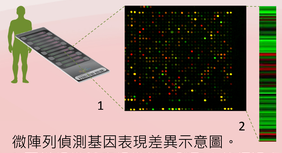
您好，我是楊欣洲，研究分子生物統計學真是帥呆了！很久很久以前，一位神機妙算的統計魔法師立志要調配出治病的配方，拯救同胞。祖傳的魔法書 – 分子生物學的中心定律 (Central Dogma) 清楚指出了細胞如何利用隱藏在去氧核醣核酸 (DNA) 中的遺傳訊息來維持運作：DNA會製造出核醣核酸 (RNA)，RNA接著製造出蛋白質，而許多稱做「酶」的蛋白質則負責指揮著細胞裡一連串複雜的化學反應，其中也包括代謝作用。於是魔法師從那游來游去的生命密碼 – DNA 著手研究起，過程中要通過層層困難的關卡，才能找到各項關鍵藥材(基因和代謝物)和配方調製的秘笈(統計分析方法)，一腳踩差就註定失敗！接下來，就讓我帶領您一同進入這趟剖析生命奧秘的冒險旅程，利用統計魔法尋找出通關的寶物，包括DNA關卡中的SNP、RNA關卡中的基因表現，以及代謝關卡中的代謝物表現，最後將這些寶物融合調配成治病妙方，解除同胞的病痛。

中央研究院統計科學研究所 楊欣洲老師實驗室

話說從頭

疾病的研究就和面對一棟大門深鎖的宅第一樣，只有奮力解開鎖栓之後，才能看見門後的各種可能；而分子生物標誌(molecular biomarker)，那些可被測量且可以輔助臨床決策的分子生物特徵，就像是各種形狀的鑰匙或解鎖方式，一直讓研究者費盡心力地尋找而且探索它們影響疾病的機制。近年來生醫技術飛速發展，基因變異(genetic variants)或蛋白質(proteins)、代謝物(metabolites)表現量上的差異等等都能大規模且有效率地衡量，許多和它們有關的生物標誌也因此被快速地定位出來。接下來讓我們進一步了解包括單一核苷酸多型性(single nucleotide polymorphism，以下簡稱SNP)、基因表現量(gene expression)等基因變異和代謝物的意義與它們經過測量後得到的數據內容，藉此幫助後續分析方法的理解。

基因表現:DNA像是電影膠片一樣，分成很多畫格；如果用正確的順序觀看，就能看到一部好戲。而基因，DNA上一段段特定的區域，則是一幕幕戲中的情節，攜帶了製造蛋白質所需要的資訊。基因表現指的是DNA的遺傳資訊傳遞到RNA再轉換成蛋白質的過程；利用微陣列晶片實驗對RNA的量測，可獲得人體組織中數萬個基因的表現程度。如果把它應用在癌症研究中，正常組織與癌組織的基因表現程度能透過微陣列晶片實驗取得，而兩組織間顯著差異表現的基因則可接著利用統計方法來偵測。

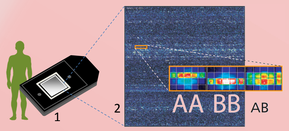


微陣列偵測基因表現差異示意圖。1.正常組織與癌組織從人體取出後經過處理即導入晶片中，2.從晶片擷取的影像裡不同的螢光顏色可以分析出各個基因在不同組織間的表現差異。

Target編號	樣本一	樣本二	樣本三	樣本四
GI_10047089-S	5.939529	5.887654	5.859701	5.934375
GI_10047091-S	6.286516	6.210589	6.277051	6.24467
GI_10047093-S	10.3538	10.34492	9.856135	10.1363
GI_10047099-S	7.59397	7.763176	7.382515	7.024285
GI_10047103-S	12.20475	12.03409	11.9565	12.17631

基因表現資料內容示意。

SNP:DNA是一條像由兩條珍珠串鍊捲繞成的長鏈狀分子，其中每條串鍊稱為DNA的一股；而鍊上的珍珠，也就是組成DNA的單元，我們稱之為核苷酸。DNA含有四種不同鹼基所形成的核苷酸，簡寫成A、T、G、C。在兩股配對的DNA中，A總是配上另外一股的T，而G則總與C配對；一組A-T或G-C的配對稱為鹼基對。遺傳資訊就透過核苷酸的特定順序儲存於DNA中。SNP指的是DNA序列在單一位置上的核苷酸發生了變異。比方說一段DNA從AATCGCC變成AATTGCC，這時第四個核苷酸從C變成了T，便是一個SNP。順帶一提，通常稱作SNP的變異必須存在族群內1%以上的個體中。不過如何從一個人三十億DNA鹼基對中衡量數以千萬計的SNPs呢？目前透過微陣列晶片實驗，我們可以一次衡量上百萬個SNPs，若是利用次世代定序技術的協助，則可以更進一步取得整段DNA序列的資訊。



微陣列量測SNP示意圖。1.血液從人體取出後經過處理即導入晶片中，2.從晶片擷取出的影像即可用作SNP量測依據。

SNP

樣本編號	SNP1	SNP2	SNP3	SNP4	SNP5
1	C C	A A	T T	A G	N N
2	A C	G A	T T	G G	C T
3	A C	G A	T T	G G	C T
4	A C	G A	T T	G G	C T

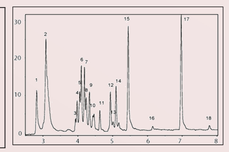
樣本　基因型

基因型(genotype)資料內容示意。

代謝物:細胞的代謝作用，簡單地說，是利用一連串的化學反應將攝入的食物轉換成能量和分子的過程。代謝和闖關遊戲的過程有些相似，在完成一個關卡後可能會得到一些寶物作為獎賞，而這些寶物同樣可能是幫助通過下道關卡的道具。代謝物，就像是這些寶物或道具，是代謝過程的中介物或產物。常見的代謝物比如說各種胺基酸、維他命等等；而在疾病研究中，研究者可以利用質譜技術量測病人與正常人的數千種代謝物數量後以統計方法找出其中有顯著數量不同的代謝物。

Peak編號	Peak3	Peak6	Peak9	Peak12
D0914_ACEI1_1	2.18914	2.13927	2.65429	2.09576
D0914_ACEI1_2	2.15588	2.17791	2.56663	2.08472
D0914_ACEI1_3	2.14121	2.17519	2.50833	2.09706
D0914_ACEI1_4	2.12428	2.17519	2.50833	2.08472
D0914_ACEI1_5	2.11421	2.18914	2.48875	2.09706
D0914_ACEI1_6	2.10097	2.16279	2.44437	2.08121
D0914_ACEI2_1	2.4699	2.41455	4.01378	2.40116
D0914_ACEI2_2	2.45272	2.47888	4.01378	2.51897
D0914_ACEI2_3	2.40116	2.42145	3.87413	2.4699
D0914_ACEI2_4	2.42145	2.44437	4.01378	2.47888

代謝物資料內容示意。其中數值表示左圖不同高點的相對強度。



利用質譜技術分析檢體成分示意圖。圖中有標號的高點即代表不同的代謝物。

圖片來源：http://depts.washington.edu/sfields/metabolites/index.html, http://www.affymetrix.com, http://microarray.org http://www.illumina.com, http://123rf.com

DNA混合池 (DNA pooling) 之統計分析工具 - MPDA

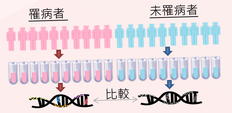
如果可知道人生病與否之間的差異，是否能夠準確對症下藥呢?

如何找出群體差異?

可利用統計方法比對兩群人的遺傳資料，找出有顯著差異的生物標誌。

如何得到群體資料?

可利用基因型鑑定技術鑑定每一個人的個體SNP資料，進而得到分別代表兩群的群體資料。



但是，收集個體資料的費用對於某些研究室而言，過於昂貴，而難以進行大樣本實驗!

例子: 目前基因型鑑定技術鑑定一個人的個體基因資料費用是16,000元，要得到1,000名未罹病者與1,000名罹病者的遺傳資料，總需花費16,000 × 2,000 = 三千兩百萬!

好…好貴！

若只要兩群的群體資料，而非個體資料…

可利用DNA混合池技術!

相同的例子，原本DNA數量是2,000管，利用DNA混合池技術，將同一群人的DNA混合在一起，就只有兩管(一群人一管)DNA需要鑑定基因資料，如此便只需花費16,000 × 2 = 3萬2千元!

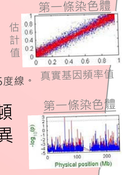
差了一千倍! **費用省很大!**

但有利也有弊，這技術同時引進了**估計誤差**!

因此，我們發展以MATLAB撰寫的軟體 - **MPDA**

除了**方便使用**的點選介面(右圖)以外，還提供以下**統計分析**的功能喔!

校正估計誤差: 利用我們發展的校正係數可使DNA混合池所得到的對偶基因頻率(AF)估計值與真實值更為接近!

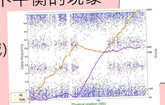


如右圖中校正後值(紅點)比未校正前(藍點)更接近45度線。

相關性分析: 使用者可經由曼哈頓圖及MPDA提供的P值找出有顯著差異的可能致病基因所在位置。

對偶基因不平衡分析: 使用者可經由標示的橘、紫線判斷是否發生對偶基因不平衡的現象。

右圖為一異常染色體的對偶基因頻率圖(正常為藍點大多集中在上中下三個區域)
橘線斜率增加: 可能發生部分倍體增加；
紫線斜率增加: 可能發生部分倍體缺失。

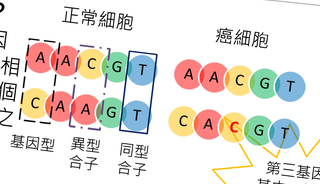


此部分的圖例皆只以23條染色體中第一條染色體為例。

DNA片段異質性缺失分析

什麼是同型合子/異型合子？

我們觀察在DNA上頭的某個基因座時，會發現兩個對偶基因可能相同亦或不同。如果基因座上的兩個對偶基因相同稱為同型合子，反之則稱為異型合子。

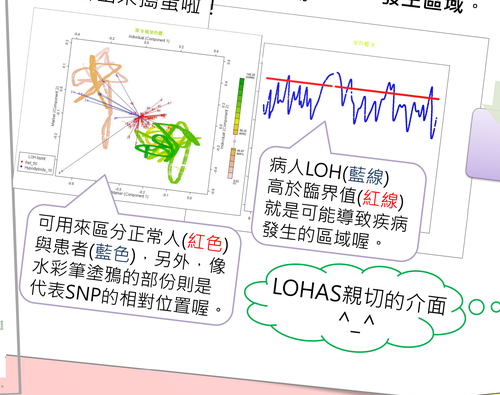


DNA片段異質性缺失 (loss of heterozygosity，簡稱LOH)

當染色體在複製的時候，萬一發生了**DNA突變**(如右上圖)、**DNA區段被刪除**等情形。使原本相對穩定的異型合子變成為同型合子。又剛好發生改變的位置是用來封印某種疾病的話，封印一旦被解開，疾病就會出來搗蛋啦！

LOHAS (Loss-of-Heterozygosity Analysis Suite)

為了找出這些封印的位置，也就是LOH強度高的區域，我們發展出**LOHAS**這套軟體，用來找出**LOH發生區域**。

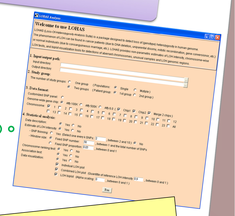


LOHAS親切的介面 ^_^

用迴歸方法估LOH強度。

用無母數方法做相關性檢定。

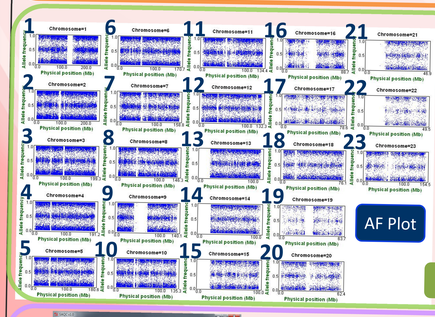
作圖幫助我們分析資料。



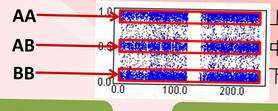
擔心生物晶片的品質嗎!? 交給 SAQC 把關就對了~

- 利用生物晶片鑑定基因型乃科技的一大進步，藉由這樣的技術，能夠讓我們更廣泛地探究基因的奧秘。
- 然而，為了得到**可靠的統計分析結果**，我們必須**確保資料的品質**。因此，開發統計方法來幫助評量生物晶片的品質便顯得勢在必行。這正是本軟體 – ***SAQC***最大的貢獻。

- 每個人在經過生物晶片的基因型鑑定後，可得到一張**對偶基因頻率圖**(Allele Frequency Plot，亦稱**AF Plot**)，其中包含23張小圖，分別代表人類的23條染色體。
- 圖中的每個藍色小點都是一個SNP的對偶基因頻率，由於每一條染色體**成對**的關係(其中一條來自父親，一條來自母親)，在**正常情況**下對偶基因頻率將**集中在3個區域**(分別於小圖內的上、中、下)。
- SAQC將利用所有23條染色體所提供的豐富訊息來進行**晶片品質的檢測**。



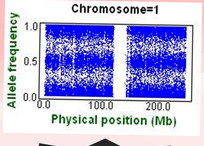
Q: 為何正常情形是集中在3個區域?(以第一條染色體為圖例)



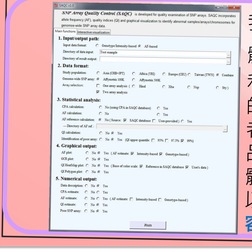
可能從父親得到對偶基因A或B ＋ 可能從母親得到對偶基因A或B

可能得到的基因型有**3**種: AA, AB 或 BB

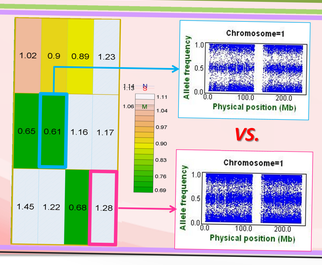
Q: 晶片品質不佳時，AF Plot可能如何呈現呢?



較不集中於3個區域!



我們發展了一套軟體，名為***SAQC***，利用對偶基因頻率的資訊來幫助使用者評估生物晶片的品質。左圖即為軟體的使用者介面，以R語言開發，**操作容易**!

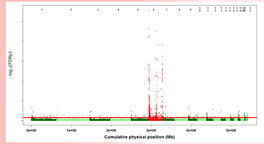


左圖為SAQC的結果之一 – **色層圖**。此圖例中一共有12張晶片，每個方框內的值各代表一張生物晶片**品質好壞的指數**。值越大，則顏色越趨近白色，代表品質較差，值越小則代表品質越好。由此例可知道SAQC能確切幫我們**分辨晶片品質的好壞**!

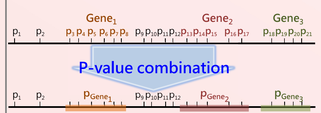
致病基因定位

治病要治本非治標，因此疾病基因定位尤顯重要，藉統計方法找出與疾病相關的基因位置，生物實驗進而驗證尋求根治方法

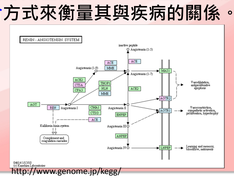
1. 早期疾病基因定位利用關聯分析(Association)找出與疾病相關的 SNP，由下面的曼哈頓圖，可看出在染色體 6 (紅色)有明顯的訊號，但其它地方呢？



2.個別 SNP 的探討，易忽略個別訊號不強，但結合起來對疾病有影響的 SNP。因此以基因(Gene)為單位，將其內各 SNP 對疾病關聯分析所得資訊(P值)結合，找出與疾病有關的基因。



3. 針對此一整合方式可以推廣應用在反應路徑(Pathway)的分析上，即將在反應路徑內的 SNP 或基因利用P值組合方式來衡量其與疾病的關係。



http://www.genome.jp/kegg/

疾病基因定位就如同警察在高樓林立的都市叢林中，試圖找出嫌犯的藏身處。

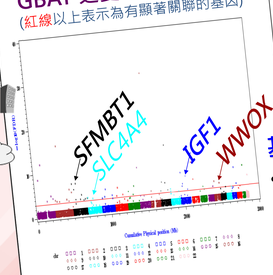
俗話說狡兔三窟，單一處所的鎖定不見得有結果，若能劃分區域來收集情報，更能有效縮小範圍並鎖定特定區域，於內展開搜索。

區域的鎖定再配合動線的考量更能洞悉嫌犯的想法，比對方早一步作安排將其手到擒來。

實際資料分析: 高血壓疾病之相關研究

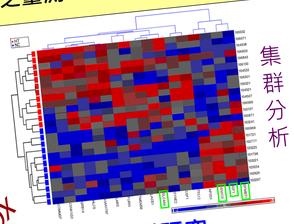
研究收集了一群罹患高血壓的病人以及一群正常人的血液，並進行其SNP鑑定、RNA表現量、代謝物濃度之量測，本研究目的是想找出與高血壓相關的致病機轉。

GBAT 之曼哈頓圖
(紅線以上表示為有顯著關聯的基因)



SNP 資料研究

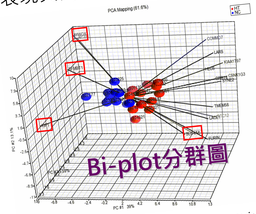
- 將**GBAT**方法應用在SNP的資料分析上，可以整合同一基因上所有SNP的資訊，藉以評估此基因是否與高血壓有關。
- 研究中找到**100**個基因與高血壓有關。



集群分析

基因表現 資料研究

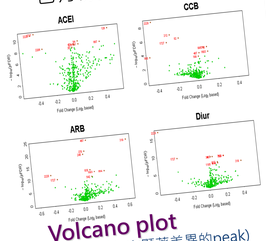
- 將兩群人的每一個基因之**RNA表現量**作一個比較，若高血壓患者的基因表現量比正常人高，表示這個基因在高血壓病人身上表現的比正常人多。代表這個基因的過多表現可能與高血壓有關。
- 由100個基因中，找到**17**個基因表現與高血壓有關。



Bi-plot分群圖

代謝物 資料研究

- 萃取兩群人血液中的代謝物，並利用質譜(MS)分析找出其可能與**高血壓相關的代謝物**，若高血壓病人中的某代謝物濃度較正常人來得高，代表罹患高血壓後，此代謝物的量較正常人來得高。表示此一代謝物與高血壓有關。
- 研究也針對有無服用高血壓藥的病人進行比較，已有初步發現。



Volcano plot
(紅點表示為有顯著差異的peak)

Reference: Identification of IGF1, SLC4A4, WWOX and SFMBT1 as hypertension susceptibility genes in Han Chinese with a genome-wide gene-based association study. (Submitted)

GENETICS