目錄

一、 概述

- **第一代臺灣人體生物資料庫基因定型晶片**

採用美國 Thermo Fisher Scientific 公司所研發之技術平台，與國家基因體研究中心（National Center for Genome Medicine, NCGM）合作設計專屬臺灣漢人的單一核苷酸多型性（Single Nucleotide Polymorphism, SNP）基因定型晶片，挑選共 653,291 個 SNPs，以臺灣人體生物資料庫(Taiwan Biobank, TWB)的英文縮寫 TWB 命名。

此晶片 SNP 挑選之來源包含：

- Axiom Genome-Wide CHB Array，共 525,652 個 SNPs。
- 已發表之癌症全基因體關聯性研究（Genome-wide Association Study, GWAS）中，具有統計顯著意義之 SNPs。
- NCGM 過去使用各種晶片所得的結果中，挑選在國人樣本中具有多型性（polymorphism）之 SNPs。
- 採用全外顯子定序（Whole Exome Sequencing, WES）與其他定序研究方法中，挑選在國人樣本中具有多型性之 SNPs。
- 其他與藥物反應、藥物代謝相關，如 MHC，PGX 等基因上的 SNPs。

- **基因定型方法**

臺灣人體生物資料庫每年隨機挑選參與者使用此晶片進行基因型鑑定，個案之挑選依據內政部戶政司人口統計之性別、5 歲年齡組別及縣市別分布為基準。委託 NCGM 使用 GeneTitan Multi-Chanel instrument 自動化操作平台進行基因型鑑定實驗，每一個 96 孔盤包含 95 位參與者樣本，及 1 位 Thermo Fisher Scientific 公司提供之對照樣本，每一樣本取 14 ng/μL 濃度之 DNA 進行，以原廠針對 Axiom 平台所推出之 Axiom Analysis Suite 免費分析軟體之最佳分析流程（Best Practice Workflow）進行資料分析及結果輸出，技術平台及原理可參考 http://ncgm.sinica.edu.tw/affymetrix_tech_01.html。

實驗依批次分別進行資料的品質控制（Quality Control, QC），當某批次的某 SNP 未符合原廠所設立之 QC 條件時，該批次中所有樣本的此 SNP 基因型將會紀錄為遺漏值。更多關於基因型判讀及品質控制的資訊可參考 http://tools.thermofisher.com/content/sfs/manuals/axiom_genotyping_solution_analysis_guide.pdf。

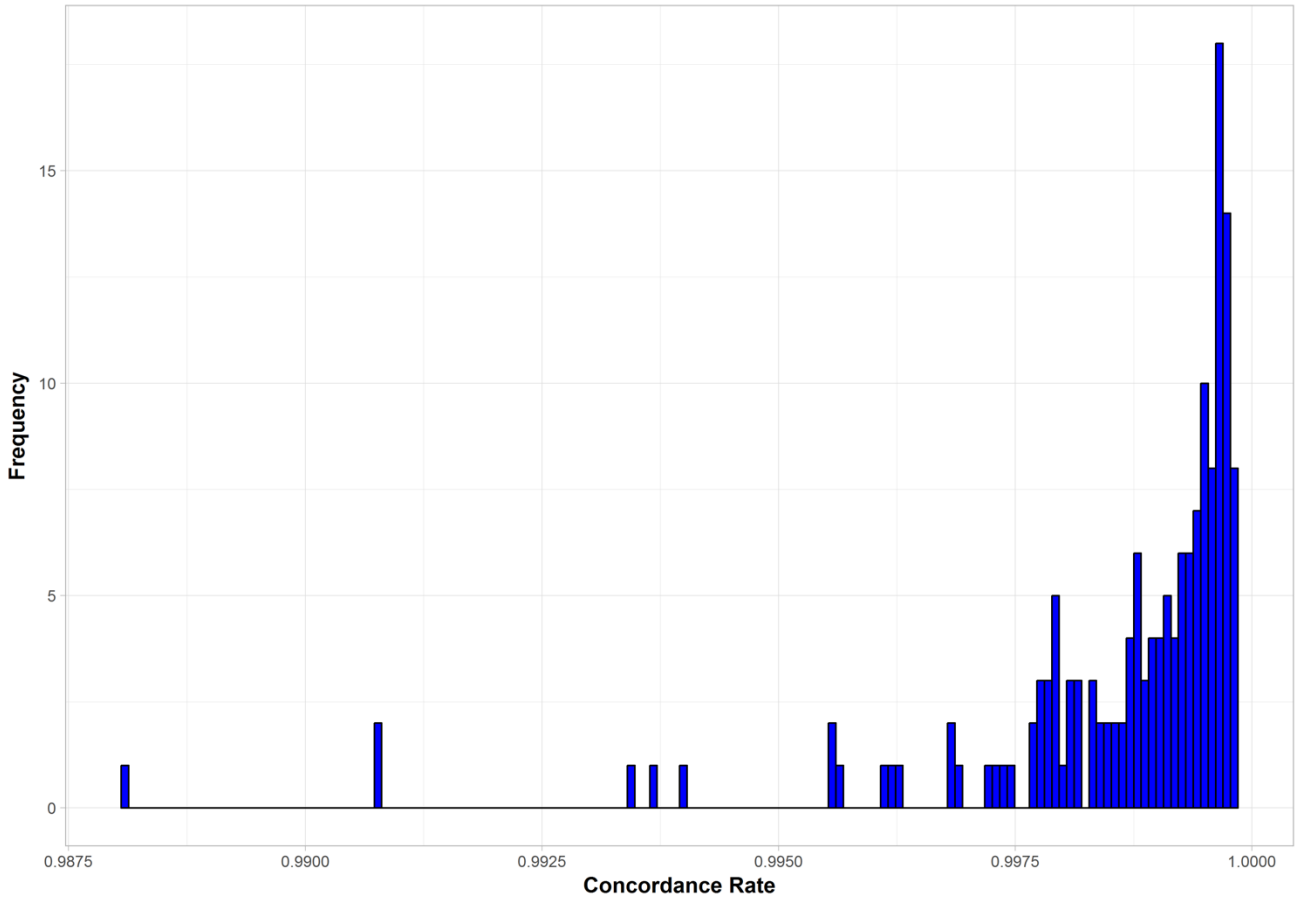
下表呈現目前可釋出之 27,751 個案中，每批次（batch）使用的批號（lot）數量、使用的盤數（plate）、樣本數及通過原廠 QC 建議之 SNP 數量。SNP 所屬染色體位置基於人類基因組序列第 37 版（Genome Reference Consortium Human Reference 37, GRCh37）座標。

批次	批號數量	盤數	樣本數	通過原廠 QC 建議 SNP 數
1	3	22	1,849	633,828
2	4	22	2,062	639,761
3	5	23	2,156	644,898
4	2	20	1,843	636,564
5	2	23	2,169	641,050
6	3	20	1,893	639,579
7	2	17	1,604	638,551
8	4	26	2,454	643,334
9	4	22	2,079	643,555
10	4	21	1,982	643,063
11	2	21	1,979	643,969
12	3	22	2,088	637,035
13	3	23	1,907	641,374

14	3	20	1,686	632,552
Total	44	302	27,751	646,973

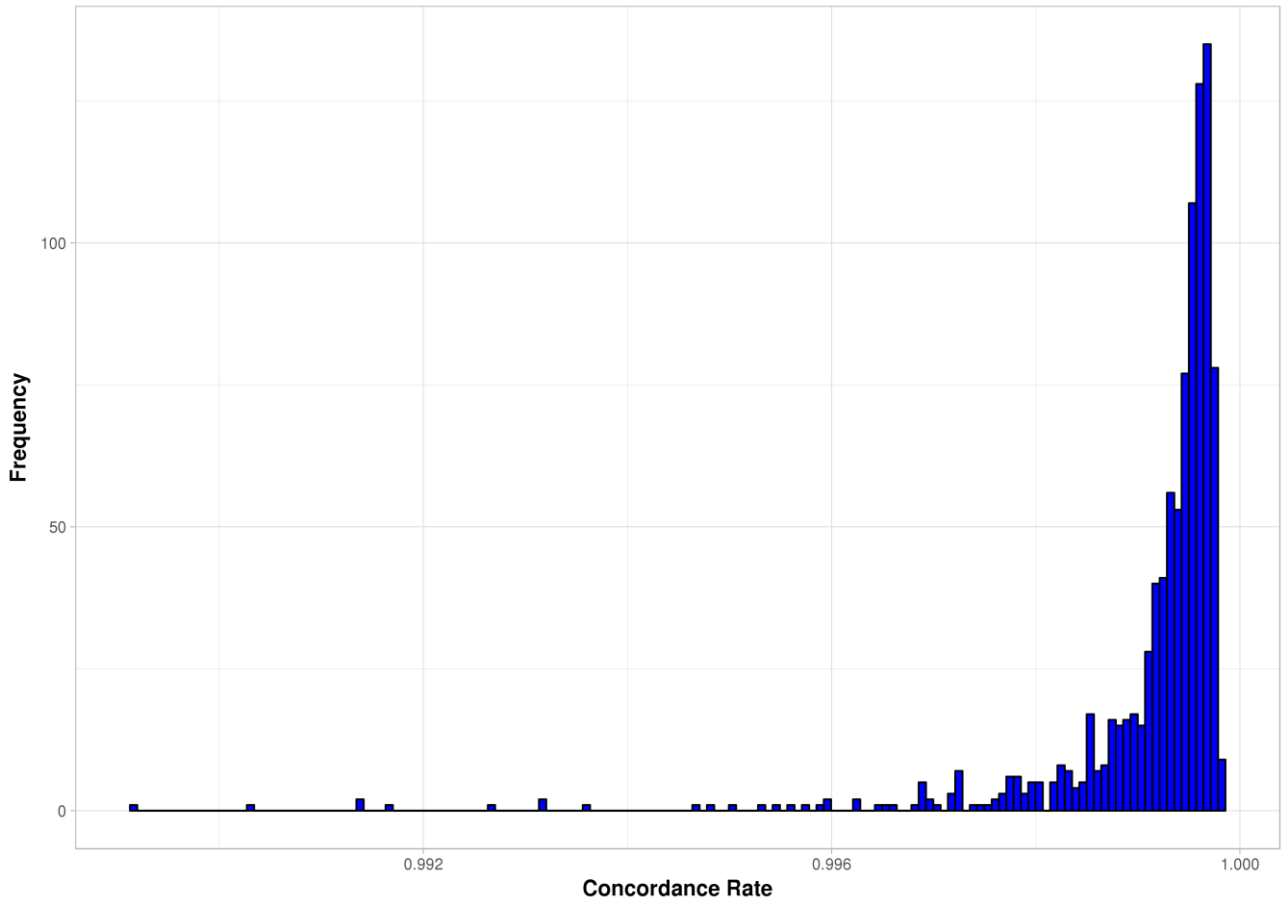
- **重複個案之基因型一致性**

我們利用 bcftools 軟體的 ”gtcheck” 指令評估重複個案之基因型一致性。在第一批次完成個案中，包含 157 對重複樣本，基因定型結果呈現相當高的一致性，如下圖所示。629,538 個通過原廠建議 QC 的 SNP 中，平均一對重複樣本帶有相同基因型之一致性為 99.86%，最低的一致性仍有 98.81%，下圖呈現此 157 對重複樣本一致性之分布情形。



我們亦使用 968 位同時具有全基因體定型資料與全基因體定序（Whole Genome Sequencing, WGS）資料之樣本，比對體染色體（autosome）上約 59 萬個 SNPs 的一致性。兩種實驗平台整體的一致性平均為 99.92%，最小值也有 98.92%，下圖呈現此 968 位個案之基因

型在兩種平台一致性之分布。



二、 基因定型資料批次效應

臺灣人體生物資料庫目前可釋出的全基因體定型資料分為 14 個批次的實驗完成，由於樣本均來自相同族群，且每個批次內的樣本組成是隨機的，因此預期某 SNP 之基因型頻率在不同批次間不會有統計上的顯著差異。同樣地，某 SNP 之基因型頻率在同一實驗批次內的不同實驗因子（批號, 盤數）間也應該沒有統計上的顯著差異。若某 SNP 於不同實驗批次或同批次內不同實驗因子間，存在統計上的顯著差異差異，代表此 SNP 沒有正確的判讀。我們使用 Fisher's exact test 來比較某實驗批次的基因型頻率與合併其他剩餘批次後的基因型頻率，當某 SNP 的 p 值小於所定義之閥值（1*e-10）時，該 SNP 視為受批次效應影響（batch effect），該批次中的此 SNP 基因型將會調整為遺漏值。同樣地，我們也比較同一實驗批次內某盤的基因型頻率與同一實驗批次內合併其他剩餘盤後的基因型頻率，當某 SNP 的 p 值小於所定義之閥值（1*e-10）時，將視為因不同盤所造成的批次效應（plate effect），該實驗批次所有盤中的此 SNP 基因型將會調整為遺漏值。同樣的方式也被用於判斷批號的批次效應（lot effect）。

我們分別進行上述這些統計檢定，因此某些 SNPs 可能同時存在實驗批次效應、盤批次效應及批號批次效應。針對體染色體與 X 染色體上的 SNPs，使用 2*3 列聯表進行分析；粒腺體上的 SNPs，使用 2*2 列聯表進行分析；Y 染色體上的 SNPs，僅針對男性個案使用 2*2 列聯表進行分析。下表列出每個 batch 裡具有 batch effect、plate effect 或 lot effect 的 SNP 數量。

批次	因實驗批次效應影響的 SNP 數量	因盤批次效應影響的 SNP 數量	因批號批次效應影響的 SNP 數量	有批次效應影響的 SNP 數量
1	656	835	323	1,245
2	297	20	9	317
3	366	6	3	370

4	286	17	4	301
5	332	15	5	347
6	269	24	27	272
7	181	4	1	184
8	269	11	8	278
9	323	16	16	329
10	293	0	0	293
11	277	12	0	288
12	340	83	91	365
13	323	909	28	1212
14	398	75	70	471
Total	2,418	2,013	577	4,035

三、 基因定型資料品質管控

我們使用經過 batch effect 處理後之 SNPs，共 27,751 位個案及 646,973 個 SNPs 進行後續的資料品質管控。

- **樣本祖源（diverse ancestry）**

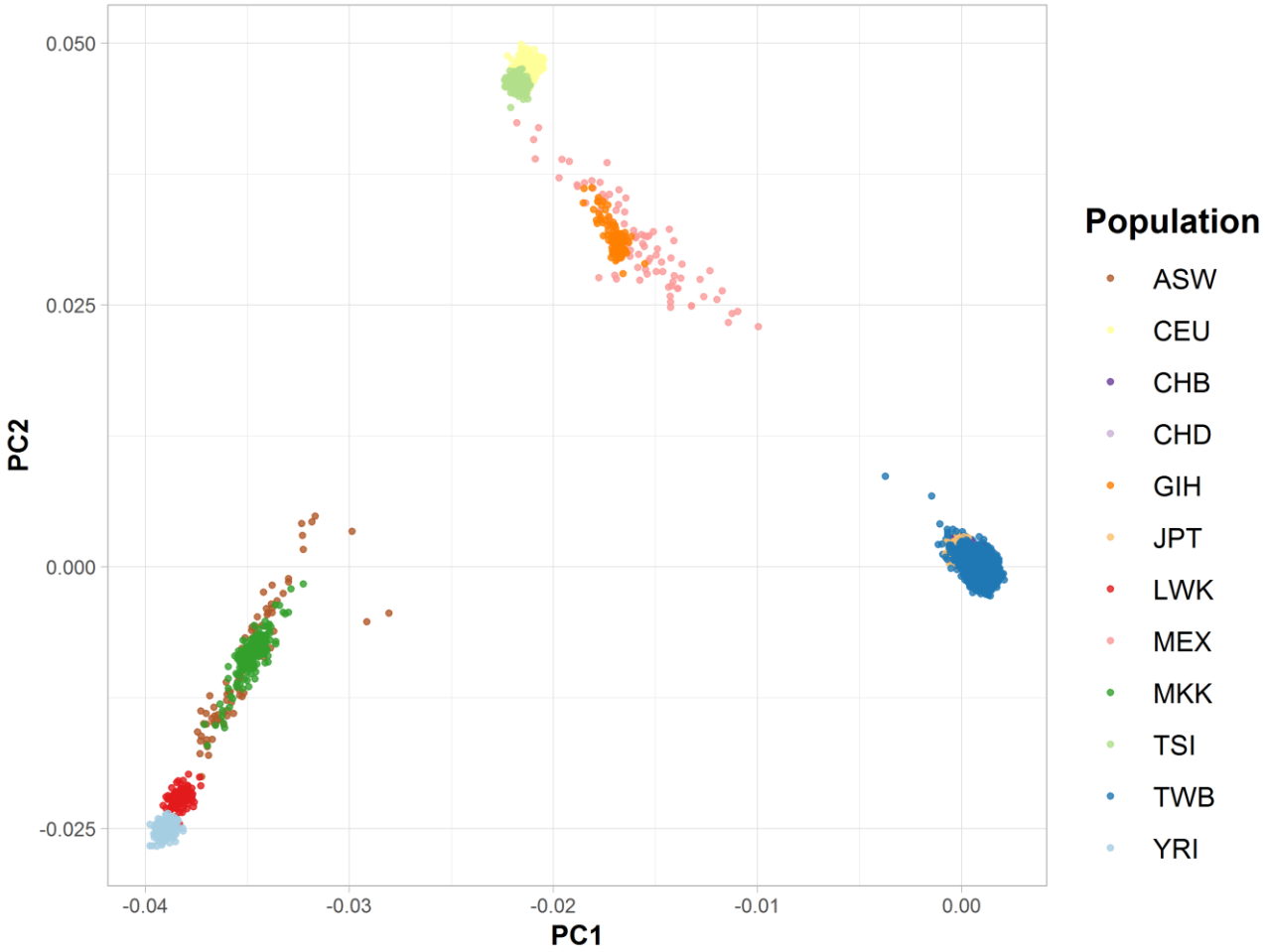
為確保我們的樣本皆來自於臺灣的參與者，我們採用主成份分析（principle component analysis, PCA）檢視資料中是否具有不同祖源之個案。首先排除符合下列特性之樣本：

- 缺失率大於 5%之個案
- 異質率超過平均異質率 5 倍標準差之個案
- 三等親內成對樣本中帶有較高缺失率之個案

並且納入來自國際人類基因組單倍型圖譜計劃（International Haplotype Map Project, HapMap）第三階段的 1,397 人作為參考族群，截取與第一代臺灣人體生物資料庫基因定型晶片重疊但不滿足下列特性之 SNPs：

- 缺失率大於 5%之 SNPs
- MAF 小於 5%之 SNPs
- 位在遠距離連鎖不平衡（long-range linkage disequilibrium）區域之 SNPs

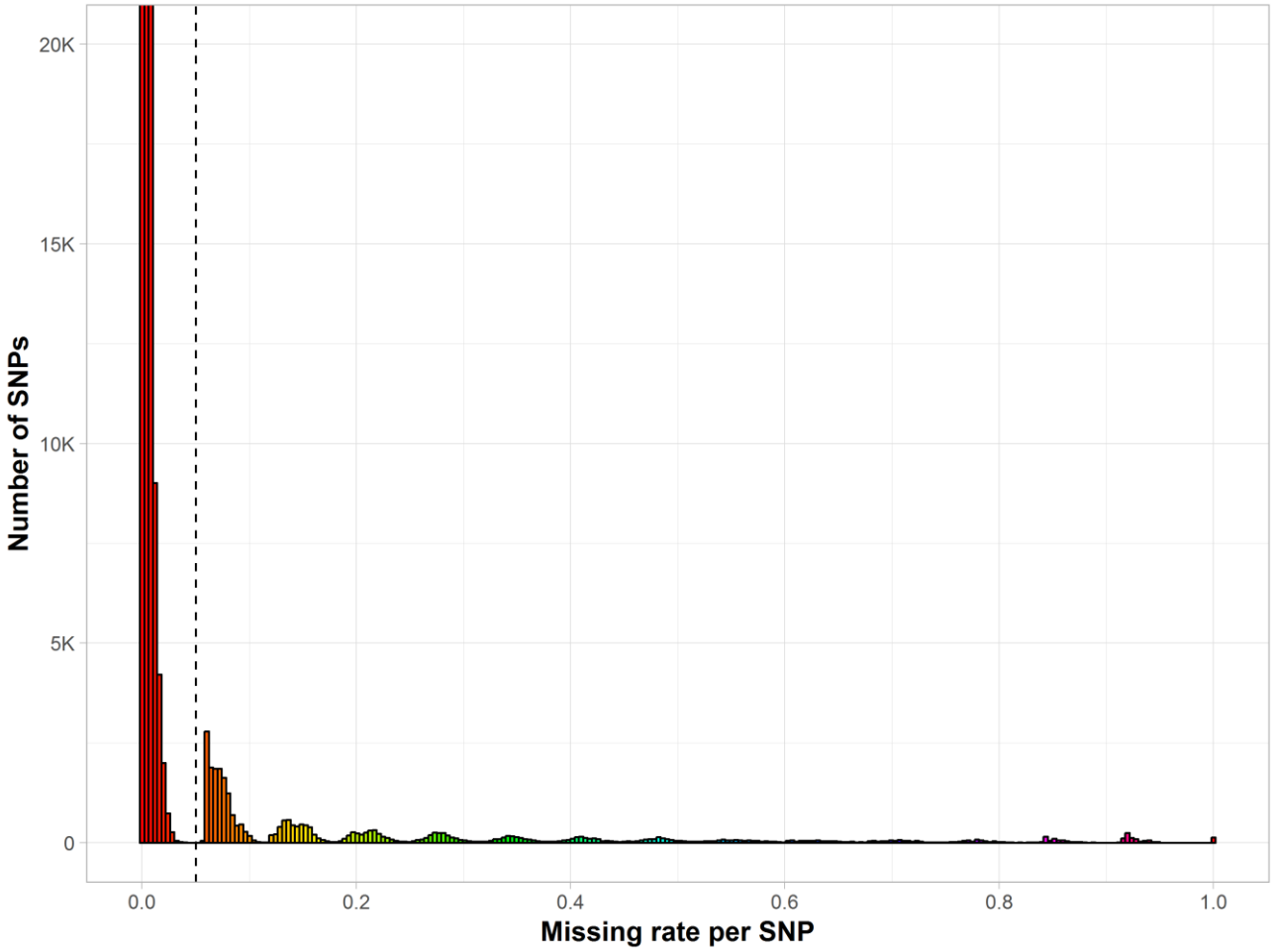
接著利用 PLINK 中的”--indep-pairwise 200 5 0.2”指令篩選一組獨立的 SNPs，最終獲得 25,873 位個案及 54,631 個 SNPs 進行 PCA 分析。下圖為 PCA 分析結果，不同顏色表示不同參考族群，可明顯觀察到 Taiwan biobank 參與者跟 CHB（Han Chinese in Beijing, China）、CHD（Chinese in Metropolitan Denver, Colorado）、JPT（Japanese in Tokyo, Japan）等亞洲族群均聚集在右邊區塊，顯示 Taiwan biobank 參與者均來自相同的祖源。



ASW, African ancestry in Southwest USA; CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; CHB, Han Chinese in Beijin, China; CHD, Chinese in Metropolitan Denver, Colorado; GIH, Gujarati Indians in Huston, Texas; JPT, Japanese in Tokyo, Japan; LWK, Luhya in Webuye, Kenya; MEX, Mexican ancestry in Los Angeles, California; MKK, Maasai in Kinyawa, Kenya; TSI, Toscans in Italy; TWB, Taiwan Biobank; YRI, Yoruba in Ibadan, Nigeria

- **SNP 缺失率（missing rate）**

SNP 的缺失率是指在所有樣本中，有多少比例的樣本沒有被判讀，過高的缺失率可能導致偽陽性及降低發現與疾病相關 SNP 的機會。如下圖如示，垂直虛線代表缺失率為 5%，共 27,543 個 SNPs 的缺失率大於 5%，會於後續加以排除。



哈溫平衡（Hardy-Weinberg equilibrium, HWE）

假設一個 SNP 的兩個對偶基因（a 及 A）的頻率分別為 p 與 q 時，哈溫平衡是在檢測 aa、aA 及 AA 是否偏離了 p^2、2pq 及 q^2 分佈，偏離了此平衡可能代表實驗上的誤差或判讀基因型時的誤差，會使後續的分析產生偽陽性。然而，與疾病相關的 SNP 也可能不符合哈溫平衡，因此，為了避免疾病狀態可能影響基因型的分佈，我們只選取一組自述無任何疾病史的

個案執行此項檢定，共 10,300 人。以 PLINK 中的”--hardy”指令進行 HWE 檢定，共 7,613 個 SNPs 的 p 值小於 1*e-5，會於後續加以排除。

次要對偶基因頻率（Minor Allele Frequency, MAF）

低頻率的 SNP 往往沒有足夠的統計檢力（statistical power），以 PLINK 中的”--freq”指令進行計算，共 8,571 個 SNPs 的 MAF 小於 5%。

四、 基因定型資料品質綜述

下表列出不符合每一 QC 條件下的樣本數或 SNP 數，及其佔所有 27,751 位個案或 646,973 個 SNPs 的百分比。

Quality control	Number of samples or SNPs	Percentage of all samples or all recommended SNPs
Sample with mean heterozygosity rate > 5 SD	145	0.52%
Sample with missing rate > 5%	0	0%
Sample with the highest missing rate from each of the 3nd degree pair	3,130	11.28%
SNP with HWE p-value < 1*e-5 in super control subset	7,613	1.18%
SNP with missing rate > 5%	27,543	4.26%
SNP with MAF < 5%	8,571	1.32%

五、 確立單倍型及基因差補

基因型插補（imputation）是預測樣本中未直接定型的基因型的一個過程，這些經由電腦模擬出來的基因型能夠增加 SNP 與疾病相關性檢定的 SNP 數量，增加研究的檢定力。插補的過程分成兩步驟：（一）單倍型推估（pre-phasing）與（二）基因插補。第一步驟中，針對樣本中已完成定型的 SNPs 進行事先的基因分型，即推論每個樣本的單倍型（haplotype），第二步驟是結合推論出的單倍型與基因體差補參考模板（reference panel）的單倍型，來插補樣本中未經實驗獲得的基因型。

- 單倍型推估（**pre-phasing**）

此步驟中，我們使用經過 batch effect 處理的檔案，並保留所有 27,751 位個案，但排除具有下列特性之 SNPs：

- MAF 小於 1%之 SNPs
- HWE 檢定 p 值小於 1*e-5 之 SNPs
- 超過 1 個 batch 為缺失之 SNPs

最後留下 607,779 個位在體染色體上之 SNPs，接著使用 SHAPEIT2（v2.r790）軟體進行基因分型。

參考序列集（reference panel）

有許多因素會影響基因型插補的精確度，但一般而言，精確度會隨著參考序列集的單倍型數量的增加而提高，以及研究對象的祖源是否與參考序列集的祖源接近。在上述的祖源分析中，Taiwan biobank 參與者與鄰近的亞洲族群聚集在一起，顯示參考序列集中應有較多帶有亞洲祖源的單倍型數量，此外 Taiwan biobank 持續在進行 WGS 實驗，目前已完成 1000 位個案，我們使用其中 973 位通過 QC 的個案（TWB panel），再加上千人基因組計劃（1000 Genome Project）中 504 位來自東亞國家的個案（EAS panel）做為參考序列集，並排除具有下

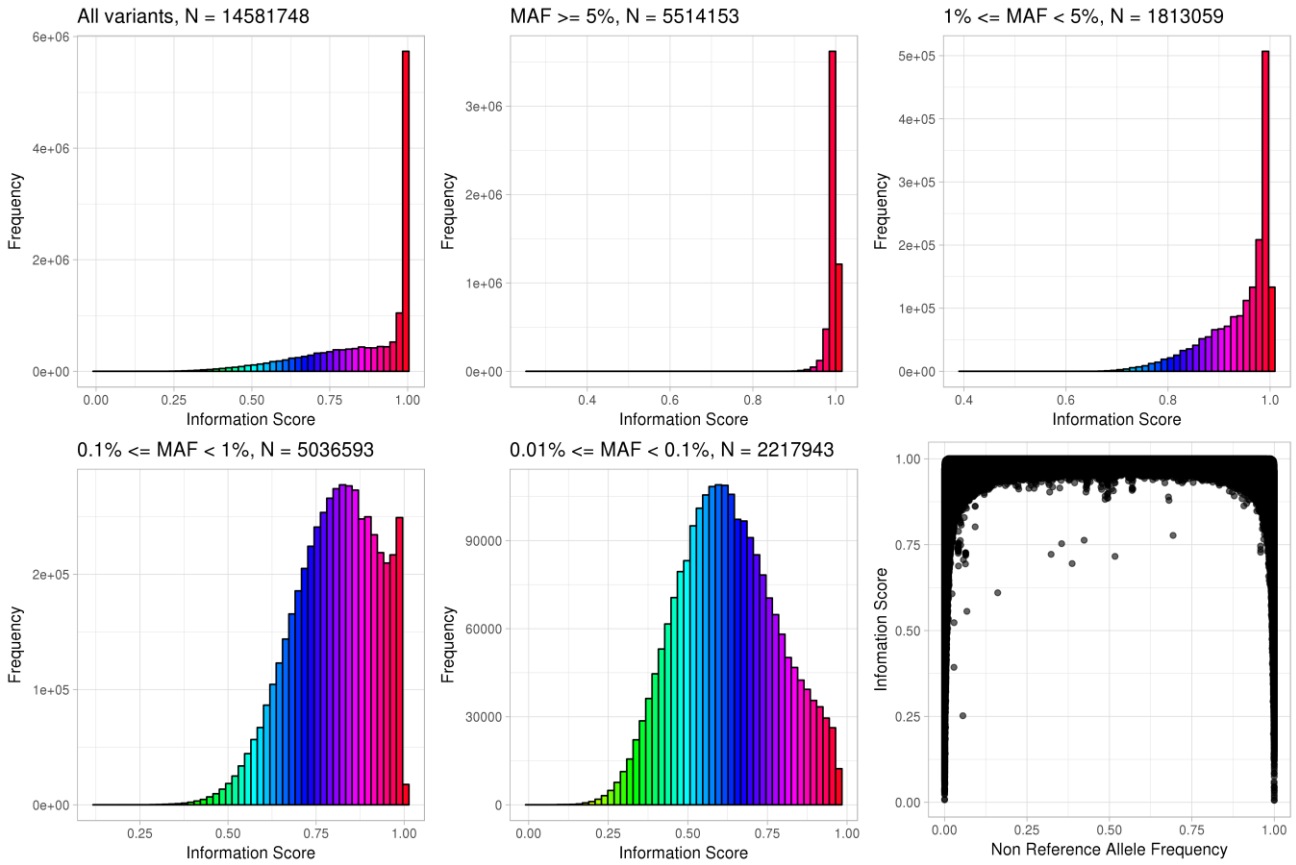
列特性之位點：

- 次要對偶基因次數小於 3 之位點
- 多等位基因（multi-allelic）之位點
- 兩組參考序列集的次要對偶基因不一致之位點
- 兩組參考序列集的次要對偶基因與 TWB 晶片不一致之位點
- EAS panel 中，重覆位置及 HWE 檢定 p 值小於 1*e-10 之位點

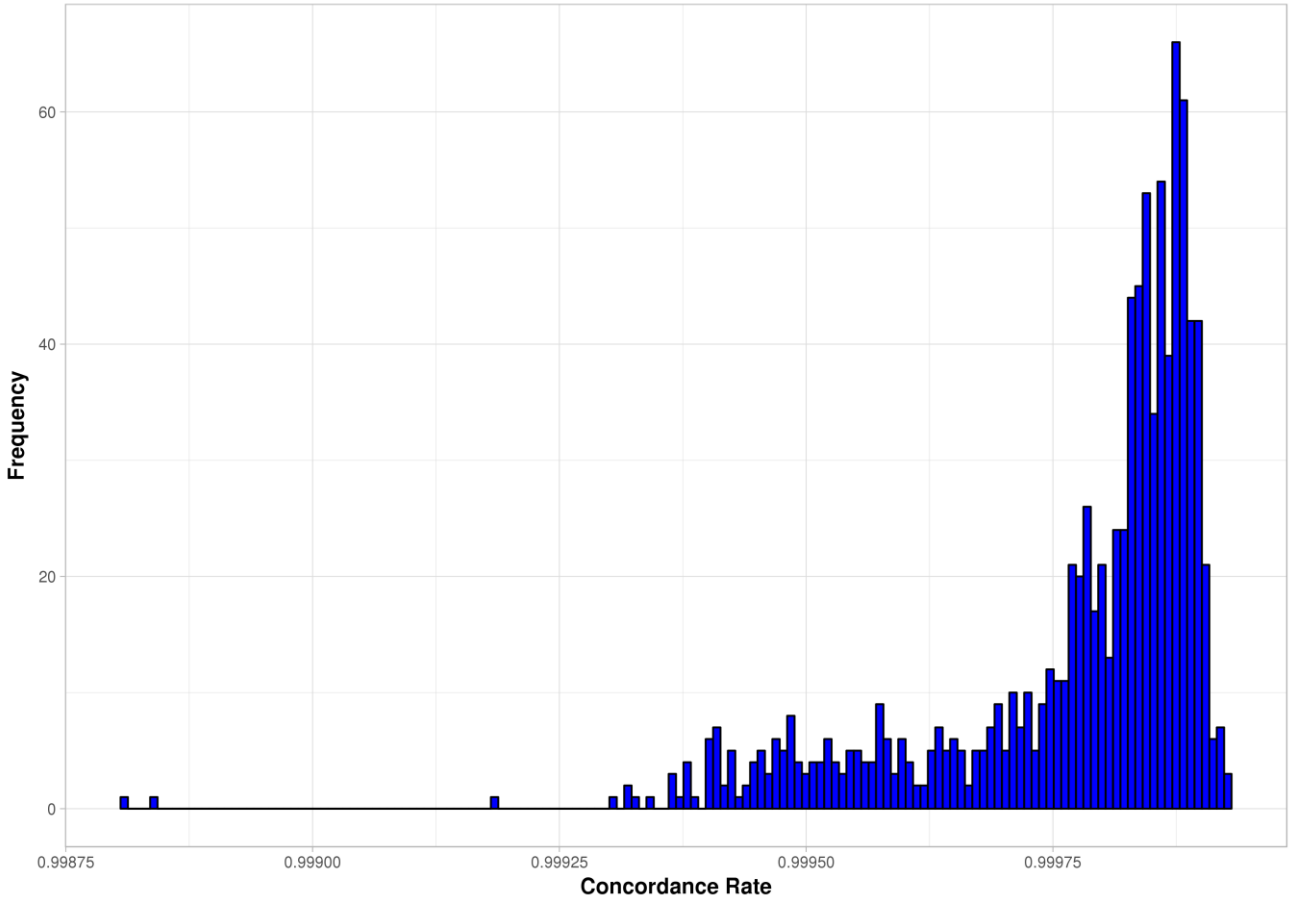
最後，TWB panel 和 EAS panel 分別留下 13,600,770 個和 11,669,571 個體染色體上之雙等位基因（bi-allelic）位點，同樣使用 SHAPEIT2 軟體進行基因分型，再以 IMPUTE2 軟體的"-merge_ref_panels"指令合併兩組序列集。

插補（imputation）

我們使用 IMPUTE2（v2.3.1）軟體進行插補，以 500 Mb 為一個區塊（chunk），前後加上 500 kb 的重疊緩衝區域，最後產出 15,866,355 個插補後的位點。針對每一位個案的每一個位點的三種基因型，插補過程會產出一個機率分布，如假設一個位點的兩個對偶基因分別為 a 及 A，三種基因型分別為 aa, Aa 及 AA，經過插補後，一位個案會產出基因型為 aa 的機率 P_{aa}、基因型為 Aa 的機率 P_{Aa} 及基因型為 AA 的機率 P_{AA}，而三個機率的加總會是 1。接著我們使用 PLINK 軟體中的"--hard-call-threshold 0.1"指令將機率轉換成真正的基因型，但僅有在機率大於等於 0.9 時才會進行基因型的判讀，若三個基因型的機率皆小於 0.9 時，則此位個案的此位點會判讀為遺漏值。另一方面，許多插補後的位點為單型性（monomorphic），或是接近單型性，因此我們進一步排除缺失率大於 5%及 MAF 小於 0.01%的位點，最後剩餘 14,581,748 個位點。下圖呈現在不同 MAF 分層下，14,581,748 個位點的 information score 分布。information score 是一個介於 0 至 1 的數值，0 代表某位點在插補過程是完全地不確定，然而 1 是代表沒有不確定性，此圖可觀察到頻率大於 0.1%的位點大部分都帶有較高的 information score。



在進行後續分析時通常會排除插補不精確的位點，這可透過要求 information score 需大於某一切點來達成，普遍使用的切點是 0.3 或更大。因此我們進一步篩選出 information score 大於 0.3 的位點，最後過濾出 14,555,421 個位點供資料釋出，並且註解上單核苷酸變異資料庫（The Database of Short Genetic Variation, dbSNP）147 版本的 reference SNP (rs) 編號。另外，我們以 bcftools 軟體的“gtcheck”指令比對 968 位具有 WGS 資料的個案，比較插補後的位點與 WGS 資料間的一致性，結果如下圖所示，在平均 1,283 萬個位點中，平均一致性約 99.98%，最小值為 99.88%。


Frequency
60
40
20
0
0.99875
0.99900
0.99925
0.99950
0.99975
Concordance Rate