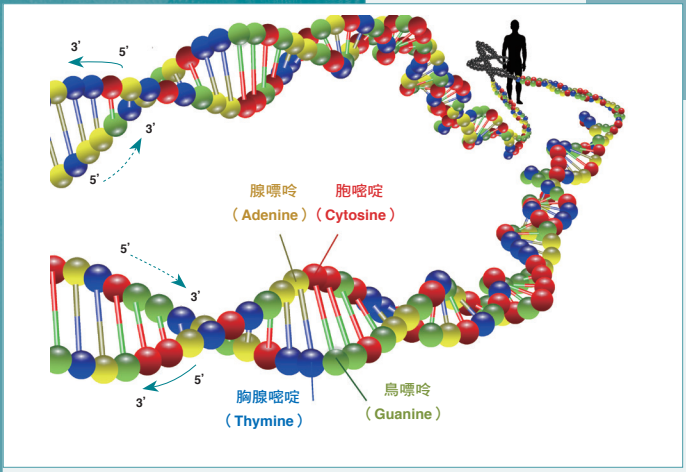
◎ 楊欣洲、陳佳煒

統計基因定位大戰

要戰勝敵人（疾病），必須先找出敵人（致病基因）。
人類與疾病間的戰爭未曾止歇，要戰勝疾病，
首要之務是找出致病基因在染色體上的位置。
統計資料科學可以透過分析大量基因體資料，
告訴你致病基因在哪裡。



生命密碼 DNA

戰勝疾病必須找出致病基因

人類與疾病間的戰爭已經纏鬥許久，許多疾病的歷史難以考據，而留有紀錄的，如早期的痲瘋病，最近的嚴重急性呼吸道症候群（SARS）、中東呼吸症候群冠狀病毒感染症（MERS）等，都是人類與疾病間一場場血淋淋的戰役。要在這些戰役中戰勝敵人（疾病），就必須先找出敵人（致病基因）。

疾病的類型不同，受基因影響的程度也不同。有些疾病是先天遺傳的，甚至是一個基因的缺陷就造成疾病發生，這類疾病稱為「孟德爾型疾病」（Mendelian diseases），通常十分罕見。例如，龐貝氏症是由於第 17 號染色體上酸性 α－葡萄糖苷酶基因突變所造成的隱性遺傳疾病。好萊塢電影〈愛的代價〉（*Extraordinary Measures*）中那位不服傳統的名醫，為了拯救因罹患龐貝氏症而生命岌岌可危的小生命的無私付出，這動人故事所描述的正是陳垣崇院士歷經 15 年研發出龐貝氏症藥物 Myozyme 的奮鬥歷程。

相對於罕見的孟德爾型疾病，常見的疾病往往是多基因遺傳，也稱為「複雜型疾病」。例如，詞曲作家黃舒駿先生筆下的〈戀愛症候群〉，就是一種典型的複雜型疾病，歌詞描述「一般發病後的初期反應……半夜突然爬起來彈鋼琴（失眠或夢遊症？），有人每天站在陽台上對著路人傻笑（失智？），有人突然瘋瘋癲癲突然很安靜（躁鬱症？），有人一臉癡呆對著鏡子咬指甲打噴嚏（癡呆症＋強迫症＋感冒？），有人對小狗罵三字經（心理疾病？）……食慾不振 歇斯底里 四肢萎縮 神經過敏 發抖抽筋（中毒？）……」

這類疾病的特性就是：表徵很多、成因複雜、致病基因很多但效果都很微弱，且還受基因以外的環境因素，甚至是基因與環境的交互作用的影響。

針對孟德爾型疾病和複雜型疾病的統計致病基因定位的方法雖不相同，但都必須仰賴「遺傳標記」這項武器的火力支援。人類有 23 對染色體，是由生命密碼—去氧核糖核酸（DNA）纏繞組織蛋白而組成，每一條染色體上都布滿許多的遺傳標記。就像高速公路上的里程牌協助標定出車輛的實際位置，遺傳標記是標定出染色體上的實際位置，遺傳標記布得愈多愈密，致病基因定位的結果就可以更精確。

生物科技產生大量基因體資料

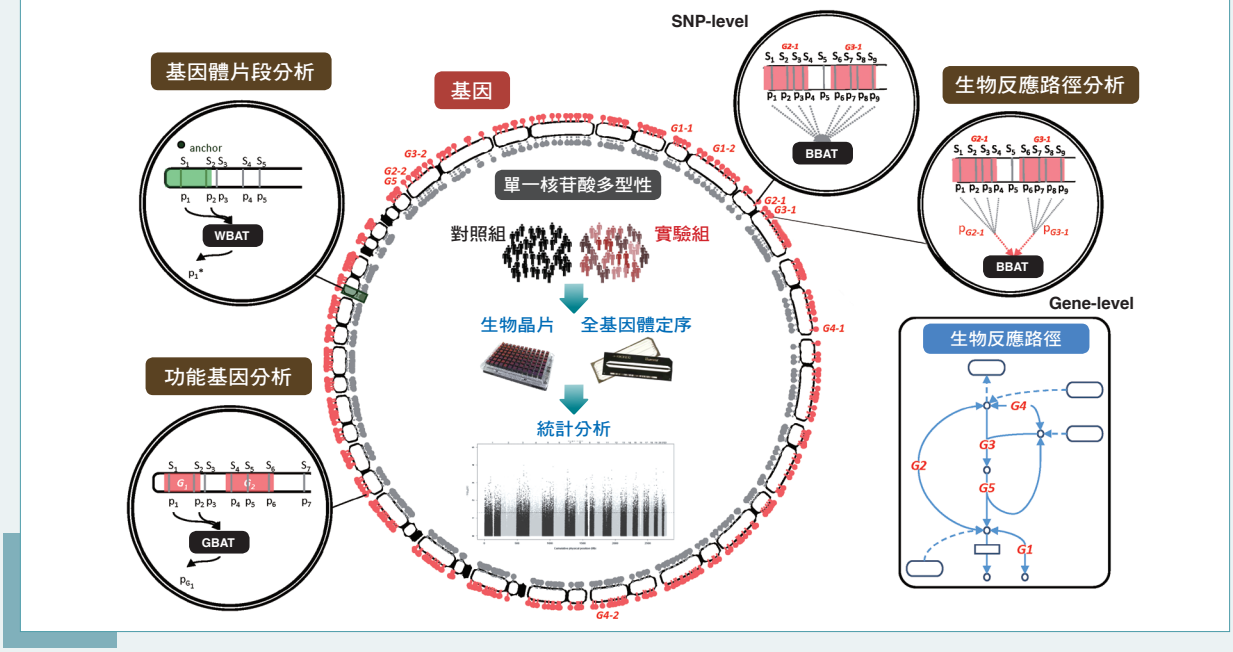
生命科學的登月計畫「人類基因體計畫」（Human Genome Project）從西元 1990 年開始，2003 年結束，歷時 13 年，完成人類基因體圖譜的草圖，正式進入後基因體時代。隨著生物科技的日新月異，當年世界多國合作耗費 13 年完成的約 30 億鹼基對的定序，現今只要單一實驗室短短幾日就可以完成再定序，也因此發現了大量新的遺傳標記，窺探人類基因體的奧妙已經不再是遙不可及的夢想！

利用生物晶片或全基因體定序實驗，可以快速且大量地產出龐大基因體數據和遺傳標記資料，裡面蘊藏豐富的生命密碼 DNA 及其與

遺傳標記是標定出染色體上的實際位置，
遺傳標記布得愈多愈密，致病基因定位的結果就可以更精確。

隨著生物科技的日新月異，
現今只要單一實驗室短短幾日就可以完成再定序，
也因此發現了大量新的遺傳標記。

統計致病基因定位的流程與方法



以病例對照研究設計來說明，在取得病例組（紅色人群）與健康對照組（黑色人群）的 DNA 檢體後，透過生物晶片實驗或全基因體定序實驗，可以獲得大量的單一核苷酸多型性的資料。利用單一遺傳標記分析或多遺傳標記分析（例如基因體片段分析、功能基因分析、生物反應路徑分析等），計算出統計顯著性（p 值），最後結果繪製成「曼哈頓圖」（Manhattan plot）。以單一遺傳標記分析的結果為例來解釋「曼哈頓圖」，圖中的一個點代表一個單一核苷酸多型性，橫軸是單一核苷酸多型性在各染色體上的實際物理位置，縱軸則是經過 –log10 轉換後的 p 值，–log10（p）值很大表示該遺傳標記的資料在病例組和健康對照組有統計顯著差異，也就是該遺傳標記與疾病有關，可能本身就是致病基因所在或很靠近致病基因，利用這統計致病基因定位方法可以快速找出致病基因在染色體上的位置。

疾病之間的關係。然而，從龐雜的基因體數據中窺探出生命的奧祕，找出與疾病的發生相關的染色體位置，難如大海撈針，統計資料科學正是克服這難題的利器。

生物晶片與基因體定序

統計基因定位有捷報也有警報

後基因體時代的基因體數據量呈現爆炸性成長，統計基因體學和生物資訊學變成顯學，百家爭鳴，許多科學家投注心力開發各種新穎的統計致病基因定位方法。除了僅針對單一遺傳標記的傳統統計關聯分析外，也有多遺傳標記的方法，包括基因體片段、功能基因、生物反應路徑的分析方法，統計致病基因定位快速發展，各種方法百花齊放。

統計致病基因定位就像官兵捉強盜一樣，統計方法（官兵）要在人類的 23 對染色體（台灣 6 個直轄市和 16 縣市）找出致病基因（強盜），要仰賴遺傳標記（全球定位系統）來定位致病基因（強盜）。針對複雜型疾病，目前最廣泛使用的統計致病基因定位方法稱為「全基因體關聯性研究」

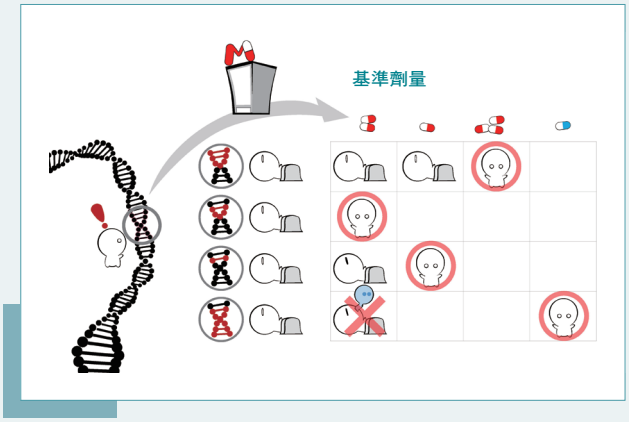
（genome-wide association study, GWAS），搭配的基因標記稱為「單一核苷酸多型性」（single nucleotide polymorphism, SNP）。

全基因體關聯性研究是一種利用統計推論方法判斷疾病與遺傳標記之間是否有關聯，進而定位出致病基因的方法。自西元 2005 年第一個成功的全基因體關聯性研究，找到重要的影響老年黃斑性病變的致病遺傳變異以來，到 2018 年 5 月已經有超過 5,000 個全基因體關聯性研究，發現約 69,000 個單一核苷酸多型性和疾病的關聯性，結果發表在超過 3,378 篇文章中。在這些重要的致病基因定位研究中，統計資料科學扮演著至關重要的角色。

在人類與疾病的對抗中，統計致病基因定位大戰烽火遍地。雖屢傳捷報，成功定位出許多疾病的致病基因，並能解釋相當比率的疾病遺傳度，例如老年黃斑性病變、第一型糖尿病等。但也有令人沮喪的消息，許多疾病所發現的致病基因僅能解釋該疾病很少的比率，很多重要的致病基因尚未找到，例如高血壓、躁鬱症等，亟需更多火力支援（生物科技與資訊科技）與幫手（統計資料科學家）加入這場聖戰，繼續搜捕之前逃脫的敵人的位置。此外，找到敵人離戰勝敵人還有一段路程，需要大家一起努力！

統計資料科學的黃金時代

「精準醫療」或「個人化醫療」的目標是「在對的時間，給對的病患對的藥物、對的劑量」，量身訂做適合每位病患的醫療與照護。要達到這理想，有賴妥善蒐集病患個人資訊（例如身高、體重、疾病家族史等）、生化檢測（例如膽固醇、糖化血色素等）、分子生物檢測（例如基因輪廓、代謝反應等）等資料。



精準醫療—根據個人 DNA 序列的組成，在對的時間，給予對的病患對的藥物、對的劑量，目標是避免藥物副作用，並達到最佳藥效反應。

圖中由左而右，由上至下，依序呈現人類基因體中的 23 條染色體示意圖（第 23 條是性染色體，本圖僅呈現 X 染色體上的結果）。每條染色體的不同位置釘上了不同顏色的大頭釘，不同顏色代表不同的疾病（例如咖啡色代表消化系統疾病、紅色代表心血管疾病、橘色代表代謝相關疾病等），若某位置上釘上了大頭釘，表示過去曾有全基因體關聯性研究報導該位置上的標誌基因與某研究的疾病有顯著統計相關。（圖片來源：由 EMBL-EBI 授權使用。參考文獻：Buniello et al (2019) Nucleic Acids Research, Vol. 47 （Database issue）: D1005-D1012.）

從龐雜的基因體數據中找出與疾病的發生相關的染色體位置，難如大海撈針，統計資料科學正是克服這難題的利器。

利用大規模的資料庫資料，進行統計資料科學的分析與探勘，將使致病基因定位更為精密，進一步往精準醫療的目標邁進。

更重要的是，必須深入了解以上資料與疾病發生和醫療反應之間的關係（例如致病基因定位的結果）。

台灣一個非常成功的精準醫療典範是史帝文強生症候群的研究。陳垣崇院士團隊利用統計基因定位方法（費雪正確檢定），成功地從眾多遺傳標記中找出位在第六號染色體上 HLA-B 這個遺傳危險因子，發現攜帶 HLA-B*1502 等位基因的病患在使用神經性疾病止痛藥物 Carbamazepine 後，會產生嚴重皮膚過敏副作用，也稱為史帝文強生症候群，嚴重可能致死，這項研究成果於西元 2004 年發表在《自然》期刊中。

在進一步經過大規模前瞻性研究的成功驗證後，西元 2010 年，健保局通過新增給付項目「HLA-B*1502 基因檢測」，給首次使用 Carbamazepine 的病患一次免費 HLA-B*1502 基因檢測。這項醫療政策不僅大幅降低史帝文強生症候群的發生，增進病人醫療安全與健康福祉，更可為國家省下大筆的醫療支出。

近年來，各國爭相投入大量經費發展精準醫療。西元 2015 年，美國總統歐巴馬批准「精準醫療啟動計畫」，要招募超過一百萬位美國志願參與者建立大型的人體生物資料庫。除了蒐集醫療紀錄、生活習慣、社會環境等相關的電子健康數據外，還包括精準醫療啟動計畫中最至關重要的基因資料。

西元 2018 年，美國國家衛生研究院正式宣布啟動「全民健康研究計畫」，將對精準醫療啟動計畫的參與者進行全基因體定序實驗，建立龐大的基因庫，探索致病相關和醫療反應相關的基因，加速實現透過基因體研究的成果進行精準醫療的目標，以改善美國人民的健康。除了美國以外，各先進國也先後建立國家級的人體生物資料庫，包括冰島、英國、日本等。

「臺灣人體生物資料庫」（Taiwan Biobank）於西元 2012 年經衛生福利部許可後設置，

臺灣人體生物資料庫（圖片來源：由中央研究院臺灣人體生物資料庫授權使用）

是台灣生命科學與精準醫療的重大基礎建設（https://www.twbiobank.org.tw）。將招募 20 萬名 30 ～ 70 歲，未經醫生確診罹患癌症的本國一般社區民眾加入，也與各醫學中心合作，邀請特定疾病組的 10 萬位病患加入。除了蒐集問卷資料和一般身體檢測資料外，也採集 DNA、血液、尿液等檢體檢驗。

另外，也進行檢體加值，包括全基因體基因型鑑定或定序、DNA 甲基化測量、代謝體實驗等，將產生大量生活型態、環境因子、疾病史、臨床檢驗、分子生物標誌資料等。利用這大規模的資料庫資料，進行統計資料科學的深度分析與探勘，將使統計致病基因定位更為精密，對影響國人的疾病的致病基因有更多、更完整的了解，讓台灣進一步往基因體醫學與精準醫療的目標邁進，以增進國人的健康福祉。

楊欣洲、陳佳煒
中央研究院統計科學研究所