基因基于 R 语言在染色体位置作图

李旭凯

山西农业大学生命科学学院，山西太谷 030801

摘 要:R 语言是一种由统计学家开发的统计计算和绘图的语言和环境。基因家族分析是目前生信分析的热点，其中基因的染色体分布是基因家族分析首先要分析的内容。Gbrowse、MapInspect、MapViewer 等软件虽然能绘制相似形式的基因位置图，但使用时数据需提前设置好，输出结果无法灵活修改，有很大的局限性。鉴于此，文章基于 R 语言，开发了一款生物辅助作图软件 Genes_on_Chr 的本地版，该软件能够依据输入数据快速绘制相应的基因在染色体分布图。该软件输入数据格式简单，输出结果易于修改，图片格式为 PDF 矢量图，具有很好的移植性，为研究人员提供便利。软件下载及操作见网址：https://github.com/xukaili/Genes_on_Chr

关键词:R 语言；基因分布图；作图；生物软件；可视化

A tool to draw genes on chromosome based on R language

LI Xukai

College of Life Sciences，Shanxi Agricultural University，Taigu 030801，China

Abstract: R language is developed by a statistician with the capability of data analysis and visualization. Genes on chromosome map is helpful for analysis on gene families. Although some tools like GBrowse, MapInspect and MapViewer could draw genes on chromosome map, there are limitations for them: (1) the data need to be decorated in advance; (2) user can't modify results. Therefore, we developed a mapping software: Genes_on_Chr with PC, which is based on R languages. The software can be used to draw the corresponding genes on chromosome map quickly in PDF format based on the input data. It will become a useful tool for drawing genes on chromosome map with the advantages of simple input data format, easily modified output and very good portability. The software download and operation location can be found at the website: https://github.com/xukaili/Genes_on_Chr

Key words: R language; gene distribution map; draw map; bio-software; visualization

数据分析的关键一步是对数据的可视化。R 软件是自由、免费、开源的软件，具有统计分析功能和绘图功能，其简单明了的命令让用户易于理解和操作，同时可编程的函数语言环境也为需要个性化定义的用户提供了极大便利，现已是国内外众多统计学者、信息学者、可视化学者喜爱的科学数据分析的强大工具[1]。

在绘制遗传图谱方面，已有多个软件利用分子标记之间的遗传距离绘制遗传图谱，例如 MAPMAKER[2]，JoinMap[3~5]，Mapplotter[6]，MapDraw[7]和 MapGene2Chrom[8]。但是在基因在染色体位置图绘制方面，虽然一些模式物种可以借助于现有公共数据库，如 TAIR 中的 Chromosome Map Tool （http://www.arabidopsis.org/jsp/ChromosomeMap/tool.jsp）、Oryzabase 中的 Chromosome Map Tool （http://viewer.shigen.info/oryzavw/maptool/MapTool.do）等可以绘制模式生物基因在染色体位置图。而绝大多数物种并没有这样方便的工具，很多研究人员无奈只能使用 PPT、Photoshop 等软件手工做图，大大限制了科研效率和结果精确性。鉴于此，本文基于 R 语言，开发了一款辅助软件 Genes_on_Chr 的本地版,输入指定格式的数据，即可快速绘制一张简洁、美观的基因在染色体位置图。

1 前沿

1.1 Genes_on_Chr 的原理与特点

Genes_on_Chr 基于 R 语言，用户需要安装 R 语言基本环境，并且无需安装额外的 R 包，避免了有些 R 版本 R 包无法安装的问题。根据染色体长度和基因的位置，可以快速画出不同染色体上基因的分布图。基因的位置信息通过常用的记事本或 Microsoft Excel 办公软件，进行简单处理即可作为输入文件。之后，运行软件，用户将得到一张基因分布 PDF 矢量图。

1.2 Genes_on_Chr 本地版使用方法

以拟南芥物种为例，从 Ensembl 数据库（ftp://ftp.ensemblgenomes.org/pub/）下载拟南芥（*Arabidopsis thaliana*）的基因注释数据（ftp://ftp.ensemblgenomes.org/pub/plants/release-37/gff3/arabidopsis_thaliana/），并随机选取基因数据作为测试数据。本文以 Windows 操作系统为例，分别从运行环境、数据准备、运行软件、相关参数等方面介绍其使用方法。

2 运行

2.1 运行环境

R 是一种自由软件编程语言与操作环境，主要用于统计分析、绘图和数据挖掘，属于 GNU 系统的一个自由、免费、源代码开放的软件，提供了大量的数据处理、统计和图形函数，可以在 Windows、Mac OS、Linux 等各大操作系统下免费使用[9]。R 语言是一种区分大小写的解释性语言，其具有的强大统计计算及绘图能力，是从大数据中获取有用信息的绝佳工具，R 软件在基本安装中提供了大量的数据处理、统计和图形函数，此外各社区还开发了数以千计的扩展包（packages）为 R 增加了更多令人惊叹的功能[10]。

Windows 系统下 R 软件的下载与安装。可在 R 的 CRAN 社区网页 http://cran.r-project.org/网站下载 R。对于 Windows 用户，单击 Download R for Windows 进入下载窗口，按照 Windows 的提示安装即可。

2.2 数据准备

绘制基因在染色体上的分布图，需要收集基因的相关信息有：染色体的数目、长度及各染色体上的基因数据，基因数据包括基因所在染色体名称、基因在染色体上的位置及相应的基因名称，将信息汇总至染色体文本文件 chr.txt 和基因文本文件 gene.txt 中；chr.txt 文件包括：各染色体名称及长度，gene.txt 文件包括：基因所在染色体名称、基因起始位置、基因名称。具体可以从 Ensembl 数据库下载基因注释数据。

chr.txt 数据的结构为：第一列为染色体编号，第二列是染色体总长(bp)。

1	Length_1
2	Length_2
3	Length_3
…	…

gene.txt 数据的结构为：第一列为染色体编号，第二列是基因的位置，第三列是基因名称。

1	site_1	Gene_name_1
1	site_2	Gene_name_2
2	site_3	Gene_name_3
...	…	...

本文通过自编写的 R 程序 Genes_on_Chr（其源码见本文）根据输入文件从每条染色体中获得基因数据样本进行绘图；为了方便并快速的从原始文件中提取用于作图的基因数据，用户可以使用 perl 语言，通过 perl 语言对下载的原始文件中提取染色体长度、基因数据信息，依次存入文本文件

chr.txt 和 gene.txt 文件中，作为 R 程序 Genes_on_Chr 的输入文件。提取染色体长度、基因数据信息的 perl 语言代码如下：

1. 提取染色体长度

   用法： perl -e '$/ = ">"; while (<>) {($head,$seq) = split(/\n/,$_,2); $len = length $seq; $head =~ /(\d+)/; print "$1\t$len\n";}' genome.fa > chr.txt

2. 提取基因数据

   用法：perl Search.pl list.txt GFF3.gff > gene.txt

   list.txt 文件内容为选定的用于作图的基因名称。

2.3 运行软件

本地版软件在运行前需要做如下操作：

第一步：将 2.2 中准备好的数据放在与 Genes_on_Chr.R 同一目录下。

第二步：打开 DOS 命令窗口（桌面左下角开始按钮→输入框中输入 cmd 即可打开 DOS 命令窗口），在 DOS 命令窗口输入“cd 文件所在路径”，即可在该文件目录下进行操作。

第三步：在 DOS 命令窗口输入“Rscript Genes_on_Chr.R chr.txt gene.txt”即可获得相关信息并得到结果文件 Gene_on_Chr_plot.pdf 的 PDF 图，存储在相同的目录下以便查看。结果如图 1、图 2 所示。

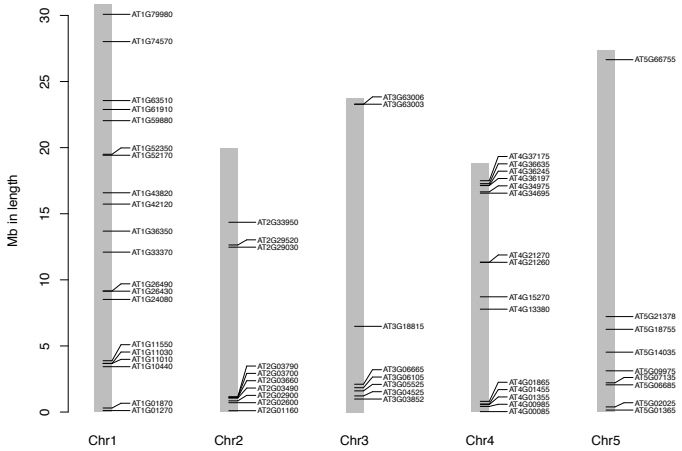

图 1 拟南芥的基因在染色体分布图

Fig.1 *Arabidopsis* genes in the chromosome map

注：图中横坐标代表各条染色体；纵坐标代表染色体对应的长度，单位为 Mb。

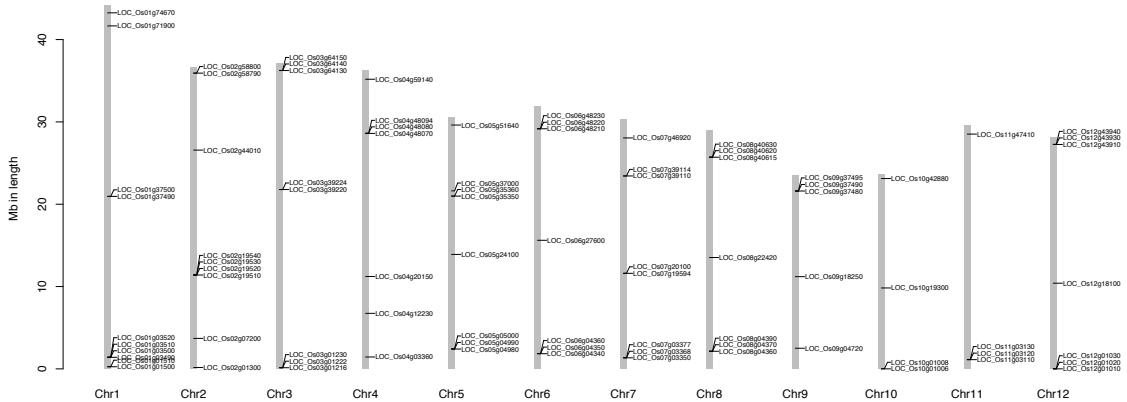

图 2 水稻的基因在染色体分布图

Fig.2 Rice genes in the chromosome map

2.4 Genes_on_Chr.R 的源代码

可将 Genes_on_Chr.R 的源代码复制到记事本中，命名(如：Genes_on_Chr.txt)并保存，然后再重新命名为 Genes_on_Chr.R，继而按 2.2 的操作流程运行使用。

2.5 参数说明

绘图效果的参数有标题的字体、字体大小及颜色;基因名称的字体、字体大小及颜色；染色体边框线宽度、颜色；连接线的宽度、颜色等。可通过在程序中修改实现，表 1 为代码中的部分参数。

在 R 语言里用命令 barplot()函数绘制条形图：

```
barplot(x, width=1, space=NULL,
        names.arg=NULL,legend.text=NULL,beside=FALSE,
    horiz=FALSE,density=NULL,angle=45,
    col=NULL,border=par(``fg``),
    main=NULL,sub=NULL,xlab=NULL,ylab=NULL,
    xlim=NULL,ylim=NULL,xpd=TRUE,log=````,
    axes=TRUE,axisname=TRUE,
    inside=TRUE,plot=TRUE,axis.lty=0,offset=0,
    add=FALSE,args.legend=NULL,...)
```

其主要参数解释如下:

angle:设置底纹的斜率。

xlim 和 ylim:设置柱形的 x 轴与 y 轴的分布范围。

xlab 和 ylab:设置 x 轴与 y 轴的标签。

axes:设置柱形是否显示 x 轴或 y 轴。

plot:设置是否显示柱形图。

col:设置条形底纹或者填充什么颜色。

border:设置柱形边缘颜色。

width:设置柱形的宽度。

space:描述条形之间空白的宽度。

axis.lty:设置图形 x 轴的类型。

names.arg:设置柱形的标签(bar labels)。

表 1：参数说明

Table1 Parameter description

参数名称	默认参数	参数说明
h1	0.5	连接线的第一条线段的长度
h3	0.5	连接线的第三条线段的长度
h	1.5	连接线的三条线段的长度之和，确定第二条线段的长度
hh1	1.5	第一条线段在染色体上的起始位置
cex	0.7	基因名称的字号
dec	1.8*max(size)/100	当前位点名称与上一位点名称之间相对距离的最小值

3 Genes_on_Chr 核心算法

Genes_on_Chr 软件绘制基因在染色体上的分布图的大致流程为:

(1)读取输入文件，分析共有几条染色体、每条染色体上的基因数；

(2)读取每条染色体的长度信息，根据各染色体中的最大值确定绘图的纵轴单位；

(3)确定基因名称的位置，根据两个相邻位点的位置差与两位点之间相对距离的最小差判断基因

名称是否重叠。若不重叠，则基因名称位置不变；若重叠，则为基因名称分配新的位置；

(4)绘制标注线，根据基因位点、基因位点名称的最终位置绘制位点和名称之间的连接线。

软件下载及操作见网址：https://github.com/xukaili/Genes_on_Chr

4 讨 论

本文根据生物研究的需要，基于 R 语言设计了一款基因在染色体上作图的生物辅助软件，根据输入数据快速绘制相应的基因在染色体上的分布图，实现了基因位置的可视化，利用 R 语言的免费开源、数据分析和绘图功能强大等特点设计一套软件分析工具为生物研究人员提供了便利。R 语言具有多方面优异的特性，在绘图和数据处理方面具有广泛的应用前景。R 软件是一款集成了数据操作、统计和可视化功能的优秀开源软件，分析人员可利用简单的 R 程序语言描述处理过程。

对某物种的基因家族分析包括基因结构、染色体分布、蛋白结构域、系统进化关系、基因表达谱等，染色体分布是基因家族分析首要的分析内容，利用绘图软件将家族基因定位在染色体上，可以清晰直观地看出该家族基因在该物种各染色体上分布的多少。该研究的生物作图软件 Genes_on_Chr 为绘制基因在染色体上的分布提供了帮助，为家族基因的功能研究及生物信息学分析提供了有用参考信息。

Genes_on_Chr 本地版软件的优点是数据量不受限制，本地版软件在 DOS 命令窗口使用时如果计算机已安装 R 软件但得到的反馈信息为：“不是内部或外部命令，也不是可运行的程序或批处理文件”，这时需要对计算机设置环境变量，具体操作如下：控制面板→系统和安全→系统→更改设置→高级→环境变量→新建系统变量，将安装 R 软件所在的路径设为变量值。操作完成后即可正常运行软件。

Genes_on_Chr 输出结果为 PDF 矢量图，便于研究人员使用。该软件尤其是为大多非模式生物绘制基因在染色体的位置图提供了方便的工具，该项工作解决了以往绘制绝大多数物种的基因在染色体分布图的困难，提高了研究人员的科研效率。

参 考 文 献

[1] 王怀亮.基于 R 语言的统计数据柱形图的实现.电子设计技术与应用[J]. 2013,8: 72-74.

[2] Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations [J]. *Genomics*, 1987, 1(2): 174-181.

[3] VAN Ooijen JW. Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species [J]. Genet Res, 2011, 93(5): 343-349.

[4] Stam P. Construction of integrated genetic linkage maps by means of a new computer package: Join Map [J]. Plant J, 1993, 3(5): 739-744.

[5] Stam P. JoinMap 2. 0 deals with all types of plant mapping populations [M]. In: Plant Genome III Abstracts. San Diego, USA, 1995.

[6] 沈利爽，郑先武，朱立煌. Mapplotter——个输出遗传图谱、图示基因型和 QTL 曲线图形的软件[J]. 遗传，2000，22(3): 172-174.

[7] 刘仁虎，孟金陵. MapDraw, 在 Excel 中绘制遗传连锁图的宏[J]. 遗传，2003，25(3): 317-321.

[8] 晁江涛，孔英珍，王倩，等. MapGene2Chrom 基于 Perl 和 SVG 语言绘因物理图谱[J]. 遗传，2015，37(1): 91-97.

[9] Kabacoff RI.R 语言实战[M]. 高涛，肖楠，陈钢译. 北京:人民邮电出版社, 2013.

[10] 袁佳.R 语言及 ggplot2 在环境空气监测数据可视化中的应用[J]. 中国高新技术企业.2015,16:89-91.