生物晶片的基本分析流程(3hr)

R的package安裝及基本操作(2hr)

如何用R操作生物晶片的基本分析流程(3hr)

與生物資料庫的連結(3hr)

生物晶片的基本分析流程

注意事項

1. 有問題可以立刻舉手

2. 如果覺得講的太快，或哪裡聽不懂，也是直接跟我說

3. 如果覺得很累…下課休息一下，也是一樣…反正教材沒有很多

4. 如果想上廁所…就直接走出去…不用說喔!!

Bioinformatics!!

Data

project: **gingival_h**

What's MOLAS?

only for lab use now

Select your database

name	type	version	taxid	description
wssv	agilent	shrimp	6687	p.m. infected by wssv
gingival_h	agilent	human	9606	2 patients of gingival_hyperplasia and 5 Normal sample pool
human_sc	agilent	human	9606	condition A,B,AP
mouse_sc	affy	mouse	10090	compare C4,C8 and ICM transcriptome

Why??

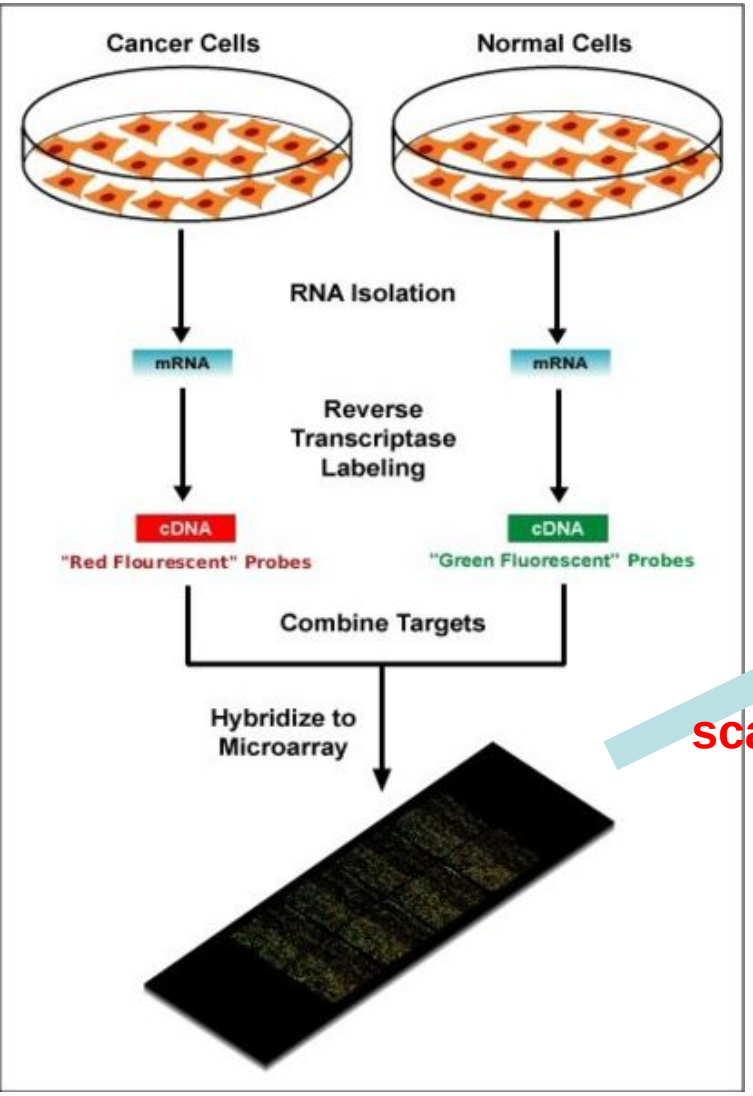
The R environment

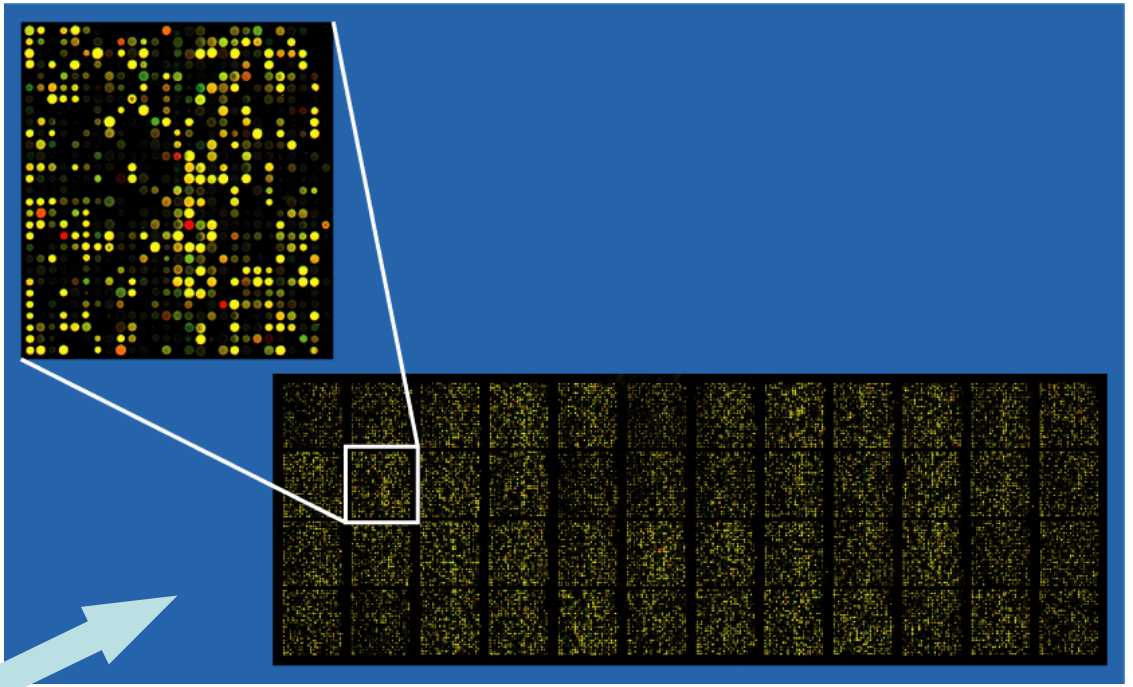
1. R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes
2. an effective data handling and storage facility,
3. a suite of operators for calculations on arrays, in particular matrices,
4. a large, coherent, integrated collection of intermediate tools for data analysis,
5. graphical facilities for data analysis and display either on-screen or on hardcopy, and
6. a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R.

BIOCONDUCTOR
open source software for bioinformatics

Two-color cDNA microarray





scanner

Image processing (Ex. GenePix)

.gpr files(table format)

Import to *R*

Two data in one spot.

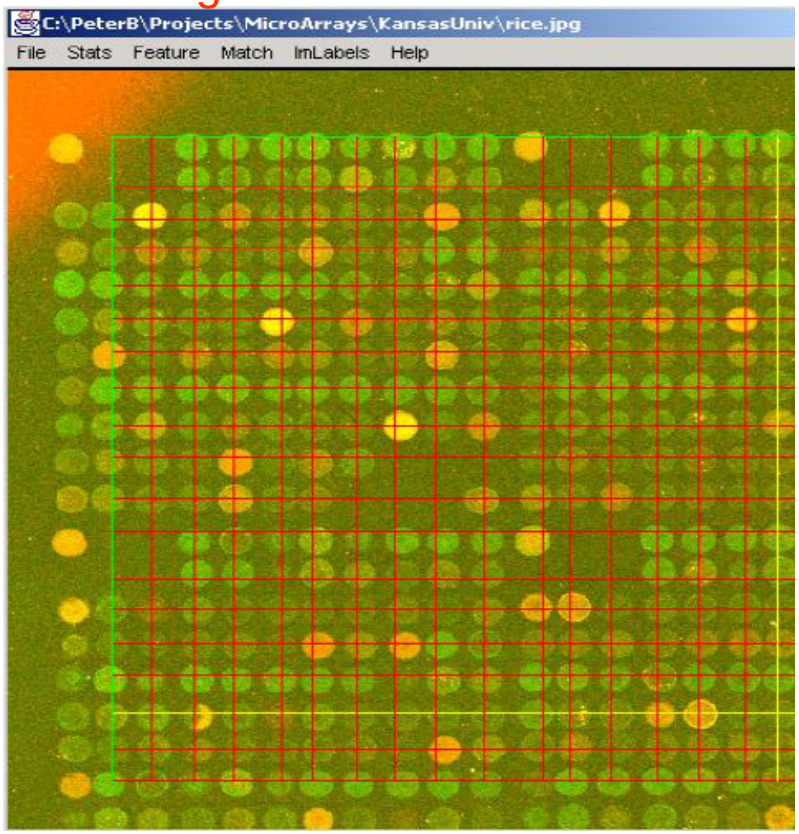
Red channel (Cy5) from “Experimental set”

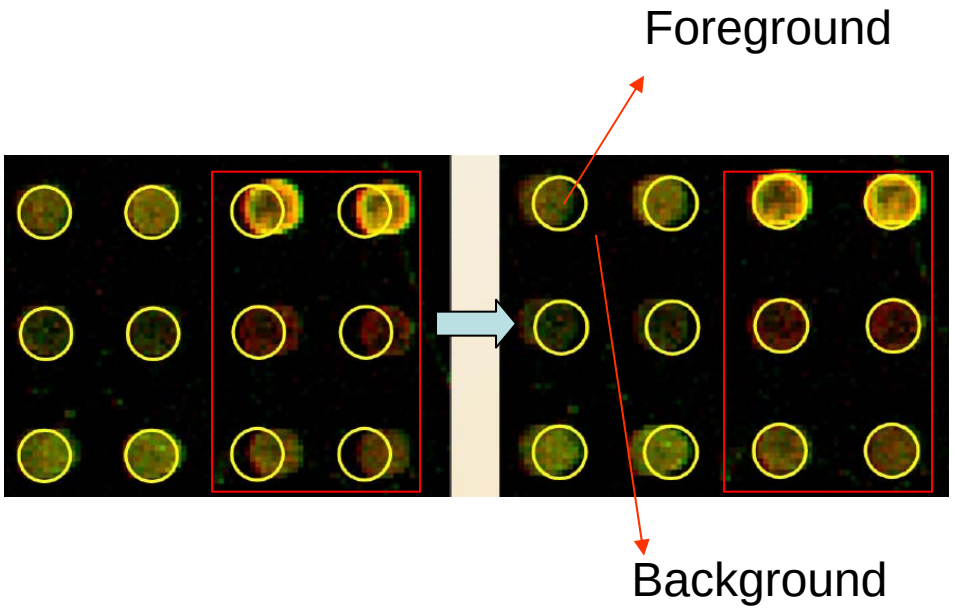
Green channel (Cy3) from “Control set”

Probes from cDNA (complementary) that be amplified by PCR

Image processing

1. Grid alignment





2. Color filter: Separate Cy5 and Cy3 signals

3. Count the foreground and background intensity of each channel

Medium
Mean
Standard deviation
Coefficient of variation

From EURASIP Journal on ASP Vol. 2006, Pages 1-13

Data processing

Raw data files: gpr (GenePix); .spot (Spot), .xls (SMD), .txt (Agilent)

Read raw data to R:

Table format data

資料編輯器

	site_index	hr10_F	hr10_B	hr0_F	hr0_B	hr12_F
1	1	3352.983	58.88606	1933.845	76.30885	2642.689
2	2	63.2931	58.88105	89.55172	75.45968	58.74627
3	3	68.66667	58.81323	104.9	75.15175	56.16667
4	4	57.39683	58.56445	85.66667	74.76758	56.52542
5	5	63.01587	59.67131	94.2381	76.99203	65.05172
6	6	63.96429	59.55289	93.60714	78.06986	57.40678
7	7	59.58929	58.90982	86.625	76.81563	61.56923
8	8	59.81034	57.9498	84.25862	74.67068	59.21053
9	9	64.85965	59.38416	96.5614	75.40396	60.57627
10	10	61.98182	58.2835	84.03636	73.63107	59.58182
11	11	64.54386	58.17255	89.49123	72.38824	58.56667
12	12	147	58.87174	169.4262	74.59118	138.6032
13	13	121.4333	58.93359	168.8	75.56641	93.44444

marrayRaw

marray package

RGList:

limma package

```
@maLabels
[1] "control" "control" "control" "control" "control"
8443 more elements ...
@maInfo
             ID  Name
control   control geno1
control.1 control geno2
control.2 control geno3
@maNotes
[1] ""
```

1. 容易對資料做較廣泛的運算
2. 完成分析的程式碼較長
3. 容易匯出在其他軟體下作業

1. 可用有限的簡單指令完成分析
2. 完成分析的程式碼較短
3. 不容易匯出在其他軟體下作業

Microarray analysis steps:

Biological experiments

Raw file_list	*condition_R and condition_G*

Preprocess microarray data

Remove statistical error

Statistical test to find DEGs

(Differentially Expressed genes)

Target

Biological meaning

Outline

Biological experiments
(WSSV infection of P. vannamei)

file_list	condition_list
5489_251841210001_1.txt	10hr_0hr
5491_251841210001_2.txt	12hr_1hr
5493_251841210001_3.txt	14hr_2hr
5495_251841210001_4.txt	18hr_3hr
5497_251841210002_1.txt	24hr_4hr
5499_251841210002_2.txt	30hr_5hr
5501_251841210002_3.txt	36hr_6hr
5503_251841210002_4.txt	48hr_8hr

Preprocess microarray data

1. Choice intensity index (Mean or Median)
2. Filtration (0hr_virus intensity)
3. Background adjustment (No substrate)
4. Normalization (Quantile Normalization)

Statistical test to find DEGs (Differentially Expressed genes)

ANOVA
Cluster

Target

1. Find the expression pattern of virus gene:
   →try to find the mechanism of virus attach

2. Find the changed genes of host:
   →which pathway that virus use it
   →which pathway that virus shutdown it

Choice intensity index (Mean or Median)

既然，intensity有mean及medium兩種數值，那我們應該採取哪一種呢??

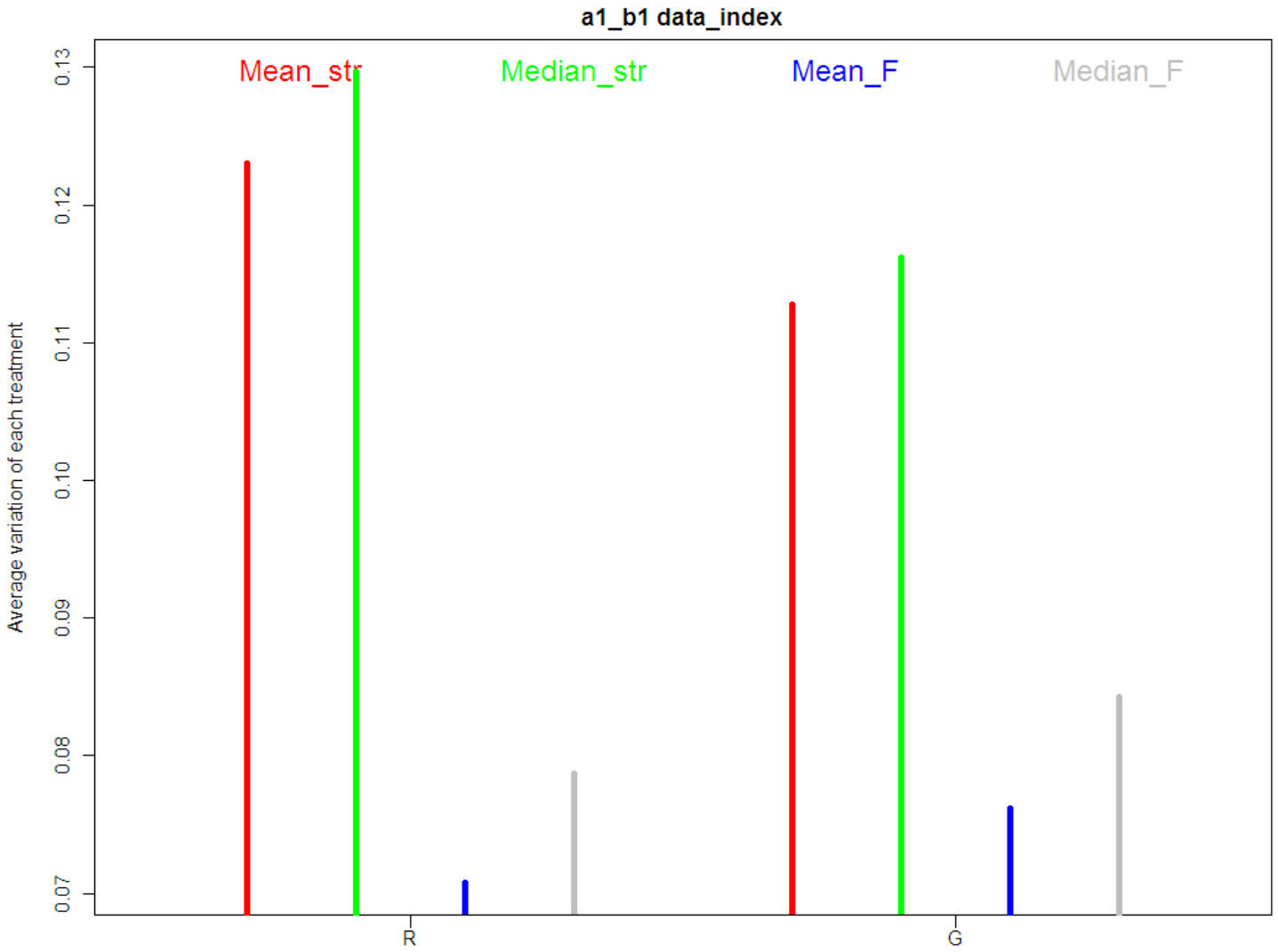
通常microarray上的probe會有重複的設計(這是在gene量較少的物種中)，而每一個點的intensity又有mean與median兩種數值，我們可以利用重複probe intensity在mean及median的再現性程度來決定我們要選用mean或median(如果以obeject的方式讀入，通常是以mean來做probe的intensity)

Name	F635 Median	F635 Mean	B635 Median	B635 Mean	F532 Median	F532 Mean	B532 Median	B532 Mean
PmTwI08H09	3355	3112	101	104	2818	2738	101	104
PmTwI08H09	3000	2899	102	106	2553	2539	108	110
PmTwI08H09	3121	2957	103	108	2598	2494	111	113
PmTwI11F10	480	541	110	111	564	581	113	116
PmTwI11F10	467	530	114	120	563	601	114	123
PmTwI11F10	488	510	114	120	550	579	112	121
PmTwI13H08	195	199	114	118	246	250	110	115
PmTwI13H08	180	191	114	119	232	236	107	114
PmTwI13H08	186	188	113	116	223	234	104	107
PmTwI09A08	356	355	116	119	403	426	102	106
PmTwI09A08	322	310	115	117	356	365	100	103
PmTwI09A08	350	358	116	119	398	418	101	104

Foreground background Foreground background

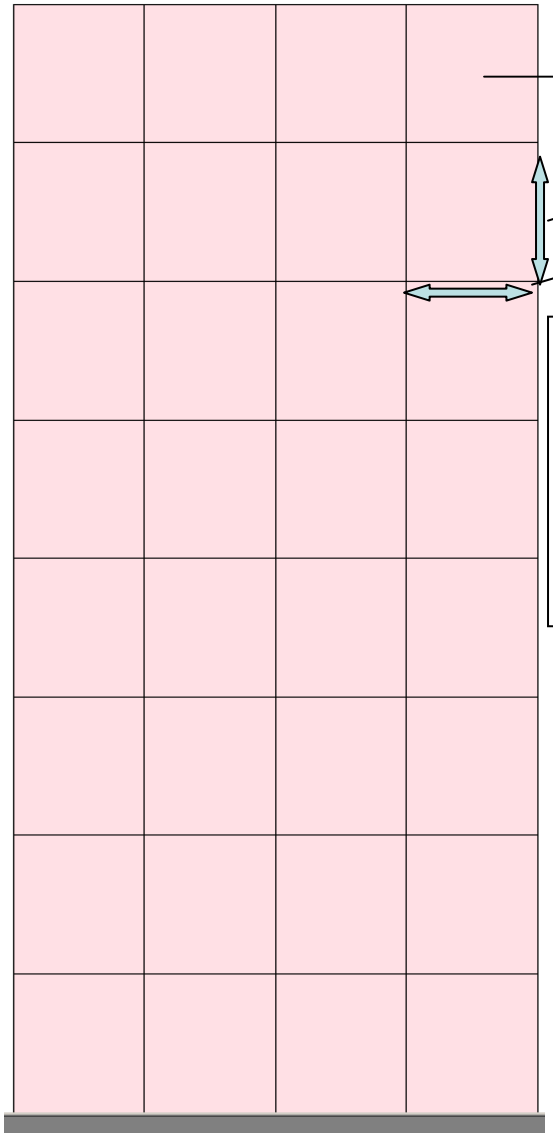
variation

Choice intensity index (Mean or Median)



Channel	Mean_str	Median_str	Mean_F	Median_F
a1	0.123006	0.129668	0.070779	0.07868
b1	0.112745	0.116168	0.076187	0.084245

Quality Assessment(1)



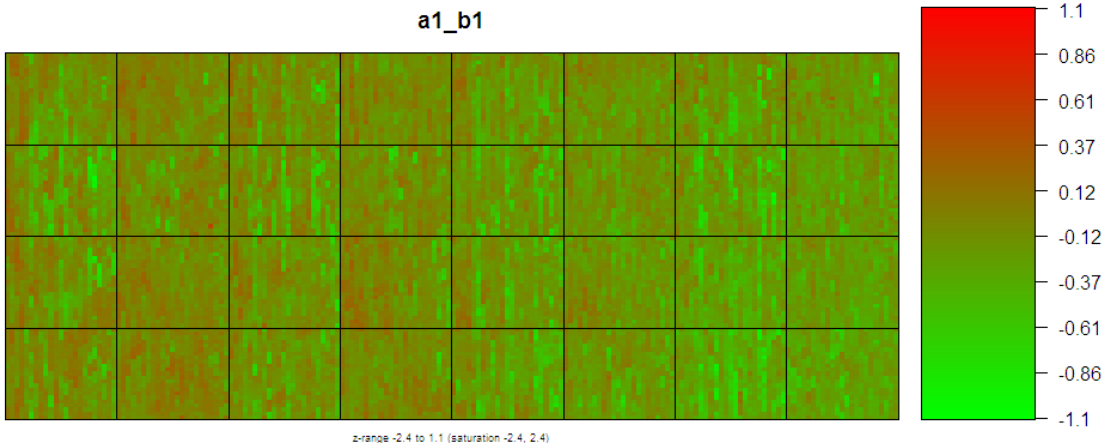
Block=32
Row=23
Column=23
Spot=32*23*23=16928

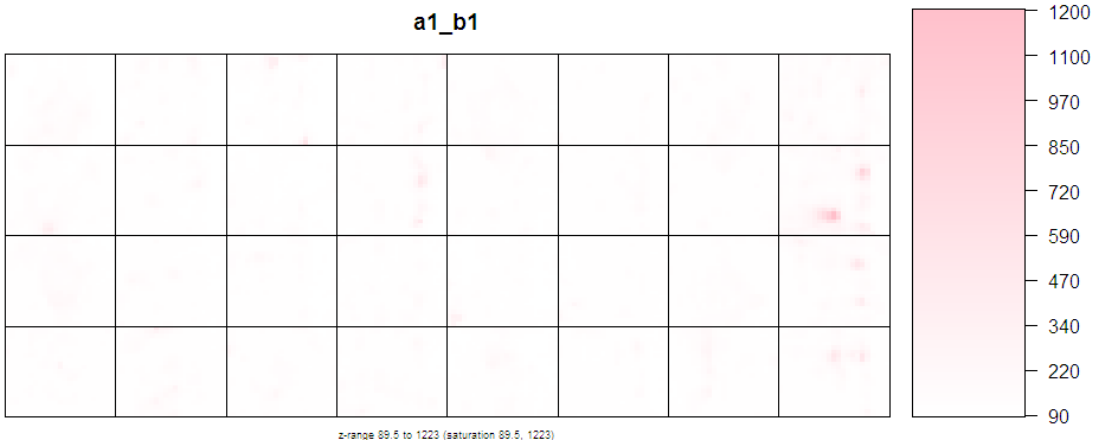
右圖是genepix的microarray format，基本上是由32個blocks(***Print-tip***)，每個block又有529個probe點所組成
因為我們已經決定了intensity的參考值，所以接下來便是檢視image會不會有什麼錯誤，同時可以將array intensty做輸出

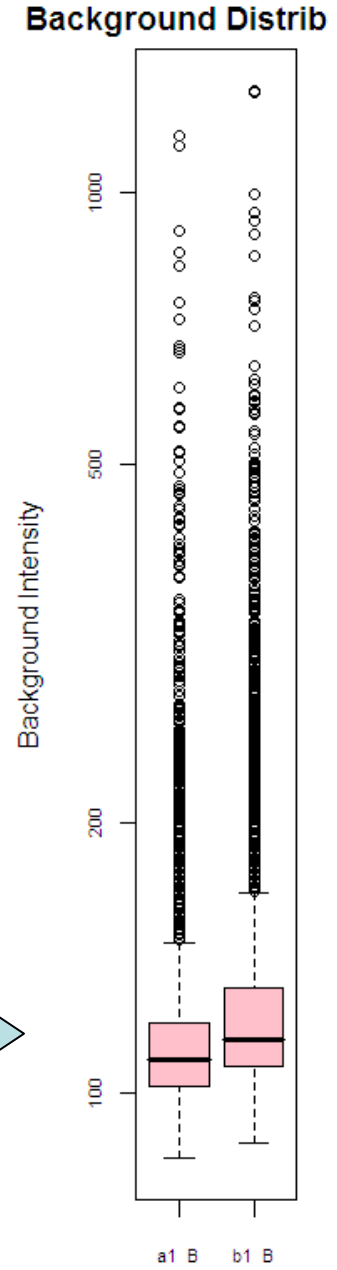
Quality Assessment(1)

檢視重點:

1. 在image中可一看一下有沒有區域性相同顏色或缺值，如果有可能是遭受的物理性的破壞
2. 另外在background的intensity範圍也不能過大，如果太大可能是因為probe spot的定位可能不甚準確

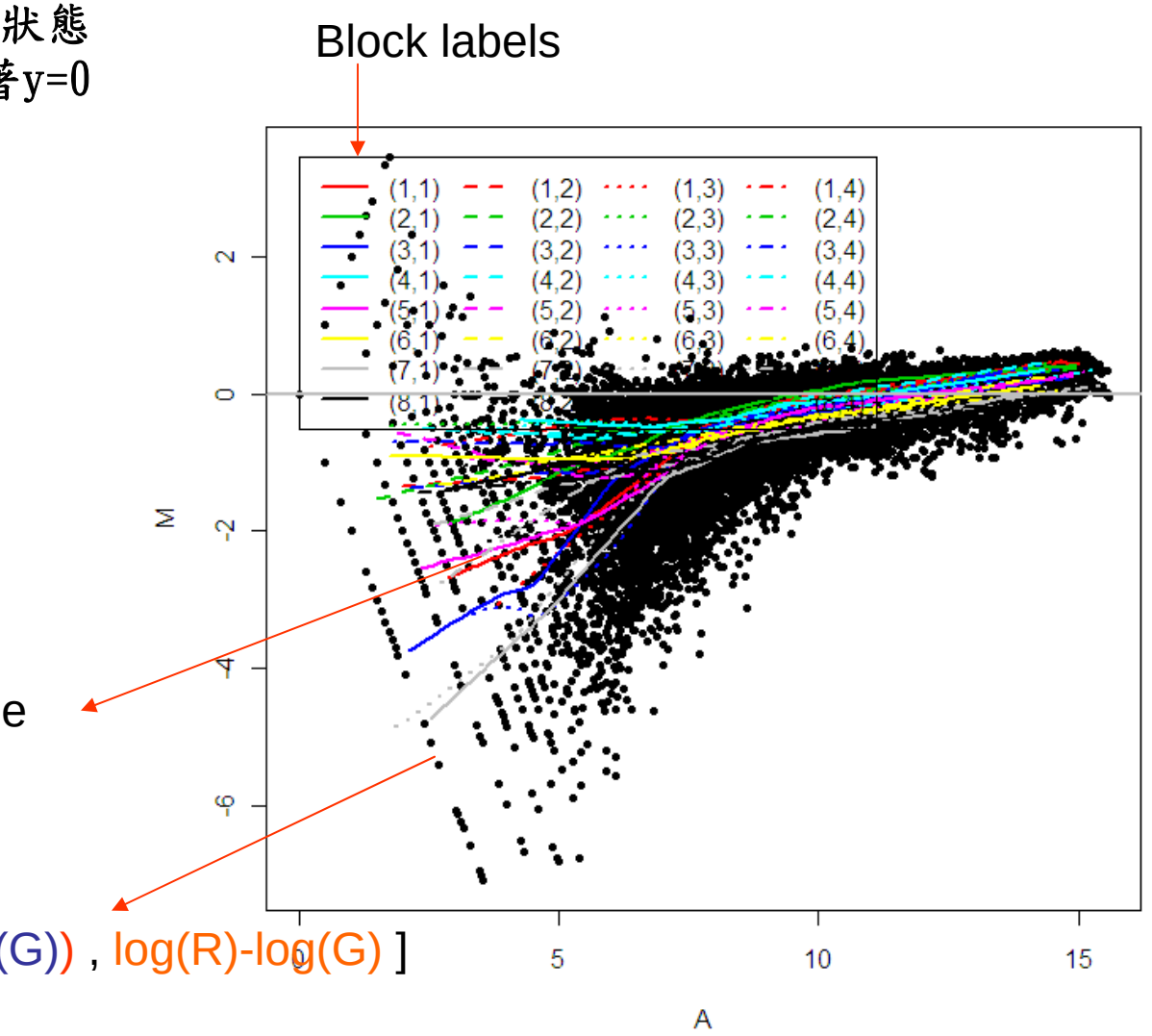






Quality Assessment(2)

最基本的maplot，這是在array分析中常看到的圖，主要是在看R/G(M)是否會隨著intensity的總強度(A)而改變，最好的狀態是大部分的gene分布是隨著y=0的直線分布



每一個block的loess line

Spot intensity[½(log(R)+log(G)) , log(R)-log(G)]

Filtration

可以根據raw data的狀況先濾掉不準確的點，例如: 1.foreground接近background的點，或是2.在重覆點之中variation過大的點，以避免影響後續進行test的結果。

Normalization

Preprocess的最後步驟也是最重要的步驟就是**Normalization**，**Normalization**的主要目的是在於將操作**array**所產生的**error**減到最低，而主要的**error**有下列幾項**:**

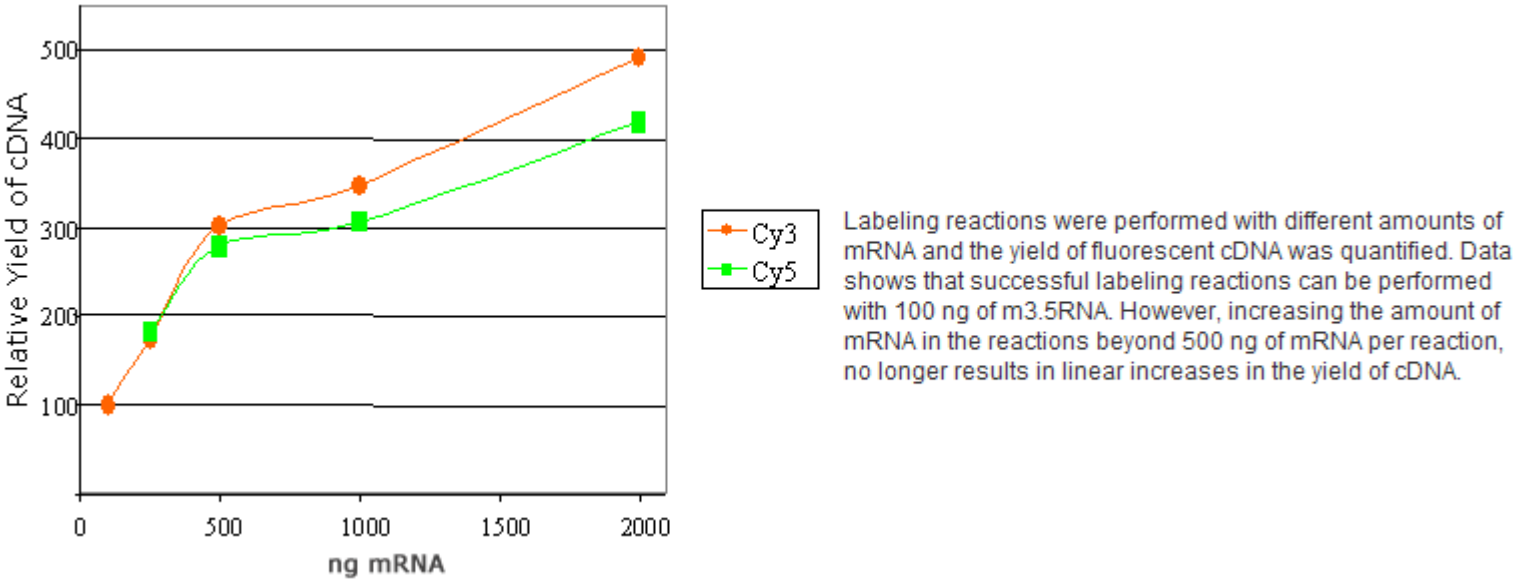
1. Dyes effects

2. Scanning parameter…between arrays

3. Print-tip different

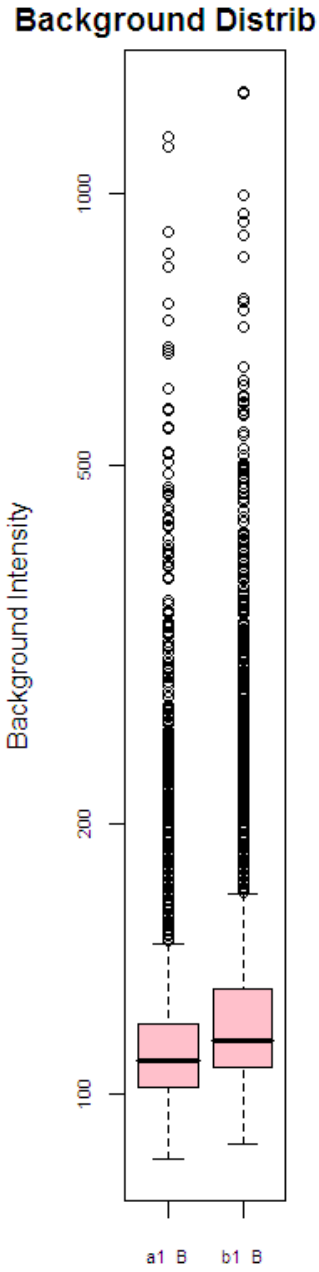
4. Spatial effects

Dyes effects

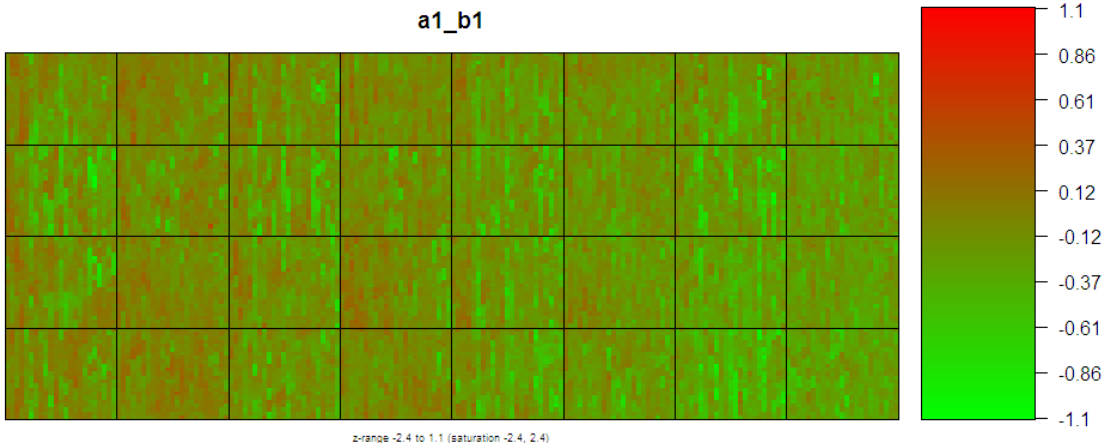


Labeling reactions were performed with different amounts of mRNA and the yield of fluorescent cDNA was quantified. Data shows that successful labeling reactions can be performed with 100 ng of m3.5RNA. However, increasing the amount of mRNA in the reactions beyond 500 ng of mRNA per reaction, no longer results in linear increases in the yield of cDNA.

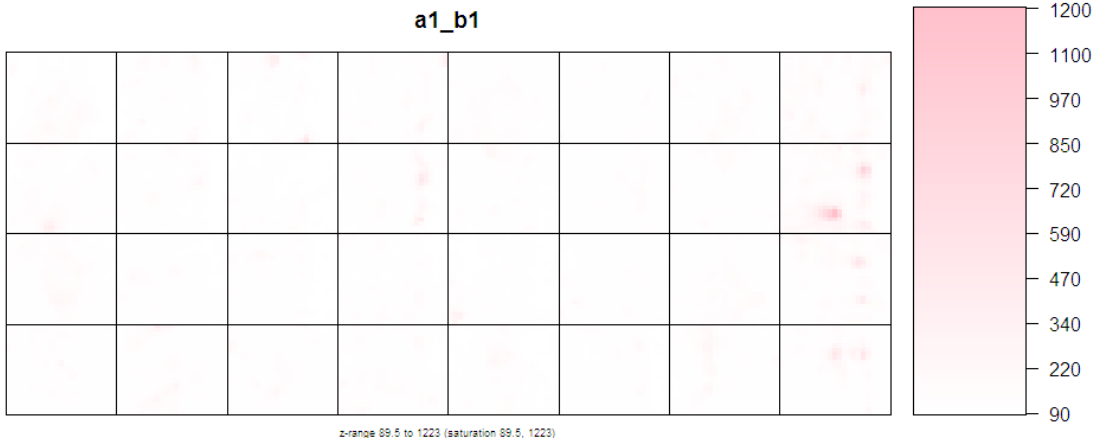
Scanning parameter…between arrays



Print-tip different

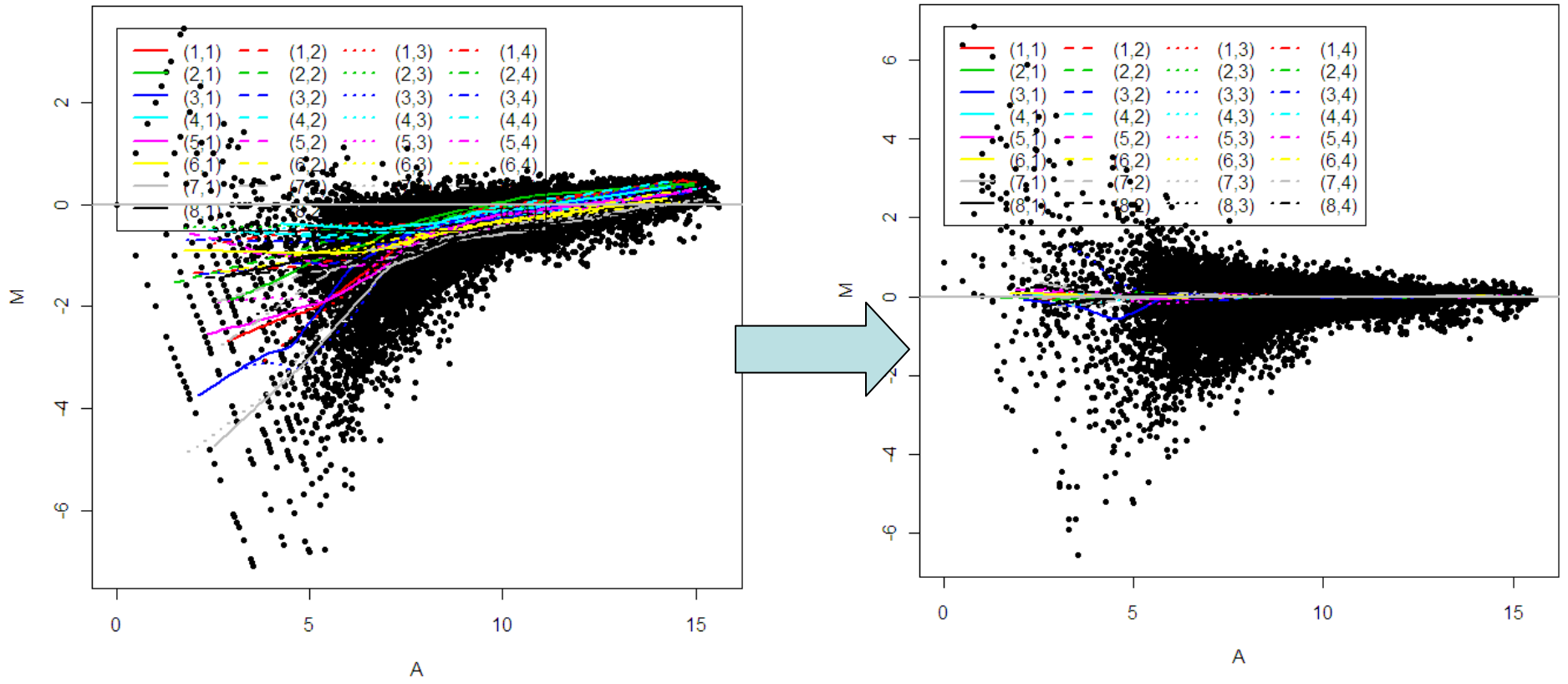


Spatial effects



Normalization type

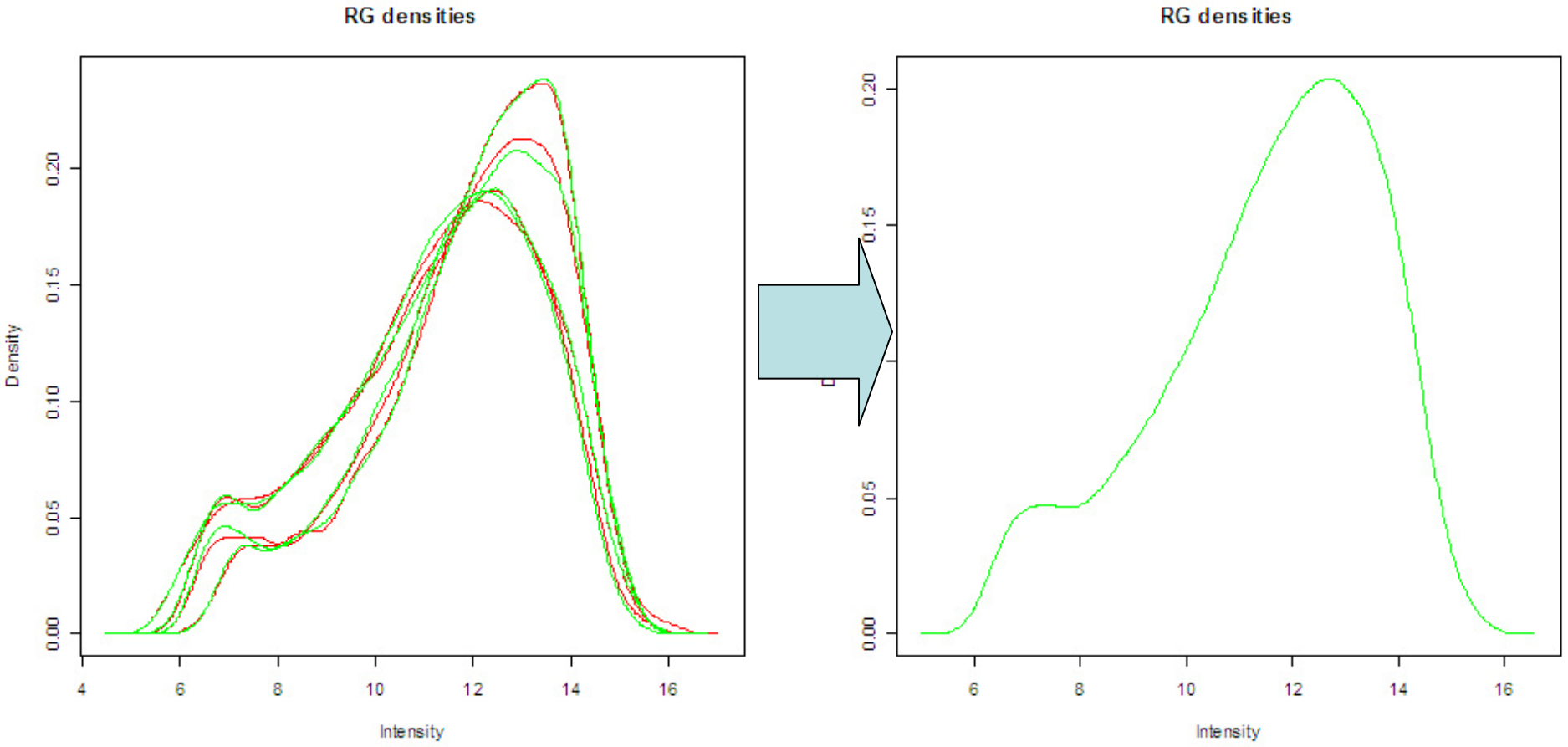
1. Within array　侷限在單張array裡面spot intensity的調整，在預設的假設上做單片的調整。
   下面的MA圖即是做Loess Normalization的array其改變狀況，所做的假設便是所有的print-tip的loess line應該都跟y=0這條線重合



Normalization type

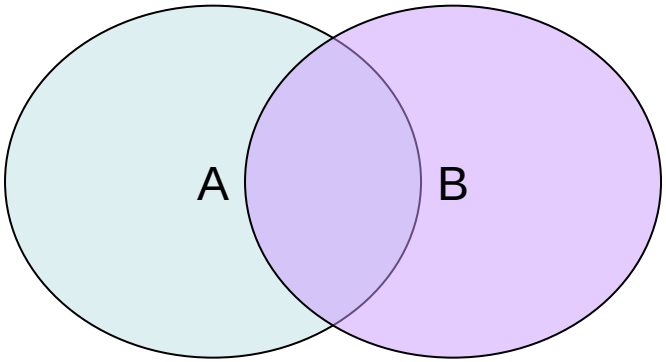
2. Between array　就多張array所做的整體性調整，通常是因為不同array在互相比較時，raw data中的control intensity就已經有差別而用以互相調整的方式，可以使後續的test不會受到data distribution的差別而失去篩選效果

Quantile Normalization:假設不同array的intensity分布相同



Statistical test to find DEGs

Independent T test

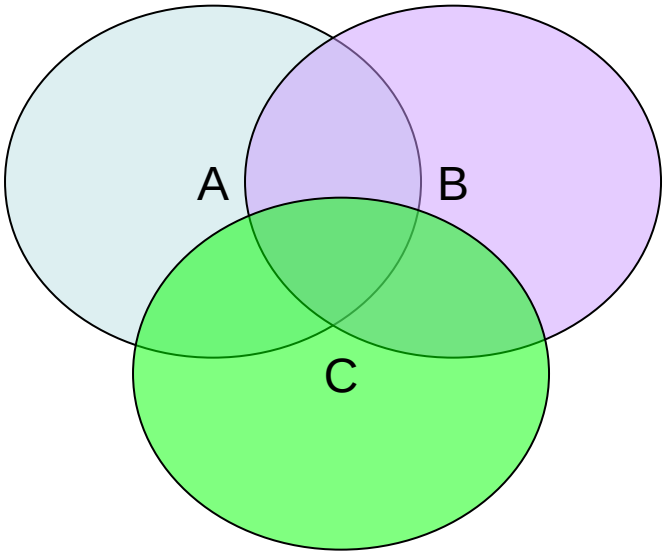


基於A取樣資料群及B取樣資料群的分布狀況，來計算A $_{mean}$及B $_{mean}$相等的機率

在microarray上則是依據實驗變因可以分成兩組data的數據，在經過Normalization之後，分別取出各gene的intensity值，計算他們可能為同一分布的機率…

Statistical test to find DEGs

ANOVA (**An**alysis **o**f **va**riance)



基於三群以上的取樣資料分布狀況，來計算這些資料相等的機率

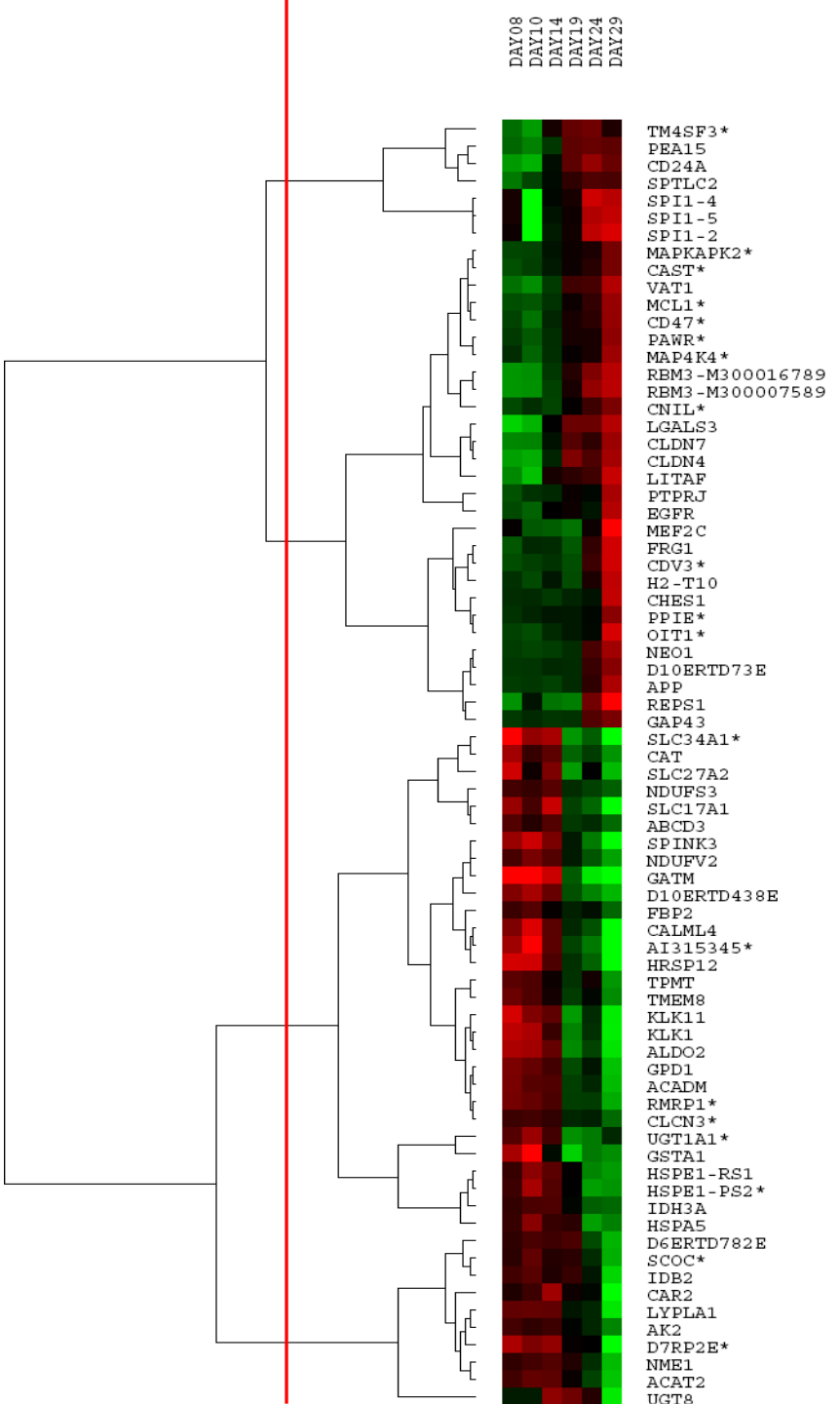
在microarray上則通常用在時間序列資料或是三種以上不同處理的資料，計算變因對基因不產生影響的機率

利用合理的p value threshold 去產生DGEs，再進行之後的annotation

Clustering

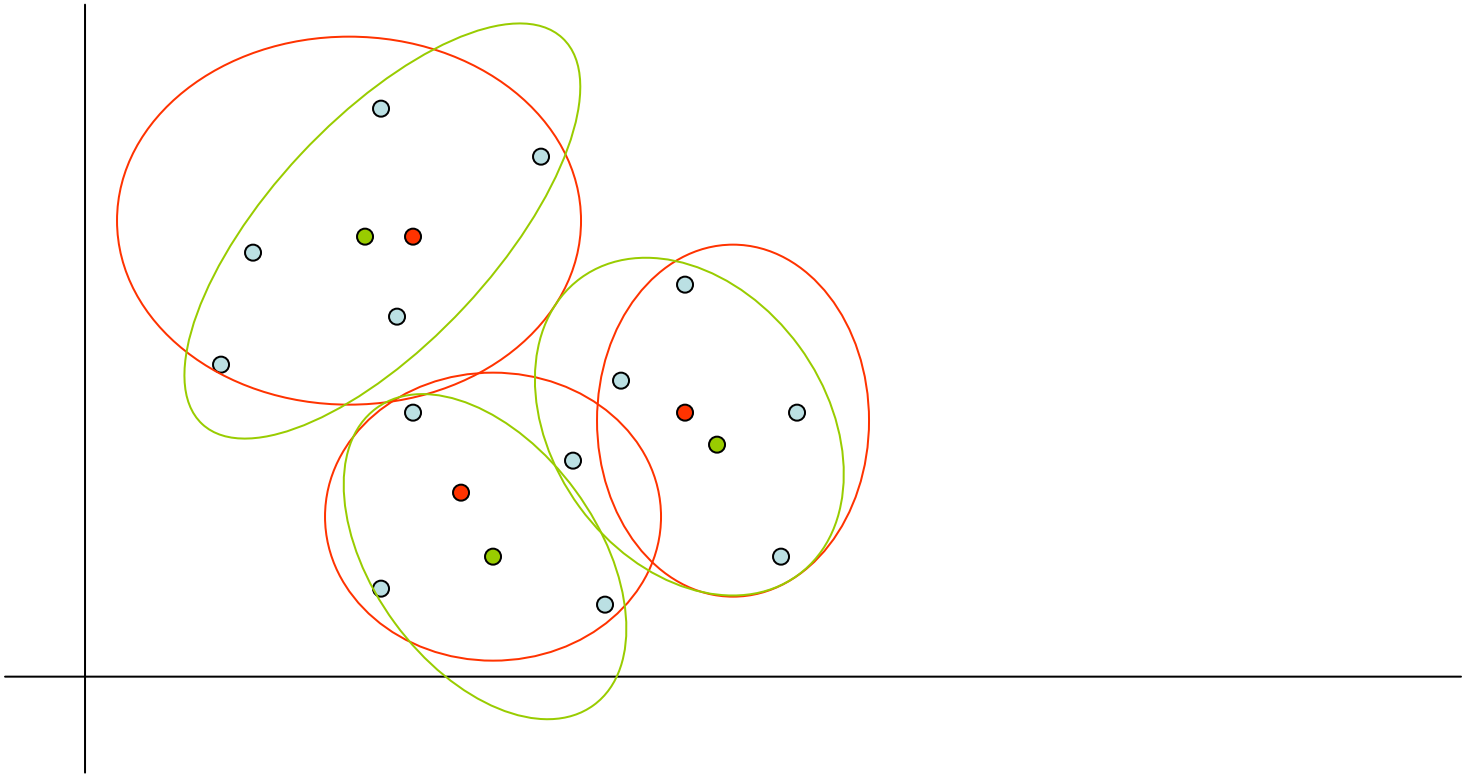
Hierarchical clustering

依照資料的相似程度直接歸類，可以畫出像右方的樹狀圖，分群時可以直接以一定相似程度的threshold將觀測點分類



K-means clustering

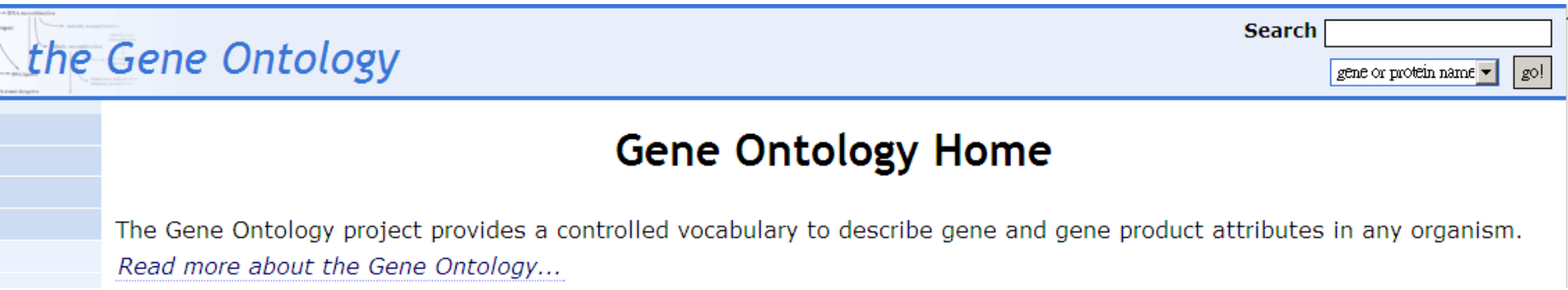
1. 在觀測點(藍點)的多維空間中隨機產生數個點(紅點)當作群組中心
2. 分別將距離群組中心最近的觀測點歸群(紅圈)，
3. 再計算歸群的觀測點之中心距離(綠點)當作新的群組中心
4. 重複2. 3. 直到群組中心收斂



Annotation

通常在分群之後，我們便要依照這些不同表現趨勢的gene做解釋，說明這些表現顯著的基因，在不同的分群中，分別表示了實驗的變因造成了哪些結果，其中最常用以解釋的就是Gene Ontology 及KEGG pathway的資料了

GO…



http://www.geneontology.org

GO database主要是紀錄gene product annotation的功能，是把annotation的記錄方式統一化，方便在查詢的時候，可以簡單將**annotation**做歸類，甚至是跨物種的查詢。收集這些資訊的方式則是從各物種的database上作統一歸類。

主要把annotation分成三大類，biological process，cellular component，molecular function，現在已經分出28XXX個GOID，大部分已知的gene在GO中都被貼上許多GOID來代表他們所具有的屬性，而這些GOID所代表的生物意義也是容易被理解的

GO web demo

KEGG…

Search KEGG | Get Entry

KEGG Home
Introduction
Overview
Release notes
Current statistics

KEGG Identifiers

KGML

KEGG API

KEGG FTP

KEGG: Kyoto Encyclopedia of Genes and Genomes

A grand challenge in the post-genomic era is a complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information. Towards this end we have been developing a bioinformatics resource named KEGG as part of the research projects of the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo.

http://www.genome.jp/kegg/

KEGG則主要是紀錄gene product pathway，將pathways區分為許多大類，記錄每個**pathway**包含的**genes**，**pathway figure**，並給一個**id**，而不同的物種可能會有相似的**pathway(**有同一個**pathway id)**及**ortholog gene**。

KEGG…

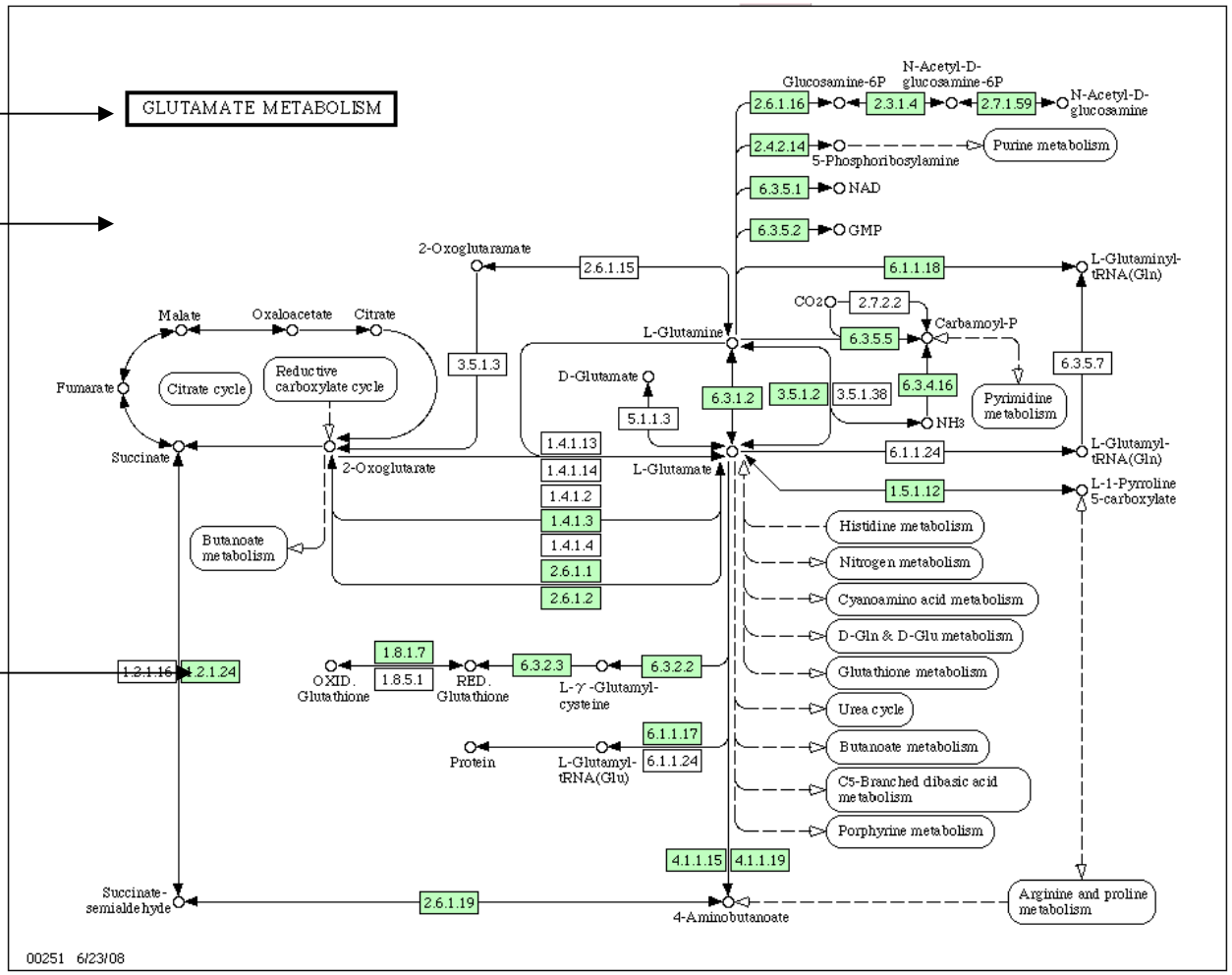
KEGG Release 51.0+/07-01, Jul 09

KEGG PATHWAY	94,266 pathways generated from 332 reference pathways
KEGG BRITE	22,260 hierarchies generated from 64 reference hierarchies
KEGG GENES	12,270 KO groups 4,617,442 genes in 102 eukaryotes + 853 bacteria + 64 archaea
KEGG LIGAND	15,758 compounds, 8,926 drugs, 10,969 glycans, 7,965 reactions, 11,495 reactant pairs

Pathway名稱

樣版圖

選擇物種所包含的gene



Fisher's exact test

	men	women	total
dieting	*a*	*b*	*a* + *b*
not dieting	*c*	*d*	*c* + *d*
totals	*a* + *c*	*b* + *d*	

用來檢驗樣本是否為男生，跟肥胖沒有關係的機率

For annotation

	In GroupA	Out GroupA	Total Gene
Have annoA	*a*	*b*	*a* + *b*
Doesn't have annoA	*c*	*d*	*c* + *d*
total	*a* + *c*	*b* + *d*	

用來檢驗樣本是否為GroupA，跟是不是annoA沒有關係的機率

GO web demo