最基本的数据类型包括数字和字符串，数字包括整数、小数等，字符串是任何字符的组合，由单引号或双引号包括。它们是其它数据类型的基本组成部分，比如逻辑符号本质就是数字。科学计数法：1e-3 1e2，1L表示1这个整数

2 ^ 6 阶乘

5 %% 2 取余

数学函数log, log2, log10 exp(n)-e的n次方 sin, cos, tan sqrt

|或 &且 ≠0的数都是T

赋值：x<- 1 x = 1 1 -> x assign( "x", 1)

ls()显示当前环境下所有变量 rm( x )删除一个变量 rm(list=ls())删除当前环境下所有变量

特殊值：NULL：值未定义的表达式和函数；Inf：无限值（数字），分为+inf和-inf；NA：不可得到的值，不是有限值；NaN：不是数字（不存在）,属于NA的一种

c(-1, 0, 1) / 0 -Inf NaN Inf

判定特别值：is.finite()是不是有限值 is.infinite() is.na() is.nan() is.null()

判断数据类型：is.numeric() is.list()等，isTRUE函数只在参数是一个逻辑运算值true时才返回true，若传入的不是一个逻辑运算值，而是一个数字/字符等就都返回false

(x)中括号作用：如果不加括号，就不会输出x的值。括号将(x)整体作为一个表达式，这个表达式的值等于x的值，因此这一行实际上就是一个值，执行时会直接输出这个值

Vector类型转换优先级：逻辑类型->数字类型 逻辑类型->字符串 数字类型->字符串

矩阵内的数值必须是同一种类型，输入混合有多种基本数据类型时，矩阵会按上面提到的规则进行强制转换

ab <- c(a,b)合并ab数组 length()-长度 rev()-将数组倒序

矩阵matrix( c(20, 30.1, 2, 45.8, 23, 14), nrow = 2, byrow = T, dimnames = list( c("row\_A", "row\_B"), c("A", "B", "C") ) ) 矩阵的指定长度：nrow行 × ncol列。矩阵长度较小时，输入数据会被截短；而矩阵长度较大时，输入数据则会被重复使用，byrow-是否按行将数据转化为matrix，dimnames指定矩阵的行名和列名

vector和矩阵都是数组。vector是一维数组，矩阵是二维数组。还可以有更多维的数组，高维数组与vector和矩阵一样，只能包含一种基本数据类型。高维数组可以由函数 array() 定array(data=LETTERS[1:16],dim=c(2,4,2),dimnames=list(c("A","B"), c("one","two","three","four"), c("一", "二") ))

m[1:2, ] 按行数取行 把1:2放逗号后面就是取列，结果都是vector

m[c( "row\_B", "row\_A" ), ] 按行名取行

m[1:2, 2:3] 取其中一部分

m[ 1:2 , ] <- matrix( 1:6, nrow = 2 ) 替换

a[is.na(a)] <- 0 将a中所有的NA替换为0

常用基础函数：

which()-遍历传入的vector，当vector中某元素满足判断条件时，which函数返回其索引

判断数据类型：str（）、class（）、typeof（）、mode();其中class、typeof显示整体数据类型，str显示变量内部每个元素类。mode函数只查看数据的大类，class函数查看数据的类，typeof函数则更细化，查看数据的细类

抽样：sample(x,size,replace = FALSE) x:数据size:抽样数replace:F，即抽样数size不能大于x长度

排序：sort()-输出排序后新序列 order()-输出新序列的每个元素在原序列中的位置

数据计算：range()-输出最大值和最小值 summary()-min,max,median,mean,1、3等分点 t()取二维矩阵的转置

改变数据类型：names( ab) <- as.character( ab) names把某个参数/行/列的名称改变，as.character将数字对象转为字符串类型

seq(0,10,along.with=c(1,10,2)), along.with=Id就是说seq生成数据序列的长度与Id序列长度相同，输出1 5 10

unique()-返回一个没有任何重复元素或行的向量、数据框或数组

subset(a,logical expression)根据逻辑表达式筛选a中的数据，返回令表达式为T的a中数据

vectorisation：R最重要的一个概念，核心在于数据自动循环使用

ab[3:6] <- c("2","3") 遍历赋值

c(1,2)\*c(2,3)就是c(1\*2,2\*3)

x<- c(4,6,5,7,10,9,4,15) ;x < c( 8, 4 )- x中的元素依次与8 4 8 4 8 4 8 4进行比较

data.frame与tibble: 二维表格，由不同列组成；每列是一个vector，不同列的数据类型可以不同，但一列只包括一种数据类型（int, num, chr …）各列的长度相同

data.frame( data = sample( 1:100, 10 ), group = sample( LETTERS[1:3], 10, replace = TRUE), data2 = 0.1 )创建3列，分别为data、group、data2

tribble(

~x, ~y, ~z, #~表示列名

"a", 2, 3.6,

"b", 1, 8.5)

tb<- tibble( x = character(), y = integer(), z = double() )规定这一列的类型

tb <- add\_row( .before=2)增加行, 在第二行之前插入 add\_column增加列

相互转换：as\_tibble()和as.data.frame

区别：

（1）tibble可以按顺序计算列值（将前一列的值用于创建下一列的值）tibble(x = 1:5, y = x ^ 2)，而data.frame不行

（2）tibble取一行/一列时(t[1,])返回的仍是一个tibble，而data.frame取一列(df[1,])时返回的是一个vector

（3）tibble可以进行可控的数据类型转换（使用[[]]或$时返回一个vector）

（4）tibble在根据列名取某列时不会部分匹配，更准确（data.frame(abc = 1)$ab不会报错，而tibble会报错）

（5）构造自动循环补全：tibble-仅限于长度为1或等长,而data.frame则为整除（data.frame(a = 1:6, b = LETTERS[1:2])可以），而tibble不行，只能tibble(a = 1, b = 1:3)

常用函数：

head(n=),tail(n=)显示前/后n行数据

nrow()行数 ncol()列数 dim()维度 rownames()取出所有行名

str()表格结构：每列都是什么数据类型、有多少数据、数据值

rbind()：根据行进行合并 cbind():增加列 如果一个df的行数少于另一处，就循环调用行数不足的vector

bind\_rows()：合并tibble merge():默认按照相同的列名合并

tibble::enframe: 将vector或list转换为一列或两列的tibble。如enframe(c(1,2,3))，结果是两列的tibble，列名分别为name和value，列值都是(1,2,3)

res$Total=rowSums(res[,col\_list])求指定列的和，同理有rowMeans()

attach(iris)：之后可以直接用sepal直接代指iris$sepal detach(iris)取消该操作，with(iris, head(Sepal.Length)也可

within(airquality,{rm(Day)})：删除airquality中Day列

knitr::kable()：以更清晰的形式展示tibble

readr包：read\_csv()逗号分隔符文件

read\_tsv(): tab separated files用tab分隔的文件

read\_delim(): general delimited files

read\_fwf(): fixed width files

read\_table(): tabular files where columns are separated by white-space.读取txt文件

read\_log(): web log files

使用：

read\_csv("data/talk03/iris.csv", col\_types = cols(Sepal.Length = col\_double(),Species = col\_character()))col\_character()指定读取的列类型, col\_names = FALSE不读取文件中的列名, col\_names = col\_name指定列名,n\_max = 7是从开始读取往后7行, quote = FALSE使输出的文本string两边没有引号, skip跳过前n行读取9

write\_tsv(iris, "iris.tsv", quote\_escape = "none")将iris写入"iris.tsv"中，同理有write\_csv、write\_delim、write\_excel\_csv、write\_file、write\_lines、write\_rds、write\_tsv

系统自带读取函数read.table、read.csv: header = TRUE显示列名, skip跳过前n行读取，quote消除string数据前后的字符, na.strings = NA让空缺数据显示为NA,col.names = col\_name指定列名,nrows = 7是从开始读取往后7行,sep=”;”让读取时按分号捕获数据

R session工作空间：save.image(file = "xx.RData")保存 load(file = "xx.RData")载入 save(city, country, file="1.RData")保存指定变量

向量：as.factor()，ordered = T参数时factors排序。Factors会限制输入数据的选择范围：新加入的值必须是levels里已存在的值，levels不可更改，若更改需调用函数levels(x) <- c(levels(x), "widowed”), levels的顺序决定了排序的顺序。用as.factor创建factor时，得到的levels按字母表排列；但是，用levels(y)方式指定levels时，则按照指定的顺序，sort函数也不会改变已指定的levels顺序

reorder(a,b,func)就是根据b对a:b这个键值对进行排序，排序方法为func（可为median等）

forcars包：

fct\_c：合并两个factor

fct\_inorder：按照第一次出现的次序排列

fct\_infreq：按照出现的频率排列（从大到小）

fct\_inseq：按照数字大小排列（从小到大）

fct\_count( gender );#统计gender里的levels数量

fct\_collapse(gender,Male=c("m", "male "))把gender中c("m", "male ")这些levels都改成Male，或使用fct\_relabel

magrittr包提供管道操作，导入tidyverse包时会自动导入该包，%>%在使用时放在两个操作之间，前一个操作产生的数据将作为后面操作的输入数据，此外还有%$%等价于attach、%<>%代表双向数据传递等其他形式的pipe、%T>%返回上游的值，所有函数都支持pipe。通常需要用 . 指代传递来的数据，并以参数的形式赋予下游函数。

res <- matrix(ncol = 2) %T>% plot() 此时res是matrix

mtcars %$% cor.test( cyl, mpg ) 等效于with( mtcars, cor.test( cyl, mpg ) )

#a %in% b 判断a是否在b中，若在就返回true

dplyr包：

select() 选择列，根据列名规则，select(CHR = `Chromosome/scaffold name`)

filter() 按规则过滤行，filter( `Transcript type` %in% c( "protein\_coding", "miRNA", "lincRNA" ) )

mutate() 增加新列，从其它列计算而得（不改变行数）

summarise() 将多个值转换为单个值（通过mean, median, sd等操作），生成新列（总行数减少，通常与 group\_by配合使用 ），summarise( count = n\_distinct( GENE\_ID ))

arrange() 对行进行排序，arrange(-grade) 按成绩由高到低排序

group\_by() 分组，group\_by( CHR, TYPE )或group\_by( CHR )

gather( course, grade, -Name, -Occupation, na.rm = T ) 除了name和occupation列，都删去，转成course和grade列，其中第一个参数即course做键（接收原列的列名），第二个参数即grade做值（接收原列的列值）

mutate() 创建新列，mutate(bmi= mass/((height/100)^2))

行操作：

distinct()：删除重复行

sample\_frac(a,0.5)：随机选出50%的行

sample\_n(a,10)：随机选出10行

slice(a,1:5)：切片，选1-5行

top\_n(a,n)：选择并排序前n个条目

tidyr包：

宽数据（列多）--优点：自然，易理解；缺点：不易处理，稀疏时问题较大

pivot\_longer() 代替 gather，pivot\_longer( - name, names\_to = "course", values\_to = "grade"，values\_drop\_na = T)第一个参数要改变的列（- name表示除了name的剩下列），第二个参数是把原列名变成一个新列的列名，第三个参数是原列对应的值变成的新列的列名，values\_drop\_na可消除值为NA的行；或pivot\_longer(!c( name, class ),..)，!加上一个vector，元素是想保留的列，!相当于前面的-name中减号的作用

pivot\_wider() 代替 spread，pivot\_wider( names\_from = "course", values\_from = "grade" )列名来自哪列，列的值来自哪列

complete( name, course )：用于把隐式的缺失值转换成正常显示的缺失值。把grades按name和course分组，因为有3种name和5种course，分组后应有15行，而原数据只有9行；所以按此分组，原数据存在隐式的缺失值，complete可以将隐式缺失显示出来，使结果有15行，多出来的6行的grade列用NA代替

PPI network：使用igraph包，除此之外还有ggnet、interactive networkD3 R package、plotly.js、D3等

toppart <- ppi %>%

filter( gene1 == "KIF23" ) %>% arrange( desc( score ) ) %>%

slice( 1:10 )#取前10行（与KIF23最有关的10个基因），要不数据太多

genes <- unique(c("KIF23", toppart$gene2)) #得到"KIF23"与gene2的并集genes

netdata <- ppi %>% filter( gene1 %in% genes & gene2 %in% genes )#挑选出gene1和gene2列都在genes里的行（11个基因的相互联系）.因为是网络图，不能只是KIF23与其它基因的关系，还有其它基因之间的联系，所以要再在ppi总集里取10个基因

netdata.nr <- netdata %>% mutate( group =

if\_else( gene1 > gene2, #相当于if(gene1 > gene2) {str\_c( gene1, gene2, sep = "-" )} else {str\_c( gene2, gene1, sep = "-" ) )}

str\_c( gene1, gene2, sep = "-" ),str\_c( gene2, gene1, sep = "-" ) ) ) %>%

group\_by(group)%>%slice(1) #按group排完序之后，每组有2个元素，slice（1）只取第一个元素

netnet.nr <- graph\_from\_data\_frame( netdata.nr, directed = FALSE ); plot(netnet.nr)

计算每个基因的连接数： netdata %>% group\_by( gene1 ) %>% summarise( degree = n() ) #group\_by( gene1 )后每个基因都是一组

宏基因组：对直接从环境样本中回收的遗传物质的研究

排完序后将后面的丰度累加在一起，汇总为"Others"分类：使用tidytidbits:: lump\_rows( scientific\_name, relative\_abundance, n = 10, other\_level = "Others" ) ，(relative\_abundance, n = 10)保留前10行的相对丰度，并把剩下列的相对丰度加和，(scientific\_name,other\_level = "Others")把加和结果的scientific\_name列命名为Others

把 Others 和 Unknown 放在最后：fct\_relevel( fct\_reorder( abu.dat$scientific\_name, abu.dat$relative\_abundance, .desc = F), "Unknown", "Others" )，fct\_reorde的结果中虽然有"Unknown"和"Others"，但fct\_relevel会优先读取"Unknown", "Others"这两个参数放到scientific\_name列的最前面并锁定位置，之后再把reorder的结果排列

数据的整合：

使用 pivot\_wider：先合并，再pivot\_wider用bind\_rows合并两个tibble时，列名需要一致

expr %>% mutate( group = "rkpm" ) %>% #expr数据的group都为rkpm

select( gene, group, value=rkpm ) %>%

group\_by( gene ) %>% slice( 1 );#如果一个基因出现多次，取第一个为准

bind\_rows( meth2, expr2 )

也可使用left\_join(x,y):返回x中的所有行，以及x和y中的所有列。x中没有匹配项的行在新列中将具有NA值。如果x和y之间有多个匹配项，则返回所有匹配项的组合。

meth %>% pivot\_wider( names\_from = "site", values\_from = "methylation\_score" )

left\_join( meth3, expr2, by = "gene" ) #按gene列进行合并，结果是增加列

list相关内容：

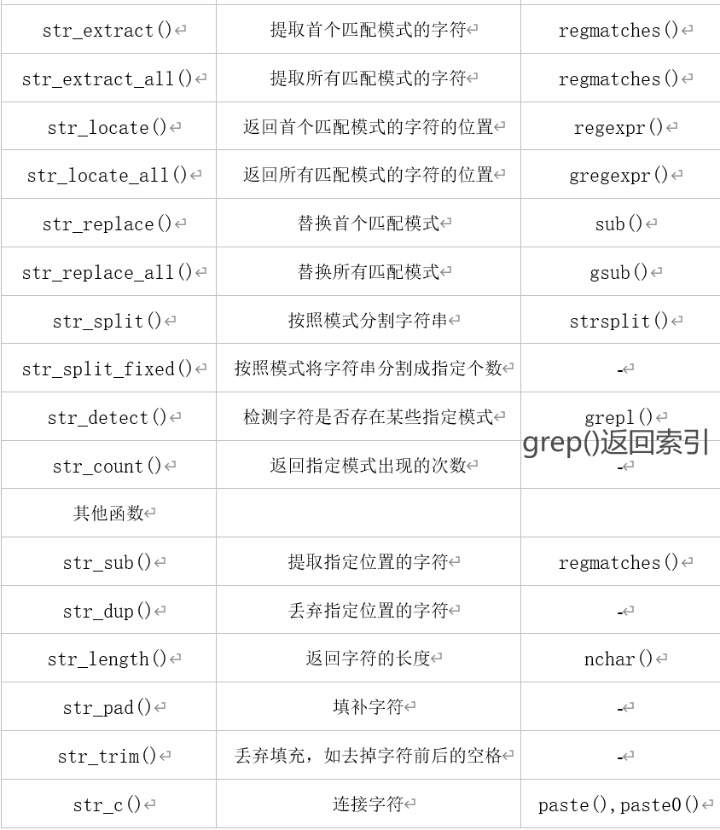
list[[1]]是取该list的第一个元素，可以是vector，若是[1]会返回一个list

data中films列的元素是一个list，该list里面的元素是一个chr型的vector，该列在rmd中输出的是<chr [5]>的形式，[5]表示该chr型vector中有5个元素。在rstudio中此句输出结果与上面tibble的相同，但生成的PDF中输出是里面包含的电影名称，因此data.frame不能实现

stringr和stringi

字符串中要想打单引号，需在前面加上\，即\'输出为‘；字符串中若换行，输出时会用\n替代换行，若使用writeLines()函数，就把字符串在换行处切开为多个字串进行输出（会换行）

"\""输出"\""，'\''输出"'"，"\\"输出"\\"，若用writeLines()函数，会输出转义后形式，且两端无引号



str\_subset 则是返回包含子字符串的字符串

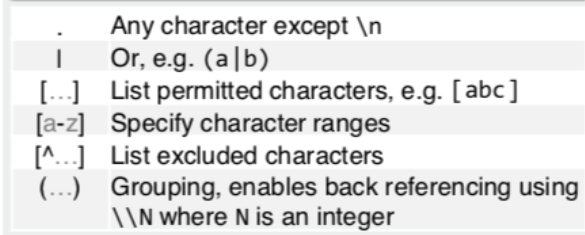
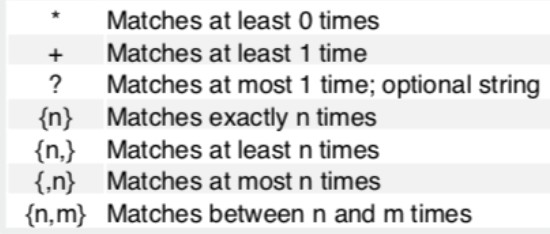
str\_extract 会返回匹配的字符串

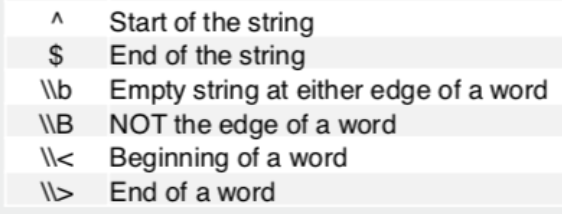
str\_match 抽取匹配的字串，第一列是整体匹配结果，后面是每一组(正则)匹配结果。

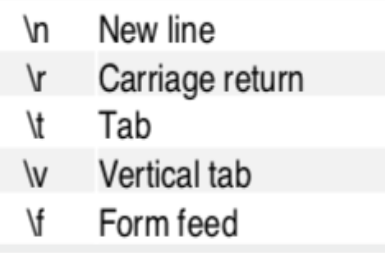
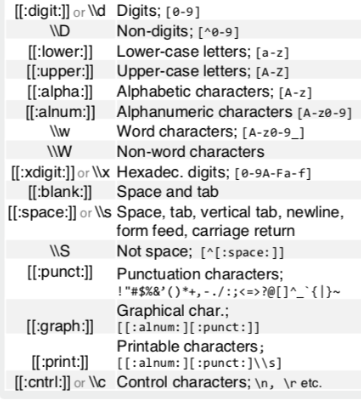
字符串比较：系统-直接"A" > "abc" pracma-strcmp()和strcmpi()，i是忽略大小写进行比较

toupper()小写变大写 tolower()大写变小写（系统自带）

倒序输出：stringi-stri\_reverse字符规则







stringr的其它函数

fixed(): match exact bytes

coll(): match human letters

boundary(): match boundaries

str\_to\_title()将首字母大写

统计元音个数：str\_count(x, "[aeiou]")

反向互补序列：seq %>% str\_replace\_all("A","t") %>% str\_replace\_all("T","a") %>%

str\_replace\_all("C","g") %>% str\_replace\_all("G","c") %>% toupper() %>% stri\_reverse()

填充字符串：str\_pad(处理类型,width,side=c("left","right","both"),pad="")，处理类型:字符串/向量 width:字符串填充后的宽度 side:从哪边开始填充，默认从左边 pad:用什么字符串填充

提取字符串中的大于等于2次的重复字段：str\_extract(x,"(\\w{2}){2,}")

有理数识别（123、-1.5等）： "^([\\-|\\+)?\\d+(\\.\\d+)?$](file:///\\-|\\+)%3f\\d+(\\.\\d+)%3f$)"，

^(\-|\+)?中\-和\+分别代表-和+，|是“或”，?表示-和+可有可无，整体表示以1个-或+或什么也不加开头（\\为转义符号，实际上是\），\d+表示至少要有1个数字，(\.\d+)?中\.表示小数点，\d+表示小数点后至少要有1个数字，?表示小数点和小数点后数字这部分可有可无