研究成果（2008.01.01 — 2012.07.31）

楊欣洲

在中研院統計所、中研院前瞻計畫、國科會研究計畫與國家型基因體醫學研究計畫的支持下，我們致力於發展創新的統計和生物資訊方法以分析大量人類基因體資料，並與生物學家密切合作以定位年輕型高血壓的易感基因，以下簡述近年主要的研究成果。

一、**混合去氧核醣核酸分析：**我們發展系統性分析混合去氧核醣核酸實驗資料的統計方法，並開發了一套分析工具 MPDA (Yang et al., *BMC Bioinformatics*, 2008)，功能含括訊息的擷取、不等雜交放大率的定量、等位基因頻率的估計、單點與多點關聯性基因定位、單點與多點等位不平衡分析等等，並已成功應用於多組實際數據的分析。

二、**遺傳關聯性檢定：**我們發展了核基礎的相關性檢定法以定位疾病易感基因，實證研究中找到重要的酗酒症易感基因,並開發了一套容易操作的共享軟體 KBAT (Yang et al., *Genetics*, 2008)。另外，我們也發展了一套以基因為基礎的遺傳相關性檢定法，其可用以評估整個基因對所研究性狀的效果，此方法也被推廣到可以評估整個反應路徑對所研究性狀的效果，並在實證研究中找到重要的類風濕性關節炎易感基因 (Yang et al., *BMC Proceedings*, 2009; Yang and Chen, *BMC Proceedings*, 2011) 。

三、**等位基因頻率研究：**我們發展了一套分析單一核苷酸多態性基因晶片所產生的雜交強度資料的分析工具 ALOHA，利用此分析工具，我們可正確地估計個人等位基因頻率，進而剖析等位基因頻率在基因體的特性，偵測染色體異常片段，對樣本進行分群，尋找異常樣本，以及推估腫瘤樣本的純度 (Yang et al., *BMC Genomics*, 2010)。

四、**異合型缺失偵測：**我們發展了一套以移動視窗為基礎的異合型缺失偵測法，用來決定病例組和對照組間異合型缺失強度的差異 (Huggins et al., *Journal of Human*

Genetics, 2008)。我們也開發了一套方便使用的共享軟體軟體 LOHAS 以進行異合型缺失或長片段同合型片段的定位與分析，協助族群遺傳研究、癌症基因定位研究和複雜疾病遺傳與基因研究。同時，實證研究中找到重要的急性白血病致癌和抑癌基因和類風濕性關節炎易感基因（Yang et al., *Genetic Epidemiology*, 2011; Yang et al., *PLoS One*, 2012)。

五、**單一核苷酸多態性基因晶片品質管制：**我們提出透過檢查單一核苷酸多態性基因晶片之雜交強度資料所呈現的異常，評估基因晶片的品質，並開發了一套分析工具 SAQC，可提供晶片品質的快速檢測（Yang et al., *BMC Bioinformatics*, 2011)。

六、**單一核苷酸多態性和基因表現整合分析：**我們創新提出整合單一核苷酸多態性和基因表現兩種標誌基因以鑑別親源接近樣本的概念，並發展一套尋找少數關鍵祖先訊息標誌基因即可正確分類不同種族樣本的方法，也開發分析工具 BIASLESS，可應用於族群遺傳學和基因體醫學研究（Yang et al., *BMC Genomics*, 2012)。

七、**年輕型高血壓全基因體相關性致病基因定位研究：**除了統計和生物資訊方法的發展，我們也與生物學家密切合作，定位年輕型高血壓的易感基因。我們完成了華人首件年輕型高血壓全基因體相關性致病基因定位研究，成功地定位出兩個重要的年輕型高血壓的易感基因，分別座落於二號和六號染色體上（Yang et al., *PLoS One*, 2009)。之後，更以基因為基礎的全基因體掃瞄，並輔以基因表現資料分析的佐證，以及香港和歐裔高加索人兩組高血壓研究資料分析的再確認，有效率地定位出與高血壓、心血管疾病、新陳代謝症候群相關的基因（Yang et al., *PLoS One*, 2012)。

各項研究成果，進一步說明如下：

第一、微陣列生物晶片型的混合去氧核醣核酸實驗，同時結合了混合去氧核醣核酸和生物晶片實驗的優點，前者節省了同時考量許多實驗樣本的經費，後者節省了同時鑑定大量標誌基因的成本，這樣的實驗提供給從事全基因體研究的研究者一個十分節省成本的好方法。微陣列生物晶片型的混合去氧核醣核酸實驗產生的資料，其屬性與結構都遠比傳統的局部實驗或是低解析度實驗所得到的資料來得大量且複雜許多，這類型的資

料目前正快速地成長與累積中。分析這樣高密度的實驗資料時，會涉及一連串十分複雜的程序，包括大量的資料處理、統計估計以及假設檢定，至今還沒有任何一個資源共享的平台，可以系統化地分析這類型的基因體資料。因此，我們發展了統計和生物資訊的理論與方法，可以有效地分析混合去氧核醣核酸實驗產生的資料，功能含括訊息的擷取、不等雜交放大率的定量、等位基因頻率的估計、單點與多點關聯性基因定位、單點與多點等位不平衡分析等等。為了方便讀者使用此方法，我們也發展了一套整合系統 MPDA (Microarray Pooled DNA Analyzer)，並提供操作手冊，讀者可在作者網頁下載 (http://www.stat.sinica.edu.tw/hsinchou/genetics/pooledDNA/mpda.htm)，本研究成果已經發表在國際期刊 *BMC Bioinformatics*。

第二、我們發展了一套核基礎的相關性檢定法，其結合「單點相關性檢定的 p 值」和「反應標誌基因間物理距離或連鎖不平衡強度的核權重」，充分利用鄰近標誌基因的整合訊息來定位疾病易感基因，不僅可用以研究候選基因，也可結合移動平均的方法以研究完整基因體。我們透過大量的模擬實驗來檢驗這個新方法，其中考量了演化的參數、疾病模式、樣本大小、各式核函數、檢定統計量、視窗特性、經驗 p 值演算法、基因圖譜等等不同的因素。模擬結果顯示，所提出的新方法不僅可以有效地控制型一誤差，同時具備很好的檢定能力。我們將此方法應用於分析一組研究酗酒症易感基因的全基因體資料，此分析成功地找到一些重要的酗酒症相關基因。綜而言之，所提出的核基礎的相關性檢定有以下幾項優點：一、其有不受干擾標誌基因影響的穩健性。二、不受所使用各式基因圖譜影響的不變性。三、可用以分析來自不同實驗設計和各種類型基因數據的通用性。為了方便讀者使用此方法，我們也發展了一套容易操作的軟體 KBAT (Kernel-based Association Test)，並提供操作手冊，讀者可在作者網頁免費下載 (http://www.stat.sinica.edu.tw/hsinchou/genetics/association/KBAT.htm)，本研究成果已經發表在國際期刊 *Genetics*。

第三、我們也發展了一套以基因為基礎的遺傳相關性偵測法。有別於過去單一核苷酸多態性、單套基因型、交互作用分析等等的相關分析，此新方法可以直間評估整個基因的效果。此方法有以下幾項優點：一、方便定義出具有生物意義的基因體片段來進行

相關性檢定。二、有不錯的統計檢定力。三、有助減緩多重檢定的問題。四、研究結果較易賦予生物上的解釋。透過此方法，我們成功地定位出年輕型高血壓、類風濕性關節炎等等疾病的致病基因 (Yang et al., *BMC Proceedings*, 2009)。我們進而推廣此方法以評估整個反應路徑對研究性狀的效果，透過分析 200 組由遺傳分析會議所提供的，由病例對照樣本設計下產生的第二代去氧核醣核酸外顯子序列資料，結果顯示我們所提出的方法確實可有效地定位出重要的疾病易感基因與疾病相關的反應路徑。本方法不只適用於處理常見遺傳變異，對於攜帶有大量罕見遺傳變異的基因序列資料的分析也同樣有效 (Yang and Chen, *BMC Proceedings*, 2011)。

第四、等位基因頻率是個重要的遺傳特性測度，已經被廣泛地應用於遺傳研究與基因體研究之中。用基因型資料來估計等位基因頻率是個方便的方法，但是當染色體有特殊結構變異或是基因型資料有鑑定錯誤的疑慮時，這樣的方法可能會忽略掉一些重要的訊息。因此我們發展了一套整合的雜交強度測量方法來估計個人的等位基因頻率，以克服過去只用基因型資料分析的不足，我們用此方法分析了 1,104 和 1,270 個分別利用 Affymetrix 100K 和 Affymetrix 500K 單一核苷酸多態性基因晶片進行鑑定的樣本。值得一提的，在估計等位基因頻率的過程中，我們矯正了不等雜交放大的現象，我們發現，矯正不等雜交放大率確實顯著地增加了等位基因頻率估計的正確性。進一步研究不等雜交放大率後發現，其不太受到基因型鑑定實驗所進行的時間、實驗室、晶片的種類和樣本的外表型等等因素的影響，但在不同種族，分佈卻會有所不同。因此，我們為不同的種族，分別建立各自的不等雜交放大率資料庫與等位基因頻率資料庫，以供使用者依其需要來使用。透過所得到的正確的等位基因頻率估計，我們可有效地剖析等位基因頻率在基因體的特性，偵測染色體異常片段，對樣本進行分群，尋找異常樣本，以及推估腫瘤樣本的純度。同時，我們也發展了一套方便使用的分析軟體 ALOHA (Allele-frequency/Loss-of-heterozygosity/Allele-imbalance)，並提供操作手冊，透過這套軟體，上述的分析都可以很容易地進行。讀者可在作者網頁免費下載 (http://www.stat.sinica.edu.tw/hsinchou/genetics/aloha/ALOHA.htm)，本研究成果已經發表在國際期刊 *BMC Genomics*。

第五、我們發展了一套以移動視窗為基礎的異合型缺失偵測法。本方法的第一步是利用無母數的方法，先決定在全基因體中，哪些染色體在病例組和正常組有異合型缺失強度的差異，第二步則再利用雙極圖的視覺畫圖形和異合型缺失強度估計以定位出有問題的染色體片段。此方法不僅在模擬研究中有很好的表現，我們也已利用此方法成功地找到急性白血球疾病的抑癌基因，本研究成果已經發表在國際期刊 *Journal of Human Genetics*。另外，我們也發展生物計算的方法來分別估計病例組和對照組中連續出現同合型基因型的比率，以協助找出具有特殊染色體結構變異或不尋常基因型趨勢的樣本，將具有相似的異合型缺失型態的樣本歸群，並定位出與疾病或癌症相關的異合型缺失基因片段，這是首次有方法可協助同時回答這些重要的課題。透過分析 304 位急性白血球疾病病患和 50 位正常的對照樣本的全基因體單一核苷酸多態性資料，我們成功地找到急性白血球疾病的致病基因。另外，此方法也被應用於研究一般正常（非疾病）族群。透過分析 60 位非洲人、60 位歐裔高加索人和 90 位亞洲人的全基因體單一核苷酸多態性資料，先估計每個樣本在基因體中長片段同合型的比率，再利用此資訊，即可將來自不同種族的樣本正確地歸群。為了方便讀者使用此方法，我們也發展了一套容易操作的共享軟體 LOHAS (Loss-Of-Heterozygosity Analysis Suite) 以進行異合型缺失或長片段同合型片段的定位與分析，讀者可在作者網頁免費下載 (http://www.stat.sinica.edu.tw/hsinchou/genetics/loh/LOHAS.htm)，本研究成果已經發表在國際期刊 *Genetic Epidemiology*。

第六、全基因體單一核苷酸多態性基因晶片一次可提供數十萬到百萬的單一核苷酸多態性資料，目前已經成功地應用於人類基因研究。然而，需要注意的是，基因晶片所產出的資料的品質，會對後續統計資料分析的精確性與準確性造成重大的影響。如何評估基因晶片所產出的資料的品質？好的品質指標仍有待發展。因此，我們發展新的品質指標來量測基因晶片或是去氧核醣核酸樣本的品質，並且仔細研究其機率分布與統計特性。所提出的品質指標是透過計算「估計的個別等位基因頻率」與「期望的個別等位基因頻率」之間的馬氏距離來得到。透過國內外幾個大型基因體計畫的資料加以佐證，我們發現所提出的品質指標會服從對數常態機率分布，而機率分布中的參數則會隨種族

的不同而有所不同。因此，我們分別為不同種族（亞洲人、歐裔高加索人、非洲人和我們自己台灣族群），建立其等位基因頻率資料庫和品質指標資料庫。同時，我們也發展了一套信賴區間的方法，用以鑑別所測試的基因晶片和樣本的品質是否有問題。透過實際基因體資料的分析與模擬基因體資料的分析，結果都顯示我們的方法有很好的敏感度與特異度。綜而言之，我們發展了新的品質指標，建立了等位基因頻率資料庫與品質指標資料庫，探討了品質指標的統計特性，開發了偵測品質不佳的晶片或是去氧核醣核酸樣本的統計方法。同時，整合以上的功能，我們完成了一套以 R 語言和 R 圖形介面所發展而成的生物資訊軟體 SAQC (SNP Array Quality Control)，這套軟體將可為未來的基因晶片研究，提供方便且有效的品質評估。讀者可在作者網頁免費下載(http://www.stat.sinica.edu.tw/hsinchou/genetics/quality/SAQC.htm)，本研究成果已經發表在國際期刊 *BMC Bioinformatics*。

第七、：「祖先訊息標誌基因」是泛指可提供回溯所研究個體之祖先族裔歷史訊息的一種標誌基因，已廣泛地被應用於鑑識科學、族群遺傳學和基因體醫學。單一核苷酸多型性常被用來做為建構祖先訊息標誌基因的材料，然而，研究顯示，其雖可有效地鑑別來自不同洲際或是親源關係遠離的個體，卻難以有效地鑑別族裔親源接近的個體。最近的研究發現，基因表現也是一種遺傳性狀，來自不同族裔的個體的基因表現會有所差異。有鑑於此，我們提出整合單一核苷酸多型性和基因表現的訊息，透過只選用少數但富含祖先訊息的標誌基因，達到可精確預測個體所屬族裔的功效，過去未有研究曾利用這樣的整合來鑑別和分類不同族裔的個體。因此，我們整合了鑑別分析和正向變數選擇程序，來鑑別和分類不同族裔的個體，並搭配交互驗證程序，以選取可達到最高交互驗證預測正確率的關鍵單一核苷酸多型性和基因表現標誌基因。實證研究中，我們分析來自國際 HapMap II 計畫中的四個族群的 210 個樣本，利用其所提供的全基因體單一核苷酸多型性和基因表現晶片的資料，進行樣本的鑑別與分群，所進行的三項鑑別分析包括有只使用單一核苷酸多型性資料、只使用基因表現資料，以及整合兩類標誌基因資料的分析。分析結果顯示，大多我們的鑑別分析的族群預測正確性都相當高，唯一例外情形是在只用單一核苷酸多型性資料的鑑別分析來分類兩個同為亞洲族群的樣本（日本東京

人與北京漢人）時的正確率偏低，約為 0.53 到 0.79，在整合單一核苷酸多型性和基因表現資料後，正確率可上升到 0.9，相較於只使用基因表現資料的結果，整合分析有十分接近的正確率，但所需的祖先訊息標誌基因的個數卻較為經濟節省。總結來說，單一核苷酸多型性和基因表現整合分析提供了一套高正確且經濟的方法來進行族群鑑別與分群，有鑑於此研究課題的實際重要性，除了方法論外，我們也發展了一套方便使用的分析工具 BIASLESS (Biomarkers Identification and Samples Subdivision)，可自龐雜的人類基因體中搜尋到關鍵的單一核苷酸多型性和基因表現作為祖先訊息標誌基因，進而協助將個體樣本正確歸類於所屬的族群。軟體與操作手冊可在作者網頁下載 http://www.stat.sinica.edu.tw/hsinchou/genetics/classification/BIASLESS.htm ，本研究成果已經發表在國際期刊 *BMC Genomics*。

最後、除了發展創新的統計和生物資訊方法外，我們與生物學家共同合作定位台灣年輕型高血壓的易感基因。高血壓是一種複雜性疾病，在全世界都有很高的盛行率，特別是在工業化程度高的國家尤為嚴重。我們首先發表了年輕型高血壓病例對照之 Affymetrix 100K 全基因體資料的研究。本分析採用一個兩階段的致病基因定位策略，第一階段利用一百七十五組成對的病例對照樣本，進行全基因體掃瞄，先找出年輕型高血壓的候選易感基因，第二階段再利用超過兩千位的病例對照樣本，進行驗證分析。在此分析中，我們成功地定位出兩個重要的年輕型高血壓的易感基因，分別座落於二號和六號染色體上 (Yang et al., *PLoS One*, 2009)。隨後，我們以所蒐集的四百對年齡與性別配對的台灣漢人病例對照樣本，每個樣本都先利用 Illumina HumanHap550-Duo 的單一核苷酸多態性基因晶片，進行基因型鑑定，然後透過全基因體以基因為基礎的相關性研究，以定位出高血壓之致病基因。分析中，我們首先找出一百個經過重排程序檢定仍然達到統計相關性顯著的基因，再透過分析基因表現晶片資料，從中找到了十七個在疾病組與對照組間有顯著表現量差異的基因。這十七個基因提供了一個 96% 正確率的統計模式來預測高血壓的疾病狀態。十七個基因中，基因 *IGF1*、*SLC4A4* 和 *WWOX* 可以被另一組來自香港的高血壓全基因體相關性研究所驗證。基因 *COMMD7* 則可以被另一組來自 Wellcome Trust 的歐裔樣本的高血壓全基因體相關性研究所驗證。另外兩個基因，

TMEM56 和 *KIAA1797*，雖然無法被香港或是 Wellcome Trust 的研究所驗證，但此兩個基因在台灣族群的疾病組與對照組間呈現出十分顯著的等位基因頻率差異與表現量差異，有可能是台灣族群特有的高血壓易感基因（Yang et al., *PLoS One*, 2012）。定位出這些高血壓基因，不僅讓現有的高血壓基因資料庫更加完整，更有助我們對漢人族群高血壓的致病與機轉有更多的瞭解。

本團隊過去的先導研究已對族群遺傳、致病基因定位和染色體異常偵測等課題多所探討，在此基礎之上，未來，我們將繼續研發新穎且有效的統計與生物資訊的方法與分析工具，同時應用我們所過去研發的工具和新發展的方法，分析大量生物實驗產生的和模擬研究產生的單一核苷酸多型性、基因序列和基因表現等基因體資料，來回答與研究人類基因體學和人類疾病基因體學中重要的生物假設與課題，並且評估我們所開發出來的新方法和新工具的表現。

論文發表（*通訊作者）：

1. **Yang, H.-C.***, Huang, M.-C., Li, L.-H., Lin, C.-H., Yu, L. T., Diccianni, M. B., Wu, J.-Y., Chen, Y.-T. and Fann, C. S. J.* (2008/04). MPDA: microarray-based pooled DNA analyzer. *BMC Bioinformatics* 9, 196. SCI.
2. **Yang, H.-C.***, Hsieh, H.-Y. and Fann, C. S. J. (2008/06). KBAT: Kernel-based association test. *Genetics* **179**, 1057-1068. SCI.
3. Huggins, R., Li, L.-H., Lin, Y.-C., Yu, A.L.T. and **Yang, H.-C.*** (2008/12). Nonparametric estimation of LOH using Affymetrix SNP genotyping arrays for unpaired samples. *Journal of Human Genetics* **53**, 983-990. SCI.
4. **Yang, H.-C.**, Liang, Y.-J., Wu, Y.-L., Chung, C.-M., Chiang, K.-M., Ho, H.-Y., Ting, C.-T., Lin, T.-H., Sheu, S.-H., Tsai, W.-C., Chen, J.-H., Leu, H.-B., Yin, W.-H., Chiu, T.-Y., Chen, C.-I., Fann, C.S.J., Wu, J.-Y., Lin, T.-N., Lin, S.-J, Chen, Y.-T., Chen, J.-W.* and Pan, W.-H.* (2009/05). Genome-wide association study of young-onset hypertension in the Han Chinese Population of Taiwan. *PLoS One* **4**, e5459. SCI.

5. **Yang, H.-C.**[*], Liang, Y.-J., Chung, C.-M., Chen, J.-W. and Pan, W.-H. (2009/12). Genome-wide gene-based association study. *BMC Proceedings* **3**, S135. PubMed.
6. **Yang, H.-C.**[*], Lin, H.-C., Huang, M.-C., Li, L.-H., Pan, W.-H., Wu, J.-Y. and Chen, Y.-T. (2010/07). A new analysis tool for individual-level allele frequency for genomic studies. *BMC Genomics* **11**, 415. SCI.
7. **Yang, H.-C.**[*], Lin, H.-C., Kang, M., Chen, C.-H., Lin, C.-W., Li, L.-H., Wu, J.-Y., Chen, Y.-T. and Pan, W.-H. (2011/04). SAQC: SNP array quality control. *BMC Bioinformatics* **12**, 100. SCI.
8. **Yang, H.-C.**[*], Chang, L.-C., Huggins, R. M., Chen, C.-H. and Mullighan, C. G. (2011/05). LOHAS: Loss-of-heterozygosity analysis suite. *Genetic Epidemiology* **35**, 247-260. SCI
9. **Yang, H.-C.**[*] and Chen, C.-W. (2011/11). Region-based and pathway-based QTL mapping using a p-value combination method. *BMC Proceedings* **5**, S43. PubMed.
10. **Yang, H.-C.**, Liang, Y.-J, Chen, J.-W., Chiang, K.-M., Chung, C.-M., Ho, H.-Y., Ting, C.-T., Lin, T.-H., Sheu, S.-H., Tsai, W.-C., Chen, J.-H., Leu, H.-B., Yin, W.-H., Chiu, T.-Y., Chern, C.-I., Lin, S.-J., Tomlinson, B., Guo, Y., Sham, P. C., Cherny S. S., Lam, T. H., Thomas, G. N. and Pan, W.-H.[*] (2012/03). A genome-wide gene-based association study identifies *IGF1*, *SLC4A4*, *WWOX* and *SFMBT1* as hypertension susceptibility genes in a Han Chinese population. *PLoS One* **7**, e32907. SCI.
11. **Yang, H.-C.**[*], Chang, L.-C., Liang, Y.-J., Lin, C.-H. and Wang, P.-L. (2012/04). A genome-wide homozygosity association study identifies runs of homozygosity associated with rheumatoid arthritis in the human Major Histocompatibility Complex. *PLoS One* **7**, e34840. SCI.
12. **Yang, H.-C.**[*], Wang, P.-L., Lin, C.-W., Chen, C.-H. and Chen, C.-H. (2012/07). Integrative analysis of single nucleotide polymorphisms and gene expression efficiently distinguishes samples from closely related ethnic populations. *BMC Genomics* **13**, 346. SCI.