中華大學生物資訊學系系統開發專題報告

脂肪細胞分化之主要調控因子鑑定研究

Identification of Key Regulatory Factors for Adipocyte Differentiation.

專題組員:簡品柔、潘潔瑩

專題編號：PROJ2018-BIOINFO-104001

指導老師:黃俊燕老師

1. 摘要

於基因表現資料庫 GEO 及 ArrayExpress 蒐集人類一般器官之脂肪細胞、脂肪幹細胞、胚胎幹細胞的微陣列基因表現數據，探討胚胎幹細胞、脂肪成體幹細胞及脂肪細胞之間的蛋白質調控群以找出調控細胞類型轉換的主要關鍵轉錄因子(Transcription Factor)。透過主成分分析(Principal components analysis，PCA)、變異數分析 (Analysis of variance，ANOVA)階層式分群法(Hierarchical clustering)等方式，將不同樣本間有高度變異性的基因進行分群，接著利用統計學及生物意義的角度去檢視分群的穩定度及驗證其生物功能。

2. 簡介

日本科學家山中伸彌 (Shinya Yamanaka)及高橋和利 (Kazutoshi Takahashi)發表了關於誘導性多能幹細胞(簡稱 iPSC)的研究[5]，於細胞中導入轉錄因子 Oct4、Sox2、c-Myc、Klf4，經過一段時間目標細胞即可轉變成與胚胎幹細胞功能相近的 iPSC，而胚胎幹細胞是目前再生醫學上重點研究，運用胚胎幹細胞極強的分化潛能及自我更新的能力，以解決人類疾病，而本專題以胚胎幹細胞、脂肪幹細胞及脂肪細胞做為研究題材，運用主成分分析等統計學分析，找出這三種細胞類型間分化的主要調控轉錄因子，若是可運用此方式衡量各基因表現量的差異並鑑定更多人類成體細胞，或許未來即可運用身體上的任何一個細胞轉換為近似於胚胎幹細胞性質的細胞，以快速修復人體組織、器官或是治療細胞功能異常的疾病。

3. 專題進行方式

3.1. 資料來源

取收錄在 ArrayExpress 的一個微陣列基因資料集，”E-MTAB-62 - Human gene expression atlas of 5372 samples representing 369 different cell and tissue types, disease states and cell lines”（ https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-62/）該資料庫的微陣列平台為 Affymetrix GeneChip Human Genome HG-U133A [HG-U133A]，資料庫中記錄人類的369種不同的細胞組織，例如：一般組織、成體幹細胞、胚胎幹細胞及疾病狀態組織等的基因表達譜，共有5372筆資料，其中取胚胎幹細胞、脂肪成體幹細胞、脂肪細胞等資料進行分類分析研究。

3.2. 研究方法

上述的資料進行數據前處理及主成分分析以篩選出在不同樣本中表現量高度變異的基因，進行變異數分析及分類分析，整理出胚胎幹細胞、脂肪成體幹細胞、脂肪細胞三種細胞類型之高表達量的蛋白質調控群。

3.2.1. 數據前處理─資料標準化

由於微陣列的探針數據用於表達對應的基因於細胞中的表現量高低，在探針的表現量上通常有高有低，在衡量資料的距離上為了不要讓極大的數字主宰整筆資料，於是我們要先將探針數據進行資料標準化（將資料 X 轉換為 Z 分數之資料矩陣）。

將資料統整為一資料 X，有 M 個樣本、有 N 個基因，X_{ij}代表第 i 個樣本中第 j 個基因的表現量

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{M1} & \cdots & x_{MN} \end{pmatrix}$$

z_{ij}為資料矩陣 X 之 Z 分數數據，μ_j為 j 基因於各樣本表現量之平均值，σ_j為 j 基因於各樣本表現量之標準差

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

$$\mu_j = \frac{x_{1j} + x_{2j} + \cdots + x_{Mj}}{M}$$

$$\sigma_j = \sqrt{\frac{\sum_{k=1}^{M}(x_{ij} - \mu_j)^2}{M-1}}$$

3.2.2. 主成分分析(Principal components analysis，PCA)

主成分分析是一種簡化數據的技術，用於找出資料間最大變異量的特徵，以減少數據資料的維度數(變數)，並保留大部分資料特性，最主要是透過共變異數矩陣(The covariance Matrix)進行特徵分析

$$C = \begin{pmatrix} cov(X_1, X_1) & \cdots & cov(X_1, X_N) \\ \vdots & \ddots & \vdots \\ cov(X_N, X_1) & \cdots & cov(X_N, X_N) \end{pmatrix}$$

其中X_i為資料矩陣X之第i行行向量，得出主成分向量，也就是特徵向量(Eigenvectors)及特徵值(Eigenvalues)。

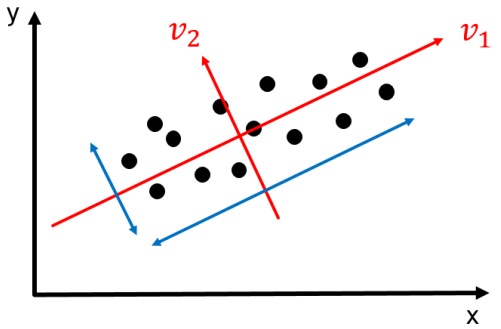
[圖1]為二維資料的資料點散布圖，經過主成分分析後可得出v_1,v_2兩特徵向量，資料點分別投影至v_1,v_2得出該向量對於資料的變異量，目的就是為了找出新的座標軸，進行降維，【圖1】v_1,v_2比較後會發現v_1可得出資料最大變異量。

$Cv = \lambda v$

C為$N \times N$矩陣

λ是特徵值 (Eigenvalues)

v為特徵向量 (Eigenvectors)



【圖1】主成分分析後特徵向量示意圖。

利用統計軟體 R 內建 PCA 涵式

{prcomp}對資料前處理後的數據進行主成分分析後可得出17個主成分，再根據

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}\lambda_i} \geq 0.95$$

取前7個主成分即可表達出原數據的百分之九十五。

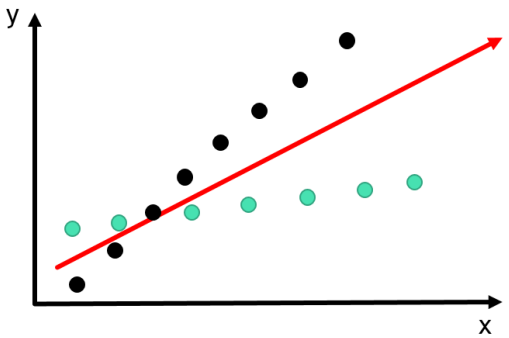
得出主要成分軸 $V = a_1e_1 + a_2e_2 + \cdots + a_Ne_N$ 後，由於主要成分軸大部分的係數a_i與最大係數數值相較之下小了許多，可達10^9的數量級差異，所以我們只保留

$a_i/\max\{a_i\} \geq \frac{1}{4}$ 的係數項。

3.2.2.1. 主成分分析的特性及限制

3.2.2.1.1. 多筆資料下的特徵向量

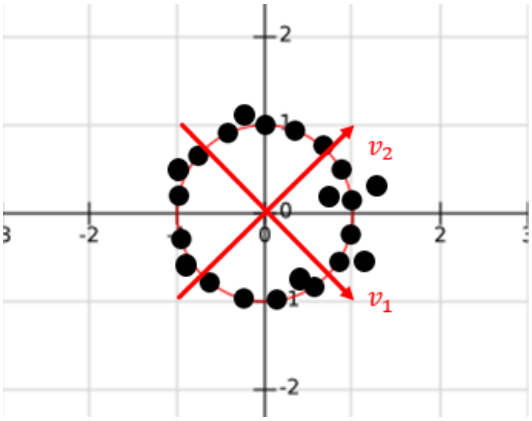
在多變數資料下，利用 PCA 可得到整個數據資料的特徵向量，所得出的最大變異量也是針對整體數據，對於多組線性組合之資料集，PCA 是無法解析出個別組別之線性資料。



【圖2】雙組別線性資料下主成分分析所呈現出的特徵向量示意圖。

3.2.2.1.2. 非線性資料

若資料點散布如【圖3】成非線性圖形，則主成分分析將無法進行資料降維，不論是v_1,v_2，所呈現的數據變異量都相同。

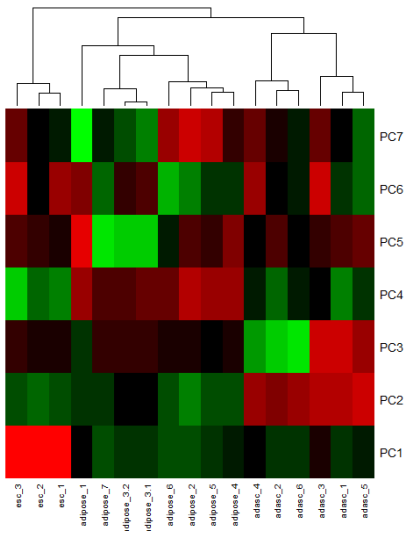


【圖3】非線性函數資料主成分分析特徵向量示意圖。

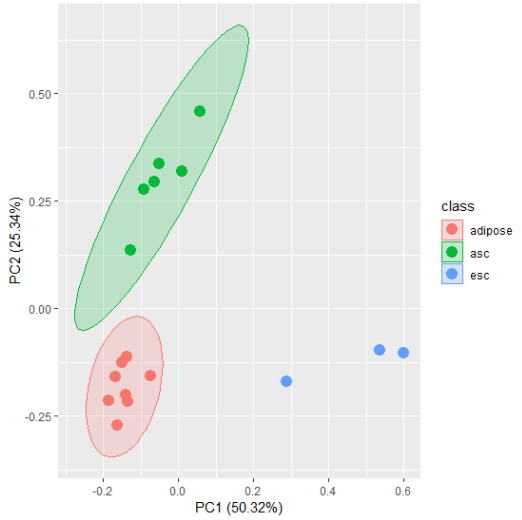
主成分分析後的結果，完全是依照所給的數據資料進行分析，因此 PCA 所獲得的結果與採用的數據有高度相關性。

4. 主要成果

所蒐集的胚胎幹細胞、脂肪成體幹細胞及脂肪細胞經過 PCA 處理分析後得出以下結果。

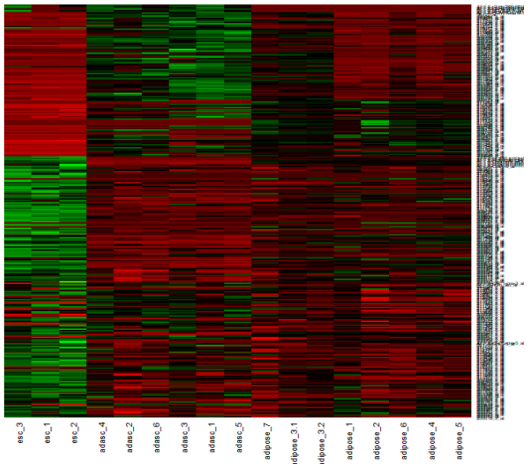


【圖4】主成分分析第1~7主成分熱圖，從熱圖可得知利用第一主成分及第二主成分即可區分出胚胎幹細胞、脂肪成體幹細胞及脂肪細胞三種細胞類型。



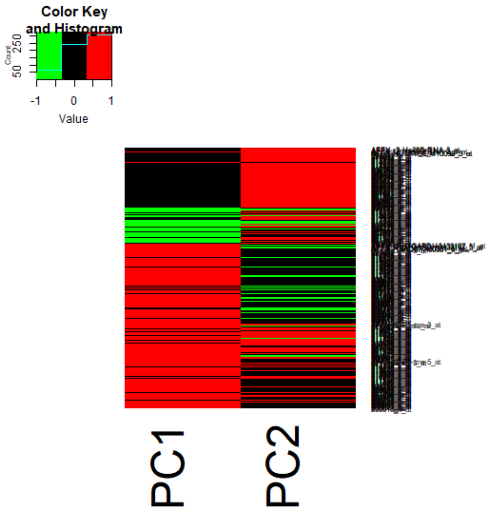
【圖5】利用主成分分析所得出的特徵向量乘上標準化的數據得出點狀圖。根據圖示發現經過主成分分析降維過後，如預期般各個樣本依照類型區分成三個區塊。

從【圖5】可以確定在不同的細胞型態中，各有主要的基因調控網路，故進行階層式分群法對基因進行分類並排序呈現出【圖6】。



【圖6】基因表現熱圖。

確定【圖5】中各個細胞所代表的區域後，三種細胞有不同的表示方式分別為第一主成分及第二主成分數值皆高(脂肪幹細胞)、第一主成分數值偏高而第二主成分數值低(胚胎幹細胞)及第一主成分及第二主成分數值皆低(脂肪細胞)，而細胞分化的主要調控方式可能是因為某個主要調控因子的開啟或關閉導致細胞改變型態分化成不同的細胞，所以我們計算出器官每個探針的係數繪製【圖7】。

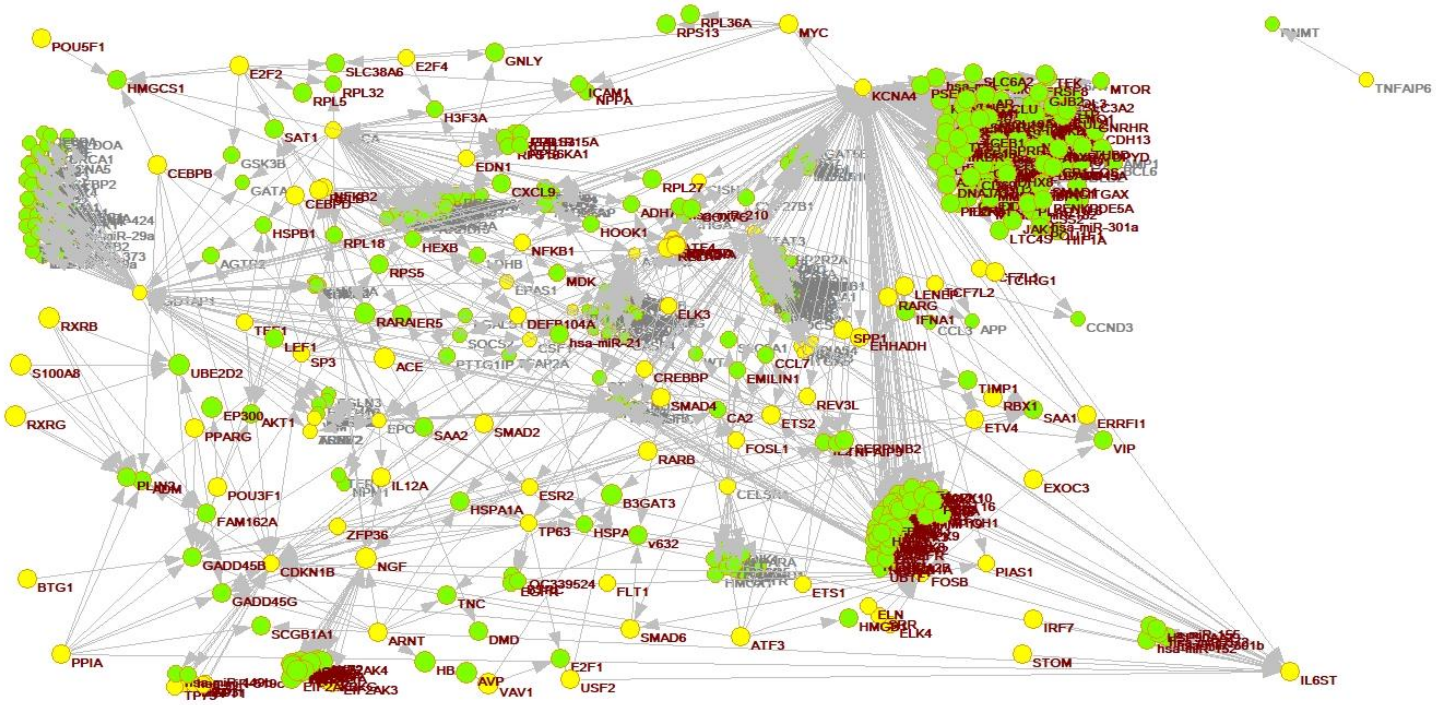


【圖7】根據主成分分析得出的變異系數大小，紅色代表高表現量，綠色代表低表現量，黑色則表示表現量差異不大，根據系數表現可區分成三種類型。

【圖6】與【圖7】互相比對可得知胚胎幹細胞、脂肪成體幹細胞及脂肪細胞受哪些基因調控。

利用篩選後的探針資料比對資料表GPL96得出各探針對應的基因蛋白質名稱為何，GPL96為 Affymetrix Human Genome U133A Array [HG-U133A]的註釋表，記載於 NCBI (https://www.ncbi.nlm.nih.gov/geo/quer

y/acc.cgi?acc=GPL96)，接著利用 RegNetwork: Regulatory Network Repository of Transcription Factor and microRNA Mediated Gene Regulations(http://www.regnetworkweb.org/search.jsp?searchItem=&searchType=all&organism=human&database=all&evidence=all&confidence=High&resultsPerPage=30&prevValidPN=1&orderBy=RegSymbol_Asc&pageNumber=1)搜尋基因主要調控因子，該數據庫記錄人類及小鼠調控網路，其中我們設定調控網路驗證有高信賴度的資料，與我們篩選後所得的基因名稱進行比對，繪製出【圖8】，



【圖8】基因與轉錄因子調控網路圖。

	主要調控轉錄因子
胚胎幹細胞分化至脂肪幹細胞	AR,CEBPA,CEBPB,CREB1,CREBBP,E2F1,E2F2,E2F4,ELK1,EP300,ERG,ESR1,ETV4,HIF1A,JUN,LEF1,MYC,NFIC,NFKB1,PPARD,PPARG,RARA,RARB,RARG,RELA,SP1,STAT1,STAT3,TFAP2A,TP53
脂肪幹細胞分化至脂肪細胞	ATF1,CEBPA,CEBPB,CEBPD,CREB1,E2F1,E2F4,HIF1A,JUN,MYC,NFIC,NFKB1,POU2F1,PPARG,SMAD1,SMAD3,SP1,TFAP2A,TP53,USF1

【表1】調控分化的轉錄因子。

5. 評估與展望

現階段只有對脂肪細胞進行研究，未來將增加更多不同的器官進行分析比對，例:神經細胞、造血細胞...等希望可以從內胚層一直慢慢拓展至中胚層甚至外胚層，將生物體上大大小小的細胞進行進一步的分類分析。

6. 結語

從本專題的資料而言我們發現胚胎幹細胞分化至脂肪幹細胞及脂肪細胞的蛋白質調控群，就可知道細胞的分化調控網路，後續將繼續探討調控蛋白質的轉錄因子，若鑑定結果與科學根據相符，未來將可進行更多細胞樣本的轉錄因子鑑定研究。

7. 銘謝

感謝黃俊燕老師這兩年來的耐心指導，讓有如白紙的我們，開始有了不同的色彩，也感謝組員間一直以來的互相幫助溝通，才能有這樣的成果。

8. 參考文獻

[1]Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen & Alvis Brazma, ”A global map of human gene expression”, Nat. Biotech. 28, 322–324, (2010).

[2] Lindsay I Smith,” A tutorial on Principal Components Analysis”, February 26, (2002).

[3] S. Niemelä, S. Miettinen, J.R. Sarkanen and N. Ashammakhi” Adipose Tissue and Adipocyte Differentiation: Molecular and Cellular Aspects and Tissue Engineering Applications” Tissue Engineering, Vol. 4. Eds. N Ashammakhi, R Reis, & F Chiellini, (2008).

[4] Christopher E. Lowe, Stephen O'Rahilly, Justin J. Rochford” Adipogenesis at a glance”, Journal of Cell Science 124, 2681-2686, (2011).

[5] Takahashi K, Yamanaka S. “Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.” Cell 126, 663-676, (2006).