

# Introduction to statistical analysis with R software for cancer scientists

Expert: **Prof Livio Fenga**, Senior Lecturer in Analytics, University of Exeter, Exeter, UK

Discussant: **Dr Luca Bertolaccini**, IEO European Institute of Oncology IRCCS, Milan, Italy

## Extract from the e-ESO policy

The website contains presentations aimed at providing new knowledge and competences, and is intended as an informational and educational tool mainly designed for oncology professionals and other physicians interested in oncology.

These materials remain property of the authors or ESO respectively.

ESO is not responsible for any injury and/or damage to persons or property as a matter of a products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material published in these presentations. Because of the rapid advances in medical sciences, we recommend that independent verification of diagnoses and drugs dosages should be made. Furthermore, patients and the general public visiting the website should always seek professional medical advice.

Finally, please note that ESO does not endorse any opinions expressed in the presentations.

# **Introduction to statistical analysis with R software for Cancer Scientists**

Livio Fenga: Senior Lecturer in Analytics, University of Exeter (UK)

Webinar e-ESO, June 16, 2022

# Outline

1. R overview
2. Mastering R...no thanks. Let's use it
3. *Changing paradigm: From plots to LATTICES*  
(Tumor growth is inherently a three-dimensional process).
5. How can R help us (PRACTICALLY)
  - 5A **Neural Network with cancer data with R**
  - 5B **R and the power of simulation for cancer data**  
(Creating and visualizing spatial simulations of tumor growth)

## R overview but, before we start, some Q&As

As an MD *and* Researcher do I need to know R?

As an MD *and* Researcher can I learn R?

**How?**

Tutorials AND other *ad hoc* resources;  
learning by doing (COPY AND PASTE!)  
interdisciplinary links among different depts.;

As an MD *and* Researcher should I know Statistics?

YES  
YES

Being a good practitioner is usually enough  
Again interdisciplinary links  
(statisticians,  
mathematicians, physicists)  
are vital

# Download R and RStudio

- Download R :

<http://cran.r-project.org/bin/>

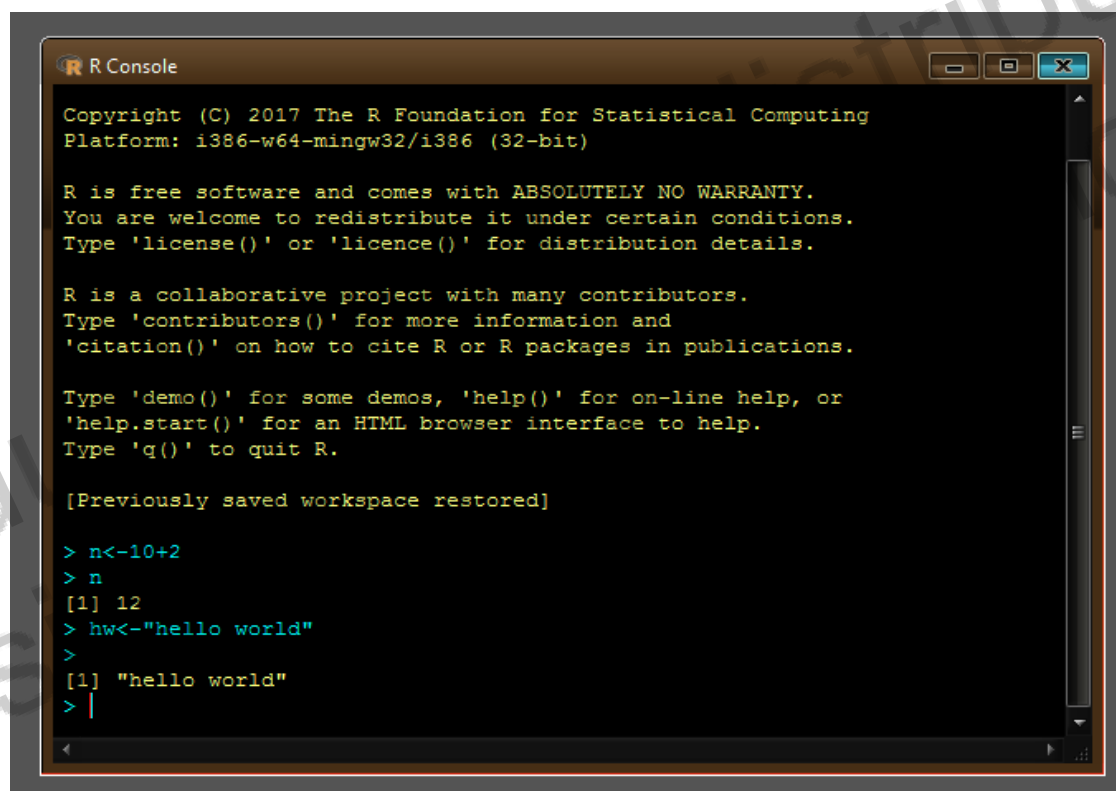
# Installation

## Installing R on windows PC :

- Use internet browser to point to : <http://mirror.aarnet.edu.au/pub/CRAN>
- Under the heading Precompiled Binary Distributions, choose the link Windows.
- Next heading is R for Windows; choose the link base.
- Click on download option(R xxxxx for windows).
- Save this to the folder C:\R on your PC.
- When downloading is complete, close or minimize the Internet browser.
- Double click on R 3.4.1-win32.exe in C:\R to install.

# R is Relatively Easy

- Double click the R icon on the Desktop and the R Console will open.
- Wait while the program loads. You observe something like this.



```
R Console

Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> n<-10+2
> n
[1] 12
> hw<-"hello world"
>
[1] "hello world"
> |
```

- You can type your own program at the prompt line >.

# R command in integrated environment

The screenshot displays the RStudio interface with the following components:

- Script Editor:** Contains an R script with the following code:

```
1 1+1
2 x=c(1,2,3,4)
3 x
4 y=c(3,4,5)
5 y
6 z=prod(x,y)
7 2==2
8 a<-x>3
9 a
10 b<-mean(c(1,2,3,4))
11 b
12 x<-c("apple",
13       "banana")
14
```
- Console:** Shows the execution of the script, with an error message for the non-existent object 'x.y':

```
D:/arpita/data analytics/my work/
length
> x.y
Error: object 'x.y' not found
> prod(x,y)
[1] 1440
> z=prod(x,y)
> 1+1
[1] 2
> x=c(1,2,3,4)
> x
[1] 1 2 3 4
> y=c(3,4,5)
> y
[1] 3 4 5
> z=prod(x,y)
> 2==2
```
- Environment Pane:** Displays the current environment with the following data:

Variable	Value
data	149 obs. of 5 variables
x5.1	num 4.9 4.7 4.6 5 5.4 4.6 5.2 4.4 4.9 5.4 ...
X3.5	num 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 3.7 ...
X1.4	num 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 1.5 ...
X0.2	num 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 0.2 ...
Iris.setosa	Factor w/ 3 levels "Iris-setosa",...: 1 1 1 1 1 1 1 1 1 1 ...
a	logi [1:4] FALSE FALSE FALSE TRUE
b	2.5
x	num [1:4] 1 2 3 4



# How to use R for simple maths

- `> 3+5`
- `> 12 + 3 / 4 - 5 + 3*8`
- `> (12 + 3 / 4 - 5) + 3*8`
- `> pi * 2^3 - sqrt(4)`
- `> factorial(4)`
- `> log(2,10)`
- `> log(2, base=10)`
- `> log10(2)`
- `> log(2)`

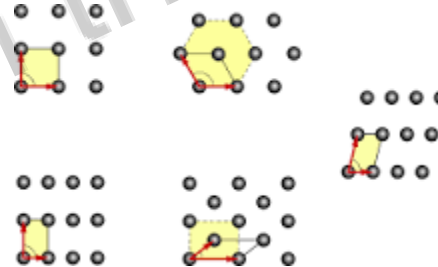
## Note

- R ignores spaces

# Lattice

Lattice data are **observations from a random process observed over a countable collection of spatial regions, and supplemented by a neighborhood structure.**

The observation locations can be regular (equally spaced grid) or irregular, and data at a particular location typically represent the entire region.



a regular geometrical arrangement of points or objects over an area or in space *specifically* : the arrangement of atoms in a crystal

# HOW R PRACTICALLY HELP US

Do not duplicate or distribute without permission from the author and ESO

[Home](#) » [Bioconductor 3.15](#) » [Software Packages](#) » [cancerR](#)

## cancerR

platforms **all** rank **1021 / 2140** support **1 / 4** in Bioc **7 years**  
build **ok** updated **< 3 months** dependencies **155**

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

### A Graphical User Interface for accessing and modeling the Cancer Genomics Data of MSKCC

Bioconductor version: Release (3.15)

The package is user friendly interface based on the `cgdsr` and other modeling packages to explore, compare, and analyse all available Cancer Data (Clinical data, Gene Mutation, Gene Methylation, Gene Expression, Protein Phosphorylation, Copy-Number Alteration) hosted by the Computational Biology Center at Memorial-Sloan-Kettering Cancer Center (MSKCC).

Author: Karim Mezhoud, Nuclear Safety &amp; Security Department, Nuclear Science Center of Tunisia.

Maintainer: Karim Mezhoud &lt;kmezhoud at gmail.com&gt;

Citation (from within R, enter `citation("cancerR")`):

Tunisia, KMNS&SDNSCo (2022), *cancerR: A Graphical User Interface for accessing and modeling the Cancer Genomics Data of MSKCC*, R package version 1.30.01.

### Installation

To install this package, start R (version "4.2") and enter:

```
if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("cancerR")
```

For older versions of R, please refer to the appropriate [Bioconductor release](#).

### Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("cancerR")
```

### 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](#)

### Support »

Please read the [posting guide](#). Post questions about Bioconductor to one of the following locations:

- [Support site](#) - for questions about Bioconductor packages
- [Bioc-devel](#) mailing list - for package developers

1. Neural network for cancer research

**2. simulations of tumor growth**

Practical sessions – share screen mode

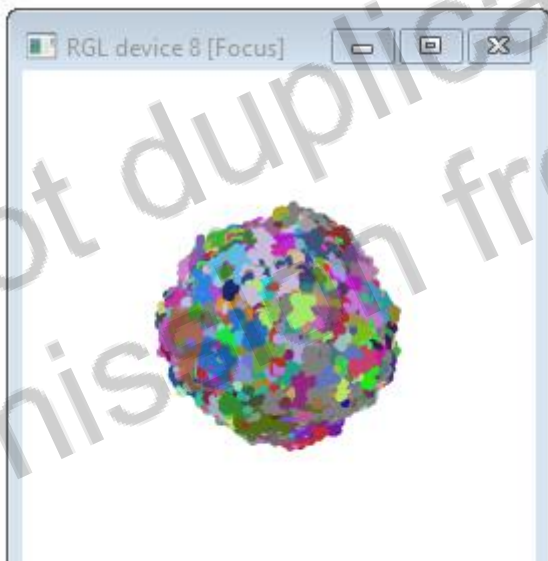
## Creating and visualizing spatial simulations of tumor growth using SITH

The cells within a cancerous tumor are usually highly diverse. An average tumor contains **hundreds of thousands of mutations** spread throughout billions of cancer cells, although it is thought that only a **small** percentage of these mutations are “drivers” which facilitate the progression of cancer into later stages (Greaves and Maley [2012](#)). A lack of understanding about the **evolutionary** process which results in the observed intratumor heterogeneity is a major obstacle preventing the development of effective cancer therapies (Stanta and Bonin [2018](#)).

```

> out <- simulateTumor(max_pop = 250000, verbose = FALSE)
> visualizeTumor(out, background = "white")
>
> names(out)
[1] "cell_ids"      "genotypes"      "muta"           "phylo_tree"     "color_scheme"   "drivers"
[7] "time"          "params"
> head(out$cell_ids)
  x   y   z genotype nmuta distance
1 15 -16 17    6358      2 27.74987
2  1 -23 32      16      2 39.42081
3 -11 19 28    1685      3 35.58089
4 13  -5 34   27691      2 36.74235
5 32 17 -3    64270      7 36.35932
6 43  1 -9    34824      4 43.94315
> head(out$muta)      ## mutation allele frequency (MAF)
  id count MAF
1  0 250000  1
2  1      0  0
3  2      0  0
4  3      0  0
5  4      0  0
6  5      0  0
>
>
> visualizeTumor(out, background = "white")      ## 3D plot of the simulated tumor
> visualizeTumor(out, background = "white")
>
> par(mfrow = c(1,2))
> plotSlice(tumor = out)
> plotSlice(tumor = out, plot.type = "heat")
> plotSlice(tumor = out)
> visualizeTumor(out, background = "white")
> sp <- spatialDistribution(tumor = out)

```



```

library("RGL")
set.seed(1126490984)

out <- simulateTumor(max_pop = 250000, verbose = FALSE)
visualizeTumor(out, background = "white")

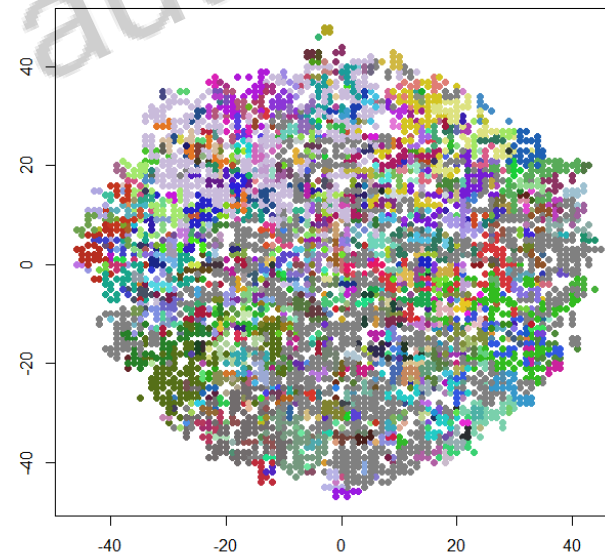
names(out)
head(out$cell_ids)
head(out$muta)      ## mutation allele frequency (MAF)

visualizeTumor(out, background = "white")      ## 3D plot of the simulated tumor
visualizeTumor(out, background = "white", plot.type = "heat") ## which colors cells on a scale from blue to red,
                                                                ## depending on the number of mutations within the cell.

par(mfrow = c(1,2))
plotSlice(tumor = out)
plotSlice(tumor = out, plot.type = "heat")

# Spatial distribution of mutants
sp <- spatialDistribution(tumor = out)

```



## Take home concepts

Use R knowing that it is not your core business

Interdisciplinary approach!  
Alone we are the  
“nothing” proceeding in the  
“Nihil”

Use grants, collaborations, students exchanges to  
establish a statistical R-based community within your  
research groups

R is free and the information/help  
Available on the net can solve most of the  
issues



Questions? Comments?

THANK YOU!!

Do not duplicate or distribute without  
permission from the author and ESO