National Translational Medicine and
Clinical Trial Resource Center
國家轉譯醫學與臨床試驗資源中心

Clinical Translational Science: Principles of Human Research

Chapter 20: Epidemiologic and Population Genetic Studies

Author: Angela J. Rogers and Scott Weiss

Mei-Hsin Su

2011-12-08

Abstract

本章主要在介紹遺傳相關性之流行病學研究。分別從研究設計、描述遺傳與疾病之相關性、以及遺傳研究之未來展望及可能的研究方向這三個部分來介紹。

第一部分

Design issue in genetic association studies

遺傳相關性研究之研究設計

Outline

- Define a phenotype
- Epidemiologic study design
- Genetic study design
- Summary on study design

Definition of phenotype -1

- Phenotype的定義不清時，會導致其他研究者在複製相同的研究時無法得到相同的結果

  - 舉例：研究糖尿病時，其中一個研究是將hemoglobin a1c為negative的樣本當作control組，但另一個研究團隊卻是以自陳沒有糖尿病的人當作control組

Definition of phenotype -2

- 除了將phenotype以二分類(有病/沒病)變項來分類之外，尚可將intermediate phenotypes (中間物)合併分析
  - 舉例：探討氣喘時，肺活量(肺功能的指標)以及methacholine 激發的程度(呼吸道過度反應的指標)即是所謂的intermediate phenotype，可將之合併納入分析

Definition of phenotype -3

- 將intermediate phenotype納入分析可提高power，因為此類的連續性變項會比原本的二分類變項含有較高的變異
- Intermediate phenotype可以將疾病的症狀分類成較為均質性的subgroup

Definition of phenotype -4

舉個例子來說明：

- 過去有許多研究在探討遺傳對貧血的貢獻有多大時，都試圖以「是否貧血」當作phenotype，結果都發現，結果卻發現此phenotype與遺傳的相關性是低的 (此時便稱貧血為distant phenotype)
- 若欲知貧血的遺傳基礎，intermediate phenotype是很重要的，不能只考量「是否貧血」尚需考慮「紅血球細胞是否為鐮刀狀(或其他遺傳性貧血的狀態)」
- 此時，「鐮刀形紅血球」就稱作 intermediate phenotype

使用**intermediate phenotype**的優點

- 可以控制distant phenotype與intermediate phenotype各別所受遺傳因素的影響

  Eg. 承上述貧血的例子，若以intermediate phenotype而非distant phenotype來分析，那麼患有其他種貧血的樣本便可作為對照組

- 當所研究的世代同質性較高時 (more homogeneous) 只需較少的樣本數即可達相當的power
- 當研究世代的特徵與實際上的基因表現的擾動(genetic perturbation)較相近時，就適用以孟德爾遺傳理倫為基礎的分析

使用**intermediate phenotype**的缺點

- 對某些疾病而言，要定義intermediate phenotype是較為困難的

  例如欲找出高血壓的intermediate phenotype，就常會受環境因素干擾。像是鹽分的攝取量會與血壓有關，但卻不是每位高血壓病患的血壓都受鹽份攝取量的影響。因此在定義intermediate phenotype時，就會受鹽份攝取量干擾

- Define a phenotype
- Epidemiologic study design
  - Case-control study
  - Population study
  - Family-based study
- Genetic study design
- Summary on study design

1. Case-control studies

- 在遺傳與疾病的相關性研究中最常被使用的研究設計
- Case與control需來自同一個族群(種族)

  若是納入分析的樣本來自過度分散的族群，將會使得全樣本的遺傳背景變異過大，如此一來便會導致false positive。對於來自許多不同族群的遺傳研究建議在分析時使用分層分析方可避免上述問題。分層分析會在後面章節提到。

2. Cohort or population studies

- Cohort是指：

  任何一群被定義納入研究的個體，且該個體需經過一段時間的追蹤或回溯。 (Last and Abramson, 1995)

- 根據追蹤的時間，可分為prospective(由目前的時間點開始隨時間的演進來追蹤)和retrospective(從目前的時間點進行回溯)
- Cohort study也是遺傳相關研究中常用的設計

Cohort study的優點

- 追蹤到最後不論發病與否，這群樣本都來自於同一個cohort，因此會比case-control study更能避免假相關的出現
- 由於cohort study是在疾病發生前便開始進行調查，因此可避免recall bias

Cohort study的缺點

- 費用昂貴
- 不適用於罕見疾病
- 若是針對特殊族群所做的(例如有特殊工作暴露者)，則所得到的結果就無法外推至其他人口

3. 家族研究 (FBAT)
Family-based genetic association studies

- 家族研究是指：當有一小孩(proband)患病時，同時也將其家庭成員(父母)一同納入當作一個研究的單位，稱為trio
- 家族研究的基礎是 transmission-disequilibrium test (TDT)，也就是觀察親代對於患病的子代是否有過度遺傳某種allele給子代

TDT

Family	患病小孩的基因型	父親基因型	母親基因型
1	AA	Aa	Aa
2	Aa	Aa	Aa
3	Aa	aa	Aa
4	AA	Aa	AA
5	aa	aa	Aa
……	……	…	…

假設有一筆家族資料如上表，則可推算父母親傳給子女的allele之個數及種類，列出一個2-by-2 table如下頁

	不傳遞	
	x1	x2
傳遞		
x1		86
x2	42	

假設父母雙親中，有傳x1但沒有傳x2給子女的family共有86個，而有傳x2但沒有傳x1給子女的family共有42個，則可套入卡方檢定公式：

$$\chi^2 = \frac{(b-c)^2}{b+c} = (86\text{-}42)^2/(86+42) = 12{,}125$$

查表後得$p<0.05$，表示父母親傳給子女的allele並不符合隨機(即x1與x2各占50%)，表示此疾病與父母親所傳遞的某個allele是有相關性的

若是雙親中的genotype為homozygous，則將無法提供訊息

family-based studies的優點

- 因為每個小孩的基因型都源自於親代，因此家族的基因相關性研究不受種族差異影響
- 在每一個trio當中，只需一個人有phenotype資料即可，此特質對於一些phenotype需要昂貴花費或是需要侵入性技術才能定義phenotype的研究而言是較為有利的

family-based studies的缺點

- 不適用於年老時才發病的研究(如Alzheimer's disease)，因為屆時已收不到proband的雙親的genotype
- 因為家族研究需要proband與雙親，共計3人的基因型資料，因此需花費較多的金額在做genotyping。
- 若是雙親的基因型不是heterozygous，那麼就無法提供計算TDT的訊息。

- Define a phenotype
- Epidemiologic study design
- Genetic study design

  GWAS vs. candidate gene approaches
- Summary on study design

GWAS (A hypothesis-free testing)

- GWAS是一個不需要任何假設的association analysis，也就是說，不需要假設任何SNP或region有可能會與疾病有關，而是直接針對全基因進行分析
- 目前做GWAS的兩大主要平台為
  1. Affymetrix (以random distribution的方式在整個genome上挑選SNP)
  2. Illumina (以haplotype-tagging的方式在整個genome上挑選SNP)

Candidate gene testing

- 選定SNP後針對特定的SNP進行分析
- 可用於小型的基因遺傳研究和已知pathway的intermediate phenotyping studies
- Candidate gene test在選擇SNP時，強烈地受到預算限制。有時可只做1至2個SNP，但有時可針對一整個region做sequencing
- Candidate gene test在選擇SNP時，也受平台影響，有些平台擅長少數SNP(如Taqman)，而有些平台較擅長做大量SNP的genotyping(如SEQUENOM)

Reviewing the past literature

- 在進行association study時，最重要的工作就是選擇SNP，而其中一個選擇SNP的方法就是透過文獻回顧
- 若某一個SNP曾被報導過與研究者有興趣的疾病有關，那麼就需將該SNP納入分析
- 此處推薦一個NCBI的網站:
  www.ncbi.nlm.nih.gov

  此網站包含許多資料庫，可查詢與特定疾病相關的SNP

NCBI共包含以下資料庫：

- OMIN (Online Mendelian Inheritance in Man,人類遺傳疾病資料庫)
  - www.ncbi.nlm.nih.gov/omim
  - 是一個集合基因以及基因相關疾病的資料庫
  - 包含了基因結構及功能等概觀
  - 會將每個gene有關的研究(曾發表的)進行網路連結，也會連結到與每個gene有關的疾病(曾被研究且發表的)
- Pubmed
  - www.ncbi.nlm.nih.gov/pubmed
  - 內含最新發表的文章資訊，且有較多免費的文章，提供完整的associated SNP相關研究，適合進行文獻回顧。

Locating SNP within the gene

介紹2個基因搜尋以及SNP搜尋的網站：

- UCSC (genome bioinformatics site)
  - http://genome.ucsc.edu/
  - 是一基因搜尋(genome browser)的工具
  - 除gene browser之外，尚提供許多生物資訊的工具
  - 可搜尋SNP的位置、功能等資訊
- SNPper
  - http://snpper.chip.org/bio/snpper
  - 可透過基因名稱或 physical position來查詢相關的SNP資料
  - 也可以透過 SNP來查詢相關基因的訊息

Identifying LD-tagging SNP

- HapMap (www.hapmap.org)
  - 此網站包含8個種族的genome-wide資料，除了independent individual之外，尚包含trio的資料
  - 提供 LD-tagging SNP的資訊，可直接將Hapmap資料上載至Haploview軟體進行LD分析

註：LD = linkage disequilibrium

- Define a phenotype
- Epidemiologic study design
- Genetic study design
- Summary on study design

- 一個遺傳相關性研究(genetic association study)需要好的研究設計，包含以下元素：
  1. 選擇一個合適的phenotype
  2. 透過廣泛已知的相關文獻搜尋選擇適當的SNP
  3. 決定分析方式(是要用GWAS或是candidate gene-based approach)

第二部分

Interpreting results of genetic association studies

描述/解釋研究結果

- 所有流行病學研究都會有無法克服的bias出現，而遺傳研究也不例外。因此，這個部分將提出4個在遺傳相關性研究中的重要觀念

1. 資料品質的管理(quality control)
2. 利用多重比較來確認分析結果的正確性(Correction for multiple comparisons)
3. 樣本的分層分析(Population stratification)可排除干擾因子影響結果
4. 合適的樣本數與Power 可避免false negative的結果

Outline

- Genotype quality control
- Correction for multiple comparisons
- Population stratification
- Power
- Conclusion

Genotype quality control

Genotyping的品質常會影響分析的結果：

- genotyping error會導致false positive和false negative
- Genotype的錯誤分組(misclassification)會增加case-control study的false negative
- 由此可見data quality control的重要性。以下就介紹幾個quality control的要點

1. Genotyping的完成率 (Genotype completion rate)

- Genotyping completion rate是指在所有marker中，有多少比例的marker是有被定序出來的(沒有missing的)，又稱作call rate
- 目前的平台技術都可達到90%至95%的call rate
- 若call rate未達上述表準，則可能會降低association analysis的power
- 因此，未達上述標準的marker應該要被剔除而不能納入後續的分析

2. 可重現的genotyping
(Reproducible genotyping)

- Reproducible genotyping有許多不同方法：
  1. 針對特定的locus重複做genotyping
  2. 從樣本中挑出5-10%的樣本重複做genotyping
  3. 利用不同的平台重複做genotyping
- 在重複genotyping時，可接受的不一致率約在1%，超過時即表示可能有genotyping error

3. 哈溫平衡
(Hardy-Weinberg equilibrium, HWE)

- 假設一個locus的兩個allele(a及A)的frequency分別為p與q時，哈溫平衡就是在檢測aa, aA, 及AA是否偏離了p^2, 2pq, 及q^2分佈
- 若是在control或founder中發現有marker偏離了哈溫平衡，即表示這些marker可能會使後續的association analysis產生false positive，因此需將這些marker予以刪除

4. Mendelian errors and non-paternity
(in family-based studies)

- Mendelian error是指，假設父母親的genotype是aa及AA，則子代正常而言應該是要出現Aa的genotype，但卻除現Aa以外的基因型稱之
- 若Mendelian error超過1%，就表示有genotyping error，因此需將Mendelian error >1% 的marker予以刪除不納入後續分析
- 透過Mendelian error的檢測亦可得知每一個trio中的父母是否為proband的親生父母。若發現在trio中有Mendelian error，則須將該trio予以刪除

- Genotype quality control
- Correction for multiple comparisons
- Population stratification
- Power
- Conclusion

多重比較的校正
(correction for multiple comparisons)

- 一般流行病學的研究中，是以p值小於0.05當作統計顯著的標準(可接受的type I error rate為0.05)，但是在GWAS的研究中，由於所納入分析的marker數過多(可能有50萬至100萬)，因此在決定顯注水準時需考慮這種多重比較的情況下，是否仍能以傳統的0.05當作顯著標準
- 以下將介紹幾種用來校正多重比較之p值的方法

1. Traditional method: controlling family-wise error rate

- 最廣為人知的校正方法就是Bonferroni correction
- 其概念是將0.05除以多重比較的次數，也就是說，假設共分析了50萬個marker，那麼顯著水準就訂在$0.05/500000=10^{-7}$
- 此校正方法曾被認為太過保守，因為這個方法當每個marker不是獨立(有linkage)時，或是當樣本數不夠多(marker數很多使得須檢定的次數過多，但樣本數卻不足以配合那麼多次的檢定)時，Bonferroni的校正法有時會將可能顯著的marker給拒絕了

2. False discovery rate (FDR)

- FDR可以幫助我們知道哪些變項(SNP)有可能是false positive的。
- FDR=0.05時，是表示在發現有達顯著的SNP當中，有5%實際上是negative的，也就是只有95%的SNP是確實有顯著相關的

3. Permutation testing/empirical distribution

- Permutation testing是指，將原始genotype所對應的phenotype先打亂後再重新編排。得到新的phenotype與genotype組合之後，進行統計檢定，然後得到一個新的p值。
- 將上述步驟針對所有的loci重複做若干次，於是便會得到一個p值的empirical distribution
- 再將實際的p值與permuted p值作比較

Permutation的舉例

- 假設SNP A經檢定之後的p-value為0.0001
- 在經過permutation一萬次之後，共有150次的p-value是小於/等於0.0001的
- 則此時SNP A的empirical p-value就是150/10000=0.015
- 此時就可以利用empirical的分佈去訂定一個顯著水準，然後再用實際p值的分佈與empirical的分佈相比，若兩者相差的程度小於某一個threshold，就表示落在此region的p-value並非by chance
- 此時，就需將這個SNP作replication testing

4. Conditional power in family-based testing

- Conditional power的方法是為了解決某些family-based study的問題
- 在family study中，某些檢定(如TDT)必須要雙親帶heterozygous者才能提供訊息，如此一來，若研究中有許多的SNP都是low minor allele frequency的話，就會使得power嚴重不足
- 於是Laird和Lange就發展了conditional power screening method試圖解決這個問題

5. 條件檢定力 (Conditional power)

- 針對每一個SNP去計算conditional power，若得到的conditional power都偏低，就可將該SNP予以剔除
- 目前並無一通用的數值來界定condition power要低於多少才需刪除。一般而言，以一個共500,000SNPs的GWAS為例，會挑選conditional power最大的前1000個SNP(最powerful的)
- 如此一來，當SNP數減少時，就可減少檢定的次數而避免多重比較時p值的不正確

6. Replication

1. 第一步
   - 將sample分為testing cohort和replication cohort兩組
   - 針對testing cohort，檢定大量的SNP
   - 然後將最promising的SNP挑出來，並在replication cohort中重複檢定這些SNP
2. 第二步
   - 在進行replication cohort testing時，會同時使用一些多重比較的校正方法(如Bonferroni)來校正之

Summary for multiple comparison

- 一般而言，在針對某一族群做遺傳相關研究(genetic association study)時，會先透過上述的各種方法找出promising SNP
- 在找promising SNP時，通常會同時合併多種multiple comparison correction methods，然後再利用一或多個族群做replication
- 而這些promising SNP就可套用至其他的族群
- 目前有許多期刊都會要求在做GWAS時，需要將SNP進行replicate至少一個族群

- Genotype quality control
- Correction for multiple comparisons
- Population stratification
- Power
- Conclusion

Population stratification

- 在GWAS中，若是case與control來自不同的種族，則SNP frequency便會在兩組中有差異，此時所發現的相關性有可能是種族的genotype frequency所導致而非真的與遺傳有關
- 舉例：

  Knowler et al.於1998年發現某SNP跟type II diabetes有關，但在將case分成Indian vs. European ancestry時，就發現那其實是假相關
- 此章節將介紹GWAS中，幾個校正 population stratification的方法

1. Ethnic matching of cases and controls

- 在探討研究族群的種族背景時，除了要將case組依種族作分層外，也需針對control組做種族的分層，如此才能縮小case與control組在genotype之間的差異
- 當case與control都按種族作分層時，可限制這兩組會因為種族的差異而在相關性分析中發現假相關結果

2. Structured association

- 要解決樣本中有族群混合的情形時，可在染色體上利用隨機分佈、在遺傳上為獨立(unlinked)、且與疾病無關的的marker
- 利用這些marker(30個SNP即足夠)的差異即足以將全樣本分成數個subpopulations
- 然後在這幾個subpopulation中各自進行association analysis，此即為structured association的概念
- 相關的軟體及分析可參考：

http://pritch.bsd.uchicago.edu/structure.html

3. Genomic control

- 概念與structured association相似：利用獨立(unlinked)且與疾病無關的marker，以卡方檢定去檢定每一個locus並得到一卡方分配
- 假若data中確實有族群的分層存在時，卡方分配就會與期望值有所差異，此時的差異就以λ來表示，稱作”variance inflation factor”
- 而當真正要進行association analysis時(針對all markers)，就利用所算出來的λ來校正卡方檢定值以及p值

4. Family-based study design

- 家族研究(family-based study)通常不會有族群分層的問題，這是因為在家族研究中，所分析的都是根據allele從父母的傳遞情形，因此不會有種族分層的問題
- 因此，家族研究是一個避免因族群分層而造成干擾的最佳方法

- Genotype quality control
- Correction for multiple comparisons
- Population stratification
- Power
- Conclusion

Power

- 除了先前所介紹的genotyping error、多重比較、和族群分層會導致false negative results外，尚有另一個原因會使分析得到false negative的結果：
  - sample size不足而導致power過低
- 遺傳研究中與power有關之因素：
  1. Minor allele frequency
  2. 某一個region裡的LD
  3. Genotyping和phenotyping的正確性

- 介紹一個專門使用在family-based testing的軟體：
  - QUANTO (http://hydra.usc.edu/GxE/)
  - 此軟體加上一個PBAT之套件 (http://www.goldenhelix.com) 之後，便可計算power

- Genotype quality control
- Correction for multiple comparisons
- Population stratification
- Power
- Conclusion

Conclusion

- 本章節介紹了在遺傳相關性研究中，會導致不正確之結果的幾個因素：
  - Genotyping error
  - 無法校正多重比較
  - 族群有分層
  - Power不夠(會造成false negative)
- 然而，再完善的研究設計都還是有可能會發生false positive或negative的情形

第三部分

Future directions

未來的研究方向

- 目前在遺傳相關研究中，最常使用的指標就是SNP (焦點都放在polymorphism的獨特性)
- 對於未來，科學家將會開始強調關於novel biology的方向，分別介紹如下

1. Whole-genome sequencing

- 目前在做GWAS時，所使用的SNP都是minor allele frequency>5%者(common variation)，但這樣就無法探討rare variants，但這些rare variants更有可能是具有某種功能或是跟疾病有相關的
- 因此下一步可能就會計畫針對人類的whole-genome做sequencing，而不是只挑某些數量的SNP來分析相關性

2. Structural variation

- 遺傳結構變異最常發生的型態就是copy number variation (CNV)
- CNV發生的長度大小可能跟它所對應的gene的功能有關，可能會影響基因的功能或表現
- CNV已漸漸被發現與人類的疾病有關，但若要以CNV來取代SNP來探討遺傳與疾病之相關性的話，其研究方法目前尚在發展的階段

3. Expression

- mRNA或protein的表現量是遺傳相關研究所欲探討的另一個重點
- DNA序列的改變會影響到mRNA和protein的表現進而影響到基因的功能，而此影響可能與疾病有關

4. Epigenetics

- 另一個在遺傳相關性研究中所欲探討的新議題就是epigenetic
- Epigenetic是指不改變核苷酸序列，而改變生物體phenotype的調控方式，也就是非遺傳因子而引起的基因表現差異(例如甲基化…等)

Integrative statistical approaches

- 若是要採用新的biological targets來進行遺傳相關性研究(genetic association analysis)，則除了上述所介紹的測量方式和研究方法需改變外，統計方法的部分也需作整合，如此才能配合所測量的變項做最恰當的分析

The End~