生信分析人员如何系统入门R(2019更新版)

> 五年前作为一个初出茅庐的菜鸟生信工程师苦于没有专业交流社群，遂自建了生信菜鸟团QQ群和博客，一点一滴积累了数万人气，进而和若干圈内好友组建了生信技能树联盟，三年前的直播生物信息学编程活动细节还历历在目，QQ群微信群记账录制视频忙的不亦乐乎，因此产生了编程语言系统入门系列教程，如下：

- 生信分析人员如何系统入门python?
- 生信分析人员如何系统入门perl?
- 生信分析人员如何系统入门R?
- 生信分析人员如何系统入门Linux?

现在回过头来看，很多教程已然过时，当然并不是说的知识点过时，其实linux基本上几十年都没有怎么变动过基础知识的，哪怕你现在搜索到十几年前的linux教学视频，也不会觉得尴尬。主要是其中一些资源链接，一些小技巧都过时了，比如R语言安装包，需求切换适合的镜像，或者某些配套书籍课程的URL肯定也会成为死链啦，所以非常有必要系统性整理一下，最新生信分析人员如何系统入门R

写在前面

R语言不仅在生物信息数（主要体现在bioconductor系列包）据处理中发挥着重要作用，其实也是其他主流数据处理人士（包括互联网，金融，游戏行业）的首选工具。所以基本上找到我来咨询如何入门生物信息学的，我都是推荐他必须学的就是R。但是实际上呢，我作为老一辈的生信工程师，所以喜欢perl一点，排斥python，我也稍微看过一些python的语法，个人认

为**R和python**呢almostly 几乎 一模一样的。R的特点就是内置了大量的函数，基本上你认识的英文单词都可以是一个函数，即使不是，你也可以自定义为函数。搞清楚了函数和变量，就可以看懂大部分的R代码了。通常我给初学者的**知识点路线图**如下：

- 了解常量和变量概念
- 加减乘除等运算（计算器）
- 多种数据类型（数值，字符，逻辑，因子）
- 多种数据结构（向量，矩阵，数组，数据框，列表）
- 文件读取和写出
- 简单统计可视化
- 无限量函数学习

六步系统入门R语言

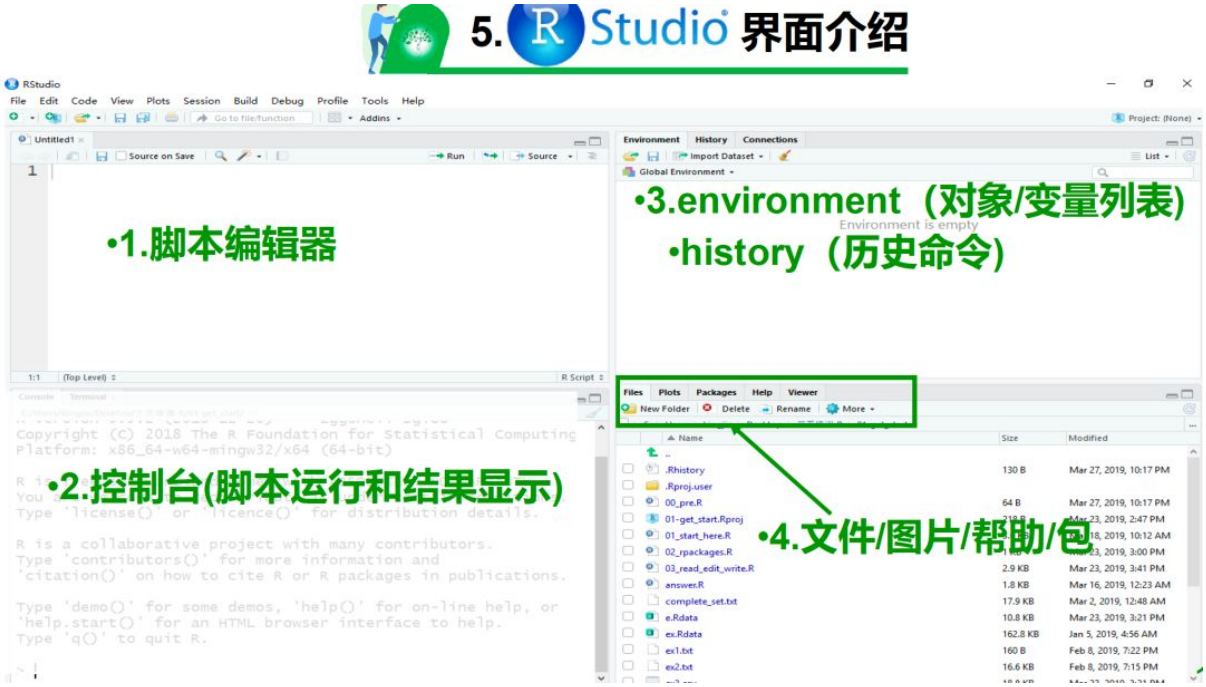
第一步：工欲善其事必先利其器

既然要学习R语言，配套的软件工具必须要摸索到位，掌握一些神操作，大幅度提高你的生产力！

- **下载适合你操作系统的R语言的软件**：https://cran.r-project.org/bin/windows/base/ 或者 https://cran.r-project.org/bin/macosx/
- **下载Rstudio这个R编辑器**：https://www.rstudio.com/products/rstudio/download/（在Rstudio里面写代码会比较方便）

这里借用我们生信技能树优秀讲师（小洁）的R语言入门课程培训PPT内

容，给大家展现Rstudio界面：



几个神技巧

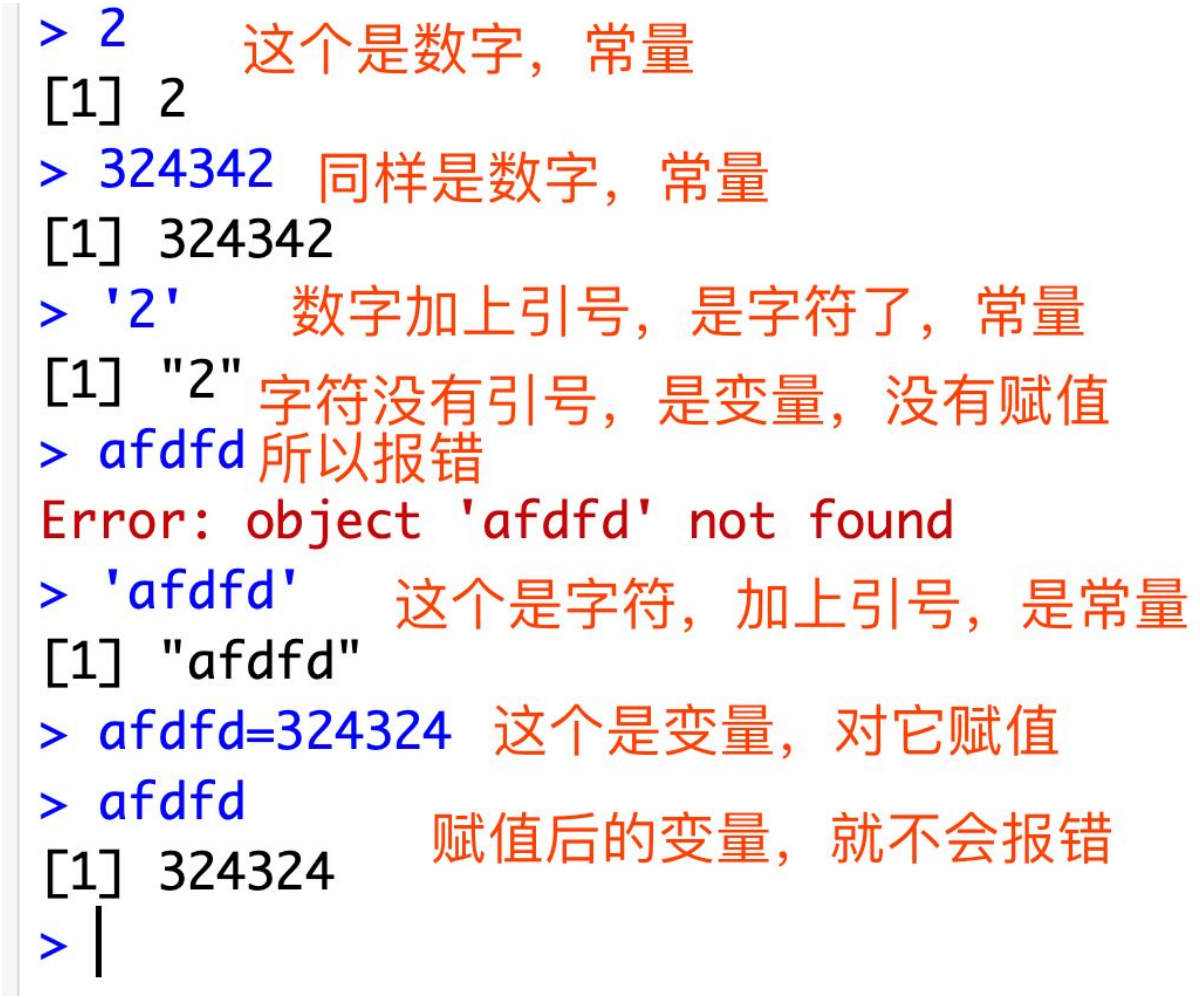
这些技巧真的是一般人我不告诉他，仅仅是因为大家是生信技能树的粉丝我才透露一下下，当然更多技巧你需要去B站看我的10个小时的R语言视频自行摸索啦！

- 用项目的方式管理你的代码！
- 用rdata文件来保存和加载（save和load）变量，类似于Excel表格软件才能打开的xlxs后缀文件。
- 注意你的代码编辑的字符编码格式，否则会出现乱码。
- 如果是Windows用户，注意修改环境变量，否则你的Rstudio会频频报错。
- 善用tab键补全，哪怕把键盘的tab键按坏也无所谓！

第二步：变量和常量

R语言跟python一样，并不是perl或者shell那样需要一些特殊字符($,@等等)开头来代表它是一个变量，只需遵循变量命名规则的字符组合即可，所以呢，如果同样的字符要昨晚常量就必须使用单双引号这样才能区分变量和常量，当然了，数字本身就只能是常量了，变量也不能以数字开头，如果一个数字加上了引号，它就是常量的字符了。

更为详细的介绍如下：



多种数据类型（数值，字符，逻辑，因子）

上面我们讲解了最简单的常量和变量，它们有数字和字符的差异，其实还有其它两个常用的数据类型，就是逻辑和因子。（请注意，我没有讲解因子哦）

```
> name='xiaoming'
> name
[1] "xiaoming"
> score=95
> score
[1] 95
> score>60
[1] TRUE
>
```

字符向量，变量

数字向量，变量

逻辑向量，变量

多种数据结构（向量，矩阵，数组，数据框，列表）

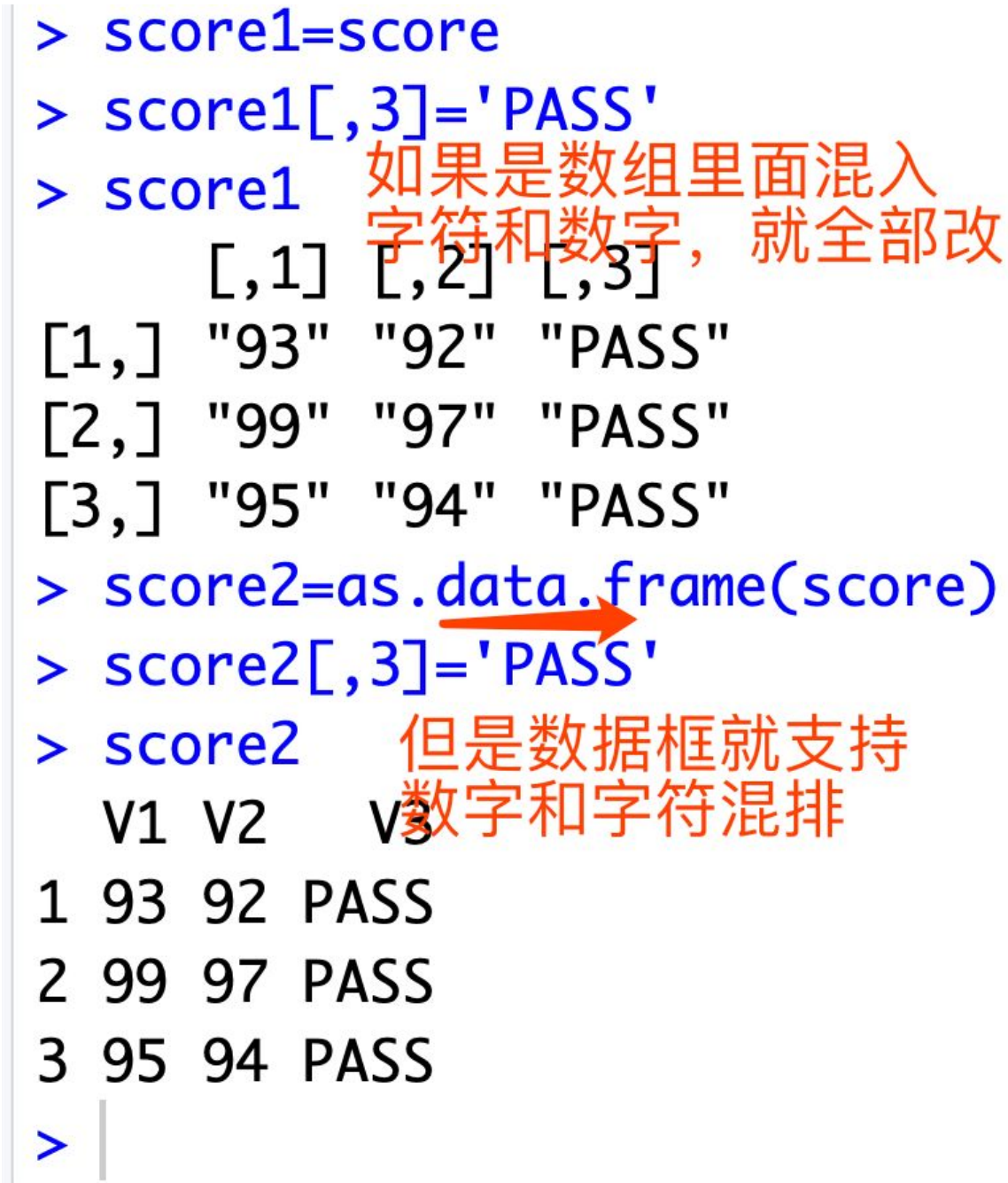
前面我们看到的常量和变量，**都只有一个元素**，是最简单的向量，实际上向量可以有多个元素，比如小明同学这个变量，他可以有名字（通常是字符），也可以有语数外的考试成绩（通常是数值），这就是长度不等的向量，如下：

但是通常一个班级不可能只有一个学生，如果有多个同学（还有小红和小绿），他们都有语数外成绩，就是一个矩阵了，一个二维矩阵，属于数组的范畴。

同样的道理，小明，小红和小绿这3个同学的语数外三门考试的成绩，可以跨越**多个年度多次统计**，这样就多了一个时间的维度，就是三维数组啦！

同样的道理，维度可以无限增加，比如增加一**期中、期末考试的分类维度**，就是4维数组。但这个不是重点。

重点是有些时候，向量和数组是不足以满足现实需求的，比如考试成绩里面，语数外都是得分，是数字，当然没有问题，但是呢，他们有一个政治成绩是PASS和failed这样的简单分类，我们的数组描述起来就有点困难了。



不过，即使我们解决了数字与字符混排这个麻烦，通过引入数据框这个知识点，比如病人的临床信息，就可以有不同的列，分别是数字的年龄，字符的TNM分期等等。但是仍然是有个问题，因为它太规范了，只能是多少行多少列的规则格式，实际上有可能是同一个班级里面的不同学生，他们可能是选修不同的课程，这样他们记录的成绩本来就是不一样的。

```
> score=list(xiaoming=c(92,95,97),xiaohong=c(98,95))
> score
$xiaoming
[1] 92 95 97

$xiaohong
[1] 98 95

> data.frame(xiaoming=c(92,95,97),xiaohong=c(98,95))
Error in data.frame(xiaoming = c(92, 95, 97), xiaohong = c(98, 95)) :
  arguments imply differing number of rows: 3, 2
>
```

这种情况，不同的学生的选修课程不一样，就只能是使用list结构
它允许里面的不同元素不同的长度

如果是数据框，就会报错

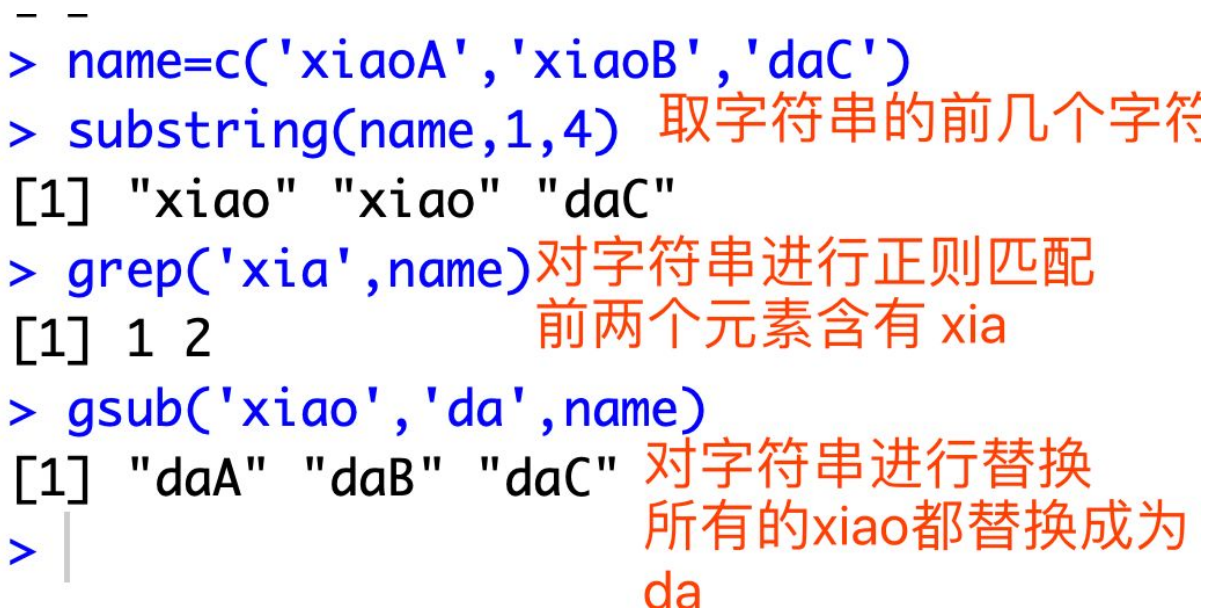
第三步：了解变量的基础操作函数

有了变量和常量的概念，数据类型和数据结构的概念，我们就相当于是半只脚踏入R语言的大门了，**因为我们有了数值型向量**，所以可以对他们进行加减乘除的基本数学运算，如下：

```
> score-1
[1] 96 94 85
> score+1
[1] 98 96 87
> score/2
[1] 48.5 47.5 43.0
> score*3
[1] 291 285 258
> score^2
[1] 9409 9025 7396
> sqrt(score)
[1] 9.848858 9.746794 9.273618
> log2(score)
[1] 6.599913 6.569856 6.426265
>
```

一系列的数学计算

因为我们有了字符型向量，也可以对其进行一系列的字符串操作



到这里，大家相当于把一本R语言教材书籍看了一半了，因为对没有编程思维的人来说，仅仅是理解这些知识点就耗费了她几个月的力气，不过对于有编程思维的人来说，也就是一杯咖啡的功夫过去了而已。

还有更多的函数，有待大家自行摸索和练习，一般来说需要入门R，掌握的函数起码得超过200个，更多函数，大家可以在我的生信五周年演讲素材看到400行基础代码：
https://github.com/jmzeng1314/5years/blob/master/learn-R/tasks/1-guozi-400.R

```
str,
```

`class,names,row.names,col.names,length,unique,view,min,max,summay,table`

这里需要重点强调的就是学习help函数（你必须要把help函数用一百次以上，不然你不可能入门的！）还需要仔细观察你的变量被你操作过程的变化，所以str,class两个函数也需要敲一百次以上。

值得提醒的是，通常我们的函数所操作的这些变量，都不是来自于我们自己

创造，我们处理生物信息学数据一般很少会手动创建这些对象，都是从文本里面读取，比如kegg数据库文件，差异分析结果，RNA-seq的表达量矩阵，但是读入之后，我们的**重点就是知道它们变成了什么**，该如何去**一步步的转换它们**。（就是大家通常说的数据清洗）

第四步：对变量的统计可视化

清洗好的变量就可以进行**统计可视化**啦，实际上大家只需要关心两大类统计概念，即：

- 一：描述性统计，充分了解你的数据，分析数据的集中趋势和离散趋势等统计学指标并且可视化
- 二：推断统计学，根据样本数据去推断总体数量特征的方法。它是在对样本数据进行描述的基础上，对统计总体的未知数量特征做出以概率形式表述的推断。

数据总体来说可以分为以下三种类型：

- 一：分类数据，又名**定性数据**或者品质数据。
- 二：顺序数据。它其实是是**分类数据**的一种
- 三：数值型数据，又名**定量数据**，这个才是重点。又可以分成**离散型和连续型**

定量数据的**集中趋势**指标主要是：众数、分位数和平均数，**定量数据**的**离散趋势**指标主要是：极差，方差和标准差，标准分数，相对离散系数（变异系数），偏态系数与峰态系数

如果大家的英语还不错，就可以很容易看懂R里面的**简单统计函数**，无非就是 **mean,sd,mad,cv,max,min,median**等等。所以我针对性的出了30题

考核大家的统计学：30道练习题带你玩转统计学的R语言版

在R里面，基本上任何数据都可以进行可视化，简简单单的**几行绘图代码**就可以画一大堆的图,**plot,boxplot,barplot,pie,hist,pair**,它们每个绘图函数都有自己要求的输入数据，特定的可视化结果，请务必在还没熟练使用之前help一下它们，自己主动查看它们好玩的地方，好好自学，**下面我简单列出部分**，如果大家确实感兴趣可以在我的生信五周年演讲素材看到600行的基础绘图练习代码：
https://github.com/jmzeng1314/5years/blob/master/learn-R/tasks/2-chunjuan-600.R

```
dev.
```

```
new
```

```
()新建画板
```

```
plot()绘制点线图,条形图,散点图.
```

```
barplot( ) 绘制条形图
```

```
dotchart( ) 绘制点图
```

```
pie( )绘制饼图.
```

```
pair( )绘制散点图阵
```

```
boxplot( )绘制箱线图
```

```
hist( )绘制直方图
```

```
scatterplot3D( )绘制
```

```
3
```

```
D散点图.
```

低级绘图函数:

par() 可以添加很多参数来修改图形

title( ) 添加标题

axis( ) 调整刻度

rug( ) 添加轴密度

grid( ) 添加网格线

abline( ) 添加直线

lines( ) 添加曲线

text
( ) 添加标签

legend() 添加图例

上面提到的这些函数基本上都有一系列的绘图参数**(坐标轴、图例，颜色，性状，大小，空白，布局)**， 非常繁琐，想掌握，花费的时间会非常多，但是很多人直接跳到**ggplot的绘图世界**了，不想搞那么多底层绘图代码。但是我看过一个底层R绘图集大成者，就**Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes**文章的作者的github里面有。**但是对大部分人来说，生信的绘图，都是有套路的，其实都被别人包装成函数了，做好数据，一个函数就出了所有复杂的图。比如热图，cluster等等。**

至于高级可视化，就不得不提ggplot2了

- 如何通过Google来使用ggplot2可视化

如何通过Google来使用ggplot2可视化

因为ggplot2本身包含数据映射到图形元素的思想，不会适合所有人，而且讲解起来差不多也可以写一本书了，大家熟知的 http://www.cookbook-r.com/Graphs/ 就是这样的一本书，把它从头到尾代码练习一遍就差不多了。

可能对大部分初学者来说，ggpubr加上http://www.sthda.com/ 会更方便，基本上你想绘制的图都被高手写过教程，一搜索就可以获得。

同样的，我也针对性的出了30个考核题给到大家：基于R的可视化习题30个

第五步：数据对象的高级操作

普通数据（向量，数据框，数组，列表）的高级操作，主要是apply家族函数，以及**aggregate，merge，split，by** 等函数的用法。这是一个**分水岭**，用好了你就才可能是R入门了。也可以用一些包，比如reshape2，dplyr，可以做数据的高级操作。值得一提的是对大部分**编程思维不够**的朋友来说，可能需要有人讲解会方便入门。

数据对象的高级操作，主要是**S3，S4对象**，更多的是学习提出这个对象的R包的文档，比如我最近在**单细胞天地**集中介绍的5大单细胞流程R包：**scater,monocle,Seurat,scran,M3Drop** 需要熟练掌握它们的对象，：一些单细胞转录组R包的对象 而且分析流程也大同小异：

- step1：创建对象
- step2：质量控制

- step3: 表达量的标准化和归一化
- step4: 去除干扰因素(多个样本整合)
- step5: 判断重要的基因
- step6: 多种降维算法
- step7: 可视化降维结果
- step8: 多种聚类算法
- step9: 聚类后找每个细胞亚群的标志基因
- step10: 继续分类

所谓对象，就是嵌套了不同的基础数据，包括向量，数据框，数组，列表，打包在了一起，这样可以**统一操作，方便数据管理。**

关于循环，实际上这里需要一个动图来展现，在我的生信五周年演讲我都提到过，恰好在郑州演讲被粉丝录屏了，所以感兴趣的可以自行去听那20分钟的视频，录屏，素材和视频都在微云，
https://share.weiyun.com/5NRQO8t

第六步：无限R包或函数

前面我们提到过，掌握了变量和常量的概念，数据类型和数据结构的概念，我们就相当于是半只脚踏入R语言的大门了，然后加上这些数据的高级操作，就相当于入门了，可是R语言远不止那么简单，不同的人使用它的效率千差万别，比如因为可以自定义函数，所以我一行代码就可以完成3个R包的RNA-seq差异分析，见
https://mp.weixin.qq.com/s/a5oAjpmnz1Jyx1fX89G1kw 因为持续学习，所以我可以通过现成的R包来自动化完成临床三线表，见
https://mp.weixin.qq.com/s/N-ZIwz-IDG7p07A-NdAVJg

对于生信工程师来说呢，主要是遨游R的bioconductor世界，这也是为什么我要求搞生物信息学数据处理的人必须学习R，就是为了应用大量的bioconductor包。每学一个bioconductor的包，都是自己R水平的提升。大家可以参考我的博客：http://www.bio-info-trainee.com/tag/bioconductor 我就是这样学习过来的。我还创建了bioconductor中国这个社区，可惜效果不好，有志者可以继续联系我，我们看看有没有可能做起来。

R语言的应用方向当然R肯定不只生物信息学啦，其实它在**非常多的地方都有应用**，尤其是金融和地理。在如何一个方向学习R，就不仅仅是R本身的语法了，你需要学习的东西太多了，我简单列出几个我接触过的方向吧：**统计，科学计算，数据挖掘，文本挖掘，基础绘图，ggplot绘图，高级编程**，都有着丰富的书籍和视频资料。

学习资源推荐

部分优秀导航贴：

- R语言学习入门导航-特别版
- R语言入门学习路径+资源集(生信篇)
- R语言的最好资源，一个就够！
- 生信人的R语言书单（文末赠书
- 生信人应该这样学R语言系列视频学习心得笔记分享
- 如何通过Google来使用ggplot2可视化

一些电子书推荐：

```
《A Handbook
```

of

Statistical Analyses_Using_R》

《Modern Applied Statistics

With

S》

《Introduction

to

Scientific Programming

and

Simulation

Using

R》

《Mastering Scientific Computing

with

R》

《Practical Data Science

with

R》

《Data Mining explain

using

R》

《ggplot2 Elegant Graphics

```
for
  Data Analysis》

  《R Graphics Cookbook》

  《R Cookbook》

  《R
in
  a Nutshell》

  《R Programming
for
  Bioinformatics》

  《software
for
  data analysis programming
with
  R》
```

看完以上这些，你就是**R大神**了。当然，前提是你看懂了也会**灵活应用**。

有小伙伴建议我继续以送视频送书籍的方式来增加浏览量，比如我网盘里面有**几千本R语言的PDF书籍，也有十几套视频**，但是，我这一篇总结写的太好了，我不想被利益被污染了，**希望你可以转发给有需要的人，你的朋友会感激你的转发，让他结识了这么多生信前辈的经验分享公众号！**

写在最后

借鉴一个生物信息学习方法：**知识和耐心，是成为强者的唯一方法**。

- 通过阅读来学习：阅读经典的教材、代码、论文、公开课。
- 通过牛人来学习：同行的聚会（比如生信五周年）、讨论（QQ群或者微信群）、大牛的博客、微博、twitter、RSS。
- 通过练习来学习：代码练习题、参加编程比赛、解决实际工作中的难题。
- 通过分享来学习：自己写笔记、写博客、写书、翻译书，和同伴分享交流、培训新人

留给大家一些作业

生信基石之R语言，**B站的10个小时教学视频**务必看完，参考 GitHub 仓库存放的相关学习路线指导资料：
https://github.com/jmzeng1314/R_bilibili

- 初级10 个题目：http://www.bio-info-trainee.com/3793.html
- 中级要求是：http://www.bio-info-trainee.com/3750.html
- 高级要求是完成20题：http://www.bio-info-trainee.com/3415.html
- 统计专题 30题：http://www.bio-info-trainee.com/4385.html
- 可视化专题30题：http://www.bio-info-trainee.com/4387.html

那么，点击阅读原文就值得我的B站10小时R语言教学视频吧！