

Team 3

類風濕關節炎單核甘酸多型性資料

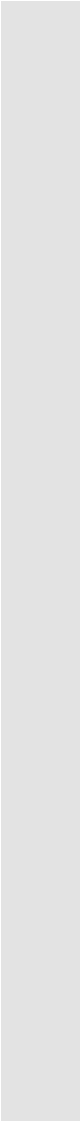

組員：
統計系107級
陳育婷
彭科榮
廖柏鈞
王嘉嘉
吳竹祐

指導教授：
鄭順林 教授

目錄

- 資料介紹
- 文獻回顧
- 資料分析
 - ✓ 1st 資料分析
 - ✓ 2nd 資料分析
 - ✓ 3rd 資料分析
 - 1. recode snps
 - 2. cluster trial
- 分析內容補充及結論-
 - 1. cluster additive/dominant coding
 - 2. 試著找出交互作用
- Appendix(User Time)
- Reference

資料介紹

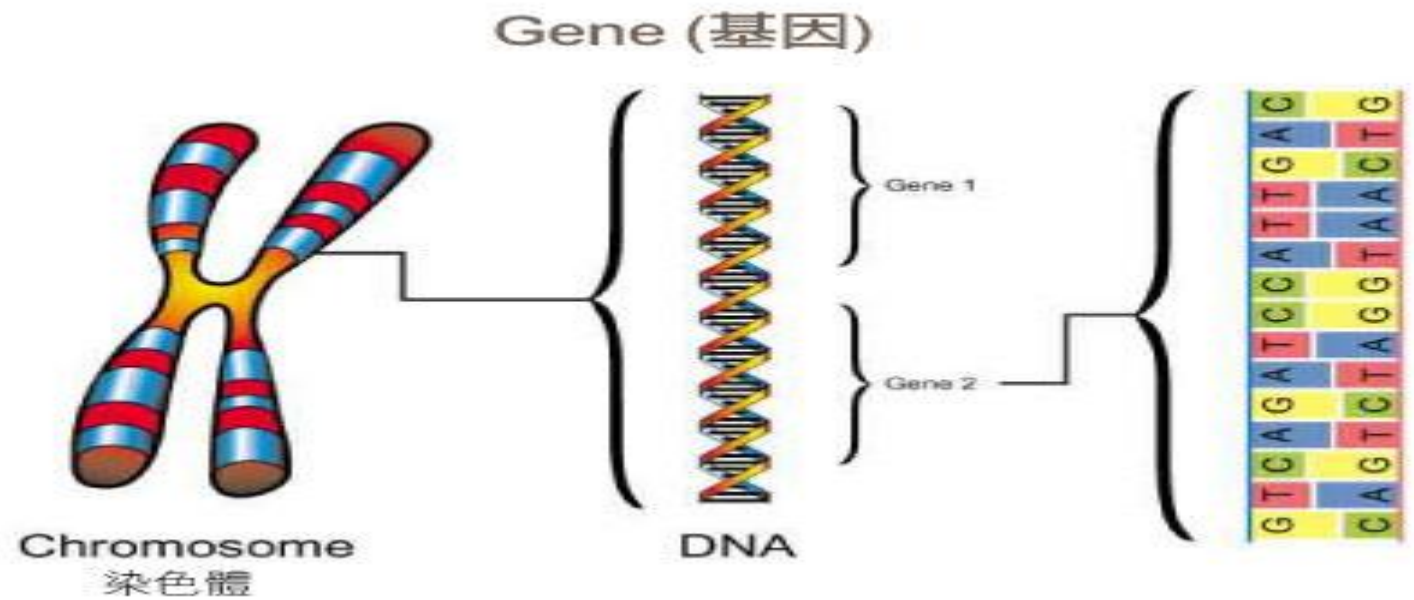


單核苷酸多型性 (Single-Nucleotide Polymorphism, SNP)

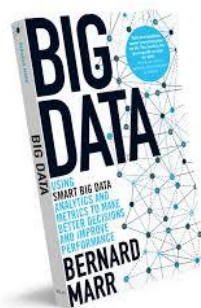
單核苷酸多型性 SNP (Single- Nucleotide Polymorphism)

- 在一族群中，基因組是存在著DNA序列差異性，而單核苷酸多型性是最普遍發生的一種遺傳變異。
- SNP意指DNA序列中的單一鹼基對變異
→ DNA序列中A、T、C、G的改變

DNA序列



圖書館



英文書



A B C D E F G
H I J K L M
N O P Q R S T
U V W X Y Z

英文字母

SNP的DNA 序列差異性

基因上的一位點出現兩個或多個的核苷酸
如圖：

序列多型性：

5'----AGAC**T**AG**A**CATT----3'

5'----AGAT**T**AG**G**CATT----3'

單核苷酸多型性 SNP (Single -Nucleotide Polymorphism)

- 藉由檢測SNP或SNPs之組合與特定疾病、病患的顯著相關性，使疾病的診療將變得更針對性、更個別化。

資料型態介紹

| | ID | Affection | Sex | DRB1_1 | DRB1_2 | SENum | SEStatus | AntiCCP | RFUW | MitoT217C | MitoG228A | MitoG247A | rs3094315 | rs12562034 |
|----|----------|-----------|-----|--------|--------|-------|----------|---------|------|-----------|-----------|-----------|-----------|------------|
| 1 | 10001201 | 1 | F | 0101 | 3 | SN | yes | 27.7 | 54 | A_A | G_G | G_G | A_A | A_A |
| 2 | 10002 | 1 | F | ? | ? | ? | ? | 381.313 | 34 | A_A | G_G | G_G | A_A | A_G |
| 3 | 10004201 | 1 | M | 0401 | 0401 | SS | yes | 103.7 | 207 | A_A | G_G | G_G | A_G | G_G |
| 4 | 10005201 | 1 | F | 0404 | 8 | SN | yes | 42 | 428 | A_A | G_G | G_G | A_A | G_G |
| 5 | 10006201 | 1 | M | 0404 | 2 | SN | yes | 157.6 | 886 | A_A | G_G | G_G | A_A | G_G |
| 6 | 10007202 | 1 | F | 0401 | 0404 | SS | yes | 91.1 | 10 | A_A | G_G | G_G | A_A | G_G |
| 7 | 10009201 | 1 | F | 0101 | 0404 | SS | yes | 82.9 | 1510 | A_A | G_G | G_G | A_G | G_G |
| 8 | 10010201 | 1 | F | 0401 | 3 | SN | yes | 95.8 | 2235 | A_A | G_G | ?_? | A_A | G_G |
| 9 | 10011201 | 1 | F | 0401 | 7 | SN | yes | 55.3 | 994 | A_A | G_G | G_G | A_A | A_G |
| 10 | 10012201 | 1 | F | 0401 | 0404 | SS | yes | 157.4 | 81 | G_G | G_G | G_G | A_G | G_G |
| 11 | 10013201 | 1 | F | 0401 | 0404 | SS | yes | 85.9 | 179 | A_A | G_G | G_G | A_A | G_G |
| 12 | 10014201 | 1 | F | 0401 | 1001 | SS | yes | 149.3 | 267 | A_A | G_G | G_G | A_A | G_G |
| 13 | 10017201 | 1 | F | 0401 | 1401 | SN | yes | 344.5 | 670 | A_A | G_G | G_G | A_A | G_G |
| 14 | 10018200 | 1 | F | 0401 | 3 | SN | yes | 20.5 | 63 | ?_? | G_G | G_G | G_G | G_G |
| 15 | 10020201 | 1 | F | 0401 | 3 | SN | yes | 21.9 | 12 | A_A | G_G | G_G | A_A | G_G |

Showing 1 to 16 of 2,062 entries

2062名觀察者 vs 545089之變數

變數種類

| | scale of measure | Data |
|-----------|------------------|---|
| Affection | ordinal | "0" , "1" |
| Sex | nominal | "F" , "M" |
| DRB1_1 | nominal | "0101", "0102", "0401" , "?" |
| DRB1_2 | nominal | "0101", "0102", "0401" , "?" |
| SENum | ordinal | "NN" , "SN" , "SS" , "?" |
| SEStatus | ordinal | "yes" , "no" , "?" |
| AntiCCP | ratio | 10~3255.00 , "?" |
| RFUW | ratio | 10~6920.00 , "?" |
| MitoX***X | nominal | "?_?" , "A_A" , "G_G" |
| rs***** | Nominal | "?_?" , "A_A" , "A_C" , "C_C" , "C_G".... |

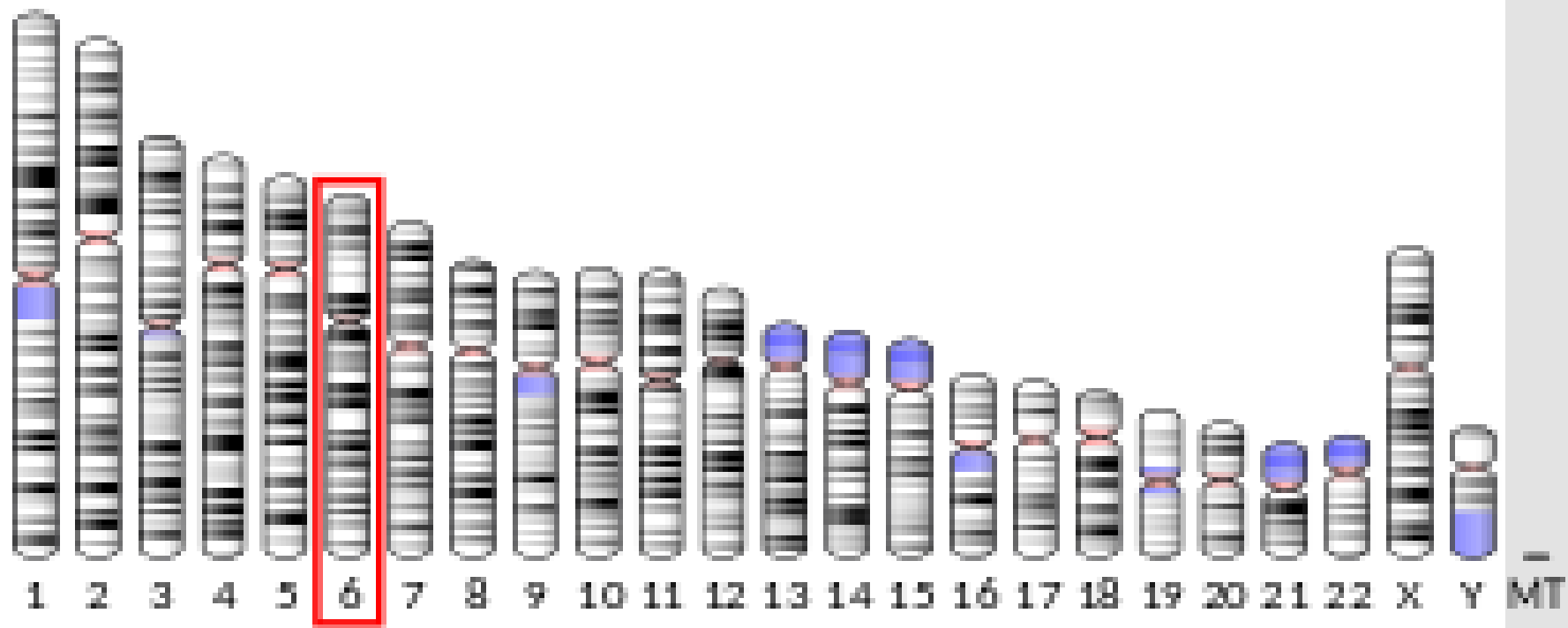
| 類風溼關節炎 | 沒有 | 有 | 總數 |
|--------|------|-----|------|
| 人數 | 1194 | 868 | 2062 |
| 性別 | 女 | 男 | 總數 |
| 人數 | 1493 | 569 | 2062 |

| 性別/有無得病 | 沒有 | 有 | 總數 |
|---------|------|-----|------|
| 女 | 852 | 641 | 1493 |
| 男 | 342 | 227 | 569 |
| 總數 | 1194 | 868 | 2062 |

| 性別/有無得病 | 沒有 | 有 |
|------------|-------|------|
| 女 | 0.57 | 0.43 |
| 男 | 0.60 | 0.40 |
| Odds ratio | 0.882 | |

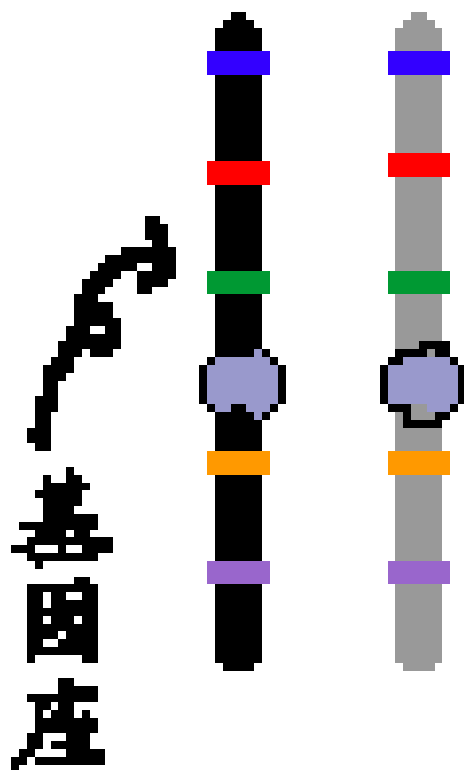
DRB1_1
DRB1_2

● 於染色體第六號基因座上的等位基因編碼



等位基因編碼

一對同源染色體



(A, a)
(B, b)
(C, C2, C3)
(D, d)
(E, e)

對偶基因

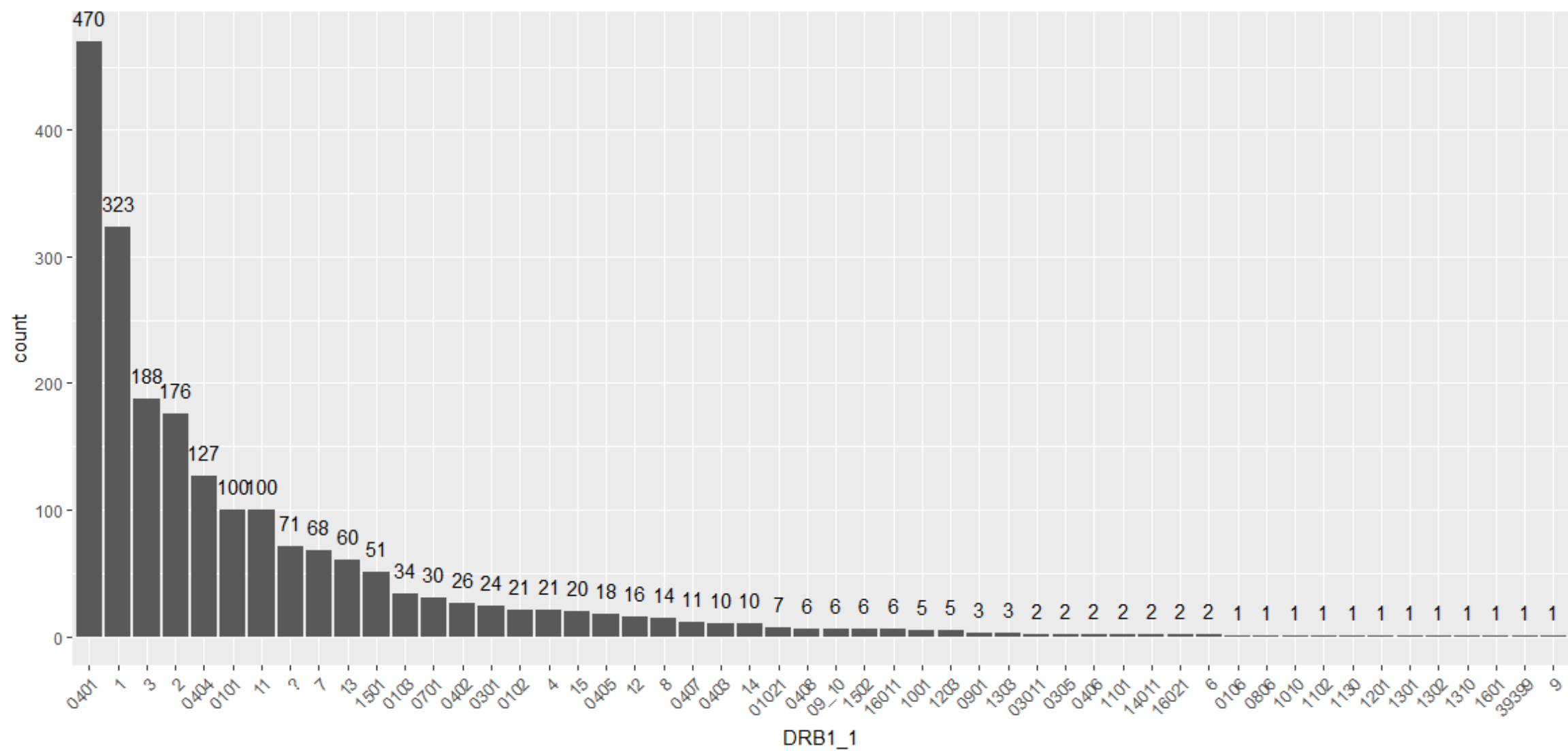


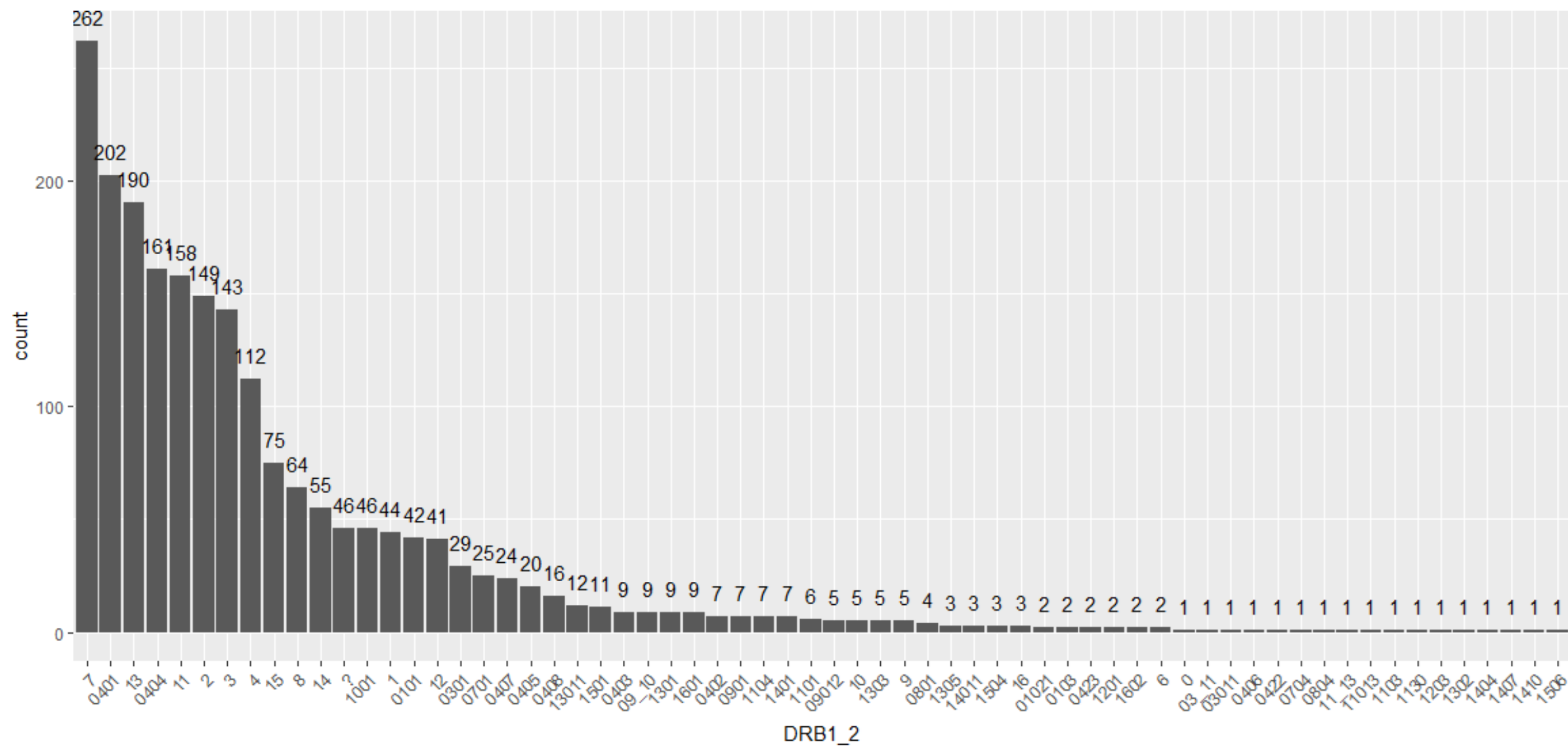
| DRB1_1 | DRB1_2 |
|--------|--------|
| 0101 | 0401 |
| 0101 | 7 |
| 0101 | 11 |
| 0101 | 2 |
| 0101 | 7 |
| 0101 | 0403 |
| 0101 | 3 |
| 0101 | 0101 |
| 0101 | 2 |
| 0101 | 0404 |
| 0101 | 7 |
| 0101 | 0101 |
| 0101 | 14 |
| 0101 | 0101 |



⋮

表現型





SENum

- SENum：基因座具有危險基因型之數量
- 危險基因型等位基因編碼：0101, 0102, 0104, 0105, 0401, 0404, 0405, 0408, 0409, 1001, 1402, 1406

| SENum/Affection | No | Yes | 總數 |
|-----------------|----------------|----------------|------|
| NN | 652 (0.973) | 18 (0.0267) | 670 |
| SN | 464 (0.495) | 474 (0.505) | 938 |
| SS | 77 (0.186) | 336 (0.814) | 413 |
| 總數 | 1193 | 828 | 2021 |
| ?: 41 | | | |

(NN = 0, SN = 1, SS = 2)

SEStatus

- 表現型，症狀之表現狀況(Yes, No)

| SEStatus/Affection | No | Yes | 總數 |
|--------------------|----------------|---------------|------|
| No | 652 (0.024) | 18 (0.976) | 670 |
| Yes | 541 (0.4) | 810 (0.6) | 1351 |
| 總數 | 1193 | 828 | 2021 |

AntiCCP

● 適用於類風濕性關節炎鑑別診斷的一種血清學檢查。

● 小於Negative

介於7-10 Equivocal

大於10 Positive

Min. : 20.05

1st Qu.: 75.78

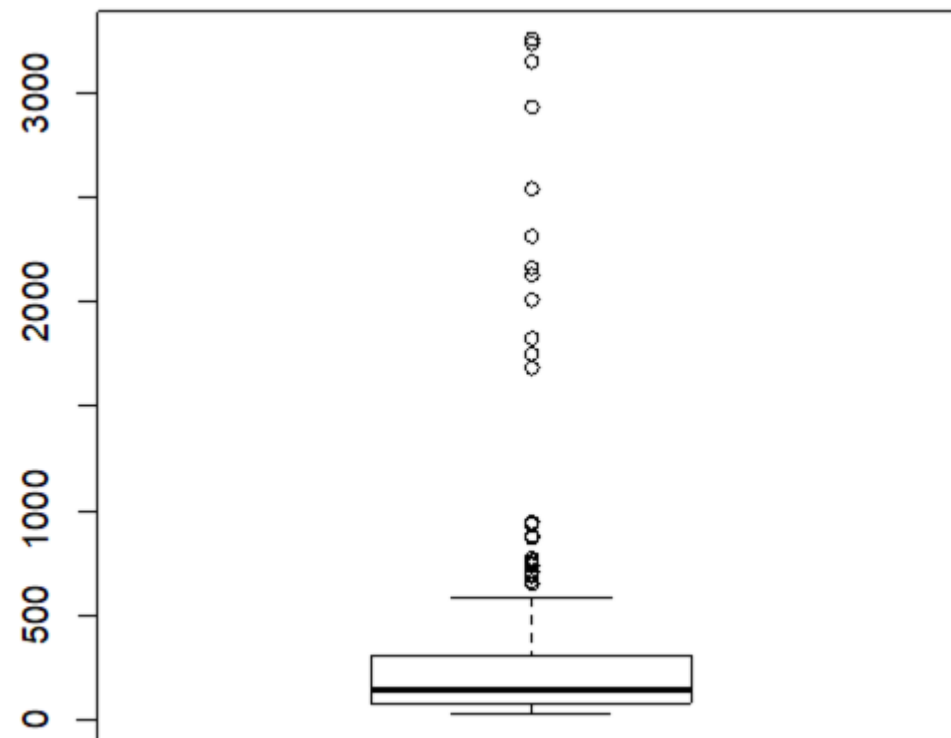
Median : 135.00

Mean : 217.58

3rd Qu.: 295.57

Max. : 3255.00

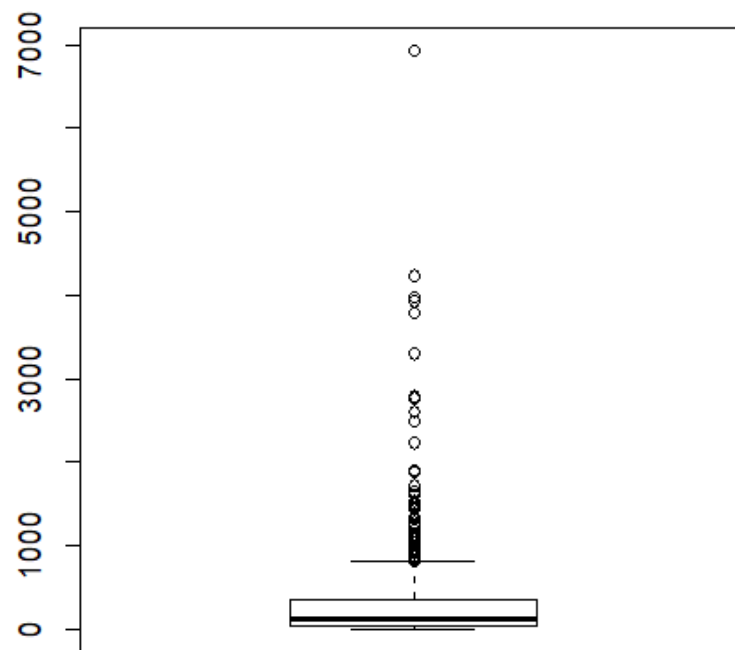
將遺失值(未罹患者)刪除後之盒狀圖



RFUW

- 類風濕性關節炎因子
- 大約有75%的類風溼性關節炎病人可偵測到IgM型的類風溼性因子。

Min. : 9.0
1st Qu.: 50.0
Median : 124.0
Mean : 299.5
3rd Qu.: 354.0
Max. : 6920.0



MitoX***X

- 粒線體具有自己的一套遺傳密碼 (稱為Mitochondria DNA，或mtDNA)，與人類個體細胞核中的遺傳訊息 (細胞核DNA，稱為Nuclear DNA，簡稱nDNA) 不太一樣，主要差異如下表：

| | nDNA | mtDNA |
|------------|--------------|---------------|
| 存在處 | 細胞核 | 粒線體 |
| 數量 | 染色體數,固定 | 數個-千個,視細胞種類而異 |
| 基因數 | 約3萬 | 37 |
| 核苷酸(鹼基)對數 | 約30億 | 約16500 |
| DNA形狀 | 線狀 | 環狀 |
| AUA 密碼子 | Isoleucine | Methionine |
| AGA/AGG密碼子 | Arginine | 終止密碼子 |
| UGA密碼子 | 終止密碼子 | Tryptophan |
| 突變頻率 | 較少 | 較多 |
| 分裂 | 有絲分裂或減數分裂 | 隨機複製 |
| 遺傳方式 | 父母各半 | 母親 |
| 遺傳異常疾病 | 依孟德爾遺傳律,有顯隱性 | 影響不同細胞,多樣性 |

rs*****

| rs3094315 | rs1256203 | rs3934834 | rs9442372 | rs3737728 | rs1126058 | rs6687776 | rs9651273 | rs4970405 | rs1272625 | rs1180784 | rs9442373 |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| A_A | A_A | G_G | G_G | A_G | G_G | G_G | G_G | A_A | A_A | A_G | A_C |
| A_A | G_G | A_G | ?_? | A_G | G_G | G_G | A_A | A_A | A_A | G_G | C_C |
| A_A | G_G | G_G | G_G | G_G | G_G | G_G | G_G | A_A | A_A | A_A | A_A |
| A_A | G_G | A_G | A_A | A_G | G_G | A_G | G_G | A_A | A_A | A_A | A_A |
| G_G | G_G | A_G | A_A | A_G | G_G | G_G | A_G | A_A | A_A | A_G | A_C |
| A_A | G_G | G_G | ?_? | A_G | G_G | A_G | A_G | A_G | A_G | G_G | C_C |
| A_G | A_G | G_G | G_G | G_G | G_G | G_G | G_G | A_A | A_A | A_A | A_A |
| A_G | G_G | G_G | G_G | G_G | G_G | G_G | G_G | A_A | A_A | A_A | A_A |
| A_A | G_G | G_G | G_G | G_G | G_G | G_G | G_G | A_A | A_A | A_A | A_A |
| A_A | A_G | A_G | A_G | G_G | G_G | A_G | G_G | A_G | A_G | A_G | A_C |
| A_A | G_G | A_G | A_G | A_G | G_G | G_G | A_G | A_A | A_A | A_G | A_C |
| A_G | A_G | G_G | A_G | A_G | G_G | G_G | A_G | A_A | A_A | A_A | A_A |

文獻回顧

相關文獻閱讀與報告

大數據的傲慢與偏見

STAT107 吳竹祐

國立成功大學統計學研究所

碩士論文

高維度資料的交互作用模型建立

- 應用於單一核甘酸多型性的全基因相關性研究和
拉曼光譜研究

研究生: 傅馨儀

指導教授: 鄭順林 博士

問題和資料

- Discrete Type: SNP data
- SNP: a site in the DNA sequence where individuals differ at a single DNA base.

For example:

AAGCCTA



AAGCTTA

問題和資料

- Small n (observations) but large p (variables)
- Data Preprocessing
- SNP data(Discrete Variables)
 - Chromosome 6 is considered that it contains the most signals for RA (rheumatoid arthritis)
- RS data(Continuous Variables)

使用方法

- SNP data(discrete variables):
 - First Stage: MDR and logicFS
 - Second Stage: logistic group LASSO
 - Goal: select important terms of the model and build a prediction model

Note that adjusted multi-factor dimensionality reduction (MDR) is abbreviated as “MDR”.

Note that feature selection with logic regression is abbreviated as “logicFS”.

Note that logistic group LASSO regression is abbreviated as “logistic group LASSO”.

使用方法

- RS data(continuous variables):
 - First Stage: MARS
 - Second Stage: logistic group LASSO regression (logistic group LASSO)
 - Goal: select important terms of the model and build a prediction model

Note that multivariate adaptive regression splines is abbreviated as "MARS".

Note that logistic group LASSO regression is abbreviated as "logistic group LASSO".

使用方法

- **Multi-Factor Dimensionality Reduction**

- Reduces the dimensionality of multi-locus information

- **Logic Regression**

- Adaptive regression methodology
- Let Y be a response variable, the logic regression model is

$$g(E[Y|\mathbf{X}]) = \beta_0 + \sum_{i=1}^k \beta_i L_i(\mathbf{X}),$$

where g is an appropriate link function, \mathbf{X} are variables, β_i , $i = 0, 1, \dots, k$, are unknown regression parameters, and L_i , $i = 1, 2, \dots, k$, represents a logic combination of binaries. Since our interest is a case-control study, we assume g to be the logit function, i.e., $g(E[Y|\mathbf{X}]) = \log(E[Y|\mathbf{X}]/(1 - E[Y|\mathbf{X}]))$.

使用方法

- **Multivariate Adaptive Regression Splines**

- $y = f(x_1, x_2, \dots, x_p) + \varepsilon_i$

- where the error term is a mean zero random variable.

- **Logistic Group Lasso Regression**

- **BIMBAM**

- For SNPs with missing data, we used BIMBAM v0.99 to impute them.

效果評估 (4.1.2)

- They select features with the average training error less than 0.35.
- Because of computational limit, we considered the first 1,000 features with the smaller average testing error as our variables at the second stage.

結論

- It is an interesting but hard work to extract useful information from different types of high dimensional data.

結論

- In genome-wide association studies, Ho applied these two methods with sliding window in which only the interaction of the SNPs from neighbor physical positions are considered.
 - Split combination terms into smaller sets.
 - Use sliding windows and combined blocks

結論

- In the continuous type variables, they convert the type of variable from continuous type to discrete type, and then analyze data with the same statistical methods. When they convert a continuous variable to discrete type, it may be not enough to cut it into 3 parts. If they cut it into more parts, the prediction may promote.

結論

- It means that they could get a more accurate judgment when determining some tissue type is normal or abnormal.
- Future Studies
 - SNP data
 - Our results are significant in statistics. To determine whether these results are also significant in biology need more domain knowledge.
 - RS data
 - More random seeds could be used for cross validation.

文獻回顧

國立高雄應用科技大學 資訊工程研究所 碩士論文
運用二元決策樹建構多重單核苷酸多態性

研究生：蔡嘉丞

指導教授：張雲龍 博士、鐘文鈺 博士

名詞解釋

1. 基因頻率

a. 基因型頻率

b. 對偶基因頻率

2. 基因外顯率

3. 哈溫平衡定律

例題

假設針對A/a型SNP進行鑑定者有100人，其中基因型AA者有30人、基因型Aa者有40人、基因型aa者有30人

1. 求AA、Aa、aa基因型頻率？

Ans: $P(AA)=0.3$ 、 $P(Aa)=0.4$ 、 $P(aa)=0.3$

2. 求A、a對偶基因頻率？

Ans: $P(A)=0.5$ 、 $P(a)=0.5$

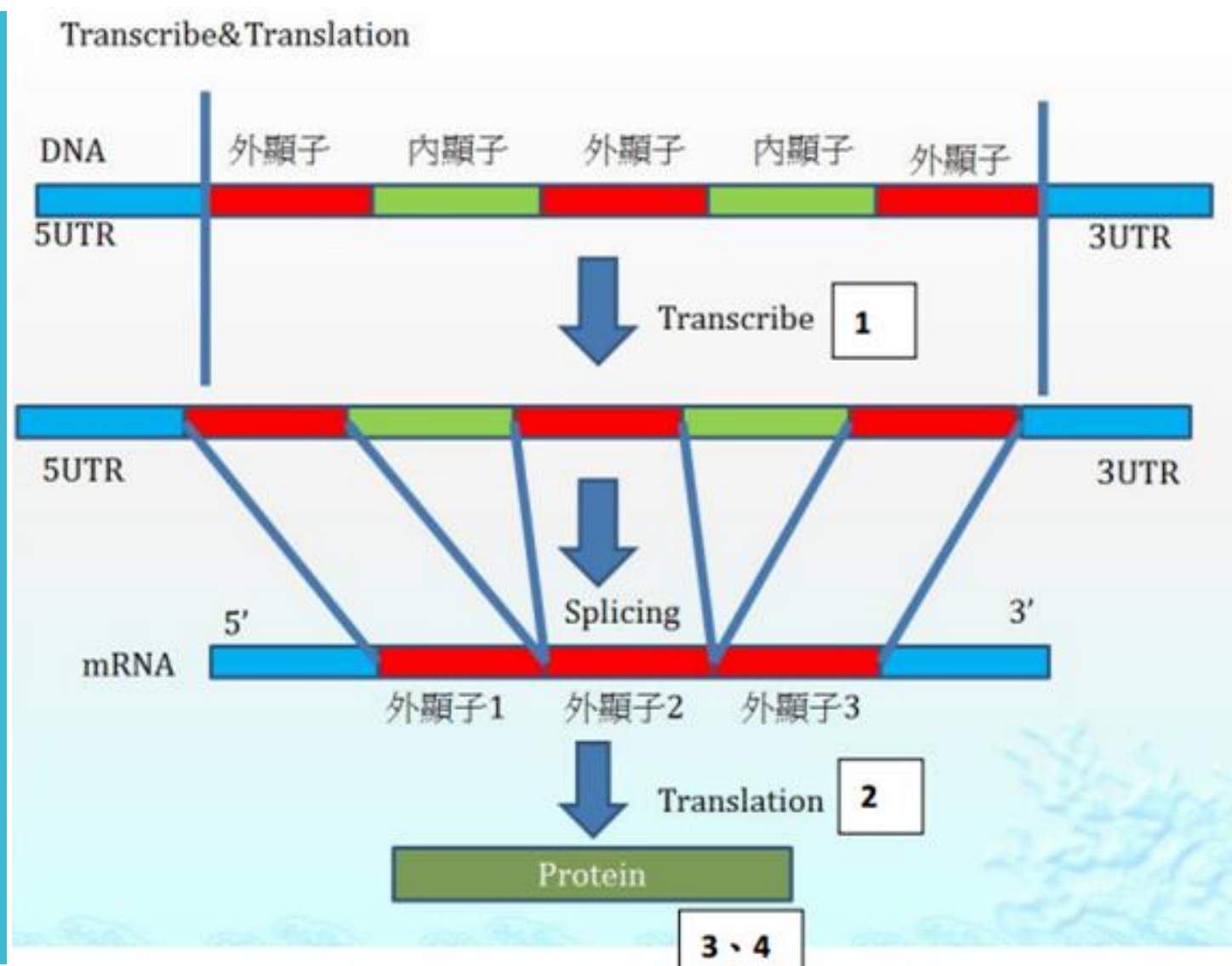
3. 求根據哈溫定律估計的基因型頻率？

Ans: $P(AA)=0.5*0.5=0.25$ 、 $P(aa)=0.5*0.5=0.25$ 、 $P(Aa)=2*0.5*0.5=0.5$

4. 若基因型外顯率 $P(+|AA)=0.8$ 、 $P(+|Aa)=0.8$ 、 $P(+|aa)=0.4$ ，試估計表現出性狀的人數？

Ans: $30*0.8+40*0.8+30*0.4=68$

轉錄與轉譯



次世代定序 (NGS)

- 4大階段：
1. 片斷化
 2. 建庫：
運用聚合酶連鎖反應(PCR)
 3. 高通量定序
 4. 分析

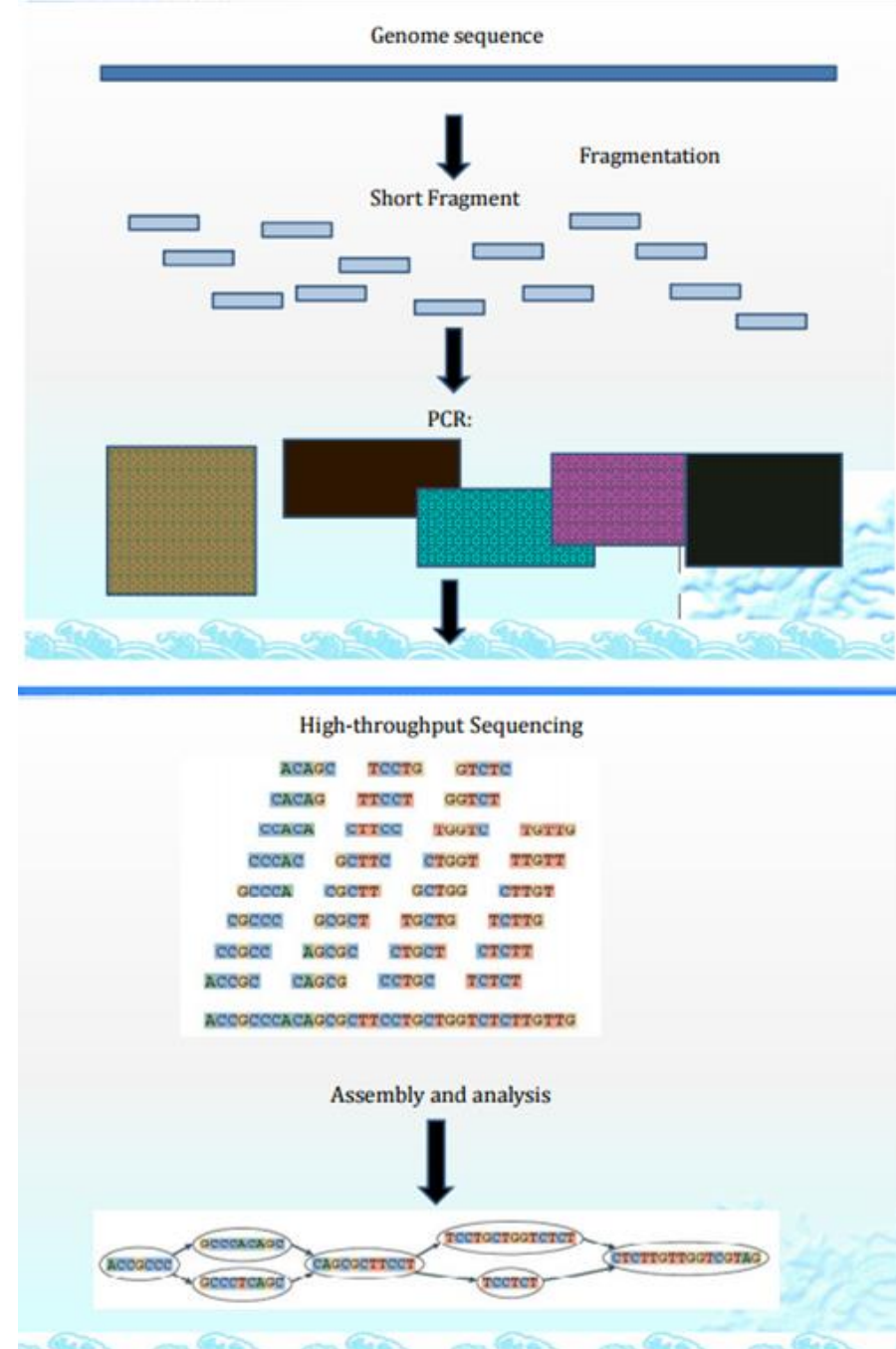


圖 2-5:次世代定序流程圖(NGS)

SNP與單倍型

- 1.SNP：如某段 DNA 序列由 TCGGGCTAC 改變為 TTGGGCTAC，即第二個位置由原本的 C 變成 T，並且此情況在整個族群的發生率大於 1%時，也就是 100 人中至少有一例，便是所謂的 SNP。
- 2.單倍型：單倍體基因型的簡稱，是指在同一染色體上進行共同遺傳的多個基因座上等位基因的組合；通俗的說法就是若干個決定同一性狀的緊密連鎖的基因構成的基因性。

連鎖不平衡

LD

(Linkage Disequilibrium)

- 連鎖不平衡是種描述不同SNP，其核苷酸之間的關聯性，受到染色體影響重組程度。
- 連鎖不平衡指數越高，代表兩組核苷酸序列間重組發生率越低；反之則代表兩組核苷酸間重組率極高。
- 理論上，通常相對距離(單位：分摩根)比較近之SNP，其中間比較不容易有染色體重組，故有較高的連鎖不平衡係數。

公式

1. $D = P_{AB} - P_A P_B$
2. $r^2 = D^2 \div [P_A(1-P_A)P_B(1-P_B)]$
3. TDT檢定(McNemar檢定)：用於親子遺傳

| 染病小孩 基因型 | 父親基因型 | 母親基因型 |
|-------------|-------|-------|
| Aa | AA | Aa |
| AA | Aa | Aa |
| ... | ... | ... |

| | 傳遞 | | |
|-------|-------|-------|-------|
| 不傳遞 | A | a | total |
| A | m1 | m2 | m1+m2 |
| a | m3 | m4 | m3+m4 |
| total | m1+m3 | m2+m4 | 2n |

TDT檢定統計量

H_0 : 兩組基因無連鎖且平衡

(父母傳遞至染病小孩A和a的機率相等)

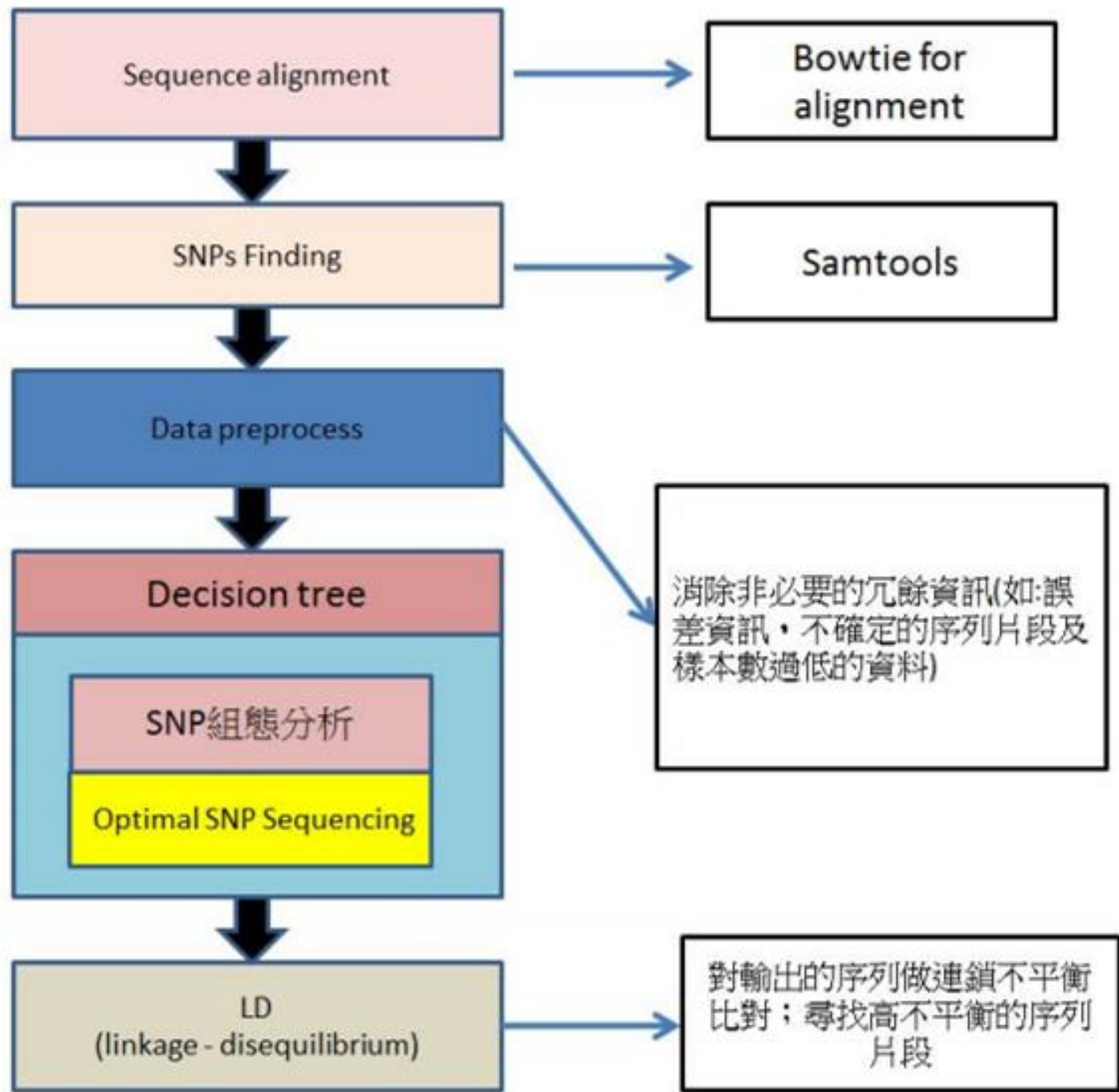
H_1 : 兩組基因有連鎖不平衡

(父母傳遞至染病小孩 A 和 a 的機率不相等)

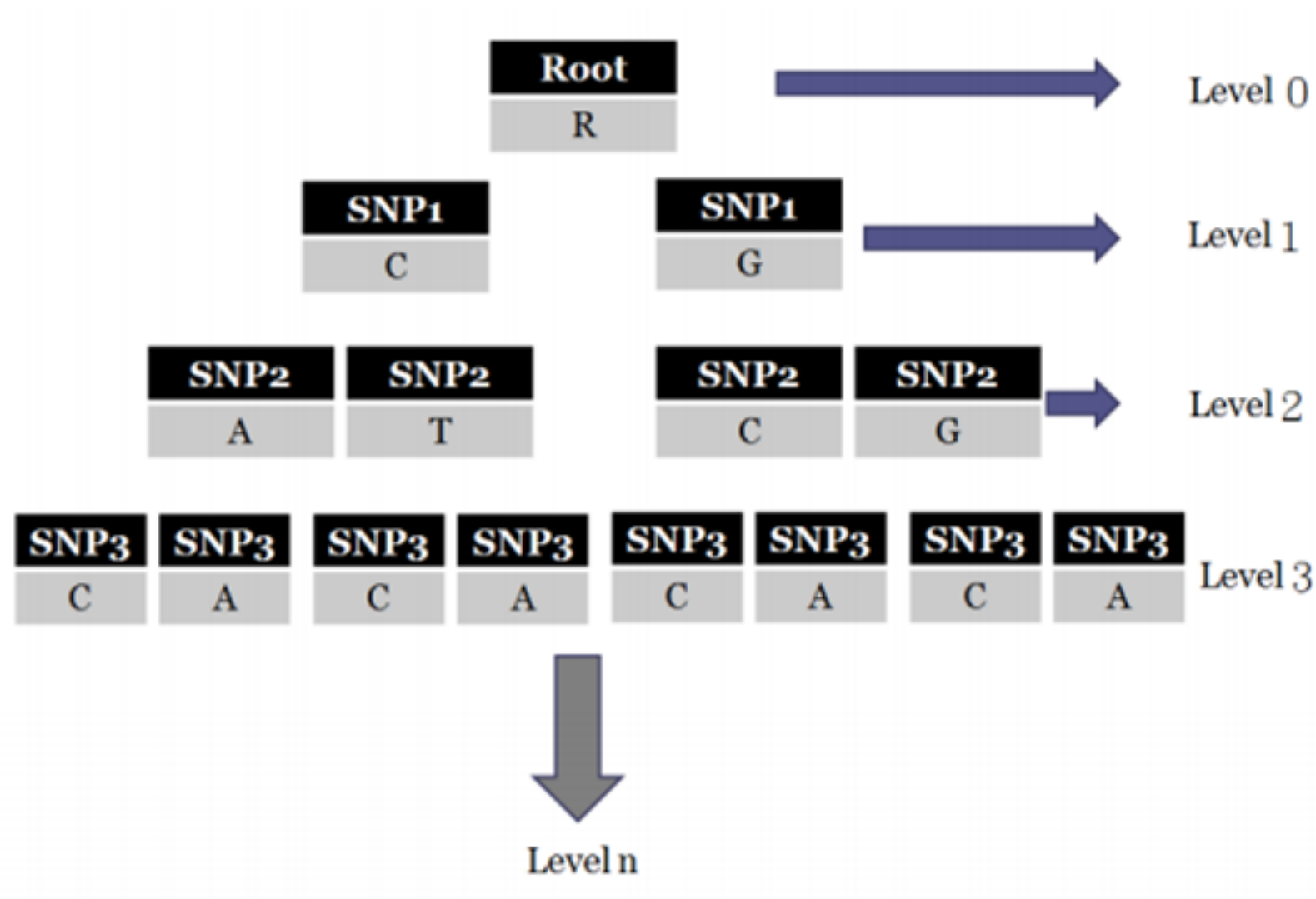
$$TDT = (m_2 - m_3)^2 / (m_2 + m_3)$$

在虛無假設下TDT服從自由度為1的卡方分配

研究方法



決策樹

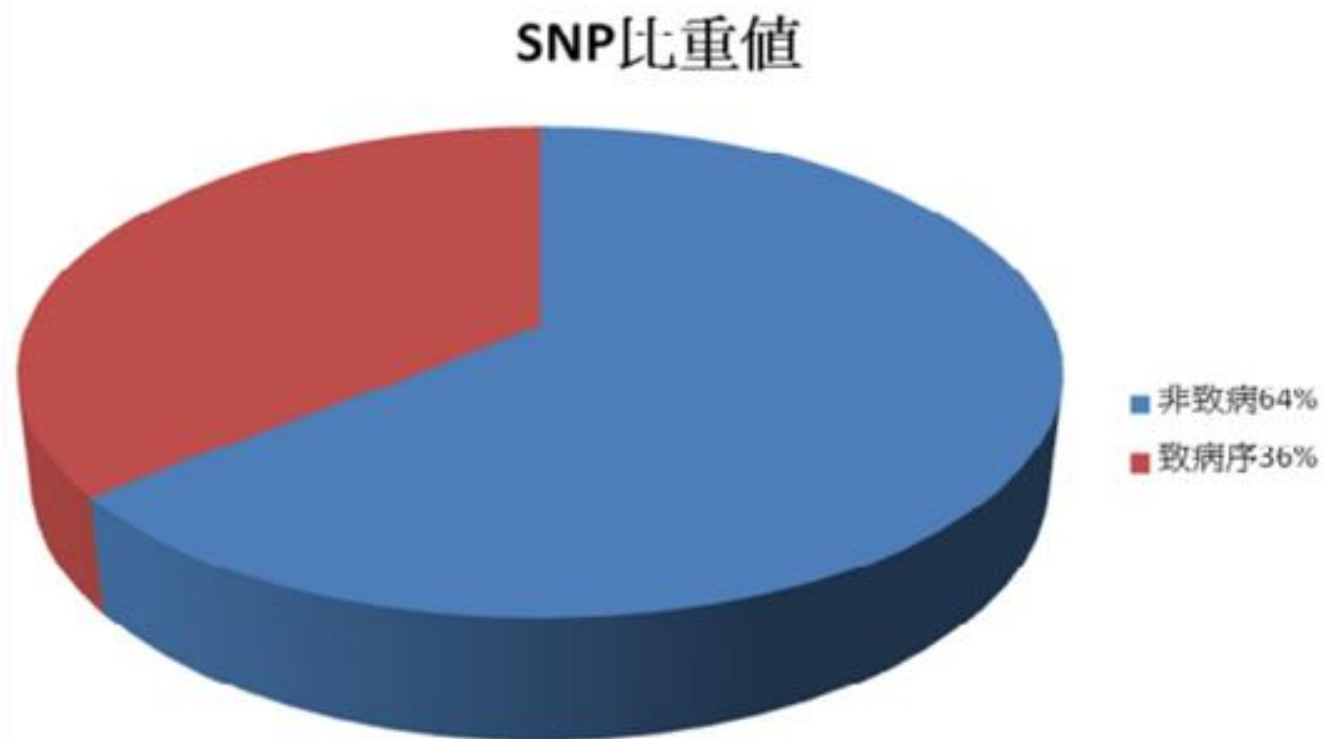



連鎖不平衡分析

我們對決策樹輸出的最佳 SNP 序列，對其依序做連鎖不平衡評估，我們將序列中具 $R^2 > 0.8$ 次數較多之序列做為我們欲搜尋的最佳序列，若沒有或相同次數則找尋 R^2 最大值做為我們的輸出。最後再將序列組合即本實驗的最終輸出結果。

實驗結果

實驗進行後，我們得以得出一條最有可能引起疾病的 **SNP** 序列，經過分析後我們可以知道該樣本資料內的 **DNA** 內發生的 **SNP** 有 64%是非引起疾病的單核苷酸多型性，剩下的 36%則是經由實驗判定的與疾病有所關聯的致病 單型





THANK YOU !

A Genome-Wide Association Study Reveals ARL15, a Novel Non-HLA Susceptibility Gene for Rheumatoid Arthritis in North Indians

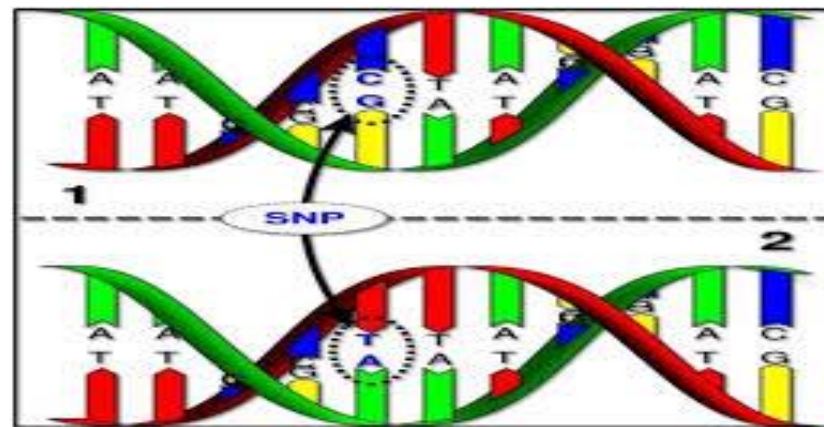
Sapna, Negi, Garima Juyal, Sabyasachi Senapati, Pusuplata Prasad, and
other 17 co-authors

巨量資料分析課程
Team 3 – 相關文獻閱讀與報告
NCKU STAT 107 陳育婷

報告大綱

1. 領域相關知識&專有名詞解釋
2. 研究目的
3. 資料與實驗介紹
4. 資料前處理方法
 - ▣ 基因定序 Genotyping
 - ▣ Quality Control-(1.)MAF<5%
 - (2.)call rate<95%
 - (3.)deviation from Hardy-Weinberg equilibrium
 - (4.)control inflation factor for pop. substructure
5. 分析方法
 - ▣ GAWs-(1.)Cochran-Armitage trend test
 - (2.)Additive Dominant and Recessive Test
 - ▣ SVM
6. 資料處理與分析結果
7. 結論與未來展望

名詞解釋



- **SNP(單核苷酸多態性):**

由單個核苷酸—A,T,C或G的改變而引起的DNA序列的改變，造成染色體基因組的多樣性，且此情況在整個族群的發生率 $>1\%$ 時。例如，來自兩個不同個體的DNA片段，AAGCCTA和AAGCTTA為等位基因。

- **GWAS (Genome-wide association study全基因組關聯分析):**

指在人類全基因組範圍內找出存在的序列變異，即單核苷酸多態性，從中篩選出與疾病相關的SNPs。

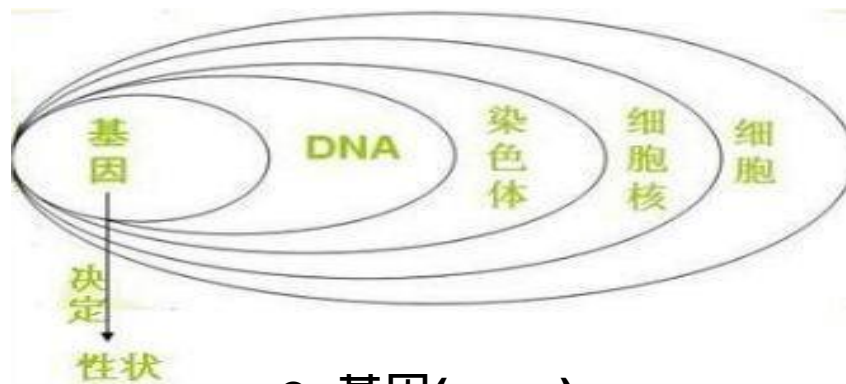
- **Genetic Heterogeneity(遺傳異質性):**

某一種遺傳疾病或表型可以由不同的等位基因或基因突變所引起的現象。

- **Heritability(遺傳力):**

某一性狀受到遺傳控制的程度，介於 0-1之間。 $=1$ ，表型變異完全是由遺傳因素決定； $=0$ ，完全是由環境決定。

遺傳物質 相關名詞解釋



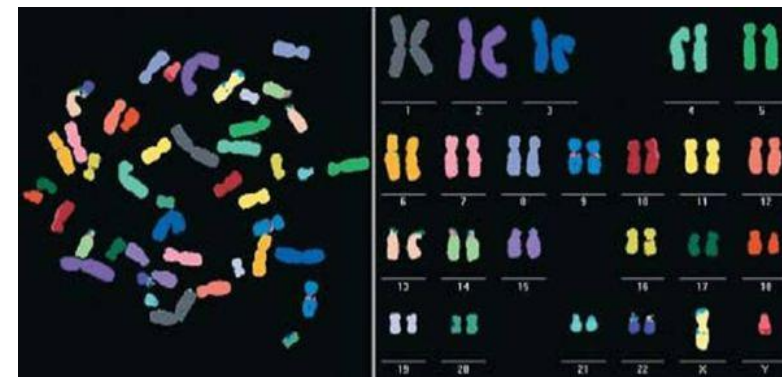
● 基因(gene)

每條DNA分子上存在很多具有遺傳效應的DNA片段，稱為基因



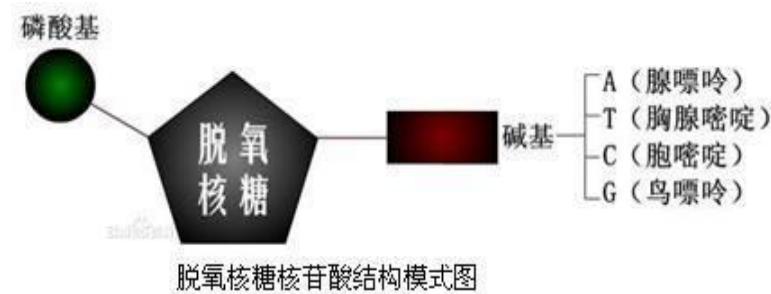
● DNA(去氧核糖核酸)

每條染色體上有一條DNA，DNA是儲存、複製和傳遞遺傳信息的主要物質基礎



● 染色體(chromosome)

細胞核內具有遺傳性質的物體，主要由脫氧核糖核酸 (DNA)與蛋白質組成，每個細胞核內有46條染色體



● 核苷酸(nucleotide)

是基因的基本結構和功能單位，每個基因上含有成百上千個核苷酸分子，一個核苷酸分子由三個分子組成：含氮鹼基、核糖、磷酸

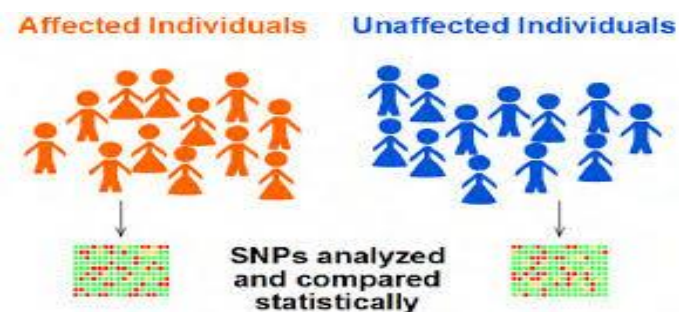
基因組? 基因? 基因座? 基因型? 等位基因?

- **基因組(基因體)(genome)**: 包含在該生物的DNA中的全部遺傳信息，基因組包括基因和非編碼DNA
- **基因(gene)**: 攜帶有遺傳信息的DNA序列，是控制性狀的基本遺傳單位。通過指導蛋白質的合成來表現所攜帶的遺傳信息，從而控制生物個體的性狀(差異)表現
- **基因座(locus)**: 染色體上的固定位置，例如某個基因的所在。而基因座上的DNA序列可能有許多不同變化，各種變化形式稱為等位基因
- **等位基因 (allele)**: 染色體內的基因座的可以複製的DNA序列，其在細胞有絲分裂時的染色體上的兩個基因座是對應排列的，故在細胞遺傳學裡稱其為等位
- **基因型(genotype)**: 該基因所擁有的一對等位基因所決定

A a

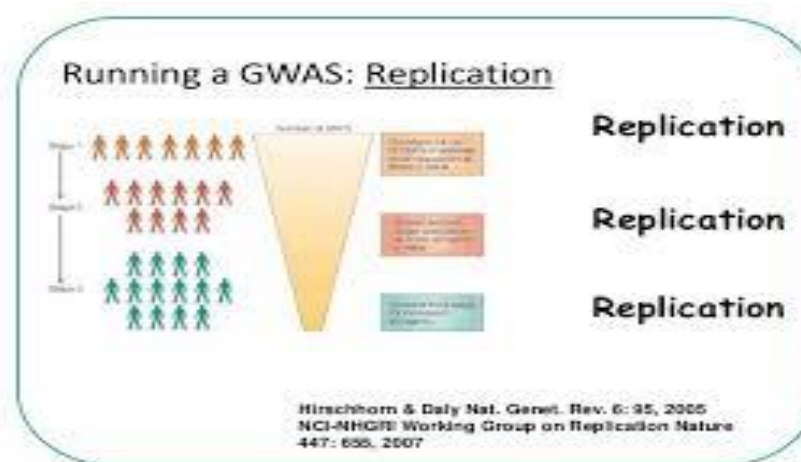
AA Aa aa

研究目的



1. **GWAS**跟之後的**meta-analysis**藉由接連發現與類風溼關節炎相關的致病基因而改變了類風濕關節炎的遺傳學研究角度。
 2. 但是，多數的研究的實驗對象都是**白種人**，只解釋了一小部分的遺傳異質性。為了理解**不同種族間致病基因的差異與共通點**，這份研究是針對**北印地安人**族進行研究。檢查有無新的治病基因，是沒有在之前的多數研究中找到的。
 3. 之前的研究發現，只有**4%的GWAS實驗單位不是白種人**。另外，發現 **PAD14 & PTPN22**，在歐洲人與亞洲人之間，對類風溼關節炎的影響有很大的差異。
- 綜合以上了解到:研究**白種人以外的致病基因非常重要**。不只可以找到種族特有的致病基因，也可以解釋更多的遺傳變異。

資料與實驗 介紹



● 實驗方法:Replication

✓ 第一步

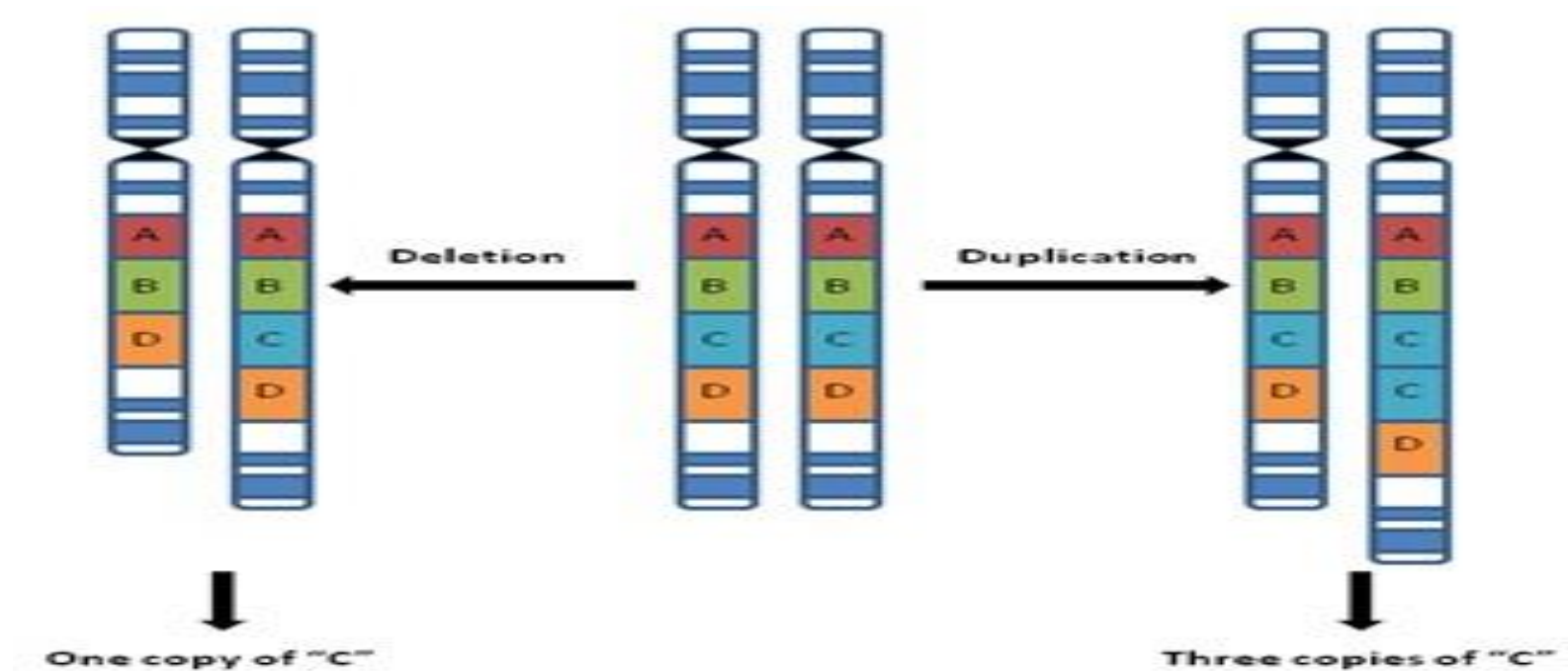
- 將sample分為testing set和replication set兩組
- 針對testing set，檢定大量的SNP
- 然後將最promising的SNP挑出來，並在replication set中重複檢定這些SNP

✓ 第二步

- 在進行replication testing時，會同時使用一些多重比較的校正方法 (如Bonferroni)來校正之

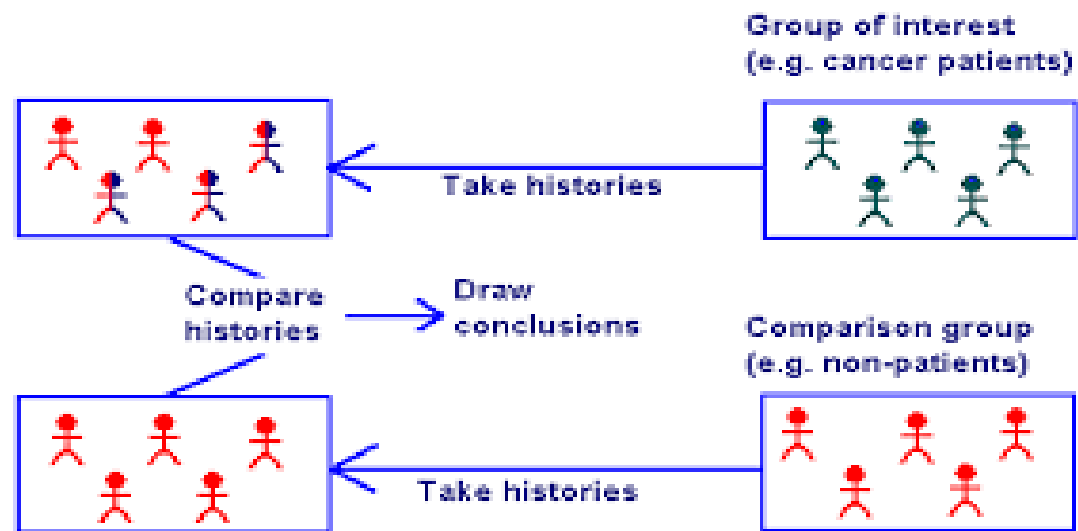
- 目前有許多期刊都會要求在做GWAS時，需要將SNP進行replicate至少一個族群

資料前處理- Genotyping (基因型分型)



1. 用末梢血白細胞(peripheral whole blood)做檢驗。
2. Discovery stage-用Illumina Human660W platform
Replication stage-用多功能核酸質譜分析儀(Sequenom)
3. CNV:基因组變異的一種形式，通常使基因组中大片段的DNA形成非正常的拷貝數量

資料與實驗 介紹



- 試驗類型: Case Control Study
- -discovery stage: 706 RA patients & 761 controls
- -replication stage: 927 RA patients & 1148 controls



資料前處理- Quality Control

1: 缺失比例(SNP with call rate<95%)

舉例: 假如89人， $89 \times 0.05 = 4.45$ 人，假如一SNP在超過 4.45個人以上都是沒被標記出來(marked)，把這些SNP從資料中刪除。

2: 小等位基因頻率 (Minor Allele frequencies) (MAF<0.05)

可以用來排除 資訊不足的SNP，因為它們在樣本的變異太小。

例如: 一SNP中只有1個人與其他人在那個點的鹼基對不同，則將SNP刪除。

3: 移除偏離哈溫平衡的SNP(in the controls, pvalue<10⁻⁷)

哈代-溫伯格定律--一個群體在理想情況，經過多個世代，基因頻率與基因型頻率會保持恆定。

最簡單的例子:

位於單一位點的兩個等位基因：顯性等位基因記為A而隱性等位基因記為a，它們的頻率分別記為p和q。頻率(A) = p；頻率(a) = q； $p + q = 1$ 。如果群體處於平衡狀態，則我們可以得到

群體中AA的頻率(AA) = p^2

群體中aa的頻率(aa) = q^2

群體中Aa的頻率(Aa) = $2pq$

資料前處理- 種族分層

- 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
- Case and Control都要做分層，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

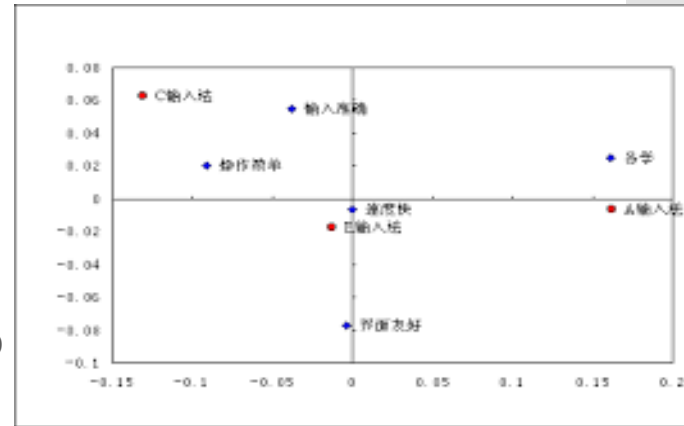
資料前處理- 分層方法

1.多維尺度分析(MDS)

- 一種資料簡化方式，利用資料點之距離與相似性找出最低度的空間重構資料點之相對關係。
- 點與點間距離:變數相似程度；距離越小，越相似。

2.QQplot (每一個 SNP 之expected vs observed p value)

- 先把P值從小到大排序。縱坐標(y):-log₁₀ observed p value，
- 橫坐標(x):-log₁₀(i/n+1)i:第i個大的p value，n:SNP總數。
- 如果出現了偏離的情況說明期望與觀察值有偏差。
- 該SNP點出現了較大的偏離-由該SNP突變的遺傳作用造成的。



↑
MDS
plot

3.GIF λ_{GC} (genomic control inflation factor) chi-square test statistics的中位數/chi square檢定統計量分布之中位數

例如:自由度為 1 的卡方分配中位數為 0.4549，我們得到的資料中檢定出來的所有服從卡方分配的檢定統計量之中位數(ex. $Z^2 \times 2$)為 0.45，那 λ_{GC} 約等於 1。

>1:有問題

分析方法介紹-趨勢分析

1.Cochran-Armitage Trend Test(趨勢分析)

- 用於類別資料分析，目的在檢驗兩變數的關係
- 分析資料型態：
一變數有2類，另一變數為順序尺度，有k類時
- 2*k的列聯表:

| | B = 1 | B = 2 | B = 3 | Sum |
|-------|----------|----------|----------|-------|
| A = 1 | N_{11} | N_{12} | N_{13} | R_1 |
| A = 2 | N_{21} | N_{22} | N_{23} | R_2 |
| Sum | C_1 | C_2 | C_3 | N |

The trend test statistic is

$$T \equiv \sum_{i=1}^k t_i (N_{1i} R_2 - N_{2i} R_1),$$
$$\text{Var}(T) = \frac{R_1 R_2}{N} \left(\sum_{i=1}^k t_i^2 C_i (N - C_i) - 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^k t_i t_j C_i C_j \right)$$
$$\frac{T}{\sqrt{\text{Var}(T)}} \sim N(0, 1)$$

分析方法介紹- 趨勢分析應用

Cochran-Armitage Trend Test(趨勢分析)在遺傳學上的應用:

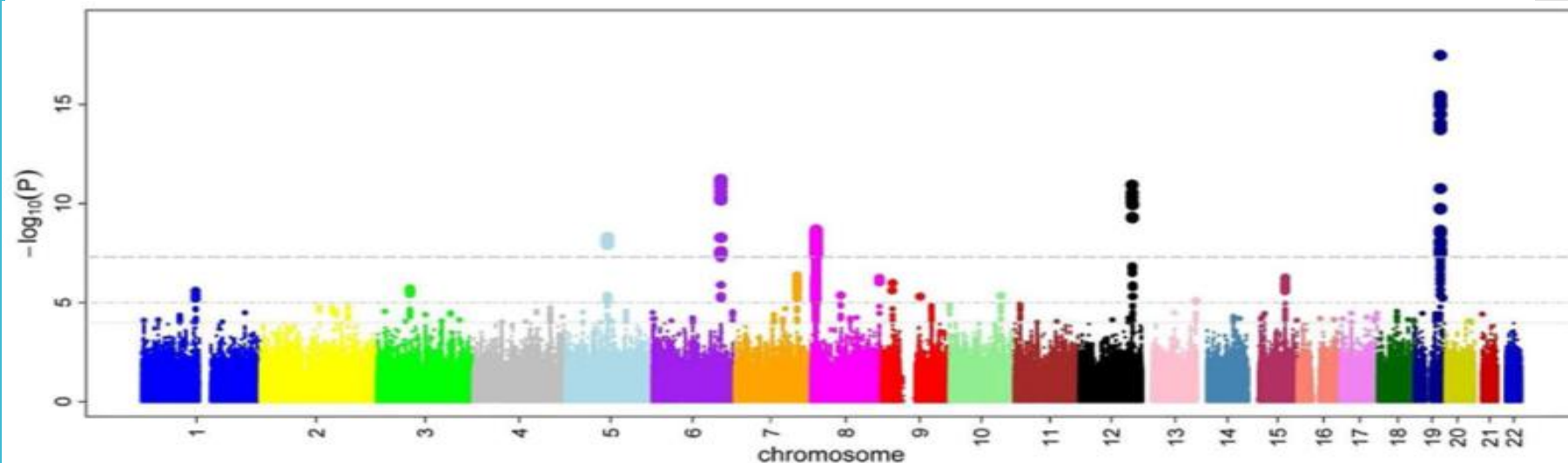
- 三種基因型在實驗組與對照組中的次數做成 3×2 的列聯表。

| | Genotype aa | Genotype Aa | Genotype AA | Sum |
|----------|-------------|-------------|-------------|-----|
| Controls | 20 | 20 | 20 | 60 |
| Cases | 10 | 20 | 30 | 60 |
| Sum | 30 | 40 | 50 | 120 |

- ti的選擇很重要
- 看SNP (GT突變, G 是minor的, 占百分比少; T是Major的, 占百分比高)
對這個位點會得到三種基因分型: GG、GT、TT。
- 分別考慮這三種基因分型或者是G、T的攜帶者是否增加患病風險。
 - 1.當**GG、GT、TT**($t=2,1,0$)在疾病中都有差異的時候我們叫做
additive model累加模型 (有一个G或T, 就增加發病可能)
 - 2.以**GG+GT vs. TT** 這種組合的方式進行分析時叫做
G的dominant mode($t=1,1,0$) ;
 - 3.以**GGvsTT+GT** 為**G的recessive model**($t=0,1,1$) 。
- 多數情況, **minor**的核苷酸是疾病的發病基因。

分析方法介紹-曼哈頓圖

- 幾乎所有 **GWAS** 研究都會附帶一張曼哈頓圖，圖中每個點對應一個 **SNP**，**x 軸** 是 **SNP** 在基因組中的位置，**y 軸** 是 **$-\log_{10}(P \text{ value})$** ，因此點越高越顯著
- 由來:大家覺得高高低低的點很像是曼哈頓的天際線



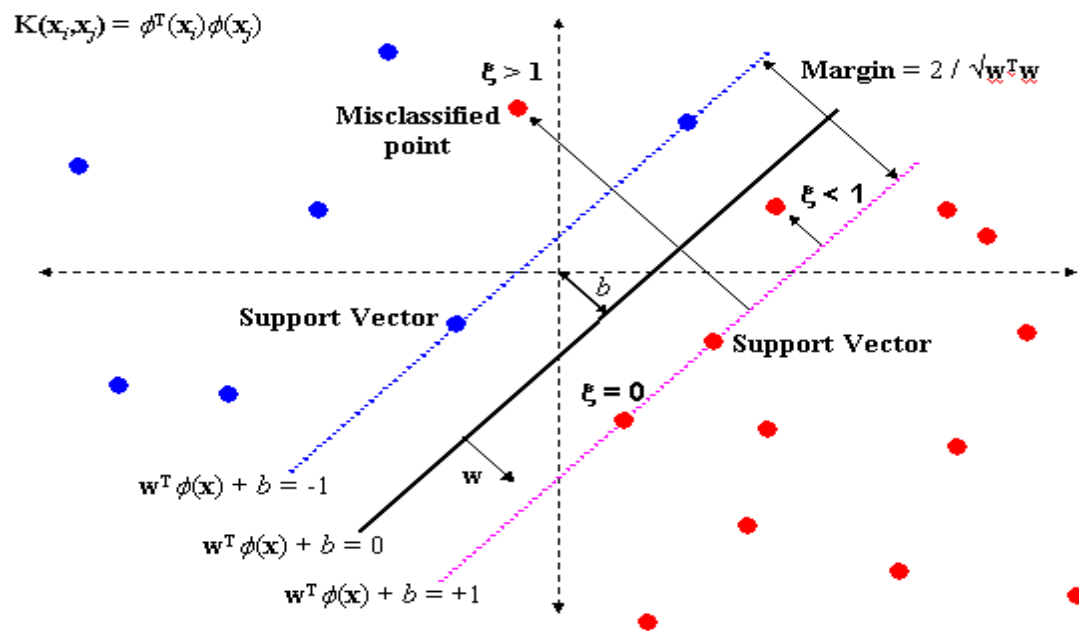
本研究:

Significant association
cutoff value: 5×10^{-8}

Suggestive association
cutoff value: 5×10^{-5}

分析方法介紹-SVM

2.SVM(Support Vector Machine)(支援向量機):



真正最好的「分割超平面」，應該是要離兩邊的點都夠遠才好，也就是所謂的「邊距」(margin) 最大。

SVM簡介

一種監督式學習方法，廣泛地應用於統計分類。在解決小樣本、非線性及高維模式識別問題中表現出許多特有的優勢

SVM想要解決的問題

找出一個超平面(hyperplane)，使之將兩個不同的集合分開。以二維平面來說，我們希望找出一條線能夠將兩種不同的點分開，而且我們還希望這條線距離這兩個集合的邊界越大越好。

分析方法介紹- SVM

- P維的空間裡(變數p個) , $\dim(\text{hyperplane})=p-1$

- Equation: $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p = 0$

1) **Maximum Margin Classifier:** $\underset{\beta_0, \beta_1, \dots, \beta_p}{\text{maximize}} M$

--Seperable Case

$$\text{subject to } \sum_{j=1}^p \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) \geq M$$

for all $i = 1, \dots, N$.

2) **Support Vector Classifier(Soft Margin Classifier):**

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^n \epsilon_i \leq C,$$

3) **Support Vector Machine:**

如果是「非線性可分集合」，可運用核函數(kernel function) 幫我們造出分割超平面

分析方法: SVM應用

SNP data:

●Features:

dim(feature space): n (SNP個數), $\dim(\text{hyperplane})=n-1$

●Response:

patients and controls (2 classes)

●模型:

1)篩SNP(p value <0.001) = 517SNPs

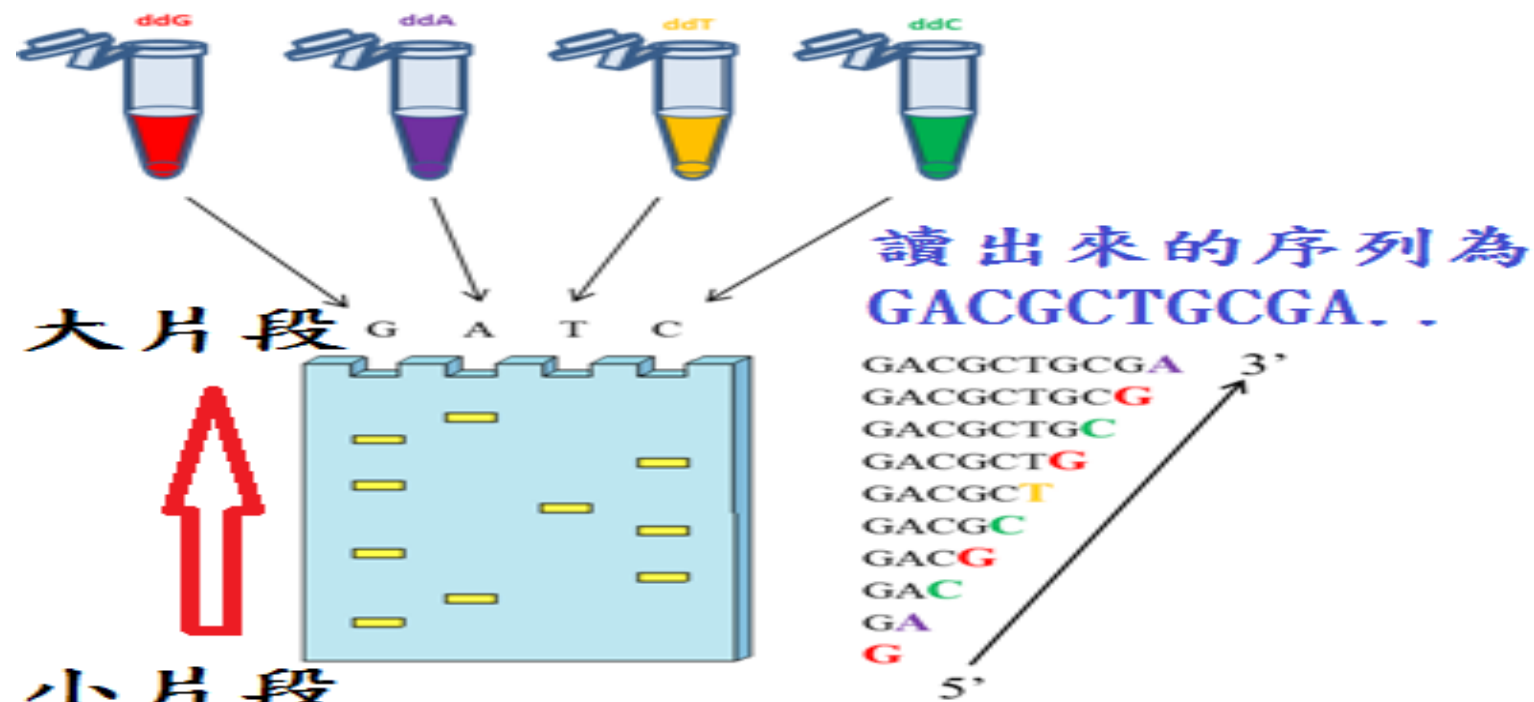
2)100 subsets of data(90% 樣本數): 100 models

3)each model:每一個SNP給一個權重 w_i ，代表此SNP對患病預測的重要性。

4)得到 $|w_i|_{\text{avg}}$, find $|w|_{\text{max}}$ and corresponding σ

資料處理結果- Genotyping(基因型分型)

- 5% SNP 沒有被標記- 定序失誤，或者是因為定序是參考白種人的資料，北印度人有些基因難被定序出來
- 10% SNP 為 CNV，移除
- 剩 559,348 SNPs

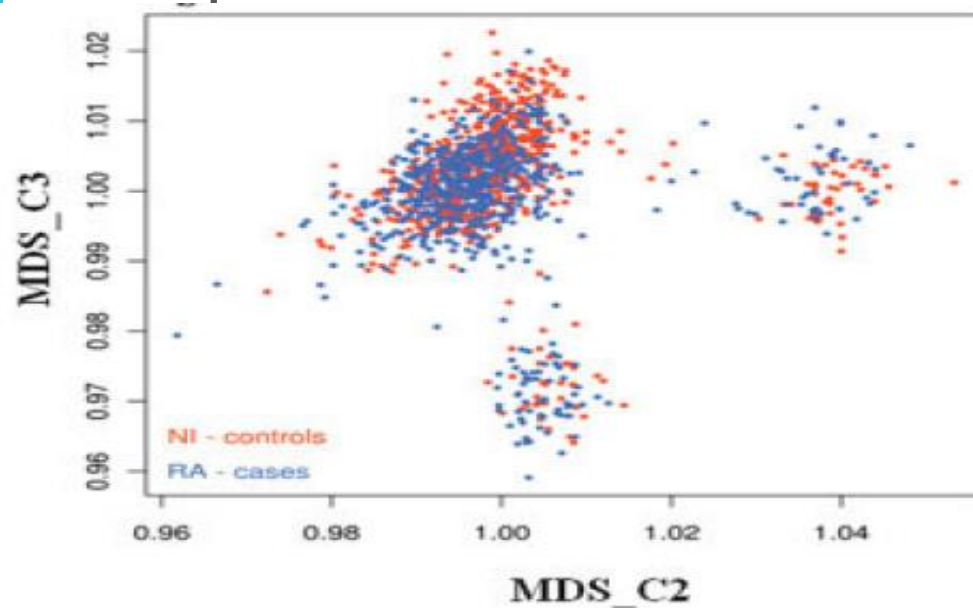


<http://ppt.cc/Xc6V>

資料處理結果- Quality Control

- 移除在第23對染色體(XX or XY)及粒線體上的基因
 - 缺失比例(SNP with call rate<95%)
 - 小等位基因頻率 (Minor Allele frequencies) (MAF<0.05)
 - 移除偏離哈溫平衡的SNP(in the controls, $pvalue < 10^{-7}$)
- 475,771 SNPs in 664 RA patients and 666 controls

- 分層-
- MDS plot



1 主要群-556 patients, 590 controls

2 小群-

- (1) 47 patients and 44 controls
- (2) 61 patients and 32 controls

資料處理結果- Quality Control

●因應之道:

- 1)每個群體在用趨勢分析檢驗與疾病關聯性時分開處理。最後再將 subgroups的p value pooled with main group 的p value，得到pooled p value
- exact-effect meta analysis method
- 2)Correct pooled p value by genomic inflation(λ_{GC})-using
-standard application of PCA or MDS

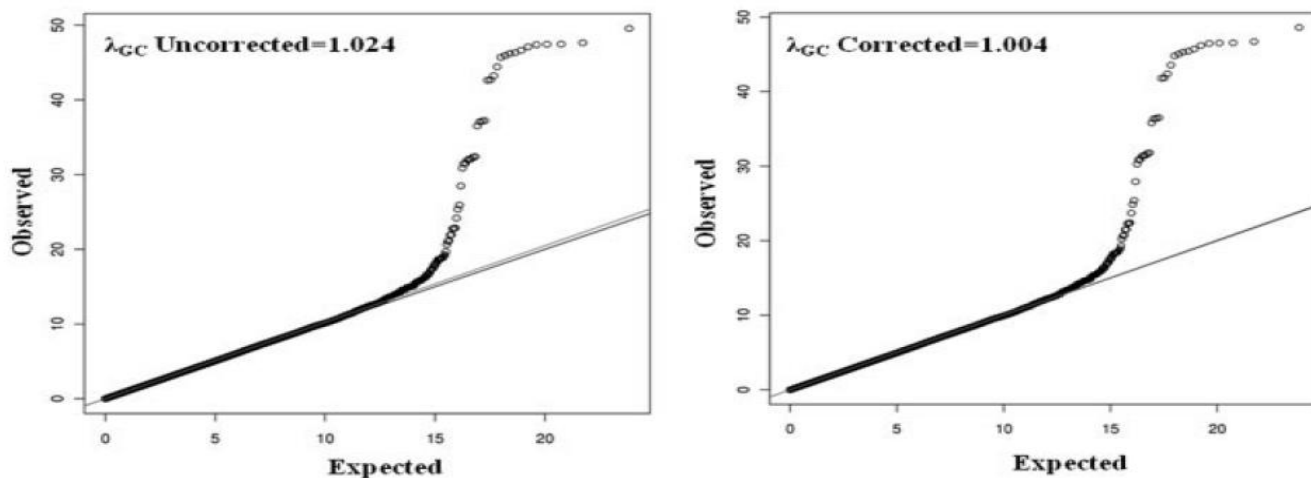
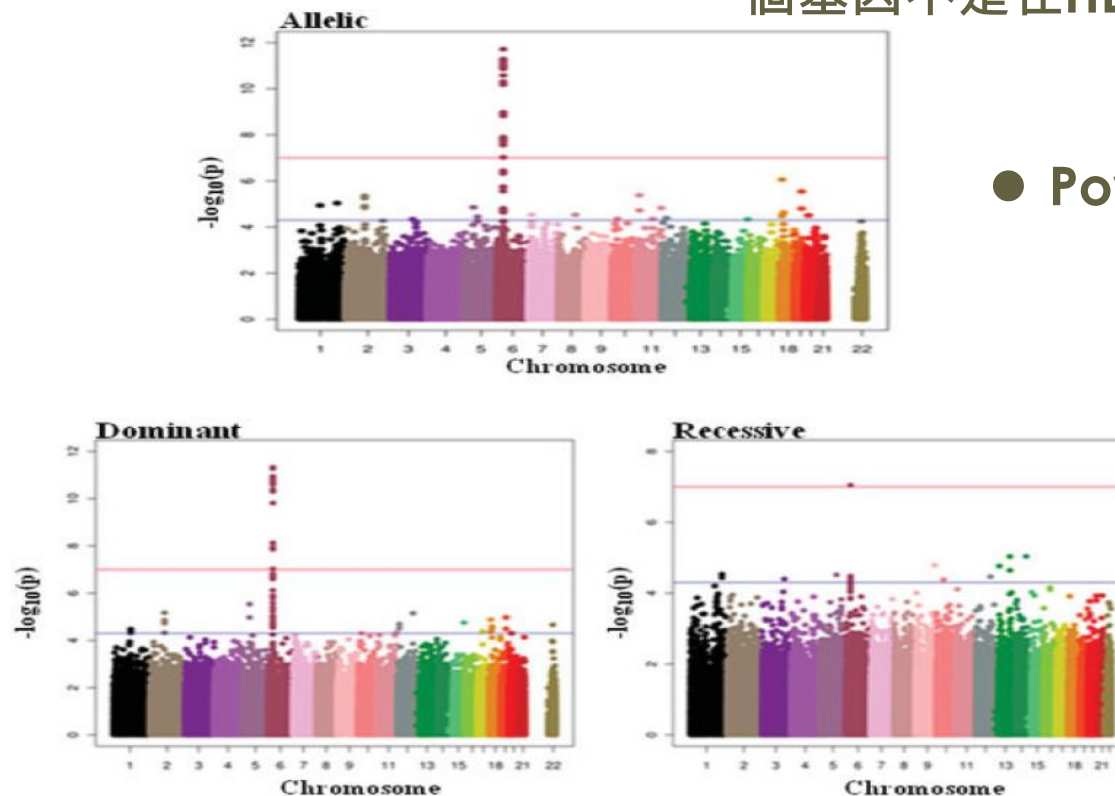


Figure 2. Q-Q plots showing the distribution of all single-nucleotide polymorphisms ($\sim 476,000$) in chi-square statistical analysis and the genomic control inflation factor (λ_{GC}) in the allelic association model.

用Cochran-Armitage Trend Test(趨勢分析)做檢驗，算出每個SNP在additive、dominant、recessive model中的p value

分析結果- Manhattan Plot

- 在第六號染色體上MHC(主要組織相容性複合體基因)是影響發病最重要的基因座。
- 非HLA (MHC醣蛋白，又稱為人類白血球抗原)的基因座對疾病的影響相對較小
- 27 SNPs 是 genome-wide significance(>紅色線)全部都在HLA基因座上。
- 19 SNPs 有suggestive association(>藍色線)，有17個基因不是在HLA基因座上



● Power:80%

Figure 3. Manhattan plots depicting chromosomal distribution of P values from allelic, dominant, and recessive association models. The red line indicates the genome-wide significance cutoff. The blue line indicates the genome-wide suggestive cutoff for P values.

分析結果- replication stage

- 選了 12個 non-HLA SNPs，加上 rs1673649 (HLA基因上的一個SNP) 當作 positive control。
- 12個 SNPs在 GWAS、replication stage、combined analysis中 做檢驗，如果都顯著才被考慮。
- 也跟 實驗單位為歐洲人的 CEU meta- analysis 做比較。

Table 1. Replicated and novel genes/loci in the discovery and replication stages and in the combined analysis in the North Indian rheumatoid arthritis cohort, and comparison with CEU meta-analysis findings (3)*

| SNP† | Chr | Chr. position | Gene | Location | Alleles‡ | MAF in North Indian cohort | P in GWAS | OR (95% CI) in GWAS | P in replication stage | OR (95% CI) in replication stage | P in combined analysis | OR (95% CI) in combined analysis | Frequency of predisposing allele in CEU meta-analysis | P in CEU meta-analysis | OR (95% CI) in CEU meta-analysis |
|--------------|-----|---------------|-----------|-----------------|----------|----------------------------|-----------------------|---------------------|------------------------|----------------------------------|------------------------|----------------------------------|---|------------------------|----------------------------------|
| rs4851269 | 2 | 100830348 | LOC150577 | Intron | G/A | 0.46 | 5.52×10^{-6} | 1.54 (1.27–1.86) | 0.47 | 0.95 (0.83–1.09) | 0.0007 | 0.83 (0.75–0.93) | 0.4 | 6.89×10^{-5} | 1.1 (1.05–1.15) |
| rs6542920 | 2 | 100845088 | LOC150577 | Intron | G/A | 0.44 | 6.88×10^{-6} | 1.53 (1.28–1.87) | 0.237 | 1.09 (0.95–1.24) | 0.0003 | 1.21 (1.09–1.35) | 0.33 | 4.84×10^{-3} | 1.11 (1.06–1.17) |
| rs1160542 | 2 | 100832155 | LOC150577 | Intron | G/A | 0.43 | 7.71×10^{-6} | 0.54 (0.40–0.71) | 1.00 | 1 (0.87–1.15) | 0.0009 | 0.77 (0.66–0.9) | 0.46 (G)¶ | 1.45×10^{-6} | 1.12 (1.07–1.17) |
| rs2557588 | 5 | 53311502 | ARL15# | Intron | A/C | 0.18 | 3.36×10^{-6} | 1.92 (1.39–2.43) | 0.05 | 1.21 (1.00–1.47) | 6.57×10^{-8} | 1.42 (1.22–1.66) | 0.25 | 0.81 | 1.01 (0.93–1.1) |
| rs1573649 | 6 | 32731258 | HLADQB2 | Coding | A/G | 0.44 | 2.75×10^{-6} | 1.46 (1.17–1.72) | 0.003 | 1.22 (1.07–1.39) | 0.000003 | 1.28 (1.16–1.42) | 0.48 (A)¶ | 0.28 | 1.03 (0.98–1.09) |
| rs561041†† | 9 | 129658678 | ZBTB34 | Flanking 3'-UTR | A/G | 0.31 | 9.42×10^{-6} | 0.34 (0.17–0.50) | 0.03†† | 1.42 (1.04–1.94) | 0.84 | 0.99 (0.88–1.11) | 0.28 | 0.90 | 1 (1–1) |
| rs4910287 | 11 | 11260163 | GALNTL4 | Flanking 3'-UTR | G/A | 0.4 | 5.20×10^{-6} | 0.65 (0.55–0.83) | 0.40 | 0.94 (0.82–1.08) | 0.0002 | 0.75 (0.64–0.87) | 0.25 | 0.48 | 0.98 (0.93–1.04) |
| rs1037013§ | 12 | 107219308 | RIC8B | Intron | A/G | 0.46 | 8.08×10^{-6} | 1.91 (1.24–2.29) | 0.42 | 1.05 (0.93–1.2) | 0.02 | 1.13 (1.02–1.25) | 0.43 | 0.84 | 1 (1–1) |
| rs7328282†† | 13 | 99554955 | DOCK9 | Intron | A/G | 0.44 | 5.19×10^{-6} | 2.02 (1.62–3.12) | 0.03 | 0.86 (0.75–0.98) | 0.82 | 1.01 (0.91–1.12) | 0.4 | 0.05 | 1.05 (1–1.1) |
| rs2094497†† | 13 | 27910122 | RASL11A | Flanking 3'-UTR | A/G | 0.39 | 9.97×10^{-6} | 2.16 (1.45–3.04) | 0.58 | 0.96 (0.84–1.10) | 0.02†† | 1.28 (1.04–1.56) | 0.41 (A)¶ | 0.81 | 1.01 (0.93–1.1) |
| rs12881250†† | 14 | 95425711 | LOC730118 | Flanking 3'-UTR | A/C | 0.29 | 5.15×10^{-6} | 2.60 (1.65–4.09) | 0.49 | 1.05 (0.91–1.21) | 0.003 | 1.19 (1.06–1.33) | 0.4 | 0.77 | 1.01 (0.94–1.08) |
| rs2002212 | 18 | 12006035 | IMPA2 | Intron | G/A | 0.08 | 1.12×10^{-6} | 1.96 (1.47–2.58) | 0.91 | 0.99 (0.8–1.22) | 0.001 | 1.30 (1.11–1.53) | 0.05 | 0.83 | 1.01 (0.92–1.11) |
| rs9941467 | 19 | 38629027 | SIPA1L3 | Intron | G/A | 0.1 | 3.59×10^{-6} | 1.88 (1.38–2.46) | 0.21 | 0.88 (0.71–1.08) | 0.063 | 1.16 (0.99–1.36) | 0.15 | 0.97 | 1 (1–1) |

分析結果-
replication
stage

**ARL15 gene
(ADP-ribosylation factor like 15)
上的rs255758 SNP 是唯一保留下來與RA
有suggestive association 的SNP.**

藉由發現ARL15上的 SNP，檢測位於ARL15上的SNP的 p
value(非之前選的12個SNP)

發現rs697109 ($P = 4.74 \times 10^{-6}$), rs697108 ($P = 4.82 \times 10^{-6}$), and rs31127($P = 9.28 \times 10^{-6}$)

這三個SNPs跟疾病都有suggestive association

這四個SNPs有很強的連鎖不平衡($r^2 > 0.9$)

分析結果- RS2555758 (c) IN ar115 與脂聯素濃 度的關係

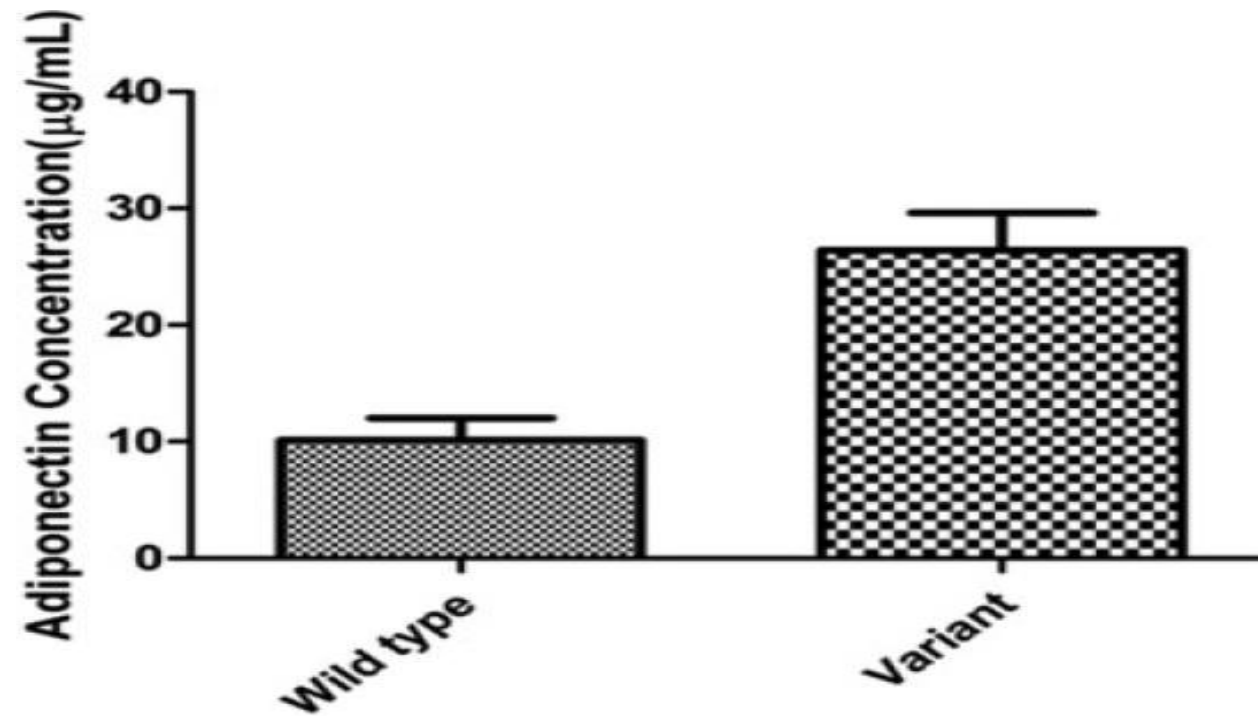


Figure 4. Adiponectin levels in individuals with wild-type (AA) and variant (CC) genotypes of rs255758 in *ARL15*. Values are the mean \pm SD. A significant difference ($P < 0.0001$) in adiponectin levels was found between rheumatoid arthritis patients harboring the wild-type genotype and those harboring the variant genotype.

- Wild type(AA)野生型等位基因型: 23/57 RA patients
- Variant(CC)突變型等位基因型: 34/57 RA patients
- 用 Mann-Whitney nonparametric test 去決定有無顯著差異
 - P value < 0.0001

分析結果-SVM

●模型:

- 1)篩SNPs($p \text{ value} < 0.001$)=517SNPs
- 2)100 subsets of data(90% 樣本數): 100 models
- 3)each model:每一個SNP給一個權重 w_i ，代表此SNP對患病預測的重要性。
- 4)得到 $|w_i|_{\text{avg}}$, find $|w|_{\text{max}}$ and corresponding σ

●發現: Cross Validation Accuracy:82.9%-95.1%, AUC:0.89-0.96

$|w|_{\text{max}}=0.572$, 對應到 rs10059065 SNP

$|w_i|_{\text{avg}}$ 在 $|w|_{\text{max}}$ 一倍 σ 以內的SNPs:

rs10059065, rs2218970, rs17023457, rs2199998, rs1892458, and rs11605437

對應到的基因是

LOC391845, NRP1, HAO2, HS3ST3B1, LY86, and ETS1(RA)

-多數都與自體免疫性疾病有關

分析結果-SVM

造成類風濕性關節炎的原因

- 造成類風濕性關節炎的原因是複雜的，包括遺傳、環境因子都有影響。

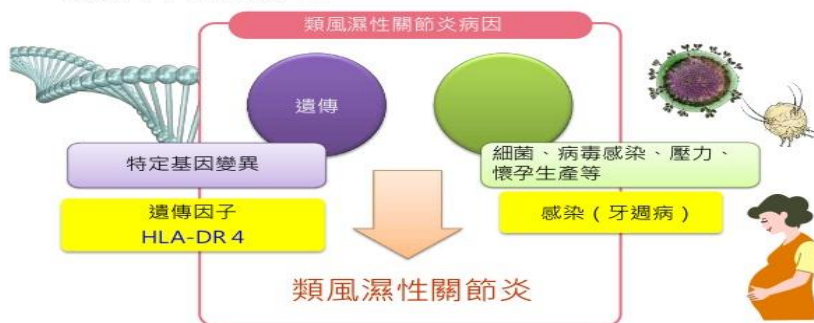


Table 2. Known disease associations of top genes identified in support vector machine analysis

| Gene (gene name) | Association with disease |
|--|--|
| NRP1 (neuropilin 1) | Inflammatory bowel disease, Crohn's disease, type 1 diabetes mellitus, osteonecrosis, ankylosing spondylitis |
| HAO2 (hydroxyacid oxidase 2) | Acquired immunodeficiency syndrome, urinary bladder neoplasms |
| HS3ST3B1 (heparan sulfate [glucosamine] 3-O-sulfotransferase 3B1) | Liver disease |
| LY86 (lymphocyte antigen 86) | Asthma, atherosclerosis, Crohn's disease, nasopharyngeal neoplasms, venous thromboembolism, mite-sensitive allergy |
| ETS1 (v-ets erythroblastosis virus E26 oncogene homolog 1 [avian]) | Lupus erythematosus, systemic lupus nephritis, celiac disease, <u>rheumatoid arthritis</u> |

- $|w_i|_{avg}$ 在 $|w|_{max}$ 3倍 σ 內的 SNPs: rs31127 對應 ARL15 基因

結論與未來展望

- 1.到現在為止我們還不知道，是否在白人種族內發現的致病基因可以類推到其他種族，所以在類風溼關節炎的GWAS中第一次針對北印地安人種族進行研究。
- 2.在第一階段實驗中，與RA最顯著相關的都是HLA gene上的SNPs，證明HLA 對RA的影響是跨越種族、非常重要的。
- 3.第二階段中，選了第一階段有suggestive association的12個non-HLA 區域之SNPs，只有 rs255758(在 ARL15基因座上)通過檢驗，為可能影響RA的SNP。
- 4.發現 ARL 15的rs255778中的 突變基因型CC，與脂聯素濃度有顯著關聯。研究證實:脂聯素的濃度越高，有RA的機率越高。

結論與未來展望

5. 利用**SVM**找出更多潛在的致病基因。找到六個可能的致病基因，也分別影響著多種自體免疫性疾病。
發現**影響自體免疫性疾病的基因是可能是共通的，有一套機制。**
6. 對照實驗對象為**白種人**的RA GWAS，找到的**致病基因有差異。**
 - **LOC 150577(rs1160542)**在兩族群中與RA關聯性有顯著差異
 - **ARL 15(rs255778)**對RA的影響並沒有在白種人族群中發現
7. 在北印地安人種族中找到一個新的可能致病基因-**ARL15**。
對於RA的遺傳學有更多的理解，也說明了在不同種族中進行 GWAS 的重要性。

Team3

類風濕關節炎單核甘酸多型性資料

1st 資料分析

資料前處理

- 篩選變數

1. 變異:

變異 $< 5\%$ 則刪掉

2. 遺漏值:

遺漏值的總數若大於 1% 刪掉

- 545080 SNPs \rightarrow 433188 SNPs

遭遇問題

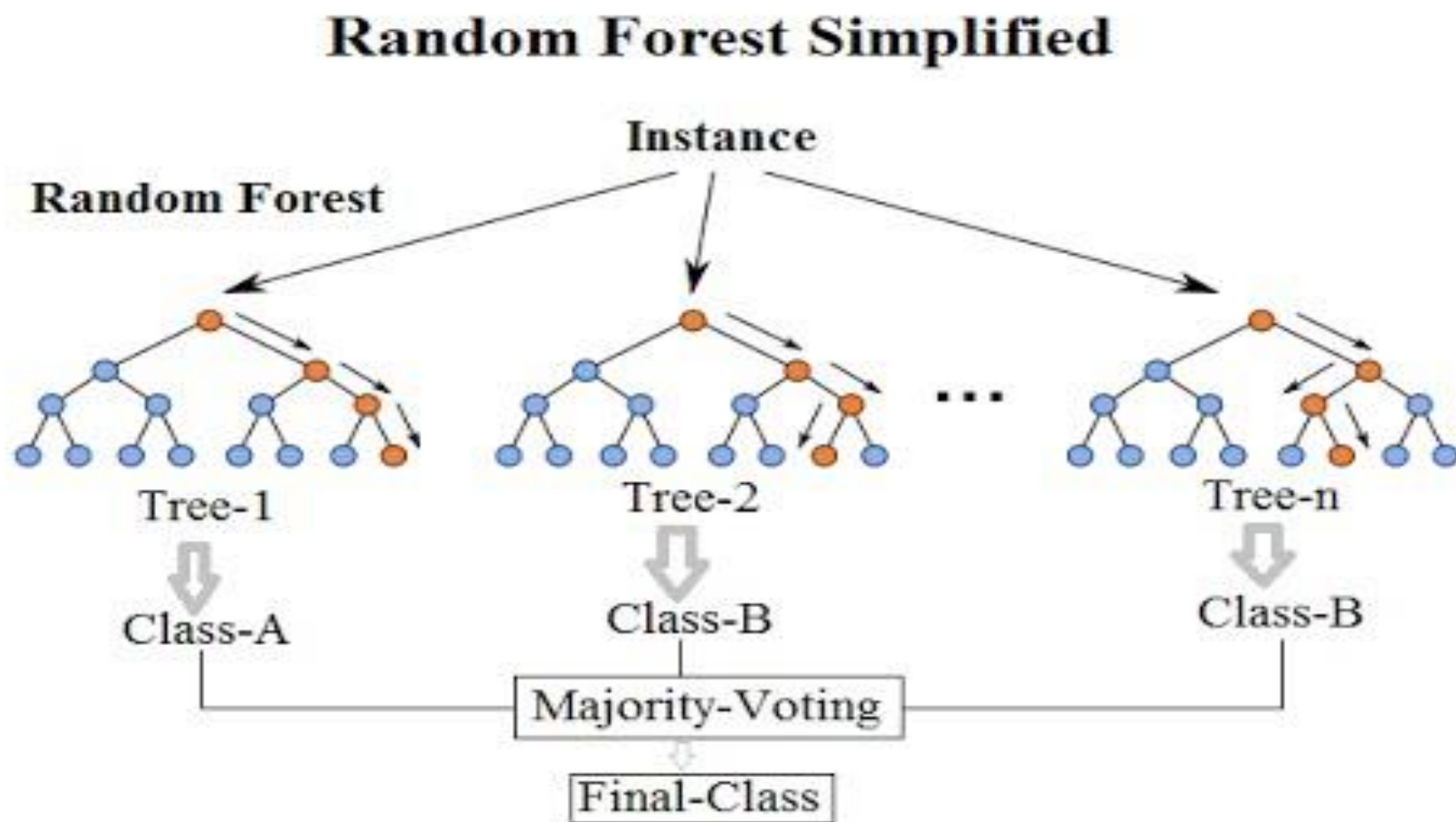
- 高維度資料，解釋變數的維度很高，遠超過樣本數
- 考慮變數間的高階交互作用下，選出適當的解釋變數

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_{443318} x_{443318} +$$

龐大交互作用



採用方法



Random Forest 隨機森林

- 所需做的假設較少
(變數的獨立性、與常態性等)
- 其算法中考慮了變數間的交互作用
- 計算效率高

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \beta_{433188} x_{433188} \\ + \beta_{433189} x_1 x_2 + \beta_{433190} x_1 x_3 + \beta_{433191} x_1 x_4 + \cdots$$

$\beta_3 \beta_5 \beta_8$ 個別計算效果不顯著

$\beta_3 \beta_5 \beta_8$ 同時考慮其效果才顯著

2nd 資料分析

DATASET

每次分割...

433188筆SNPs

44個1萬筆SNPs的子集

By Random Forest
method

選出100個重要的SNPs

100

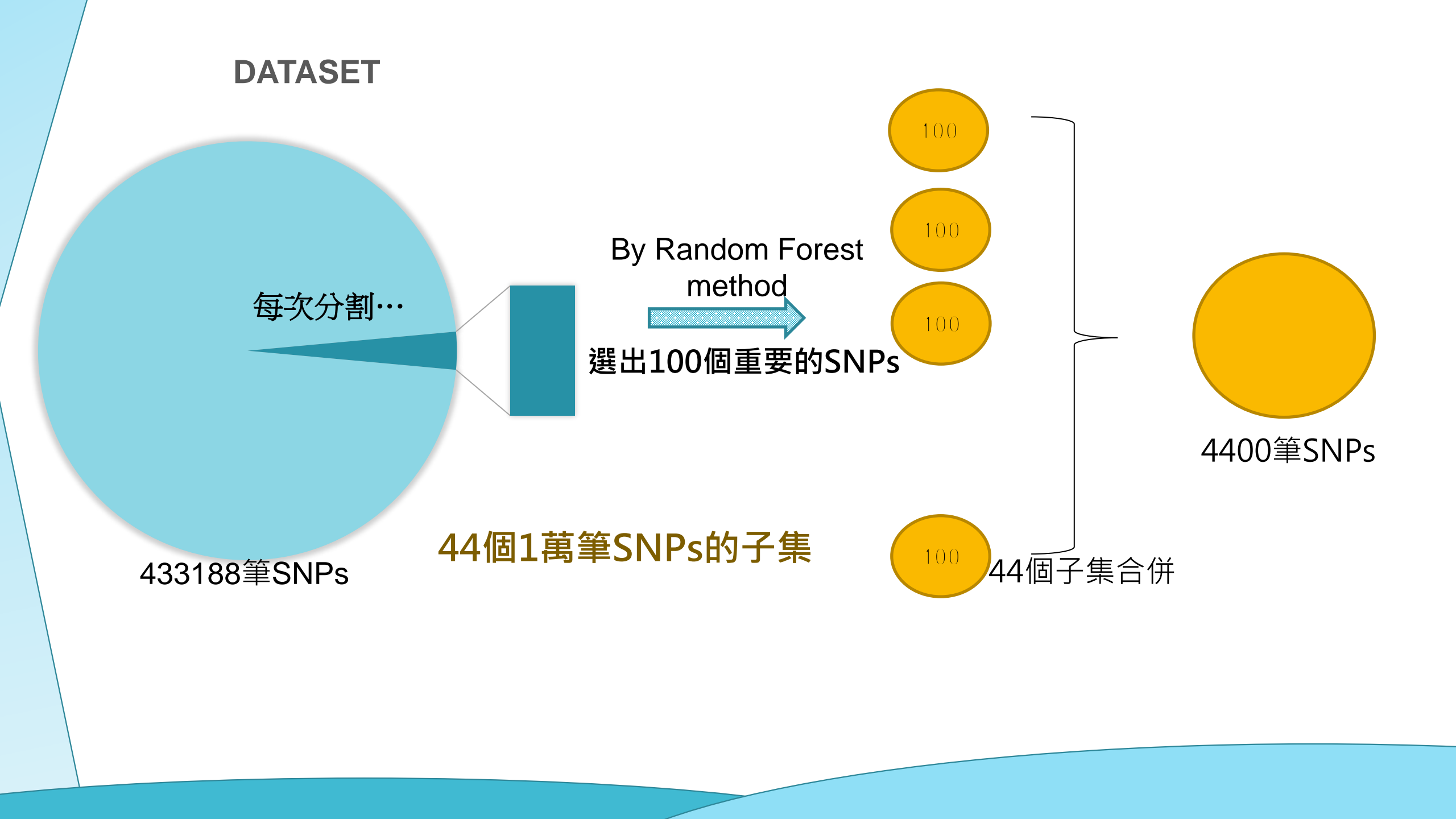
100

100

100

4400筆SNPs

44個子集合併



一萬筆SNPs中 做Random Forest

● 為求運算上能夠執行

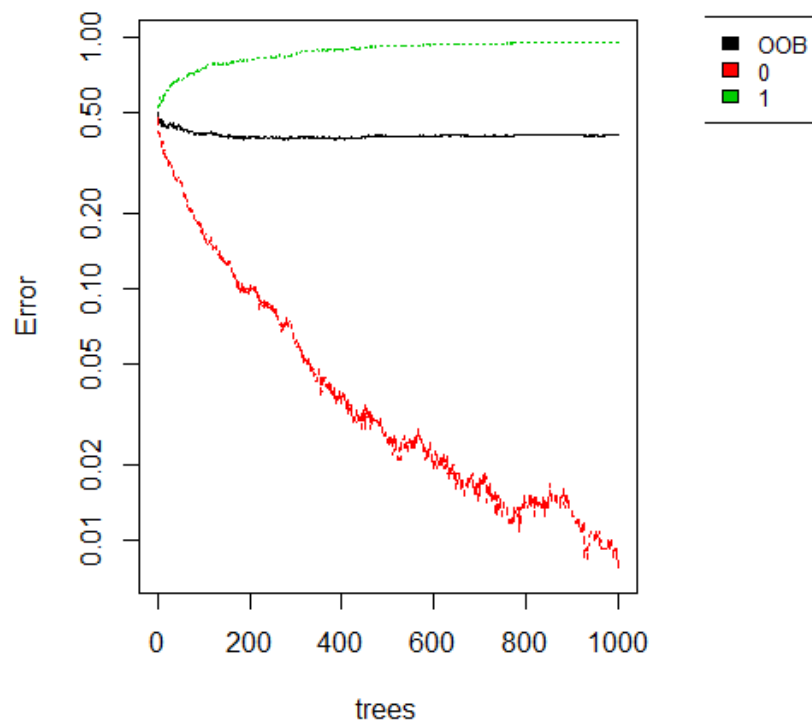
1. Procedure:
2. 將資料分割成44個資料集，
每個資料集有10000個 snps
(最後一個資料集3188個snps)
3. 每個資料集做隨機森林

參數:

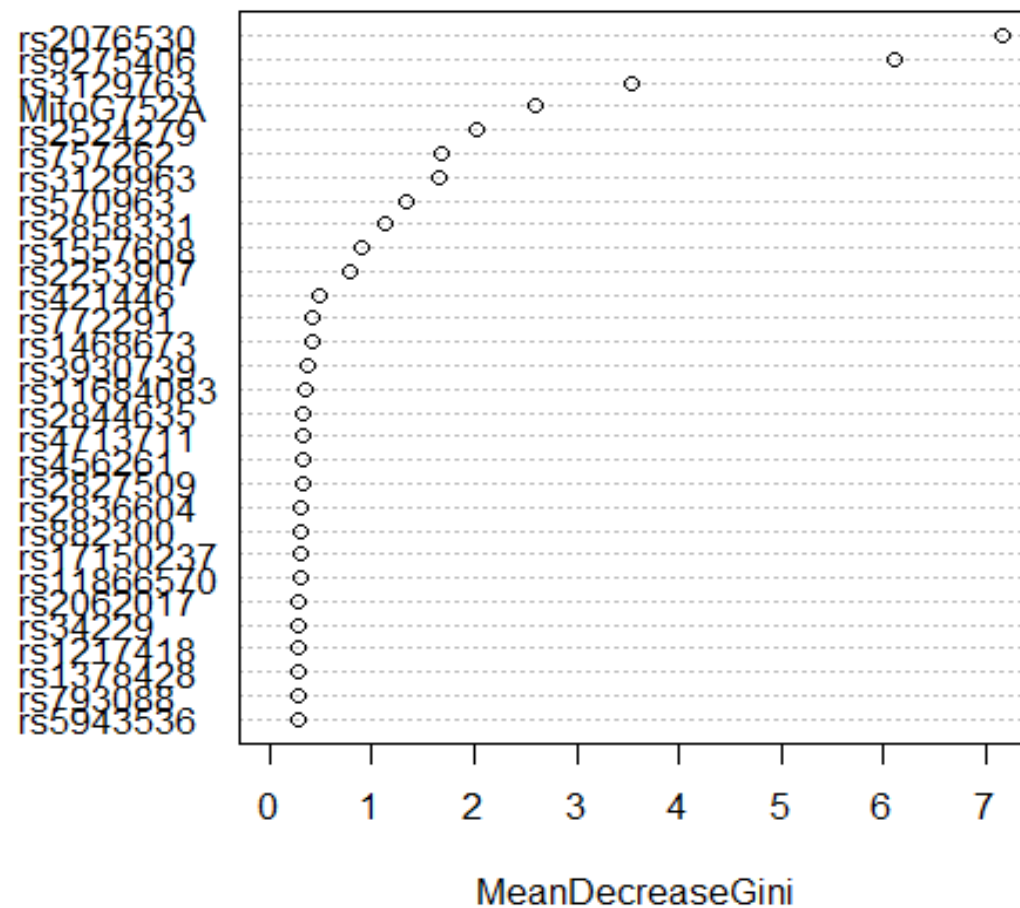
- 每次選取變數個數 $m = \sqrt{p} = \sqrt{10000} = 100$
- 每次種樹數目 $ntrees = 1000$

● 以下為隨機抽四個資料集的圖表闡釋

sample.1 10000 snps set

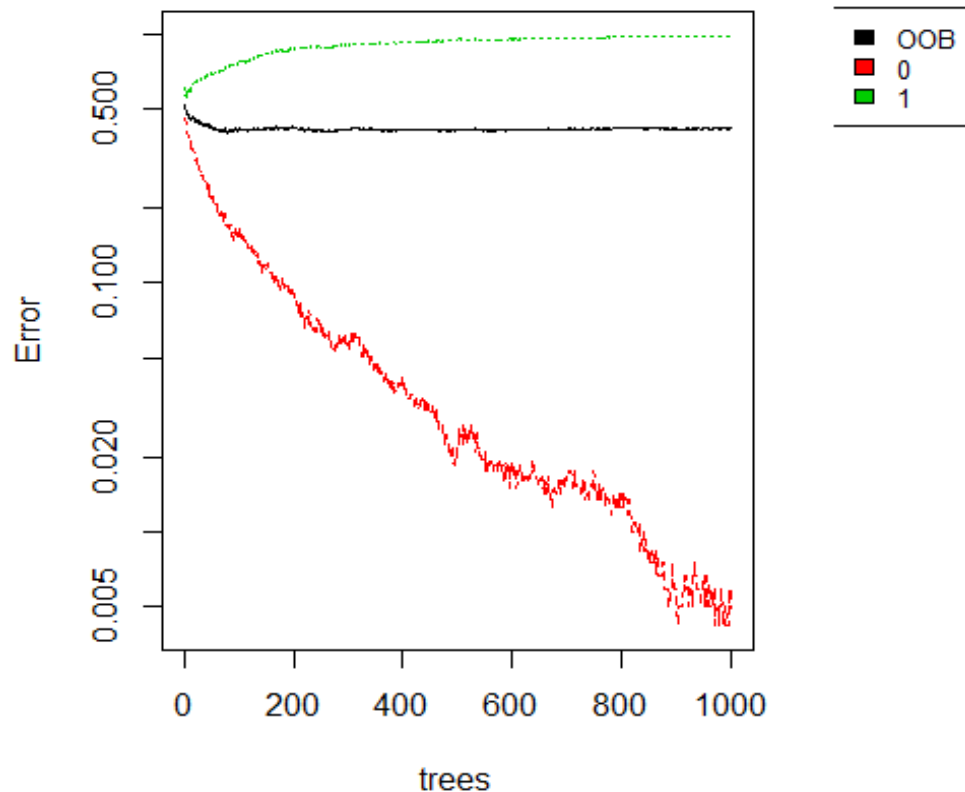


Variable Importance sample.1 10000 snps

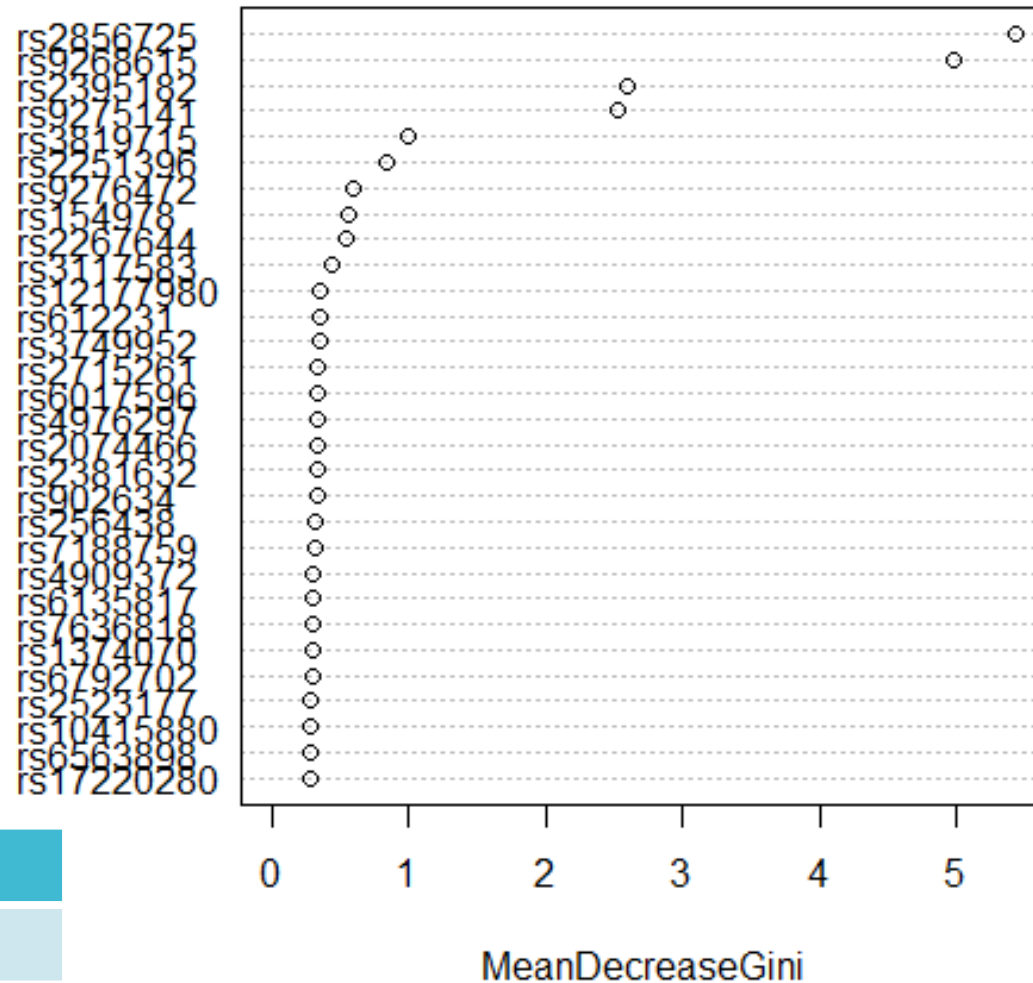


| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|-------|
| 實際沒患病 | 1184 | 10 | 0.008 |
| 實際有患病 | 830 | 38 | 0.956 |
| 整體錯誤率:40.74% | | | |

sample.2 10000 snps set

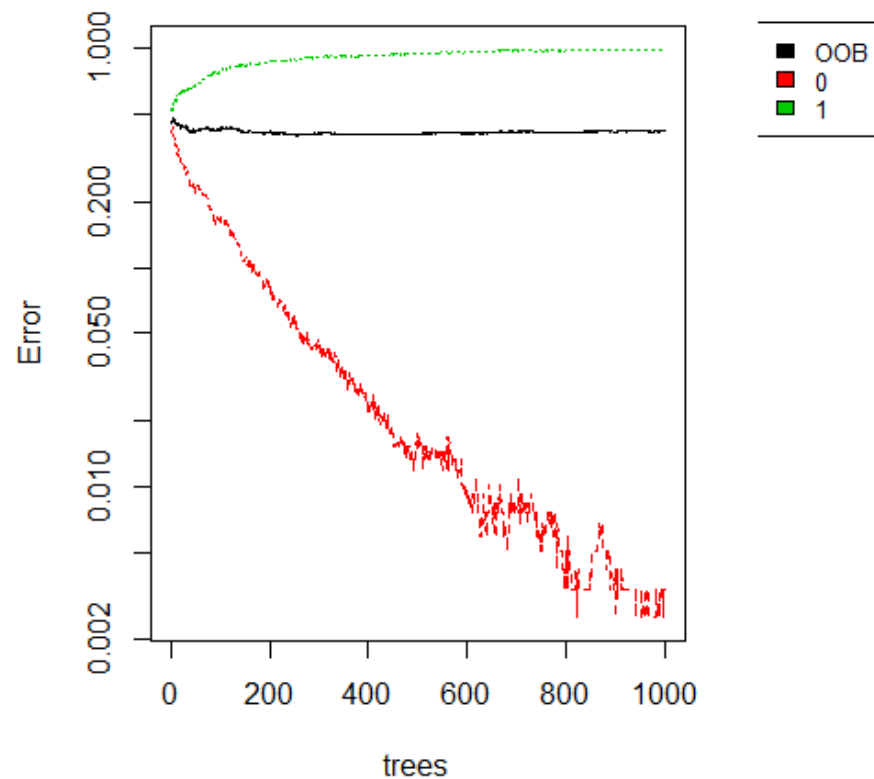


Variable Importance sample.2 10000 snps

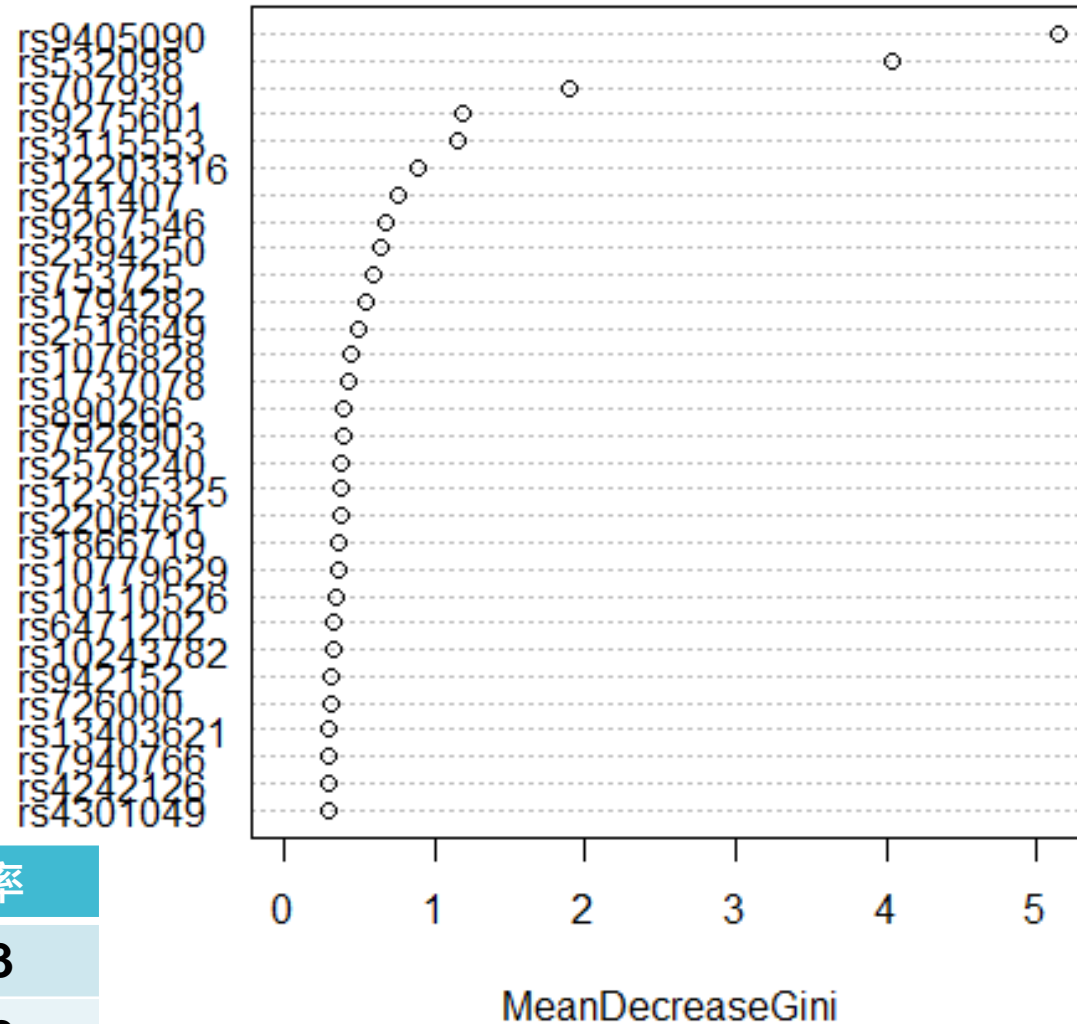


| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|--------|
| 實際沒患病 | 1187 | 7 | 0.0059 |
| 實際有患病 | 850 | 18 | 0.9793 |
| 整體錯誤率:41.56% | | | |

sample.3 10000 snps set

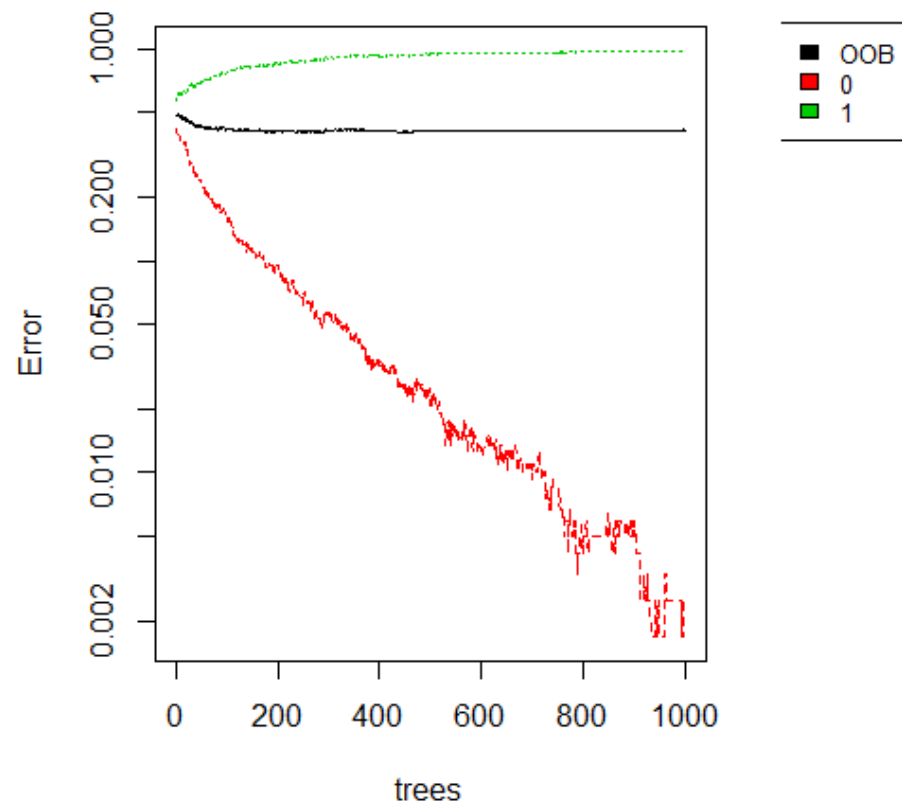


Variable Importance sample.3 10000 snps

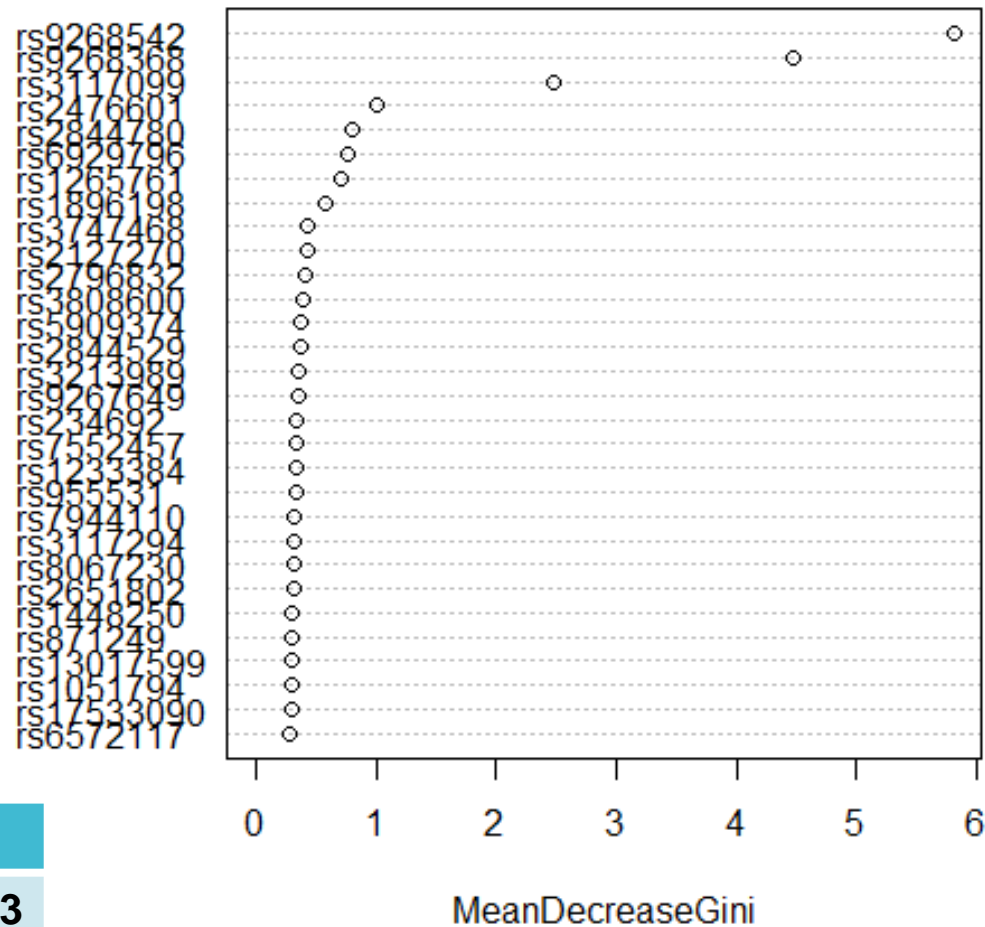


| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|-------|
| 實際沒患病 | 1190 | 4 | 0.003 |
| 實際有患病 | 861 | 7 | 0.992 |
| 整體錯誤率:41.95% | | | |

sample.4 10000 snps set



Variable Importance sample.4 10000 snps



| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|-------------|
| 實際沒患病 | 1191 | 3 | 0.002512563 |
| 實際有患病 | 853 | 15 | 0.982718894 |
| 整體錯誤率:41.51% | | | |

從44個資料集
中選取重要變數

- 每個資料集選100個最重要的SNPs
-根據variable importance

- 合併共4400 SNPs做隨機森林

參數:

- 每次選取變數個數

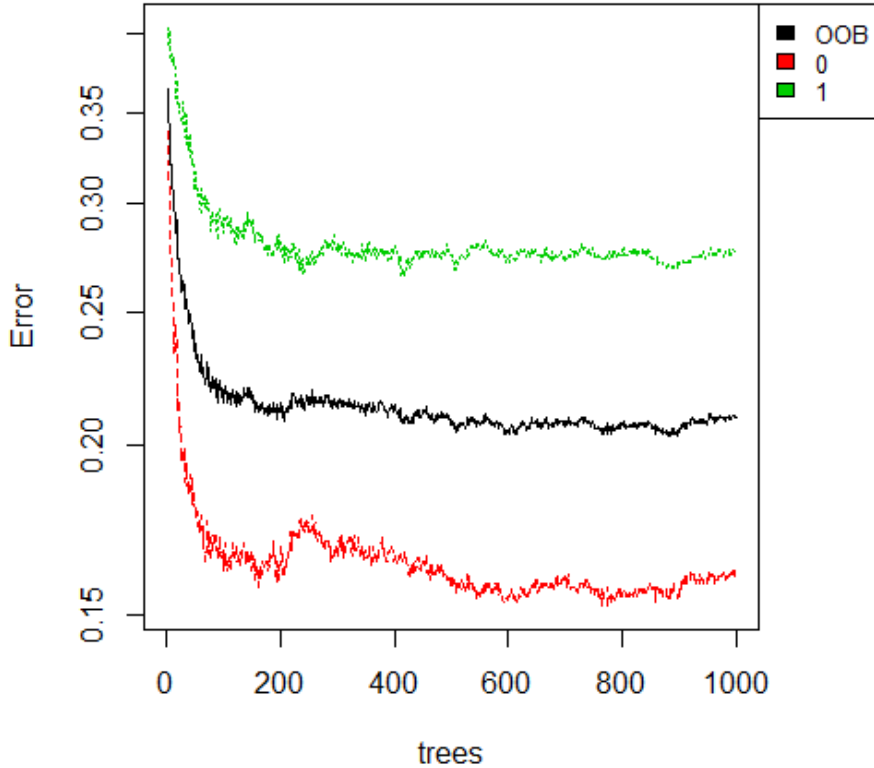
$$m = \sqrt{p} = \sqrt{4400} = 66$$

- 每次種樹數目

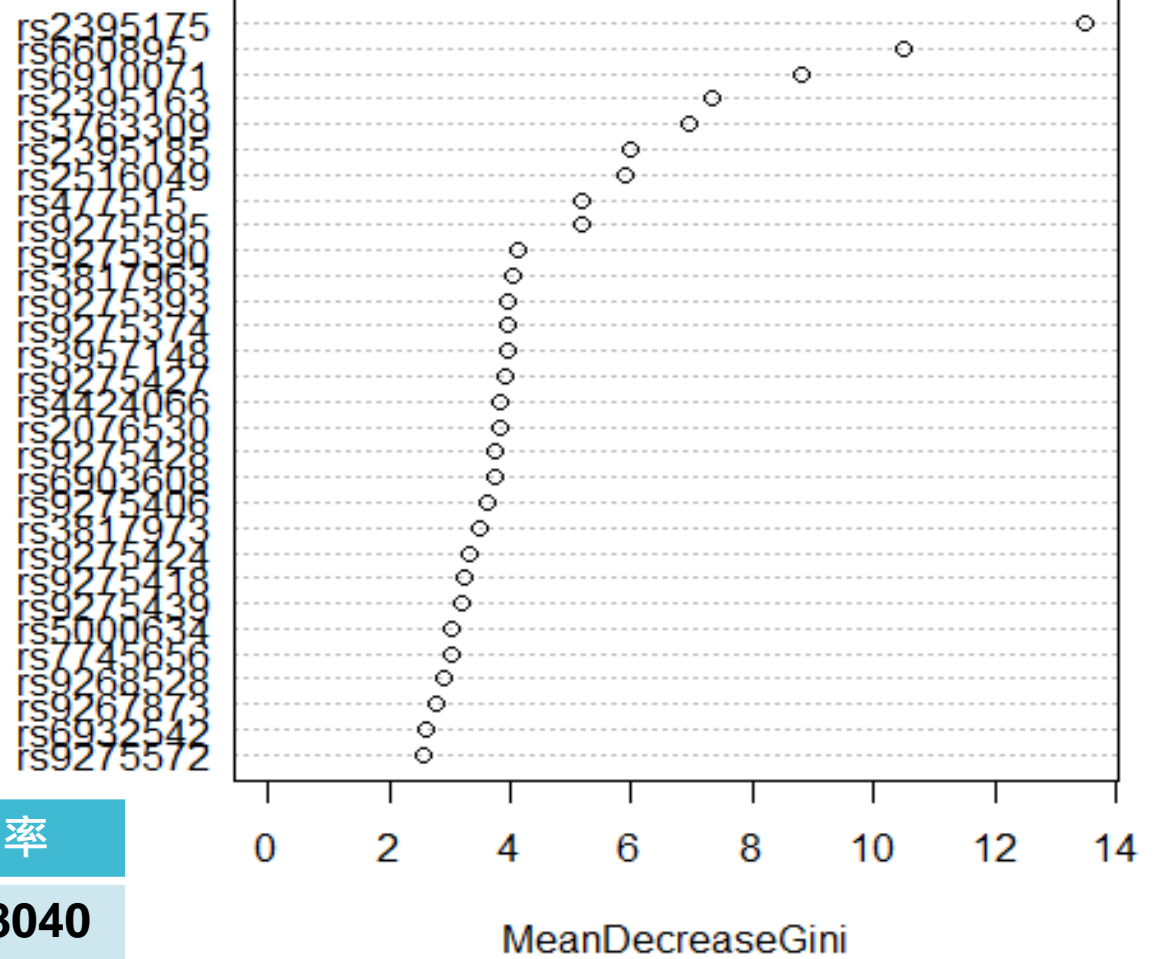
$$ntrees = 1000$$

- 圖表闡釋如下

selected snps set



Variable Importance selected snps



| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|-----------|
| 實際沒患病 | 1002 | 192 | 0.1608040 |
| 實際有患病 | 240 | 628 | 0.2764977 |
| 整體錯誤率:20.95% | | | |

100

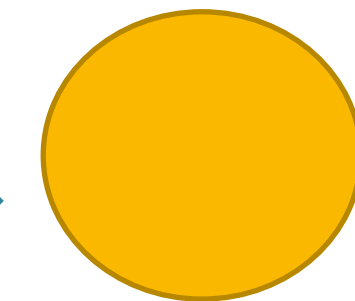
100

100

100

44個子集合併

藉由隨機森林
在每個資料集選100
個最重要的SNPs

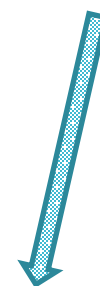


4400筆SNPs

再做隨機森林



?



如何決定篩選多少個SNPs?

決定最終 篩選變數 個數

- 採用wrapper variable selection

- 參考自

< Variable Selection in Random Forest with Application to Quantitative Structure-Activity Relationship >

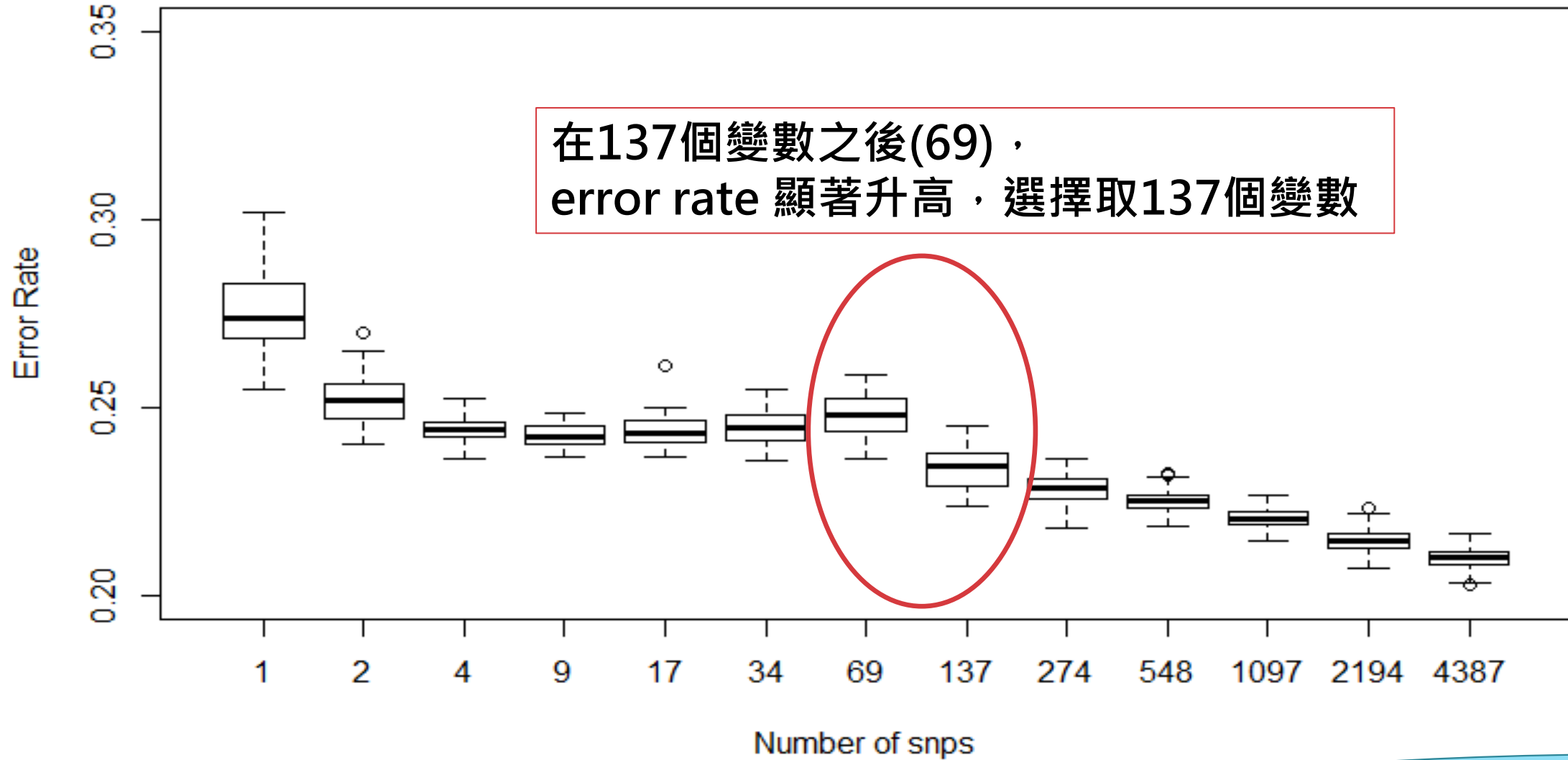
--Vladimir Svetnik, Andy Liaw, and Christopher Tong

決定篩選 變數個數

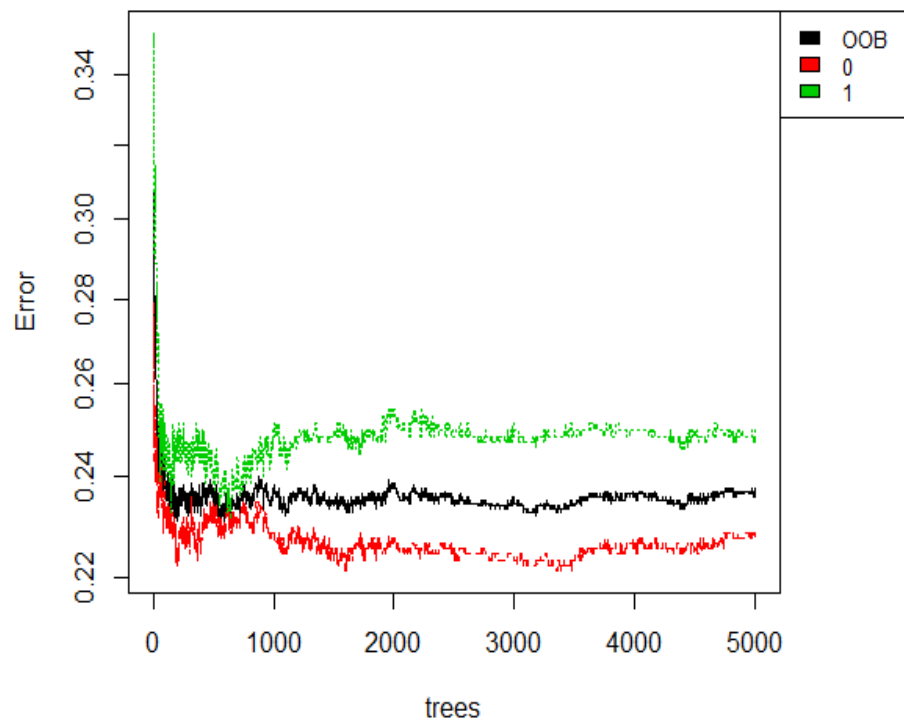
wrapper variable
selection

1. Partition the data for k-fold cross-validation.
我們選 $k = 5$
2. Remove the least important fraction.
We remove half of the variables
3. Repeat this step of removing half of the variables until a small number (1) remain
4. 重複50次的k-fold cross-validation以降低變異

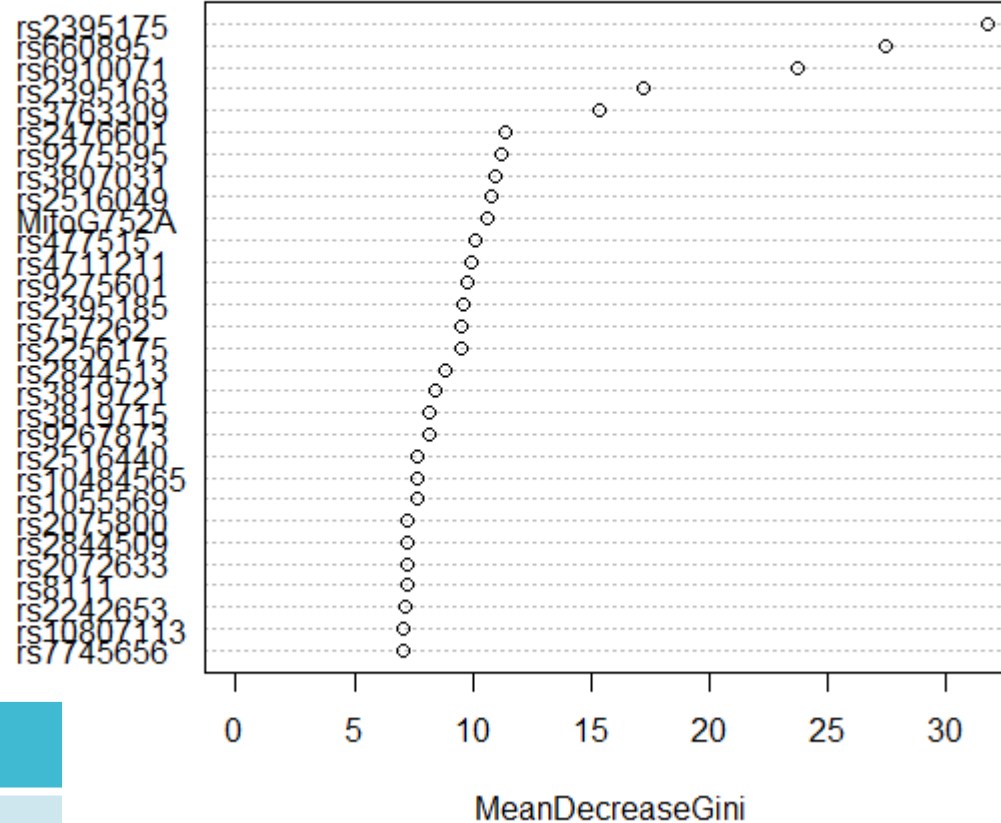
snps select by cv.error /50 times each var.num.(wrapper method)



select 137 snps set



Variable Importance selected 137 snps set



| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|-------|
| 實際沒患病 | 732 | 222 | 0.233 |
| 實際有患病 | 176 | 520 | 0.253 |
| 整體錯誤率:24.12% | | | |

決策樹與 隨機森林 預測率比較

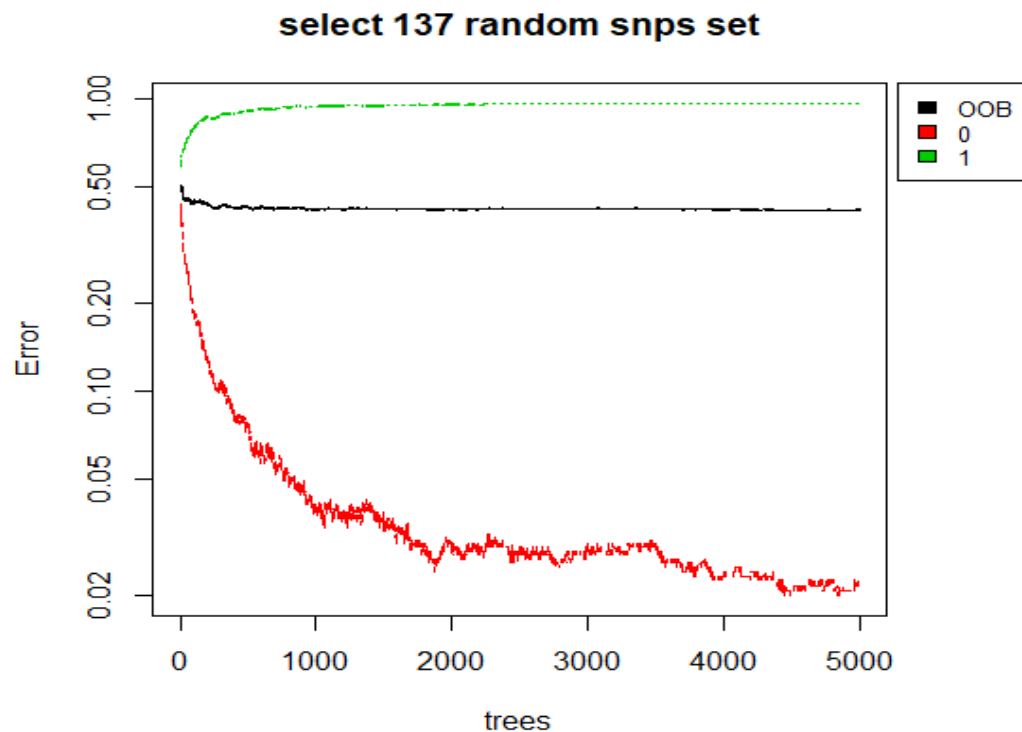
137 final SNPs,
4/5 training set
(1650 patients)
1/5 testing set
(412 patients)

Win!

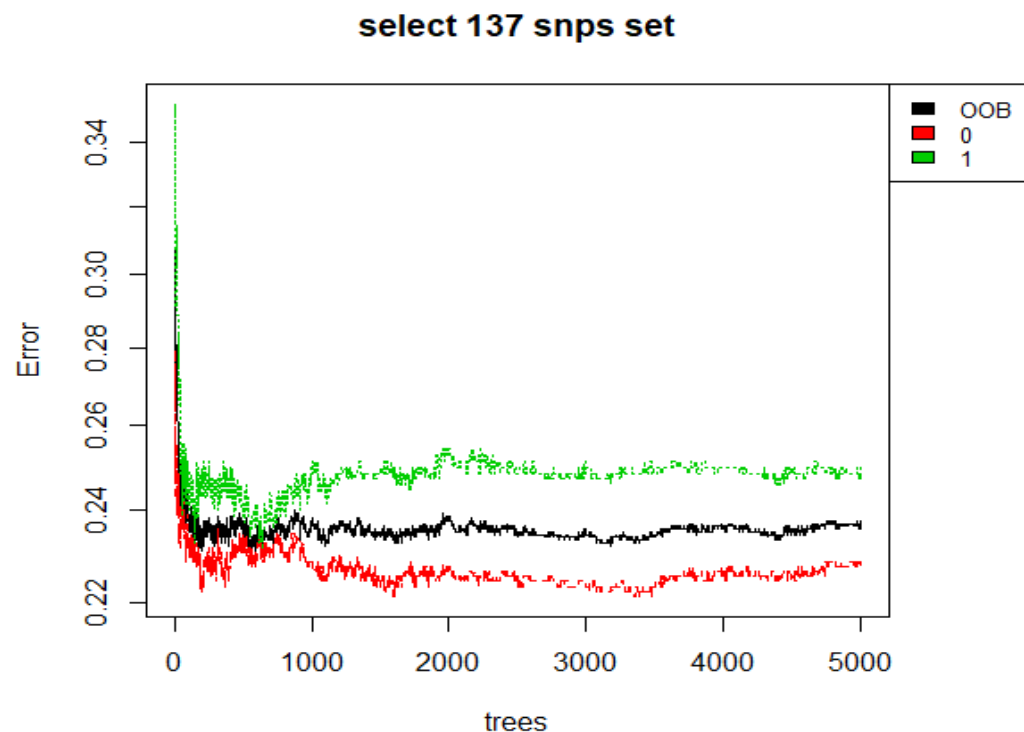
| 隨機森林 | 實際沒患病 | 實際有患病 | 準確率 |
|-------------|-------|-------|--------|
| 實際沒患病 | 229 | 15 | 0.9385 |
| 實際有患病 | 12 | 156 | 0.9285 |
| 整體準確率:0.934 | | | |

| 決策樹 | 預測沒患病 | 預測有患病 | 準確率 |
|-------------|-------|-------|--------|
| 實際沒患病 | 181 | 63 | 0.7418 |
| 實際有患病 | 47 | 121 | 0.7202 |
| 整體準確率:0.733 | | | |

比較隨機選出的 snps 跟最終選定的snps預測能力



| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|------------|
| 實際沒患病 | 1167 | 27 | 0.02261307 |
| 實際有患病 | 836 | 32 | 0.96313364 |
| 整體錯誤率:41.85% | | | |



| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|-------|
| 實際沒患病 | 732 | 222 | 0.233 |
| 實際有患病 | 176 | 520 | 0.253 |
| 整體錯誤率:24.12% | | | |

Rank of first 18 SNPs out of 137 selecte SNPs

| rsID | Chromosome | SNP | Associated |
|-------------|------------|-----|-----------------------------|
| "rs2395175" | 6 | A/G | |
| "rs660895" | 6 | A/G | 類風濕關節炎、免疫球蛋白腎病 |
| "rs6910071" | 6 | A/G | 類風濕關節炎 |
| "rs2395163" | 6 | C/T | 帕金森氏症 |
| "rs3763309" | 6 | A/C | 類風濕關節炎 |
| "rs9275595" | 6 | C/T | BMI、肥胖 |
| "rs477515" | 6 | C/T | B型肝炎疫苗免疫應答、EB病毒免疫應答、發炎性腸道疾病 |
| "rs2395185" | 6 | G/T | 霍奇金氏淋巴瘤、肺癌、抗核抗體水平、潰瘍性結腸炎 |
| "rs2516049" | 6 | A/G | EB病毒核抗原 1 IgG水平 |
| "rs3817963" | 6 | A/G | 肺腺癌、C型肝炎引起的肝硬化、多發性硬化症 |
| "rs9275427" | 6 | C/T | |
| "rs3817973" | 6 | A/G | |
| "rs9275393" | 6 | A/G | |
| "rs9275390" | 6 | C/T | 系統性硬化症 |
| "rs9275428" | 6 | A/G | |
| "rs9275424" | 6 | A/G | |
| "rs9275406" | 6 | G/T | 類風濕關節炎 |

發現

- 篩選出來的137個SNPs全落在第六號染色體基因座上。
- 實驗結果與先前文獻所論述相同，危險因子座落在第六號染色體上
- 找到的前幾個snps已被證實與類風溼關節炎有關
- 找到的前6個最重要的snps跟
- A genome-wide association scan for rheumatoid arthritis data by Hotelling's T^2 tests 這篇文章的結果一樣!

3rd 資料分析

Goal:
增加預測
準確率

1.Recode SNPs

→ 出現比較少次的含氮鹼基數量

ex.出現比較少次之含氮鹼基:G,

0個G:0

1個G:1

2個G:2

2.到cluster上用原始資料
(433188 SNPs)進行分析

→ 找到global的optimum以及其
交互作用

Goal:
增加預測
準確率

1.Recode SNPs

→ 出現比較少次的含氮鹼基數量

ex.出現比較少次之含氮鹼基:G

0個G:0

1個G:1

2個G:2

DATASET

將資料重新
編碼為出現較少次之
含氮鹼基數量:0,1,2
3 levels

每次分割...

433188筆SNPs

By Random Forest
method

選出100個重要的SNPs

100

100

100

4400筆SNPs

44個1萬筆SNPs的子集

100

44個子集合併

從44個資料集
中選取重要變數

- 每個資料集選100個最重要的SNPs
-根據variable importance

- 合併共4400 SNPs做隨機森林

參數:

- 每次選取變數個數

$$m = \sqrt{p} = \sqrt{4400} = 66$$

- 每次種樹數目

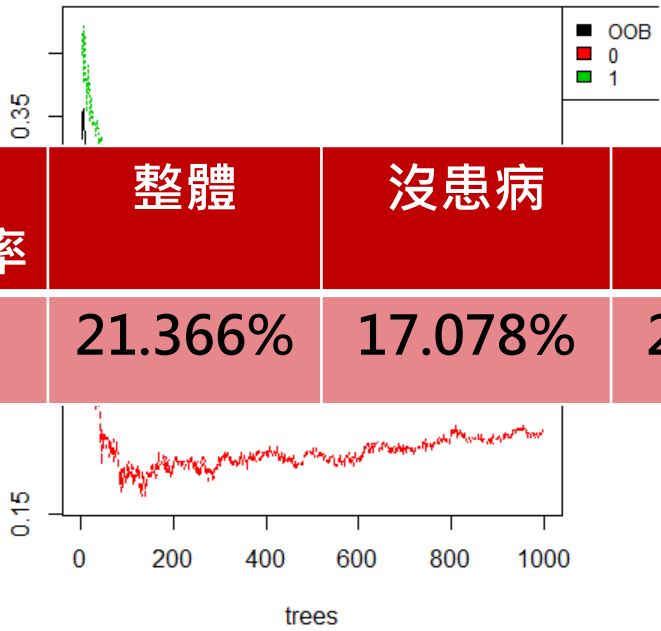
$$ntrees = 1000$$

比較未編碼與有編碼之4400 SNPs 之預測能力

Win!

未
編
碼
結
果

select 4400 snps set

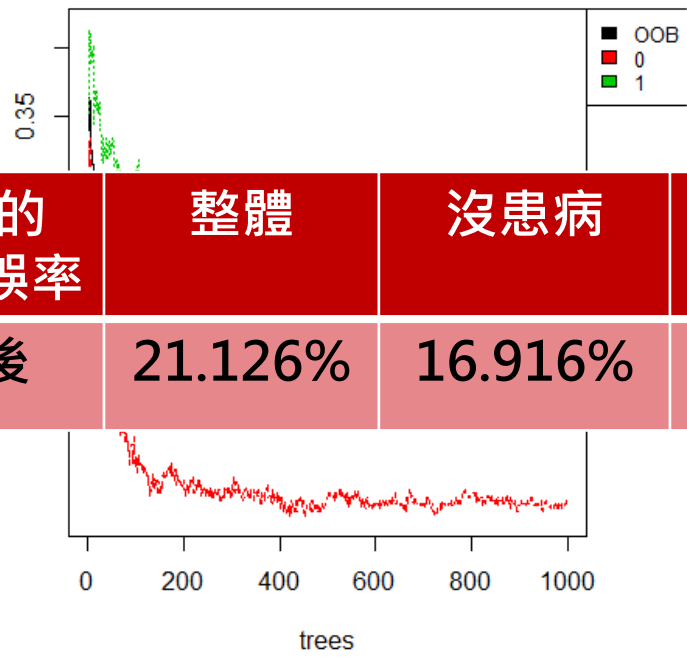


| 做5次的 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|---------------|---------|---------|---------|
| 沒有轉換 | 21.366% | 17.078% | 27.298% |

| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|-------|
| 實際沒患病 | 982 | 212 | 0.178 |
| 實際有患病 | 232 | 636 | 0.267 |
| 整體錯誤率:21.53% | | | |

有
編
碼
結
果

selected 4400(recode)snps set



| 做5次的 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|---------------|---------|---------|---------|
| 轉換後 | 21.126% | 16.916% | 26.912% |

| | 預測沒病 | 預測有患病 | 錯誤率 |
|-------------|------|-------|-------|
| 實際沒患病 | 998 | 196 | 0.164 |
| 實際有患病 | 237 | 631 | 0.273 |
| 整體錯誤率:21.0% | | | |

Dominant、
Recessive
的編碼轉換
與分析結果

4400SNPs

- Additive Coding:
編碼為:0,1,2=出現較少次之含氮鹼基數量
3 levels
- Dominant Coding:
編碼為:1,1,0=是否有出現較少次之含氮鹼基
2 levels
- Recessive Coding:
編碼為:2,1,1=是否有兩個出現較少次之含氮
鹼基 2 levels

Additive Dominante Recessive

編碼分析結果比較

4400 SNPs

2

Additive Coding(2,1,0)

| 做5次 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|--------------|---------|---------|---------|
| | 21.126% | 16.196% | 26.912% |

3

Recessive Coding(2,1,1)

| 做5次 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|--------------|---------|-------|---------|
| | 24.606% | 22.9% | 26.958% |

1

Dominant Coding(1,1,0)

| 做5次 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|--------------|--------|--------|--------|
| | 20.73% | 12.59% | 31.93% |

決定最終 篩選變數 個數

- 採用wrapper variable selection

- 參考自

< Variable Selection in Random Forest with Application to Quantitative Structure-Activity Relationship >

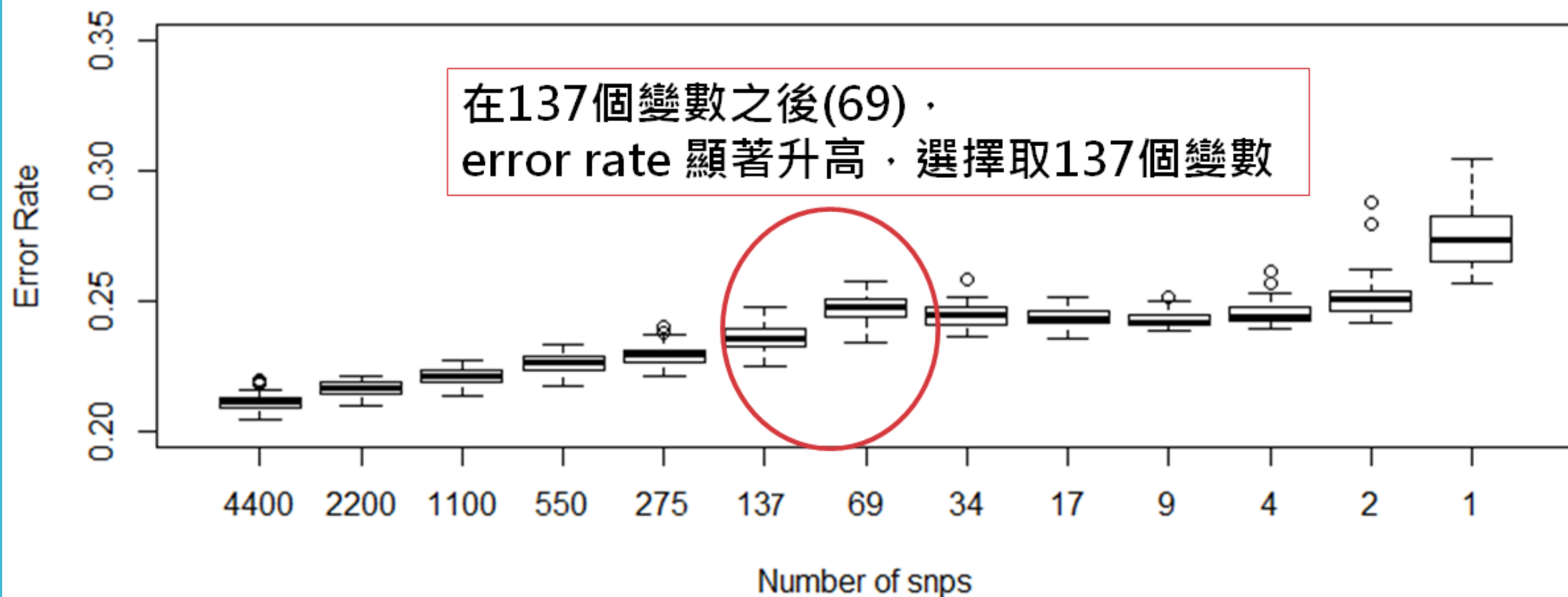
--Vladimir Svetnik, Andy Liaw, and Christopher Tong

決定篩選 變數個數

wrapper variable
selection

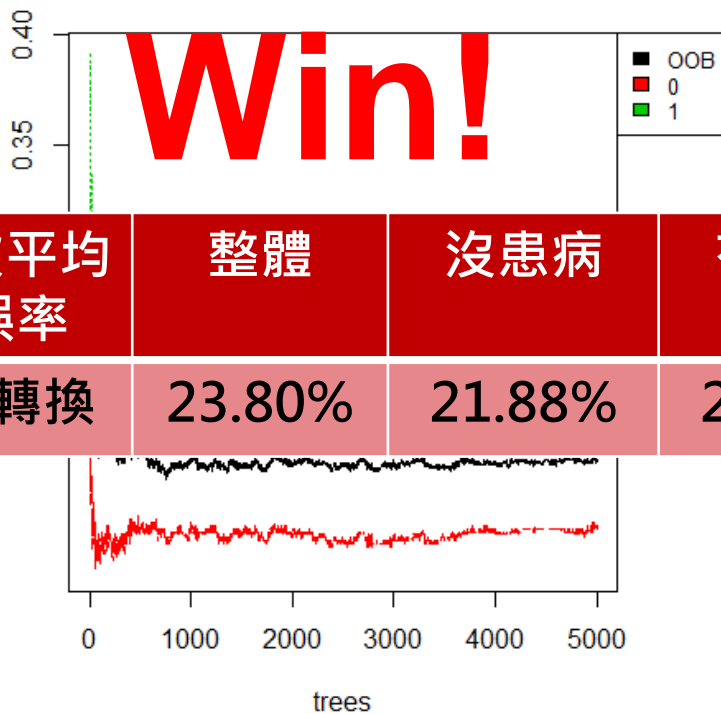
1. Partition the data for k-fold cross-validation.
我們選 $k = 5$
2. Remove the least important fraction.
We remove half of the variables
3. Repeat this step of removing half of the variables until a small number (1) remain
4. 重複50次的k-fold cross-validation以降低變異

snps select by cv.error(recode)/50 times each var.num.(wrapper method)



未編碼結果

select 137 snps set

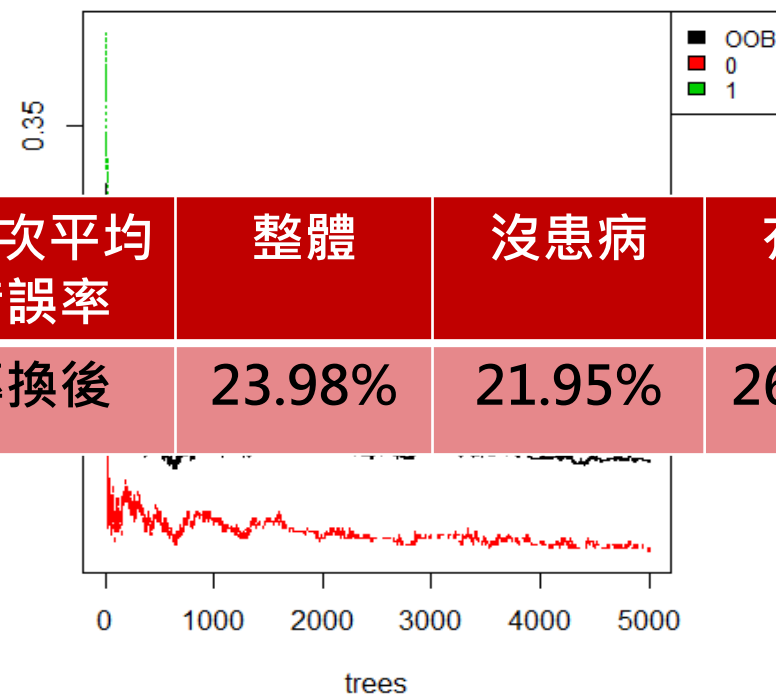


| 做5次平均 錯誤率 | 整體 | 沒患病 | 有患病 |
|--------------|--------|--------|--------|
| 沒有轉換 | 23.80% | 21.88% | 26.43% |

| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|-------|
| 實際沒患病 | 932 | 262 | 0.219 |
| 實際有患病 | 229 | 639 | 0.264 |
| 整體錯誤率:23.81% | | | |

有編碼結果

select 137 snps(recode) set



| 做5次平均 錯誤率 | 整體 | 沒患病 | 有患病 |
|--------------|--------|--------|---------|
| 轉換後 | 23.98% | 21.95% | 26.758% |

| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|-------|
| 實際沒患病 | 932 | 262 | 0.219 |
| 實際有患病 | 235 | 633 | 0.270 |
| 整體錯誤率:24.10% | | | |

Dominant、
Recessive
的編碼轉換
與分析結果

137SNPs

- Additive Coding:
編碼為:0,1,2=出現較少次之含氮鹼基數量
3 levels
- Dominant Coding:
編碼為:1,1,0=是否有出現較少次之含氮鹼基
2 levels
- Recessive Coding:
編碼為:2,1,1=是否有兩個出現較少次之含氮
鹼基 2 levels

Additive Dominante Recessive

編碼分析結果比較

137snps

2

Additive Coding(2,1,0)

| 做5次 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|--------------|--------|--------|--------|
| | 23.98% | 21.95% | 26.76% |

3

Recessive Coding(2,1,1)

| 做5次 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|--------------|---------|--------|--------|
| | 24.59 % | 22.90% | 26.96% |

1

Dominant Coding(1,1,0)

| 做5次 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|--------------|--------|--------|--------|
| | 23.76% | 21.30% | 27.14% |

Goal:
增加預測
準確率

2.到cluster上用原始資料
(433188 SNPs)進行分析
→找到global的optimum以及其
交互作用

Result

- Package:ranger-
fast version of 隨機森林
- 參數: ntree=5000
 mtry=5000
 自變數:433188 SNPs

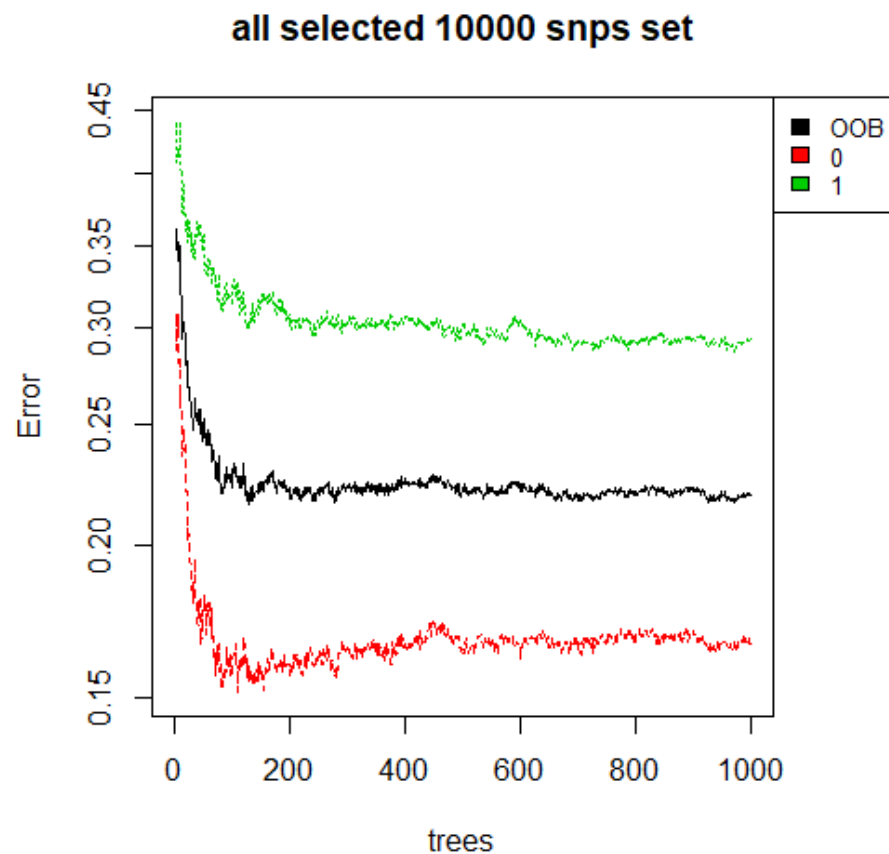
| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|--------|
| 實際沒患病 | 978 | 216 | 0.1809 |
| 實際有患病 | 274 | 594 | 0.3156 |
| 整體錯誤率:23.76% | | | |

Result

- 參數: ntree=1000
mtry=100
10000 SNPs

| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|-------|
| 實際沒患病 | 996 | 198 | 0.165 |
| 實際有患病 | 255 | 613 | 0.293 |
| 整體錯誤率:21.97% | | | |

選出最重要的10000個SNPs
再進行一次隨機森林分析。



未編碼(切割)4400SNPs

| 做5次 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|--------------|--------|--------|--------|
| | 21.37% | 17.08% | 27.30% |

未編碼(cluster)4400SNPs

| 做5次 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|--------------|--------|--------|-------|
| | 21.47% | 17.38% | 27.2% |

4400 SNPs錯誤率比較

2

Additive(切割)4400SNPs

| 做5次 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|--------------|--------|--------|--------|
| | 21.13% | 16.92% | 26.91% |

1

Dominant (切割)4400SNPs

| 做5次 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|--------------|--------|--------|--------|
| | 20.73% | 12.59% | 31.93% |

5

Recessive (切割)4400SNPs

| 做5次 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|--------------|--------|-------|--------|
| | 24.61% | 22.9% | 26.96% |

前10個重要的SNPs差異

| SNPs | Cluster 4400 SNPs | 已被證實 影響之疾病 | 切割 4400SNPs | 已被證實影 響之疾病 | 切割 Recode 0,1,2 4400 SNPs | 已被證實 影響之疾病 |
|------|-------------------|--------------------|----------------|--------------------|---------------------------------|----------------|
| 1 | "rs2395175" | 類風濕關節炎 | "rs2395175" | 類風濕關節炎 | "rs2395175" | 類風濕關節炎 |
| 2 | "rs660895" | 類風濕關節炎 | "rs660895" | 類風濕關節炎 | "rs660895" | 類風濕關節炎 |
| 3 | "rs6910071" | 類風濕關節炎 | "rs6910071" | 類風濕關節炎 | "rs2395163" | 類風濕關節炎 |
| 4 | "rs2395163" | 類風濕關節炎 | "rs2395163" | 類風濕關節炎 | "rs6910071" | 類風濕關節炎 |
| 5 | "rs3763309" | 類風濕關節炎 | "rs3763309" | 類風濕關節炎 | "rs3763309" | 類風濕關節炎 |
| 6 | "rs2395185" | 類風濕關節炎 | "rs2395185" | 類風濕關節炎 | "rs2516049" | 類風濕關節炎 |
| 7 | "rs2516049" | EB病毒核抗原 1 IgG水平 | "rs2516049" | EB病毒核抗原 1 IgG水平 | "rs2395185" | 霍奇金氏淋巴瘤、 肺癌 |
| 8 | "rs477515" | 發炎性腸道病 | "rs477515" | 發炎性腸道病 | "rs477515" | 發炎性腸道病 |
| 9 | "rs9275595" | BMI、肥胖 | "rs9275595" | BMI、肥胖 | "rs9275595" | BMI、肥胖 |
| 10 | "rs9275425" | | "rs9275390" | 系統性硬化症 | "rs9275424" | |

分析內容補充及結論

1. 試著找出交互作用

Global前4400 snps- decision tree

分支節點
(terminal node)10個

rs2395175 =GG

rs9275595= AA

rs660895 = AA

rs2306420=GG

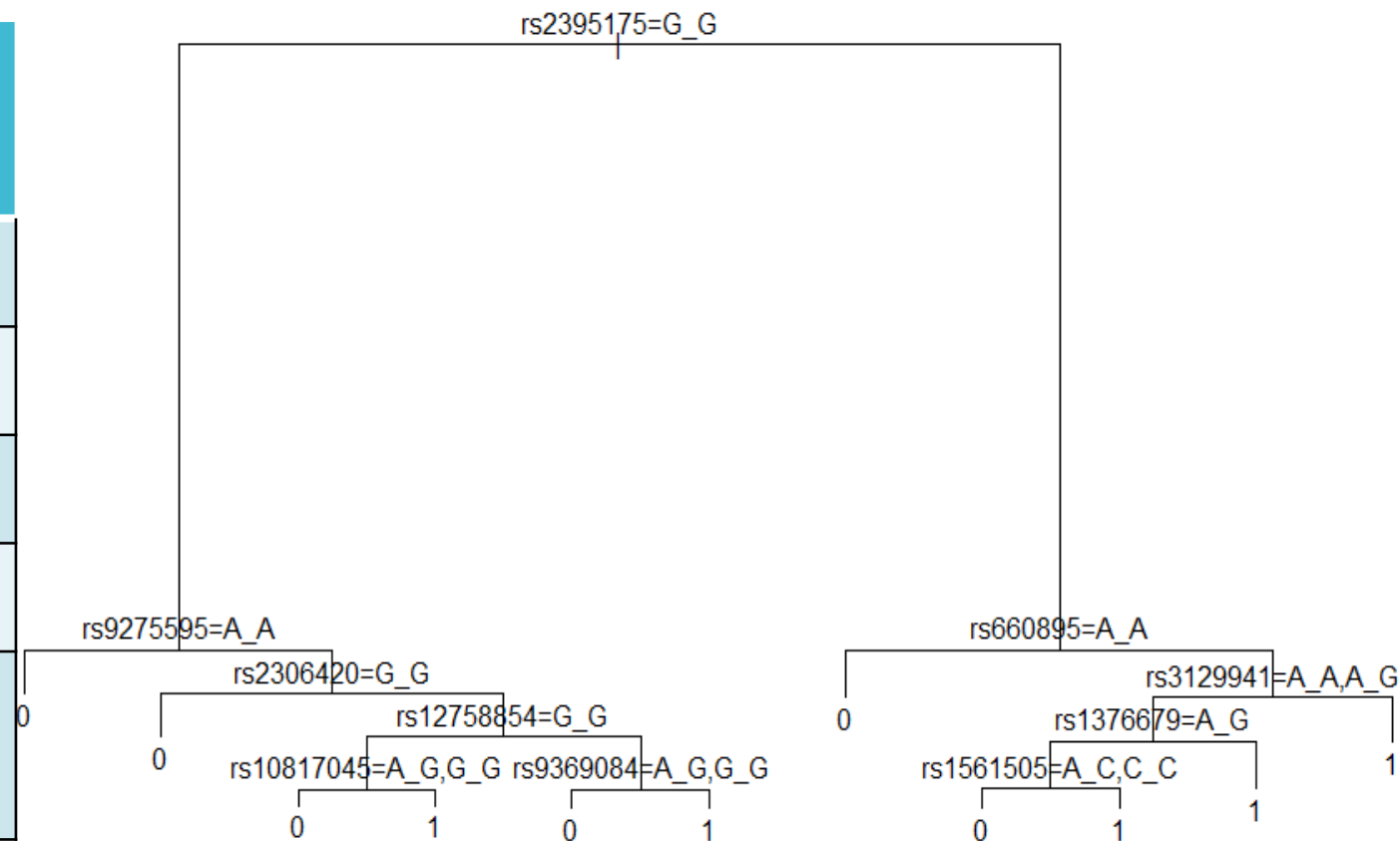
rs3129941=AA,AG

rs12758854=GG

rs1376679=AG

rs1087145=AG,GG
rs9369084=AG,GG

rs1561505=AC,CC



Problem of Detecting Interaction effect by decision tree

●Interaction Term Detect Reference: Methods for Interaction Detection in Predictive Modeling Using SAS (2012) Doug Thompson

- 檢驗分支準則的依據：
哪個屬性作為分枝變數可以帶來最大的純度
- 雖然決策樹的結構會自動形成交互作用，
但一個由A和B分支變數形成的節點，
不一定表示A和B有交互作用，
可能只是代表A和B這兩個變數的主效果都很重要。
- 用決策樹偵測交互作用時，要記得所選的分枝變數可能是主效果都很顯著，或彼此之間有交互作用
- Because leaves are constructed by identifying subgroups of observations that are as pure as possible with respect to the dependent variable, this essentially means that Y depends on both X1 and X2.
- However, a leaf defined by two predictors X1 and X2 does not necessarily point to an interaction between X1 and X2. It means that X1 and X2 are both important for the prediction of Y, but this is consistent with either main effects of X1 and X2 and no X1*X2 interaction, or an interaction between X1 and X2
- When using the decision tree method to detect interactions, it is worthwhile to keep in mind that tree leaves may point to either main effects or interactions.

Logistic Regression

(包含1.主效果
2.二階交互作用項)

反應變數y :Affection(0/1)

自變數xi： 18項

| 主效果 | 二階交互作用 |
|------------|-----------------------|
| rs2395175 | rs2395175*rs9275595 |
| rs9275595 | rs2395175*rs660895 |
| rs660895 | rs9275595*rs2306420 |
| rs2306420 | rs660895*rs3129941 |
| rs3129941 | rs2306420*rs12758854 |
| rs12758854 | rs3129941*rs1376679 |
| rs1376679 | rs12758854*rs10817045 |
| rs10817045 | rs12758854*rs9369084 |
| rs9369084 | rs1376679*rs1561505 |

| Variable Selection Method | Final Selected Model |
|---------------------------|---|
| Stepwise | $ \begin{aligned} &\text{rs2395175} + \text{rs9275595} + \text{rs660895} + \text{rs2306420} + \text{rs3129941} + \\ &\text{rs9369084} + \text{rs12758854} + \text{rs1376679} + \text{rs10817045} + \\ &\text{rs2395175:rs9275595} + \text{rs2395175:rs660895} + \\ &\text{rs9369084:rs12758854} + \text{rs3129941:rs1376679} \end{aligned} $ |
| Backward | Same |
| Forward | Same |

Logistic Regression selected model

| 變數 | 估計值 | P value | 變數 | 估計值 | P value |
|--|-------|-----------------------|----------------------------|-------|-----------------------|
| Intercept | 3.34 | 0.008 | rs2395175A_G:rs9275595A_G | 0.8 | 0.12 |
| rs2395175A_G | -4.19 | 0.0002 | rs2395175G_G:rs9275595A_G | 2.81 | 1.98*10 ⁻⁷ |
| rs2395175G_G | -4.8 | 6.11*10 ⁻⁷ | rs2395175A_G:rs9275595G_G | 1.9 | 0.003 |
| rs9275595A_G | -0.39 | 0.42 | rs2395175G_G:rs9275595G_G | 3.48 | 1.41*10 ⁻⁷ |
| rs9275595G_G | -0.6 | 0.28 | rs2395175A_G:rs660895A_G | 4.58 | 0.0001 |
| rs660895A_G | -1.88 | 0.06 | rs2395175G_G:rs660895A_G | 2.08 | 0.04 |
| rs660895G_G | 0.33 | 0.68 | rs2395175A_G:rs660895G_G | 2.64 | 0.01 |
| rs2306420A_G | -0.13 | 0.55 | rs2395175G_G:rs660895G_G | NA | NA |
| rs2306420G_G | -0.6 | 0.006 | rs9369084A_G:rs12758854A_G | -0.18 | 0.66 |
| rs3129941A_G | -0.3 | 0.69 | rs9369084G_G:rs12758854A_G | -2.26 | 0.01 |
| rs3129941G_G | -0.41 | 0.58 | rs9369084A_G:rs12758854G_G | -0.16 | 0.7 |
| rs9369084A_G | -0.18 | 0.64 | rs9369084G_G:rs12758854G_G | -2.47 | 0.004 |
| rs9369084G_G | 1.66 | 0.04 | rs3129941A_G:rs1376679A_G | -0.06 | 0.94 |
| rs12758854A_G | 0.11 | 0.69 | rs3129941G_G:rs1376679A_G | 0.87 | 0.34 |
| rs12758854G_G | -0.19 | 0.52 | rs3129941A_G:rs1376679G_G | 0.04 | 0.96 |
| rs1376679A_G | -0.95 | 0.29 | rs3129941G_G:rs1376679G_G | 0.53 | 0.59 |
| rs1376679G_G | -0.55 | 0.57 | | | |
| rs10817045A_G | -0.41 | 0.02 | | | |
| rs10817045G_G | -0.41 | 0.02 | | | |
| Null Deviance: 2806.8 2160 df Residual Deviance: 1886.2 2028 df AIC:1954.2 | | | | | |

Logistic
Regression:
remove
rs9369084*
rs12758854
interaction
term model

| 變數 | 估計值 | P value | 變數 | 估計值 | P value |
|---|-------|-----------------------|----------------------------|-------|-----------------------|
| Intercept | 2.87 | 0.008 | rs2395175A_G:rs9275595A_G | 0.85 | 0.1 |
| rs2395175A_G | -4.13 | 0.0002 | rs2395175G_G:rs9275595A_G | 2.81 | 1.77*10 ⁻⁷ |
| rs2395175G_G | -4.76 | 9.20*10 ⁻⁷ | rs2395175A_G:rs9275595G_G | 1.93 | 0.002 |
| rs9275595A_G | -0.42 | 0.38 | rs2395175G_G:rs9275595G_G | 3.46 | 1.46*10 ⁻⁷ |
| rs9275595G_G | -0.61 | 0.27 | rs2395175A_G:rs660895A_G | 4.47 | 0.0002 |
| rs660895A_G | -1.81 | 0.07 | rs2395175G_G:rs660895A_G | 2.02 | 0.05 |
| rs660895G_G | 0.38 | 0.65 | rs2395175A_G:rs660895G_G | 2.56 | 0.01 |
| rs2306420A_G | -0.11 | 0.59 | rs2395175G_G:rs660895G_G | NA | NA |
| rs2306420G_G | -0.6 | 0.006 | rs9369084A_G:rs12758854A_G | -0.23 | 0.58 |
| rs3129941A_G | -0.3 | 0.43 | rs9369084G_G:rs12758854A_G | -0.22 | 0.6 |
| rs3129941G_G | 0.16 | 0.66 | rs9369084A_G:rs12758854G_G | -2.34 | 0.007 |
| rs12758854A_G | 0.14 | 0.63 | rs9369084G_G:rs12758854G_G | -2.53 | 0.003 |
| rs12758854G_G | -0.17 | 0.56 | | | |
| rs1376679A_G | -0.35 | 0.01 | | | |
| rs1376679G_G | -0.17 | 0.26 | | | |
| rs10817045A_G | -0.4 | 0.03 | | | |
| rs10817045G_G | -0.41 | 0.02 | | | |
| rs9369084A_G | -0.13 | 0.72 | | | |
| rs9369084G_G | 1.71 | 0.03 | | | |
| Null Deviance: 2806.8 2160 df Residual Deviance: 1895.4 2032 df AIC:1955.4 | | | | | |

Test Fit of Model

1. H_0 : simpler model holds
 H_a : more complex model holds
(include rs3129941:rs1376679 interaction term)

2. $\alpha = 0.05$

3. Rejection Region: Chi Square Test
Statistic > 9.488 (chi square statistic with right tail prob. 0.05 value (df=4))

4. LR stat = difference in deviance =
 $1895.4 - 1886.2 = 9.2 < 9.488$

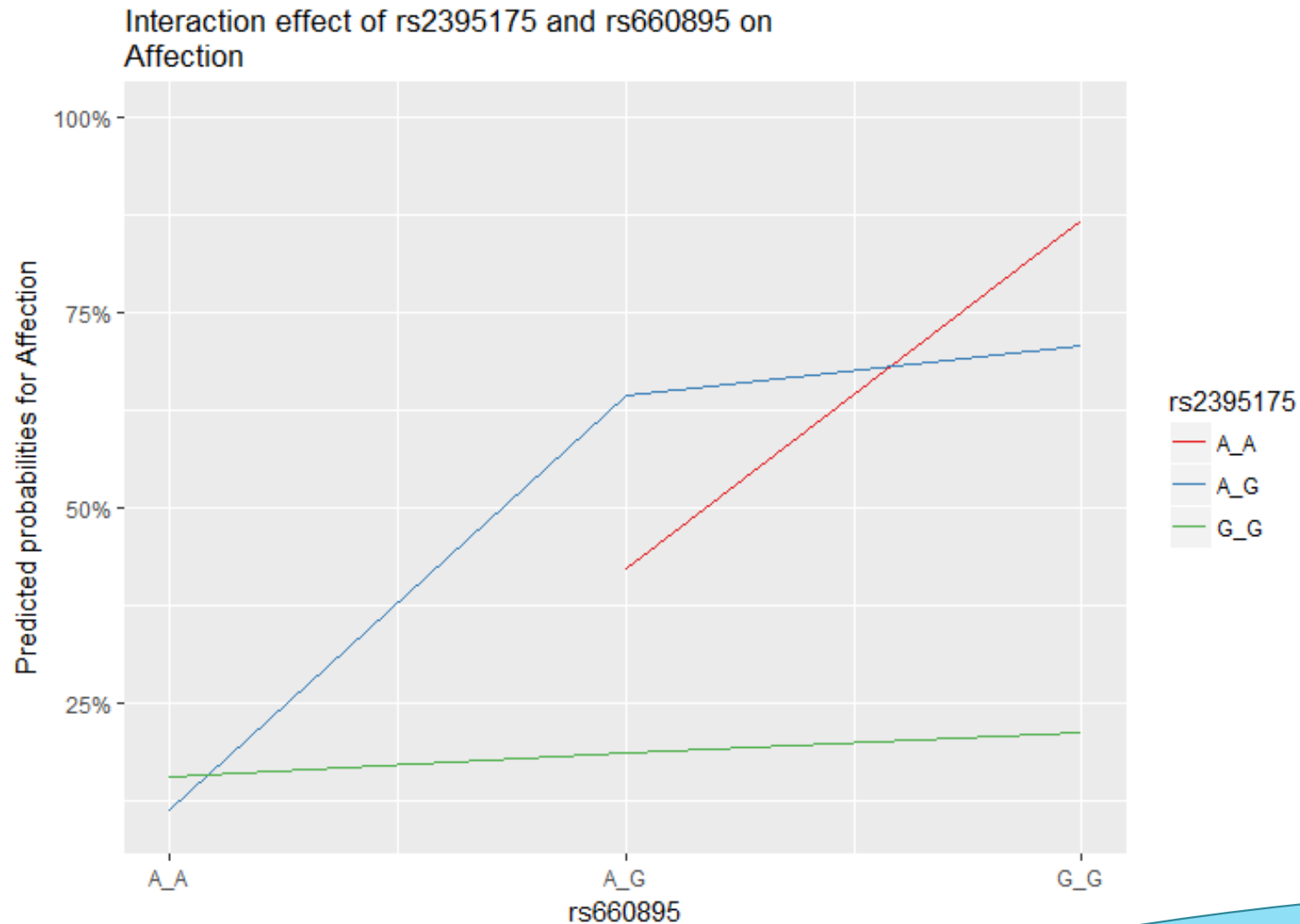
5. Conclusion: Do not reject H_0 , simpler model is adequate.

Logistic Regression Final model

| 變數 | 估計值 | P value | 變數 | 估計值 | P value |
|---|----------------------|------------|----------------------------|-------|------------|
| Intercept | 2.87 | 0.008 | rs2395175A_G:rs9275595A_G | 0.85 | 0.1 |
| rs2395175A_G | -4.13 | 0.0002 | rs2395175G_G:rs9275595A_G | 2.81 | 1.77*10^-7 |
| rs2395175G_G | -4.76 | 9.20*10^-7 | rs2395175A_G:rs9275595G_G | 1.93 | 0.002 |
| rs9275595A_G | -0.42 | 0.38 | rs2395175G_G:rs9275595G_G | 3.46 | 1.46*10^-7 |
| rs9275595G_G | -0.61 | 0.27 | rs2395175A_G:rs660895A_G | 4.47 | 0.0002 |
| rs660895A_G | -1.81 | 0.07 | rs2395175G_G:rs660895A_G | 2.02 | 0.05 |
| rs660895G_G | 0.38 | 0.65 | rs2395175A_G:rs660895G_G | 2.56 | 0.01 |
| rs2306420A_G | -0.11 | 0.59 | rs2395175G_G:rs660895G_G | NA | NA |
| rs2306420G_G | -0.6 | 0.006 | rs9369084A_G:rs12758854A_G | -0.23 | 0.58 |
| rs3129941A_G | -0.3 | 0.43 | rs9369084G_G:rs12758854A_G | -0.22 | 0.6 |
| rs3129941G_G | 0.16 | 0.66 | rs9369084A_G:rs12758854G_G | -2.34 | 0.007 |
| rs12758854A_G | 0.14 | 0.63 | rs9369084G_G:rs12758854G_G | -2.53 | 0.003 |
| rs12758854G_G | -0.17 | 0.56 | | | |
| rs1376679A_G | rs2395175*rs9275595 | | | | |
| rs1376679G_G | | | | | |
| rs10817045A_G | rs2395175*rs660895 | | | | |
| rs10817045G_G | | | | | |
| rs9369084A_G | rs9369084*rs12758854 | | | | |
| rs9369084G_G | | | | | |
| Null Deviance: 2806.8 2160 df Residual Deviance: 1895.4 2032 df AIC:1955.4 | | | | | |

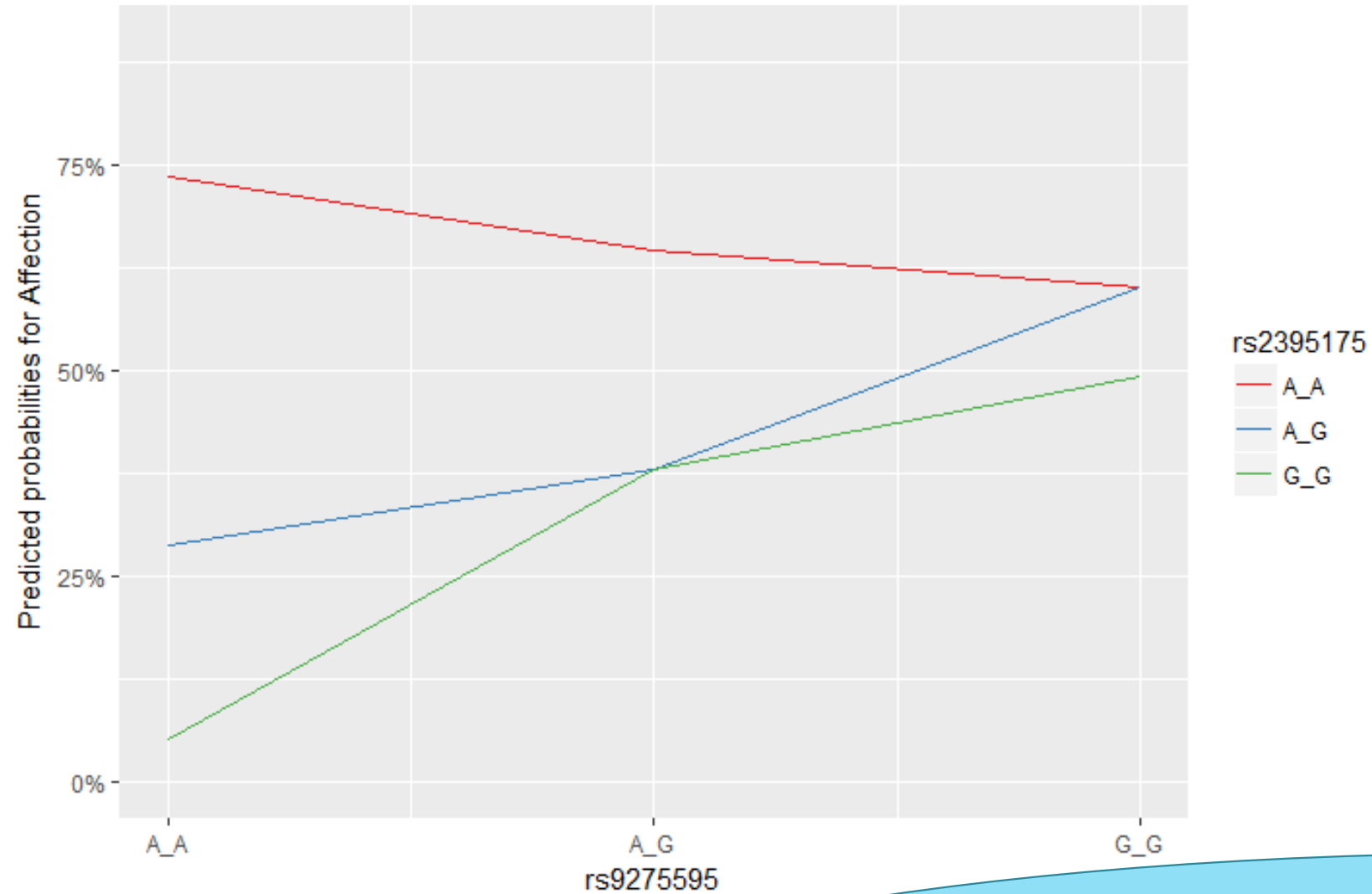
Interaction Plot

| rs2395175/ rs660895 | AA | AG | GG |
|------------------------|-----|-----|-----|
| AA | 0 | 16 | 186 |
| AG | 33 | 680 | 64 |
| GG | 907 | 170 | 6 |



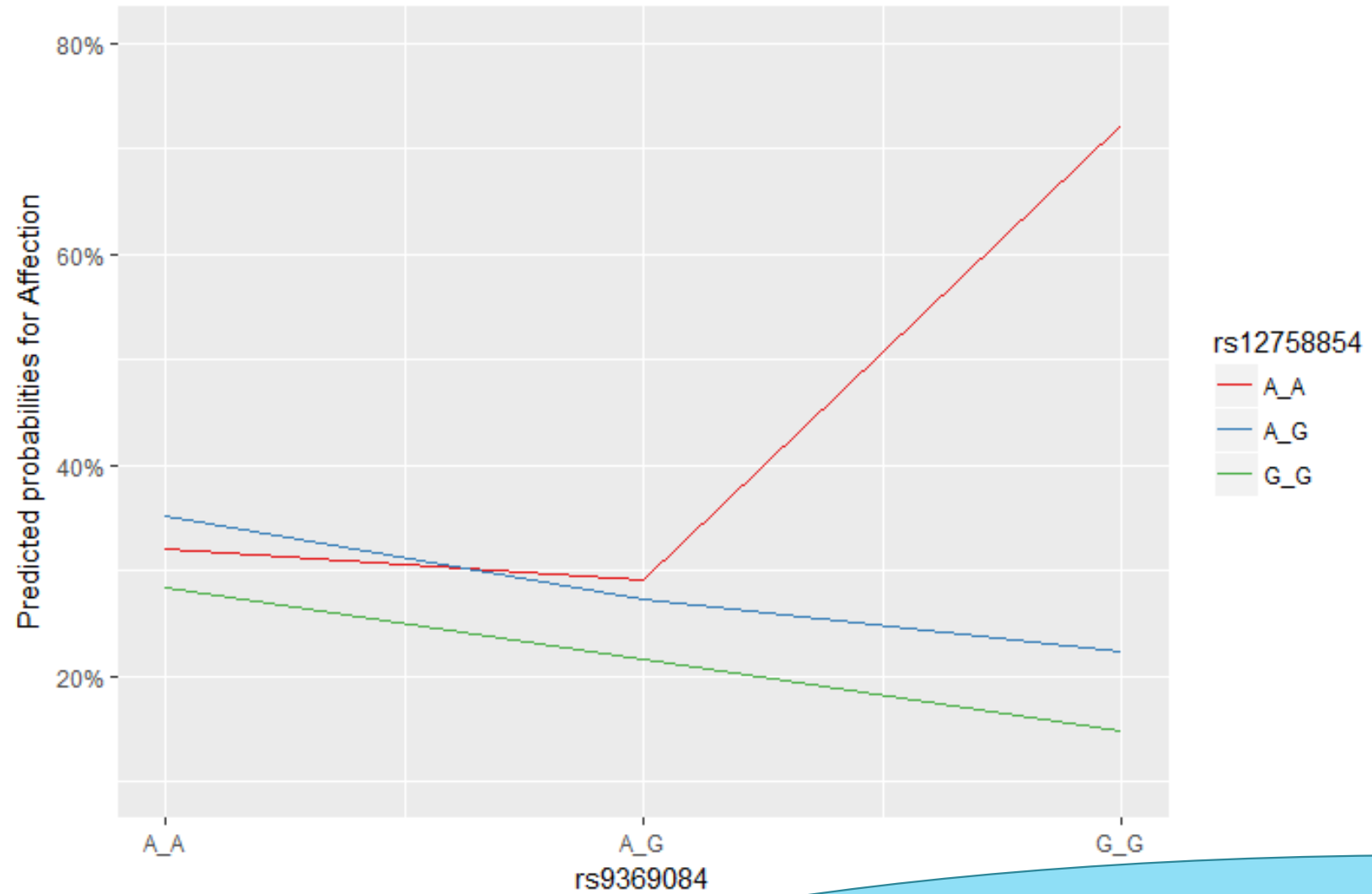
Interaction Plot

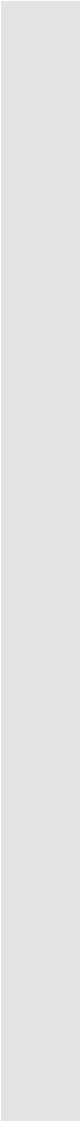

Interaction effect of rs2395175 and rs9275595 on Affection



Interaction Plot

Interaction effect of rs12758854 and rs9369084 on Affection





2.Cluster additive/ dominant coding

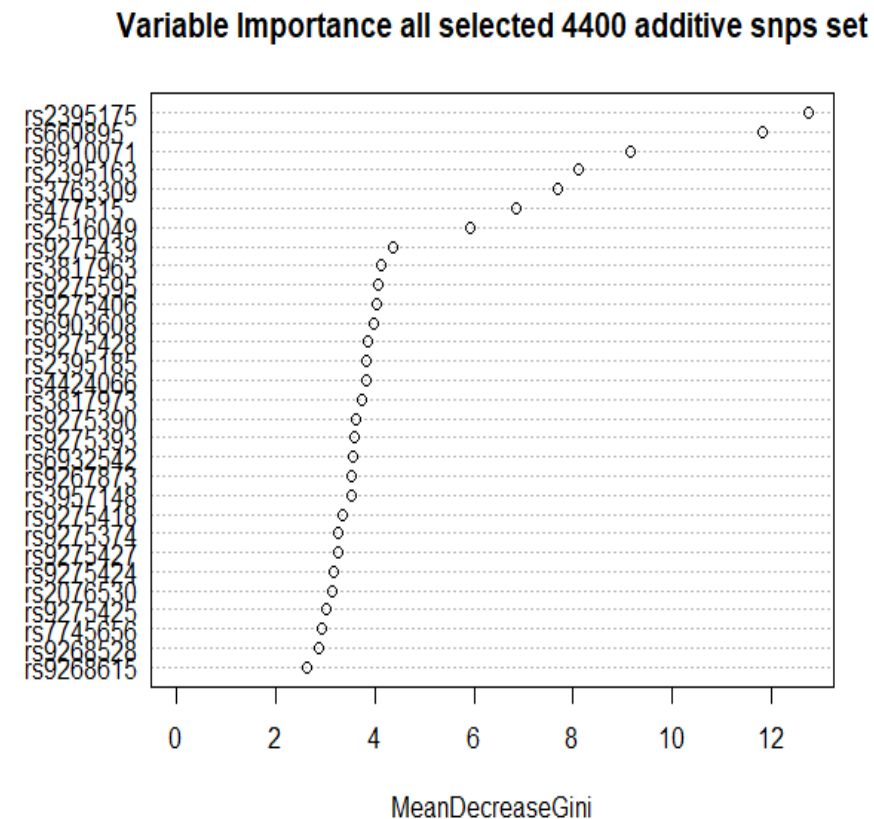
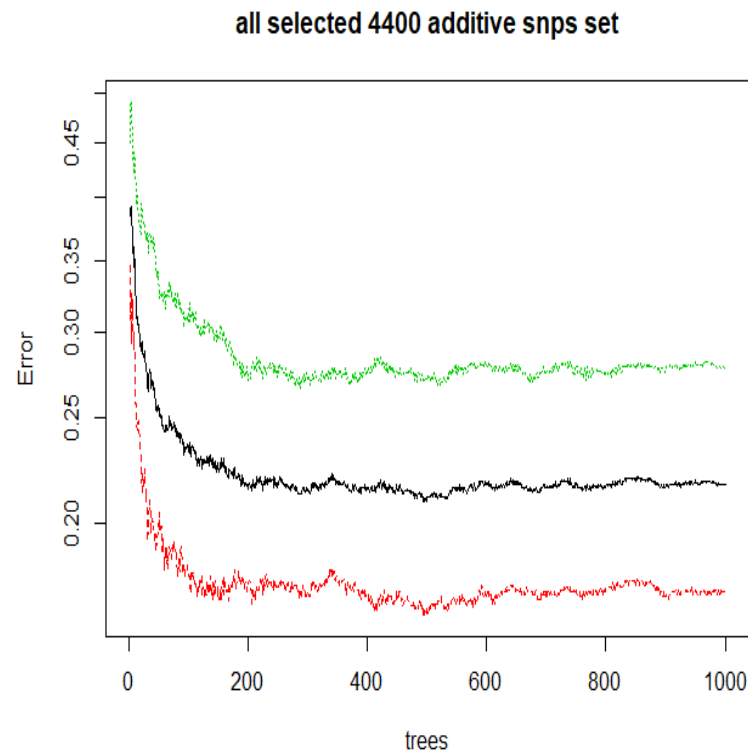
Result-additive coding

- Package:ranger-
fast version of 隨機森林
- 參數: ntree=5000
 mtry=5000
 自變數:433188 SNPs

| | 預測沒病 | 預測有患病 | 錯誤率 |
|-------------|------|-------|--------|
| 實際沒患病 | 976 | 218 | 0.1825 |
| 實際有患病 | 277 | 591 | 0.3191 |
| 整體錯誤率:24.0% | | | |

Global 4400
snps
Additive(2,1,0)
Coding
Accuracy=78.9%

Parameters:
1000 trees
m=66



| | 預測沒病 | 預測有患病 | 錯誤率 |
|-------------|------|-------|-------|
| 實際沒患病 | 990 | 204 | 0.178 |
| 實際有患病 | 231 | 637 | 0.267 |
| 整體錯誤率:21.1% | | | |

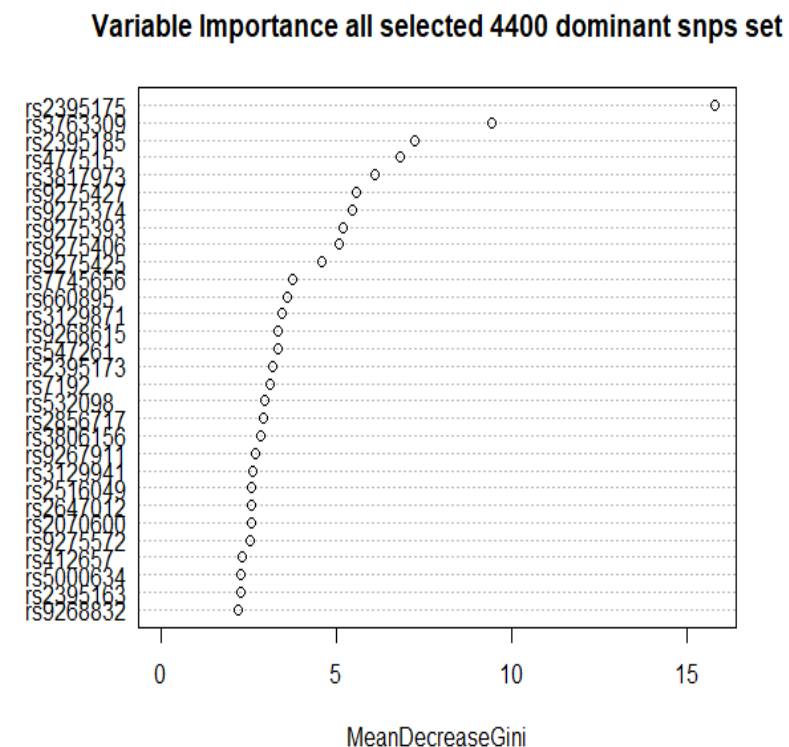
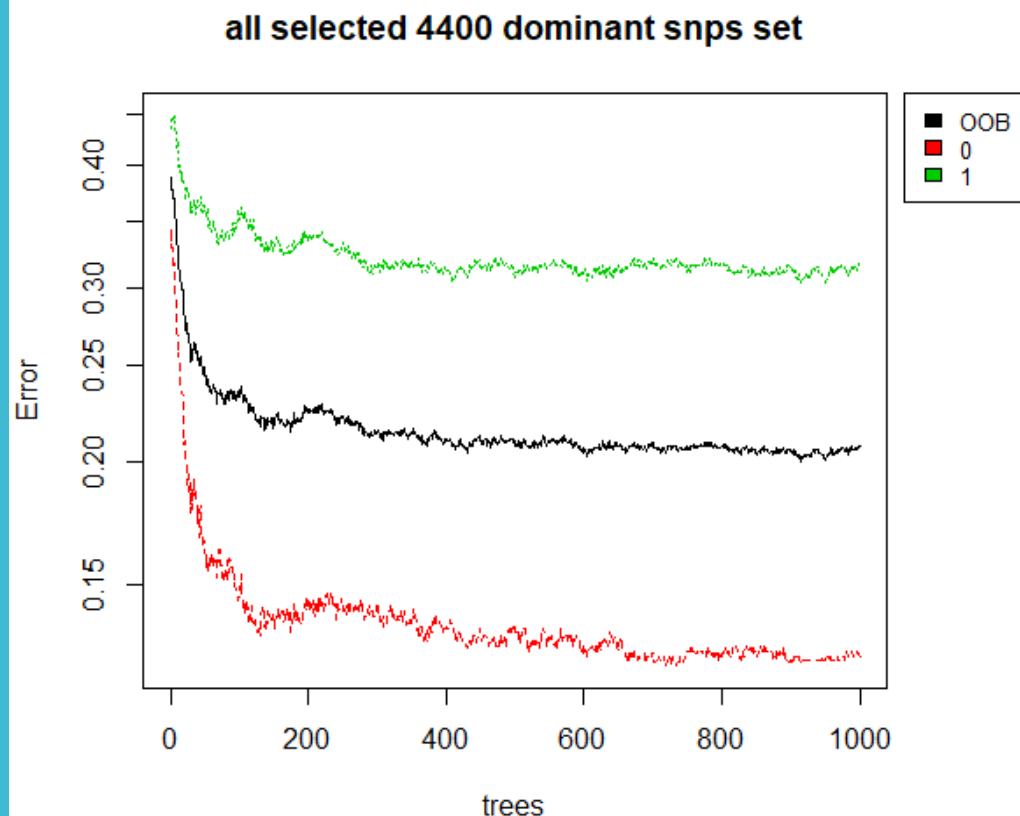
Result-Dominant coding

- Package:ranger-
fast version of 隨機森林
- 參數: ntree=5000
 mtry=5000
 自變數:433188 SNPs

| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|--------|
| 實際沒患病 | 986 | 208 | 0.1742 |
| 實際有患病 | 242 | 626 | 0.2788 |
| 整體錯誤率:21.82% | | | |

Global 4400
snps
Dominant(1,1,0)
Coding
Accuracy=
80.41%!
Finally>80%

Parameters:
1000 trees
m=66



| | 預測沒病 | 預測有患病 | 錯誤率 |
|--------------|------|-------|-------|
| 實際沒患病 | 1058 | 136 | 0.113 |
| 實際有患病 | 268 | 600 | 0.308 |
| 整體錯誤率:19.59% | | | |

結論

流程闡釋

Goal: 找到可能影響類風溼關節炎的snps，

提升預測準確率

●Preprocessing Data:

✓1.變異:

變異 < 5% 則刪掉

✓遺漏值:

遺漏值若大於1%刪掉

545080 SNPs → 433188 SNPs

●Analysis Tool:

✓Random Forest

1. 自動偵測交互作用
2. 高維資料計算效率高

●Ways:

✓資料切割或合併

1. Slice: 10000 筆資料
一單位 / 每單位取 100 snps
2. Cluster

✓Coding:

1. Original
2. Additive(2,1,0)
3. Dominant(1,1,0)
4. Recessive (2,1,1)

●Results Comparison

4400 SNPs錯誤率比較

Parameters:
1000 trees
m=66

未編碼(切割)4400SNPs

| 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|-------|--------|--------|--------|
| | 21.37% | 17.08% | 27.30% |

3

未編碼(cluster)4400SNPs

| 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|-------|--------|--------|-------|
| | 21.47% | 17.38% | 27.2% |

Additive(切割)4400SNPs

| 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|-------|--------|-------|-------|
| | 21.63% | 17.5% | 27.3% |

2

Additive(cluster)4400SNPs

| 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|-------|--------|--------|--------|
| | 21.13% | 16.92% | 26.91% |

Dominant (切割)4400SNPs

| 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|-------|--------|--------|--------|
| | 20.73% | 12.59% | 31.93% |

1

Dominant(cluster)4400SNPs

| 平均錯誤率 | 整體 | 沒患病 | 有患病 |
|-------|--------|--------|--------|
| | 19.86% | 11.87% | 30.81% |

前5重要 snps 與患病與否 列聯表 (two way contingency table)

| rs2395175/ Affection | Yes | No |
|-------------------------|------|------|
| AA | 0.83 | 0.17 |
| AG | 0.64 | 0.36 |
| GG | 0.19 | 0.81 |

| Odds Ratio(A) rs2395175 | |
|----------------------------|-------|
| AA/GG | 20.81 |
| AG/GG | 7.578 |
| AA/AG | 2.74 |

| rs660895/ Affection | Yes | No |
|------------------------|------|------|
| AA | 0.17 | 0.83 |
| AG | 0.57 | 0.43 |
| GG | 0.82 | 0.18 |

| Odds Ratio(G) rs660895 | |
|---------------------------|-------|
| GG/AA | 22.24 |
| AG/AA | 6.47 |
| GG/AG | 3.43 |

| rs6910071/ Affection | Yes | No |
|-------------------------|------|------|
| AA | 0.19 | 0.81 |
| AG | 0.58 | 0.42 |
| GG | 0.79 | 0.21 |

| Odds Ratio(G) rs6910071 | |
|----------------------------|-------|
| GG/AA | 16.03 |
| AG/AA | 5.88 |
| GG/AG | 2.72 |

| rs2395163/ Affection | Yes | No |
|-------------------------|------|------|
| AA | 0.19 | 0.81 |
| AG | 0.57 | 0.43 |
| GG | 0.79 | 0.21 |

| Odds Ratio(G) rs2395163 | |
|-------------------------|-------|
| GG/AA | 16.03 |
| AG/AA | 5.65 |
| GG/AG | 2.83 |

| rs3763309/ Affection | Yes | No |
|-------------------------|------|------|
| AA | 0.77 | 0.23 |
| AC | 0.57 | 0.43 |
| CC | 0.19 | 0.81 |

| Odds Ratio(A) rs3763309 | |
|-------------------------|-------|
| AA/CC | 14.27 |
| AC/CC | 5.65 |
| AA/AC | 2.52 |

| Dominant Coding | 預測沒病 | 預測有患病 | 錯誤率 |
|-----------------|------|-------|-------|
| 實際沒患病 | 1058 | 136 | 0.113 |
| 實際有患病 | 268 | 600 | 0.308 |
| 整體準確率:80.41% | | | |

符合之前分析結果:

Dominant Coding(只要有一個出現較少次的含氮鹼基就分一類)最好
Recessive Coding(是否出現兩個較少次的含氮鹼基才會影響)最差

影響疾病

| rsID | Chromosome | SNP | Associated |
|-------------|------------|-----|----------------|
| "rs2395175" | 6 | A/G | 類風濕關節炎 |
| "rs660895" | 6 | A/G | 類風濕關節炎、免疫球蛋白腎病 |
| "rs6910071" | 6 | A/G | 類風濕關節炎 |
| "rs2395163" | 6 | C/T | 類風濕關節炎、帕金森氏症 |
| "rs3763309" | 6 | A/C | 類風濕關節炎 |
| "rs9275595" | 6 | C/T | 類風濕關節炎、BMI、肥胖 |

- 找到的前幾個snps
已被證實與類風溼關節炎有關
- 找到的前6個最重要的snps跟
- A genome-wide association scan for rheumatoid arthritis data by Hotelling's T^2 tests
這篇文章的結果一樣，且與許多篇文章找到的 snps有大量重疊。

Appendix User Time 2nd analyzation

User time:讀進整筆資料:96secs.

User time:分割資料:520secs.

User time:讀進分割資料:60secs.

User time:分割資料做隨機森林:137secs./each.
總共做了44次

User time:合併選出來共4400SNPs做隨機森林:40 secs.

User time:每個切半的變數個數做50次cv總共做650次
(13項變數個數): 2hr:59min:44secs.

User time:用前137個變數做隨機森林:20" 97secs.

Appendix
User time-
3rd analyzation

● User time:recode SNPs /10000筆資料:90secs.

● User time:每個切半的變數個數做50次cv
總共做650次(13項變數個數):10468.6secs.

● User time:cluster(433188 SNPs)
random forest:76797.556 secs.

Appendix
User time-
Final
analyzation

● User time:cluster(433188 SNPs dominant coding) random forest:142625.247 secs.

● User time:cluster(433188 SNPs additive coding) random forest:134647.439 secs.

References

- **Textbook:** An introduction to statistical learning with Applications in R-Chapter 8 Tree-Based Methods

- **Variable Selection Article:** Variable Selection in Random Forest with Application to Quantitative Structure-Activity Relationship

- **Websites:** GWAS Catalog

<https://www.ebi.ac.uk/gwas/search?query=Rheumatoid%20arthritis>

- **Recode snps:** R package: "SNPassoc"

- **Interaction Term Detection:** Methods for Interaction Detection in Predictive Modeling Using SAS-2012