

# AI、大數據與品質研討會

中華民國品質學會



主題：品質控制在生物資訊的應用

主講人：李明昌 博士

2021年4月26日

# 大綱

- R語言與RStudio軟體簡介
- Bioconductor實務應用
- 品質控制技術
- 結論與未來展望

# 個人簡介

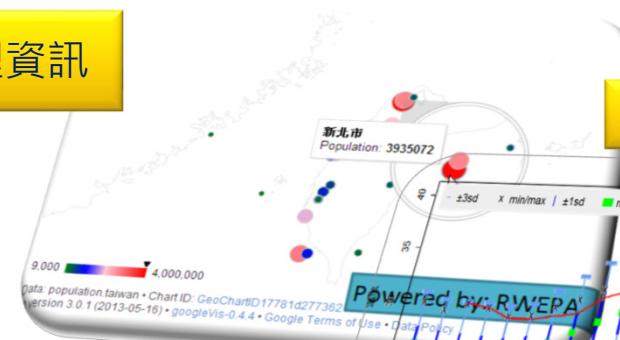
- 姓名：李明昌 (ALAN LEE)
- 現職：中華R軟體學會 常務理事  
臺灣資料科學與商業應用協會 常務理事
- 學歷：中原大學 工業與系統工程所 博士
- 經歷：
  - 育達科技大學 資訊管理系(所) 專任助理教授
  - 佛光大學 兼任教師
  - 國立台北商業大學 兼任教師
  - 東吳大學 兼任教師
  - 崇友實業 行銷企劃專員
  - 國航船務代理股份有限公司 海運市場運籌管理員
- 國內外各大專院校、資策會、工業技術研究院、國家發展委員會、中央氣象局、公平交易委員會、各縣市政府與日本名古屋產業大學等公民營單位演講達280多場，2460小時以上。
- 連絡資訊：
  - RWEPA網站：<http://rwepa.blogspot.com/>
  - E-MAIL: [alan9956@gmail.com](mailto:alan9956@gmail.com)



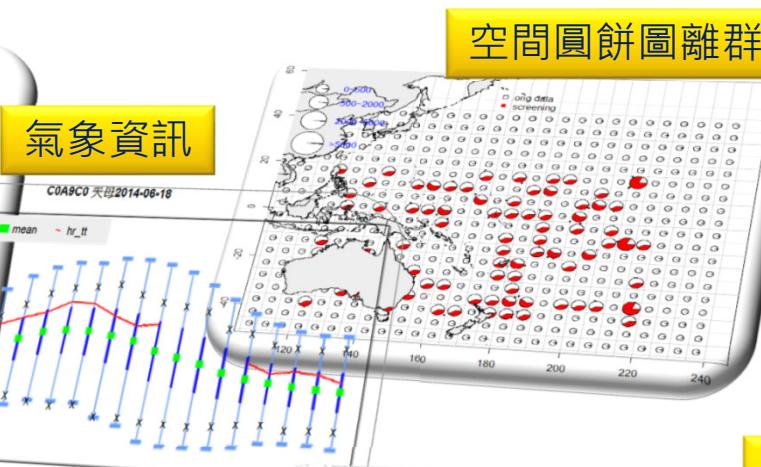
# 資料分析/視覺化成果?

R + shiny → 互動式網頁

地理資訊



氣象資訊

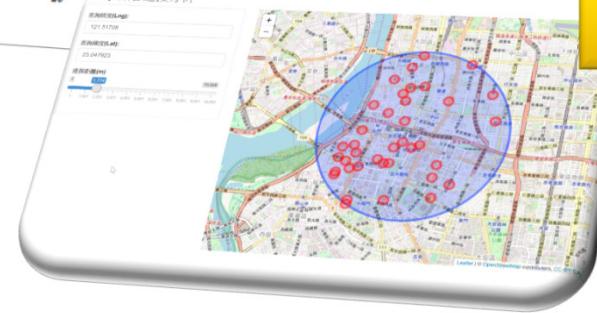


空間圓餅圖離群值分析

保險預測



顧客連結資訊



網頁呈現

# 中央氣象局 1,600萬筆資料



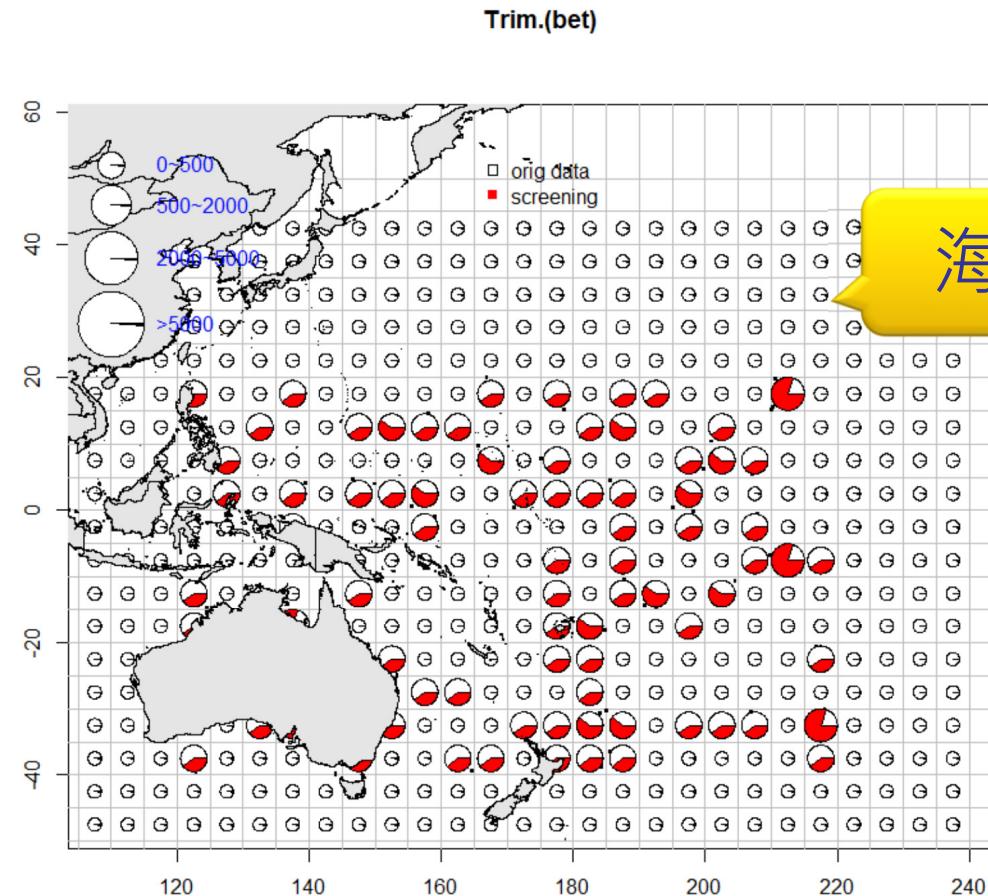
# 保險預測模型

機率模型閾值調整

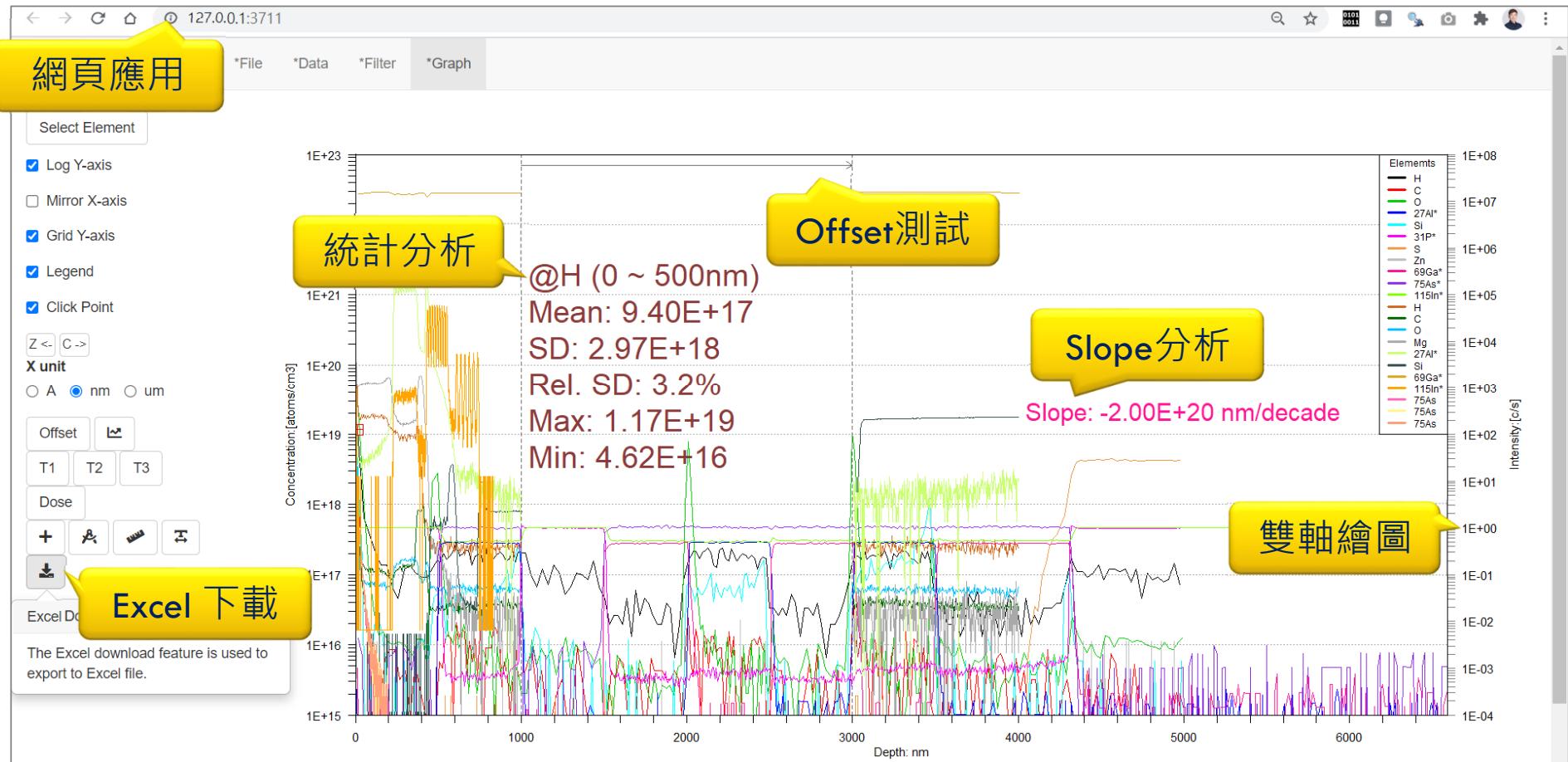
預測結果

性別	女性	車輛種類	私家車	曝露風險		被保險人年齡	無索償折扣	私家車 一車齡 0	私家車 一車齡 1	私家車 一車齡 2	私家車 -車齡 0_1_2 組合	車齡 組合	車齡 0_1_2 組合	預測機率	理賠	
				曝露風險對數	無索償折扣											
M	0	A	1	0.9144422	-0.08944106	50	4	1	0	0	1	0	2	0.1069	有	
M	0	A	1	0.8158795	-0.20348856	20	4	0	0	1	1	2	2	0.1441	有	
3	M	0	A	1	0.8377823	-0.17699695	50	3	0	0	1	1	2	2	0.1866	有
4	M	0	A	1	0.4325804	-0.83798702	50	6	0	1	0	1	1	2	0.0944	無
5	M	0	A	1	0.7173169	-0.33223755	50	4	0	0	1	1	2	2	0.1218	有
6	M	0	A	1	0.8377823	-0.17699695	50	4	0	0	1	1	2	2	0.1495	有
7	M	0	A	1	0.8487337	-0.16400975	50	5	0	0	1	1	2	2	0.1422	有
8	F	1	A	1	0.8268309	-0.19015503	10	3	0	0	1	1	2	2	0.1733	有
9	M	0	A	1	0.7145791	-0.33606164	0	5	1	0	0	1	0	2	0.0694	無
10	M	0	A	1	0.3340178	-1.09656101	0	3	0	0	1	1	2	2	0.0783	無

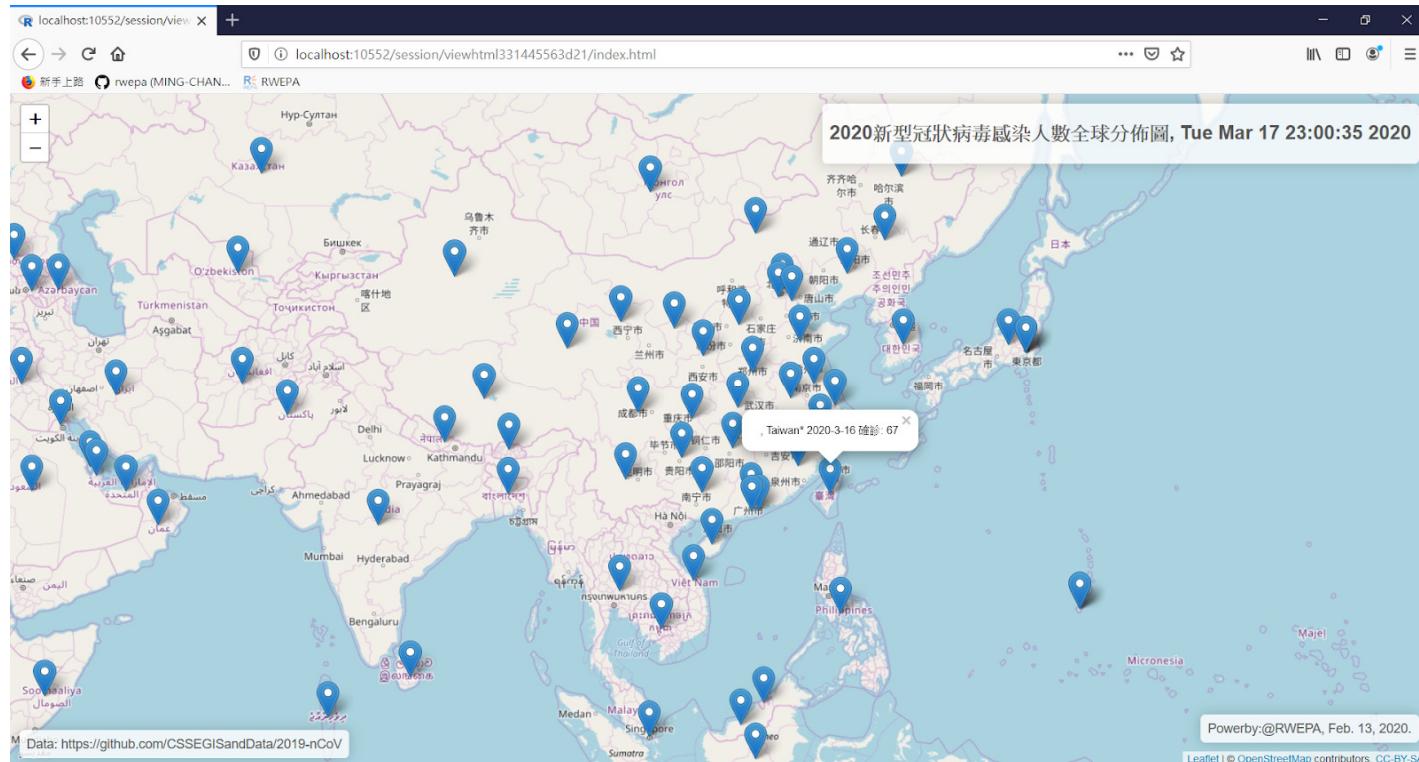
# 空間圓餅圖離群值分析



# 離子資料分析與視覺化應用



# 2020新型冠狀病毒視覺化



<http://rwepa.blogspot.com/2020/02/2019nCoV.html>

# 大數據分析工具



- Microsoft Excel 2019: 104萬餘筆資料限制

A	B	C	D	E	F	G	
1	WEEK_END_DATE	STORE_NUM	UPC	UNITS	VISITS	HHS	SPEND
1048572	14-Jan-09	367	1111009477	13	13	13	18.07
1048573	14-Jan-09	367	1111009497	20	18	18	27.8
1048574	14-Jan-09	367	1111009507	14	14	14	19.32
1048575	14-Jan-09	367	1111035398	4	3	3	14
1048576	14-Jan-09	367	1111038078	3	3	3	7.5

1,048,576筆資料限制

- 免費: 核心程式 + 套件(模組) + IDE



# 大數據分析免費工具



軟體	Python	R	Julia
Released	1991	2000	2012
用途	程式語言 系統結合	統計,繪圖,視覺化 程式語言	科學計算 程式語言
版本	自由軟體 物件導向	自由軟體 物件導向	自由軟體 物件導向
附加功能	免費模組	免費套件	免費模組
使用者	工科+ 商管	商管+ 工科	商管+ 工科

# 1.R語言與RStudio軟體簡介

---

# R 安裝與簡介

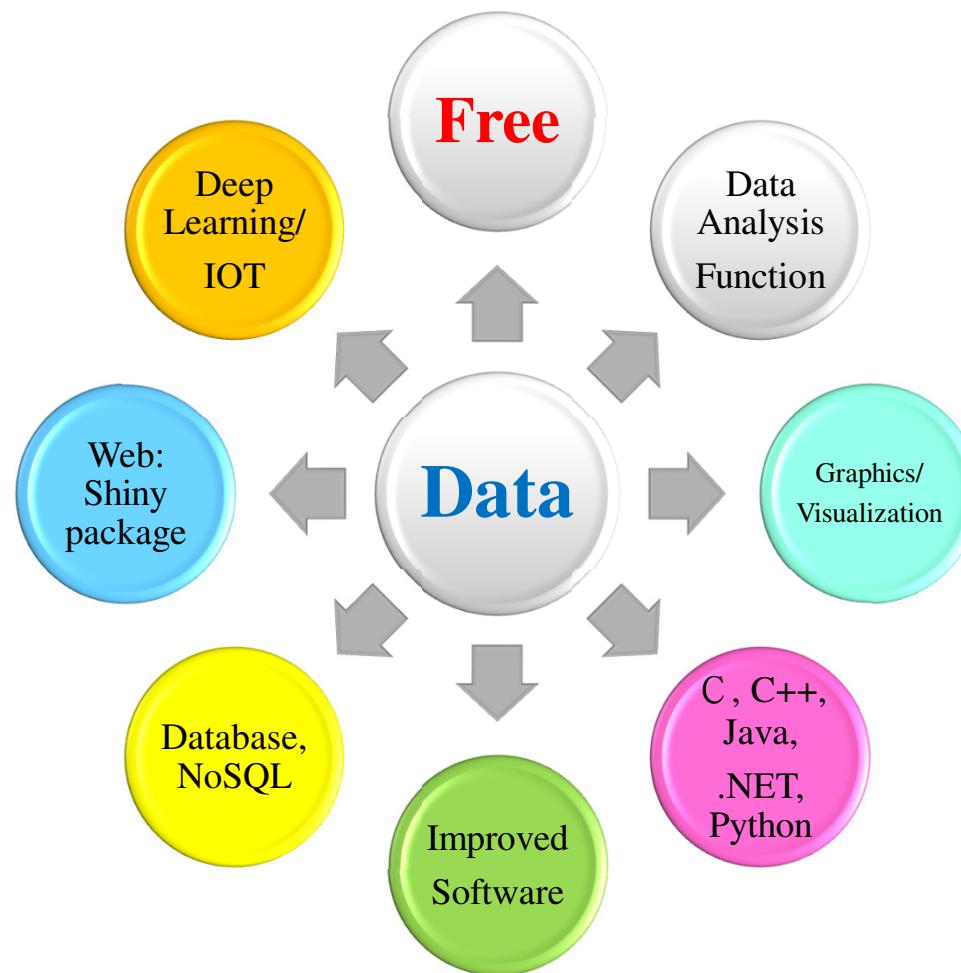
---

## 認識R

- 1976 - 貝爾實驗室 John Chambers, Rick Becker, and Allan Wilks研發S 語言。
- 1993 - Ross Ihaka and Robert Gentleman, University of Auckland, New Zealand 研發R 語言。
  - R 是一種基於 S 語言所發展出具備統計分析、繪圖與資料視覺化的程式語言。
- 1997年—R的核心開發團隊 (R development core team) 成立，專責R原始碼的修改與編寫。
  - 2000年2月 – R 1.0.0
  - 2013年3月 – R 2.15.3
  - 2021年3月 – R 4.0.5



# R-八大功能



## R-下載

- 官網: <http://www.r-project.org/>
- 選取左側 Download \ CRAN
- 選取 Taiwan CRAN

Taiwan  
<https://cran.csie.ntu.edu.tw/>

- 選取 Download R for Windows

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)



## R-下載 (續)

- 選取 base → 下載 [R-4.0.5-win.exe]



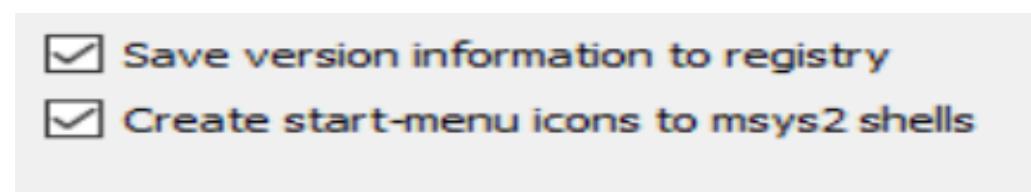
- R安裝路徑: 保留原路徑,不要修改
- [https://github.com/rwepa/DataDemo/blob/master/windows\\_intall\\_R.pdf](https://github.com/rwepa/DataDemo/blob/master/windows_intall_R.pdf)

# 下載並安裝 Rtools

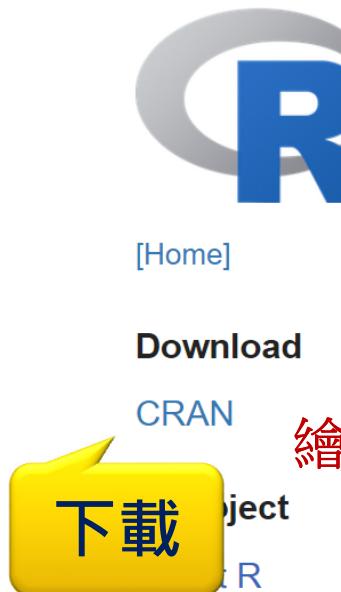
- Rtools for Windows: 一定要保留原路徑 C:\rtools40



On Windows 64-bit: [rtools40-x86\\_64.exe](#) (recommended: includes both i386 and x64 compilers)



# R官方網頁



## The R Project for Statistical Computing

### Getting Started

統計計算

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

# R Manuals

## The R Manuals

edited by the R Development Core Team.

The following manuals for R were created on Debian Linux and may differ from the manuals for Mac or Windows on platform-specific pages, but most parts version of the manuals for each platform are part of the respective R installations. The manuals change with R, hence we provide versions for the most recent version for the patched release version (R-patched) and finally a version for the forthcoming R version that is still in development (R-devel).

Here they can be downloaded as PDF files, EPUB files, or directly browsed as HTML:

Manual	R-release	R-patched
<b>An Introduction to R</b> is based on the former "Notes on R", gives an introduction to the language and how to use R for doing statistical analysis and graphics.	<a href="#">HTML</a>   <a href="#">PDF</a>   <a href="#">EPUB</a>	<a href="#">HTML</a>   <a href="#">PDF</a>   <a href="#">EPUB</a>
<b>R Data Import/Export</b> describes the import and export facilities available either in R itself or via packages which are available from CRAN.	<a href="#">HTML</a>   <a href="#">PDF</a>   <a href="#">EPUB</a>	<a href="#">HTML</a>   <a href="#">PDF</a>   <a href="#">EPUB</a>
<b>R Installation and Administration</b>	<a href="#">HTML</a>   <a href="#">PDF</a>   <a href="#">EPUB</a>	<a href="#">HTML</a>   <a href="#">PDF</a>   <a href="#">EPUB</a>
<b>Writing R Extensions</b> covers how to create your own packages, write R help files, and the foreign language (C, C++, Fortran, ...) interfaces.	<a href="#">HTML</a>   <a href="#">PDF</a>   <a href="#">EPUB</a>	<a href="#">HTML</a>   <a href="#">PDF</a>   <a href="#">EPUB</a>
A draft of <b>The R language definition</b> documents the language <i>per se</i> . That is, the objects that it works on, and the details of the expression evaluation process, which are useful to know when programming R functions.		
<b>R Internals</b> : a guide to the internal structures of R and coding standards the core team working on R itself.		
<b>The R Reference Index</b> : contains all help files of the R standard and recommended packages in printable form. (9MB, approx. 3500 pages)		

**contributed documentation**  
(貢獻文件, 免費啦)

Translations of manuals into other languages than English are available from the [contributed documentation](#) section (only a few translations are available).

# R Manuals (續)

## Contributed Documentation

[English](#) --- [Other Languages](#)

Manuals, tutorials, etc. provided by users of R. The R core team does not take any responsibility for contents, but we appreciate the effort very much and encourage everybody to contribute to this list! To submit, follow the submission instructions on the [CRAN main page](#). All material below is available directly from CRAN, you may also want to look at the list of [other R documentation](#) available on the Internet.

**Note:** Please use the [directory listing](#) to sort by name, size or date (e.g., to see which documents have been updated lately).

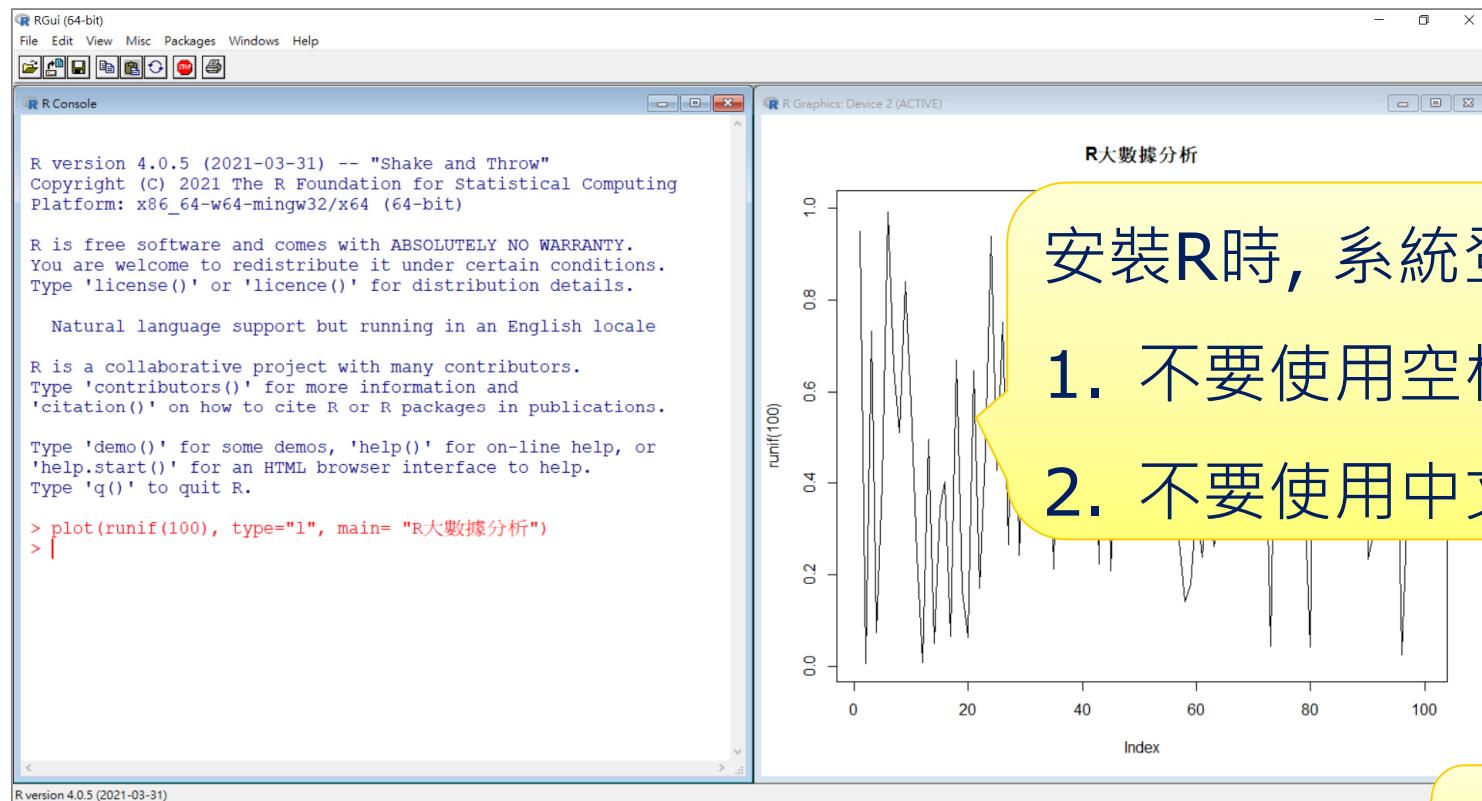
### English Documents

Documents with more than 100 pages:

好書!

- “**Visual Statistics. Use R!**” by Alexey Shipunov ([PDF](#), 2016-06-06, 301 pages) are accessible from [Alexey Shipunov's English R page](#).
- “**Using R for Data Analysis and Graphics - Introduction, Examples and Commentary**” by John Maindonald ([PDF](#), data sets and scripts are available at [JM's homepage](#)).
- “**Practical Regression and Anova using R**” by Julian Faraway ([PDF](#), data sets and scripts are available at the [book homepage](#)).

# R 執行畫面 - Windows



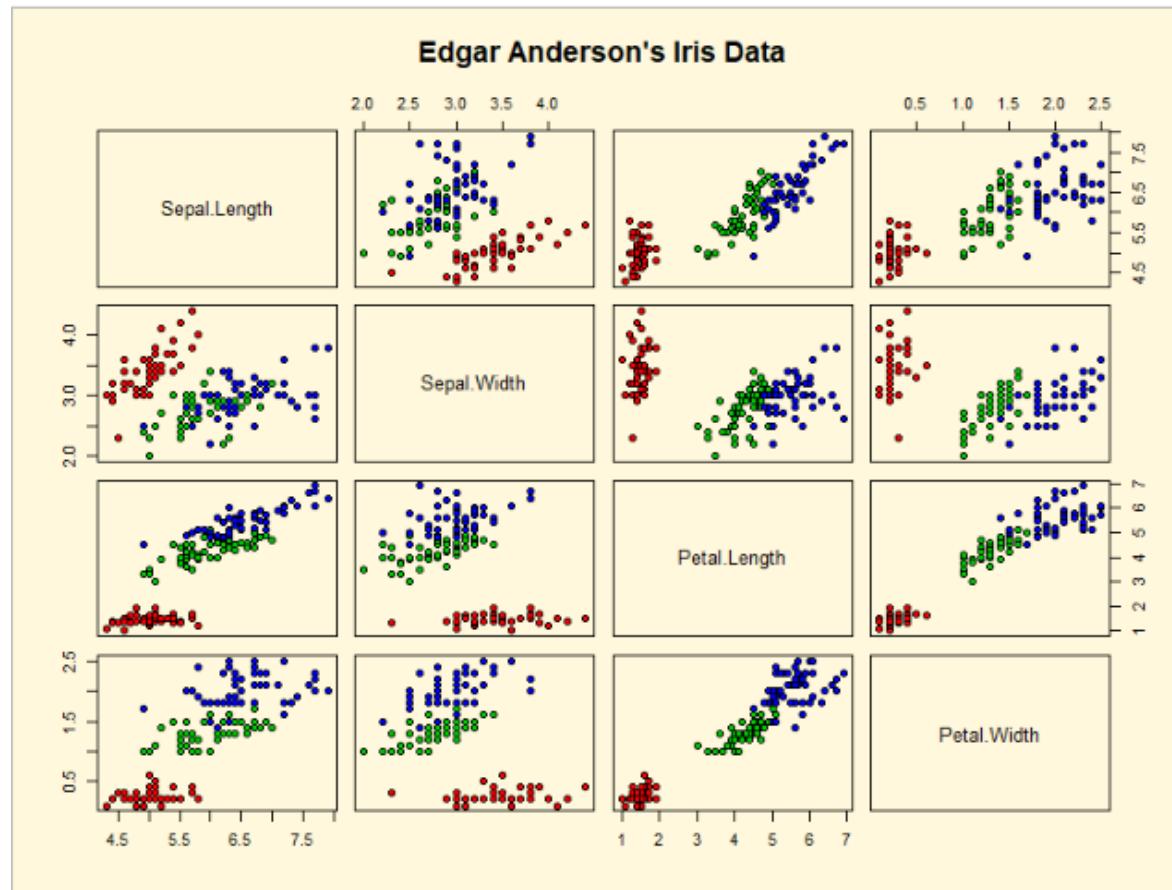
安裝R時，系統登入名稱：

1. 不要使用空格
2. 不要使用中文字

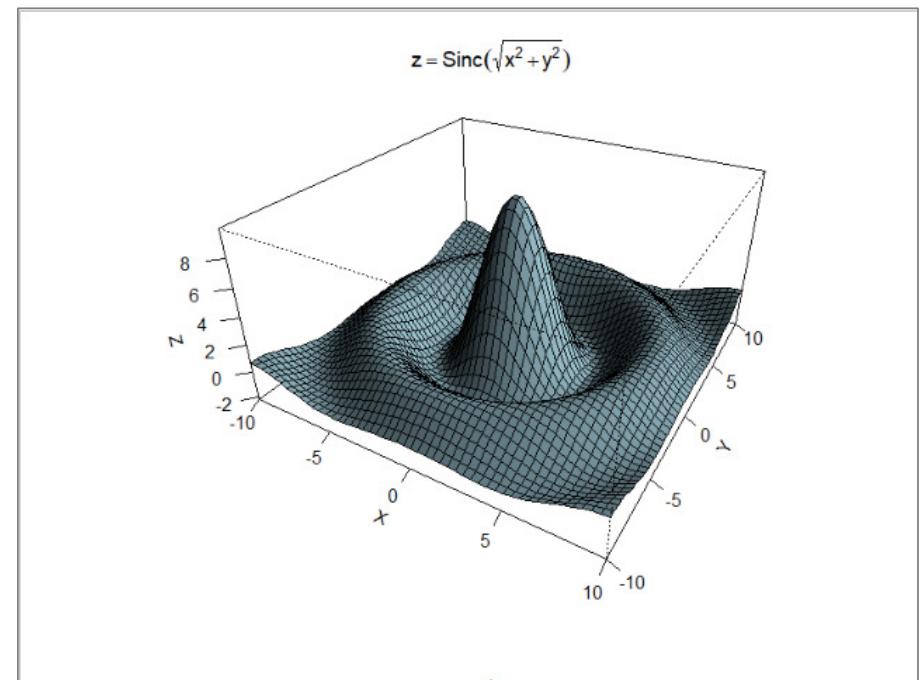
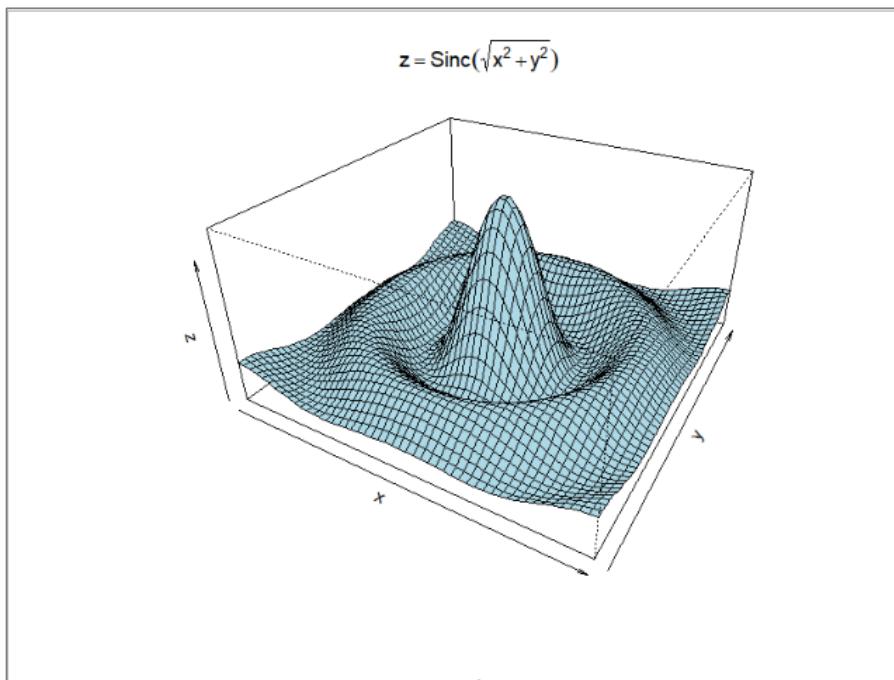
```
plot(runif(100), type="l", main= "R大數據分析")
```

大小寫  
須一致

# demo(graphics)



# demo(persp)



# R for Mac

- <https://youtu.be/72MYRBNo5Bk>



**Mac OS X 安裝 R 軟體**

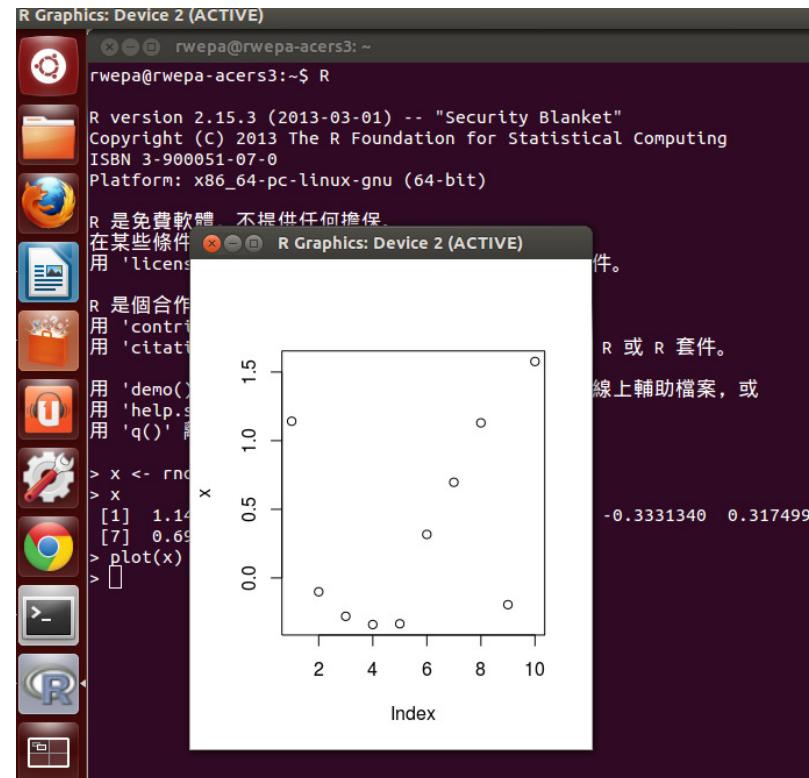
育達科技大學  
資訊管理系  
李明昌  
*alan9956@gmail.com*

**R**  
WEPA  
Since 2013

更多訊息 <http://rwepa.blogspot.tw/>

# R for Ubuntu

- <http://rwepa.blogspot.com/2013/05/ubuntu-r.html>



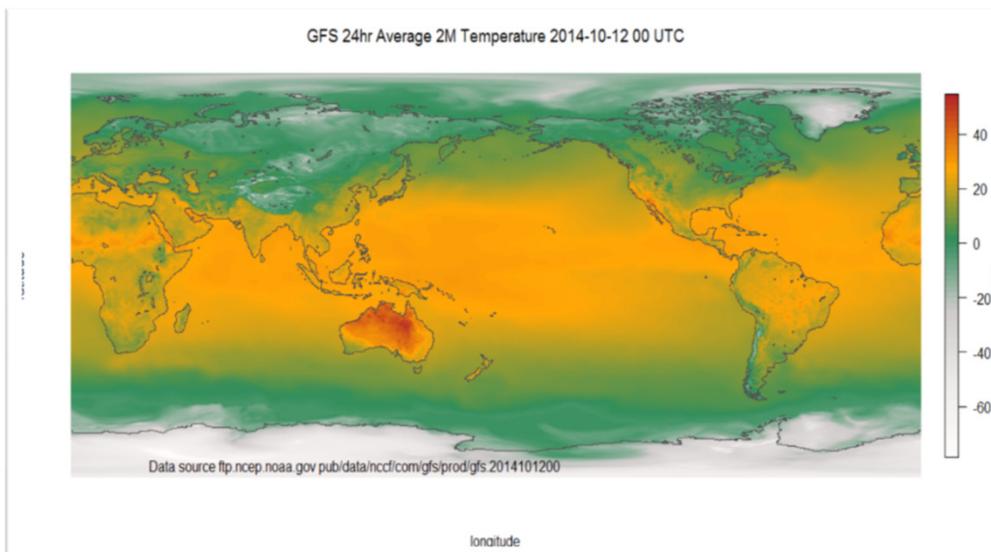


# RStudio 安裝與簡介

---

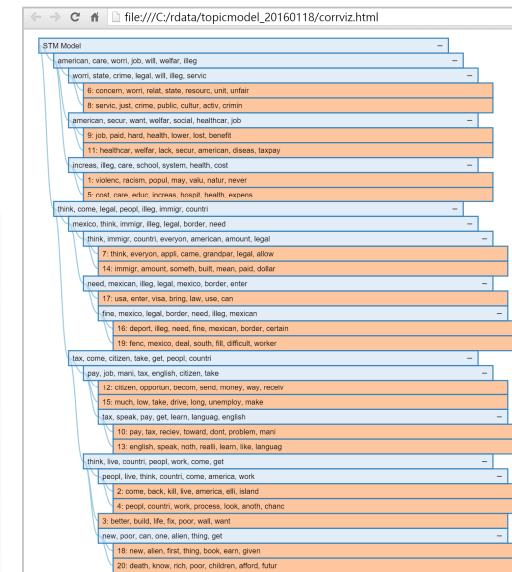
# 整合式開發環境 - RStudio

- <http://www.rstudio.com/>



視覺化應用

(全球2M氣溫圖)



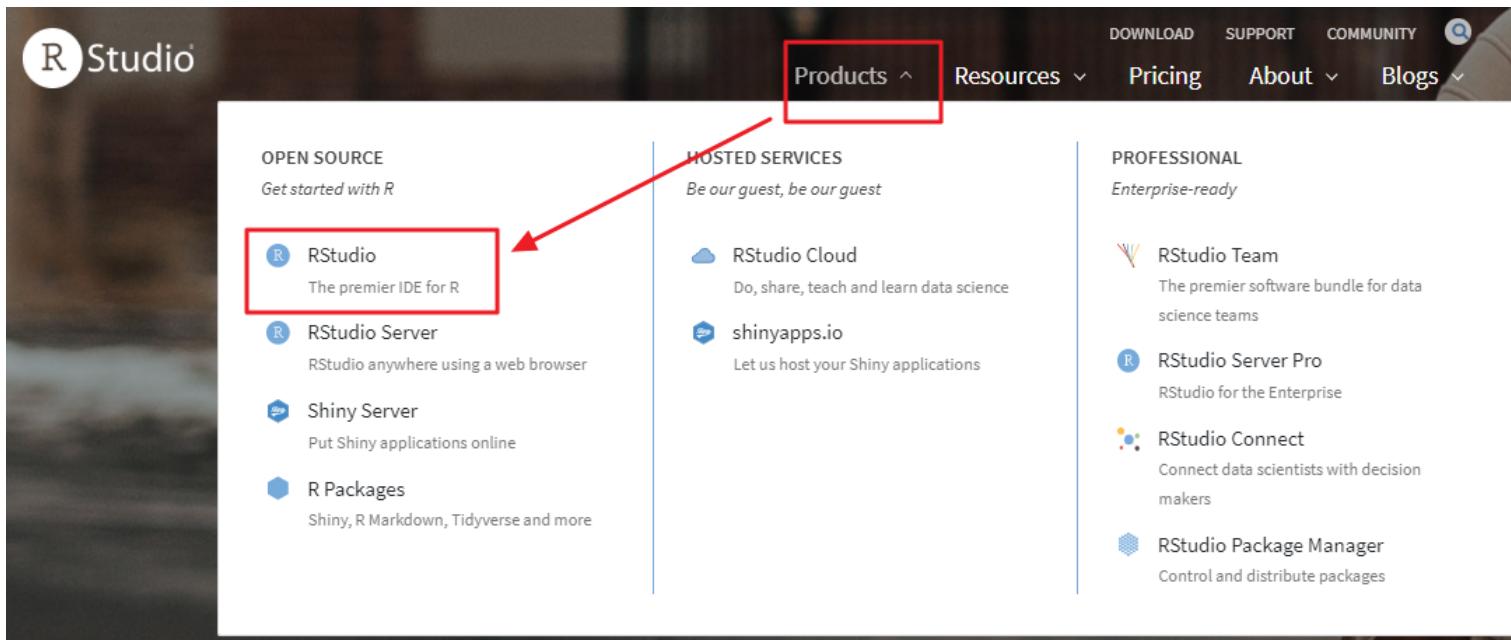
主題模型

# RStudio - 特性

- 支援智慧輸入 (按Tab)
- 高亮度顯示程式碼
- 整合R程式, 控制台, 變數清單, 繪圖視窗
- 整合資料庫匯入 SQL, Spark
- 整合R套件: shiny, rmarkdown
- 安裝注意:
  - 先安裝R, 再安裝 RStudio
  - 安裝 RStudio時, 請先關閉R

# RStudio 下載

- <http://www.rstudio.com/>



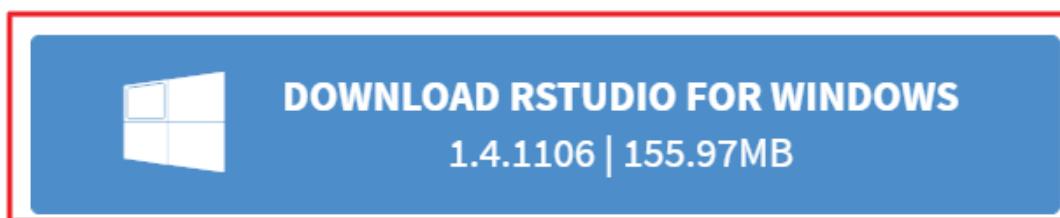


# RStudio 下載 (續)

# RStudio 下載 (續)

RStudio Desktop 1.4.1106 - [Release Notes](#)

- 1.** Install R. RStudio requires [R 3.0.1+](#).
  
- 2.** Download RStudio Desktop. Recommended for your system:



Requires Windows 10/8 (64-bit)

# RStudio 安裝

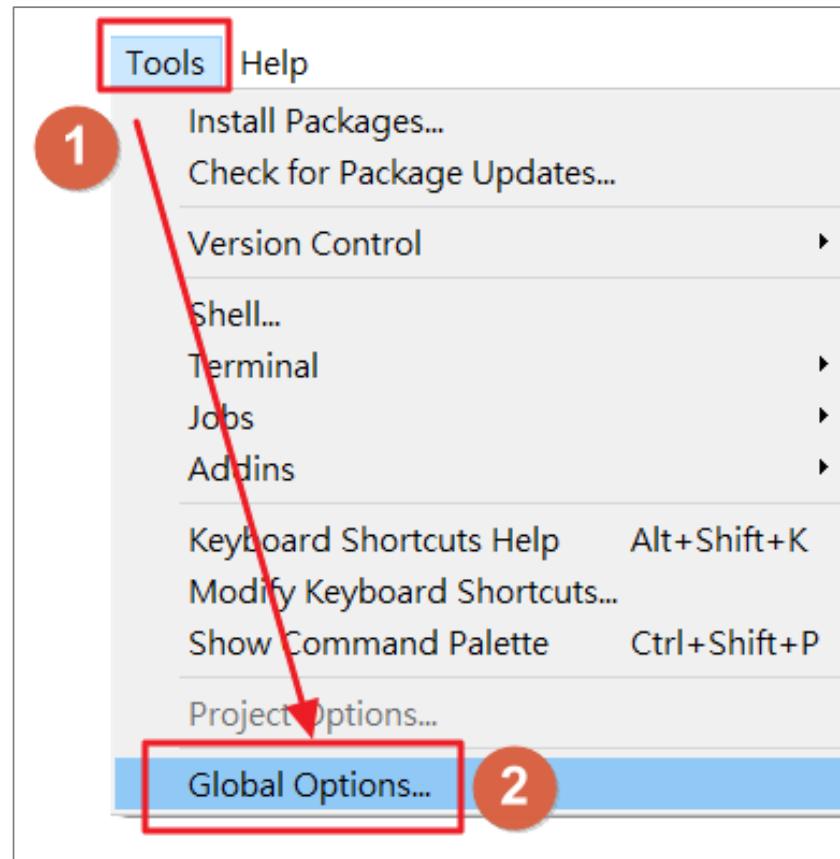


# RStudio 版本訊息

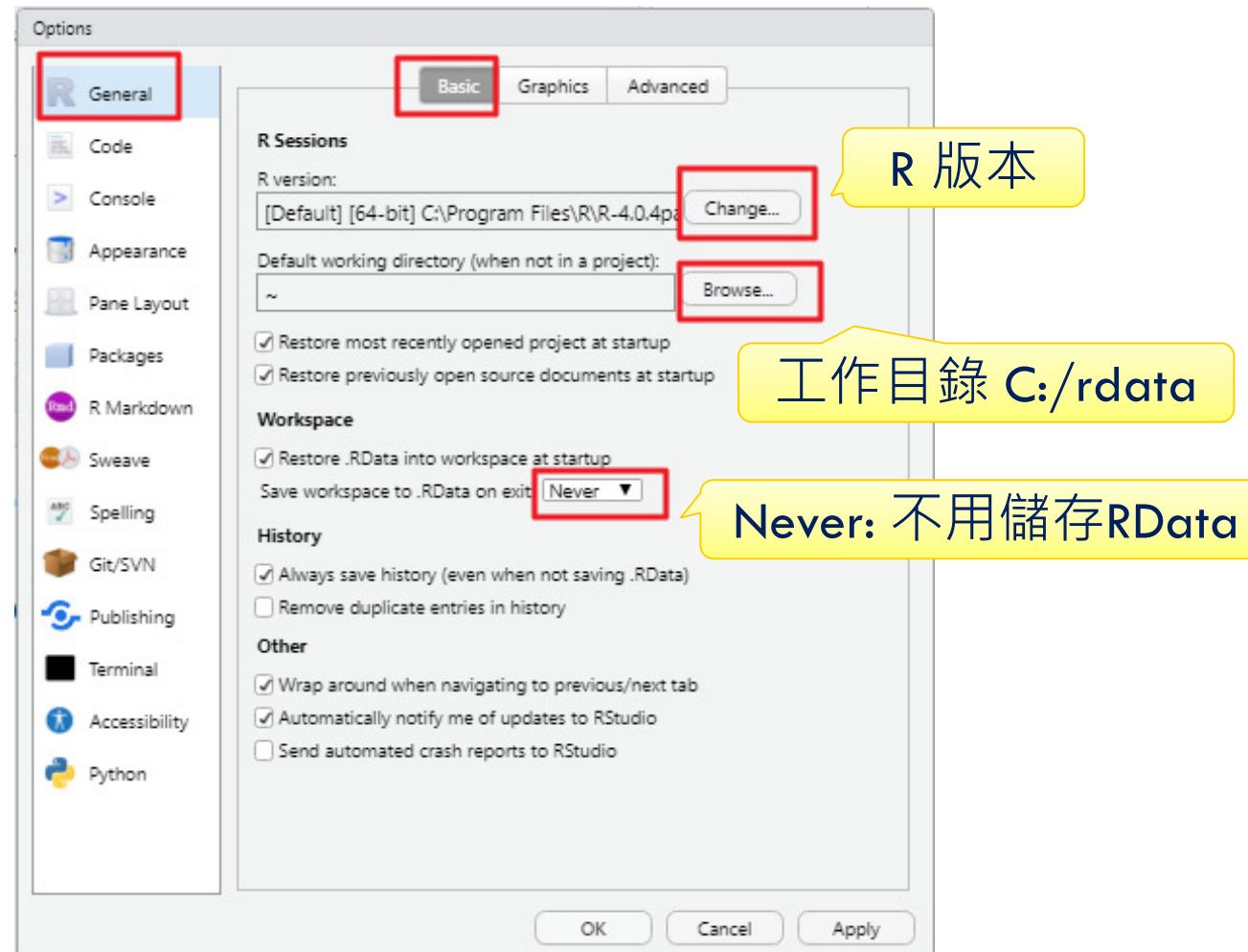


# RStudio - 選項設定

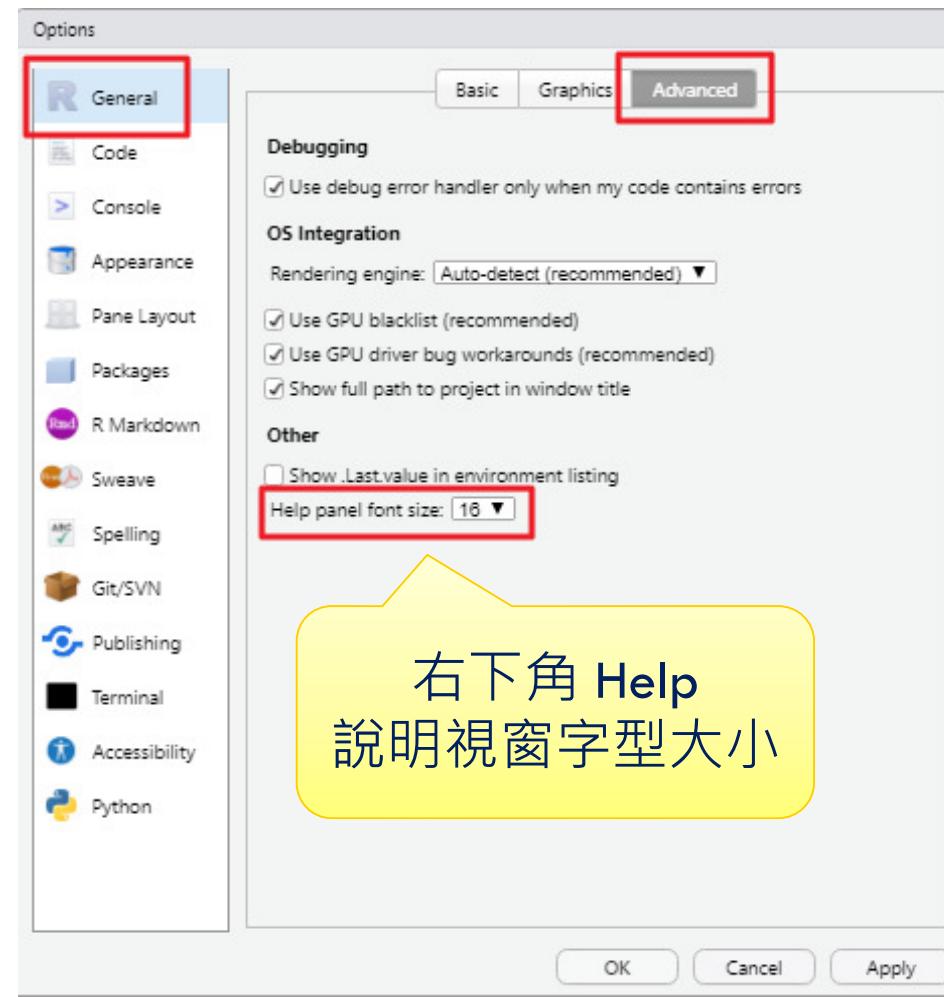
## ■ Tools \ Global Options



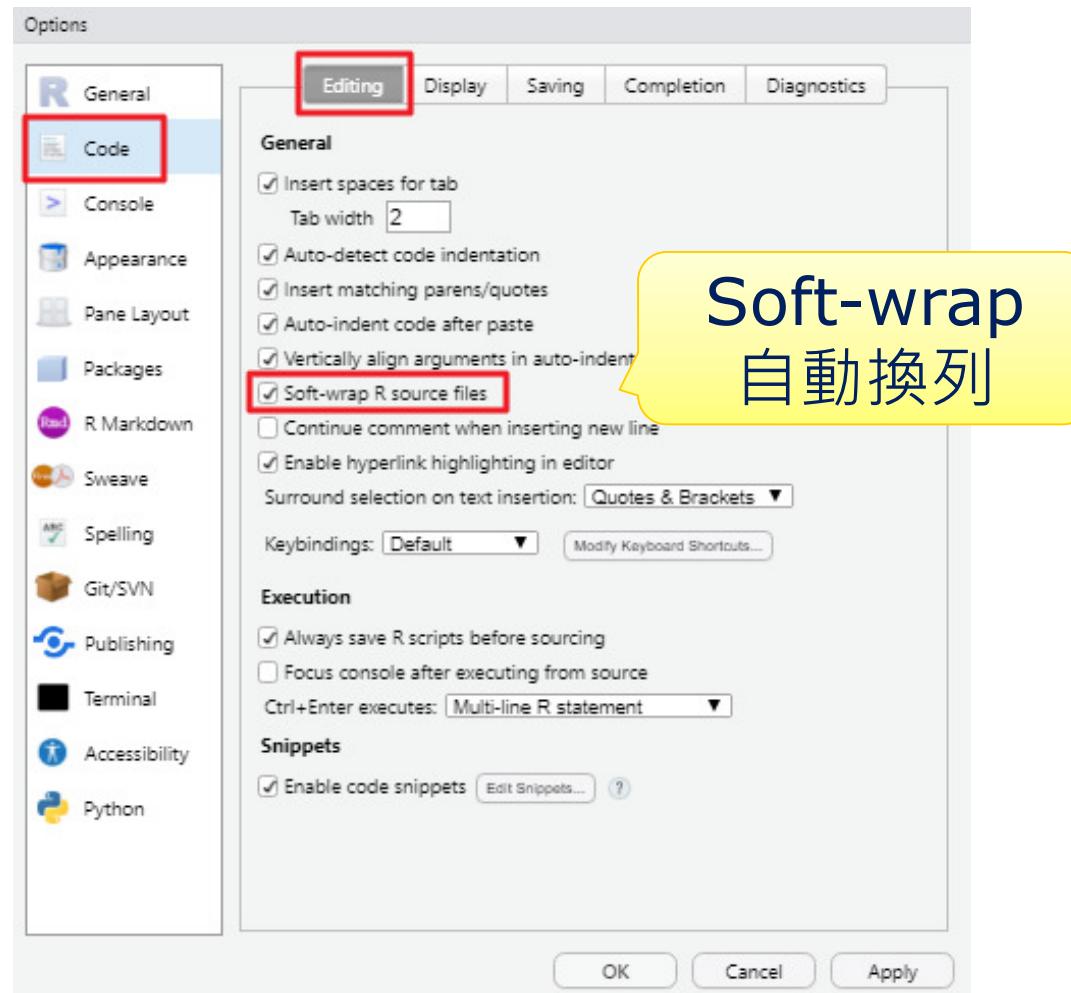
# General \ Basic



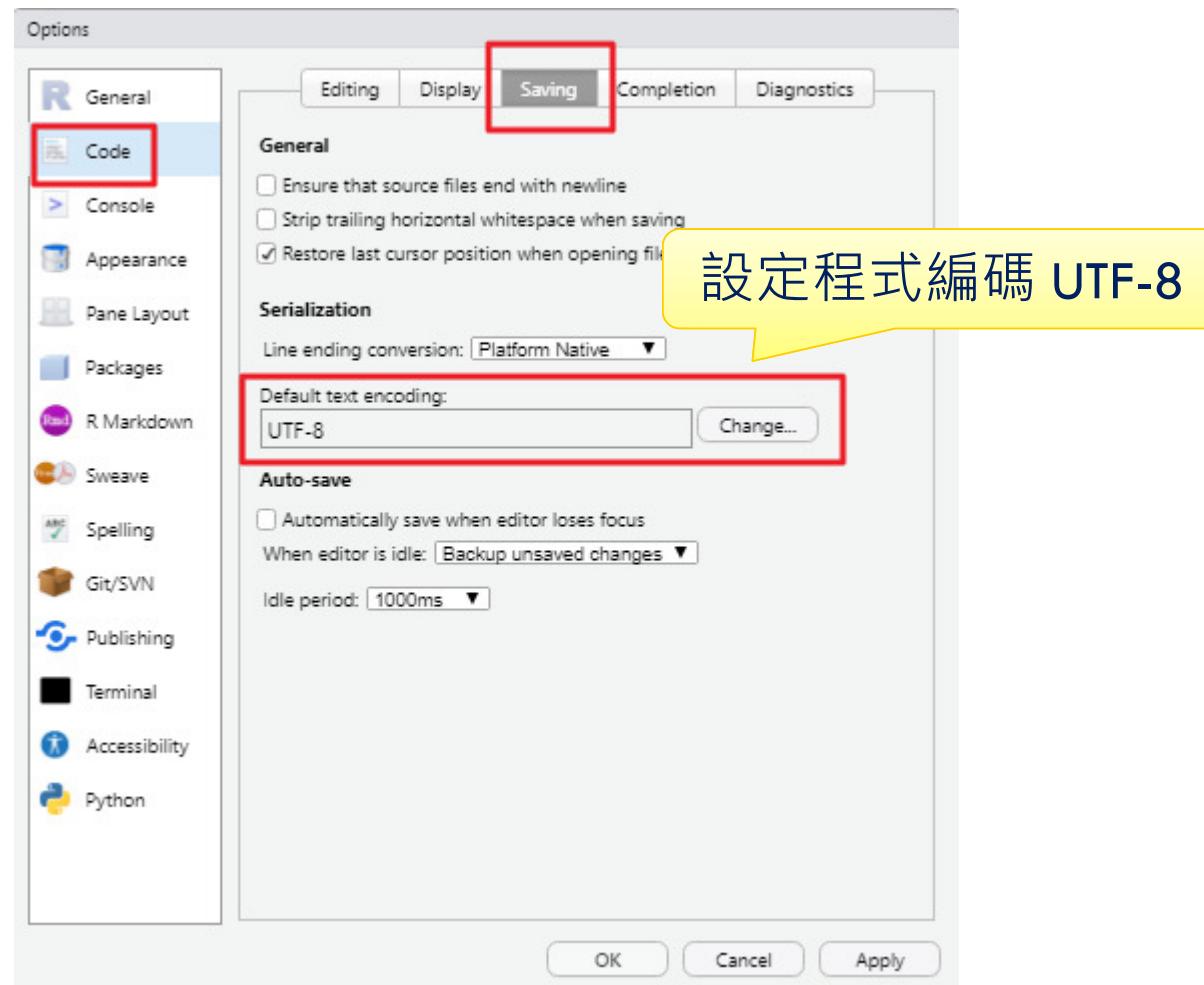
# General \ Advanced



# Code \ Editing



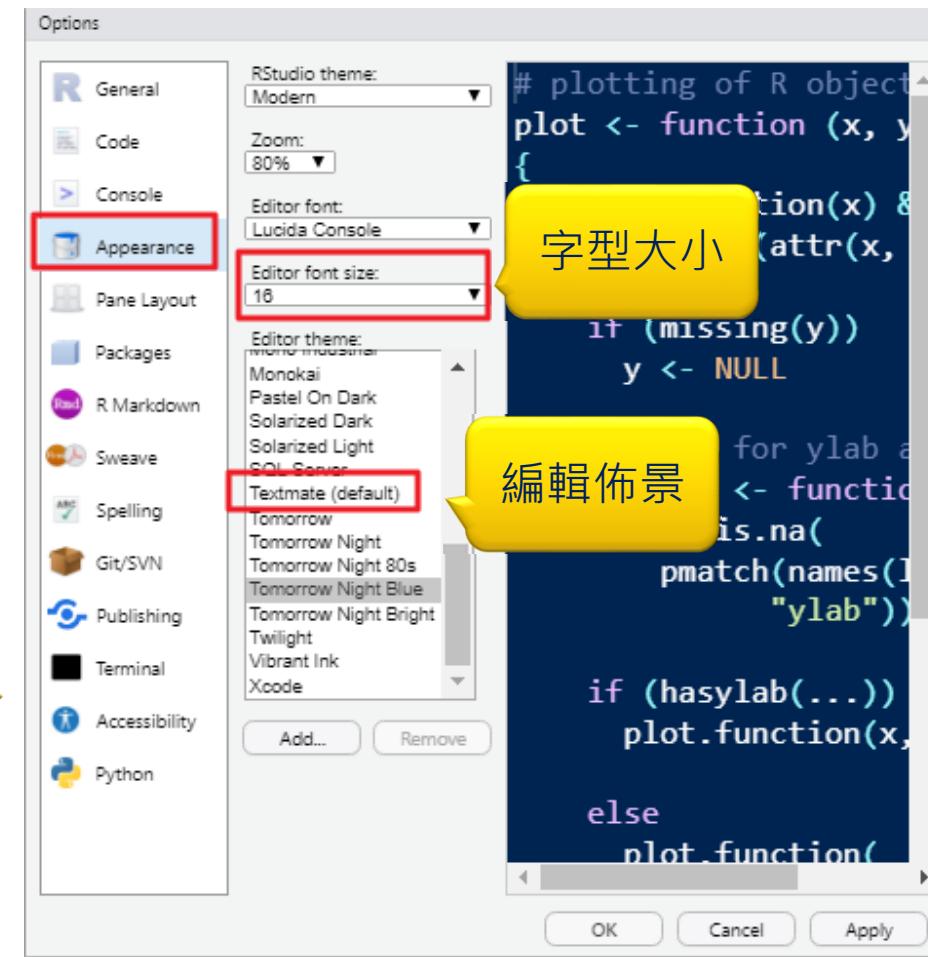
# Code \ Saving



# RStudio-選項設定(續)

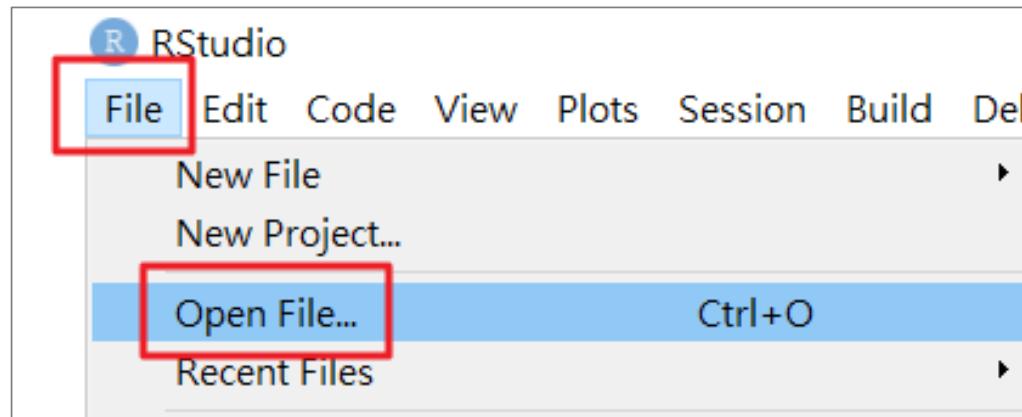
- Appearance \ Editor theme
- 預設值:  
TextMate

設定完成,須重  
新啟動RStudio



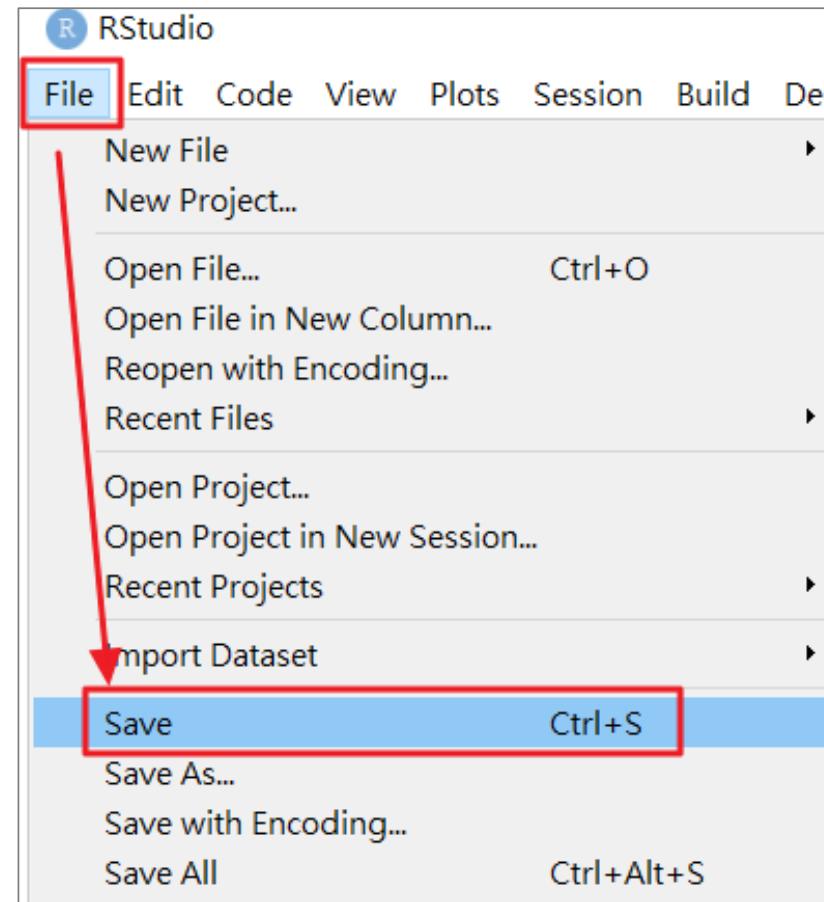
# 開啟檔案

- File \ Open File

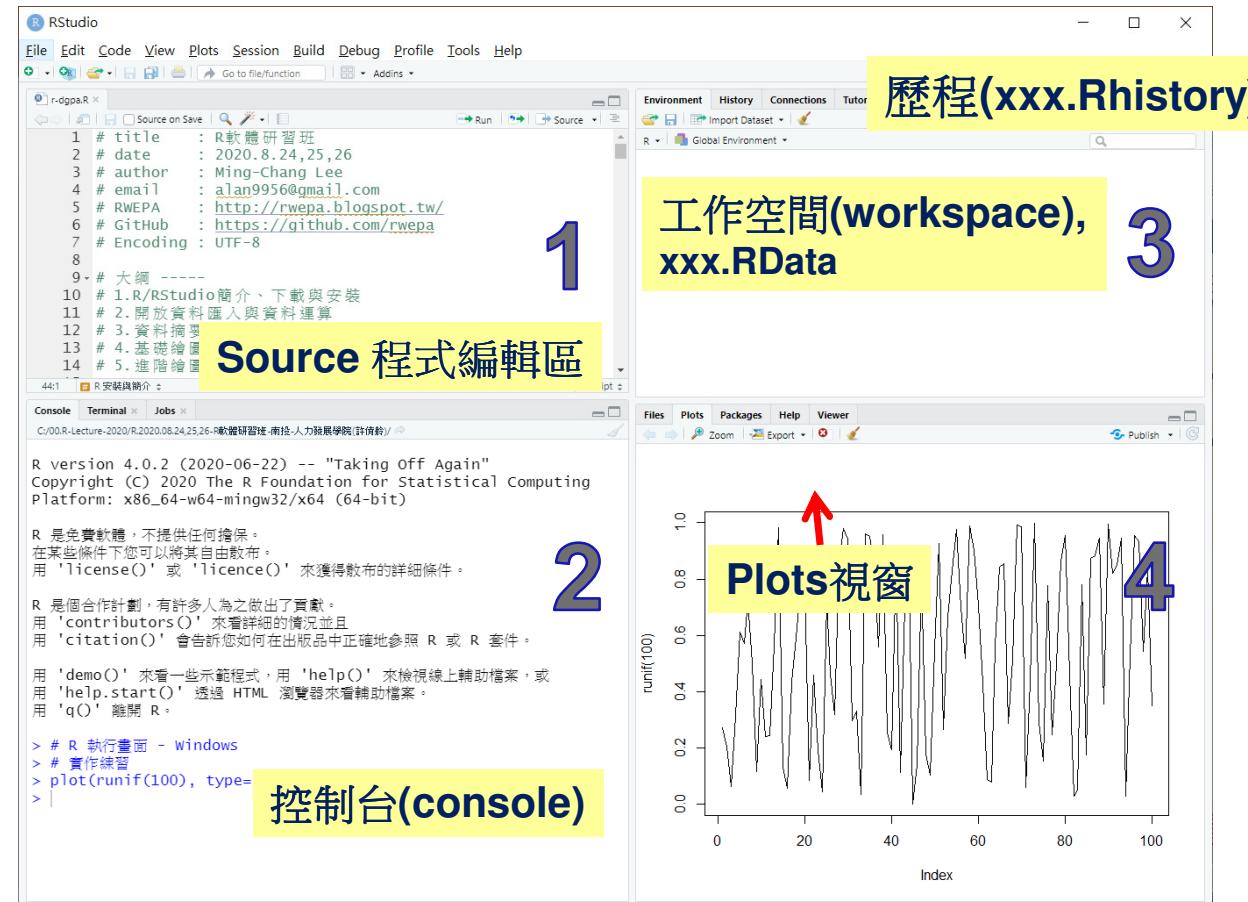


# 儲存檔案

- File \ Save  
(CTRL + S)
- 檔名: xxx.R



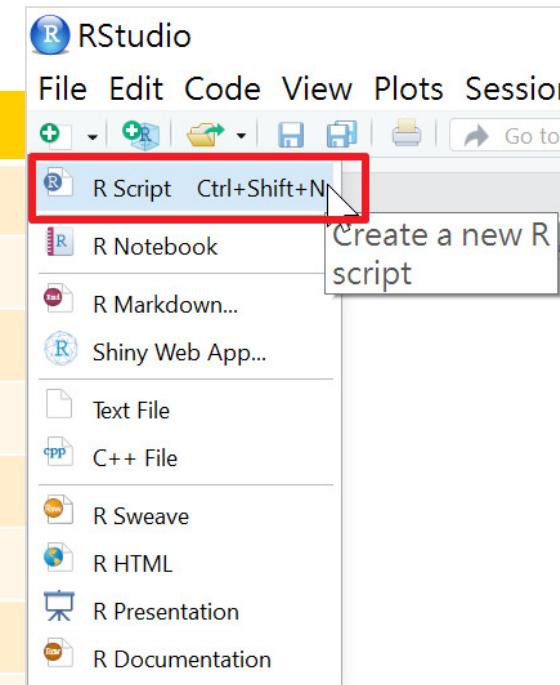
# R/RStudio環境的基礎觀念



**CTRL + SHIFT + F10: 重新啟動R**

# RStudio 快速鍵

快速鍵功能	功能
Ctrl + Shift + N	建立新的R程式
Ctrl + S	儲存檔案
Ctrl + Shift + R	建立章節 ( ----- )
Alt + -	指派符號
Ctrl + Shift + C	註解
Ctrl + Enter	執行程式
Ctrl + Shift + F10	重新啟動R
Alt + Shift + K	快速鍵總表 (Esc 退出)

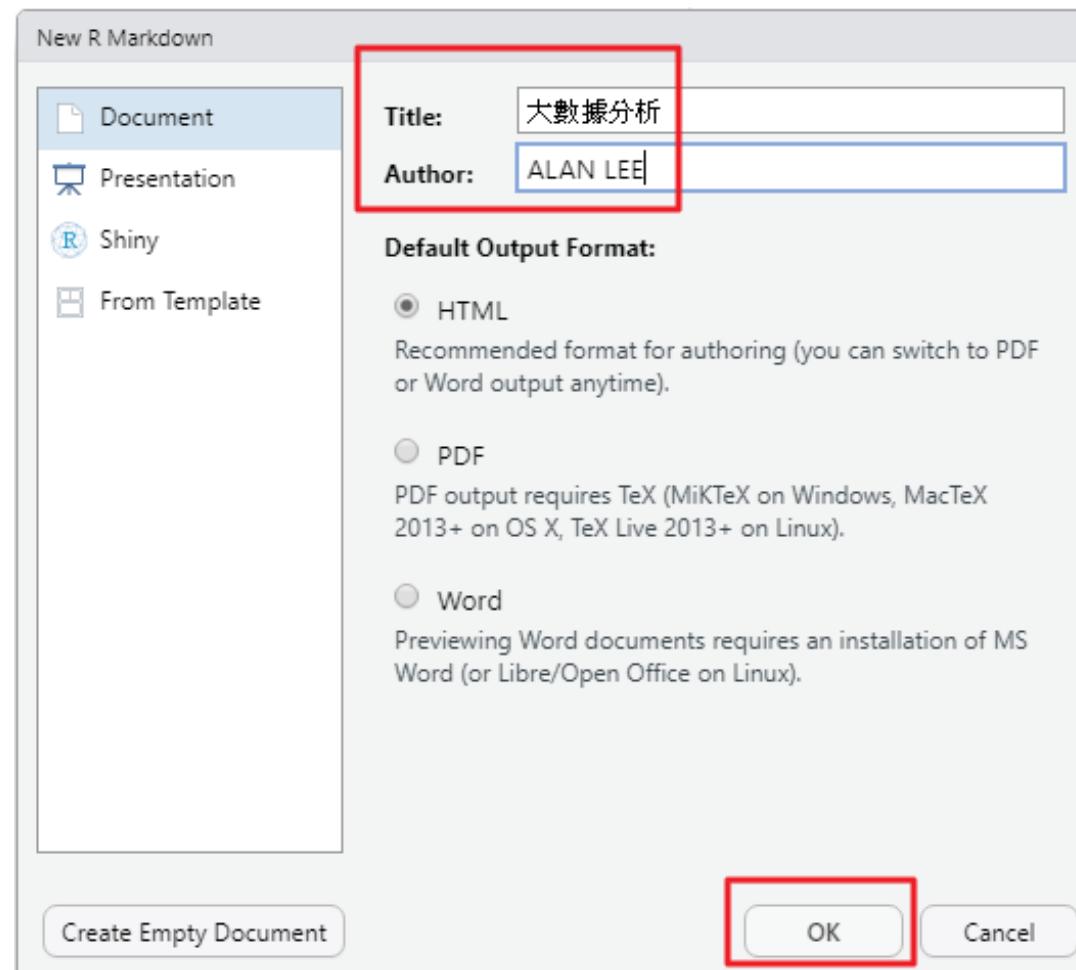


# R Markdown

---

R標記語言

# RStudio - Markdown



# RStudio - Markdown (續)

The screenshot shows the RStudio interface with a Markdown document open in the left pane. The code includes setup code for knitr and a brief introduction to R Markdown. The right pane shows an empty environment. A yellow callout box points to the 'Help' menu item in the top bar, which is highlighted with a red box and circled with a red number 1. The 'Help' menu itself is open, showing various links. Three specific items are highlighted with red boxes and circled with red numbers: 'R Resources' (circled 2), 'RStudio Cheat Sheets' (circled 3), and 'RStudio Tip of the Day'.

1. Help

2.

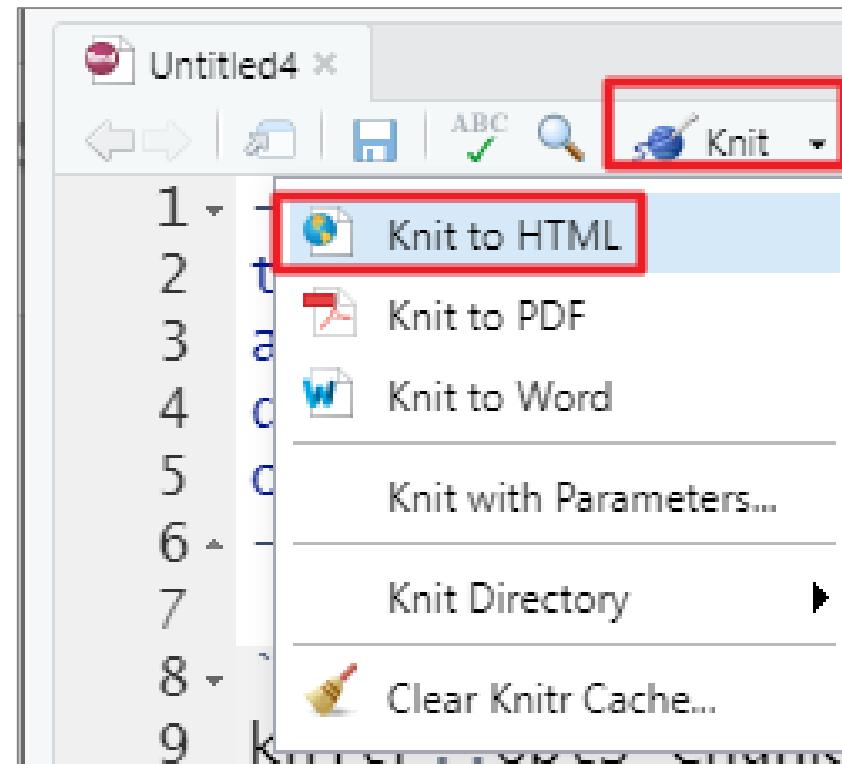
3. RStudio Cheat Sheets

# R Markdown Cheatsheet 線上說明

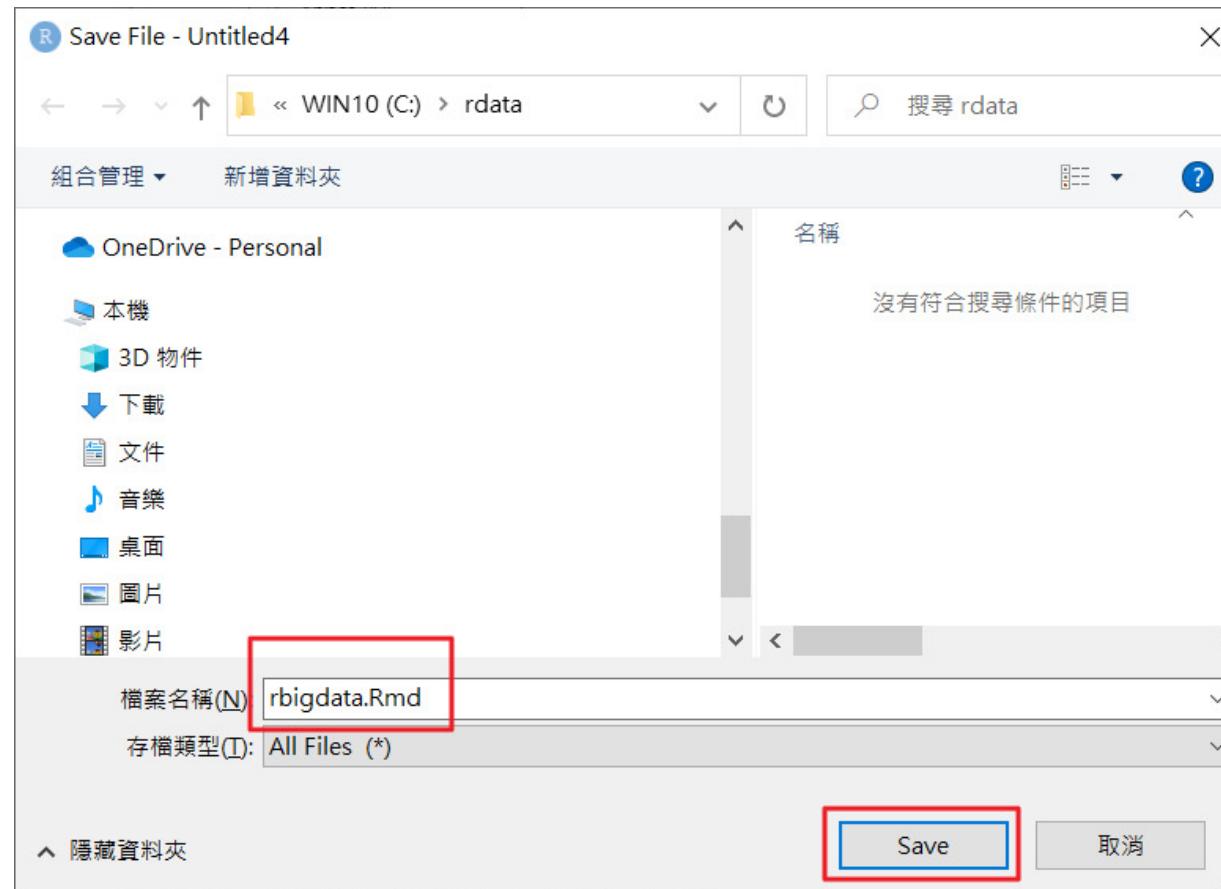
- <https://www.rstudio.com/resources/cheatsheets/>

# RStudio - Markdown (續)

- Knit HTML
- Knit PDF
- Knit Word



# RStudio - Markdown (續)



# Knit to HTML

大數據分析

ALAN LEE  
2020/7/6

R Markdown

This is an R Markdown document. You can embed an R code chunk like this: `summary(cars)`. When you click the Knit button, the document is processed into a single HTML file containing both the formatted text content as well as the output of any embedded R code chunks.

summary(cars)

	speed	dist
## Min. :	4.0	2.00
## 1st Qu.:	12.0	18.00
## Median :	15.0	36.00
## Mean :	15.4	42.98
## 3rd Qu.:	19.0	56.00
## Max. :	25.0	120.00

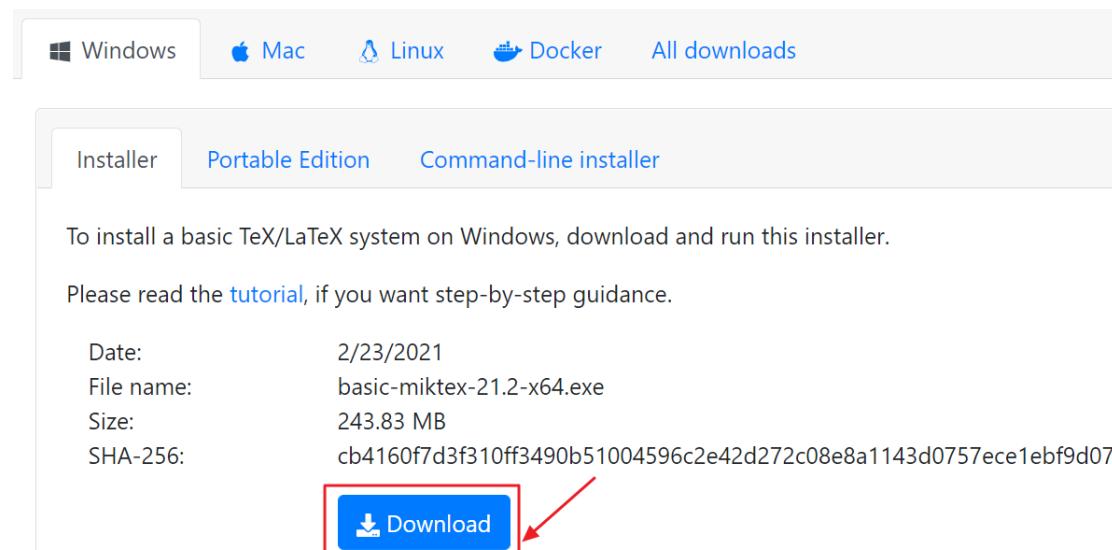
Including Plots

You can also embed plots, for example:

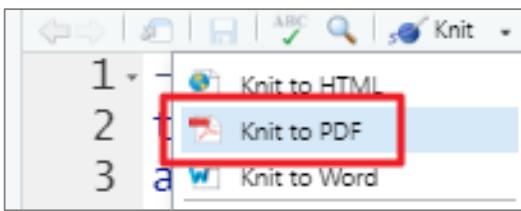
A scatter plot showing the relationship between 'pressure' (Y-axis, 0 to 800) and 'speed' (X-axis, 0 to 350). The data points show a positive correlation, with most points clustered below 100 speed and 200 pressure, and a few outliers at higher speeds and pressures.

# RStudio - Markdown : PDF

- 下載 Miktex: <https://miktex.org/download>
- basic-miktex-21.2-x64.exe (243.83MB)



# Knit to PDF

A screenshot of a PDF viewer displaying a document titled 'bigdata.pdf'. The document contains the following text:

big data  
alan lee  
2021/3/29

**R Markdown**

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

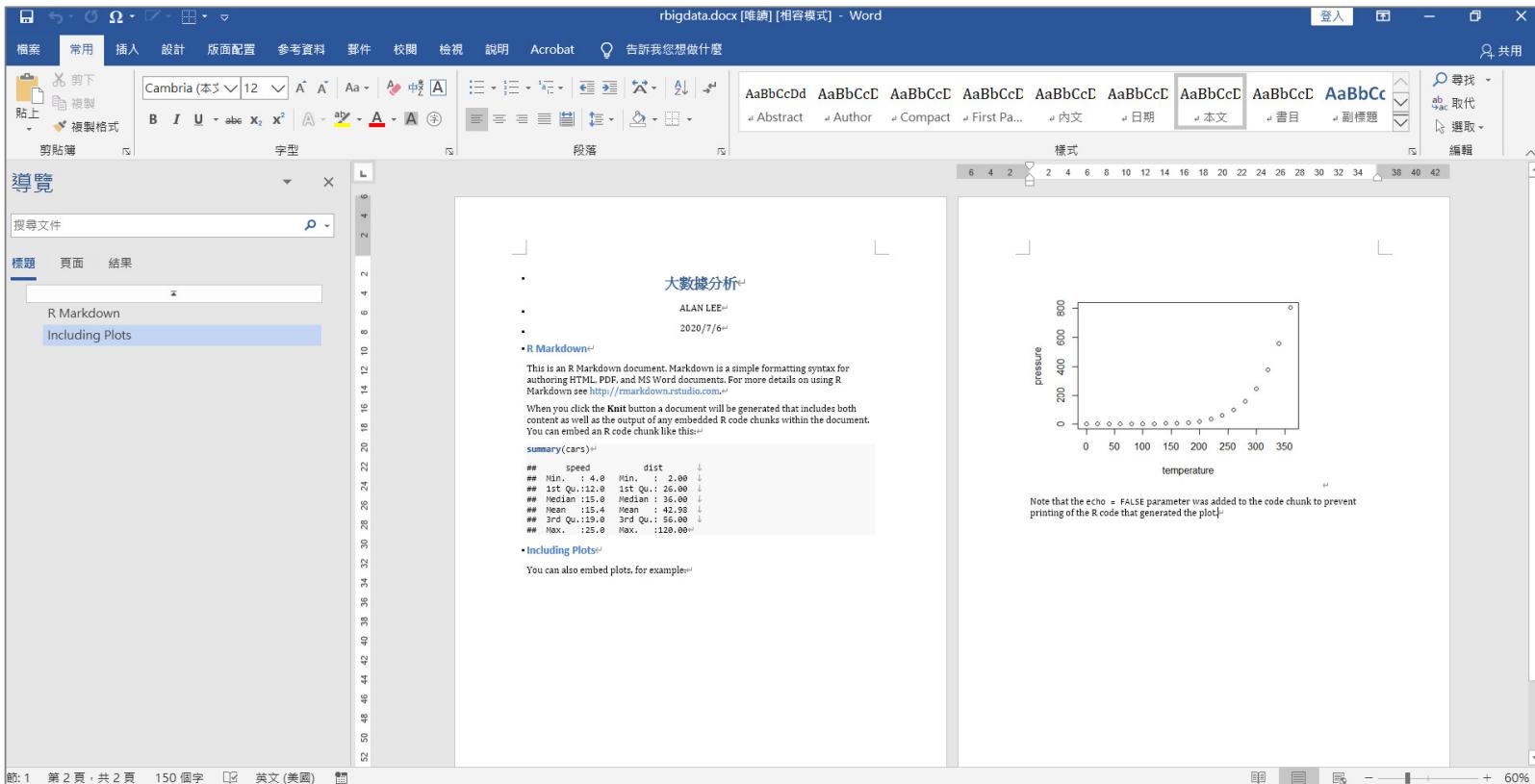
```
summary(cars)
```

speed	dist
Min. : 4.00	Min. : 2.00
1st Qu.:12.00	1st Qu.: 26.00
Median :15.0	Median : 36.00
Mean :15.4	Mean : 42.98
3rd Qu.:29.0	3rd Qu.: 56.00
Max. :25.0	Max. :120.00

**Including Plots**

You can also embed plots, for example:

# RStudio - Markdown : Word

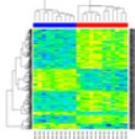
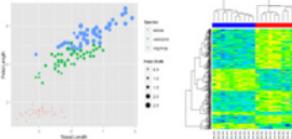


# R學習

- <http://rwepa.blogspot.com/>

**RWEPA Since 2013**

library(shiny)  
shinyUI(pageWithSidebar(library(foreach)  
headerPanel("Shiny"),  
sidebarPanel(c1 <- makecluster  
map.taiwan <- get\_map(location =  
ggmap(map.taiwan)



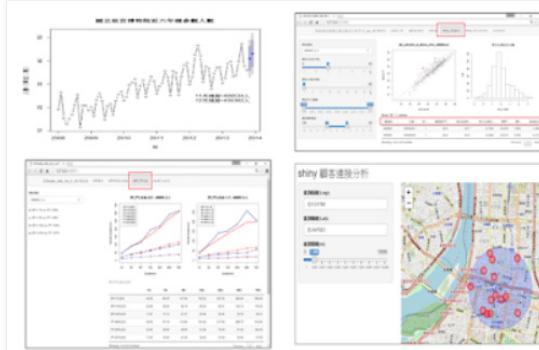
歡迎光臨 RWEPA 部落格  
Welcome to RWEPA blog

system.time(ndsindex  
# user system elapsed

歡迎來到 RWEPA blog, 成立宗旨為提供免費R軟體的相關資訊。R包括大量套件可應用於不同領域, 例如: 2D/3D互動式繪圖, 資料視覺化, 資料探勘, 線性與非線性最佳化問題, 時間序列, 空間資料, 財務分析, 多變量分析, 問卷調查, 實驗設計, 統計製程管制, 存活分析, 臨床實驗分析, 社會網絡分析, 生物資訊, 醫學統計等。

2021年4月12日 星期一

### 2021-R軟體與Shiny Web應用程式設計



免費教學

搜尋此網誌 (例: task)

搜尋

- GitHub DataDemo
- iPAS-R-tutorial
- iPAS-Python-tutorial
- ★★★R入門資料分析與視覺化(付費,中文字幕)
- ★★★★R商業預測與應用(付費,中文字幕)
- ★★★★R語言-直播課程(付費)
- R教學-基礎篇/程式碼(免費)
- Python程式設計PDF(免費)
- ★R 4.0.5-Windows下載
- ★RStudio-1.4.1106下載
- ★RStudio Daily
- R-bloggers

下載 R,RStudio

# 工作目錄

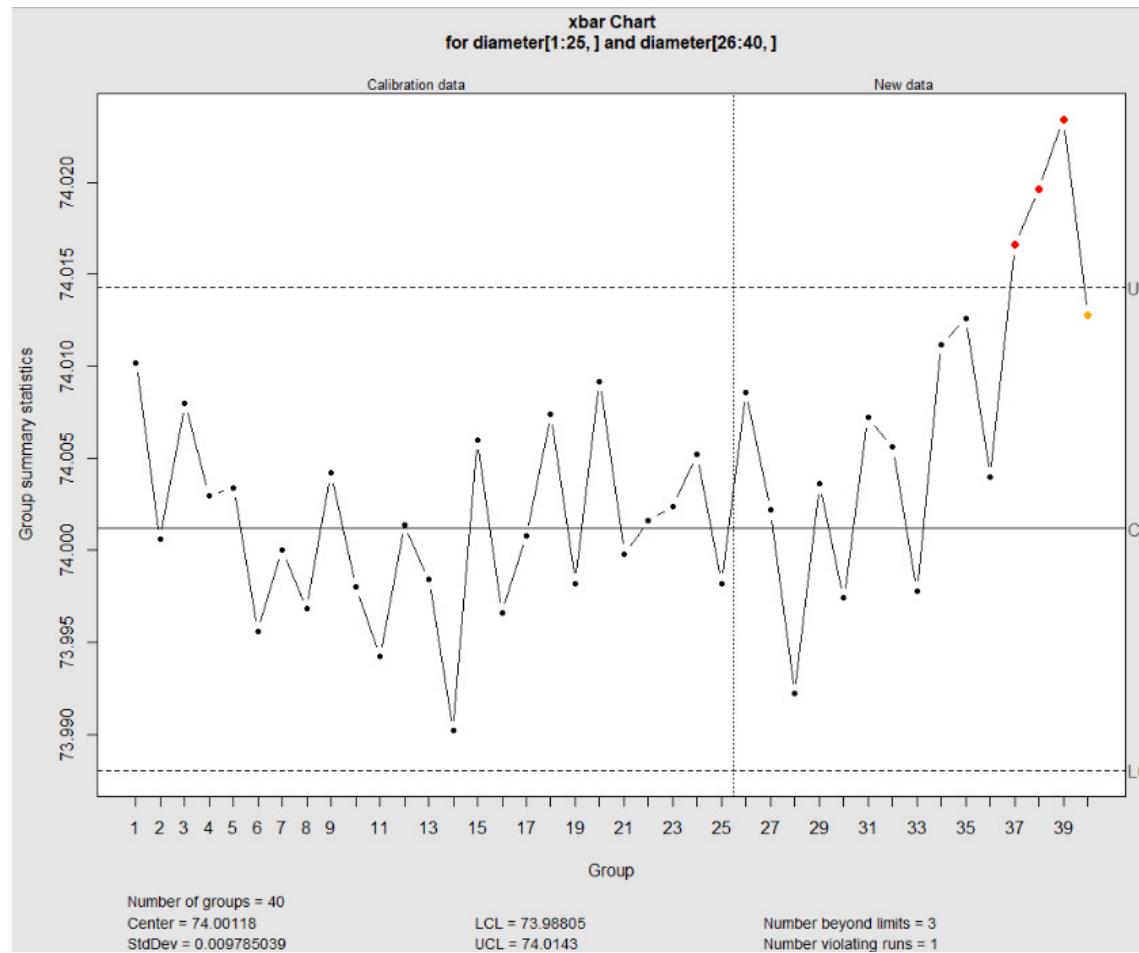
```
> # 取得工作目錄
> getwd()
[1] "C:/Users/88697/Documents"
>
> # 設定工作目錄
> # 先建立 C:/rdata 資料夾
> setwd("C:/rdata")
>
> getwd()
[1] "C:/rdata"
>
```

# qcc: Quality Control Charts

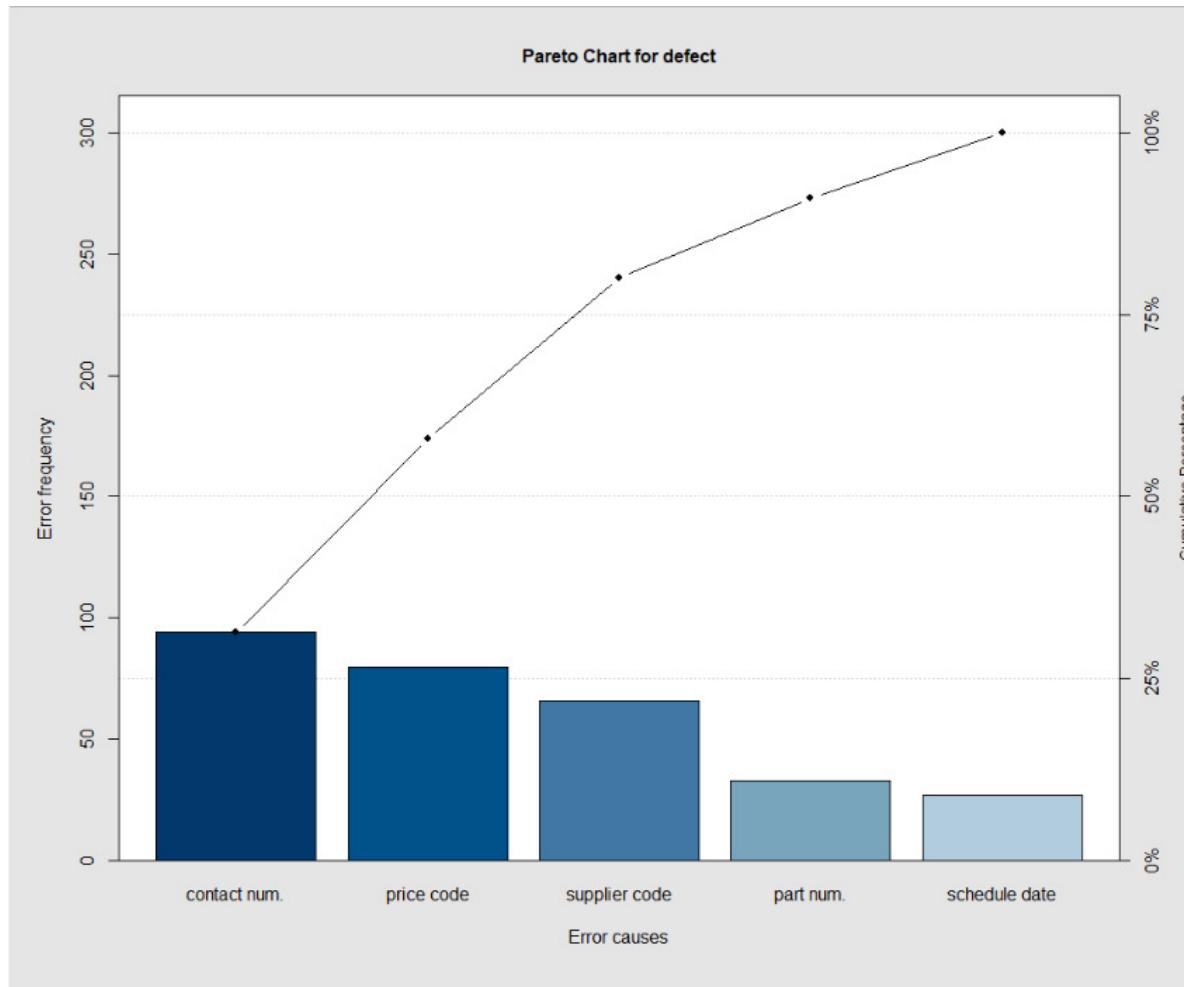
- <https://cran.r-project.org/web/packages/qcc/index.html>

```
> # qcc package
> # https://cran.r-project.org/web/packages/qcc/index.html
>
> library(qcc)
[1] "Quality Control Charts and"
[2] "Statistical Process Control"
[3] "version 2.7"
Type 'citation("qcc")' for citing this R package in publications.
>
> # x-bar chart
> data(pistonrings)
> diameter = with(pistonrings, qcc.groups(diameter, sample))
> head(diameter)
     [,1]   [,2]   [,3]   [,4]   [,5]
1 74.030 74.002 74.019 73.992 74.008
2 73.995 73.992 74.001 74.011 74.004
3 73.988 74.024 74.021 74.005 74.002
4 74.002 73.996 73.993 74.015 74.009
5 73.992 74.007 74.015 73.989 74.014
6 74.009 73.994 73.997 73.985 73.993
> qcc(diameter[1:25,], type="xbar", newdata=diameter[26:40,])
List of 15
 $ call      : language qcc(data = diameter[1:25, ], type = "xbar", newdata = diameter[26:40, ])
 $ type      : chr "xbar"
 $ data.name : chr "diameter[1:25, ]"
 $ data      : num [1:25, 1:5] 74 74 74 74 74 ...
   .. attr(*, "dimnames")=List of 2
 $ statistics : Named num [1:25] 74 74 74 74 74 ...
   .. attr(*, "names")= chr [1:25] "1" "2" "3" "4" ...
```

# $\bar{x}$ chart (平均值管製圖)

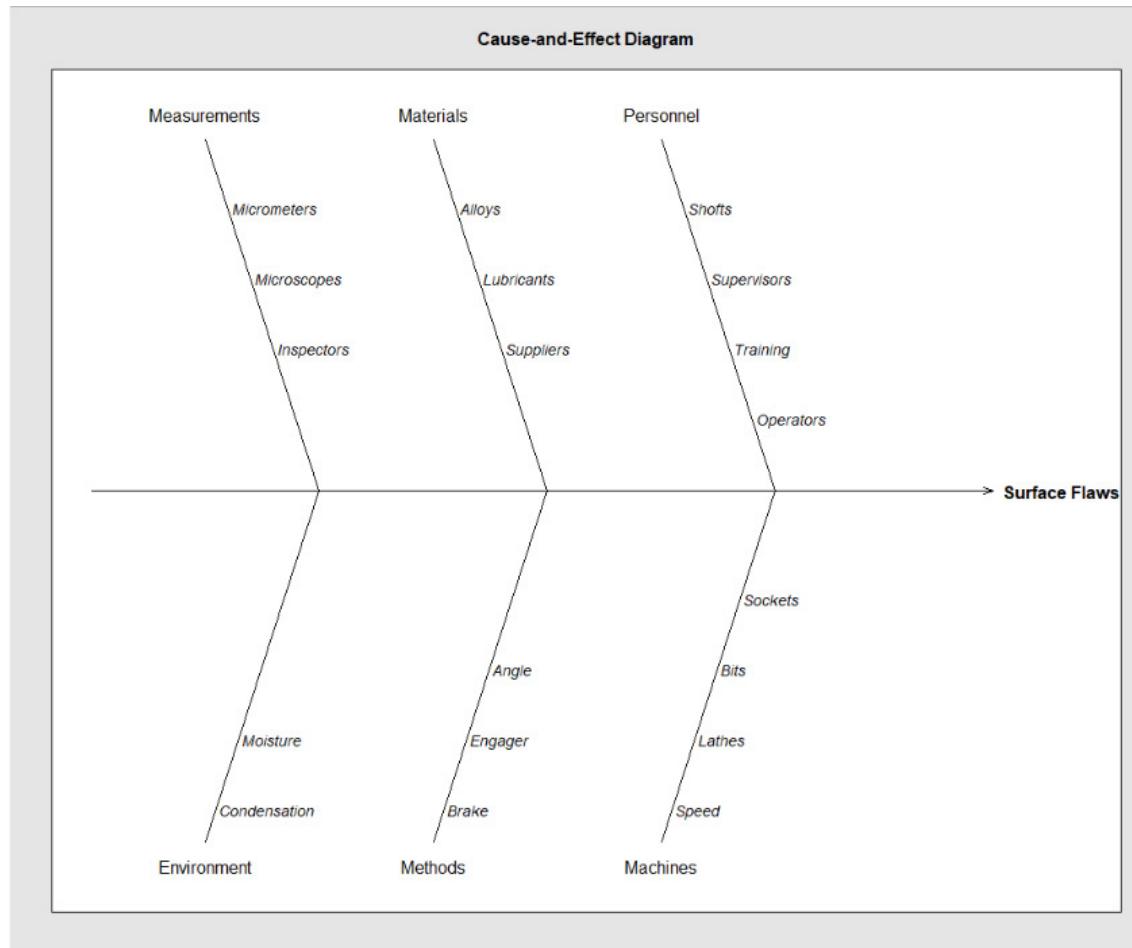


# Pareto chart (柏拉圖)

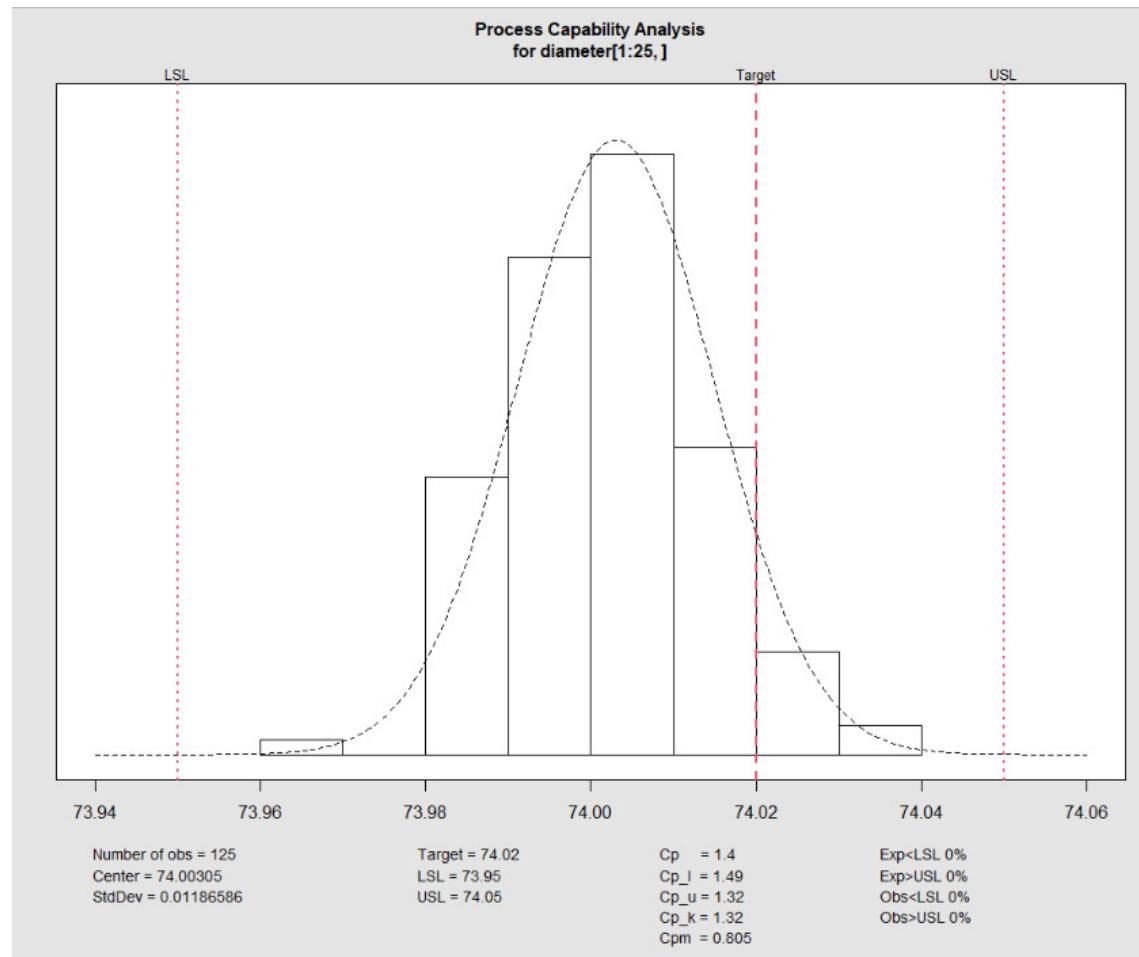


# Cause and effect diagram (要因分析圖)

## 魚骨圖(Fishbone Diagram), 石川圖(Isakiawa Diagram)



process capability index (製程能力 Cp)



## 2.Bioconductor實務應用

---

<https://www.bioconductor.org/>



The screenshot shows the Bioconductor website homepage. A yellow callout bubble on the left points to the "功能" (Features) section of the "About Bioconductor" page. Another yellow callout bubble on the right points to the "研討會" (Conferences) section. A third yellow callout bubble at the bottom points to the "Docker" section of the "News" page. A fourth yellow callout bubble on the right points to the "安裝" (Installation) section of the "Learn" page.

**功能**

**About Bioconductor**

*Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data.*

*Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, and an active user community. Bioconductor is also available as an [AMI](#) (Amazon Machine Image) and [Docker](#) images.*

**News**

**Docker**

- Bioconductor announced. Please view for important deadlines.*
- Nominations for the Bioconductor 2021 Awards now Open! See [award page](#) for more details or use this [nomination form](#).*
- See our [google calendar](#) for events,*

**BioC 2021**

Visit the [BioC 2021](#) website for complete conference information! The virtual conference will be held August 4-6, 2021!

News highlights:

- Registration is Open! [Register Here](#).
- Bioconductor 2021 Award nominations now open! See [award page](#) for more details
- See the list of confirmed speakers on the [website home page](#)

**Install »**

- Discover [1974 software packages](#) available in Bioconductor release 3.12.

Get started with Bioconductor

- 安装**
- [Install Bioconductor](#)
- [Get support](#)
- [Latest newsletter](#)
- [Follow us on twitter](#)
- [Install R](#)

**Learn »**

Master Bioconductor tools

- [Courses](#)
- [Support site](#)
- [Package vignettes](#)
- [Literature citations](#)
- [Common work flows](#)
- [FAQ](#)
- [Community resources](#)
- [Videos](#)

# 安裝 Bioconductor 基礎套件

```
# 安裝 Bioconductor 基礎套件
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install(version = "3.12")
```

# Common work flows

Bioconductor version 3.12 (Release)

Autocomplete biocViews search:

- ▶ Software (1975)
- ▶ AnnotationData (971) rnaseqGene ↑
- ▶ ExperimentData (398)
- Workflow (28)
  - AnnotationWorkflow (3)
  - BasicWorkflow (5)
  - EpigeneticsWorkflow (4)
  - GeneExpressionWorkflow (11)
  - GenomicVariantsWorkflow (2)
  - ImmunoOncologyWorkflow (14)
  - ProteomicsWorkflow (2)
  - ResourceQueryingWorkflow (2)
  - SingleCellWorkflow (2)

Packages found under Workflow:

Rank based on number of downloads: lower numbers are more frequently downloaded.

Show All entries Search table:

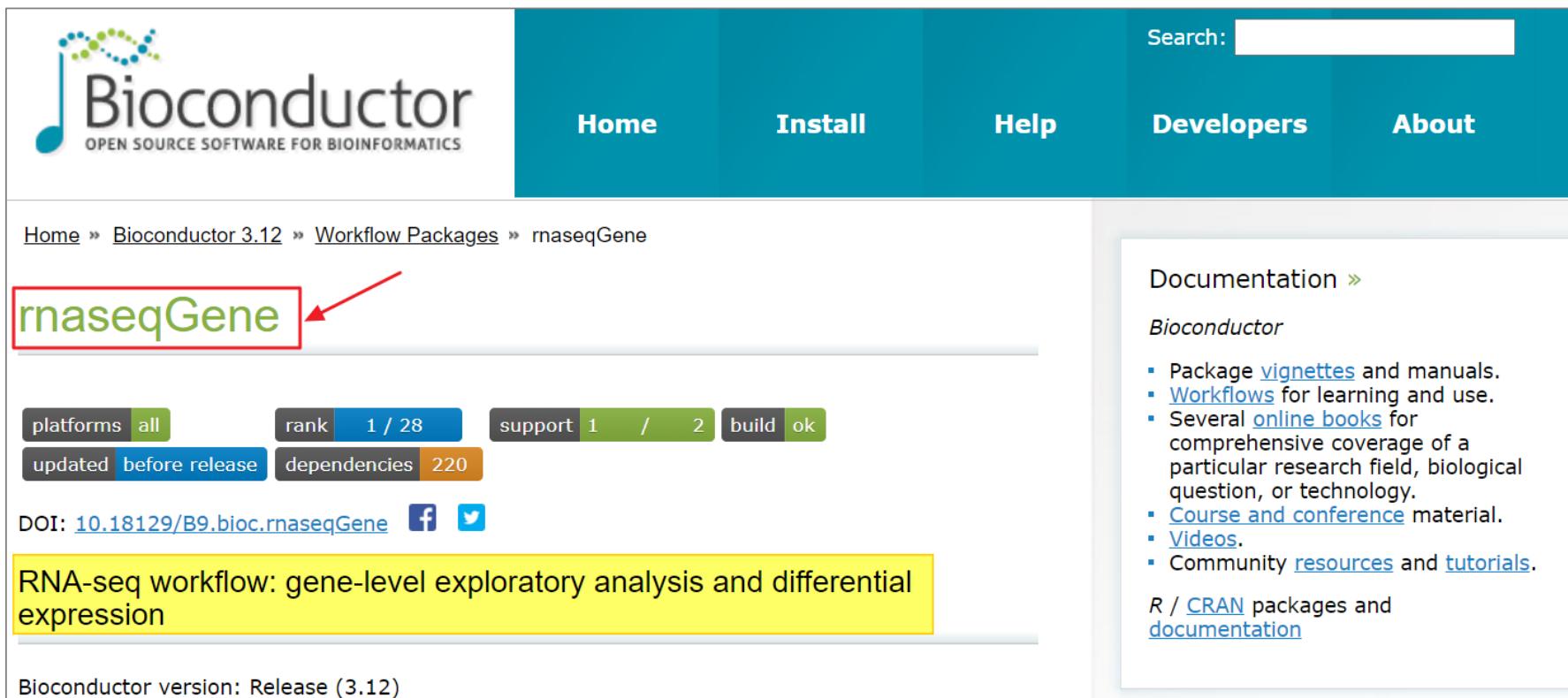
Rank	Title	Maintainer	Package
1	RNA-seq workflow: gene-level exploratory analysis and differential expression	Michael Love	<a href="#">rnaseqGene</a>
2	Changing genomic coordinate systems with rtracklayer::liftOver	Bioconductor Package Maintainer	<a href="#">liftOver</a>
3	RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR	Matthew Ritchie	<a href="#">RNAseq123</a>
4	TCGA Workflow Analyze cancer genomics and epigenomics data using Bioconductor packages	go pedraoui Silva	<a href="#">TCGAWorkflow</a>
5	A cross-package Bioconductor workflow for analysing methylation array data	Jovana Maksimovic	<a href="#">methylationArrayAnalysis</a>

第1名

含中文說明

# rnaseqGene 套件

- <https://bioconductor.org/packages/release/workflows/html/rnaseqGene.html>



The screenshot shows the Bioconductor website for the rnaseqGene package. The top navigation bar includes links for Home, Install, Help, Developers, and About. A search bar is also present. The main content area displays the package details for rnaseqGene, including its rank (1 / 28), support (1 / 2), build status (ok), and dependencies (220). It also shows the DOI (10.18129/B9.bioc.rnaseqGene) and social media links for Facebook and Twitter. A yellow box highlights the package's description: "RNA-seq workflow: gene-level exploratory analysis and differential expression". The bottom of the page indicates the Bioconductor version is Release (3.12).

rnaseqGene

platforms all rank 1 / 28 support 1 / 2 build ok  
updated before release dependencies 220

DOI: [10.18129/B9.bioc.rnaseqGene](https://doi.org/10.18129/B9.bioc.rnaseqGene) [f](#) [t](#)

RNA-seq workflow: gene-level exploratory analysis and differential expression

Bioconductor version: Release (3.12)

Documentation »  
*Bioconductor*

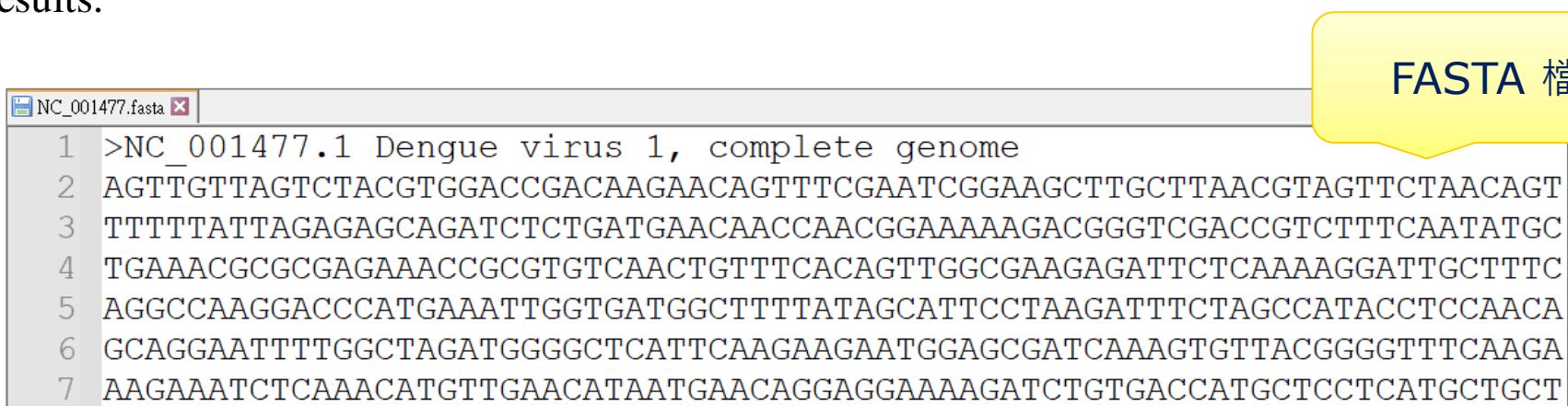
- Package [vignettes](#) and manuals.
- [Workflows](#) for learning and use.
- Several [online books](#) for comprehensive coverage of a particular research field, biological question, or technology.
- [Course and conference](#) material.
- [Videos](#).
- Community [resources](#) and [tutorials](#).

*R / CRAN* packages and documentation

# rnaseqGene 套件 - 功能

- Here we walk through an end-to-end gene-level RNA-seq differential expression workflow using Bioconductor packages.
- We will start from the **FASTQ files**, show how these were aligned to the reference genome, and prepare a count matrix which tallies the number of RNA-seq reads/fragments within each gene for each sample.
- We will perform **exploratory data analysis (EDA)** for quality assessment and to explore the relationship between samples, perform differential **gene expression analysis**, and visually explore the results.

FASTA 檔案



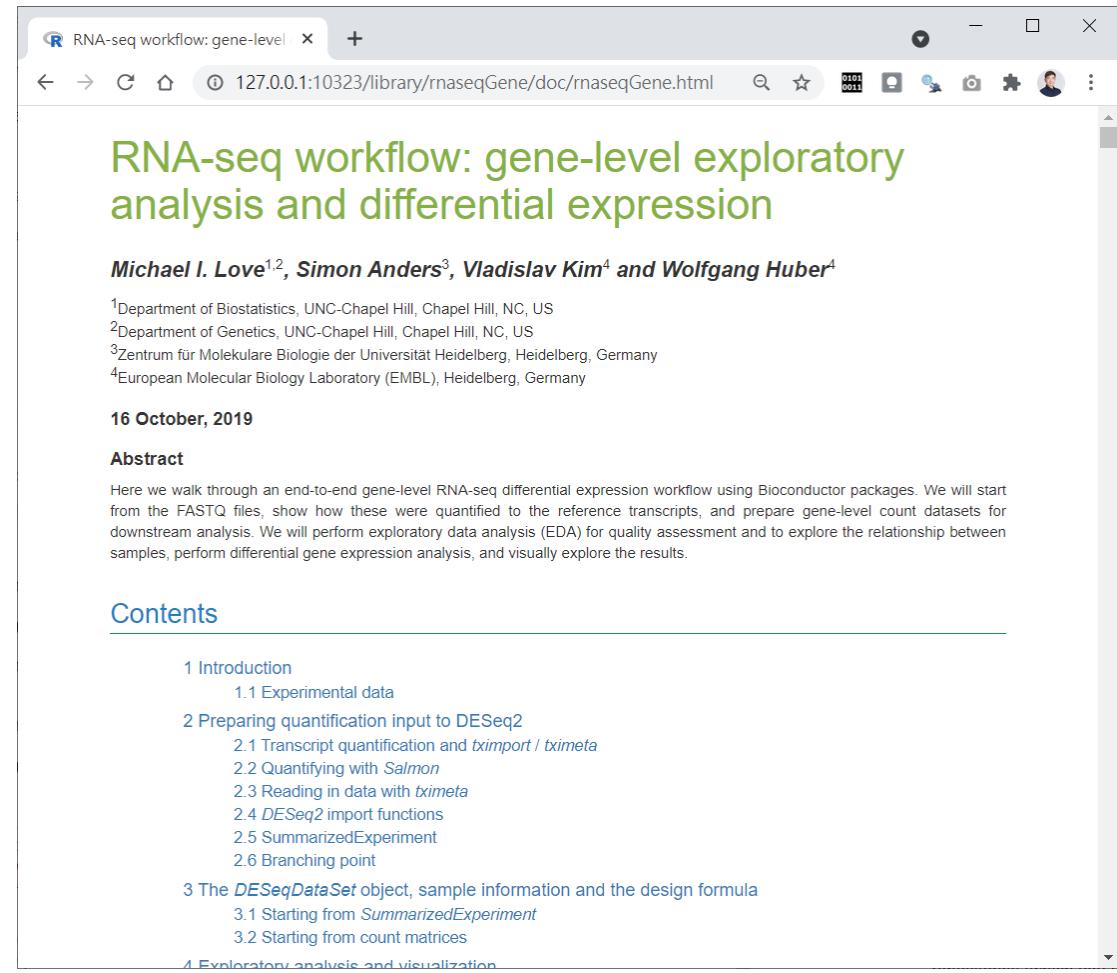
```
NC_001477.fasta
1 >NC_001477.1 Dengue virus 1, complete genome
2 AGTTGTTAGTCTACGTGGACCGACAAGAACAGTTCGAACATCGGAAGCTTGCTTAACGTAGTTCTAACAGT
3 TTTTTATTAGAGAGCAGATCTCTGATGAACAAACCAACGGAAAAAGACGGGTCGACCGTCTTCAATATGC
4 TGAAACGCCGAGAAACCCGTGTCAACTGTTCACAGTTGGCGAAGAGATTCTCAAAGGATTGCTTTC
5 AGGCCAAGGACCCATGAAATTGGTGTGGCTTTATAGCATTCTAACAGATTCTAGCCATACCTCCAACA
6 GCAGGAATTTGGCTAGATGGGGCTCATTCAAGAAGAATGGAGCGATCAAAGTGTACGGGTTCAAGA
7 AAGAAATCTCAAACATGTTAACATAATGAACAGGAGGAAAAGATCTGTGACCAGCTCCTCATGCTGCT
```

# 安裝 rnaseqGene 套件

```
> BiocManager::install("rnaseqGene")
'getOption("repos")' replaces Bioconductor standard repositories, see '?repositories' for details
replacement repositories:
  CRAN: https://cran.rstudio.com/
Bioconductor version 3.12 (BiocManager 1.30.12), R 4.0.5 (2021-03-31)
Installing package(s) 'rnaseqGene'
also installing the dependencies 'fs', 'sass', 'jquerylib', 'backports', 'httpuv', 'sourcetools', 'later', 'lazyeval', 'commonmark', 'bslib', 'htmlwidgets', 'crosstalk', 'formatR', 'checkmate', 'rstudioapi', 'cpp11', 'system', 'stringi', 'assertthat', 'shiny', 'DT', 'cli', 'utf8', 'lambda.r', 'futile.options', 'numDeriv', 'mvtnorm', 'colorspace', 'bit', 'cachem', 'Formula', 'data.table', 'htmlTable', 'viridis', 'forcats', 'rmarkdown', 'hms', 'prettyunits', 'askpass', 'htmltools', 'xfun', 'tinytex', 'evaluate', 'highr', 'markdown', 'stringr'
```

# rnaseqGene 套件 - 線上說明

```
# 線上說明  
browseVignettes("rnaseqGene")
```



The screenshot shows a web browser window displaying the "RNA-seq workflow: gene-level exploratory analysis and differential expression" vignette from the rnaseqGene package. The URL in the address bar is 127.0.0.1:10323/library/rnaseqGene/doc/rnaseqGene.html. The page title is "RNA-seq workflow: gene-level exploratory analysis and differential expression". The authors listed are Michael I. Love<sup>1,2</sup>, Simon Anders<sup>3</sup>, Vladislav Kim<sup>4</sup> and Wolfgang Huber<sup>4</sup>. The text indicates the workflow starts from FASTQ files, quantifies transcripts relative to reference transcripts, and prepares gene-level count datasets for downstream analysis. It includes exploratory data analysis (EDA) for quality assessment and exploring relationships between samples, differential gene expression analysis, and visual exploration of results. The page also lists the contents of the vignette, including sections on introduction, preparing quantification input to DESeq2, the DESeqDataSet object, sample information, design formula, and exploratory analysis and visualization.

RNA-seq workflow: gene-level exploratory analysis and differential expression

Michael I. Love<sup>1,2</sup>, Simon Anders<sup>3</sup>, Vladislav Kim<sup>4</sup> and Wolfgang Huber<sup>4</sup>

<sup>1</sup>Department of Biostatistics, UNC-Chapel Hill, Chapel Hill, NC, US  
<sup>2</sup>Department of Genetics, UNC-Chapel Hill, Chapel Hill, NC, US  
<sup>3</sup>Zentrum für Molekulare Biologie der Universität Heidelberg, Heidelberg, Germany  
<sup>4</sup>European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

16 October, 2019

**Abstract**

Here we walk through an end-to-end gene-level RNA-seq differential expression workflow using Bioconductor packages. We will start from the FASTQ files, show how these were quantified to the reference transcripts, and prepare gene-level count datasets for downstream analysis. We will perform exploratory data analysis (EDA) for quality assessment and to explore the relationship between samples, perform differential gene expression analysis, and visually explore the results.

**Contents**

- 1 Introduction
  - 1.1 Experimental data
- 2 Preparing quantification input to DESeq2
  - 2.1 Transcript quantification and *tximport* / *tximeta*
  - 2.2 Quantifying with *Salmon*
  - 2.3 Reading in data with *tximeta*
  - 2.4 *DESeq2* import functions
  - 2.5 *SummarizedExperiment*
  - 2.6 Branching point
- 3 The *DESeqDataSet* object, sample information and the design formula
  - 3.1 Starting from *SummarizedExperiment*
  - 3.2 Starting from count matrices
- 4 Exploratory analysis and visualization

### 3. 品質控制技術

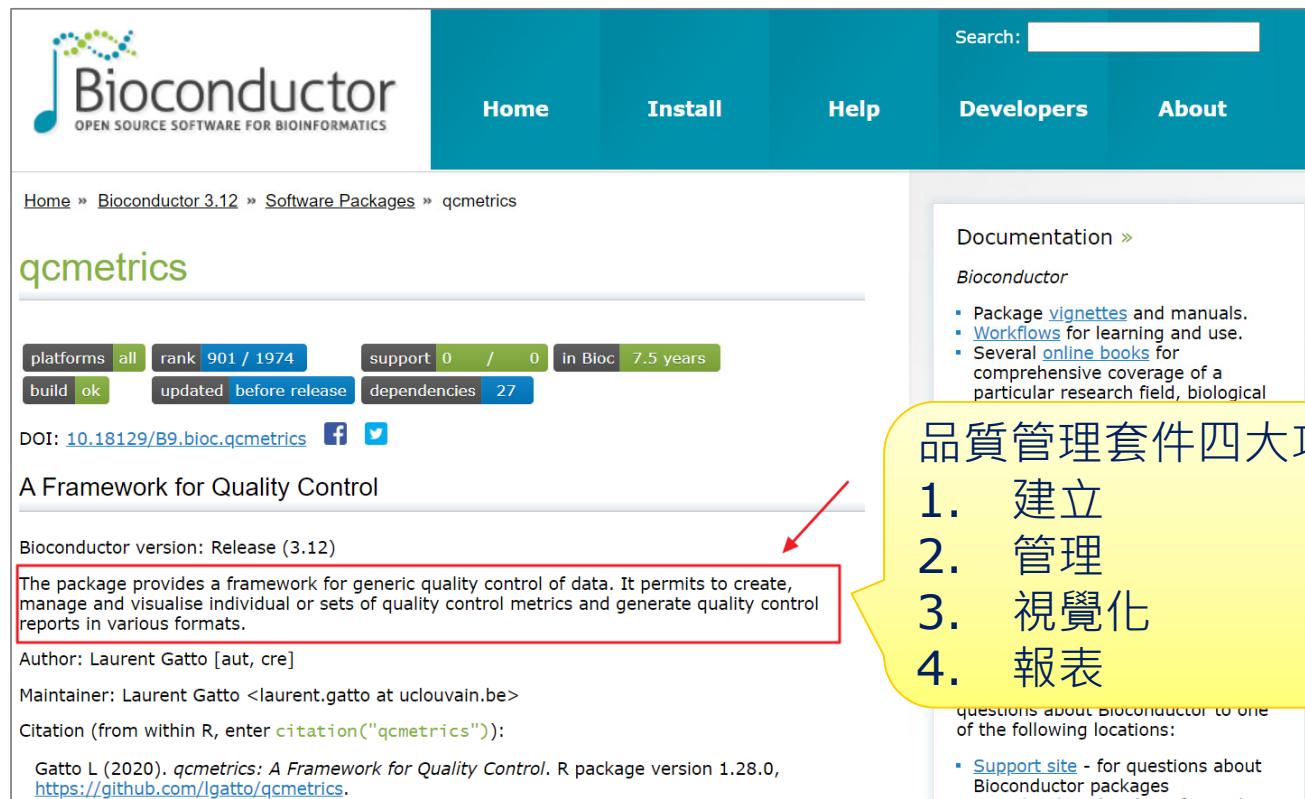
---

## 品質控制

- 在實驗或陣列中獲得的資料的質量對於資料分析和正確解釋微陣列資料是重要環結之一。
- 單個或幾個異常資料都可能破壞大型數據集的分析結果。在首先確保數據質量可以接受之前，不應進行分析。
- 如果樣品在收集後未立即正確處理，則RNA易於快速降解（degradation）而影響分析結果。
- 若樣品的品質不合格最直接影響的就是產出的資料量不足或是導致資料豐富度不夠而無法呈現最真實的分析結果。

# qcmetrics 套件

- <https://www.bioconductor.org/packages/release/bioc/html/qcmetrics.html>



The screenshot shows the Bioconductor qcmetrics package page. At the top, there's a navigation bar with links for Home, Install, Help, Developers, and About. A search bar is also present. Below the navigation, the package name 'qcmetrics' is displayed along with its Bioconductor version (3.12). Key statistics shown include platforms (all), rank (901 / 1974), support (0 / 0), dependencies (27), build status (ok), and update information (updated before release). A DOI link ([10.18129/B9.bioc.qcmetrics](https://doi.org/10.18129/B9.bioc.qcmetrics)) and social media links for Facebook and Twitter are provided. The main content area describes 'A Framework for Quality Control'. A red box highlights a paragraph about the package's purpose: 'The package provides a framework for generic quality control of data. It permits to create, manage and visualise individual or sets of quality control metrics and generate quality control reports in various formats.' An arrow points from this highlighted text to a yellow callout box containing the text '品質管理套件四大功能'. To the right of the main content, there's a sidebar with documentation links and a 'Support site' section.

Documentation »  
*Bioconductor*

- Package [vignettes](#) and manuals.
- [Workflows](#) for learning and use.
- Several [online books](#) for comprehensive coverage of a particular research field, biological

品質管理套件四大功能

1. 建立
2. 管理
3. 視覺化
4. 報表

questions about Bioconductor to one of the following locations:

- [Support site](#) - for questions about Bioconductor packages

# qcmetrics 套件 demo

```
> # qcmetrics 套件
> # https://www.bioconductor.org/packages/release/bioc/html/qcmetrics.html
>
> # BiocManager::install("qcmetrics")
> library("qcmetrics")
>
> # 建立 QcMetric 物件
> qc <- QcMetric(name = "A test metric")
> class(qc)
[1] "QcMetric"
attr(,"package")
[1] "qcmetrics"
> qc
Object of class "QcMetric"
  Name: A test metric
  Status: NA
  Data: empty
>
> # qcdata: 實際儲存資料
> qcdata(qc, "x") <- rnorm(100)
> qcdata(qc) ## all available qcdata
[1] "x"
```

# 函數 show

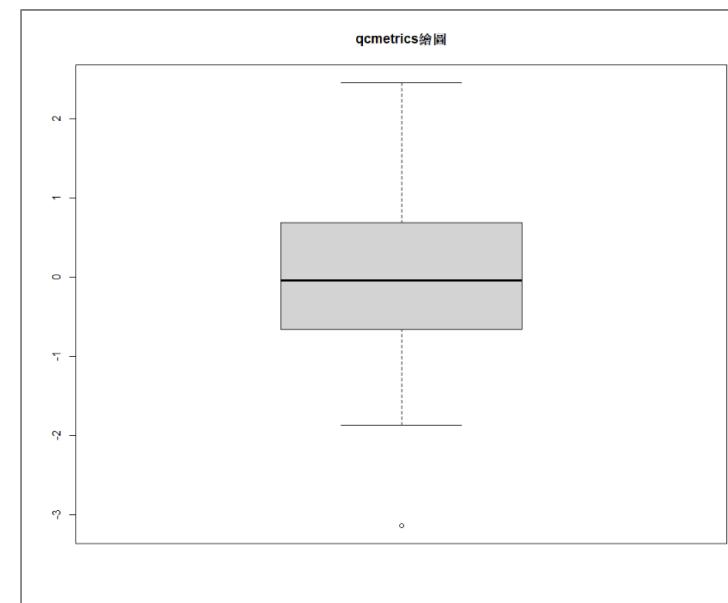
```
> qc
Object of class "QcMetric"
  Name: A test metric
  Status: NA
  Data: x
> show(qc)
Object of class "QcMetric"
  Name: A test metric
  Status: NA
  Data: x
> summary(qcdata(qc, "x"))
    Min. 1st Qu. Median      Mean 3rd Qu.      Max.
-3.13692 -0.65860 -0.04072  0.01924  0.68710  2.44919
> # status: 狀態屬性 {TRUE, FALSE}
> status(qc) <- TRUE
> qc
Object of class "QcMetric"
  Name: A test metric
  Status: TRUE
  Data: x
>
```

# 函數 boxplot

```
> # 預設繪圖有錯誤  
> plot(qc)  
Warning message:  
In x@plot(x, ...) : No specific plot function defined  
> # 新增繪圖方法  
> plot.qc <- function(object, ...) boxplot(qcdata(object, "x"), ...)  
>  
> # 繪製盒鬚圖  
> plot(qc, main = "qcmetrics繪圖")
```

1

2



# Metadata 元資料

```
> # 建立元資料 metadata
> metadata(qcm) <- list(author = "李明昌(ALAN LEE)",
+                           lab = "Big array lab")
> qcm
Object of class "QcMetrics"
containing 1 QC metrics.
and 2 metadata variables.
> mdata(qcm)
$author
[1] "李明昌(ALAN LEE)"

$lab
[1] "Big array lab"

> # 更新元資料
> metadata(qcm) <- list(author = "李明昌(ALAN LEE)",
+                           lab = "Big array lab",
+                           organization = "中華民國品質學會")
> mdata(qcm)
$author
[1] "李明昌(ALAN LEE)"

$lab
[1] "Big array lab"

$organization
[1] "中華民國品質學會"
```

# 函數 yaqc – 計算 YAQC 統計

```
# MAQCsubsetAFX 套件內建資料集 refA
# BiocManager::install("MAQCsubsetAFX")

library("MAQCsubsetAFX")
data(refA)
refA

library("affy")

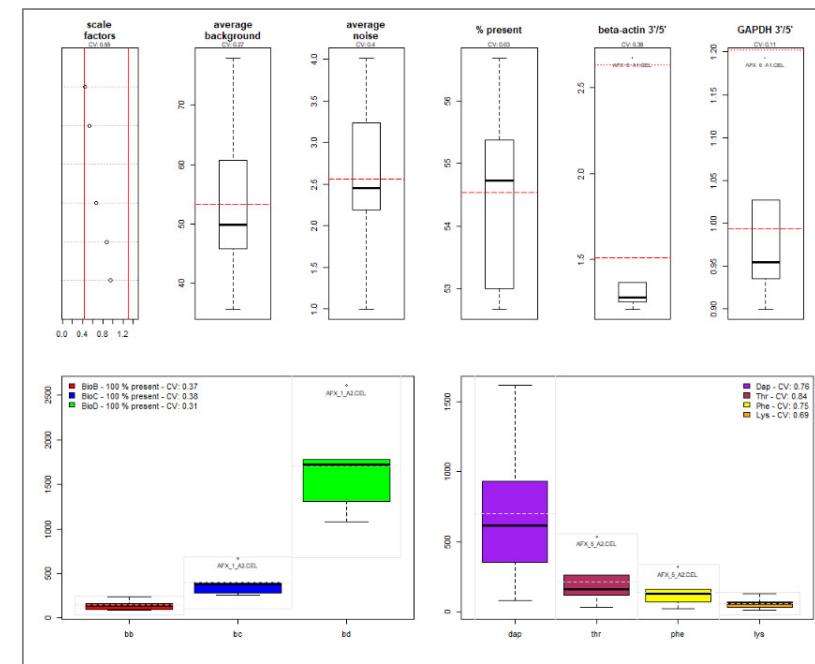
# 計算 RNA degradation
deg <- AffyRNADeg(refA)
deg

# BiocManager::install("yaqcAffy")
library("yaqcAffy")

# 計算 YAQCStats
yqc <- yaqc(refA)
show(yqc)

plot(yqc)
```

```
> refa
AffyBatch object
size of arrays=1164x1164 features (19 kb)
cdf=HG-U133_Plus_2 (54675 affyids)
number of samples=6
number of genes=54675
annotation=hgu133plus2
notes=
```



## 函數 qcReport – 建立報表

```
> # qcReport 建立報表
> qcReport(maqcm,
+           reportname = "RNA-deg",
+           type = "html",
+           author = "Ming-Chang Lee")
Report written to RNA-deg.html
>
```

# file:///C:/rdata/RNA-deg.html

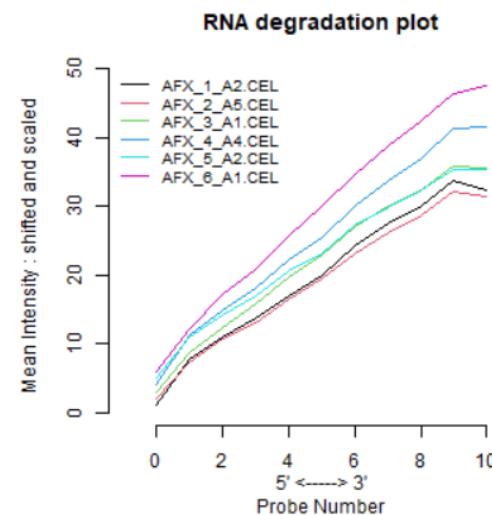
## Quality control report generated with qcmetrics

Author: Ming-Chang Lee

Date: Thu Apr 22 21:42:36 2021

### Affy RNA degradation slopes

```
## Object of class "QcMetric"  
## Name: Affy RNA degradation slopes  
## Status: TRUE  
## Data: deg
```



## 4. 結論與未來展望

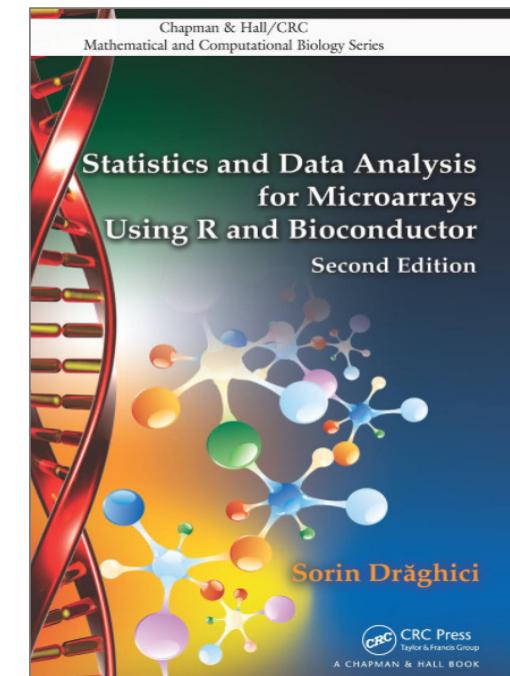
---

# Recap

- R, RStudio 簡介 - qcc 套件
- Bioconductor 實務應用 - rnaseqGene 套件
- 品質控制技術 - qcmetrics 套件
- 生物資訊研究方向
  - Tallulah S. Andrews, Vladimir Yu Kiselev, Davis McCarthy & Martin Hemberg , Tutorial: guidelines for the computational analysis of *single-cell RNA sequencing data*, Nature Protocols volume 16, pages 1–9 (2021)
  - <https://www.nature.com/articles/s41596-020-00409-w>

# 參考資料

- Sorin Drăghici (2012), [Statistics and Data Analysis for Microarrays: Using R and Bioconductor](#), Second Edition, CRC Press, 2012.
- Gatto L (2020), [qcmetrics](#): A Framework for Quality Control. R package version 1.28.0, <https://github.com/lgatto/qcmetrics>.
- R教學-基礎篇(免費)
  - <http://rwepa.blogspot.tw/2013/01/r-201174.html>
- R入門資料分析與視覺化應用教學(付費)
  - <https://courses.mastertalks.tw/courses/R-teacher>
- R商業預測與應用(付費)
  - <https://courses.mastertalks.tw/courses/R-2-teacher>



# 謝謝您的聆聽

## Q & A



李明昌

*alan9956@gmail.com*

<http://rwepa.blogspot.tw/>