胞株間結合特異性之研究，並將細胞株 MCF-7、HepG2 與 A549 所建立的轉錄因子模型應用於預測胸腺、肝與肺組織的參考與變異序列上進行轉錄因子的結合預測。

3.2.3 結果分析

本步驟介紹實驗結果分析之方法，分為兩部分(一)序列特徵分析以及(二)成對樣本 t 檢定分析，以下針對上述兩部分進行詳細說明。

(一)序列特徵分析

MEME (Bailey et al., 2015)工具分析輸入序列上是否有重複出現的區域，出現次數達設定的門檻且該區域符合所設定之序列長度，即會被認成一特徵序列，輸出所需的序列特徵數量，藉以找出生物意義。本研究利用預測模型擷取不同目標序列上可能的結合特徵之子序列 FASTA 檔案進行 MEME 序列特徵分析，並查找已知資料庫中相同轉錄因子資料進行比對，探討 ezGeno (Lin et al., 2021)於預測結合特徵的效果。

(二)成對樣本 t 檢定

成對樣本 t 檢定(Paired Sample t-test) (Kim, 2015)是比較兩組相依樣本之間的平均數差異，本研究應用此檢定方法於正、負樣本間比較參考序列與變異序列之結合預測差異，探討變異對基因表達之影響並分析是否具顯著差異，其中假設檢定必須先設定虛無假設(Null hypothesis) H_0 : 兩筆無累積差異；及對立假設(Alternative hypothesis) H_1 : 兩筆資料有累積差異，設定不同顯著性 p-value 門檻值，當顯著性 p-value 大於設定門檻，則無法拒絕虛無假說，即成對樣本間無顯著不同。

第四章 結果與討論



本論文的實驗結果分為六部分，第一部分為評估 ezGeno 使用資料大小之成果，第二部分為 ezGeno 針對五種細胞株中的 24 種轉錄因子的預測準確度比較，第三、四部分分別討論模型對結合位預測效果與模型選擇分析，第五部分則為研究變異位點可能對基因表達所造成的影響。

4.1 ezGeno 訓練資料的峰值數目對模型準確度的影響

第一部分使用兩種資料集分析測試訓練資料的峰值數目對模型準確度的影響，第一種資料集為 K562 中的 10 個轉錄因子(ATF2、ATF3、CBX2、CTCF、GATA1、GATA2、IRF1、JUN、RNF2 以及 SETDB1) 的峰值資料，第二種則使用轉錄因子 JUND 與 ELF1 分別於 HepG2、K562、GM12878、MCF-7 以及 A549 五種細胞系的峰值資料。如圖 4-1、4-2 所示，分別為附表 A-1、A-2 之數據視覺化呈現，為了說明不同訓練資料數目間的預測表現差異，因此使用成對樣本 t 檢定計算顯著性 p-value，如表 4-1 所示，其中訓練資料正樣本數目 250 與 750 間最具有顯著性，而隨著數量增加，顯著性降低，由此可知，當訓練資料正樣本數量到達 750 以上時，明顯有較佳的預測準確度表現，而隨著使用訓練資料正樣本繼續增加時，雖然所建立模型在測試資料上的預測準確度隨之上升，但上升幅度有限，預測表現會趨於穩定。

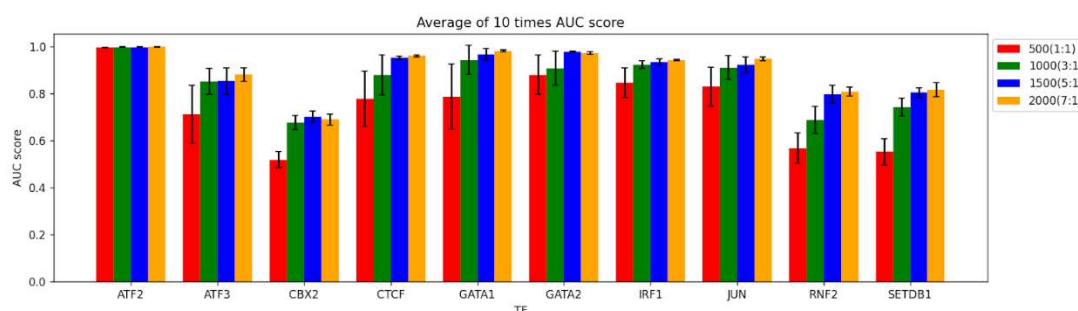


圖 4-1 針對 K562 中 10 種轉錄因子，分別使用 250、750、1250 與 1750 作為正樣本的訓練資料的峰值數目對模型準確度之影響，括號內為訓練與測試資料比

例

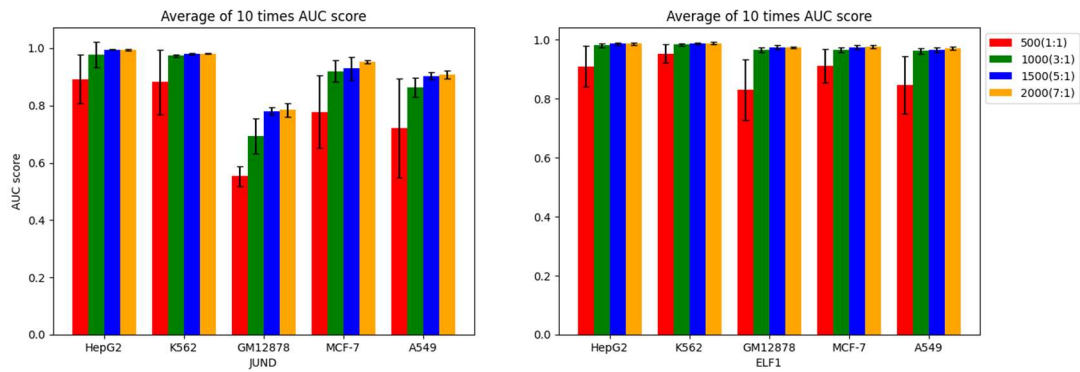


圖 4-2 針對 2 種轉錄因子於五種細胞株中，分別使用 250、750、1250 與 1750 作為正樣本的訓練資料的峰值數目對模型準確度之影響，括號內為訓練與測試資料比例

表 4-1 不同峰值數量間對預測表現的顯著性 p-value

Group1 與 Group2 為訓練資料正樣本數量; Overall 為結合兩種資料集

Group 1	Group 2	p-value		
		10 TFs in K562	2 TFs in 5 cell lines	Overall
250	750	0.00035	0.00002	1.75×10^{-8}
750	1250	0.00914	0.04761	0.00089
1250	1750	0.05458	0.12734	0.01309

第二部分使用於五種細胞株中，各別峰值數目皆大於 10,000 之轉錄因子 SIN3A，分別以 1750、3750、5750、7750 與 9750 作為訓練資料正樣本峰值數目，和第三部分使用峰值數目大於 50,000 之轉錄因子，分別為 RAD21 (MCF-7)、CEBPB (HepG2)、MAFK (HepG2)與 MAFK (A549)，以 1750、3750、5750、7750、9750、19750、29750、39750 與 49750 作為訓練資料正樣本峰值數目，分析訓練資料的峰值數目對模型準確度的影響。如圖 4-3、4-4 所示，分別為附錄 A-3、A-

4 之數據視覺化呈現，棒狀圖趨勢與第一部分相同，再次驗證了隨著使用訓練資料正樣本增加時，所建立模型在測試資料上的預測準確度隨之上升，預測表現也有所穩定的預測的情況。

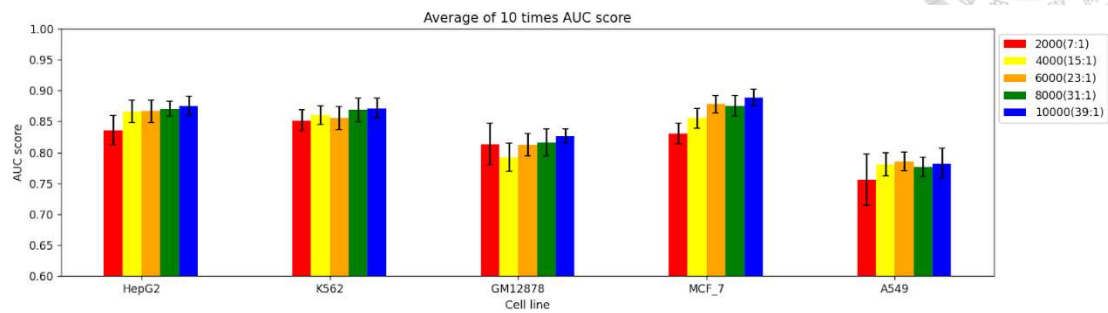


圖 4-3 分別使用 1750、3750、5750、7750 與 9750 作為正樣本的訓練資料的峰值數目對模型準確度之影響(轉錄因子為 SIN3A)，括號內為訓練與測試資料比

例

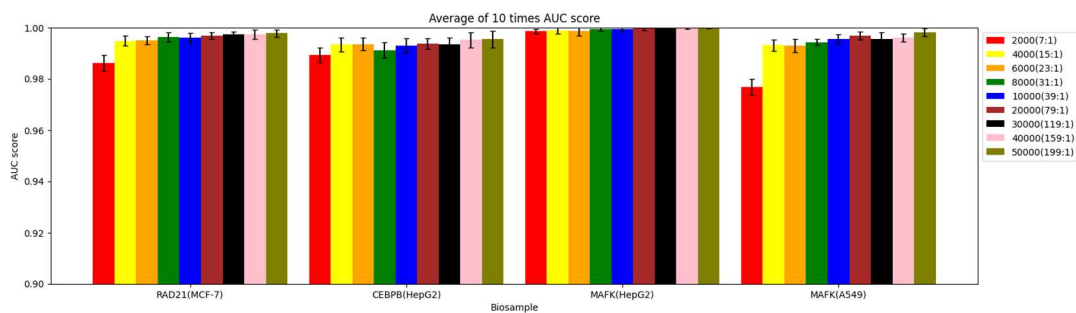


圖 4-4 分別使用 1750、3750、5750、7750、9750、19750、29750、39750 與 49750 作為正樣本的訓練資料峰值數目對模型準確度之影響，括號內為訓練與測試資料比例

4.2 跨細胞株間 ezGeno 預測轉錄因子結合位的表現差異

跨五種細胞株間的 24 種轉錄因子中有 22 種轉錄因子在五種相同細胞株間存在，圖 4-5 為表 3-4 部分數據視覺化呈現該 110 個原始下載檔案之峰值數量熱圖，圖 4-6 則為附表 B-1 部分數據視覺化，共有 110 個 ezGeno 平均預測準確度群集熱圖並參照表 4-2，表中整理了 JASPAR 資料庫(Castro-Mondragon et al., 2021)中 24 種蛋白質家族與結合特徵，由分群結果得以推測可能具有相似的結合特性，

像是 JUND 與 CEBPB 為 Basic leucine zipper factors (bZIP)、ELF1 與 GABPA 皆為 Tryptophan cluster factors 等。此外，對照兩熱圖進行觀察可發現成些微正相關，推測原始資料峰值數量越多則預測準確度越高，因此，進一步使用轉錄因子 RAD21 與 MAFK 分析排序後前 2000 取其中的 1750 與末 1750 的峰值分別作為訓練資料正樣本，並使用相同的 250 個峰值作為測試資料正樣本，預測表現結果於附表 B-1、B-2，如圖 4-7 所示，計算其顯著性 p-value 為 0.0298 (<0.05)，顯示兩群具有顯著差異，說明原始資料中的峰值品質有所差異，經排序後挑選的峰值進行訓練會影響預測表現，當原始峰值數目多時，其中品質較佳的峰值相對較多。

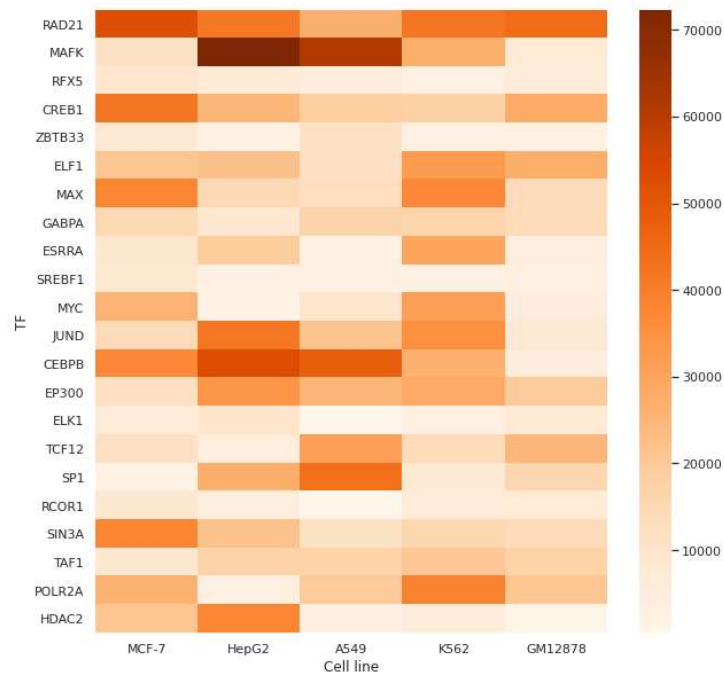


圖 4-5 五種細胞株中不同轉錄因子之峰值數量熱圖

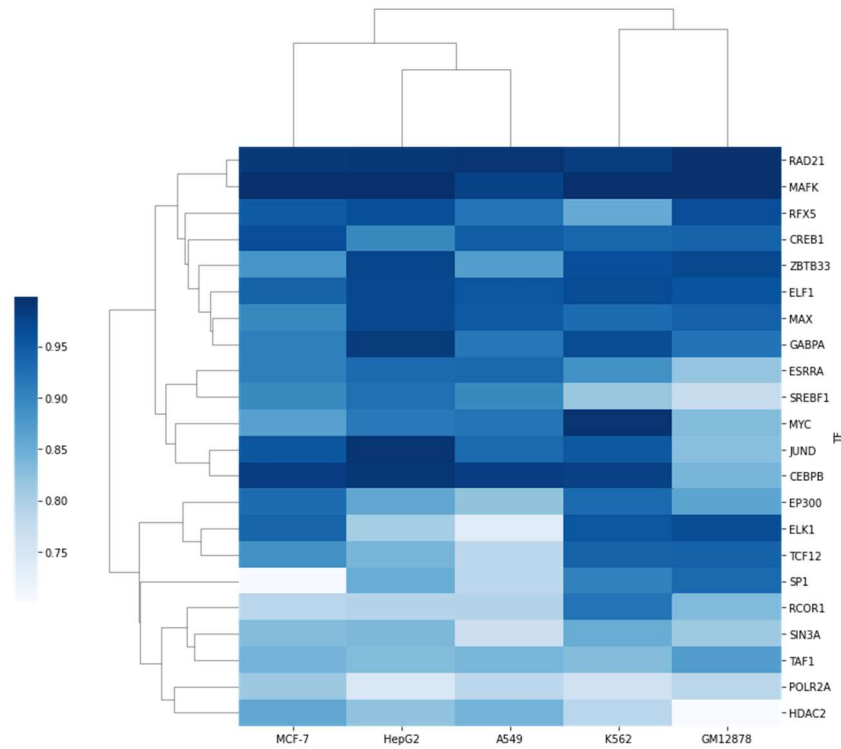


圖 4-6 模型準確度之分群熱圖，固定訓練資料中正樣本數為 1750，測試資料正樣本數為 250

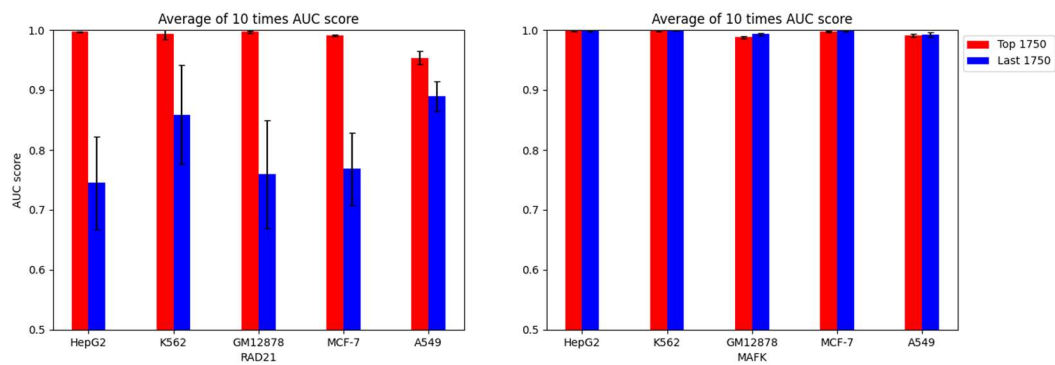


圖 4-7 使用轉錄因子 RAD21 與 MAFK 於五種細胞株的前、末 1750 峰值作為訓練正樣本的預測準確度

表 4-2 JASPAR 資料庫中 24 種轉錄因子的資料



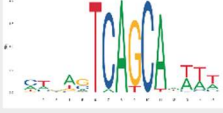














TF	TF family	JASPAR
RAD21	Rad21 family	NA
MYC	Basic helix-loop-helix factors (bHLH) bHLH-ZIP	
ELK1	Tryptophan cluster factors Ets-related	
POLR2A	RNA polymerase beta' chain family	NA
MAFK	Basic leucine zipper factors (bZIP) Maf-related	
CREB1	Basic leucine zipper factors (bZIP) CREB-related factors	
SIN3A	SIN3 transcription regulator family	NA
YY1	C2H2 zinc finger factors More than 3 adjacent zinc fingers	
JUND	Basic leucine zipper factors (bZIP) Jun-related	
TAF1	TAF1 family	NA
TCF12	Basic helix-loop-helix factors (bHLH) E2A	
EP300	Histone acetyltransferase	NA
SP1	C2H2 zinc finger factors Three-zinc finger Kruppel-related	
CEBPB	Basic leucine zipper factors (bZIP) CEBP-related	

表 4-2(續表) JASPAR 資料庫中 24 種轉錄因子的資料

TF	TF family	JASPAR
MAX	Basic helix-loop-helix factors (bHLH) bHLH-ZIP	
RCOR1	CoREST family	NA
ELF1	Tryptophan cluster factors Ets-related	
ZBTB33	C2H2 zinc finger factors Other factors with up to three adjacent zinc fingers	
HDAC2	Histone deacetylase family. HD type 1 subfamily	NA
RFX5	Fork head/winged helix factors RFX-related factors	
GABPA	Tryptophan cluster factors Ets-related	
ESRRA	Nuclear receptors with C4 zinc fingers Steroid hormone receptors (NR3)	
SREBF1	Basic helix-loop-helix factors (bHLH) bHLH-ZIP	
ZNF24	C2H2 zinc finger factors More than 3 adjacent zinc fingers	



4.3 序列特徵分析

將峰值數量超過 50,000 的轉錄因子進行序列特徵分析，使用訓練資料正樣本為 1750 個峰值的預測模型，分別針對只有 250 個峰值的測試資料正樣本、2000 個峰值含有訓練與測試資料正樣本以及前 50,000 名峰值序列進行重點位置分析並輸出，隨後使用 MEME 分析工具進行序列特徵探勘。如表 4-3 所示，分別呈現 ezGeno 挑選出的重點序列片段數目、MEME 分析後的特徵、該特徵出現的數目及該數目與 ezGeno 挑選的序列數目比值，並參照 JASPAR 資料庫的轉錄因子特徵。原峰值序列於 MEME 分析所使用的序列數量比例皆高於模型篩選，由此可推測模型學到除了主要結合序列特徵以外，模型也學到一些額外的特徵，而當可篩選之序列數量增加時，主要的特徵序列比例也隨之增加。

表 4-3 訓練資料正樣本為 1750 個峰值的預測模型應用於不同序列資料之結果

	Number of peaks in the raw data	Number of sub-sequences extracted by ezGeno	Sites discovered by MEME	Sequence usage of MEME	Motif by MEME	Motif in JASPAR database
RAD21 in MCF-7	250	623	201	0.323		NA
	2000	5527	810	0.147		
	50000	62584	33370	0.533		
	52253	NA	51929	0.994		
CEBPB in HepG2	250	343	255	0.743		
	2000	6196	1015	0.164		
	50000	135648	31286	0.231		

表 4-3(續表) 訓練資料正樣本為 1750 個峰值的預測模型應用於不同序列資料之
結果

	Number of peaks in the raw data	Number of sub-sequences extracted by ezGeno	Sites discovered by MEME	Sequence usage of MEME	Motif by MEME	Motif in JASPAR database
CEBPB in HepG2	53225	NA	45878	0.862		
MAFK in HepG2	250	440	196	0.445		
	2000	5591	1004	0.180		
	50000	54568	40516	0.742		
	72325	NA	71963	0.995		
MAFK in A549	250	300	221	0.737		
	2000	5797	661	0.114		
	50000	94787	31588	0.333		
	61039	NA	60584	0.993		

4.4 模型架構的挑選分析

此處分三部分去探討 ezGeno 所建立的預測模型架構，第一部分為針對相同訓練與測試資料模型是否具穩定性，如圖 4-8、4-9 所示，第一層模型皆傾向使用相同的模型單元，可推斷使用相同訓練及測試資料所建立的 ezGeno 預測模型具有穩定性，由此說明 ezGeno 在建立預測模型的過程中，並非隨便挑選組裝，而

是能夠根據不同資料內容提供適合的模型。

第二部分為分析轉錄因子於不同細胞株間模型架構的差異，討論 ezGeno 是否因資料性質而有模型挑選的差異。圖 4-10 至 4-12 分別呈現了轉錄因子 ELF1、ELK1 與 JUND 於五種不同細胞株中的 ezGeno 預測模型選擇 UMAP (Uniform Manifold Approximation and Projection) 作圖，在三張圖中皆呈現出不同細胞株間挑選的模型架構差異，其中 ELK1 分群最為明顯，可能也反映出該轉錄因子於這五種細胞株間結合性質有所差異。

第三部分則為分析不同轉錄因子的模型差異。圖 4-13 隨機挑選六種轉錄因子分別為 MAX、ELF1、SIN3A、ZBTB33、GABPA 和 RFX5 各別以五種細胞株各十次實驗共建構 50 個模型 UMAP 作圖，除了 SIN3A 以外，MAX 集中於左側部分而其餘集中於中央，顯示出轉錄因子因結合特性差異所需建構不同模型架構。

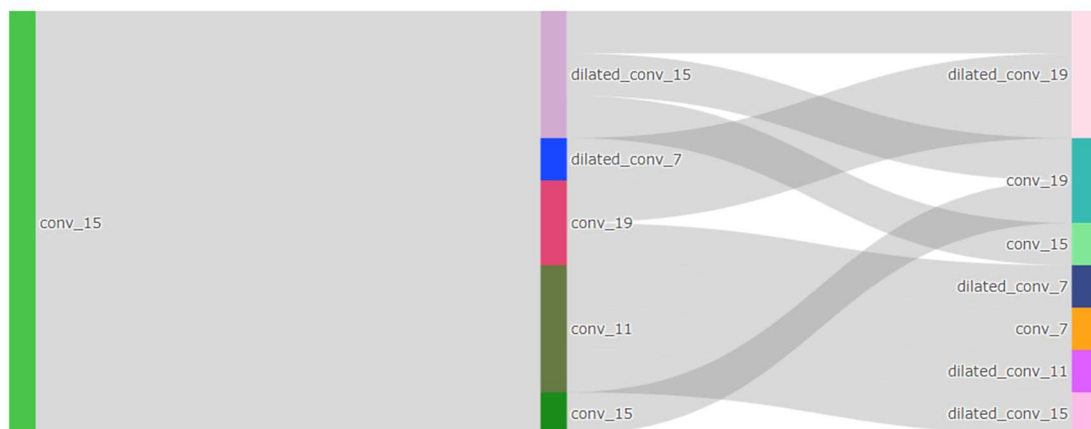


圖 4-8 ezGeno 針對 MCF-7 中轉錄因子 RAD21 所建構之模型

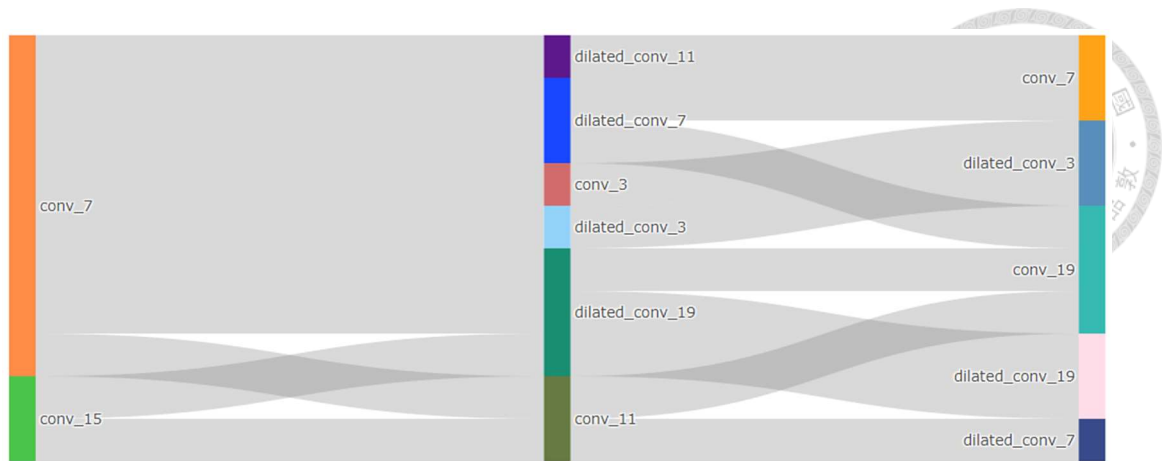


圖 4-9 ezGeno 針對 GM12878 中轉錄因子 GABPA 所建構之模型

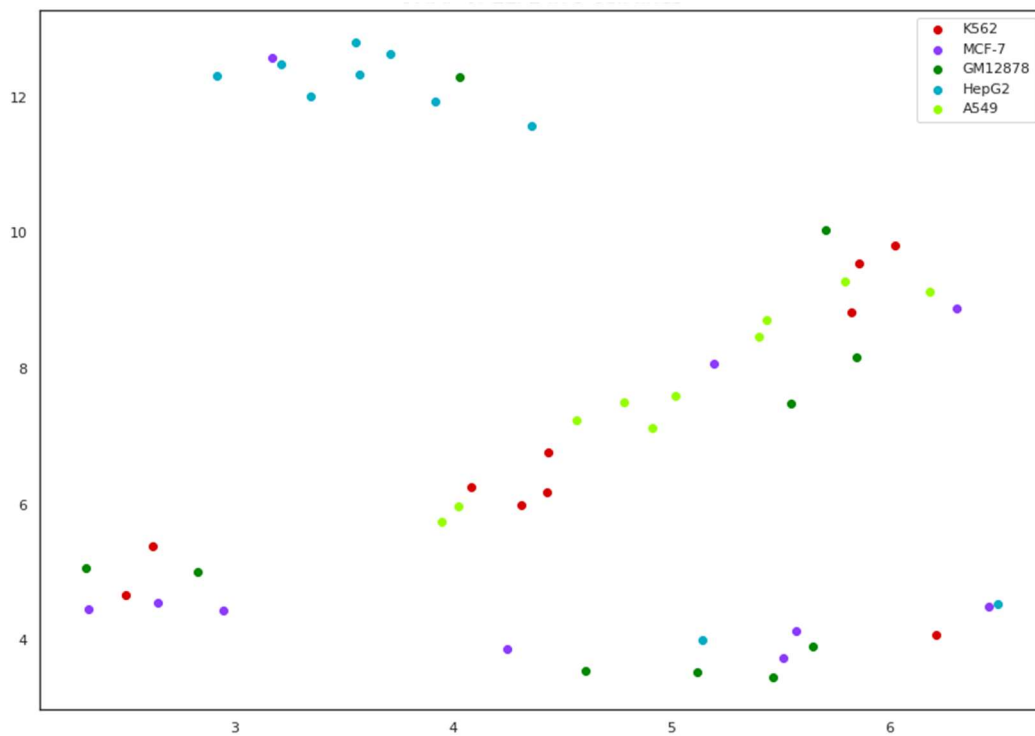


圖 4-10 轉錄因子 ELF1 於五種細胞株所建模型之 UMAP 圖

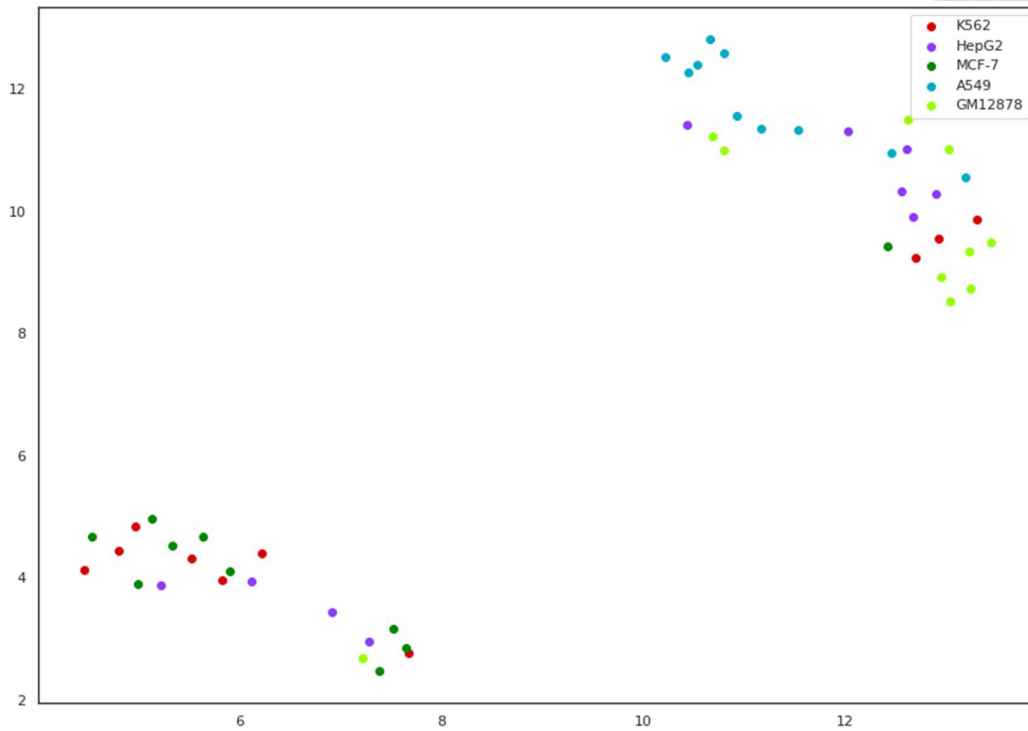


圖 4-11 轉錄因子 ELK1 於五種細胞株所建模型之 UMAP 圖

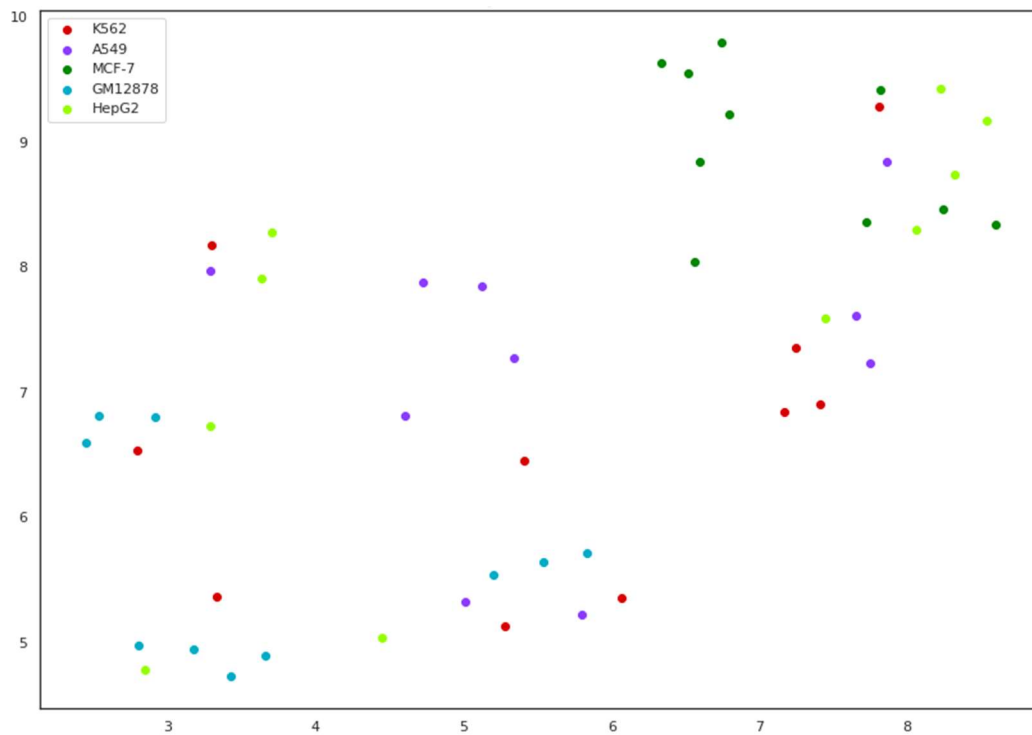


圖 4-12 轉錄因子 JUND 於五種細胞株所建模型之 UMAP 圖

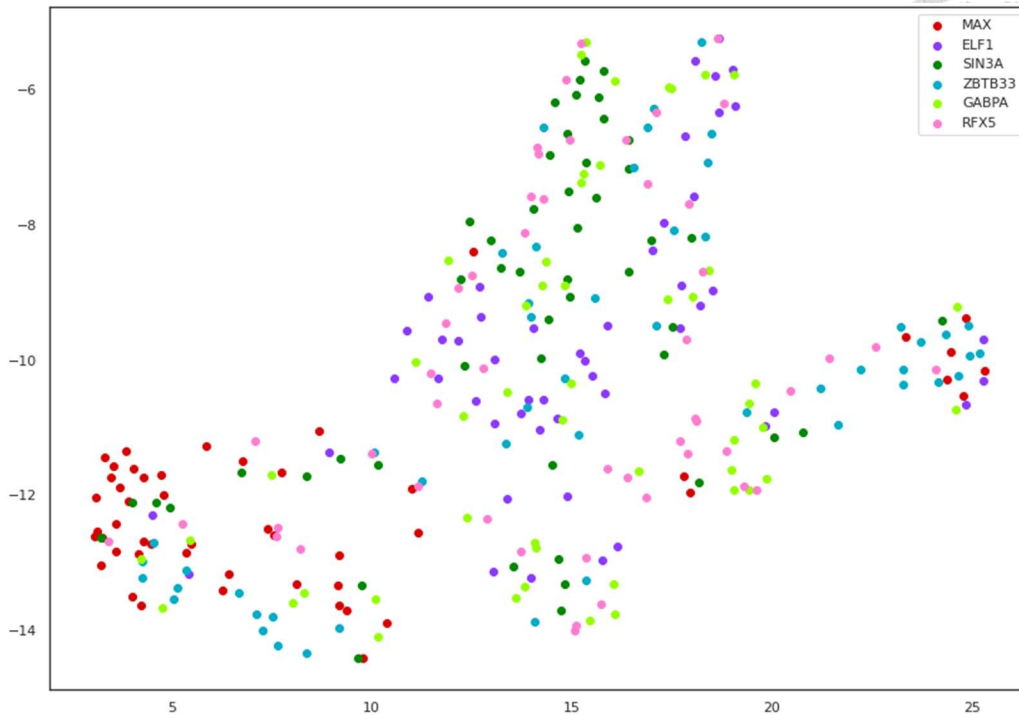


圖 4-13 隨機選取六種轉錄因子之 UMAP 圖

4.5 變異對基因表達所造成的影響

針對胸腺、肝與肺組織的變異之序列使用相關細胞株 MCF-7、HepG2 與 A549 所建立的轉錄因子預測模型分別對參考序列與變異序列進行結合位預測，以成對樣本 t 檢定計算 p-value 並設定不同閾值判斷是否有顯著差異，如表 4-4 所示，表中數字為通過閾值之變異數量，於三種組織中通過閾值之正樣本變異數量皆多於負樣本變異數量，由此可說明所建模型預測變異位點對轉錄因子結合機率具可靠性，圖 4-14、4-15 為視覺化呈現肺組織中正、負樣本於不同轉錄因子模型結合預測後通過閾值(p-value<0.005)具顯著性之變異位點，由於序列數量過多因而分成了兩張圖呈現，由兩圖觀察到大部分變異位點於轉錄因子 SIN3A 模型皆具顯著性，可供後續深入研究。附錄 C 中分別統計各轉錄因子模型於不同組織中變異位點對結合預測造成影響之序列數量，附表 C-1 可推測胸腺組織中變異位點對轉

錄因子 MYC、CEBPB 結合預測影響相對顯著、而附表 C-2 為肝組織變異位點對結合預測影響相對顯著的轉錄因子有 ELF1、GABPA，附表 C-3 則是轉錄因子 SIN3A 與 ELK1 於肺組織中變異位點對結合預測的影響較為顯著。針對正、負樣本中對 24 種轉錄因子結合預測有影響的變異位點計算平均佔比，並以成對樣本 t 檢定分析具影響的變異位點於正、負樣本間的差異性，其 p-value 為 0.0045 (<0.05)，說明兩群間的差異具有顯著性。

表 4-4 三種組織中，對結合預測有影響之變異數量，括號內為平均佔比

Variant list		p-value <0.05	p-value <0.01	p-value <0.005
Breast	Positive	11069 (0.4263)	6209 (0.2391)	4672 (0.1799)
	Negative	10681 (0.4113)	5968 (0.2298)	4458 (0.1717)
Liver	Positive	3325 (0.3936)	1782 (0.2109)	1277 (0.1512)
	Negative	3142 (0.3719)	1660 (0.1965)	1231 (0.1457)
Lung	Positive	17198 (0.4524)	10564 (0.2779)	8365 (0.2200)
	Negative	15435 (0.4060)	8648 (0.2275)	6504 (0.1711)

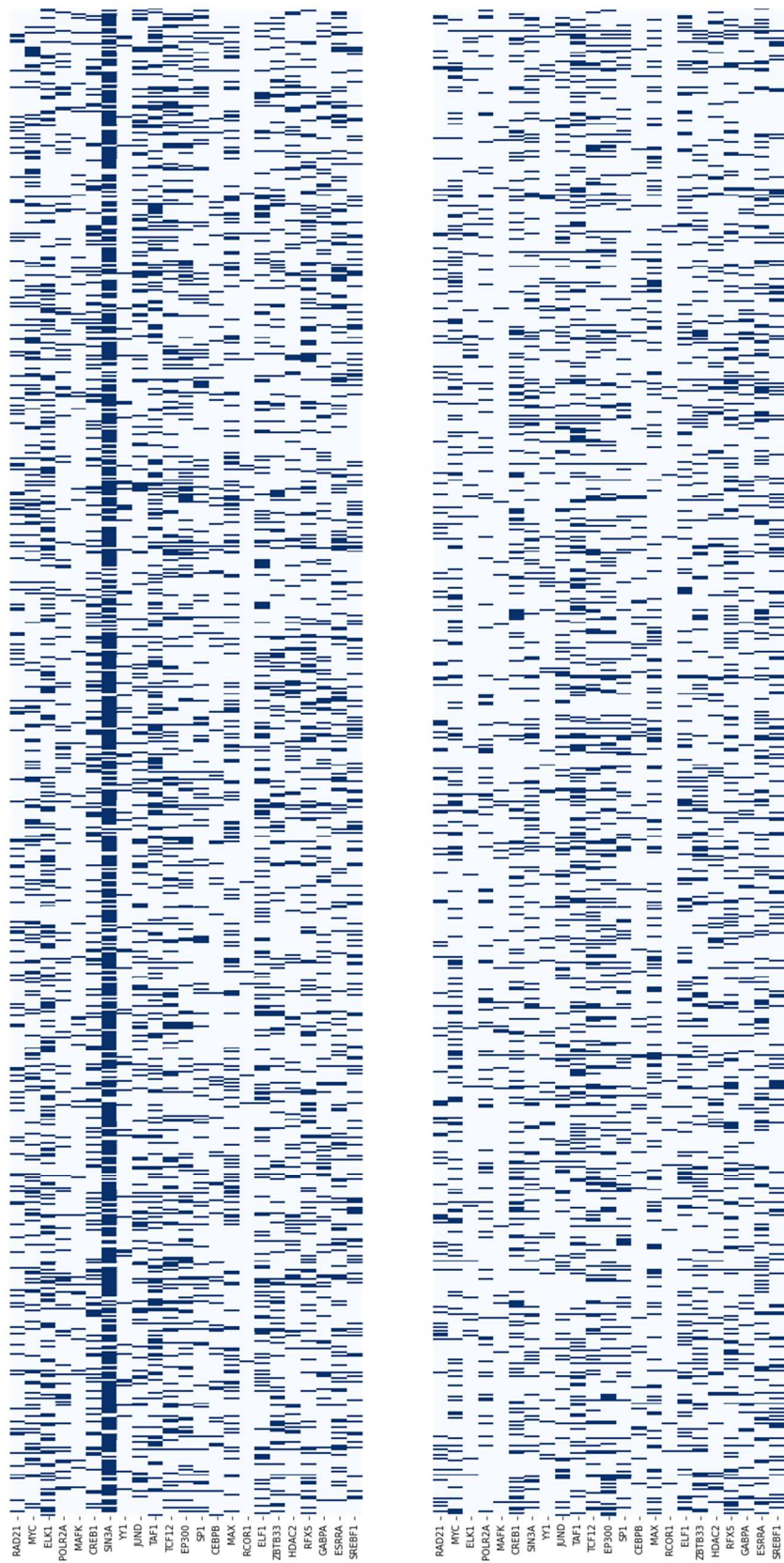


圖 4-14 肺組織中的變異位點以顯著性 $p\text{-value} < 0.005$ 標註(第一部分)

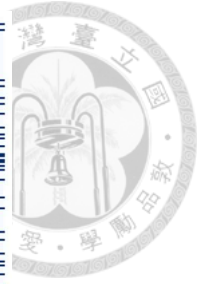
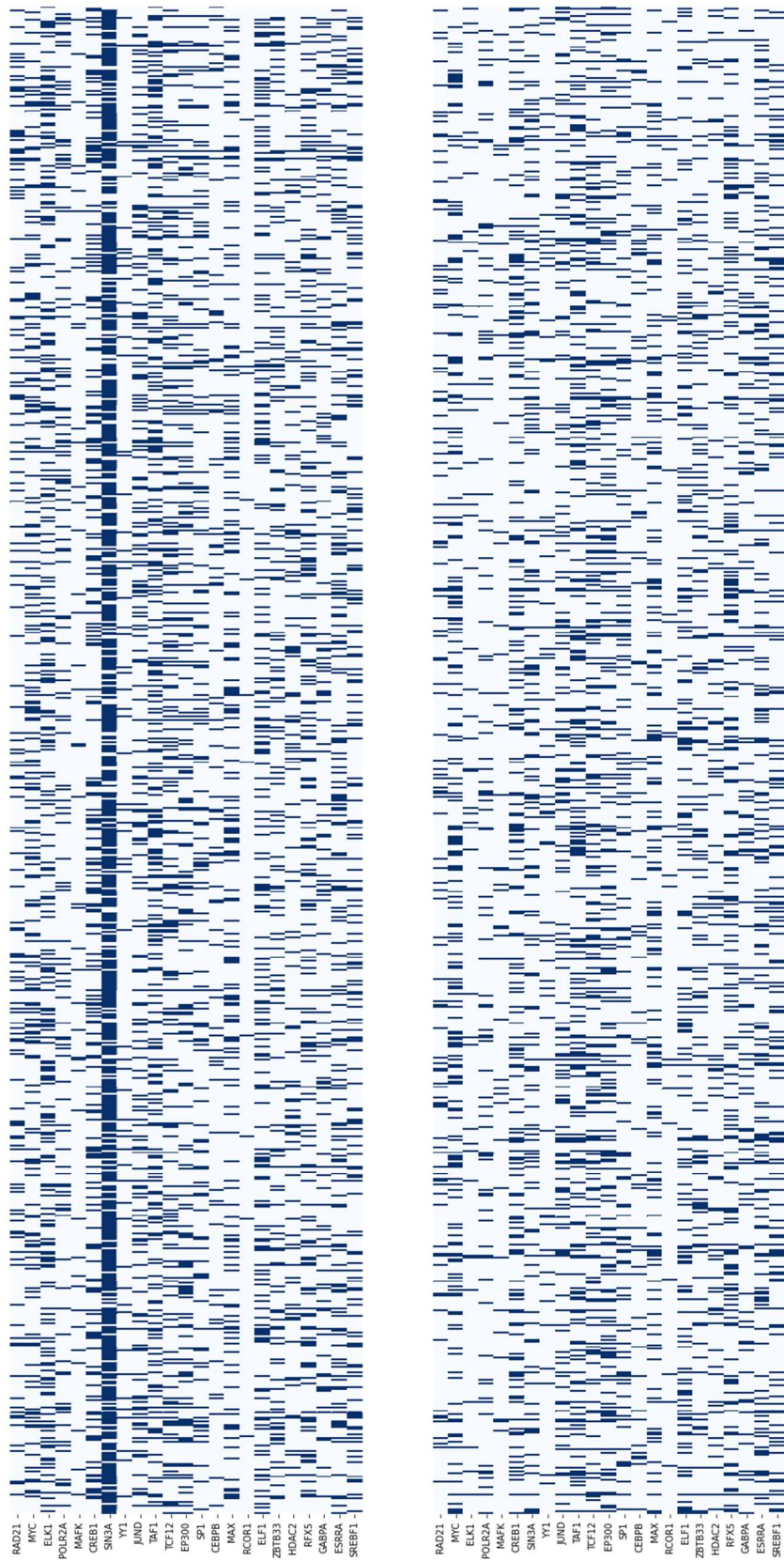


圖 4-15 肺組織中的變異位點以顯著性 $p\text{-value} < 0.005$ 標註(第二部分)

第五章 結論



本研究針對大規模轉錄因子於跨細胞株進行分析與探討，但由於轉錄因子結合特性有所不同而需要不同模型架構的問題，因而選用自動機器學習工具 ezGeno 進行實驗，先針對 ezGeno 的訓練資料大小進行評估，隨後將評估後結果應用於 ezGeno 對目標資料進行模型建立、結合預測與特徵分析，分別針對預測準確度、特徵序列、模型穩定度、不同細胞株與不同轉錄因子對建構好的模型進行比較分析，並將所建構之模型應用於變異位點造成轉錄因子結合異常進而影響基因表達的探討。

本論文首先使用 ENCODE 資料庫中 K562 細胞株的 10 種轉錄因子免疫沉澱定序資料，分別以 500、1000、1500 與 2000 的峰值序列資料量大小完成 ezGeno 訓練資料大小評估，當資料量大於 1000(也就為訓練資料大於 750)時，所建模型會有較佳的預測準確度，而當資料量越多時則預測準確度越穩定，可供往後的使用者參考。第二部分利用 ENCODE 資料庫中跨五種細胞株之 24 種轉錄因子免疫沉澱定序資料，皆以訓練資料 1750、測試資料 250 使用 ezGeno，建構 1200 個序列特徵預測模型，以模型預測準確度進行分群並查找所屬蛋白質家族，發現 JUND、CEBPB 與 ELF1、GABPA 兩組轉錄因子分別為相同家族，顯示可能具相似的結合特性，分析特徵序列部分，原資料使用 MEME 工具分析出的主要特徵其序列使用率皆高於經模型所挑選之序列，此現象說明預測模型除了主要結合特徵還學習到其他序列特徵；以外，模型架構分析再次驗證了不同轉錄因子需要不同模型架構。後續將所建構的 MCF-7、HepG2 與 A549 相關模型分別應用於分析胸腺、肝與肺組織中的變異位點對基因表達所造成的影響，藉由設定不同顯著性 p-value 門檻進行篩選具顯著性之變異位點，並進一步計算得正、負樣本間具有顯著性差異，顯示本研究所建立的預測模型在未來應用於尋找可能影響轉錄因子結合之變異位點具有可行性。變異位點對轉錄因子結合預測的分析部分，受限於目前資料庫中一般組織的實驗資料數量過少，而無法進行比較分析，希望於未來

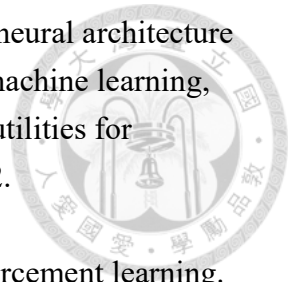
能完整該實驗。期許本研究分析流程在未來能幫助生物資訊領域解決更多基因表現與轉錄因子結合位的議題。



參考文獻

- Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*, 33(8), 831-838. <https://doi.org/10.1038/nbt.3300>
- Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods*, 18(10), 1196-1203. <https://doi.org/10.1038/s41592-021-01252-x>
- Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Res*, 43(W1), W39-49. <https://doi.org/10.1093/nar/gkv416>
- Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., Fornes, O., Leung, Tiffany Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., . . . Mathelier, A. (2021). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 50(D1), D165-D173. <https://doi.org/10.1093/nar/gkab1113>
- Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57-74. <https://doi.org/10.1038/nature11247>
- Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258), 561-563. <https://doi.org/10.1038/227561a0>
- Jin, H., Song, Q., & Hu, X. (2019). *Auto-Keras: An Efficient Neural Architecture Search System* Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA. <https://doi.org/10.1145/3292500.3330648>
- Kim, T. K. (2015). T test as a parametric statistic. *Korean J Anesthesiol*, 68(6), 540-546. <https://doi.org/10.4097/kjae.2015.68.6.540>
- Lin, J.-L., Hsieh, T.-T., Tung, Y.-A., Chen, X.-J., Hsiao, Y.-C., Yang, C.-L., Liu, T.-L., & Chen, C.-Y. (2021). ezGeno: an automatic model selection package for genomic data analysis. *Bioinformatics*, 38(1), 30-37. <https://doi.org/10.1093/bioinformatics/btab588>
- McGahon, A. J., Martin, S. J., Bissonnette, R. P., Mahboubi, A., Shi, Y., Mogil, R. J., Nishioka, W. K., & Green, D. R. (1995). The end of the (cell) line: methods for the study of apoptosis in vitro. *Methods Cell Biol*, 46, 153-185. [https://doi.org/10.1016/s0091-679x\(08\)61929-9](https://doi.org/10.1016/s0091-679x(08)61929-9)
- Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10), 669-680. <https://doi.org/10.1038/nrg2641>

- Pham, H., Guan, M., Zoph, B., Le, Q., & Dean, J. (2018). Efficient neural architecture search via parameters sharing. International conference on machine learning,
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841-842.
<https://doi.org/10.1093/bioinformatics/btq033>
- Zoph, B., & Le, Q. V. (2016). Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.



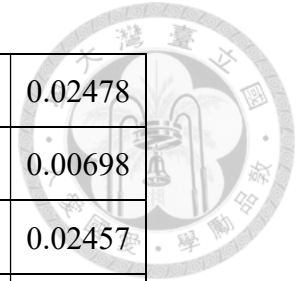


附錄A

附表 A-1 十種轉錄因子於 K562 中，使用不同正樣本峰值數目的 ezGeno 預測準確度

TF	AUC											
	1	2	3	4	5	6	7	8	9	10	AVG	STDEV
正樣本數量為 500(1:1)，訓練與測試資料之正樣本分別為 250 與 250												
ATF2	0.9966	0.9972	0.9996	0.9966	0.9965	0.9972	0.9996	0.9960	0.9973	0.9947	0.9971	0.00150
ATF3	0.6614	0.7748	0.6900	0.6678	0.8479	0.5270	0.8449	0.8703	0.5337	0.7143	0.7132	0.12299
CBX2	0.5201	0.5803	0.4991	0.4856	0.5489	0.4912	0.5001	0.5664	0.4882	0.5113	0.5191	0.03424
CTCF	0.6676	0.8895	0.9391	0.8790	0.8205	0.7168	0.6742	0.8933	0.6881	0.6203	0.7788	0.11707
GATA1	0.6110	0.8188	0.9582	0.8029	0.9611	0.6679	0.7301	0.7660	0.6061	0.9556	0.7878	0.13779
GATA2	0.8721	0.7256	0.8876	0.9390	0.9238	0.8995	0.9592	0.9494	0.9161	0.7394	0.8812	0.08292
IRF1	0.9329	0.8107	0.8141	0.7765	0.8587	0.7477	0.8312	0.9010	0.9236	0.8854	0.8482	0.06257
JUN	0.9034	0.7935	0.8470	0.9159	0.9000	0.8961	0.7678	0.8482	0.6564	0.7686	0.8297	0.08284
RNF2	0.6568	0.5221	0.6447	0.4642	0.5494	0.5237	0.5189	0.6110	0.6395	0.5569	0.5687	0.06536

SETDB1	0.6712	0.5259	0.5893	0.5006	0.5458	0.5959	0.5304	0.5343	0.4681	0.5643	0.5526	0.05662
正樣本數量為 1000(1:1)，訓練與測試資料之正樣本分別為 750 與 250												
ATF2	0.9992	0.9996	0.9980	0.9982	0.9998	0.9970	0.9946	0.9993	1.0000	0.9990	0.9985	0.00164
ATF3	0.8475	0.8417	0.8910	0.8704	0.9067	0.8888	0.7094	0.8632	0.8395	0.8654	0.8524	0.05490
CBX2	0.6684	0.6949	0.6134	0.6876	0.6929	0.6771	0.6858	0.6891	0.6419	0.7230	0.6774	0.03050
CTCF	0.9308	0.9418	0.9172	0.9359	0.7301	0.9498	0.7592	0.9108	0.9439	0.7904	0.8810	0.08557
GATA1	0.9426	0.9578	0.8662	0.8036	0.9806	0.9835	0.9686	0.9845	0.9789	0.9782	0.9445	0.06100
GATA2	0.8080	0.9515	0.9383	0.9410	0.9771	0.8050	0.9652	0.9216	0.8082	0.9734	0.9089	0.07224
IRF1	0.9197	0.9284	0.9216	0.9333	0.9330	0.9274	0.9425	0.9255	0.9300	0.8811	0.9243	0.01648
JUN	0.8065	0.9404	0.9527	0.9463	0.8657	0.9340	0.9439	0.9377	0.8547	0.9353	0.9117	0.05044
RNF2	0.7234	0.7013	0.6789	0.6475	0.6392	0.6914	0.7988	0.6019	0.6575	0.7428	0.6883	0.05693
SETDB1	0.7188	0.7385	0.8205	0.7035	0.7488	0.6995	0.7861	0.7186	0.7708	0.7220	0.7427	0.03911
正樣本數量為 1500(1:1)，訓練與測試資料之正樣本分別為 1250 與 250												
ATF2	0.9981	0.9999	0.9984	0.9999	0.9978	1.0000	0.9996	0.9975	0.9997	0.9970	0.9988	0.00115
ATF3	0.7688	0.8673	0.9150	0.8800	0.8986	0.7718	0.8821	0.8978	0.7881	0.8849	0.8554	0.05634



CBX2	0.7163	0.7125	0.6607	0.6714	0.7360	0.7132	0.7139	0.6871	0.6894	0.7299	0.7030	0.02478
CTCF	0.9626	0.9553	0.9495	0.9609	0.9548	0.9563	0.9530	0.9371	0.953	0.9554	0.9538	0.00698
GATA1	0.9805	0.976	0.9435	0.9821	0.9561	0.9093	0.9850	0.9751	0.9861	0.9808	0.9675	0.02457
GATA2	0.9783	0.9785	0.9809	0.9760	0.9843	0.9769	0.9806	0.9798	0.9778	0.9736	0.9787	0.00295
IRF1	0.9470	0.9267	0.9485	0.9524	0.9353	0.9336	0.9407	0.9083	0.9312	0.9312	0.9355	0.01279
JUN	0.9251	0.9281	0.9299	0.9330	0.9258	0.9476	0.9490	0.9358	0.8254	0.9302	0.9230	0.03529
RNF2	0.7240	0.7664	0.7967	0.8180	0.8521	0.7615	0.8160	0.8268	0.8233	0.8031	0.7988	0.03793
SETDB1	0.8159	0.7833	0.8259	0.7736	0.7741	0.8167	0.8266	0.8071	0.8213	0.8088	0.8053	0.02069
正樣本數量為 2000(1:1)，訓練與測試資料之正樣本分別為 1750 與 250												
ATF2	0.9999	0.9994	1.0000	0.9993	0.9993	1.0000	0.9981	1.0000	1.0000	0.9995	0.9996	0.00059
ATF3	0.8122	0.8835	0.8835	0.8834	0.8726	0.8695	0.9093	0.9106	0.8923	0.9103	0.8827	0.02904
CBX2	0.6752	0.7255	0.6591	0.6891	0.6815	0.7012	0.6499	0.7141	0.6978	0.7051	0.6899	0.02383
CTCF	0.9659	0.9607	0.957	0.9614	0.9624	0.9677	0.9586	0.9541	0.9587	0.9680	0.9615	0.00463
GATA1	0.9857	0.9887	0.9861	0.9739	0.9820	0.9823	0.9766	0.9838	0.9820	0.9826	0.9824	0.00438
GATA2	0.9834	0.9705	0.9826	0.9714	0.9784	0.9689	0.9713	0.9700	0.9732	0.9644	0.9734	0.00614

IRF1	0.9420	0.9463	0.9452	0.9448	0.9408	0.9412	0.9456	0.9433	0.9453	0.9393	0.9434	0.00241
JUN	0.9500	0.9501	0.9517	0.9580	0.9363	0.9388	0.9526	0.9582	0.9461	0.9409	0.9483	0.00762
RNF2	0.8152	0.8434	0.8320	0.7748	0.8038	0.8131	0.7967	0.7986	0.8044	0.8007	0.8083	0.01917
SETDB1	0.8238	0.8326	0.8606	0.8194	0.7876	0.7713	0.8598	0.7878	0.8292	0.8045	0.8177	0.03006



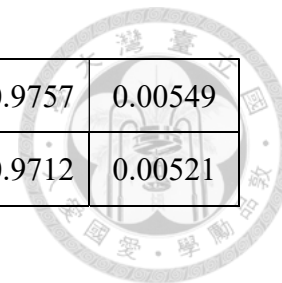
附表 A-2 兩種轉錄因子於五種細胞株中，使用不同正樣本峰值數目的 ezGeno 預測準確度

		AUC												
TF		1	2	3	4	5	6	7	8	9	10	AVG	STDEV	
正樣本數量為 500(1:1)，訓練與測試資料之正樣本分別為 250 與 250														
JUND	HepG2	0.9308	0.9700	0.9843	0.8095	0.8792	0.8524	0.9619	0.9748	0.7328	0.8276	0.8923	0.08556	
	K562	0.9596	0.9506	0.6235	0.9424	0.7710	0.8894	0.9642	0.8180	0.9470	0.9573	0.8823	0.11270	
	GM12878	0.6195	0.5502	0.5322	0.5758	0.5356	0.5221	0.5552	0.4979	0.5789	0.5671	0.5535	0.03424	
	MCF-7	0.7328	0.6803	0.8251	0.8177	0.9122	0.6913	0.8797	0.8976	0.8372	0.5039	0.7778	0.12658	
	A549	0.5965	0.5273	0.8712	0.8540	0.8220	0.8378	0.8495	0.8759	0.4724	0.5043	0.7211	0.17204	
ELF1	HepG2	0.9697	0.7978	0.9491	0.9234	0.9457	0.9105	0.9373	0.9381	0.7685	0.9532	0.9093	0.06876	
	K562	0.9676	0.9528	0.9610	0.8661	0.9633	0.9700	0.9738	0.9679	0.9591	0.9514	0.9533	0.03148	
	GM12878	0.8844	0.8998	0.8855	0.7721	0.9481	0.8151	0.9312	0.8410	0.6693	0.6564	0.8303	0.10261	
	MCF-7	0.9373	0.9377	0.9121	0.9340	0.9172	0.9539	0.9537	0.7592	0.8919	0.9153	0.9112	0.05683	
	A549	0.7793	0.7775	0.9237	0.9206	0.6599	0.9438	0.9112	0.7948	0.9496	0.8059	0.8466	0.09670	
正樣本數量為 1000(1:1)，訓練與測試資料之正樣本分別為 750 與 250														

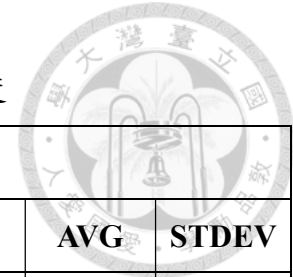
JUND	HepG2	0.9918	0.9912	0.9897	0.9947	0.8532	0.9919	0.9937	0.9914	0.9904	0.985	0.9773	0.04368
	K562	0.9746	0.9765	0.9715	0.9783	0.9726	0.9676	0.9828	0.9678	0.9680	0.9727	0.9732	0.00497
	GM12878	0.5926	0.7490	0.7511	0.7277	0.6802	0.6519	0.6450	0.6919	0.6421	0.7891	0.6921	0.06126
	MCF-7	0.9462	0.9473	0.8567	0.9484	0.9268	0.9366	0.9187	0.9356	0.9334	0.8438	0.9194	0.03770
	A549	0.8821	0.8264	0.8839	0.8443	0.8736	0.8765	0.8156	0.9095	0.8215	0.9032	0.8637	0.03420
ELF1	HepG2	0.9782	0.9664	0.9786	0.9855	0.9829	0.9878	0.9795	0.9837	0.9812	0.9821	0.9806	0.00583
	K562	0.9900	0.9909	0.9816	0.9805	0.9864	0.9838	0.9838	0.9793	0.9836	0.9828	0.9843	0.00381
	GM12878	0.9730	0.9735	0.9654	0.9612	0.9470	0.9698	0.9675	0.9741	0.9686	0.9630	0.9663	0.00807
	MCF-7	0.9706	0.9620	0.9645	0.9691	0.9531	0.9578	0.9756	0.9662	0.9739	0.9614	0.9654	0.00711
	A549	0.9700	0.9561	0.9680	0.9662	0.9576	0.9619	0.9614	0.9647	0.9416	0.9687	0.9616	0.00843
正樣本數量為 1500(1:1)，訓練與測試資料之正樣本分別為 1250 與 250													
JUND	HepG2	0.9946	0.9966	0.9931	0.9943	0.9978	0.9982	0.9922	0.9949	0.9931	0.9971	0.9952	0.00212
	K562	0.9772	0.9822	0.9761	0.9874	0.9821	0.9798	0.9836	0.9801	0.9766	0.9784	0.9804	0.00354
	GM12878	0.7695	0.7766	0.7806	0.7850	0.8015	0.7966	0.7595	0.7720	0.7812	0.7812	0.7804	0.01234
	MCF-7	0.9520	0.9526	0.8417	0.9484	0.9465	0.9497	0.9463	0.9374	0.9524	0.8613	0.9288	0.04126

	A549	0.8858	0.9007	0.8816	0.8999	0.9076	0.9123	0.9112	0.8945	0.9151	0.9196	0.9028	0.01264
ELF1	HepG2	0.9831	0.9913	0.9909	0.9858	0.9863	0.9900	0.9855	0.9853	0.9807	0.9846	0.9864	0.00344
	K562	0.9907	0.9900	0.9822	0.9808	0.9896	0.9875	0.9845	0.9900	0.9859	0.9906	0.9872	0.00366
	GM12878	0.9607	0.9737	0.9769	0.9813	0.9677	0.9704	0.9768	0.9835	0.9783	0.9720	0.9741	0.00676
	MCF-7	0.9639	0.9748	0.9763	0.9685	0.9705	0.9852	0.9731	0.9739	0.9810	0.9778	0.9745	0.00612
	A549	0.9669	0.9784	0.9520	0.9676	0.9652	0.9625	0.9783	0.9638	0.9722	0.9526	0.966	0.00907
正樣本數量為 2000(1:1)，訓練與測試資料之正樣本分別為 1750 與 250													
JUND	HepG2	0.9958	0.9922	0.9937	0.9981	0.9906	0.9946	0.9930	0.993	0.9905	0.9941	0.9936	0.00230
	K562	0.9808	0.9832	0.9819	0.9773	0.9795	0.9820	0.9830	0.9781	0.9840	0.9831	0.9813	0.00230
	GM12878	0.8072	0.7691	0.7643	0.8134	0.8071	0.7625	0.7491	0.7991	0.7713	0.8001	0.7843	0.02327
	MCF-7	0.9520	0.9515	0.9488	0.9602	0.9594	0.9463	0.9516	0.9488	0.9451	0.9542	0.9518	0.00503
	A549	0.9276	0.9118	0.9055	0.8889	0.9222	0.9119	0.8972	0.9169	0.8907	0.9039	0.9077	0.01290
ELF1	HepG2	0.9864	0.9864	0.9885	0.9888	0.9840	0.9826	0.9878	0.9839	0.9810	0.9925	0.9862	0.00341
	K562	0.9904	0.9895	0.9879	0.9857	0.9904	0.9825	0.9916	0.9908	0.9835	0.9917	0.9884	0.00337
	GM12878	0.9796	0.9737	0.9719	0.9690	0.9748	0.9752	0.9742	0.9767	0.9664	0.9739	0.9735	0.00374

	MCF-7	0.9768	0.9754	0.9775	0.9764	0.9824	0.9752	0.9742	0.9789	0.9784	0.9615	0.9757	0.00549
	A549	0.9752	0.9644	0.9675	0.9746	0.9748	0.9744	0.9628	0.9752	0.9762	0.9665	0.9712	0.00521



附表 A-3 轉錄因子 SIN3A 於五種細胞株中，使用不同正樣本峰值數目的 ezGeno 預測準確度



Cell line	Sequence	AUC											
		1	2	3	4	5	6	7	8	9	10	AVG	STDEV
HepG2	2000	0.8793	0.8598	0.7978	0.8002	0.8405	0.8439	0.8325	0.8207	0.8452	0.8387	0.8359	0.02367
	4000	0.8617	0.8910	0.8784	0.8412	0.8585	0.8681	0.8735	0.8339	0.8661	0.8914	0.8664	0.01790
	6000	0.8383	0.8770	0.8719	0.8987	0.8834	0.8554	0.8536	0.8782	0.8651	0.8431	0.8665	0.01798
	8000	0.8598	0.8462	0.8618	0.8738	0.8670	0.8792	0.8844	0.8769	0.8860	0.8704	0.8706	0.01164
	10000	0.8651	0.8630	0.8894	0.9006	0.8756	0.8859	0.8608	0.8480	0.8896	0.8759	0.8754	0.01539
K562	2000	0.8543	0.8517	0.8369	0.8845	0.8569	0.8172	0.8438	0.8518	0.8734	0.8482	0.8519	0.01750
	4000	0.8662	0.8590	0.8447	0.8448	0.8760	0.8882	0.8667	0.8342	0.8641	0.8611	0.8605	0.01511
	6000	0.8684	0.8164	0.8483	0.8547	0.8631	0.8699	0.8560	0.8648	0.8828	0.8316	0.8556	0.01848
	8000	0.8838	0.8950	0.8878	0.8551	0.8581	0.8421	0.8819	0.8550	0.8887	0.8429	0.8690	0.01928
	10000	0.8721	0.8366	0.8671	0.9017	0.8621	0.8859	0.8821	0.8751	0.8666	0.8642	0.8714	0.01627
GM12878	2000	0.8684	0.7543	0.8264	0.8081	0.8127	0.8057	0.8160	0.8057	0.8659	0.7696	0.8133	0.03393
	4000	0.7796	0.8269	0.7763	0.7676	0.7767	0.7796	0.8119	0.7665	0.8102	0.8279	0.7923	0.02295

	6000	0.8450	0.7874	0.7884	0.7942	0.8143	0.8150	0.8246	0.8107	0.8353	0.8081	0.8123	0.01817
	8000	0.8169	0.8112	0.8159	0.8328	0.8606	0.7924	0.8176	0.8268	0.8171	0.7726	0.8164	0.02203
	10000	0.8201	0.8300	0.8409	0.8321	0.8408	0.8085	0.8175	0.8394	0.8090	0.8284	0.8267	0.01174
MCF-7	2000	0.8616	0.8276	0.8392	0.8166	0.8102	0.8409	0.8180	0.8072	0.8358	0.8478	0.8305	0.01671
	4000	0.8296	0.8399	0.8529	0.8659	0.8487	0.8638	0.8406	0.8868	0.8564	0.8665	0.8551	0.01570
	6000	0.8873	0.8747	0.8743	0.8976	0.8712	0.8744	0.8865	0.8962	0.8736	0.8461	0.8782	0.01412
	8000	0.8639	0.8610	0.8764	0.8666	0.8589	0.8780	0.8595	0.8952	0.8760	0.9134	0.8749	0.01668
	10000	0.8909	0.8929	0.8882	0.8735	0.8644	0.9105	0.8757	0.8863	0.9004	0.9047	0.8888	0.01374
A549	2000	0.7480	0.7869	0.7675	0.7642	0.7641	0.7835	0.7631	0.7474	0.7748	0.7699	0.7669	0.01231
	4000	0.7743	0.8061	0.7735	0.7934	0.7567	0.7725	0.8172	0.7833	0.7783	0.7554	0.7811	0.01878
	6000	0.8068	0.7905	0.7825	0.8086	0.7788	0.7744	0.7932	0.7555	0.7911	0.7724	0.7854	0.01533
	8000	0.7516	0.7901	0.7535	0.7710	0.7898	0.7762	0.7983	0.7620	0.7877	0.7836	0.7764	0.01550
	10000	0.7318	0.8059	0.7872	0.7927	0.7691	0.8261	0.7923	0.7598	0.7706	0.7879	0.7823	0.02471

附表 A-4 RAD21(MCF-7)、CEBPB(HepG2)、MAFK(HepG2)和 MAFK(A549)，使用不同正樣本峰值數目的 ezGeno 預測準確度

TF (cell line)	Sequence	AUC score											
		1	2	3	4	5	6	7	8	9	10	AVG	STDEV
RAD21 (MCF-7)	2000	0.9900	0.9823	0.9823	0.9847	0.9873	0.9921	0.9882	0.9860	0.9840	0.9851	0.9862	0.00305
	4000	0.9944	0.9973	0.9904	0.9947	0.9944	0.9962	0.9947	0.9982	0.9953	0.9936	0.9949	0.00203
	6000	0.9938	0.9979	0.9955	0.9937	0.9952	0.9951	0.9932	0.9936	0.9976	0.9946	0.9950	0.00155
	8000	0.9969	0.9949	0.9933	0.9934	0.9971	0.9983	0.9983	0.9974	0.9970	0.9970	0.9964	0.00175
	10000	0.9963	0.9960	0.9964	0.9918	0.9979	0.9931	0.9981	0.9956	0.9981	0.9971	0.9960	0.00200
	20000	0.9970	0.9953	0.9970	0.9983	0.9984	0.9985	0.9952	0.9973	0.9949	0.9963	0.9968	0.00129
	30000	0.9973	0.9992	0.9959	0.9980	0.9975	0.9987	0.9980	0.9960	0.9965	0.9972	0.9974	0.00104
	40000	0.9984	0.9965	0.9990	0.9994	0.9958	0.9987	0.9940	0.9991	0.9946	0.9983	0.9974	0.00189
	50000	0.9982	0.9979	0.9993	0.9975	0.9984	0.9991	0.9998	0.9951	0.9977	0.9952	0.9978	0.00150
CEBPB (HepG2)	2000	0.9906	0.9848	0.9880	0.9877	0.9918	0.9895	0.9898	0.9851	0.9940	0.9916	0.9893	0.00279
	4000	0.9923	0.9959	0.9990	0.9917	0.9960	0.9885	0.9922	0.9939	0.9925	0.9918	0.9934	0.00278
	6000	0.9959	0.9922	0.9962	0.9957	0.9946	0.9934	0.9920	0.9962	0.9909	0.9887	0.9936	0.00245

	8000	0.9922	0.9896	0.9895	0.9920	0.9977	0.9861	0.9936	0.9922	0.9904	0.9888	0.9912	0.00298
	10000	0.9961	0.9919	0.9938	0.9892	0.9969	0.9917	0.9939	0.9918	0.9958	0.9876	0.9929	0.00286
	20000	0.9963	0.9968	0.9919	0.9947	0.9938	0.9907	0.9940	0.9930	0.9954	0.9910	0.9938	0.00201
	30000	0.9953	0.9949	0.9979	0.9902	0.9973	0.9909	0.9944	0.9922	0.9913	0.9919	0.9936	0.00258
	40000	0.9887	0.9952	0.9983	0.9983	0.9968	0.9962	0.9959	0.9908	0.9958	0.9961	0.9952	0.00293
	50000	0.9978	0.9972	0.9975	0.9968	0.9980	0.9966	0.9980	0.9907	0.9941	0.9881	0.9955	0.00328
MAFK (HepG2)	2000	0.9998	0.9991	0.9990	0.9998	0.9991	0.9977	0.9973	0.9994	0.9974	0.9977	0.9986	0.00094
	4000	0.9996	0.9995	0.9990	0.9997	0.9978	1.0000	0.9999	0.9994	0.9998	0.9955	0.9990	0.00132
	6000	0.9995	0.9959	0.9998	1.0000	0.9998	0.9995	0.9999	0.9997	0.9978	0.9951	0.9987	0.00172
	8000	0.9979	1.0000	1.0000	0.9998	0.9999	0.9999	1.0000	0.9998	0.9996	0.9978	0.9995	0.00082
	10000	0.9971	1.0000	1.0000	0.9998	0.9999	0.9998	0.9976	1.0000	0.9997	0.9999	0.9994	0.00103
	20000	0.9997	0.9977	1.0000	1.0000	0.9999	1.0000	0.9999	1.0000	0.9998	0.9999	0.9997	0.00067
	30000	1.0000	1.0000	0.9999	1.0000	0.9999	1.0000	0.9999	1.0000	0.9990	0.9995	0.9998	0.00031
	40000	0.9988	1.0000	0.9999	0.9997	0.9995	0.9997	1.0000	1.0000	0.9999	1.0000	0.9998	0.00036
	50000	1.0000	0.9999	1.0000	0.9999	1.0000	1.0000	0.9999	0.9995	0.9999	1.0000	0.9999	0.00014

MAFK (A549)	2000	0.9769	0.9777	0.9768	0.9771	0.9800	0.9773	0.9738	0.9741	0.9831	0.9711	0.9768	0.00316
	4000	0.9926	0.9902	0.9962	0.9929	0.9929	0.9917	0.9954	0.9894	0.9942	0.9962	0.9932	0.00224
	6000	0.9959	0.9938	0.9860	0.9937	0.9934	0.9939	0.9942	0.9947	0.9909	0.9925	0.9929	0.00262
	8000	0.9953	0.9972	0.9938	0.9936	0.9942	0.9938	0.9947	0.9922	0.9942	0.9949	0.9944	0.00124
	10000	0.9957	0.9916	0.9940	0.9949	0.9971	0.9979	0.9952	0.9977	0.9963	0.9968	0.9957	0.00182
	20000	0.9989	0.9976	0.9964	0.9949	0.9984	0.9973	0.9949	0.9990	0.9944	0.9975	0.9969	0.00162
	30000	0.9959	0.9991	0.9932	0.9939	0.9945	0.9980	0.9980	0.9911	0.9964	0.9969	0.9957	0.00237
	40000	0.9976	0.9927	0.9958	0.9960	0.9958	0.9977	0.9982	0.9962	0.9958	0.9947	0.9961	0.00152
	50000	0.9953	0.9984	0.9991	0.9985	0.9993	0.9996	0.9979	0.9954	0.9998	0.9980	0.9981	0.00152

附錄B



附表 B-1 二十四種轉錄因子在五種細胞株中，使用 2000 正樣本峰值數目 1750 作為訓練資料的 ezGeno 預測準確度

TF	Cell line	AUC											
		1	2	3	4	5	6	7	8	9	10	AVG	STDEV
RAD21	HepG2	0.9955	0.9972	0.9968	0.9961	0.9975	0.9974	0.9988	0.9972	0.9967	0.998	0.9971	0.00093
	K562	0.9673	0.9988	0.9972	0.9970	0.9961	0.9979	0.9966	0.9961	0.9954	0.996	0.9938	0.009379
	GM12878	0.9978	0.9969	0.9987	0.9936	0.9977	0.9996	0.9988	0.9934	0.9977	0.9981	0.9972	0.002101
	MCF-7	0.9911	0.9906	0.9938	0.9924	0.9911	0.9904	0.9925	0.989	0.9905	0.9901	0.9912	0.001391
	A549	0.9478	0.9424	0.9662	0.9531	0.9416	0.9513	0.9627	0.9715	0.9607	0.9397	0.9537	0.011126
MYC	HepG2	0.9285	0.9254	0.9206	0.9195	0.9231	0.9176	0.9177	0.8924	0.9222	0.9153	0.9182	0.009888
	K562	0.9106	0.9217	0.9159	0.9177	0.9346	0.9368	0.9339	0.9267	0.9352	0.9179	0.9251	0.009558
	GM12878	0.8303	0.8147	0.8206	0.8318	0.8252	0.8223	0.8201	0.8048	0.8181	0.8315	0.8219	0.008402
	MCF-7	0.9399	0.9323	0.9339	0.9457	0.9473	0.9319	0.9382	0.9380	0.9320	0.9311	0.9370	0.005870
	A549	0.8581	0.904	0.8924	0.9085	0.8988	0.8904	0.9098	0.8886	0.8884	0.8969	0.8936	0.014737

ELK1	HepG2	0.8300	0.8383	0.8066	0.8598	0.8151	0.8313	0.8307	0.8400	0.8414	0.8460	0.8339	0.015133
	K562	0.9590	0.9595	0.9574	0.9523	0.9438	0.9579	0.9507	0.9543	0.9563	0.9512	0.9542	0.004859
	GM12878	0.9287	0.9301	0.9420	0.9278	0.9456	0.9455	0.9131	0.9256	0.9295	0.9082	0.9296	0.012530
	MCF-7	0.9580	0.9502	0.9503	0.9438	0.955	0.9408	0.941	0.9584	0.9417	0.9617	0.9501	0.007955
	A549	0.6930	0.8689	0.8738	0.8154	0.5983	0.8420	0.6982	0.8326	0.8762	0.7917	0.7890	0.094528
POLR2A	HepG2	0.7432	0.7332	0.7595	0.7308	0.7219	0.7585	0.7824	0.7402	0.7618	0.7728	0.7504	0.019585
	K562	0.8134	0.8397	0.7938	0.8199	0.8215	0.8098	0.8598	0.7777	0.7957	0.8475	0.8179	0.025603
	GM12878	0.7763	0.6851	0.7434	0.7565	0.7567	0.7964	0.6819	0.7420	0.7485	0.7961	0.7483	0.039422
	MCF-7	0.8310	0.8201	0.7998	0.7973	0.7915	0.7949	0.8233	0.7993	0.8057	0.8095	0.8072	0.013385
	A549	0.7994	0.7114	0.8088	0.7440	0.7771	0.7325	0.7175	0.8079	0.7370	0.8041	0.7640	0.039465
MAFK	HepG2	0.9989	0.9996	0.9976	0.9994	0.9983	0.9983	0.9990	0.9988	0.9965	0.9989	0.9985	0.000917
	K562	0.9983	0.9990	0.9995	0.9974	0.9991	0.9976	0.9997	0.9968	0.9990	0.9991	0.9986	0.000977
	GM12878	0.988	0.9855	0.9900	0.9867	0.9899	0.9861	0.9897	0.9854	0.9902	0.9897	0.9881	0.002011
	MCF-7	0.9966	0.9983	0.9983	0.9978	0.9977	0.9965	0.9944	0.9995	0.9992	0.9969	0.9975	0.001494
	A549	0.9940	0.9908	0.9932	0.9922	0.9911	0.9925	0.9891	0.9869	0.9877	0.9922	0.9910	0.002365

CREB1	HepG2	0.9651	0.9504	0.9535	0.9752	0.9670	0.9661	0.9541	0.9646	0.9503	0.9716	0.9618	0.009008
	K562	0.9815	0.9745	0.9774	0.9763	0.9786	0.9720	0.9771	0.9614	0.9715	0.9707	0.9741	0.005617
	GM12878	0.9763	0.9749	0.9727	0.9718	0.9507	0.9724	0.9712	0.9695	0.9773	0.9710	0.9708	0.007471
	MCF-7	0.9974	0.9988	0.9960	0.9953	0.9963	0.9970	0.9979	0.9965	0.9994	0.9977	0.9972	0.001272
	A549	0.9371	0.9407	0.9395	0.9399	0.9527	0.9308	0.9340	0.9394	0.9478	0.9320	0.9394	0.006771
SIN3A	HepG2	0.7855	0.8357	0.8099	0.8496	0.8192	0.8302	0.8112	0.8293	0.8160	0.8270	0.8214	0.017417
	K562	0.8904	0.8576	0.8466	0.8691	0.8718	0.8501	0.8661	0.8577	0.8922	0.8506	0.8652	0.016072
	GM12878	0.8213	0.8022	0.8264	0.8563	0.8174	0.8663	0.8656	0.8960	0.8273	0.8791	0.8458	0.030872
	MCF-7	0.8498	0.8354	0.8458	0.8256	0.8296	0.8358	0.8775	0.8601	0.8728	0.7924	0.8425	0.024887
	A549	0.7813	0.7376	0.7249	0.6963	0.7481	0.7513	0.7643	0.7768	0.7467	0.7246	0.7452	0.025820
YY1	HepG2	0.9684	0.9694	0.9772	0.9761	0.9761	0.9772	0.9773	0.9745	0.9794	0.9776	0.9753	0.003616
	K562	0.8937	0.9040	0.8761	0.8798	0.9014	0.8968	0.9012	0.8931	0.8873	0.8819	0.8915	0.009819
	GM12878	0.9803	0.9850	0.9846	0.9816	0.9861	0.9854	0.9801	0.9837	0.9827	0.9845	0.9834	0.002130
	A549	0.9870	0.9826	0.9887	0.9854	0.9898	0.9852	0.9875	0.9841	0.9873	0.9862	0.9864	0.002146
	HEK293	0.9919	0.9930	0.9894	0.9898	0.9901	0.9903	0.9943	0.9878	0.9830	0.9910	0.9901	0.003098

JUND	HepG2	0.9958	0.9922	0.9937	0.9981	0.9906	0.9946	0.9930	0.993	0.9905	0.9941	0.9936	0.002300
	K562	0.9808	0.9832	0.9819	0.9773	0.9795	0.9820	0.9830	0.9781	0.9840	0.9831	0.9813	0.002300
	GM12878	0.8072	0.7691	0.7643	0.8134	0.8071	0.7625	0.7491	0.7991	0.7713	0.8001	0.7843	0.023271
	MCF-7	0.9520	0.9515	0.9488	0.9602	0.9594	0.9463	0.9516	0.9488	0.9451	0.9542	0.9518	0.005026
	A549	0.9276	0.9118	0.9055	0.8889	0.9222	0.9119	0.8972	0.9169	0.8907	0.9039	0.9077	0.012898
TAF1	HepG2	0.9039	0.816	0.8611	0.8243	0.7860	0.8447	0.8192	0.7745	0.8269	0.8744	0.8331	0.039286
	K562	0.8894	0.8875	0.8588	0.8951	0.8630	0.8492	0.8764	0.8920	0.9013	0.8681	0.8781	0.017560
	GM12878	0.8560	0.8840	0.8601	0.8518	0.8647	0.8967	0.8453	0.8588	0.8284	0.8684	0.8614	0.019133
	MCF-7	0.8259	0.8338	0.8468	0.8598	0.8483	0.8260	0.8607	0.8397	0.8220	0.7941	0.8357	0.020044
	A549	0.8673	0.8810	0.8943	0.8937	0.8824	0.8661	0.9007	0.8801	0.9075	0.8844	0.8858	0.013445
TCF12	HepG2	0.8451	0.757	0.8528	0.8371	0.8376	0.8408	0.8558	0.8449	0.7658	0.8468	0.8284	0.035849
	K562	0.9622	0.9721	0.9690	0.9635	0.9808	0.9738	0.9681	0.9612	0.9763	0.9669	0.9694	0.006379
	GM12878	0.8371	0.9544	0.9634	0.9416	0.9597	0.9428	0.9540	0.9412	0.9432	0.9325	0.9370	0.036386
	MCF-7	0.8468	0.8370	0.8630	0.8323	0.8669	0.8554	0.8849	0.8508	0.8845	0.8820	0.8604	0.019252
	A549	0.8864	0.8742	0.8839	0.9023	0.9168	0.8672	0.8949	0.8842	0.8562	0.8881	0.8854	0.017266

EP300	HepG2	0.7265	0.8073	0.8626	0.8468	0.8591	0.8489	0.7852	0.7308	0.7451	0.7387	0.7951	0.056753
	K562	0.9435	0.9613	0.9429	0.9397	0.9355	0.9505	0.9382	0.9363	0.9162	0.9429	0.9407	0.011508
	GM12878	0.7856	0.7739	0.8982	0.8998	0.8730	0.8872	0.8631	0.8311	0.8560	0.8376	0.8506	0.043932
	MCF-7	0.9180	0.9251	0.9130	0.9273	0.9246	0.9194	0.9249	0.9093	0.9045	0.9077	0.9174	0.008268
	A549	0.8245	0.8248	0.8694	0.8304	0.8713	0.8256	0.8223	0.8801	0.8544	0.8419	0.8445	0.022478
SP1	HepG2	0.8601	0.8407	0.8812	0.8659	0.8512	0.8447	0.8067	0.8740	0.8525	0.9111	0.8588	0.027605
	K562	0.8941	0.9004	0.8885	0.9056	0.9012	0.9011	0.9176	0.8935	0.9054	0.8882	0.8996	0.008942
	GM12878	0.9598	0.9589	0.968	0.9501	0.9501	0.9452	0.9596	0.9381	0.9571	0.9409	0.9528	0.009499
	MCF-7	0.6932	0.7134	0.6667	0.7016	0.7980	0.6868	0.7053	0.6475	0.8167	0.6736	0.7103	0.054933
	A549	0.8185	0.7139	0.8102	0.8449	0.7459	0.7090	0.8524	0.7772	0.7892	0.8412	0.7902	0.052926
CEBPB	HepG2	0.9866	0.9862	0.9927	0.9932	0.9961	0.9839	0.9940	0.9918	0.9831	0.9895	0.9897	0.004524
	K562	0.9885	0.9957	0.9914	0.9939	0.9938	0.9923	0.9909	0.9918	0.9835	0.9922	0.9914	0.003386
	GM12878	0.8920	0.8656	0.8963	0.8660	0.8365	0.8628	0.8700	0.8083	0.8844	0.7941	0.8576	0.034355
	MCF-7	0.9956	0.9957	0.9946	0.9931	0.9956	0.9960	0.9899	0.9897	0.9972	0.9926	0.9940	0.002597
	A549	0.9869	0.9843	0.9811	0.9846	0.9878	0.9866	0.9813	0.988	0.9897	0.9915	0.9862	0.003383

MAX	HepG2	0.9767	0.976	0.9752	0.9723	0.9709	0.9799	0.9779	0.9704	0.9766	0.9689	0.9745	0.003631
	K562	0.9664	0.9594	0.9679	0.9702	0.9793	0.9769	0.9746	0.9791	0.9694	0.9651	0.9708	0.006543
	GM12878	0.9538	0.9584	0.9637	0.9625	0.9554	0.9644	0.9554	0.9419	0.9676	0.9668	0.959	0.007775
	MCF-7	0.946	0.9424	0.9606	0.9419	0.9566	0.9326	0.9461	0.9431	0.9272	0.9401	0.9437	0.00987
	A549	0.9376	0.9312	0.9406	0.9528	0.9426	0.9235	0.9508	0.9508	0.9439	0.9384	0.9412	0.009205
RCOR1	HepG2	0.768	0.7991	0.7746	0.7528	0.7894	0.6542	0.7792	0.7813	0.7189	0.8099	0.7627	0.045756
	K562	0.8668	0.8631	0.8482	0.8386	0.8613	0.8314	0.8523	0.8651	0.8568	0.8401	0.8524	0.012392
	GM12878	0.7837	0.7852	0.8577	0.8391	0.8217	0.8183	0.8291	0.8291	0.8101	0.8305	0.8205	0.022834
	MCF-7	0.8389	0.779	0.7871	0.841	0.7939	0.7968	0.6834	0.8118	0.8117	0.753	0.7897	0.045778
	A549	0.7656	0.7287	0.7119	0.814	0.7196	0.7586	0.6324	0.8202	0.6025	0.7407	0.7294	0.069608
ELF1	HepG2	0.9864	0.9864	0.9885	0.9888	0.984	0.9826	0.9878	0.9839	0.981	0.9925	0.9862	0.003408
	K562	0.9904	0.9895	0.9879	0.9857	0.9904	0.9825	0.9916	0.9908	0.9835	0.9917	0.9884	0.003374
	GM12878	0.9796	0.9737	0.9719	0.969	0.9748	0.9752	0.9742	0.9767	0.9664	0.9739	0.9735	0.003741
	MCF-7	0.9768	0.9754	0.9775	0.9764	0.9824	0.9752	0.9742	0.9789	0.9784	0.9615	0.9757	0.005493
	A549	0.9752	0.9644	0.9675	0.9746	0.9748	0.9744	0.9628	0.9752	0.9762	0.9665	0.9712	0.00521

ZBTB33	HepG2	0.9496	0.9551	0.9537	0.934	0.9571	0.9378	0.9445	0.9489	0.9509	0.9323	0.9464	0.008884
	K562	0.9424	0.938	0.9407	0.9396	0.9459	0.931	0.9458	0.9346	0.9356	0.9452	0.9399	0.005113
	GM12878	0.972	0.9712	0.9633	0.9604	0.9666	0.9666	0.9597	0.9633	0.9625	0.9676	0.9653	0.004211
	MCF-7	0.9099	0.8718	0.8833	0.9022	0.8772	0.8822	0.8804	0.9135	0.9056	0.8923	0.8918	0.014925
	A549	0.8777	0.9149	0.896	0.9072	0.9128	0.9001	0.8779	0.8639	0.908	0.901	0.896	0.017139
HDAC2	HepG2	0.8384	0.8486	0.7295	0.8429	0.8196	0.8264	0.8439	0.8379	0.8116	0.8433	0.8242	0.03534
	K562	0.7757	0.7616	0.7394	0.7591	0.8311	0.8001	0.9023	0.7616	0.7604	0.8057	0.7897	0.04814
	GM12878	0.6328	0.6556	0.6412	0.6487	0.7132	0.7344	0.6999	0.7315	0.6757	0.671	0.6804	0.037292
	MCF-7	0.8206	0.8262	0.8217	0.8606	0.8293	0.7575	0.8143	0.7792	0.814	0.8029	0.8126	0.028216
	A549	0.8391	0.8301	0.8217	0.8086	0.8076	0.8331	0.7776	0.7982	0.827	0.8228	0.8166	0.018672
RFX5	HepG2	0.9804	0.9695	0.9708	0.9776	0.9719	0.9552	0.9332	0.975	0.9681	0.9549	0.9657	0.014189
	K562	0.8798	0.9082	0.8878	0.8502	0.896	0.9037	0.8698	0.9037	0.8745	0.8834	0.8857	0.018084
	GM12878	0.9536	0.9823	0.9713	0.9618	0.9674	0.962	0.9644	0.9808	0.9719	0.9653	0.9681	0.008799
	MCF-7	0.983	0.986	0.9796	0.9757	0.978	0.9801	0.971	0.9853	0.9733	0.9784	0.979	0.004887
	A549	0.9156	0.9562	0.9601	0.9517	0.9634	0.9732	0.9696	0.9515	0.9663	0.9501	0.9558	0.016206

GABPA	HepG2	0.9885	0.9885	0.9928	0.9803	0.9944	0.988	0.9859	0.9916	0.9872	0.9881	0.9885	0.00393
	K562	0.9822	0.9833	0.978	0.9916	0.9894	0.9806	0.9849	0.9922	0.9828	0.9934	0.9858	0.005398
	GM12878	0.9761	0.9669	0.9648	0.9716	0.9784	0.9786	0.9661	0.9781	0.9849	0.9696	0.9735	0.006673
	MCF-7	0.9688	0.9677	0.9712	0.9705	0.9685	0.9798	0.9701	0.9701	0.9644	0.9687	0.97	0.003941
	A549	0.9789	0.9812	0.9794	0.9799	0.9737	0.9801	0.9784	0.9783	0.9684	0.9849	0.9783	0.004454
ESRRA	HepG2	0.9384	0.9458	0.829	0.9365	0.9339	0.9251	0.898	0.9178	0.9447	0.9307	0.92	0.034945
	K562	0.9331	0.9321	0.9266	0.9414	0.9279	0.9285	0.9381	0.9328	0.9183	0.9293	0.9308	0.006381
	GM12878	0.8029	0.8271	0.8354	0.8464	0.8161	0.826	0.8321	0.8388	0.8409	0.8204	0.8286	0.013011
	MCF-7	0.9242	0.8622	0.9168	0.9179	0.9248	0.8635	0.9343	0.9247	0.8929	0.9095	0.9071	0.025799
	A549	0.9111	0.9168	0.9121	0.9146	0.9316	0.9352	0.9193	0.929	0.9365	0.9351	0.9241	0.010316
SREBF1	HepG2	0.8955	0.9177	0.8932	0.8922	0.8839	0.9006	0.9077	0.9255	0.8672	0.8914	0.8975	0.016655
	K562	0.7803	0.7888	0.7752	0.7838	0.8199	0.7801	0.808	0.7925	0.7754	0.7919	0.7896	0.014493
	GM12878	0.7989	0.7484	0.7411	0.7662	0.7769	0.7621	0.7816	0.7935	0.7773	0.7854	0.7731	0.018679
	MCF-7	0.9025	0.9304	0.9099	0.9114	0.9189	0.9029	0.9244	0.9126	0.9188	0.9251	0.9157	0.009427
	A549	0.9186	0.8532	0.8904	0.8659	0.8892	0.8909	0.8998	0.8639	0.8793	0.8901	0.8841	0.019154

ZNF24	HepG2	0.9169	0.9012	0.9275	0.8984	0.8832	0.9094	0.9088	0.8665	0.9201	0.8709	0.9003	0.020726
	K562	0.9746	0.9739	0.976	0.9779	0.9784	0.9719	0.9773	0.9759	0.9805	0.9486	0.9735	0.009085
	GM12878	0.9214	0.9312	0.8809	0.9183	0.9393	0.8632	0.9355	0.8691	0.9519	0.9317	0.9143	0.031469
	MCF-7	0.8912	0.8731	0.9299	0.8973	0.9169	0.9309	0.9201	0.9054	0.9209	0.8924	0.9078	0.019034
	HEK293	0.8536	0.8404	0.7365	0.7476	0.8675	0.8285	0.7403	0.8122	0.7894	0.8301	0.8046	0.048500

附表 B-2 兩種轉錄因子在五種細胞株中，使用 1750 末位峰值作為訓練正樣本的 ezGeno 預測準確度

TF	Cell line	AUC											
		1	2	3	4	5	6	7	8	9	10	AVG	STDEV
RAD21	HepG2	0.7985	0.7436	0.8033	0.6731	0.7983	0.6296	0.6940	0.8031	0.8548	0.6540	0.7452	0.077391
	K562	0.9209	0.9036	0.9059	0.7243	0.9147	0.7319	0.9186	0.9033	0.7641	0.9025	0.8590	0.082888
	GM12878	0.7013	0.8168	0.8061	0.8454	0.5893	0.8091	0.7669	0.8179	0.6218	0.8159	0.7591	0.090219
	MCF-7	0.7766	0.6560	0.7722	0.7571	0.8068	0.8318	0.8025	0.8076	0.8076	0.6665	0.7685	0.060513
	A549	0.9046	0.8856	0.8936	0.9070	0.9016	0.8537	0.9033	0.8436	0.8786	0.9228	0.8894	0.024764
MAFK	HepG2	0.9994	0.9996	0.9992	0.999	0.9967	0.9975	0.9993	0.9996	0.9994	0.9996	0.9989	0.001001
	K562	0.9986	0.9996	0.9998	0.9988	0.9998	0.9993	0.9997	0.9995	0.9995	0.9994	0.9994	0.000406
	GM12878	0.9912	0.9920	0.9943	0.9953	0.9956	0.9921	0.9948	0.9919	0.9922	0.9921	0.9932	0.001649
	MCF-7	0.9994	0.9988	0.9964	0.9995	0.9995	0.9992	0.9992	0.9981	0.9989	0.9991	0.9988	0.000943
	A549	0.9935	0.9903	0.9922	0.9895	0.9935	0.9953	0.9962	0.9936	0.9842	0.9924	0.9921	0.003427

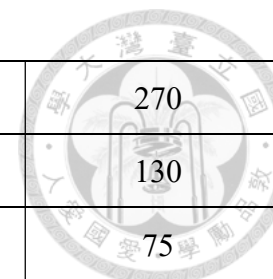
附錄C



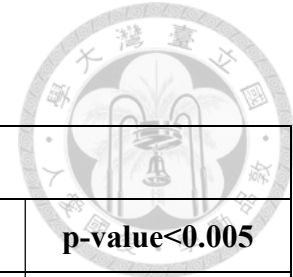
附表 C-1 胸腺組織中的變異位點對 MCF-7 相關模型於結合預測影響之序列數量

TF	Positive			Negative		
	p-value<0.05	p-value<0.01	p-value<0.005	p-value<0.05	p-value<0.01	p-value<0.005
RAD21	316	106	54	302	81	51
MYC	609	411	331	550	373	303
ELK1	518	266	192	458	255	190
POLR2A	507	303	237	483	281	202
MAFK	342	101	70	292	88	40
CREB1	422	219	151	430	189	137
SIN3A	563	347	267	578	372	293
JUND	538	309	247	531	328	248
TAF1	538	362	295	520	323	231
TCF12	605	386	295	561	367	286

EP300	573	365	302	569	343	270
SP1	380	191	133	356	183	130
CEBPB	376	153	93	349	130	75
MAX	579	377	308	580	378	299
RCOR1	464	251	175	440	244	185
ELF1	448	232	158	435	184	132
ZBTB33	546	336	266	561	360	278
HDAC2	516	302	233	512	303	227
RFX5	450	220	142	440	231	159
GABPA	441	215	152	415	203	134
ESRRA	450	275	209	446	271	219
SREBF1	483	267	202	440	261	201
ZNF24	405	215	160	433	220	168
Total	11069	6209	4672	10681	5968	4458

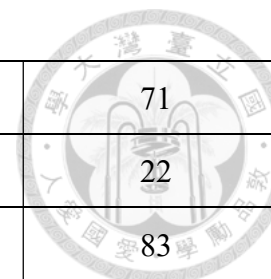


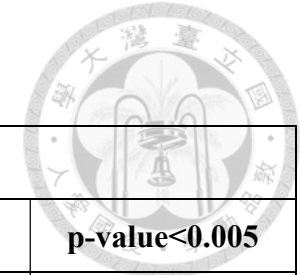
附表 C-2 肝組織中的變異位點對 HepG2 相關模型於結合預測影響之序列數量



TF	Positive			Negative		
	p-value<0.05	p-value<0.01	p-value<0.005	p-value<0.05	p-value<0.01	p-value<0.005
RAD21	74	20	8	71	15	11
MYC	152	97	75	144	89	68
ELK1	152	93	75	139	84	66
POLR2A	164	96	73	156	96	67
MAFK	105	38	20	92	28	17
CREB1	143	73	50	138	68	50
SIN3A	165	95	76	146	88	71
YY1	114	46	27	125	38	24
JUND	95	29	14	77	21	18
TAF1	123	56	38	111	58	39
TCF12	166	105	75	153	92	72
EP300	169	101	70	164	100	79

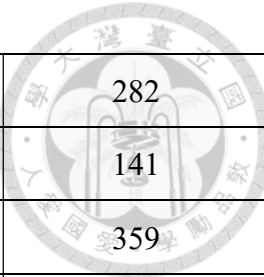
SP1	159	82	63	138	86	71
CEBPB	117	56	31	103	39	22
MAX	168	102	77	170	110	83
RCOR1	132	70	50	109	59	42
ELF1	124	49	37	104	33	18
ZBTB33	138	73	60	141	78	55
HDAC2	156	94	70	168	107	79
RFX5	137	80	51	138	75	49
GABPA	160	97	71	140	75	53
ESRRA	154	95	73	150	94	79
SREBF1	148	81	58	138	75	57
ZNF24	110	54	35	127	52	41
Total	3325	1782	1277	3142	1660	1231





附表 C-3 肺組織中的變異位點對 A549 相關模型於結合預測影響之序列數量

TF	Positive			Negative		
	p-value<0.05	p-value<0.01	p-value<0.005	p-value<0.05	p-value<0.01	p-value<0.005
RAD21	687	373	278	594	318	235
MYC	808	526	406	808	529	397
ELK1	873	600	508	478	167	90
POLR2A	634	348	253	576	292	210
MAFK	479	171	103	410	127	78
CREB1	830	523	424	798	493	385
SIN3A	1416	1326	1289	711	437	317
YY1	549	217	137	532	212	137
JUND	784	477	379	755	440	342
TAF1	890	624	528	872	556	454
TCF12	841	528	423	831	531	419
EP300	833	513	388	799	511	400



SP1	715	409	286	720	396	282
CEBPB	588	254	160	520	218	141
MAX	790	498	381	748	450	359
RCOR1	286	54	28	342	101	48
ELF1	841	531	415	792	472	354
ZBTB33	698	409	317	673	402	310
HDAC2	651	343	250	569	279	213
RFX5	776	505	391	772	484	396
GABPA	688	384	273	649	311	221
ESRRA	777	484	380	774	484	380
SREBF1	764	467	368	712	438	336
Total	17198	10564	8365	15435	8648	6504