國立臺灣大學公共衛生學院流行病學與預防醫學研究所

碩士論文

Graduate Institute of Epidemiology and Preventive Medicine
College of Public Health
National Taiwan University

Master Thesis

以 *Q* 值適性結合法來指出罕見致病變異

Pinpointing Rare Causal Variants with the Adaptive Combination of *Q*-values Method

李貞儀

Jen-Yi Li

指導教授：林菀俞 博士

Advisor: Wan-Yu Lin, Ph.D.

中華民國 105 年 7 月

July, 2016

doi:10.6342/NTU201602253

誌　謝

在論文即將完成的同時，首先要感謝我的指導教授林菀俞老師，在這一年多裡的研究過程中，每當我碰到問題時，像是程式寫錯、過去文獻理解錯誤，老師都會細心地一一教我，並給予許多觀念上的啟發，讓我有想法，使得我對研究內容能夠更進一步的了解。此外，每週的討論，老師很認真慢慢地指引我的研究方向，在這過程中，雖然我常常會在同個地方打轉，頭腦轉不過來，但老師非常有耐心一直講解給我聽，有時也會讓我回家想想再告訴她我的想法。經歷過無數個研究的日子，雖然過程有點辛苦，但也讓我對於研究有了深深的體會，且我也獲得很多關於基因方面的知識，以及提升自己撰寫程式的能力，因此我的研究能夠順利進行真的要非常感謝我的老師。另外，我的老師時常跟我說研究生要細心和謹慎，我覺得我這方面真的還需要多多加強，這樣出了社會才能夠比別人更有競爭力。

關於論文的修訂上，我要感謝我的口試委員們李文宗老師、邱燕楓老師和范盛娟老師 (依照姓氏筆畫排序)，這三位老師對於本論文提供了許多寶貴的建議，像是論文的撰寫方式，以及觀念上的釐清，讓我知道自己不足的地方，使我收益良多，而有了這些建議後也讓本論文能夠更加的完整。

再來，我要感謝我親愛的家人，因為有你們在背後默默的付出與陪伴著我，讓我無後顧之憂的專心完成學業，且在我研究論文遇到挫折心情不如意時，不時給予我精神上的鼓勵與支持。此外，我也要感謝我研究所的同學們以及身邊的朋友，大家偶爾會在研究室為彼此加油打氣，一起互相討論與幫忙，當中也特別感謝洪瑞襄、林威澤，時常與我討論和解決問題。因為有你們大家，我的研究生涯過的既充實又精采，讓我有個難忘的研究所生活。

最後，感謝所有關心過我與幫助過我的人，讓我能夠順利地完成碩士學位，非常的感謝你們！

doi:10.6342/NTU201602253

中 文 摘 要

過去十年間，全基因組關聯研究 (genome-wide association studies, GWAS) 已指出數千個與複雜疾病有關的單一核苷酸多型性 (single-nucleotide polymorphisms, SNPs)。隨著次世代定序 (next-generation sequencing, NGS) 技術的進步，遺傳學家們得以在人類的染色體上觀察到更多遺傳訊息，使得尋找次要對偶基因頻率 (minor allele frequency, MAF) 小於 1% 的罕見致病變異 (rare causal variants) 逐漸成為可能。

為了從眾多的變異中指出罕見致病變異，統計方法已陸續發展出來，例如：向後刪除法 (backward elimination procedure, 簡稱「BE」) 和 *P* 值適性結合法 (adaptive combination of *P*-values method, 簡稱「ADA」)，已有文獻指出 ADA 方法辨認變異的訊號雜訊比 (signal-to-noise ratio) 高於 BE 方法。本文提出「*Q* 值適性結合法」 (adaptive combination of *Q*-values, 簡稱「ADAQ」) 以進一步來提高發現致病變異的機率。在變異有同義/非同義註解 (synonymous / non-synonymous annotations) 的情況下，吾人首先將全部的變異分為同義變異群與非同義變異群，再使用 Benjamini-Hochberg 法分別將兩組內的 *P* 值轉換成 *Q* 值(簡稱 B-H *Q*-values)，繼而移除 *Q* 值較大者，因其較有可能真為中立變異 (neutral variants)。

經由模擬發現，ADAQ 的陽性預測值 (positive predictive value) 較 ADA 更高。此外，吾人亦將 ADAQ 應用到遺傳分析工作坊 17 (GAW 17) 的資料上，發現 ADAQ 較 ADA 更能有效地控制偽陽性 (false positives) 的個數且產生較高的陽性預測值。因此，當所研究的變異有同義/非同義註解時，吾人推薦使用 ADAQ 來指出個別罕見致病變異。

關鍵字：中立變異；罕見變異；致病變異；非同義變異；次世代定序。

doi:10.6342/NTU201602253

英 文 摘 要

In the past decade, genome-wide association analyses have identified thousands of single-nucleotide polymorphisms (SNPs) associated with complex diseases. With the improvement of next-generation sequencing technology, geneticists have observed more inherited information on human chromosome. Searching for rare causal variants (minor allele frequency < 1%) gradually becomes possible. In order to pinpoint rare causal variants in a large number of variants, statistical approaches such as the BE (backward elimination) procedure and the ADA method (the adaptive combination of *P*-values method), have been developed. It has been shown that the signal-to-noise ratio of variants identified by ADA is larger than that of variants identified by BE. In this study, we propose an ADAQ method ('adaptive combination of *Q*-values method') to further increase the probability that a finding is genuine. With synonymous / non-synonymous annotations for variants, we first allocate all variants into a non-synonymous group and a synonymous group, and transform two groups of per-site *P*-values into Benjamini-Hochberg *Q*-values, respectively. We then remove the variants with larger *Q*-values that are more likely to be neutral. Comprehensive simulations have shown that ADAQ has an even larger positive predictive value than ADA. Moreover, we applied ADAQ to the Genetic Analysis Workshop 17 (GAW 17) data sets. It controls the number of false positives more effectively and generates a larger positive predictive value than ADA. Therefore, we recommend using ADAQ to pinpoint individual rare causal variants, when synonymous / non-synonymous annotations for variants are available.

Keywords: Neutral variants; rare variants; causal variants; non-synonymous variants; next-generation sequencing.

doi:10.6342/NTU201602253

目　錄

doi:10.6342/NTU201602253

圖　目　錄

doi:10.6342/NTU201602253

表　目　錄

doi:10.6342/NTU201602253

第一章 前言

隨著生物科技技術的進步，例如：次世代定序 (Next-generation sequencing, NGS)，基因定序的技術比過去更加有效率，定序成本也大為降低。由於此技術的發展，使得去氧核醣核酸 (DNA) 序列能一一解碼。在許多研究中發現，一些複雜疾病 (complex diseases) 與 DNA 序列中鹼基對 (base pair, bp) 的變異有關聯 [Cirulli and Goldstein 2010; Shi, et al. 2009; Sullivan, et al. 2012]，與複雜疾病有關的變異可區分為兩大類，常見變異 (common variant) 與罕見變異 (rare variant)。

過去在全基因組關聯性研究 (genome-wide association studies, GWAS) 中，多以「常見變異」為主要的研究標的，此常被定義為次要對偶基因頻率 (minor allele frequency, MAF) 大於 5% 的變異。與複雜疾病有關聯的常見變異也陸續被找到，如：影響阿茲海默症的 APOE 對偶基因 [Bertram and Tanzi 2009]。此外，「罕見變異」亦可能與複雜疾病有關，例如：研究人員發現 GRM3 基因裡的罕見變異與精神分裂症、雙向情感障礙以及酒精依賴症有密切相關 [Schizophrenia Working Group of the Psychiatric Genomics 2014]，並發表在《精神遺傳病學》(Psychiatric Genetics) 雜誌上。罕見變異通常定義為次要對偶基因頻率小於 1% 的變異。有鑑於罕見變異於複雜疾病中扮演著重要角色 [Cirulli and Goldstein 2010]，本文將發展尋找罕見致病變異 (rare causal variants) 的統計分析方法。

於尋找罕見變異時，若將每個變異位點 (locus) 逐一作關聯檢定，其統計檢定力非常低。研究者為了克服這樣的難題，多數的統計方法合併一基因 (或一染色體區域) 裡的多個變異資訊，來檢定該基因 (或該染色體區域) 是否與疾病有關，此方法可提高檢定力 [Han and Pan 2010; Lee, et al. 2012; Li and Leal 2008; Madsen and Browning 2009; Morgenthaler and Thilly 2007; Morris and Zeggini 2010; Price, et al. 2010; Wu, et al. 2011]。

將一區段內多個罕見變異資訊合併的方法可粗略分為兩大類，第一大類為「burden-based methods」，此法是先求出每個人於該區段內攜帶的罕見變異個數加權和，再檢定此加權和與性狀值 (traits) 之間的相關 [Han and Pan 2010; Li and Leal 2008; Madsen and Browning 2009; Morgenthaler and Thilly 2007]。第二大類為「kernel-based methods」，此法是先為每個變異位點求出一個分數統計量 (score

doi:10.6342/NTU201602253

statistic)，再將各位點上的分數統計量平方進行加權和 [Wu, et al. 2011]。此兩種方法各有其優勢，為了將此兩種方法的優點融合在一起，遂有學者以線性組合的方式產生新的檢定統計量 [Derkach, et al. 2013; Lee, et al. 2012]。

然而，將一區段內多個變異聚集時，無可避免地也同時將中立變異 (neutral variant, 實與疾病無關的變異) 的雜訊 (noise) 納入，「burden-based methods」與「kernel-based methods」的統計檢定力會因納入雜訊而降低。Yang 與 Chen [Yang and Chen 2011] 根據費雪的 *P* 值 (*P*-value) 組合法 [Fisher 1932]，與截斷較大 *P* 值的想法 [Zaykin, et al. 2002]，將 *P* 值大於 0.05 的變異予以剔除。然 0.05 對低檢力的罕見變異而言恐過於嚴苛，Lin 等人 [Lin 2014b; Lin, et al. 2014] 遂提出「*P* 值適性結合法」 (adaptive combination of *P*-values, 簡稱為「ADA」)，利用排列法 (permutation) 的方式，求出最佳 *P* 值截斷門檻，剔除 *P* 值高於該截斷門檻的變異 (因其較可能真為中立變異)，以此來提高統計檢定力。ADA 亦被發展於處理家族資料上 [Lin 2014a; Lin and Liang 2016]。

上述的罕見變異關聯檢定多以基因或染色體區域為單位，一旦檢定出基因與疾病有關，下一步應當是在該基因中指出個別的罕見致病變異 [Lin 2016]。Ionita-Laza 等人 [Ionita-Laza, et al. 2014] 提出了向後刪除法 (backward elimination procedure, 簡稱為「BE」)，可逐步篩選出基因內貢獻程度高的罕見變異，若貢獻程度的衡量是以「burden-based methods」為準，則稱「BE-BURDEN」法；若貢獻程度的衡量是以「kernel-based methods」為準，則稱「BE-SKAT」法 (SKAT 全稱為「sequence kernel association test」)。另外，Ionita-Laza 等人也提出一階層模型 (hierarchical model, 簡稱為「HM」) 的方法，來估計個別變異對疾病的效應，但 HM 方法僅將變異依重要性排序，並未訂出截斷門檻值，亦即無法將一群變異區分為致病變異或中立變異。最近，Lin 提出以 ADA 法來篩出個別的罕見致病變異，在多種的模擬情境之下，均一致地發現 ADA 的陽性預測值 (positive predictive value, PPV) 較 BE 方法高得多，PPV 比較結果為：ADA > BE-SKAT > BE-BURDEN [Lin 2016]。

雖現有的 ADA 法已有不錯的尋找個別罕見致病變異之能力，但若能進一步將變異的功能註解 (functional annotation) 資訊納入分析中 [Byrnes, et al. 2013]，將會

doi:10.6342/NTU201602253

更有效地指出罕見致病變異以及排除中立變異。另外，在過去的研究中，Li 等人與 Lin 和 Lee [Li, et al. 2008; Lin and Lee 2012] 提出一「優先子集合分析」(Prioritized Subset Analysis, PSA) 方法，當對變異有先驗知識 (prior knowledge) 時，可將之歸於「優先子集合」(prioritized subset)，表示其較可能與疾病有關；其餘無先驗知識者則歸於「剩餘子集合」。再將此兩組各自使用 Benjamini 和 Hochberg 之「錯誤發現率」(false discovery rate) [Benjamini and Hochberg 1995] 控制將 *P* 值轉成 *Q* 值，以進行多重檢定校正。過去文獻發現 PSA 較不分群的方法更具統計檢定力 [Li, et al. 2008; Lin and Lee 2012]。其原理為，子集合內的基因變異可經「借力使力」(borrow the strength of others) 的 *P* 轉 *Q* 過程，使得「優先子集合」內的變異彼此受益，*Q* 值往 0 跑 (趨顯著)；而「剩餘子集合」的 *Q* 值則往 1 跑 (趨不顯著)。只要分群是稍有資訊的，PSA 就較具統計檢定力 [Li, et al. 2008; Lin and Lee 2012]。

為了找出真正的罕見致病變異，且在有變異的資訊和分群的概念之優勢情況下，有了此研究動機，進而本文提出一「*Q* 值適性結合法」(adaptive combination of *Q*-values)，簡稱「ADAQ」。由於對每個罕見變異位點逐一分析檢定力將很低，吾人採取兩階段方式分析，第一階段為在眾多的基因中檢定出疾病關聯基因 (將於第二章 2.1 節介紹)；第二階段為使用 ADAQ 法於疾病關聯基因內繼而指出個別的罕見致病變異 (將於第三章介紹)。此法適用於二元型態性狀或連續型態性狀，以變異的功能註解來分群，以提高找到真正罕見致病變異的機率。

在二元型態性狀與連續型態性狀下，吾人比較 ADA 方法與本文提出的 ADAQ 方法的真陽性 (true positives, TP) 個數、偽陽性 (false positives, FP) 個數和陽性預測值。此外，吾人亦將此法運用於遺傳分析工作坊 17 (GAW 17) 的資料上 [Almasy, et al. 2011]，以評估各方法之表現。

doi:10.6342/NTU201602253

第二章 文獻回顧

本章回顧既有的文獻與方法，分為兩小節，2.1 節介紹以基因為基礎 (gene-based) 的關聯檢定，2.2 節介紹如何由一基因內的眾多變異來指出個別的罕見致病變異。

2.1 以基因為分析單元的關聯檢定

探討一群罕見變異與疾病的關聯性時，最常使用的方法有兩種，一種為 BURDEN 檢定 (此為第一類的「burden-based methods」)，另一種為 SKAT 檢定 (此為第二類的「kernel-based methods」中較受矚目的方法)。此小節將介紹這兩種分析方法。

為研究一基因內 (或一染色體區域) 多個變異與疾病的關聯性，學者們利用迴歸模式來進行分析。若有 n 個個體，考慮一基因內有 m 個變異位點，令 $\boldsymbol{X}$ 為變異上「次要對偶基因」 (minor allele) 的拷貝數矩陣（維度$n \times m$），考慮廣義線性模式如下：

$$g[E(Y_i)] = \alpha_0 + \mathbf{C_i}\boldsymbol{\alpha} + \mathbf{X_i}\boldsymbol{\beta}, \quad (1)$$

其中$g(\cdot)$為連結函數 (link function)，Y_i為第 i 個體的性狀值，α_0為截距項，$\mathbf{C_i} = (C_{i1}, C_{i2}, \cdots, C_{ip})$為第 i 個體的 p 個共變項 (如：年齡，性別等調整項)，$\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)'$為對應的迴歸係數值，$\mathbf{X_i} = (X_{i1}, X_{i2}, \cdots, X_{im})$為第 i 個體於 m 個變異位點上的次要對偶基因拷貝數 (可能值：0, 1, 2)，$\boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)'$為相對應的迴歸係數值。在虛無假說下，這 m 個變異與疾病皆無關聯，$H_0: \boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)' = \mathbf{0}$。

在 BURDEN 檢定的作法之下，設定 $\beta_1, \cdots, \beta_m$ 皆相同，令其皆為 $w_j\beta_0$，其中 w_j 為第 j 個變異對疾病的影響程度之權重 (如：常用每個變異的 MAF 作為權重)，式 (1) 遂可化簡為 $g[E(Y_i)] = \alpha_0 + \mathbf{C_i}\boldsymbol{\alpha} + \beta_0 \sum_{j=1}^{m} w_j X_{ij}$，原虛無假說 $H_0: \boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)' = \mathbf{0}$，現成為 $H_0: \beta_0 = 0$。所以，BURDEN 檢定是先把每個人於該基因內帶有多少次要對偶基因進行加總後，再檢定此新變項 $(\sum_{j=1}^{m} w_j X_{ij})$ 與性狀值之間的關聯性。

而由 Wu 等人所提出的 SKAT 檢定 [Wu, et al. 2011] 則是假設 $\beta_1, \beta_2, \ldots, \beta_m$ 皆

doi:10.6342/NTU201602253

為有分布的隨機變數，期望值皆為 0，而變異數皆設為 $w_j^2\tau$，其中 w_j 為給第 j 個變異之權重，τ為變異數成份 (variance component)。在虛無假說下，這 m 個變異與疾病皆無關聯， H_0: $\boldsymbol{\beta} = \mathbf{0}$，遂可簡化成為 H_0: $\tau = 0$。由於實際情況是未知的，無法事先知道採取哪一種檢定方法較為恰當，故 Lee 等人 [Lee, et al. 2012] 綜合此兩種方法來保有其各自的優點，進而提出一個整合的統計量：

$$Q_\rho = (\mathbf{Y} - \widehat{\boldsymbol{\mu}_\mathbf{0}})'\mathbf{K}_\rho(\mathbf{Y} - \widehat{\boldsymbol{\mu}_\mathbf{0}}), \tag{2}$$

其中 $\mathbf{K}_\rho = \mathbf{XWR}_\rho\mathbf{WX}'$，而 $\boldsymbol{R}_\rho = (1-\rho)\mathbf{I} + \rho 11'$ 為一個可交換的相關係數 (exchangeable correlation) 矩陣，對角線皆為 1，非對角線皆為ρ，$\mathbf{W} = diag(w_1, \cdots, w_m)$ 為一個對角線為權重的對角矩陣。若 $\mathbf{Y}$ 為連續型態性狀，$\widehat{\boldsymbol{\mu}_\mathbf{0}}$為虛無假說下 $\mathbf{Y}$ 的估計值；若 $\mathbf{Y}$ 為二元型態性狀，$\widehat{\boldsymbol{\mu}_\mathbf{0}}$為虛無假說下 $\mathbf{Y}$=1 機率的估計值。當$\rho = 1$時，式 (2) 可寫為

$$Q_{\rho=1} = \left[\sum_{j=1}^{m} w_j \sum_{i=1}^{n} (Y_i - \widehat{\mu_{i,0}})X_{ij}\right]^2, \tag{3}$$

此即為 BURDEN 檢定的統計量。於虛無假說下，式 (3) 服從自由度 1 的卡方分布 (chi-square distribution)。由原本自由度 m 降為 1，以提高檢定力 [Lee, et al. 2012; Li and Leal 2008; Lin, et al. 2011; Madsen and Browning 2009; Morris and Zeggini 2010]。當$\rho = 0$時，式 (2) 可寫為

$$Q_{\rho=0} = \sum_{j=1}^{m} w_j^2 \left[\sum_{i=1}^{n} (Y_i - \widehat{\mu_{i,0}})X_{ij}\right]^2, \tag{4}$$

此即為 SKAT 檢定的統計量 [Wu, et al. 2011]。於虛無假說下，式 (4) 服從一混合卡方分布 (mixture of chi-square distributions)，可由 Davies 所提出的方法來計算 P 值 [Davies 1980]。

文獻上對於 BURDEN 檢定與 SKAT 檢定有諸多比較和討論 [Basu and Pan 2011]，當基因內變異方向一致時 (如：全為有害變異或全為保護變異)，BURDEN 的檢定力較SKAT高；當變異方向不全一致時 (有害與保護變異並存於一基因內)，SKAT 的檢定力較 BURDEN 高，因在 BURDEN 檢定統計量裡有害變異與保護變

doi:10.6342/NTU201602253

異效應會正負相消之故。另，若致病變異在基因內所占的比例較小時，SKAT 檢定會比 BURDEN 檢定好 [Basu and Pan 2011]。

2.2 指出個別罕見致病變異的分析方法

於 2.1 節找出與疾病有關的基因（內含多個變異）後，下一步應於該基因內找出與疾病有關的個別變異。本節將介紹兩種尋找罕見致病變異的分析方法，第一種方法為向後刪除法 (backward elimination procedure, 簡稱為「BE」) [Ionita-Laza, et al. 2014]，以 BURDEN 或 SKAT 統計量作為衡量標準，分別稱為「BE-BURDEN」和 「BE-SKAT」。第二種方法為 「P 值適性結合法」 (adaptive combination of P-values, 簡稱為「ADA」) [Lin 2016; Lin, et al. 2014]。

由 Ionita-Laza 等人提出的 BE 方法，可由一基因內的眾多變異區分出致病變異。步驟如下：

步驟一：先於基因內隨機抽出 r 個變異 (例如，$r = 20$)，構成集合 $V_c = \{v_1, \cdots, v_r\}$，使用 BURDEN 或 SKAT 計算這群變異的 P 值，簡稱為 P_{V_c}。

步驟二：依序移除 r 個變異中每一個變異，例如，移除第 j 個變異時，剩下的變異所構成的集合為 $V_{-j} = \{v_1, \cdots, v_{j-1}, v_{j+1}, \cdots, v_r\}$，再使用 BURDEN 或 SKAT 來計算這群變異的 P 值，簡稱為 $P_{V_{-j}}$。

步驟三：若 $min(P_{V_{-1}}, \cdots, P_{V_{-r}}) \leq P_{V_c}$，則移除 $V_c = \{v_1, \cdots, v_r\}$ 中第 k 個變異，其中 $k = argmin(P_{V_{-1}}, \cdots, P_{V_{-r}})$，而剩下的變異構成新的集合 $V_c = \{v_1, \cdots, v_{k-1}, v_{k+1}, \cdots, v_r\}$。

步驟四：重複步驟二和三，直到 BURDEN 或 SKAT 的 P 值無法再更小為止，將最後剩餘在集合 V_c 裡的變異傳回記錄檔。

一基因可能內含數百個變異，若這為數眾多的變異同步進行向後刪除法，計算量將會很大，所以上述步驟一才需以每次 20 個變異的方式進行之。重抽 B 次進行如上四個步驟之後(例如：$B = 1000$)，綜合 B 次傳回的記錄檔，計算每個變異的總傳回次數，再經由無母數期望值最大化 (Expectation-Maximization, EM) 方法 [Benaglia, et al. 2009] 將這些變異的總傳回次數區分為高和低兩群，進而將變異區

doi:10.6342/NTU201602253

分成「感興趣」和「不感興趣」兩群，「感興趣」這群即列為可能的致病變異。

第二種為 Lin 等人所提出的 ADA 方法，可於一基因內的眾多變異排除帶雜訊的中立變異，來尋找出致病變異。此法是將基因內的 m 個變異分別求出個別的 P 值，令為 $p_1, p_2, \ldots, p_m$，基於費雪的 P 值合併法 [Fisher 1932]，將 m 個變異的 P 值合併為 $-2\sum_{j=1}^{m} logp_j$ (此會讓致病變異的貢獻大於中立變異)，同時考量給罕見變異的權重 [Cheung, et al. 2012]，比較在病例組與對照組的頻率將變異區分為「傾向有害變異」與「傾向保護變異」，再將兩群變異的 P 值分別作整合，同時欲去除 P 值較大的變異 (因其有較大可能真為中立變異)。但 ADA 在結合變異的 P 值之前，會針對 P 值給定一截斷點，且 ADA 不若 Yang 等人使用一個固定的 P 值截斷點 0.05 [Yang and Chen 2011]，而是考慮 11 個 P 值截斷點：0.10, 0.11, 0.12, …, 0.19, 0.20。傾向有害變異與傾向保護變異其 P 值分別整合如下：

(1) 在第 f 個 P 值截斷點 (令為 θ_f) 之下，傾向有害的變異其 P 值整合為：

$$S_f^+ = -\sum_{j=1}^{m} \xi_j \cdot I\left[p_j < \theta_f\right] \cdot w_j \log p_j \ , \tag{5}$$

其中 ξ_j 為指標變數，若第 j 個變異為傾向有害的變異時，ξ_j 等於 1，否則為 0；$I[p_j < \theta_f]$ 為指標函數，若 $p_j < \theta_f$ 成立則 $I[p_j < \theta_f]$ 等於 1，w_j 為給第 j 個變異之權重，延續 SKAT 檢定 [Wu, et al. 2011] 裡設的 $Beta(MAF_j;1,25)$，此為參數 1 與 25 的貝塔分布機率密度函數，而 MAF_j 為第 j 個變異的次要對偶基因頻率。

(2) 在第 f 個 P 值截斷點 (令為 θ_f) 之下，傾向保護的變異其 P 值整合為：

$$S_f^- = -\sum_{j=1}^{m} \phi_j \cdot I\left[p_j < \theta_f\right] \cdot w_j \log p_j \ , \tag{6}$$

其中 ϕ_j 為指標變數，若第 j 個變異為傾向保護的變異時，ϕ_j 等於 1，否則為 0。其餘符號定義如式 (5)。

依式 (5) 與式 (6) 分別求出 S_f^+ 與 S_f^- 後，在無基因整體效應是傾向有害/保護的背景知識下，取其中較大值得到在第 f 個 P 值截斷點下的檢定統計量 $S_f = max(S_f^+, S_f^-)$，$f = 1, 2, \ldots, 11$。繼而進行 B 次排列 (permutation)，每次排列均隨機

doi:10.6342/NTU201602253

重排個案的性狀值 (traits)，創造出在虛無假說下基因-性狀無關聯的狀態。接著，比較在原始樣本 (未經重排的樣本) 與 *B* 個重排樣本於 11 個 *P* 值截斷點下訊號最高者，以此來檢定該基因與性狀間是否相關。值得一提的是，雖吾人給予原始樣本 11 次機會來尋找訊號最高者，但並不需作多重檢定校正，因重排樣本亦皆有 11 次機會來尋找訊號最高者。

而使得原始樣本達訊號最高的該 *P* 值截斷點，稱為「最適 *P* 值截斷點」(the optimal *P*-value truncation threshold)。所有變異的個別 *P* 值小於「最適 *P* 值截斷點」者即判斷為「可能的致病變異」(possible causal variants)。

doi:10.6342/NTU201602253

第三章 材料與方法

本章節將介紹在不同型態性狀的資料下，如何於疾病關聯基因中指出個別罕見致病變異。若資料為二元型態性狀，吾人採用費雪精確檢定 [Fisher 1922]；若資料為連續型態性狀，則吾人採用線性模式，以此來描述性狀與基因內變異之關聯。

情況(A)：二元型態性狀

設共有 n 個個體，m 個變異位點，第 j 個變異的基因型可能為 GG, Gg, gg，假如 G 為主要對偶基因 (major allele)，g 為次要對偶基因 (minor allele)，則第 j 個變異之次要對偶基因的拷貝數，可能值為 0, 1, 2。由於性狀為二元型態，吾人使用費雪精確檢定對 m 個變異逐一進行分析。

對第 j 個變異而言，若病例組中次要對偶基因的頻率大於對照組中次要對偶基因的頻率，則稱此變異為「傾向有害變異」，令 X 為一隨機變數，代表病例組中次要對偶基因總個數，a^{cs} 為病例組中觀察到的次要對偶基因總個數，a 為病例組與對照組中次要對偶基因的總個數，t^{cs} 為病例組中觀察到的對偶基因總個數，而病例組和對照組中對偶基因總個數為 t。以下表示之：

	病例組	對照組	兩組總和
次要對偶基因總個數	a^{cs}	$a - a^{cs}$	a
主要對偶基因總個數	$t^{cs} - a^{cs}$	$t - a - (t^{cs} - a^{cs})$	$t - a$
觀察到的對偶基因總個數	t^{cs}	$t - t^{cs}$	t

在虛無假說下，此變異與疾病無相關，單尾檢定的中間 P 值 (mid-P-value) 為：

$$\frac{1}{2}P(X = a^{cs}) + P(X > a^{cs}) = \frac{1}{2}\frac{\binom{a}{a^{cs}}\binom{t-a}{t^{cs}-a^{cs}}}{\binom{t}{t^{cs}}} + \sum_{x=a^{cs}+1}^{a}\frac{\binom{a}{x}\binom{t-a}{t^{cs}-x}}{\binom{t}{t^{cs}}}，(7)$$

doi:10.6342/NTU201602253

採取中間 P 值的原因是，若包含 a^{cs} 此點機率時，P 值會偏大而不易拒絕虛無假說；若不含 a^{cs} 此點機率時，P 值會偏小而容易拒絕虛無假說。故採取折衷的中間 P 值。

類似地，若對第 j 個變異而言，對照組中次要對偶基因的頻率大於病例組中次要對偶基因的頻率，則稱此變異為「傾向保護變異」，令 X 為一隨機變數，代表對照組中次要對偶基因總個數，a^{cn} 為對照組中觀察到的次要對偶基因總個數，a 為病例組與對照組中次要對偶基因的總個數，t^{cn} 為對照組中觀察到的對偶基因總個數，而病例組和對照組中對偶基因總個數為 t。在虛無假說下，此變異與疾病無相關，單尾檢定的中間 P 值 (mid-P-value) 為：

$$\frac{1}{2}P(X=a^{cn})+P(X>a^{cn})=\frac{1}{2}\frac{\binom{a}{a^{cn}}\binom{t-a}{t^{cn}-a^{cn}}}{\binom{t}{t^{cn}}}+\sum_{x=a^{cn}+1}^{a}\frac{\binom{a}{x}\binom{t-a}{t^{cn}-x}}{\binom{t}{t^{cn}}} \quad (8)$$

以式 (7) 或式 (8) 來計算出個別變異的 P 值之後，吾人將這 m 個變異區分為「非同義變異」 (non-synonymous variants) 與「同義變異」 (synonymous variants) 兩群。所謂的「非同義變異」為鹼基 A、T、C、G 的改變，會造成胺基酸序列的改變，而「同義變異」為鹼基 A、T、C、G 的改變，不致造成胺基酸序列的改變 [Ramensky, et al. 2002]。再使用 Benjamini-Hochberg 法 [Benjamini and Hochberg 1995] 將兩組內的 P 值分別轉換成 Q 值 (亦稱「B-H Q-values」)，若非同義變異有 m_{ns} 個，轉換過程可分為三個步驟：

步驟一：先將 P 值由小排到大 $P_{(1)} \le P_{(2)} \le \cdots \le P_{(m_{ns})}$。

步驟二：最大 P 值所對應的 Q 值即為自身，$\hat{Q}\{P_{(m_{ns})}\}=P_{(m_{ns})}$。

步驟三：依序轉換 P 值為 Q 值，$\hat{Q}\{P_{(i)}\}=\min[P_{(i)}\times {}^{m_{ns}}/_{i}\ ,\ \hat{Q}\{P_{(i+1)}\}]$，$i=m_{ns}-1,\ m_{ns}-2,\ ...,\ 1$。

由於 $\hat{Q}\{P_{(i)}\}$ 不會大於 $\hat{Q}\{P_{(i+1)}\}$，故同組裡的變異 Q 值會受到同儕的影響，此即「借力使力」的來由。同樣地，若同義變異有 $(m-m_{ns})$ 個，依以上三個步驟將 P 值轉換成 Q 值，再將這兩群 Q 值合為一群。繼而移除 Q 值較大者，因其較有可能真為中立變異，吾人考慮 F 個 Q 值截斷門檻 $\theta_1,\theta_2,\cdots,\theta_F$。經比較在病例

doi:10.6342/NTU201602253

組與對照組的頻率，將變異區分為「傾向有害變異」與「傾向保護變異」後，再將兩種變異的Q值依式 (5) 與式 (6) 分別作整合 (式 (5) 與式 (6) 裡的P值需替換為 Q 值)。

依式 (5) 與式 (6) 分別求出 S_f^+ 與 S_f^- 後，在無基因整體效應是傾向有害/保護的背景知識下，取其中較大值得到在第f個 Q 值截斷點下的檢定統計量 $S_f = max(S_f^+, S_f^-)$，$f = 1, 2, \dots, F$。繼而進行 B 次排列，每次排列均隨機重排個案的性狀值，創造出在虛無假說下基因-性狀的無關聯狀態。在第f個 Q 值截斷點下，觀察到的樣本其檢定統計量 S_f 與 B 次排列樣本的檢定統計量 $S_f^{(1)}, S_f^{(2)}, \dots, S_f^{(B)}$ 比較後，可估出 S_f 相對應的 P 值為 $\frac{\sum_{b=1}^{B} I\left(S_f^{(b)} \geq S_f\right)+1}{B+1}$，其中 $I(\cdot)$ 為指標函數，可能值為 0 或 1。而第 k 次排列樣本的檢定統計量 $S_f^{(k)}$ 與其它 $(B-1)$ 次排列樣本的檢定統計量比較後，可估計出 $S_f^{(k)}$ 相對應的 P 值為 $\frac{\sum_{b \neq k} I\left(S_f^{(b)} \geq S_f^{(k)}\right)+1}{B}$，$k = 1, 2, \dots, B$。

令觀察到的樣本在 F 個 Q 值截斷門檻中最小的 P 值為 $MinP = \min_{1 \leq f \leq F} \frac{\sum_{b=1}^{B} I\left(S_f^{(b)} \geq S_f\right)+1}{B+1}$，第 k 次排列樣本在 F 個 Q 值截斷門檻中最小的 P 值為 $MinP^{(k)} = \min_{1 \leq f \leq F} \frac{\sum_{b \neq k} I\left(S_f^{(b)} \geq S_f^{(k)}\right)+1}{B}$。比較 $MinP$ 與 $MinP^{(k)}$, $k = 1, 2, \dots, B$ ，吾人可得「調整後 P 值」(adjusted P-value) 為 $\frac{\sum_{k=1}^{B} I\left(MinP^{(k)} \leq MinP\right)+1}{B+1}$ 。此涵義為在 F 個 Q 值截斷門檻中尋找訊號最高者，雖吾人給予原始樣本 F 次機會來尋找訊號最高者，但並不需作多重檢定校正，因重排樣本亦皆有相同的 F 次機會。

而使得原始樣本達訊號最高的該 Q 值截斷點，稱為「最適 Q 值截斷點」(the optimal Q-value truncation threshold)。所有變異的個別 Q 值小於「最適 Q 值截斷點」者即判斷為「可能的致病變異」。

為減少計算上所耗費的時間，吾人採用 Besag 和 Clifford 提出的「序列蒙地卡羅法」(sequential Monte Carlo method) [Besag and Clifford 1991]，令最少排列次數 B_{min} 為 10 次，最大排列次數 B_{max} 為 1,000 次，停止條件為

$$\sum_{b=1}^{B} I(MinP^{(b)} \leq MinP) \geq (c^2 + B^{-1})^{-1} \text{ ，或 } B = B_{max}$$

doi:10.6342/NTU201602253

，其中常數 c 設為 0.25，表示「調整後 P 值」的標準誤大約為其值的 0.25 倍，若此「調整後 P 值」的估計要愈精確，則需設定更小的 c 值，所需排列次數將會增加。

情況(B)：連續型態性狀

設共有 n 個個體，若欲分析的性狀為連續型態，如；血壓、膽固醇值，當分析第 j 個變異時，吾人考慮的線性模式如下：

$$\mathrm{E}(Y_i) = \alpha_0 + \boldsymbol{C_i}\boldsymbol{\alpha} + X_{ji}\beta_j\,, \tag{9}$$

其中 Y_i 為第 i 個體的性狀值，α_0 為截距項，$\boldsymbol{C_i} = (C_{i1}, C_{i2}, \cdots, C_{ip})$ 為第 i 個體的 p 個共變項 (如：年齡，性別等調整項)，$\boldsymbol{\alpha} = \left(\alpha_1, \cdots, \alpha_p\right)'$ 為對應的迴歸係數值，X_{ji} 為第 i 個體於第 j 個變異位點上的次要對偶基因拷貝數 (可能值：0, 1, 2)，β_j 為相對應的迴歸係數值。在虛無假說下，第 j 個變異與疾病無關，H_0: $\beta_j = 0$。華德檢定統計量 (Wald statistic) $T_j = \frac{\widehat{\beta_j}}{SE(\widehat{\beta_j})}$，服從自由度 $n-(p+2)$ 的司徒頓 t 分布 (Student's t distribution)，其中 $\widehat{\beta_j}$ 為第 j 個變異的迴歸係數估計值，可由最小平方法或最大概似法求得，$SE(\widehat{\beta_j})$ 為 $\widehat{\beta_j}$ 的標準誤。

求出個別變異的 P 值之後，吾人將這 m 個變異區分為「非同義變異」與「同義變異」兩群。如同 (A) 小節所述使用 Benjamini-Hochberg 法 [Benjamini and Hochberg 1995] 將兩組內的 P 值分別轉換成 Q 值 (亦稱「B-H Q-values」)。若第 j 個變異的迴歸係數值大於 0，則歸類為「傾向提升性狀變異」；若小於 0，則歸類為「傾向降低性狀變異」。將兩種變異的 Q 值依式 (5) 與式 (6) 分別作整合 (式 (5) 與式 (6) 裡的 P 值需替換為 Q 值)。後續分析如同 (A) 小節所述。

於情況(A)或情況(B)之方法，吾人稱本法為「Q 值適性結合法」(adaptive combination of Q-values)，簡稱「ADAQ」。

doi:10.6342/NTU201602253

第四章 模擬

本章藉由電腦模擬，探討在不同型態性狀的資料下，使用第三章所提出的分析方法，可否能有效地排除中立變異，找出真正的罕見致病變異。4.1 節將先介紹模擬時的參數設定，4.2 節比較各方法之不同處，最後於 4.3 節呈現模擬之結果。

4.1 模擬設計

情況(A)：二元型態性狀

於模擬中，吾人使用的染色體序列資料是依據溯祖過程 (coalescent process) [Hudson 2002] 且採用 Schaffner 等人之 Cosi 程式 [Schaffner, et al. 2005] 來產生資料，此 Cosi 程式已普遍地被使用來模擬人類基因組序列 [Byrnes, et al. 2013; Lin, et al. 2013; Lin, et al. 2012; Peng 2015]。吾人使用 Cosi 生成 1,000 個序列資料集，每個資料集生成 10,000 條仿歐裔連鎖不平衡狀態且染色體區段長度 (region size) 為 20,000 個鹼基對 (20 kilo base pairs, 20kb) 的單套型 (haplotype) 資料。由於本文主要探討的是罕見變異，因此排除 MAF 大於 5%的常見變異，使用 MAF 小於 5%的變異納入分析中。

為衡量找到真正致病變異的能力，吾人於 MAF 介於 0.1% 與 1% 的變異中隨機選擇若干作為影響疾病的變異，使得有害/保護變異占總分析變異的比例 (causal percentage) 為 7.5% 或 15%，其餘即設為中立變異。

關於有害/保護變異之配置，令有害變異之比例為 r_{isk}%，r_{isk} 設置分別為 5、20、50、80 及 100，其餘 $(100-r_{isk})$% 則為保護變異之比例。分別假設每個有害/保護變異的族群可歸因危險性 (population attributable risk, PAR) 為 0.3%或 0.5%，表示罹患此疾病之比例有 0.3%或 0.5%可歸因於此有害/保護變異，族群可歸因危險性定義為在族群中疾病的發生率 (incidence) 可歸因於某一暴露的比例。

由族群可歸因危險性的定義，吾人可推得第 h 個有害/保護變異的基因型相對危險性 (genotype relative risk, GRR) 為：

$$GRR_h = \left(\frac{PAR_h}{(1-PAR_h)\cdot MAF_h}+1\right)^{(-1)^{\xi_h}}, \qquad (10)$$

其中 PAR_h 和 MAF_h 分別為第 h 個有害/保護變異的族群可歸因危險性及其次要

doi:10.6342/NTU201602253

對偶基因頻率；當第 h 個變異為保護變異時， ξ_h 等於 1；為有害變異時， ξ_h 等於0。式 (10) 也已被許多文獻所採用 [Li, et al. 2010; Lin, et al. 2013; Lin, et al. 2012; Madsen and Browning 2009]。

為生成病例或對照個體，吾人從 Cosi 軟體套件生成的歐裔群體 10,000 條單套型中，隨機選取兩條單套型 H_1, H_2 以形成一個體的基因型。個體疾病狀態沿用 Li 等人 [Li, et al. 2010; Lin, et al. 2013; Lin, et al. 2012] 之設定，罹患疾病的機率為：

$$P(affected|H_1, H_2) = f_0 \times \prod_{r=1}^{2}\prod_{h=1}^{d} GRR_h^{I(H_{r,h}=g_h)} \ , \qquad (11)$$

其中 f_0 為基準外顯率 (baseline penetrance)，吾人依照 Cheung 等人設定為 1% [Cheung, et al. 2012]，$H_{r,h}$ 為第 r 條 (r = 1, 2) 單套型上第 h 個有害/保護變異的對偶基因，g_h 為第 h 個有害/保護變異的次要對偶基因，d 為總有害/保護變異的個數，$I(\cdot)$ 為指標函數。重複收集個體直至收滿 500 個病例與 500 個對照為止。非同義/同義變異之個數配置請參見表一，此番設計沿用已發表文獻 [Byrnes, et al. 2013; Ionita-Laza, et al. 2014]，設定在有害/保護變異中，非同義變異所占的比例為 0.8 和 0.9，由於非同義變異會造成胺基酸序列的改變，其影響疾病的可能性較高；另設定在中立變異中，非同義變異所占的比例為 0.33、0.47 和 0.56。

決定 Q 值截斷門檻

因 Q 值是由 P 值作多重檢定校正而來，會比 P 值還要大，故 Q 值截斷門檻應比 P 值截斷門檻 (0.10, 0.11, 0.12, …, 0.20 [Lin 2016]) 來得寬鬆些。自 0.10 起，吾人先考慮最大 Q 值截斷門檻 (θ_{max}) 為 0.15、0.20、0.25、0.30 或 0.35，當θ_{max}=0.15 時，則有 6 個 Q 值截斷門檻：θ_1=0.10, θ_2=0.11, …, θ_6=0.15。經由本文所提出的 ADAQ 方法測試後，圖一至圖四為 1,000 次重複之下，5 種 θ_{max} 之接受者作業特徵曲線 [Receiver operator characteristic (ROC) curve] 之結果。由圖一至圖四可知當有害/保護變異占總分析變異的比例 (causal percentage)、族群可歸因危險性、r_{isk}% 愈大時，ROC 曲線愈高於 45°線，意謂表現愈好。

評估哪一個最大 Q 值截斷門檻最好，吾人考慮以下兩準則：

doi:10.6342/NTU201602253

(Ⅰ) 敏感度 (sensitivity) 加特異度 (specificity) 值愈大愈好；

(Ⅱ) ROC 曲線上的點 (θ_{max}) 距離左上角 [座標點(0,1)] 愈短愈好。

表二列出準則 (Ⅰ) 下的結果，多數情形是以最大Q值截斷門檻為0.3時表現最好；表三列出準則 (Ⅱ) 下的結果，最大 Q 值截斷門檻為 0.35 時表現最好。最後，為抉擇 0.3 與 0.35 哪一個較好，吾人模擬 100 次重複，比較平均真陽性個數、偽陽性個數和陽性預測值。從圖五可看出，最大 Q 值截斷門檻為 0.35 時，真陽性的個數雖較 Q 值截斷門檻 0.3 時多，但偽陽性的個數卻也高得多，致使陽性預測值較低，故吾人最後選用的最大 Q 值截斷門檻為 0.3。

情況(B)：連續型態性狀

染色體序列資料來源同情況(A)，考慮的染色體區段長度亦為 20kb，吾人於 MAF 介於 0.1% 與 1% 的變異中隨機選擇若干作為真正影響性狀的變異，使其占總分析變異的比例 (causal percentage) 為 7.5% 或 15%，其餘即設為中立變異。關於增加/降低性狀變異之配置，令增加性狀變異之比例為 r_{isk}%，r_{isk} 設置分別為 5、20、50、80 及 100，其餘 $(100-r_{isk})$% 則為降低性狀變異之比例。

為生成連續型態性狀，吾人沿用 Wu 等人 [Wu, et al. 2011] 文章裡的模擬設定來產生連續型態性狀：

$$y = 0.5C_1 + 0.5C_2 + \beta_1 X_1^c + \beta_2 X_2^c + \cdots + \beta_m X_m^c + \varepsilon, \qquad (12)$$

其中 C_1 模擬一連續型的共變項(如：年齡、體重等調整項)，令其服從標準常態分布，C_2 為 0 或 1 且機率各為 0.5 之二元型態共變項(如：性別等調整項)，ε為誤差項服從標準常態分布，β_h 為 $c|log_{10}MAF_h| \times (-1)^{\xi_h}$，若第 h 個變異會降低性狀值，則指標函數 ξ_h 等於 1；若會增加性狀值，ξ_h 等於 0，c 為效應值的乘數且設置為 0.2 或 0.4 (其中 0.4 為 [Wu, et al. 2011] 文章裡的模擬設定)。假如第 h 個變異會增加性狀值且若其 MAF 為 0.005，當 c 設置為 0.2 時，β_h約為 0.4602。$X_1^c, X_2^c, \cdots, X_m^c$為第 1 個到第 m 個變異上的拷貝數。模擬樣本數為 1,000 人。

doi:10.6342/NTU201602253

4.2 方法比較

情況(A)：二元型態性狀

使用本文所提出的 ADAQ 方法與 ADA 方法 [Lin 2016] 比較，其中 ADAQ 方法依表一對非同義/同義變異個數之六種配置，依序為 ADAQ-1、ADAQ-2、ADAQ-3、ADAQ-4、ADAQ-5、ADAQ-6。觀察在不同情況下(染色體區段長度為 20kb，有害/保護變異占總分析變異的比例為 7.5%或 15%，族群可歸因危險性為 0.3% 或 0.5%，有害變異占有害/保護變異之比例為 5%、20%、50%、80%或 100%)，顯著水準設為 0.01，比較七種方法的檢定力 (power)、真陽性的個數，偽陽性的個數和陽性預測值。顯著水準設為 0.01 的原因是，假設為研究五個基因的 Bonferroni 校正，在研究更多基因時，顯著水準應設得更小，此時，以排列法 (permutation) 求 *P* 值的 ADA 與 ADAQ 需較多時間，故此時吾人提倡二階段分析方法，第一階段為在眾多的基因中檢定出疾病關聯基因 (如第二章 2.1 節介紹的 BURDEN 檢定或 SKAT 檢定，皆提供解析 *P* 值，analytical *P*-value)；第二階段再使用 ADAQ 法於疾病關聯基因內指出個別的罕見致病變異 (於第三章介紹)。

ADA 方法是由 Lin 網址：http://homepage.ntu.edu.tw/~linwy/ADAprioritized.html 所提供之 R 程式以進行分析。本文 ADAQ 方法與 ADA 方法主要不同的地方在於，對每個變異逐一進行分析後，ADAQ 納入變異的資訊將分析後所得到的 *P* 值分成非同義變異 *P* 值與同義變異 *P* 值兩群，而 ADA 並未將 *P* 值作分群。

在方法設定上，ADA 方法所得的最終 *P* 值是經由 1,000 次排列而得，而 ADAQ 方法為了降低計算時間，採用 Besag 和 Clifford 所提出的「序列蒙地卡羅法」[Besag and Clifford 1991]，設定最小排列次數 10 次，最大 1,000 次。另外，ADA 方法設定 11 個 *P* 值截斷門檻為 θ_1=0.10, θ_2=0.11,…, θ_{11}=0.20；而 ADAQ 方法設定 21 個 *Q* 值截斷門檻為 θ_1=0.10, θ_2=0.11,…, θ_{21}=0.30。

情況(B)：連續型態性狀

同情況(A)，使用本文提出的 ADAQ-1、ADAQ-2、ADAQ-3、ADAQ-4，ADAQ-5 和 ADAQ-6 方法與 ADA 方法比較。觀察在不同情況下(染色體區段長度為 20kb，增加/降低性狀變異占總分析變異的比例為 7.5%或 15%，效應值乘數為 0.2 或 0.4，

doi:10.6342/NTU201602253

增加性狀變異占增加/降低性狀變異之比例為 5%、20%、50%、80%或 100%)，顯著水準設為 0.01，比較七種方法的檢定力、真陽性的個數，偽陽性的個數和陽性預測值。

doi:10.6342/NTU201602253

4.3 模擬結果

為瞭解 P 值轉換成 Q 值後的變化，吾人重複 100 次模擬，每次模擬約有數個至 14 個為真正的致病變異，以及一百多個真正的中立變異，分別觀察其 P 值至 Q 值的轉變。

情況(A)：二元型態性狀

圖六和圖七為有害/保護變異 P 值(上列)至 Q 值(下列)的變化，而圖八和圖九為中立變異 P 值至 Q 值的變化。ADA 法的 P 值截斷門檻最高為 0.2 [Lin 2016]；而本文的 ADAQ 法 Q 值截斷門檻最高為 0.3。吾人可發現 Q 值小於 0.3 的有害/保護變異雖變少 (少於 P 值小於 0.2 的有害/保護變異)，但 Q 值小於 0.3 的中立變異可變得更少 (少於 P 值小於 0.2 的中立變異)。這表示 ADAQ 法將可剔除更多的偽陽性 (false positives)。

對基因型相對危險性 (GRR) 的設定，吾人採用式 (10)，因其已被許多文獻所採用 [Li, et al. 2010; Lin, et al. 2013; Lin, et al. 2012; Madsen and Browning 2009]。然而，在此設定下，給定相同的族群可歸因危險性 (PAR) 及次要對偶基因頻率 (MAF) 時，保護變異的效應值會比有害變異來得低，證明可見 [Lin, et al. 2014]。無獨有偶，根據 Wang 等人 2015 年於美國人類遺傳學會年會上的發表，其提到罕見變異關聯檢定的模擬常見缺點之一：給保護變異與有害變異相同的效應值，如：[Lee, et al. 2012; Wu, et al. 2011]。事實上，若適當設計情境，保護變異的影響應比有害變異來得低。所以，實際上 BURDEN 檢定優於 SKAT 檢定的情形應更廣泛 [Wang, et al. 2015]。

因保護變異的效應值較低，其 P 值並不會太顯著，即使被歸類在非同義變異群裡，亦難以因「借力使力」 (borrow the information from others) 的特性使其 Q 值低於 0.3。有害變異因效應值較高，P 值會較保護變異顯著些，故轉換為 Q 值後較有機會低於 0.3。

情況(B)：連續型態性狀

圖十和圖十一為增加/降低性狀變異 P 值(上列)至 Q 值(下列)的變化，而圖十

doi:10.6342/NTU201602253

二和圖十三為中立變異 *P* 值至 *Q* 值的變化。吾人可發現中立變異的 *P* 值原約呈現均勻分布 (Uniform distribution)，但經轉換為 *Q* 值後，*Q* 值小於 0.3 的中立變異變得很少。這表示 ADAQ 法將可剔除更多的偽陽性。

當染色體區段中全為中立變異時：

情況(A)：二元型態性狀

吾人先令染色體區段中全為中立變異，以創造在虛無假說下的情形。重複收集個體直至收滿 500 個病例與 500 個對照。關於非同義/同義變異個數之配置，非同義變異個數：同義變異個數（**NS:S**）沿用 Ionita-Laza 等人 [Ionita-Laza, et al. 2014] 模擬設定之 0.6：1。吾人自 1,000 個 Cosi 序列資料集中，每個序列資料集皆進行重複模擬 100 次，故共有 10 萬次 (1000×100) 重複，訂定顯著水準為 0.01，吾人計算於這 10 萬次重複中「調整後 *P* 值」小於顯著水準 0.01 的比例，此為型一錯誤率 (type I error rates)。

由表四可知，ADAQ 與 ADA 的型一錯誤率皆與顯著水準 0.01 接近，表示此兩種檢定皆具有正確性 (validity)。就偽陽性的個數而言，ADAQ 約只有 ADA 的一半。

情況(B)：連續型態性狀

吾人令染色體區段不含任何增加/降低性狀變異，以創造在虛無假說下的情形。重複收集個體直至收滿 1,000 個個案。關於非同義/同義變異個數之配置，與情況 (A) 相同。吾人自 1,000 個 Cosi 序列資料集中，每個序列資料集皆進行重複模擬 100 次，故共有 10 萬次 (1000×100) 重複，訂定顯著水準為 0.01，吾人計算於這 10 萬次重複中「調整後 *P* 值」小於顯著水準 0.01 的比例，此為型一錯誤率。

由表四可知，ADAQ 與 ADA 的型一錯誤率皆與顯著水準 0.01 接近，表示此兩種方法皆具有正確性。就偽陽性的個數而言，ADAQ 不到 ADA 的三分之一。

doi:10.6342/NTU201602253

當染色體區段中<u>不全為</u>中立變異時：

情況(A)：二元型態性狀

吾人再令染色體區段中含有若干致病變異，以創造在對立假說下的情形。重複收集個體直至收滿 500 個病例與 500 個對照。非同義/同義變異個數之配置由表一所示，此番設計沿用已發表文獻 [Byrnes, et al. 2013; Ionita-Laza, et al. 2014]。吾人自 1,000 個 Cosi 序列資料集中，於每種模擬情境之下，每個序列資料集皆進行重複模擬 2 次，故共有 2,000 次 (1000×2) 重複，訂定顯著水準為 0.01，吾人計算於這 2,000 次重複中「調整後 P 值」小於顯著水準 0.01 的比例，此為檢定力。另外，於「調整後 P 值」小於顯著水準 0.01 的情況下，代表該基因/區域與疾病狀態有顯著相關，進而以 ADAQ 與 ADA 來指出個別的罕見致病變異，吾人列出真陽性個數、偽陽性個數及陽性預測值之中位數，因其較平均數更為穩健。

由圖十四（模擬設定有害/保護變異占總分析變異之 7.5%）與圖十五（模擬設定有害/保護變異占總分析變異之 15%），吾人可看出雖 ADAQ 找出的真陽性個數比 ADA 來得少，但偽陽性個數更少，使得 ADAQ 的陽性預測值高於 ADA (陽性預測值：由 ADAQ 判斷為致病變異者，其真為致病變異的機率)。由表一可知，若以非同義/同義變異來區分成兩群，資訊程度最高的是第 4 種情境 ADAQ-4，因其 $P_{NS|C}$（在有害/保護變異中，非同義變異所占的比例）為較高的 0.9（沿用自 [Byrnes, et al. 2013]）；而 $P_{NS|NC}$ (在中立變異中，非同義變異所占的比例）為較低的 0.33。無怪乎 ADAQ-4 的陽性預測值高於其它配置情境。

情況(B)：連續型態性狀

吾人令染色體區段中含有若干增加/降低性狀變異，以創造在對立假說下的情形。重複收集個體直至收滿 1,000 個個案。非同義/同義變異個數之配置與情況(A)相同。吾人自 1,000 個 Cosi 序列資料集中，於每種模擬情境之下，每個序列資料集皆進行重複模擬 2 次，故共有 2,000 次 (1000×2) 重複，訂定顯著水準為 0.01，吾人計算於這 2,000 次重複中「調整後 P 值」小於顯著水準 0.01 的比例，此為檢定力。同樣地，於「調整後 P 值」小於顯著水準 0.01 的情況下，代表該基因/區域與性狀值有顯著相關，進而以 ADAQ 與 ADA 來指出個別的罕見致病變異，吾人

doi:10.6342/NTU201602253

亦列出真陽性個數、偽陽性個數及陽性預測值之中位數。

由圖十六（模擬設定增加/降低性狀變異占總分析變異之 7.5%）與圖十七（模擬設定增加/降低性狀變異占總分析變異之 15%），吾人可看出雖 ADAQ 找出的真陽性個數比 ADA 來得少，但偽陽性個數更少，使得 ADAQ 的陽性預測值高於 ADA。同樣地，因第 4 種非同義/同義變異配置資訊程度最高，故 ADAQ-4 的陽性預測值高於其它配置情境。

doi:10.6342/NTU201602253

第五章 應用於遺傳分析工作坊 17 資料

吾人運用到遺傳分析工作坊 17 (Genetic Analysis Workshop 17, 簡稱為「GAW17」) 中小規模外顯子組 (mini-exome) 的序列資料 [Almasy, et al. 2011]，此資料模擬自千人基因組計畫 (1000 Genomes Project) [Abecasis, et al. 2010]，提供仿真實情境的基因變異數量及頻率。資料中共有 697 位無血緣關係之個體之年齡、抽菸與否、基因型與性狀值。共有 22 對體染色體，3,205 個基因，共 24,487 個變異，吾人分析兩個連續型態性狀 Q1 與 Q4。此外，GAW17 團隊針對每個人的基因型產生 200 次重複的性狀值，故吾人有 200 組 Q1 與 Q4 可進行方法的衡量。

在 GAW17 的模擬設定中，影響 Q1 的基因主要來自血管內皮生長因子 (vascular endothelial growth factor, 簡稱為「VEGF」) 途徑 (pathway)，實際上有 39 個影響性狀的變異。此外，GAW17 團隊為提供參加者衡量檢定方法之型一錯誤率，亦生成一性狀 Q4，而實際上無任何變異會影響該性狀的數值。吾人採取兩種策略來分析 Q4 與 Q1：

策略一：單一變異分析，就 24,487 個變異逐一分析，當分析第 j 個變異時，吾人考慮的線性模式如式 (9) 所示，共變項為年齡與抽菸狀態，由華德檢定統計量 (Wald statistic) 得到變異的個別 P 值，共有 24,487 個 P 值，再進行邦弗朗尼校正 (Bonferroni correction) 控制「整體錯誤率」(family-wise error rate) 於 0.05 (亦即顯著水準設為 $0.05/24487$) 、或 Benjamini 和 Hochberg 法控制「錯誤發現率」(false discovery rate) 於 0.05 (簡稱「FDR-BH」) [Benjamini and Hochberg 1995]。

策略二：以基因為單位執行 SKAT 或 BURDEN 檢定，並以邦弗朗尼校正 (Bonferroni correction) 控制「整體錯誤率」(family-wise error rate) 於 0.05，亦即顯著水準設為 $0.05/3205$，於顯著基因內復以 ADA 或 ADAQ 尋找影響性狀的變異，ADA 與 ADAQ 需要的個別 P 值亦由式 (9) 而得，共變項為年齡與抽菸狀態，由華德檢定統計量 (Wald statistic) 得到變異的個別 P 值。

分析 Q4 的結果列於表五，吾人可看出單一變異分析結合邦弗朗尼校正是最保守的，綜括 200 次重複而言，平均每次重複僅會出現 0.065 個偽陽性 (亦即全部 200 次重複裡僅發現 13 個偽陽性)；單一變異分析結合 FDR-BH 最不保守，綜括 200

doi:10.6342/NTU201602253

次重複而言，平均每次重複會出現 1.665 個偽陽性 (亦即全部 200 次重複裡發現了 333 個偽陽性)；而策略二介於前述二者之間，以 ADAQ 尋找訊號變異的偽陽性個數低於 ADA 法，此與上一章的模擬結果一致。

分析 Q1 的結果亦列於表五，吾人可看出單一變異分析結合邦弗朗尼校正是最保守的，真陽性個數與偽陽性個數皆最少；單一變異分析結合 FDR-BH 的真陽性個數比策略二少，但偽陽性個數卻遠比策略二多，這表示策略二較單一變異分析結合 FDR-BH 為佳；而策略二有高於策略一的真陽性個數，偽陽性個數則介於邦弗朗尼校正與 FDR-BH 之間。策略二之中，ADAQ 尋找訊號變異的真陽性與偽陽性個數皆低於 ADA 法，此與上一章的模擬結果一致。

值得一提的是，各方法於分析 Q1 時的偽陽性都為數可觀，這結果與許多 GAW17 的分析報告一致 [Lin, et al. 2011; Tintle, et al. 2011]，可能原因是 GAW17 團隊利用血管內皮生長因子途徑來指定影響 Q1 的變異，儘管只指定了 39 個影響 Q1 性狀的變異，然其它未被指定的變異 (被視為是中立變異) 卻與這 39 個變異存在相關性，畢竟，於 24,487 個變異中有近 10,000 個是私房變異 (private variants，僅於某個人或某個家族內才會被觀察到) [Tintle, et al. 2011]。

繼而，吾人將策略二的結果進行詳細解析，於影響 Q1 性狀值的九個基因當中 (答案取自 GAW17 團隊文獻 [Almasy, et al. 2011])，以基因為單位執行 SKAT 或 BURDEN 檢定時，僅 *FLT1* 與 *KDR* 基因的檢定力大於 10%，表六列出此二基因的分析結果 (其餘七個影響 Q1 的基因檢定力過低，難以呈現有意義的比較，故略)。吾人可得與上一章模擬一致的結果，亦即，ADAQ 的真陽性個數略少於 ADA，但偽陽性個數較 ADA 少更多，使得 ADAQ 的陽性預測值高於 ADA。

doi:10.6342/NTU201602253

第六章 結論與討論

於次世代定序資料中指出個別的罕見致病變異著實不易。作單一變異分析時，由於罕見變異觀察到的次數太少而使得檢定力低落，若再以施以後續保守的邦弗朗尼校正，則更難以尋找到真正的致病變異。有鑑於此，研究者可採取兩階段方式來分析。第一階段先以基因為單位的方式作關聯性檢定 (可採常用的 SKAT 或 BURDEN 檢定)，找到與複雜性疾病有關的基因，第二階段再將顯著的基因施以 ADA 或 ADAQ 分析，來指出個別罕見致病變異。本文於分析 GAW17 資料時採取此兩階段的方式，使用 BURDEN 檢定或 SKAT 檢定先將與複雜疾病有關聯的基因找出，再使用 ADAQ 來指出個別的罕見致病變異。吾人發現此兩階段的策略優於單一變異分析結合「錯誤發現率」控制 (文中簡稱為「FDR-BH」) [Benjamini and Hochberg 1995]。

本文提出的 ADAQ 方法只與 ADA 方法作比較，原因為最近的研究發現 [Lin 2016]，ADA 方法比 Ionita-Laza 等人提出的向後刪除法 [Ionita-Laza, et al. 2014] 好，ADA 有較高的陽性預測值與較短的計算時間，因此，本文只與 ADA 作比較。

於指出個別的罕見致病變異上，ADAQ 僅利用非同義/同義變異資訊，即可比 ADA 更有效地控制偽陽性的個數，並擁有比 ADA 更高的陽性預測值，亦即由 ADAQ 找出來的變異有較高的機率真為致病變異，此為研究者關心的重要指標。因此，當非同義/同義變異註解可得時，吾人推薦使用 ADAQ 來指出個別的罕見致病變異。

doi:10.6342/NTU201602253

有害/保護變異占總分析變異的比例 (causal percentage)	情境簡寫	NS:S [1]	$P_{NS\|C}$ [2]	$P_{NS\|NC}$ [3]
較低的 (~7.5%)	**ADAQ-1**	0.57	0.8	0.33
	ADAQ-2	0.97	0.8	0.47
	ADAQ-3	1.36	0.8	0.56
	ADAQ-4	0.58	0.9	0.33
	ADAQ-5	1.00	0.9	0.47
	ADAQ-6	1.40	0.9	0.56
較高的 (~15%)	**ADAQ-1**	0.66	0.8	0.33
	ADAQ-2	1.08	0.8	0.47
	ADAQ-3	1.47	0.8	0.56
	ADAQ-4	0.70	0.9	0.33
	ADAQ-5	1.14	0.9	0.47
	ADAQ-6	1.56	0.9	0.56

表一：模擬研究中，非同義/同義變異個數之配置

[1] **NS:S** 為（非同義變異個數：同義變異個數）於所有重複模擬下之平均。

[2] $P_{NS|C}$ 為在有害/保護變異中，非同義變異所占的比例，其中 0.8 沿用自 [Ionita-Laza, et al. 2014] 之模擬情境設定，0.9 沿用自 [Byrnes, et al. 2013] 之模擬情境設定。

[3] $P_{NS|NC}$ 為在中立變異中，非同義變異所占的比例，此部分設定沿用自 [Ionita-Laza, et al. 2014] 之模擬情境設定。

doi:10.6342/NTU201602253

(Ⅰ) 敏感度+特異度 (愈大愈好)						
最大Q值截斷門檻 (θmax)		0.15	0.2	0.25	0.3	0.35
	有害變異占有害/保護變異之比例 (r_{isk}%)					
染色體區段長度=20kb 有害/保護變異占總分析變異的比例=7.5% 族群可歸因危險性=0.3%	5%	1.0110	1.0161	1.0230	1.0357	1.0265
	20%	1.0174	1.0213	1.0295	1.0438	1.0306
	50%	1.0372	1.0467	1.0586	1.0708	1.0487
	80%	1.1026	1.1159	1.1331	1.1589	1.1434
	100%	1.1418	1.1589	1.1806	1.2009	1.1908
染色體區段長度=20kb 有害/保護變異占總分析變異的比例=7.5% 族群可歸因危險性=0.5%	5%	1.0298	1.0358	1.0480	1.0765	1.0938
	20%	1.0455	1.0552	1.0723	1.1008	1.1225
	50%	1.1060	1.1247	1.1463	1.1735	1.1776
	80%	1.2203	1.2461	1.2724	1.2914	1.2869
	100%	1.2754	1.3033	1.3350	1.3619	1.3636
染色體區段長度=20kb 有害/保護變異占總分析變異的比例=15% 族群可歸因危險性=0.3%	5%	1.0205	1.0294	1.0404	1.0534	1.0160
	20%	1.0243	1.0320	1.0408	1.0604	1.0108
	50%	1.0593	1.0736	1.0904	1.1089	1.0695
	80%	1.1469	1.1629	1.1859	1.2010	1.1810
	100%	1.1911	1.2112	1.2322	1.2493	1.2448
染色體區段長度=20kb 有害/保護變異占總分析變異的比例=15% 族群可歸因危險性=0.5%	5%	1.0355	1.0479	1.0682	1.1003	1.1011
	20%	1.0518	1.0693	1.0912	1.1170	1.1177
	50%	1.1333	1.1598	1.1875	1.2080	1.1974
	80%	1.2566	1.2820	1.3082	1.3297	1.3430
	100%	1.3055	1.3292	1.3560	1.3754	1.3890
染色體區段長度=10kb 有害/保護變異占總分析變異的比例=7.5% 族群可歸因危險性=0.3%	5%	1.0150	1.0216	1.0304	1.0500	1.0311
	20%	1.0252	1.0330	1.0455	1.0695	1.0451
	50%	1.0438	1.0535	1.0674	1.0948	1.0725
	80%	1.0959	1.1205	1.1410	1.1707	1.1452
	100%	1.1229	1.1469	1.1728	1.2028	1.1851
染色體區段長度=10kb 有害/保護變異占總分析變異的比例=7.5% 族群可歸因危險性=0.5%	5%	1.0402	1.0537	1.0694	1.1049	1.1001
	20%	1.0551	1.0711	1.0891	1.1260	1.1321
	50%	1.1256	1.1480	1.1774	1.2096	1.2076
	80%	1.2263	1.2604	1.2931	1.3171	1.3077
	100%	1.2789	1.3178	1.3574	1.3888	1.3798
染色體區段長度=10kb 有害/保護變異占總分析變異的比例=15% 族群可歸因危險性=0.3%	5%	1.0229	1.0330	1.0459	1.0551	1.0184
	20%	1.0242	1.0368	1.0514	1.0625	1.0336
	50%	1.0539	1.0691	1.0877	1.1057	1.0688
	80%	1.1402	1.1623	1.1909	1.2031	1.1863
	100%	1.1805	1.2044	1.2348	1.2540	1.2424
染色體區段長度=10kb 有害/保護變異占總分析變異的比例=15% 族群可歸因危險性=0.5%	5%	1.0394	1.0574	1.0797	1.1111	1.1033
	20%	1.0669	1.0880	1.1085	1.1358	1.1234
	50%	1.1485	1.1788	1.2065	1.2270	1.2108
	80%	1.2717	1.3096	1.3454	1.3625	1.3604
	100%	1.3314	1.3701	1.4035	1.4217	1.4199

表二：於各種模擬情況下，採用準則（Ⅰ）來比較 5 個最大 Q 值截斷門檻 (θ_{max}) 每個數值皆為 1,000 次重複下之平均，淺灰色底表示在 5 個最大 Q 值截斷門檻 (θ_{max}) 當中，敏感度加特異度值達到最大。

doi:10.6342/NTU201602253

(Ⅱ) 離座標(0,1)之距離 (愈小愈好)						
最大Q值截斷門檻（θmax）		0.15	0.2	0.25	0.3	0.35
	有害變異占有害/保護變異之比例（r_{isk}%）					
染色體區段長度=20kb 有害/保護變異占總分析變異的比例=7.5% 族群可歸因危險性=0.3%	5%	0.9864	0.9798	0.9715	0.9507	0.8351
	20%	0.9792	0.9742	0.9644	0.9412	0.8445
	50%	0.9551	0.9446	0.9310	0.9099	0.8303
	80%	0.8795	0.8638	0.8442	0.8114	0.7600
	100%	0.8339	0.8146	0.7905	0.7611	0.7208
染色體區段長度=20kb 有害/保護變異占總分析變異的比例=7.5% 族群可歸因危險性=0.5%	5%	0.9651	0.9571	0.9429	0.9014	0.7901
	20%	0.9487	0.9373	0.9180	0.8794	0.7804
	50%	0.8821	0.8615	0.8375	0.8010	0.7539
	80%	0.7520	0.7236	0.6944	0.6688	0.6487
	100%	0.6908	0.6603	0.6257	0.5929	0.5795
染色體區段長度=20kb 有害/保護變異占總分析變異的比例=15% 族群可歸因危險性=0.3%	5%	0.9745	0.9639	0.9503	0.9216	0.8287
	20%	0.9716	0.9623	0.9518	0.9204	0.8466
	50%	0.9300	0.9135	0.8944	0.8656	0.8306
	80%	0.8244	0.8059	0.7804	0.7574	0.7396
	100%	0.7712	0.7483	0.7246	0.7001	0.6854
染色體區段長度=20kb 有害/保護變異占總分析變異的比例=15% 族群可歸因危險性=0.5%	5%	0.9587	0.9444	0.9214	0.8729	0.7796
	20%	0.9411	0.9210	0.8968	0.8523	0.7927
	50%	0.8486	0.8195	0.7890	0.7562	0.7405
	80%	0.7045	0.6760	0.6465	0.6201	0.5965
	100%	0.6430	0.6160	0.5857	0.5614	0.5419
染色體區段長度=10kb 有害/保護變異占總分析變異的比例=7.5% 族群可歸因危險性=0.3%	5%	0.9817	0.9734	0.9627	0.9196	0.8315
	20%	0.9704	0.9613	0.9463	0.9031	0.8222
	50%	0.9515	0.9401	0.9246	0.8745	0.8096
	80%	0.8916	0.8645	0.8411	0.7964	0.7538
	100%	0.8615	0.8350	0.8061	0.7587	0.7132
染色體區段長度=10kb 有害/保護變異占總分析變異的比例=7.5% 族群可歸因危險性=0.5%	5%	0.9543	0.9387	0.9207	0.8640	0.7863
	20%	0.9387	0.9211	0.9006	0.8442	0.7699
	50%	0.8653	0.8409	0.8084	0.7588	0.7173
	80%	0.7556	0.7187	0.6832	0.6460	0.6266
	100%	0.6997	0.6584	0.6158	0.5735	0.5599
染色體區段長度=10kb 有害/保護變異占總分析變異的比例=15% 族群可歸因危險性=0.3%	5%	0.9719	0.9602	0.9446	0.8998	0.8316
	20%	0.9718	0.9571	0.9396	0.8960	0.8245
	50%	0.9390	0.9216	0.8996	0.8540	0.8148
	80%	0.8415	0.8170	0.7847	0.7546	0.7275
	100%	0.7951	0.7684	0.7339	0.7001	0.6809
染色體區段長度=10kb 有害/保護變異占總分析變異的比例=15% 族群可歸因危險性=0.5%	5%	0.9553	0.9353	0.9097	0.8469	0.7765
	20%	0.9260	0.9026	0.8792	0.8225	0.7776
	50%	0.8386	0.8055	0.7739	0.7337	0.7222
	80%	0.7013	0.6605	0.6206	0.5926	0.5830
	100%	0.6355	0.5932	0.5557	0.5294	0.5224

表三：於各種模擬情況下，採用準則 (Ⅱ) 來比較 5 個最大 Q 值截斷門檻 (θ_{max}) 每個數值皆為 1,000 次重複下之平均，淺灰色底表示在 5 個最大 Q 值截斷門檻 (θ_{max}) 當中，與座標點 (0,1) 的距離最短，深灰色底與座標點 (0,1) 的距離次短。

doi:10.6342/NTU201602253

	ADA	ADAQ
二元型態性狀		
型一錯誤率 [a]	0.010	0.009
偽陽性個數平均 [b]	23.0	12.0
連續型態性狀		
型一錯誤率 [a]	0.009	0.008
偽陽性個數平均 [b]	40.0	11.0

表四：虛無假說下(染色體區段內無任何致病變異)，ADA 方法與 ADAQ 方法之型一錯誤率及偽陽性個數平均

[a] 顯著水準訂為 0.01。

[b] 染色體區段長度為 20kb 時，平均於一次模擬中出現的偽陽性個數。

doi:10.6342/NTU201602253

	單一變異分析		以基因為單位執行SKAT或BURDEN檢定後，於顯著基因內以ADA或ADAQ尋找致病變異			
	Bonferroni	**FDR-BH**	**SKAT**		**BURDEN**	
			ADA	**ADAQ**	**ADA**	**ADAQ**
分析 Q4 偽陽性的個數 [1]	0.065	1.665	0.475	0.470	0.605	0.570
分析 Q1 真陽性的個數 [2]	4.3	8.1	9.1	8.3	11.1	10.3
分析 Q1 偽陽性的個數 [1]	64.2	435.6	244.4	221.6	265.6	246.9

表五：GAW17 資料分析：Q4 (實際上無任何影響性狀的變異) 與 Q1 (實際上有 39 個影響性狀的變異)

[1] 於每次重複下平均出現的偽陽性個數。

[2] 於每次重複下平均出現的真陽性個數。

	SKAT					**BURDEN**				
基因	**Power[1]**		**#(TP)[2]**	**#(FP)[3]**	**PPV [4]**	**Power[1]**		**#(TP)[2]**	**#(FP)[3]**	**PPV [4]**
FLT1	1	**ADA**	4.885	5.665	0.473	0.995	**ADA**	4.884	5.673	0.473
		ADAQ	4.3	2.545	0.667		**ADAQ**	4.302	2.543	0.668
KDR	0.59	**ADA**	6.390	1.619	0.807	0.95	**ADA**	6.379	1.532	0.816
		ADAQ	6.178	1.042	0.869		**ADAQ**	6.147	0.989	0.874

表六：GAW17 資料分析：Q1

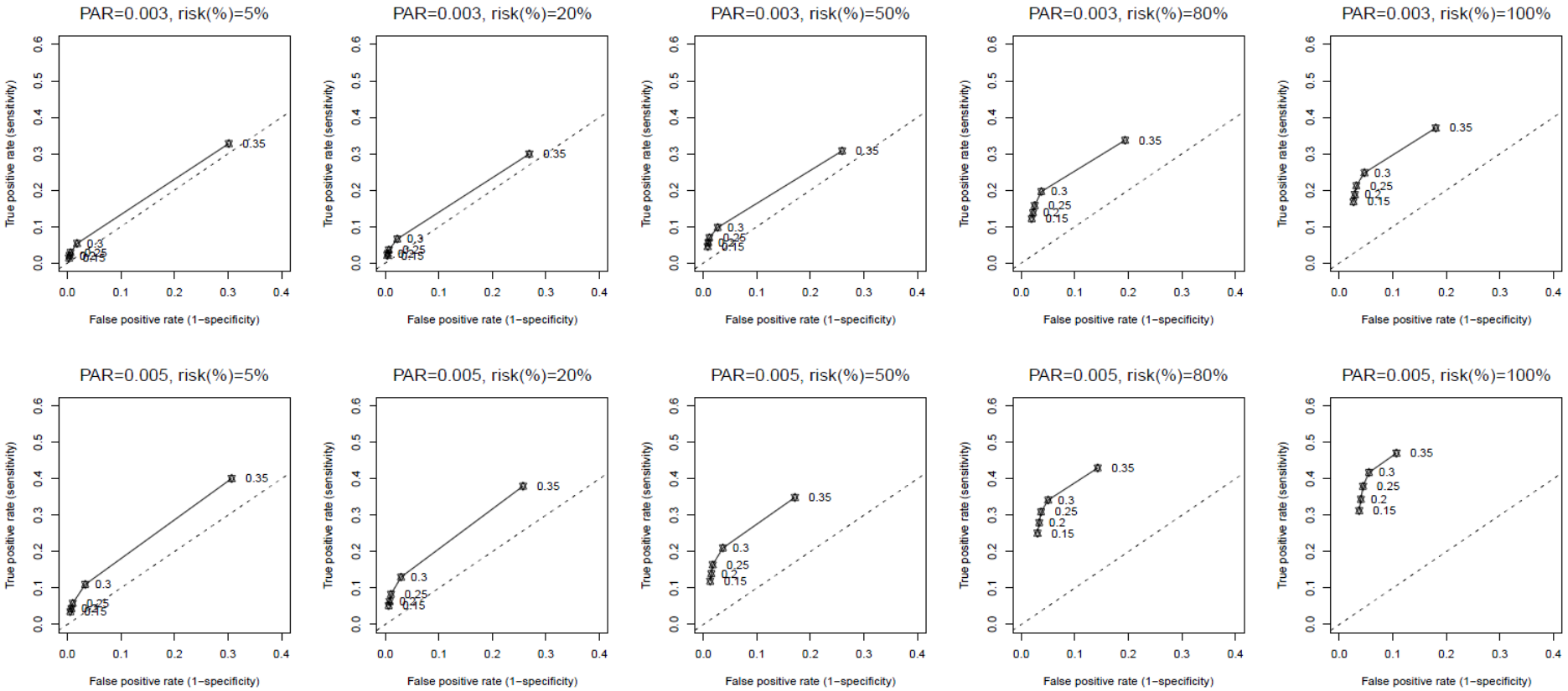
[1] SKAT/BURDEN 方法檢測基因 *FLT1* 和 *KDR* 所得的檢定力。

[2] 在該基因被 SKAT/BURDEN 判定為顯著的情形下 (顯著水準設為$0.05/3205$)，平均可於基因內找出的真陽性個數。根據 GAW17 模擬設定，*FLT1* 基因內含 11 個影響性狀的變異；*KDR* 基因內含 10 個影響性狀的變異。

[3] 在該基因被 SKAT/BURDEN 判定為顯著的情形下 (顯著水準設為$0.05/3205$)，平均於基因內找出的偽陽性個數。

[4] 在該基因被 SKAT/BURDEN 判定為顯著的情形下 (顯著水準設為$0.05/3205$)，平均於基因內獲得的陽性預測值。

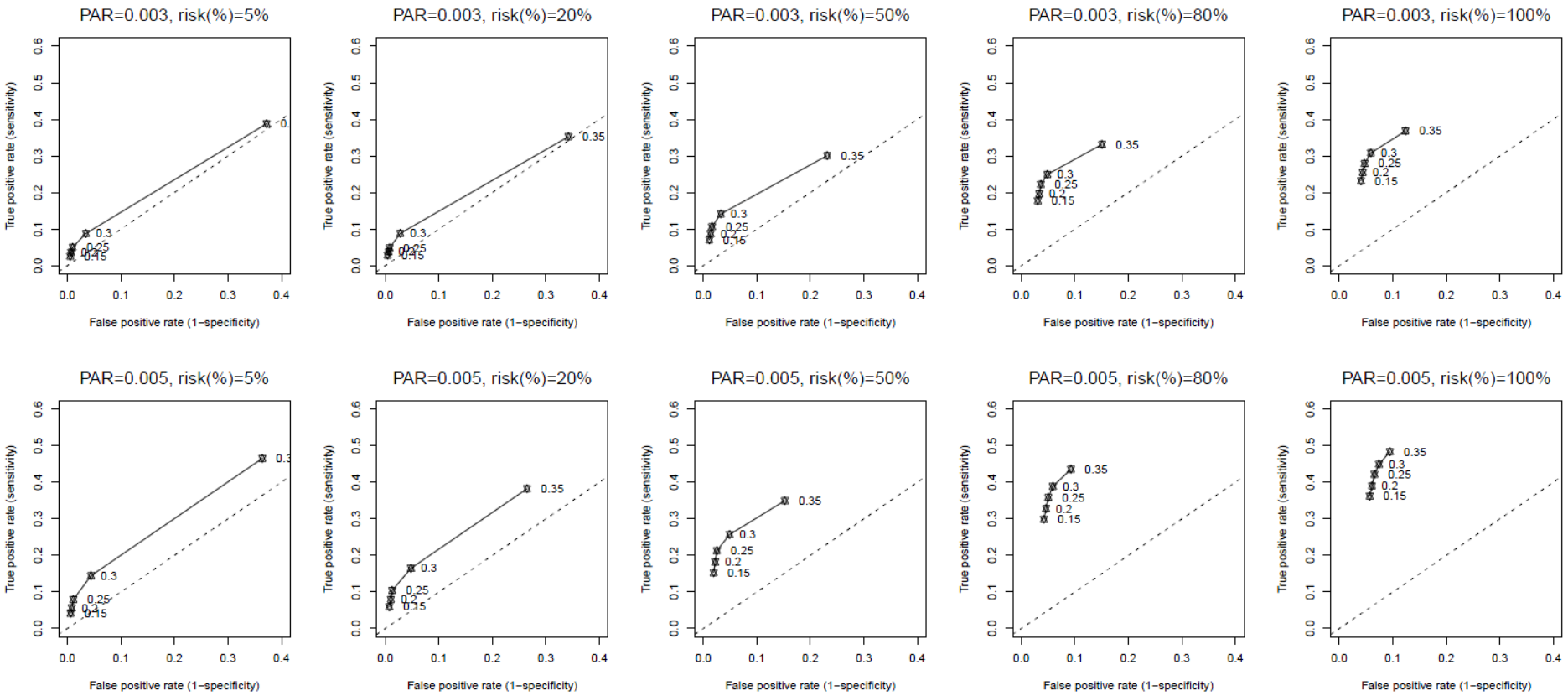
doi:10.6342/NTU201602253



圖一：接受者作業特徵曲線

染色體區段長度為 20kb，有害/保護變異占總分析變異的比例為 7.5%，上列與下列之族群可歸因危險性分別設為 0.3%與 0.5%，由左至右 $r_{isk}\%$ 分別設為 5%, 20%, 50%, 80%, 與 100%。x 軸為偽陽率(1 −特異度)，y 軸為真陽率(敏感度)，每個點標示出 1,000 次重複下偽陽率平均及真陽率平均，點旁的數值為最大 Q 值截斷門檻 (θ_{max})。

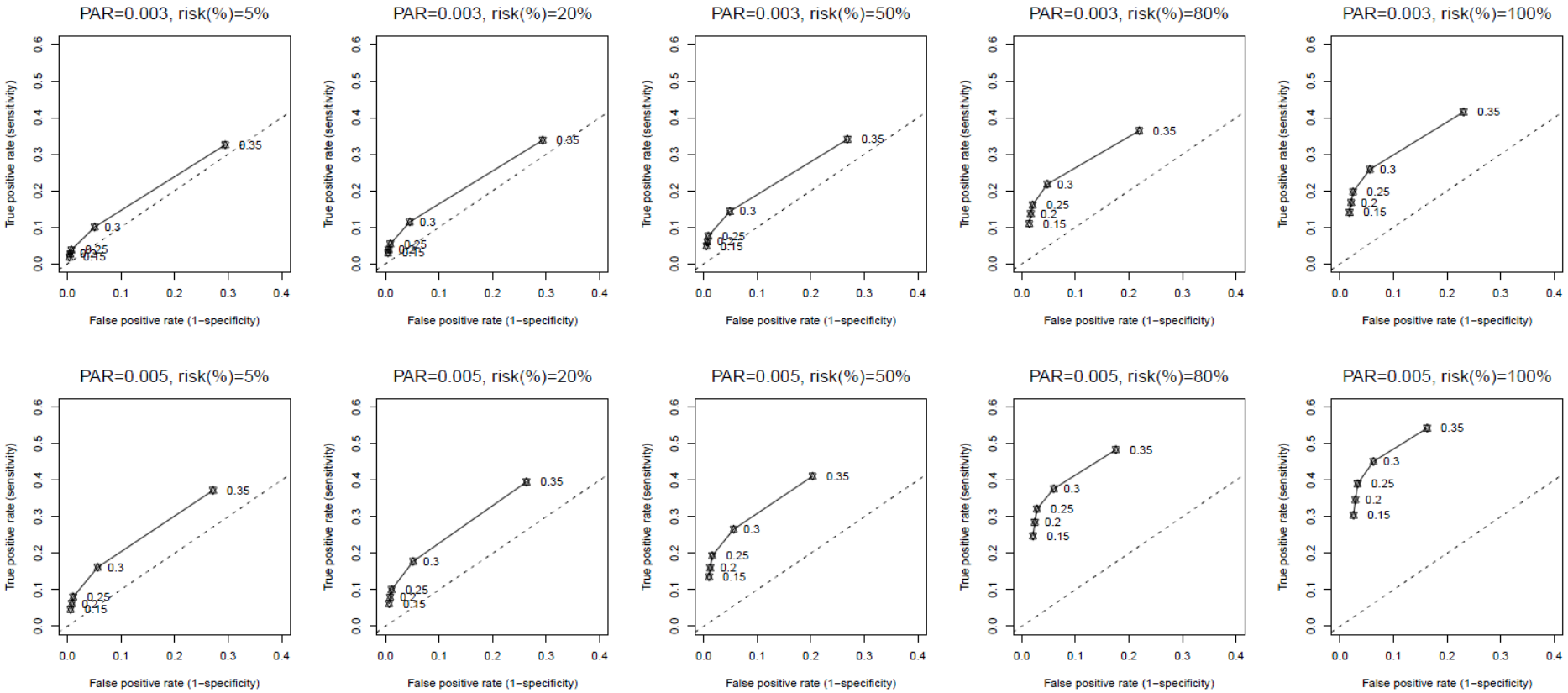
doi:10.6342/NTU201602253



圖二：接受者作業特徵曲線

染色體區段長度為 20kb，有害/保護變異占總分析變異的比例為 15%，上列與下列之族群可歸因危險性分別設為 0.3%與 0.5%，由左至右 $r_{isk}\%$ 分別設為 5%, 20%, 50%, 80%, 與 100%。x 軸為偽陽率(1 −特異度)，y 軸為真陽率(敏感度)，每個點標示出 1,000 次重複下偽陽率平均及真陽率平均，點旁的數值為最大 Q 值截斷門檻 (θ_{max})。

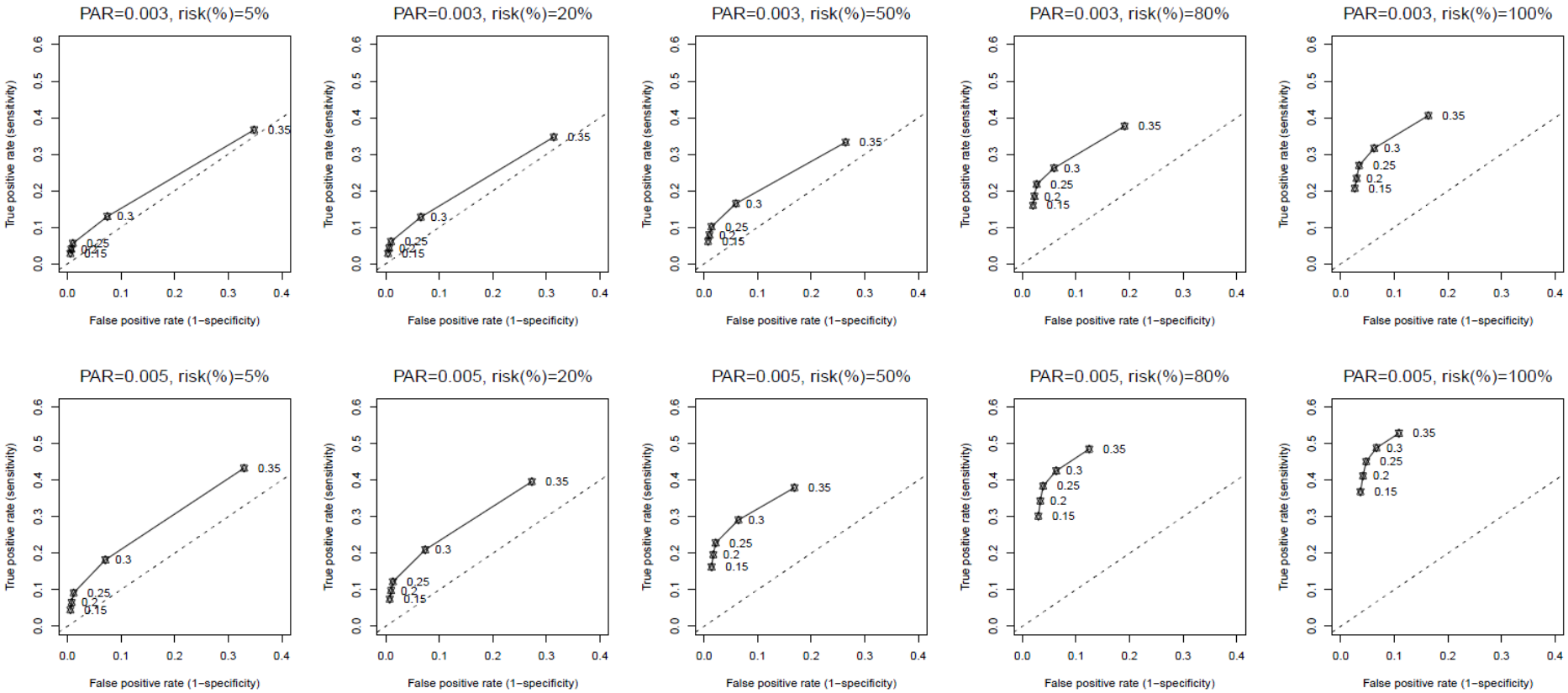
doi:10.6342/NTU201602253



圖三：接受者作業特徵曲線

染色體區段長度為 10kb，有害/保護變異占總分析變異的比例為 7.5%，上列與下列之族群可歸因危險性分別設為 0.3%與 0.5%，由左至右 r_{isk}% 分別設為 5%, 20%, 50%, 80%, 與 100%。x 軸為偽陽率(1 −特異度)，y 軸為真陽率(敏感度)，每個點標示出 1,000 次重複下偽陽率平均及真陽率平均，點旁的數值為最大 Q 值截斷門檻 (θ_{max})。

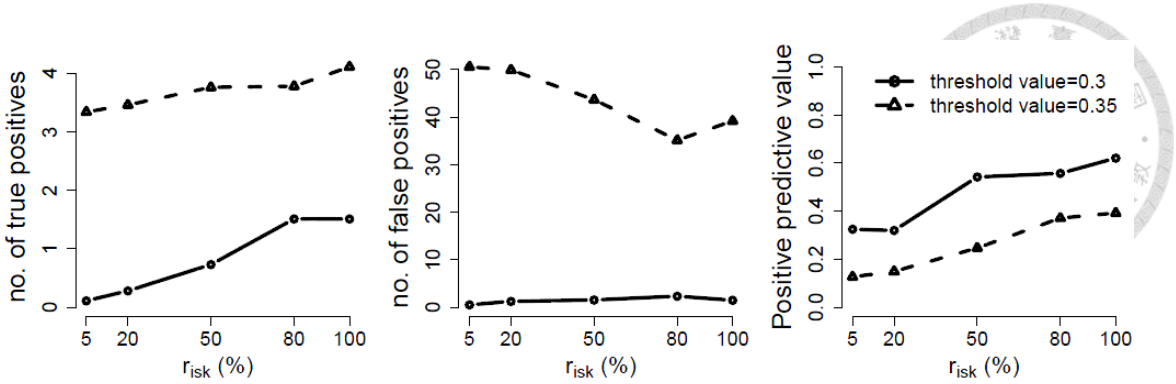
doi:10.6342/NTU201602253



圖四：接受者作業特徵曲線

染色體區段長度為 10kb，有害/保護變異占總分析變異的比例為 15%，上列與下列之族群可歸因危險性分別設為 0.3%與 0.5%，由左至右 $r_{isk}\%$ 分別設為 5%, 20%, 50%, 80%, 與 100%。x 軸為偽陽率(1 −特異度)，y 軸為真陽率(敏感度)，每個點標示出 1,000 次重複下偽陽率平均及真陽率平均，點旁的數值為最大 Q 值截斷門檻 (θ_{max})。

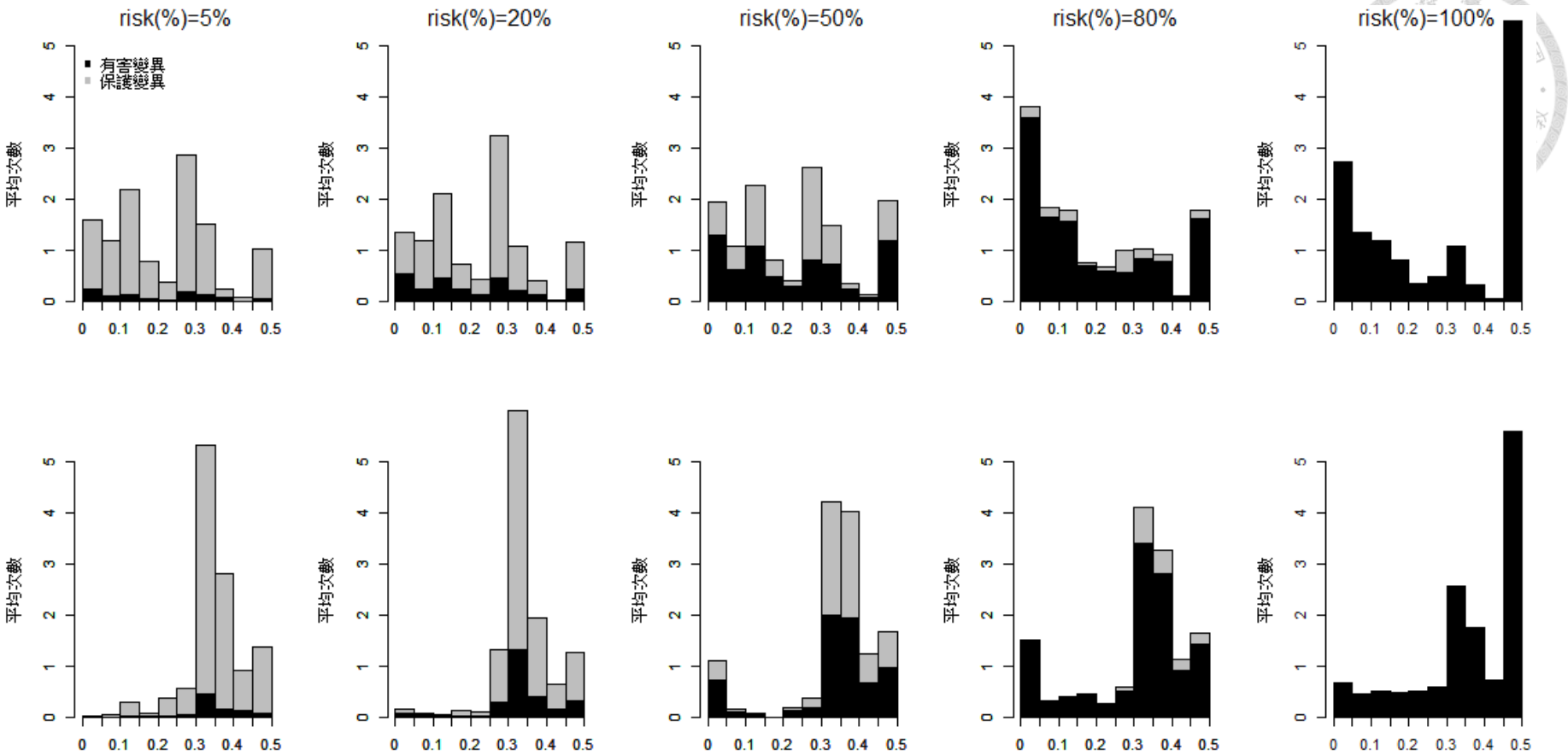
doi:10.6342/NTU201602253



圖五：比較最大 Q 值截斷門檻 (θ_{max}) 為 0.3 和 0.35 之表現

染色體區段長度為 20kb，有害/保護變異占總分析變異的比例為 7.5%，族群可歸因危險性設為 0.3%，x 軸由左至右 $r_{isk}\%$ 分別設為 5%, 20%, 50%, 80%, 與 100%。圖自左至右之 y 軸依序為真陽性的個數、偽陽性的個數與陽性預測值。圖上每個點皆為 100 次重複之平均，最大 Q 值截斷門檻 (θ_{max}) 分別設為 0.3 與 0.35。

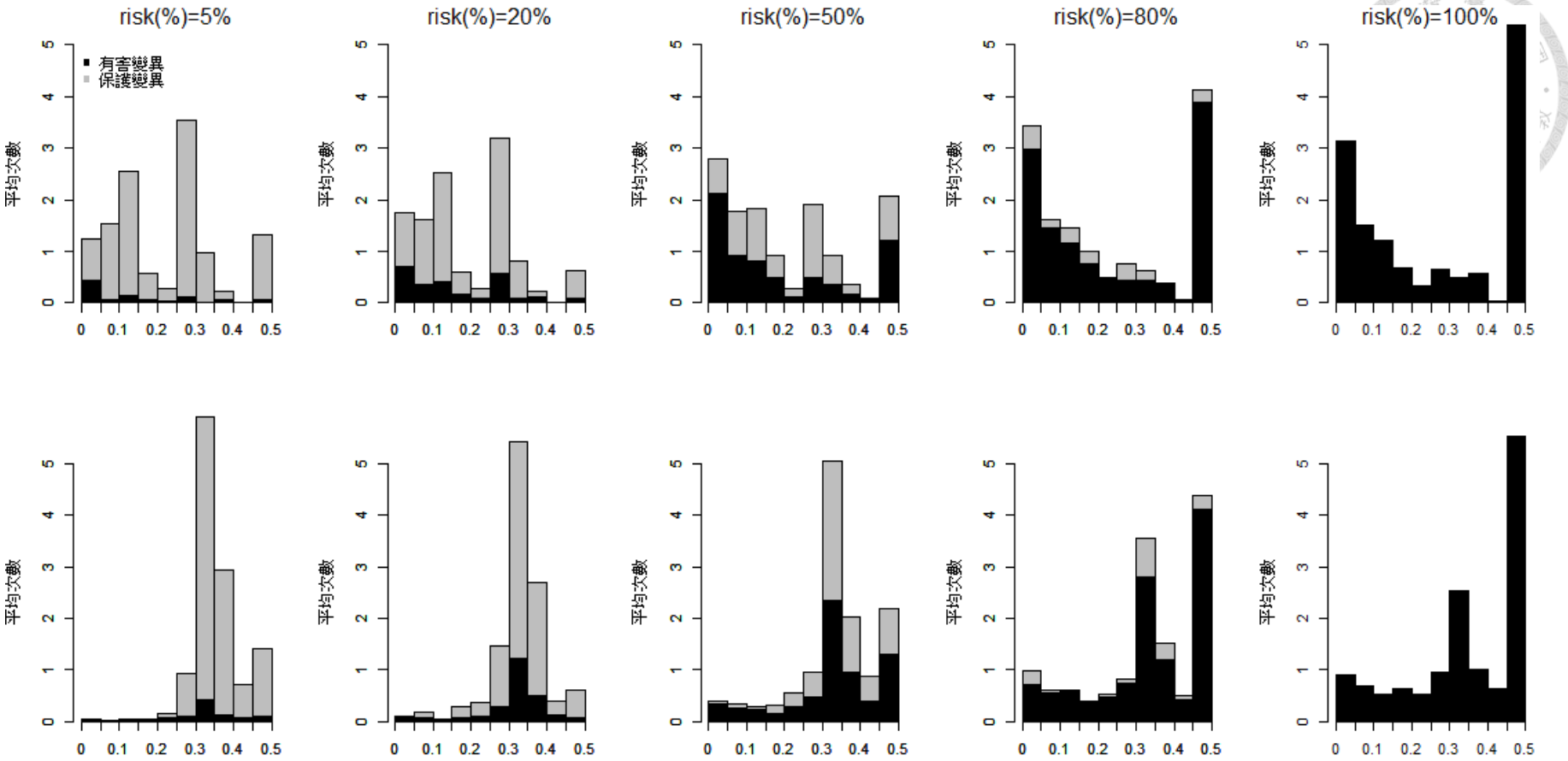
doi:10.6342/NTU201602253



圖六：二元型態性狀下，有害/保護變異 *P* 值(上列)至 *Q* 值(下列)的變化

情境：有害/保護變異占總分析變異的比例為 7.5%、族群可歸因危險性為 0.3%，此圖列出有害/保護變異 *P* 值(上列)至 *Q* 值(下列)的變化。上列 *x* 軸為 *P* 值，下列 *x* 軸為 *Q* 值，*y* 軸為重複 100 次模擬後，*P* 值/*Q* 值位於一區間內的平均次數，區間共有 10 個，從 0 到 0.5 每隔 0.05 為一區間。

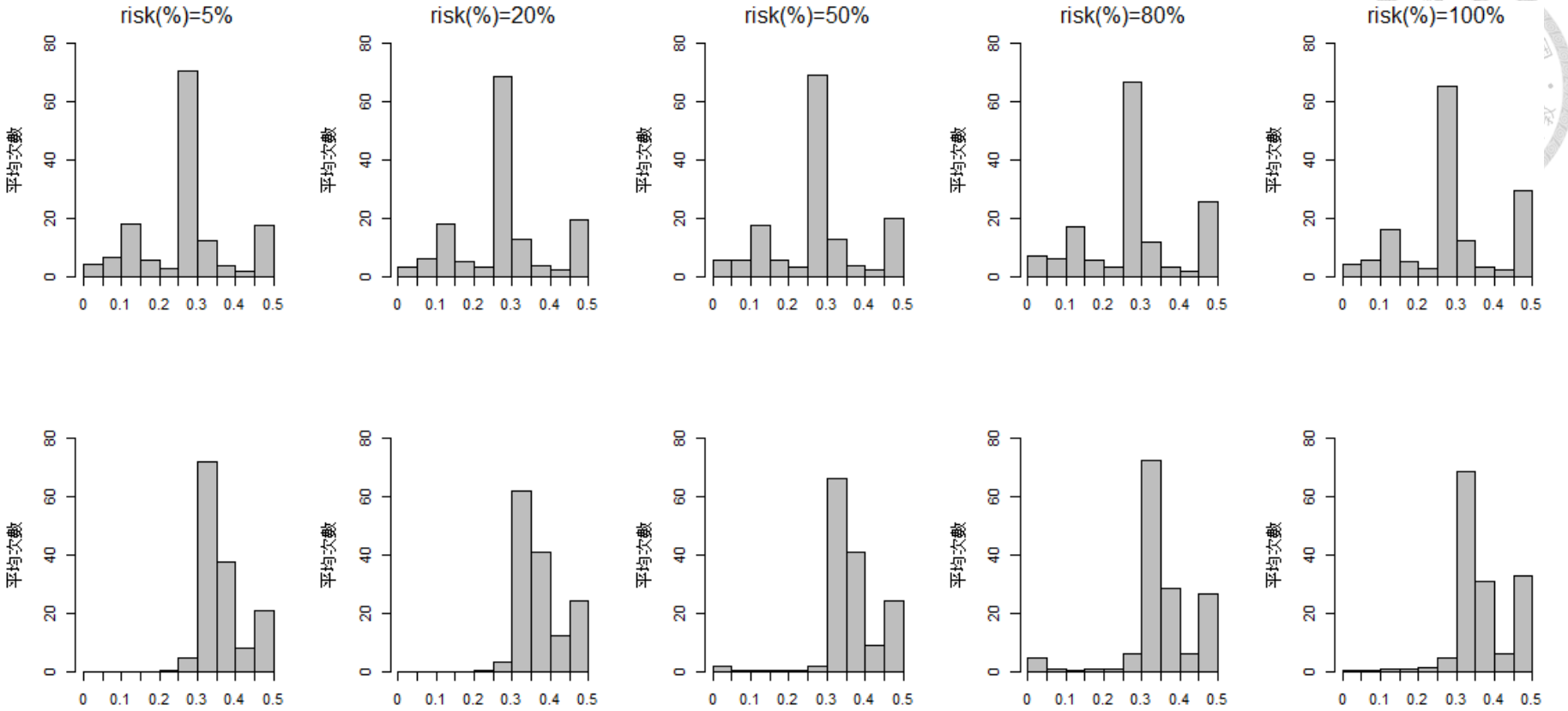
doi:10.6342/NTU201602253



圖七：二元型態性狀下，有害/保護變異 *P* 值(上列)至 *Q* 值(下列)的變化

情境：有害/保護變異占總分析變異的比例為 7.5%、族群可歸因危險性為 0.5%，此圖列出有害/保護變異 *P* 值(上列)至 *Q* 值(下列)的變化。上列 *x* 軸為 *P* 值，下列 *x* 軸為 *Q* 值，*y* 軸為重複 100 次模擬後，*P* 值/*Q* 值位於一區間內的平均次數，區間共有 10 個，從 0 到 0.5 每隔 0.05 為一區間。

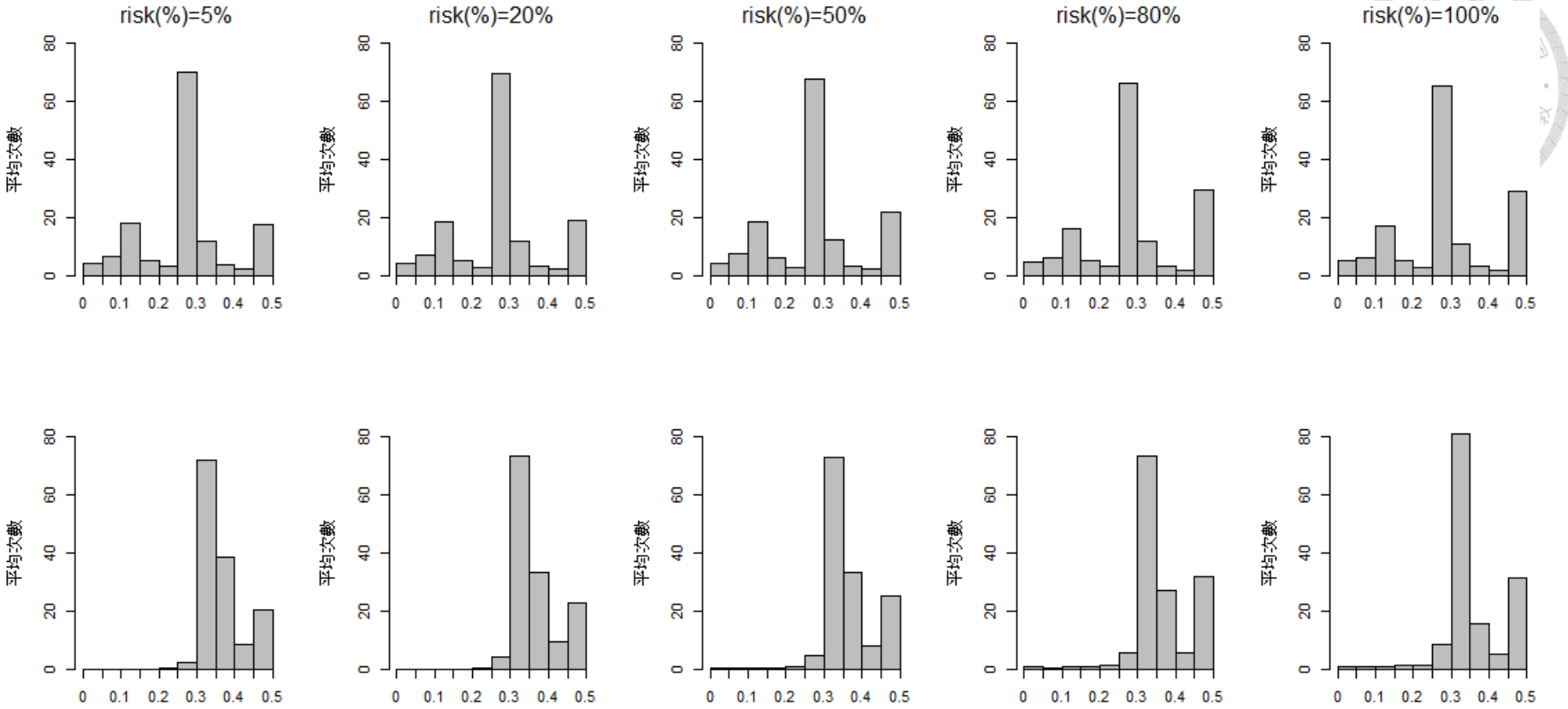
doi:10.6342/NTU201602253



圖八：二元型態性狀下，中立變異 P 值(上列)至 Q 值(下列)的變化

情境：有害/保護變異占總分析變異的比例為 7.5%、族群可歸因危險性為 0.3%，此圖列出中立變異 P 值(上列)至 Q 值(下列)的變化。上列 x 軸為 P 值，下列 x 軸為 Q 值，y 軸為重複 100 次模擬後，P 值/Q 值位於一區間內的平均次數，區間共有 10 個，從 0 到 0.5 每隔 0.05 為一區間。

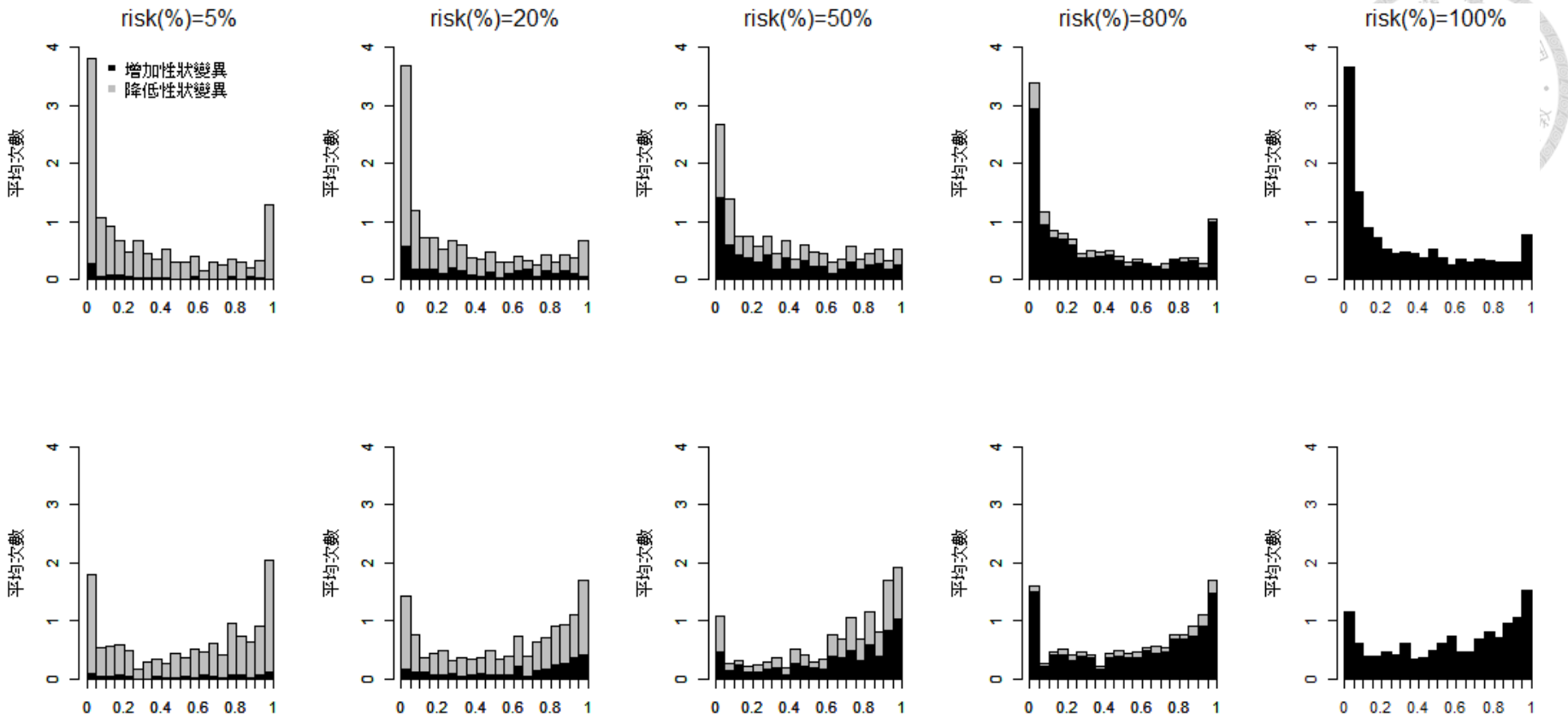
doi:10.6342/NTU201602253



圖九：二元型態性狀下，中立變異 *P* 值(上列)至 *Q* 值(下列)的變化

情境：有害/保護變異占總分析變異的比例為 7.5%、族群可歸因危險性為 0.5%，此圖列出中立變異 *P* 值(上列)至 *Q* 值(下列)的變化。上列 *x* 軸為 *P* 值，下列 *x* 軸為 *Q* 值，*y* 軸為重複 100 次模擬後，*P* 值/*Q* 值位於一區間內的平均次數，區間共有 10 個，從 0 到 0.5 每隔 0.05 為一區間。

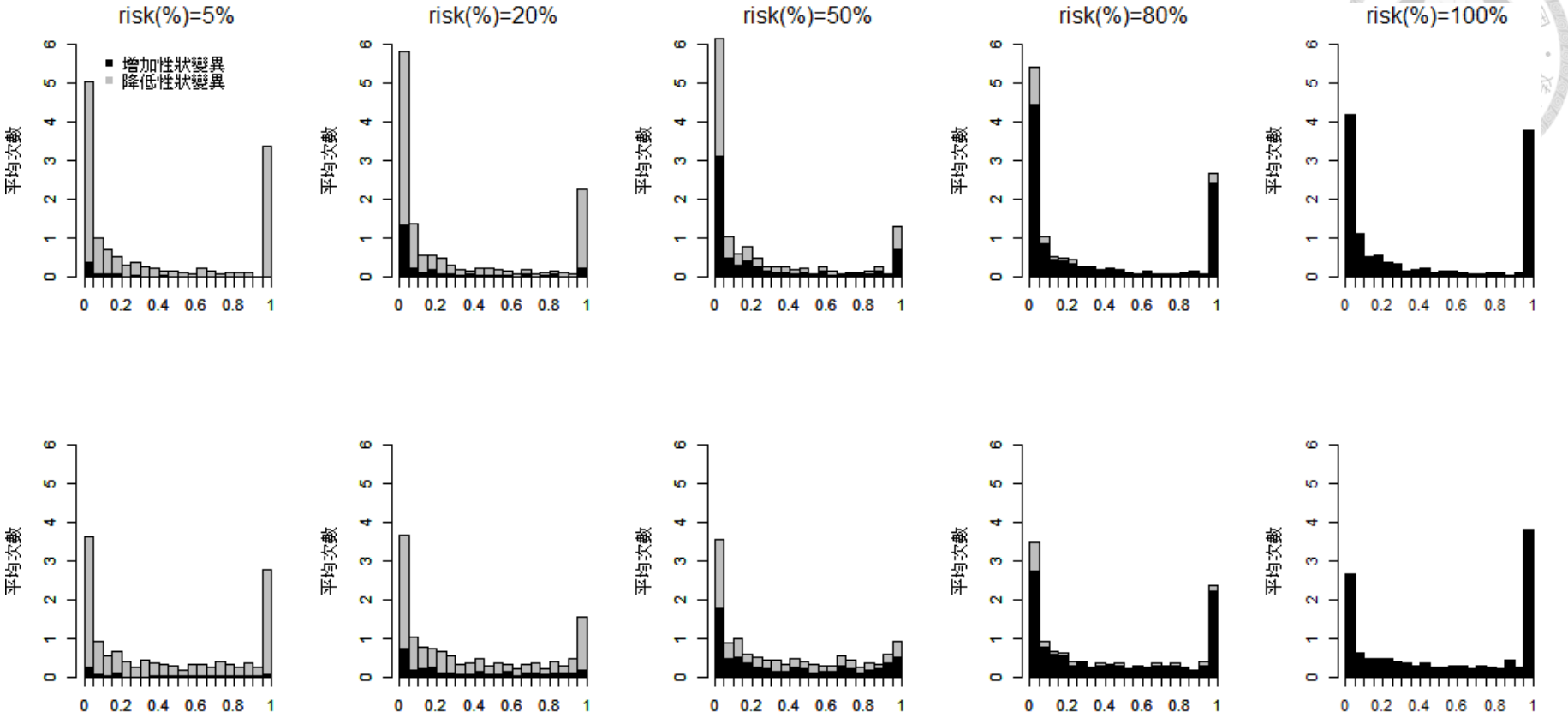
doi:10.6342/NTU201602253



圖十：連續型態性狀下，增加/降低性狀變異 *P* 值(上列)至 *Q* 值(下列)的變化

情境：增加/降低性狀變異占總分析變異的比例為 7.5%、效應值乘數 *c* 設為 0.2，此圖列出增加/降低性狀變異 *P* 值(上列)至 *Q* 值(下列)的變化。上列 *x* 軸為 *P* 值，下列 *x* 軸為 *Q* 值，*y* 軸為重複 100 次模擬後，*P* 值/*Q* 值位於一區間內的平均次數，區間共有 20 個，從 0 到 1 每隔 0.05 為一區間。

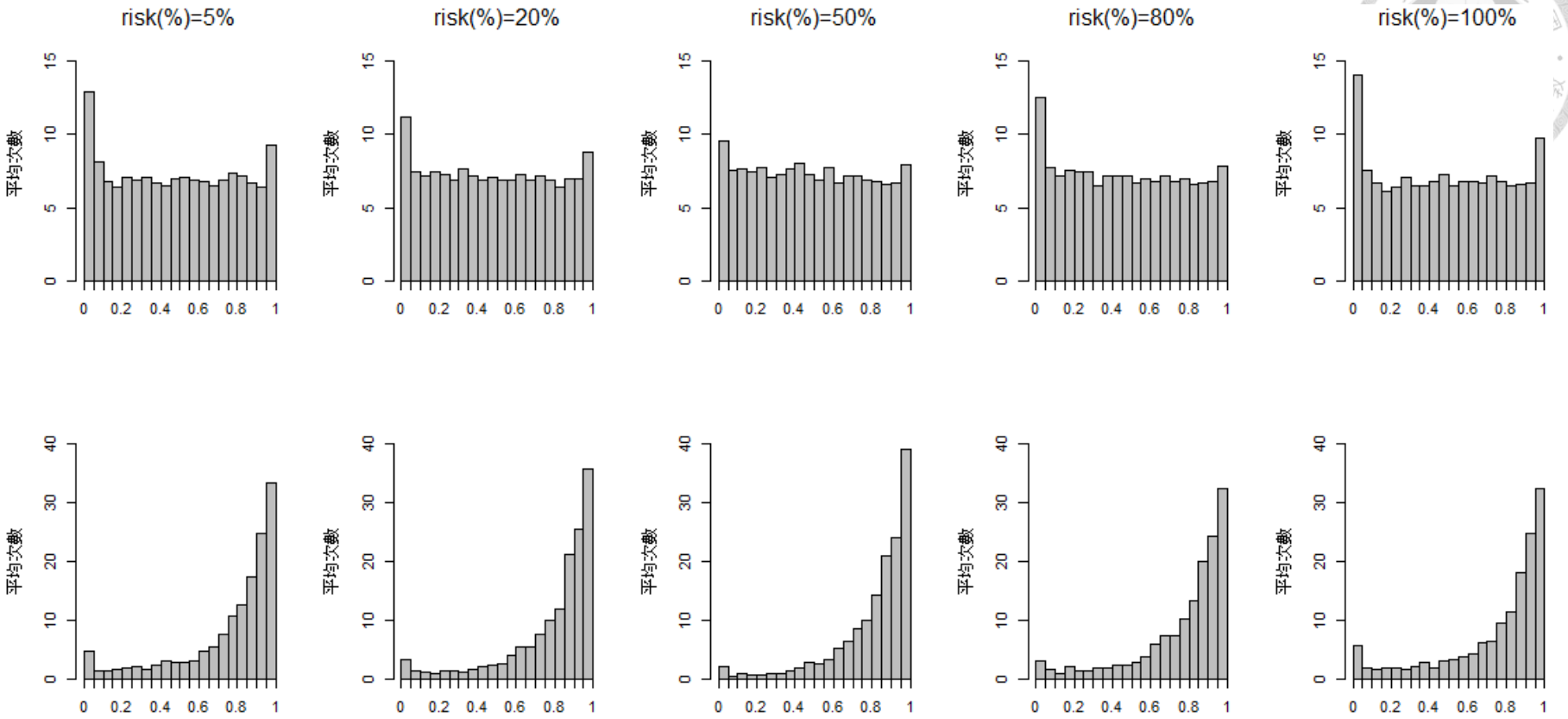
doi:10.6342/NTU201602253



圖十一：連續型態性狀下，增加/降低性狀變異 *P* 值(上列)至 *Q* 值(下列)的變化

情境：增加/降低性狀變異占總分析變異的比例為 7.5%、效應值乘數 *c* 設為 0.4，此圖列出增加/降低性狀變異 *P* 值(上列)至 *Q* 值(下列)的變化。上列 *x* 軸為 *P* 值，下列 *x* 軸為 *Q* 值，*y* 軸為重複 100 次模擬後，*P* 值/*Q* 值位於一區間內的平均次數，區間共有 20 個，從 0 到 1 每隔 0.05 為一區間。

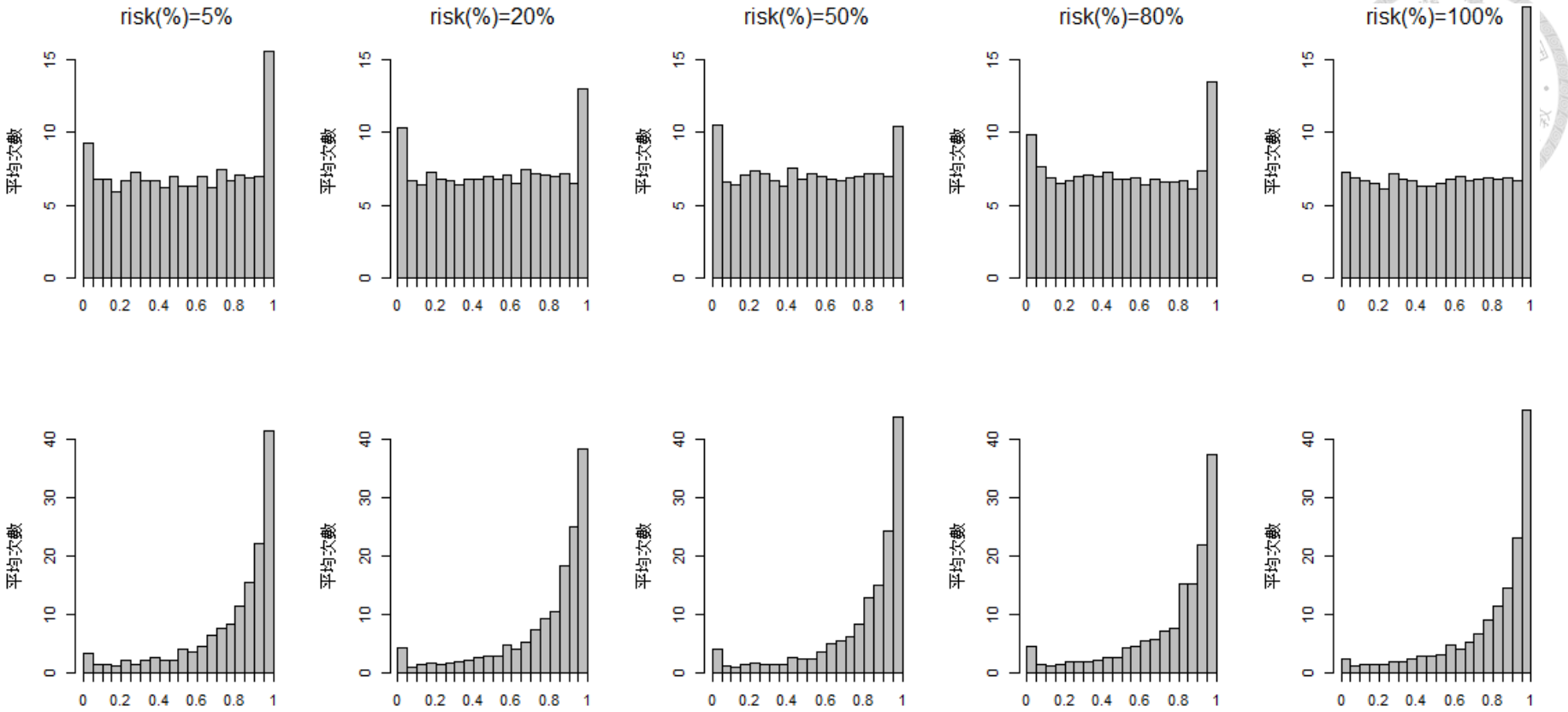
doi:10.6342/NTU201602253



圖十二：連續型態性狀下，中立變異 *P* 值(上列)至 *Q* 值(下列)的變化

情境：增加/降低性狀變異占總分析變異的比例為 7.5%、效應值乘數 *c* 設為 0.2，此圖列出中立變異 *P* 值(上列)至 *Q* 值(下列)的變化。上列 *x* 軸為 *P* 值，下列 *x* 軸為 *Q* 值，*y* 軸為重複 100 次模擬後，中立變異 *P* 值/*Q* 值位於一區間內的平均次數，區間共有 20 個，從 0 到 1 每隔 0.05 為一區間。

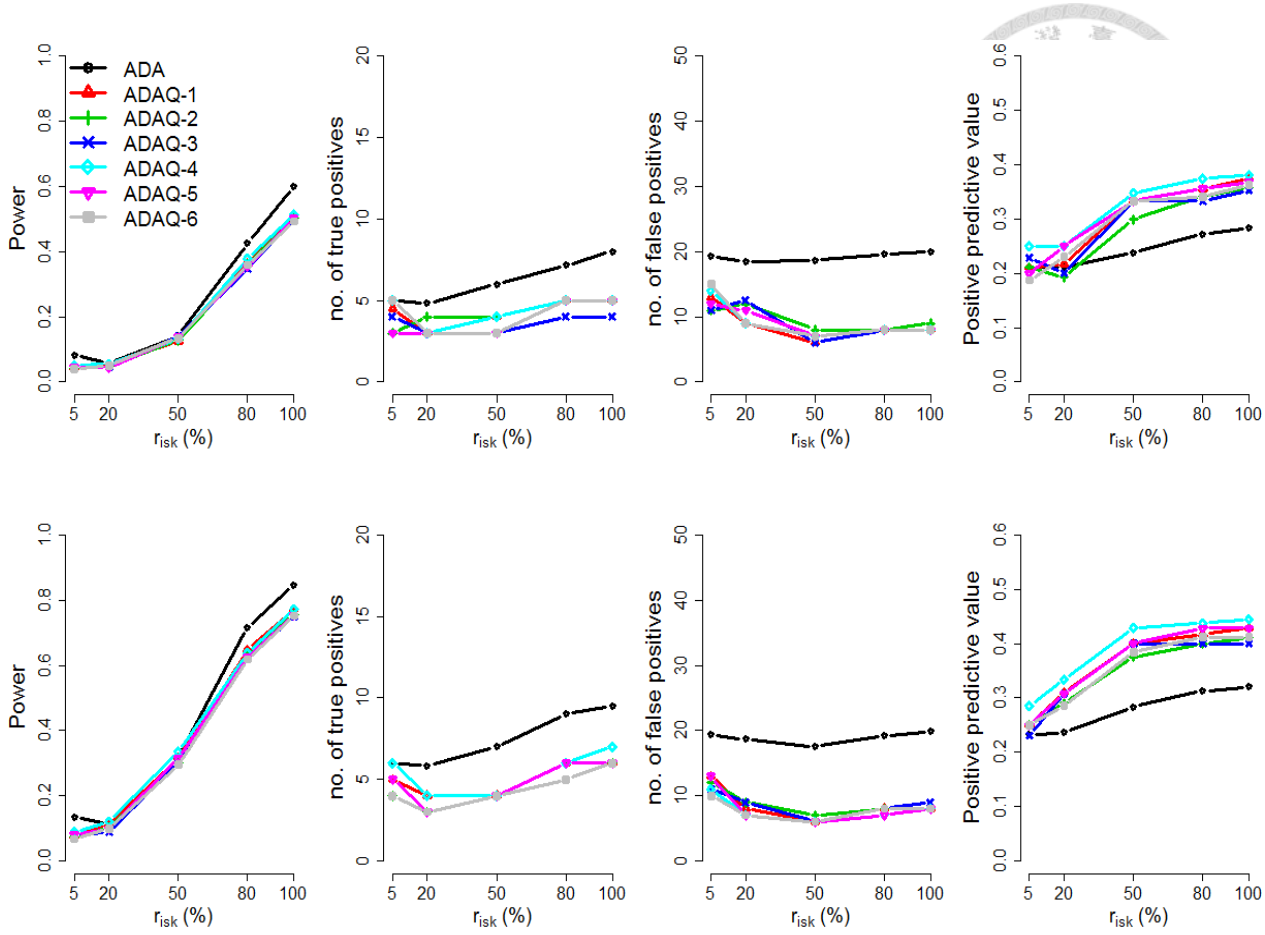
doi:10.6342/NTU201602253



圖十三：連續型態性狀下，中立變異 P 值(上列)至 Q 值(下列)的變化

情境：增加/降低性狀變異占總分析變異的比例為 7.5%、效應值乘數 c 設為 0.4，此圖列出中立變異 P 值(上列)至 Q 值(下列)的變化。上列 x 軸為 P 值，下列 x 軸為 Q 值，y 軸為重複 100 次模擬後，中立變異 P 值/Q 值位於一區間內的平均次數，區間共有 20 個，從 0 到 1 每隔 0.05 為一區間。

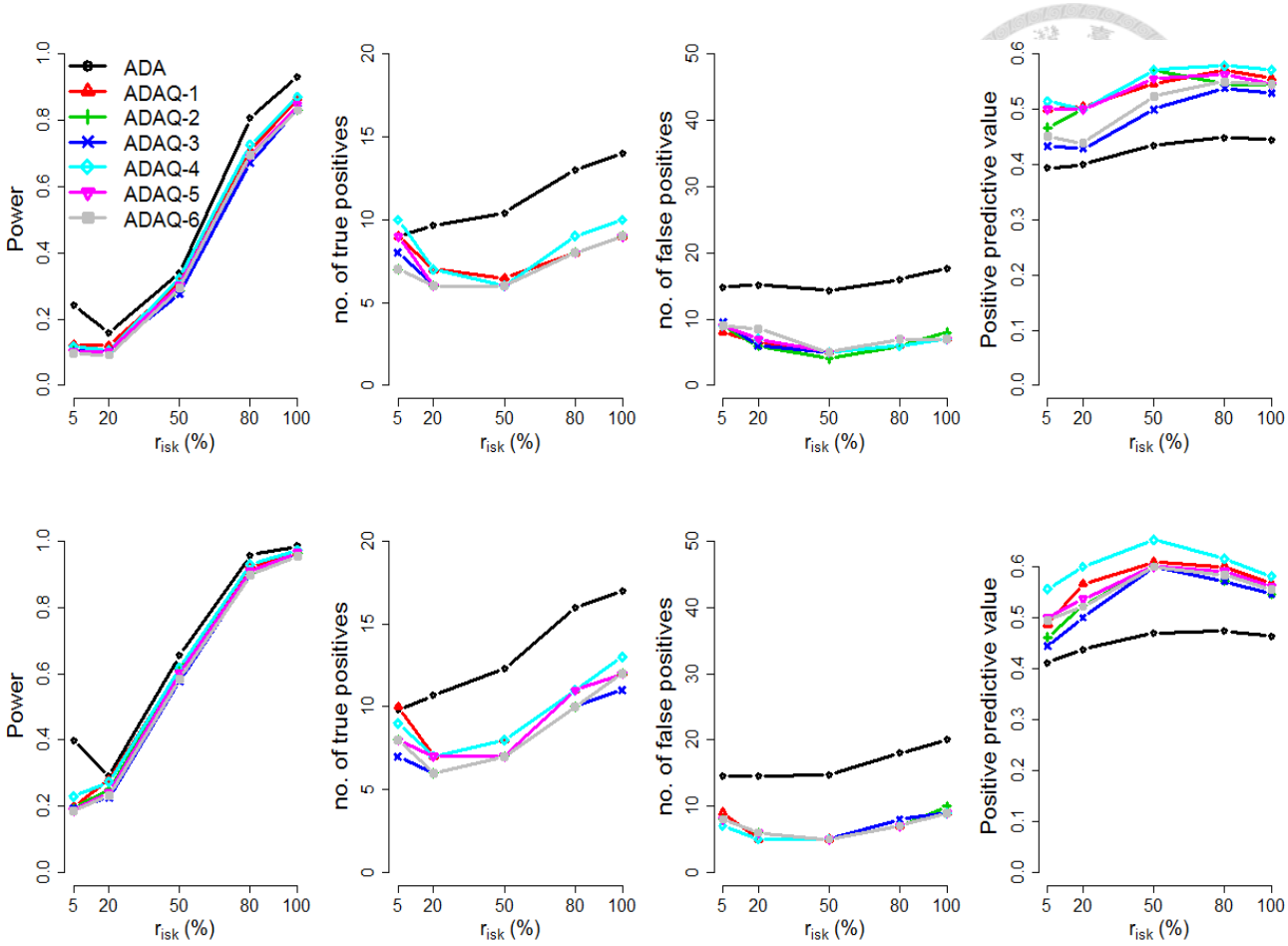
doi:10.6342/NTU201602253



圖十四：二元型態性狀下，各方法的檢定力(顯著水準訂為 0.01)、真陽性個數、偽陽性個數與陽性預測值

情境：染色體區段長度為 20kb，有害/保護變異占總分析變異之 7.5%，每個有害/保護變異之族群可歸因危險性為 0.3% (上列) 或 0.5% (下列)，*x* 軸為有害變異占有害/保護變異之比例，自 5%至 100%。吾人自 1,000 個 Cosi 序列資料集中，於每種模擬情境之下，每個序列資料集皆進行重複模擬 2 次，故共有 2,000 次 (1000×2) 重複，訂定顯著水準為 0.01，吾人計算於這 2,000 次重複中「調整後 *P* 值」小於顯著水準 0.01 的比例，此為檢定力 (第 1 欄)。另外，於「調整後 *P* 值」小於顯著水準 0.01 的情況下，代表該基因/區域與疾病狀態有顯著相關，進而以 ADAQ 與 ADA 來指出個別的罕見致病變異，第 2-4 欄分別列出真陽性個數、偽陽性個數及陽性預測值之中位數。

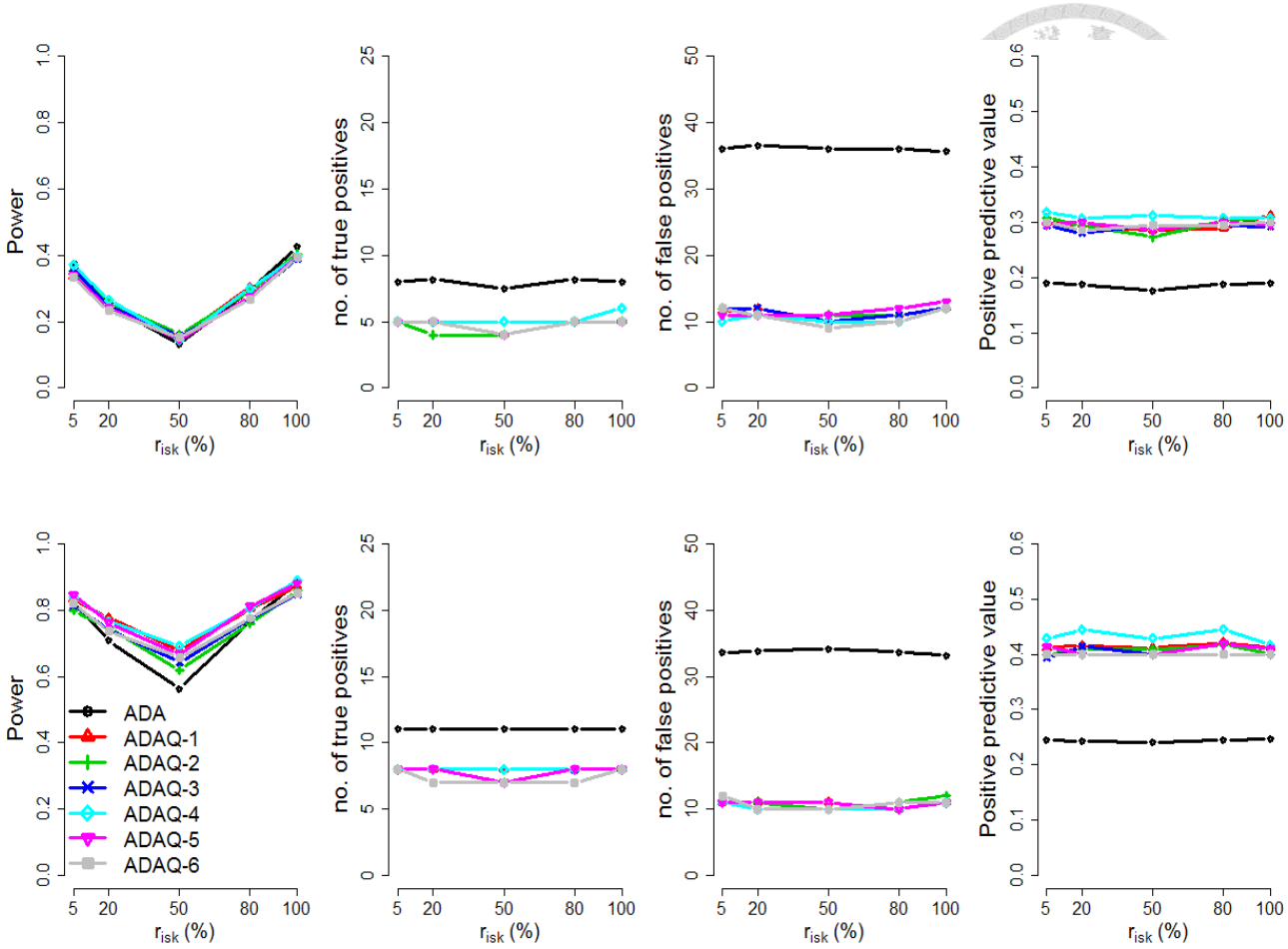
doi:10.6342/NTU201602253



圖十五：二元型態性狀下，各方法的檢定力(顯著水準訂為 0.01)、真陽性個數、偽陽性個數與陽性預測值

情境：染色體區段長度為 20kb，有害/保護變異占總分析變異之 15%，每個有害/保護變異之族群可歸因危險性為 0.3% (上列) 或 0.5% (下列)，*x* 軸為有害變異占有害/保護變異之比例，自 5%至 100%。吾人自 1,000 個 Cosi 序列資料集中，於每種模擬情境之下，每個序列資料集皆進行重複模擬 2 次，故共有 2,000 次 (1000×2) 重複，訂定顯著水準為 0.01，吾人計算於這 2,000 次重複中「調整後 *P* 值」小於顯著水準 0.01 的比例，此為檢定力 (第 1 欄)。另外，於「調整後 *P* 值」小於顯著水準 0.01 的情況下，代表該基因/區域與疾病狀態有顯著相關，進而以 ADAQ 與 ADA 來指出個別的罕見致病變異，第 2-4 欄分別列出真陽性個數、偽陽性個數及陽性預測值之中位數。

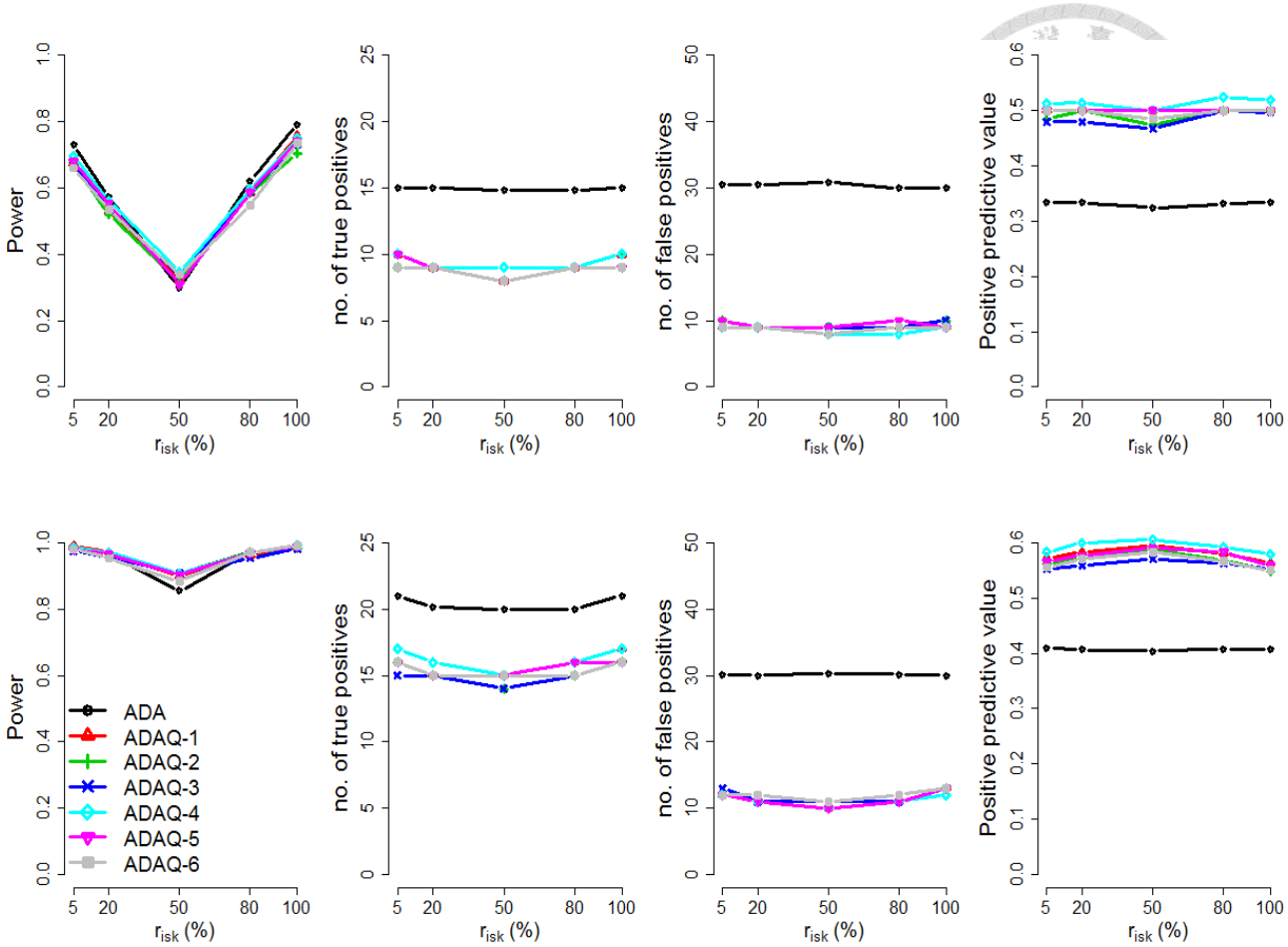
doi:10.6342/NTU201602253



圖十六：連續型態性狀下，各方法的檢定力(顯著水準訂為 0.01)、真陽性個數、偽陽性個數與陽性預測值

情境：染色體區段長度為 20kb，增加/降低性狀變異占總分析變異之 7.5%，每個增加/降低性狀變異之效應值乘數 c 為 0.2 (上列) 或 0.4 (下列)，x 軸為增加性狀變異占增加/降低性狀變異之比例，自 5%至 100%。吾人自 1,000 個 Cosi 序列資料集中，於每種模擬情境之下，每個序列資料集皆進行重複模擬 2 次，故共有 2,000 次 (1000×2) 重複，訂定顯著水準為 0.01，吾人計算於這 2,000 次重複中「調整後 P 值」小於顯著水準 0.01 的比例，此為檢定力 (第 1 欄)。另外，於「調整後 P 值」小於顯著水準 0.01 的情況下，代表該基因/區域與性狀值有顯著相關，進而以 ADAQ 與 ADA 來指出個別的罕見致病變異，第 2-4 欄分別列出真陽性個數、偽陽性個數及陽性預測值之中位數。

doi:10.6342/NTU201602253



圖十七：連續型態性狀下，各方法的檢定力(顯著水準訂為 0.01)、真陽性個數、偽陽性個數與陽性預測值

情境：染色體區段長度為 20kb，增加/降低性狀變異占總分析變異之 15%，每個增加/降低性狀變異之效應值乘數 c 為 0.2 (上列) 或 0.4 (下列)，x 軸為增加性狀變異占增加/降低性狀變異之比例，自 5%至 100%。吾人自 1,000 個 Cosi 序列資料集中，於每種模擬情境之下，每個序列資料集皆進行重複模擬 2 次，故共有 2,000 次 (1000×2) 重複，訂定顯著水準為 0.01，吾人計算於這 2,000 次重複中「調整後 P 值」小於顯著水準 0.01 的比例，此為檢定力 (第 1 欄)。另外，於「調整後 P 值」小於顯著水準 0.01 的情況下，代表該基因/區域與性狀值有顯著相關，進而以 ADAQ 與 ADA 來指出個別的罕見致病變異，第 2-4 欄分別列出真陽性個數、偽陽性個數及陽性預測值之中位數。

doi:10.6342/NTU201602253

參考文獻

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467(7319):1061-73.

Almasy L, Dyer TD, Peralta JM, Kent JW, Jr., Charlesworth JC, Curran JE, Blangero J. 2011. Genetic Analysis Workshop 17 mini-exome simulation. BMC Proc 5 Suppl 9:S2.

Basu S, Pan W. 2011. Comparison of statistical tests for disease association with rare variants. Genet Epidemiol 35(7):606-19.

Benaglia T, Chauveau D, Hunter DR. 2009. An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures. Computational and Graphical Statistics 18:505-526.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. B. 57:289-300.

Bertram L, Tanzi RE. 2009. Genome-wide association studies in Alzheimer's disease. Hum Mol Genet 18(R2):R137-45.

Besag J, Clifford P. 1991. Sequential Monte Carlo p-values. Biometrika 78:301-304.

Byrnes AE, Wu MC, Wright FA, Li M, Li Y. 2013. The value of statistical or bioinformatics annotation for rare variant association with quantitative trait. Genet Epidemiol 37(7):666-74.

Cheung YH, Wang G, Leal SM, Wang S. 2012. A fast and noise-resilient approach to detect rare-variant associations with deep sequencing data for complex disorders. Genet Epidemiol 36(7):675-85.

Cirulli ET, Goldstein DB. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11(6):415-25.

Davies R. 1980. The distribution of a linear combination of chi-square random variables. J. R. Stat. Soc. Ser. C Appl. Stat. 29:323-333.

Derkach A, Lawless JF, Sun L. 2013. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. Genet Epidemiol 37(1):110-21.

Fisher RA. 1922. On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. J. R. Stat. Soc. 85:87-94.

Fisher RA. 1932. Statistical methods for research workers. London: Oliver and Boyd.

Han F, Pan W. 2010. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered 70(1):42-54.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18(2):337-8.

Ionita-Laza I, Capanu M, De Rubeis S, McCallum K, Buxbaum JD. 2014. Identification of rare causal variants in sequence-based studies: methods and applications to VPS13B, a gene involved in Cohen syndrome and autism. PLoS Genet 10(12):e1004729.

Lee S, Wu MC, Lin X. 2012. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13(4):762-75.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83(3):311-21.

Li C, Li M, Lange EM, Watanabe RM. 2008. Prioritized subset analysis: improving

doi:10.6342/NTU201602253

power in genome-wide association studies. Hum Hered 65(3):129-41.
Li Y, Byrnes AE, Li M. 2010. To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests. Am J Hum Genet 87(5):728-35.
Lin WY. 2014a. Adaptive combination of P-values for family-based association testing with sequence data. PLoS One 9(12):e115971.
Lin WY. 2014b. Association testing of clustered rare causal variants in case-control studies. PLoS One 9(4):e94337.
Lin WY. 2016. Beyond Rare-Variant Association Testing: Pinpointing Rare Causal Variants in Case-Control Sequencing Study. Sci Rep 6:21824.
Lin WY, Lee WC. 2012. Improving power of genome-wide association studies with weighted false discovery rate control and prioritized subset analysis. PLoS One 7(4):e33716.
Lin WY, Liang YC. 2016. Conditioning adaptive combination of P-values method to analyze case-parent trios with or without population controls. Sci Rep 6:28389.
Lin WY, Lou XY, Gao G, Liu N. 2014. Rare Variant Association Testing by Adaptive Combination of P-values. PLoS One 9(1):e85728.
Lin WY, Yi N, Lou XY, Zhi D, Zhang K, Gao G, Tiwari HK, Liu N. 2013. Haplotype kernel association test as a powerful method to identify chromosomal regions harboring uncommon causal variants. Genet Epidemiol 37(6):560-70.
Lin WY, Yi N, Zhi D, Zhang K, Gao G, Tiwari HK, Liu N. 2012. Haplotype-based methods for detecting uncommon causal variants with common SNPs. Genet Epidemiol 36(6):572-82.
Lin WY, Zhang B, Yi N, Gao G, Liu N. 2011. Evaluation of pooled association tests for rare variant identification. BMC Proc 5 Suppl 9:S118.
Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5(2):e1000384.
Morgenthaler S, Thilly WG. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res 615(1-2):28-56.
Morris AP, Zeggini E. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34(2):188-93.
Peng B. 2015. Reproducible simulations of realistic samples for next-generation sequencing studies using Variant Simulation Tools. Genet Epidemiol 39(1):45-52.
Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. 2010. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86(6):832-8.
Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30(17):3894-900.
Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15(11):1576-83.
Schizophrenia Working Group of the Psychiatric Genomics C. 2014. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511(7510):421-427.
Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, Dudbridge F, Holmans PA, Whittemore AS, Mowry BJ and others. 2009. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 460(7256):753-7.
Sullivan PF, Daly MJ, O'Donovan M. 2012. Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat Rev Genet

doi:10.6342/NTU201602253

13(8):537-51.
Tintle N, Aschard H, Hu IC, Nock N, Wang HT, Pugh E. 2011. Inflated type I error rates when using aggregation methods to analyze rare variants in the 1000 Genomes Project exon sequencing data in unrelated individuals: summary results from Group 7 at Genetic Analysis Workshop 17. Genetic Epidemiology 35:S56-S60.
Wang GT, Zhang D, He Z, Hang D, Li B, Leal S. 2015. Pitfalls in development of statistical methods for rare variant association studies. Presented at the 65th Annual Meeting of The American Society of Human Genetics, October 7, 2015, Baltimore, MD.
Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82-93.
Yang HC, Chen CW. 2011. Region-based and pathway-based QTL mapping using a p-value combination method. BMC Proc 5 Suppl 9:S43.
Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. 2002. Truncated product method for combining P-values. Genet Epidemiol 22(2):170-85.

doi:10.6342/NTU201602253