

Imputation of HLA alleles

2024/12/26

人類白血球抗原（Human Leukocyte Antigen, HLA）基因位於人類第 6 號染色體短臂上（6p21.3）的主要組織相容性複合體（Major Histocompatibility Complex, MHC）區域內，對於免疫反應的調控至關重要，並與許多疾病的易感受性相關。然而，由於 HLA 區域的高度多型性（polymorphic）與連鎖不平衡（linkage disequilibrium, LD），透過定序方式直接測定 HLA 基因型的成本高昂且技術挑戰巨大，因此不易大規模地適用於大樣本。HLA 基因型的插補（imputation）透過結合低成本的單一核苷酸多型性（Single Nucleotide Polymorphism, SNP）晶片資料和 HLA 參考序列集（reference panel），提供了一種經濟又有效的替代方案，廣泛應用於疾病研究和族群遺傳學分析。

臺灣人體生物資料庫委託有勁生物科技股份有限公司使用 Ion S5 XL 定序平台搭配 NXType NGS HLA typing kit 進行 HLA 分型實驗，以 TypeStream 軟體進行資料分析與結果輸出，最高能得到 8 位元（8-digit）解析度的分型結果，共計完成 1097 位參與者。其中，1095 人同時有 TWB1 晶片資料，877 人同時有 TWB2 晶片資料，以此資料集作為參考序列集的來源，並且挑選缺失率（missing rate）小於 5%、次要對偶基因頻率（minor allele frequency, MAF）大於 0.5% 以及哈溫平衡（Hardy-Weinberg equilibrium, HWE）檢定 P 值大於 $1e-6$ 的 SNPs，透過 R 的套件 HIBAG（v1.30.2）分別為 TWB1 與 TWB2 晶片建立 8 個 HLA 基因型的預測模型，包含 HLA-A、HLA-B、HLA-C、HLA-DPA1、HLA-DPB1、HLA-DQA1、HLA-DQB1、HLA-DRB1 等基因，再將模型套用至全基因體定型數位資料集，得到每位參與者插補後的 HLA 分型資料，以及相對應的機率。

我們將參考序列集分成 5 份，利用 5 倍交叉驗證的方式來評估此參考序列集的表現能力，結果如下表所示，以 0.5 設為事後機率（posterior probability）的閾值，兩張晶片在 8 個 HLA 基因的插補準確性都在 94.8% 以上。

Performance of estimated two-field HLA types with a posterior probability call threshold of 0.5.

Array	performance	A	B	C	DPA1	DPB1	DQA1	DQB1	DRB1
TWB1	Accuracy	96.3%	94.8%	97.4%	99.5%	97.8%	98.8%	99.1%	96.2%
	Call rate	95.4%	87.9%	98.1%	98.4%	95.4%	82.7%	99.3%	89.2%
TWB2	Accuracy	96.6%	95.4%	98.2%	99.5%	98.1%	99.0%	99.1%	95.4%
	Call rate	95.8%	87.8%	98.4%	98.6%	95.2%	86.4%	99.1%	91.2%

插補後的 HLA 分型資料可以利用 HIBAG 套件與疾病或其他表現型資料進行相關性統計，此軟體提供線性迴歸與邏輯斯迴歸模型，並且能搭配不同遺傳模式的分析，如顯性模式、隱性模式或是加乘模式等，同時設定事後機率的閾值，也能依需求設定 HLA 分型的解析度，相關操作流程請參考軟體的文件

（https://www.bioconductor.org/packages/release/bioc/vignettes/HIBAG/inst/doc/HLA_Association.html）。