data

MDPI

Article

BioCPR–A Tool for Correlation Plots

Vidal Fey [1,†], Dhanaprakash Jambulingam [2,†], Henri Sara [3], Samuel Heron [2], Csilla Sipeky [2,4] and Johanna Schleutker [2,5,*]

1 Faculty of Medicine and Health Technology/BioMediTech, Tampere University, 33520 Tampere, Finland; vidal.fey@tuni.fi
2 Cancer Research Unit and FICAN West Cancer Centre, Institute of Biomedicine, University of Turku and Turku University Hospital, 20520 Turku, Finland; dhanaprakash.jambulingam@utu.fi (D.J.); samuel.heron@protonmail.com (S.H.); csilla.sipeky@utu.fi (C.S.)
3 Independent Researcher, 20500 Turku, Finland; henri.sara@gmail.com
4 UCB Pharma, Data & Translational Sciences, 1420 Braine l'Alleud, Belgium
5 Laboratory Division, Department of Medical Genetics, Genomics, Turku University Hospital, 20521 Turku, Finland
* Correspondence: johanna.schleutker@utu.fi; Tel.: +358-29-450-2726
† These authors contributed equally to this work.

Abstract: A gene is a sequence of DNA bases through which genetic information is passed on to the next generation. Most genes encode for proteins that ultimately control cellular function. Understanding the interrelation between genes without the application of statistical methods can be a daunting task. Correlation analysis is a powerful approach to determine the strength of association between two variables (e.g., gene-wise expression). Moreover, it becomes essential to visualize this data to establish patterns and derive insight. The most common method for gene expression visualization is to use correlation heatmaps in which the colors of the plot represent strength of co-expression. In order to address this requirement, we developed a visualization tool called BioCPR: Biological Correlation Plots in R. This tool performs both correlation analysis and subsequent visualization in the form of an interactive heatmap, improving both usability and interpretation of the data. BioCPR is an R Shiny-based application and can be run locally in Rstudio or a web browser.

Keywords: correlation heatmaps; gene expression; r shiny application

Citation: Fey, V.; Jambulingam, D.; Sara, H.; Heron, S.; Sipeky, C.; Schleutker, J. BioCPR–A Tool for Correlation Plots. *Data* **2021**, *6*, 97. https://doi.org/10.3390/data6090097

Academic Editor: Ren-Hua Chung

Received: 30 June 2021
Accepted: 30 August 2021
Published: 8 September 2021

1. Introduction

Correlation analysis is a statistical method used to ascertain the strength and direction of the relationship between two variables. Correlation is expressed using the correlation coefficient, which varies between −1 and +1. A correlation coefficient value of ±1 represents the perfect degree of association between variables, whereas the signs "+" and "−" represent the direction of the relationship. A value of 0 signifies no relationship between the variables.

The most commonly used statistical analysis methods in correlation analysis are Kendall rank correlation, Spearman's rank correlation, and Pearson's correlation. Spearman's rank correlation performs best for variables with skewed distributions and Kendall rank correlation is used to measure the strength of dependence between two variables, whereas Pearson's correlation is used when the variables being studied are normally distributed [1–8].

The formulas to represent the correlation are as follows:

i. Kendall rank correlation:

$$\tau = \frac{(n_c) - (n_d)}{\frac{1}{2}n(n-1)}$$

where,
τ = Kendall coefficient

n = sample size
n_c = number of concordant
n_d = number of discordant pairs

ii. Spearman's rank correlation:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where,
ρ = Spearman's rank correlation coefficient
d = difference between the ranks of corresponding values
n = number of observations

iii. Pearson correlation:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

where,
r_{xy} = Pearson correlation coefficient between x and y
n = number of observations
x_i = value of x (for ith observation)
y_i = value of y (for ith observation)

In a biological context, such as the study of gene expression, the interpretation of correlation coefficients can become a strenuous task due to the number of variables. In order to better understand the relationship and to identify clear patterns, it is therefore good practice to visualize the results. There are several notable R packages for this purpose, including corrplot [9], ggcorrplot [10], and GGally [11], which visualize correlation in the form of heatmaps. However, these packages require a user to have knowledge of R programming and to calculate the correlation matrix themselves. While this offers more flexibility when performing correlation analysis, it also implies that the user has statistical knowledge. Furthermore, there is a lack of user-friendly, web-based interactive applications that are published in peer-reviewed journals. To fill this gap, and without imposing a great deal of prior data analysis experience on the user, we developed BioCPR. BioCPR allows users to upload their expression dataset in various file formats, display the contents of the dataset for inspection and optional pre-filtering, and optionally retrieve HUGO Gene Nomenclature Committee (HGNC) gene symbols for corresponding Ensembl IDs. After data are uploaded to the app server, the dataset can be interactively filtered to optimize the correlation heatmap. Filtering can be carried out either by pre-defining genes of interest or by restricting the number of genes to be plotted as per their variance. By filtering according to variance, a standard pre-filtering technique, the data matrix is reduced to the genes with the highest variation before calculating correlation. On the other hand, filtering with genes of interest works on the precalculated correlation matrix and selects the genes adjacent to the specified genes in the matrix, where the number of "surrounding" genes can be defined by the user. The correlation plot and the matrix can be viewed, and then the heatmap can be altered interactively using the following features:

1. Filter by correlation value;
2. Filter by "constant-height tree cut" implemented in the function *cutreeStatic* in the R package *WGCNA*.

The plot can be further adjusted and customized using the controls on the "CORRELATION HEATMAP" Table. To facilitate interpretation of the depicted correlation values, significance asterisks can be added to encode significance of the correlation. In cases of large data sets and no filtering, the user can zoom in up to 800% to allow for better readability. If genes of interest are selected for filtering the data set, those can be highlighted in the heatmap. Further, the label size can be adjusted, and a custom plot title

can be added. Both significance asterisks and gene highlighting are features that are, to the authors' knowledge, not readily available in current R packages that produce correlation heatmaps.

2. Materials and Methods

BioCPR is an open-source application written in the R programming language [12]. As the R programming language is supported by all major operating systems, such as Linux, MacOS, and Windows, BioCPR benefits from cross-platform compatibility. BioCPR utilizes the Shiny library, an open-source R package that provides a framework for building interactive applications with a graphical user interface (GUI) that can be run either from RStudio or within a web browser [13]. The backend functionalities used to calculate the correlation coefficients and plot them for interactive visual exploration were developed by our team and are available in the Comprehensive R Archive Network (CRAN).

The app heavily relies of the functionality of the R package *heatmapFlex* implementing a novel graphics engine designed for drawing heatmaps in a flexible way and adding new features as compared with existing heatmap libraries. Two major explorative functions of that package are the significance asterisks showing statistical significance of the correlations and the highlighting of rows and columns for genes of interest in the heatmap, both available in the app.

A list of the packages used and short descriptions of their functions are presented in Table 1.

Table 1. A list of in-house R packages used by BioCPR and their functions in brief.

R Packages	Description of Function
coreheat	Calculates and clusters correlation matrices, tests the strength of two correlations and plots correlation heatmaps
convertid	Convert Gene IDs between each other and fetch annotations from Biomart
readmoRe	Tools for importing and manipulating bio-medical data files
heatmapFlex	Tools for producing flexible heatmaps with functions including, but not limited, to zooming, splitting, and legends

Density distribution of gene expression data is mostly normal or log-normal (around 80%), since log-normal distributions can be converted to a normal distribution; gene expression data is considered as normal distribution in most simple analyses [14]. BioCPR therefore uses Pearson's correlation for the calculation of correlation coefficients from the gene expression data with the resultant values used in the creation of an interactive heatmap [15]. Our tool helps users to plot heatmaps in a dynamic manner, allowing them control over visualization parameters, from viewing the variables in the uploaded file and sorting the values based on sample identifiers to the number of genes to be included, clustering and filtering genes based on their correlation, and cutting dendrograms based on heatmap clusters. After generating the heatmap, it is possible to adjust the resolution of the image and gene labels, as well as indicate significant co-expression using asterisks, for which *p*-values are calculated from the correlation coefficients using a Student's *t*-test.

In order to test for significance, t-test statistics are calculated as follows [16]:

$$t = \frac{r * \sqrt{n-2}}{\sqrt{1-r^2}}$$

where,

r = correlation coefficient

n = number of samples with no missing values for each gene

The distribution function "pt" from the Student t Distribution in the R *stats* package is used to calculate the probability cumulative densities [12]. The pt function provides

p-values for each associated t-score and is corrected to account for the two-tailed nature of the test, i.e., the possibility of positive as well as negative correlation. The *p*-values are then used to indicate the significance levels, as described in Table 2.

Table 2. *p*-values and their corresponding significance asterisks.

p-Value	Number of Asterisks
p-value < 0.05	one asterisk (*)
p-value < 0.01	two asterisks (**)
p-value < 0.001	three asterisks (***)

BioCPR is intended to be a straightforward tool that provides an easy and user-friendly environment for users to explore gene expression data in a graphical format. To aid in usability, the tool features an inbuilt tab explaining usage instructions and example input format. Typical input for the tool is a single file in tab-delimited ASCII text format containing a record for each gene and sample-wise expression values as fields.

We compared BioCPR with the most commonly available tools for the visualization of correlation data. The following key features were examined: GUI availability, correlation coefficient calculation, and compatibility with gene expression data.

Data Acquisition and Processing

The functionality of the tool was examined using The Cancer Genome Atlas (TCGA) dataset consisting of 333 patients with primary prostate cancer [17]. The dataset was obtained via the cBioPortal for Cancer Genomics [18], and two subsets of the data were created; one contained genes commonly associated with cancer and the other contained 5000 genes selected at random. The subsets were named PrCaTCGASample.txt and PrCaTCGASample_5000.txt respectively, and are provided in the tool repository.

3. Results and Discussion

We developed BioCPR, a shiny application that enables users without programming skills to generate effective visualizations from gene expression data. Our tool demonstrates integrated compatibility with BioMart and the option to select genes based on their variance, in addition to functionalities such as clustering, dendrogram construction, and quality image production.

We compared the BioCPR tool with other widely used alternatives for correlation analysis, and this comparison is detailed in Table 3. When utilizing these alternative methods, knowledge of programming with a statistical language like R, and its libraries such as corrplot, ggcorrplot or Ggally, is a prerequisite and presents a problem for users without prior programming experience. In contrast to these approaches, BioCPR presents a more user-friendly approach whilst incorporating inbuilt support for biological nomenclature.

Table 3. A comparison of the features of BioCPR with other widely used tools for plotting correlation heatmaps.

	Corrplot	Ggcorrplot	GGally	BioCPR
Programming language	R	R	R	R
Availability	Free	Free	Free	Free
Correlation co-efficient calculation	Manual	Manual	Manual	Automatic
Graphical User Interface (GUI)	No	No	No	Yes
Support for gene expression data	No	No	No	Yes

The most common situations in which BioCPR would be useful would be when:

1. There is a gene of interest and one would like to explore which genes the gene of interest clusters to and what the correlation significance of those clusters is.
2. There is a large dataset and BioCPR would be used to filter down the dataset to a handful of genes from which gene-wise interactions of statistical significance could be studied further.

We demonstrated the usability of BioCPR with the following use cases. In the correlation plots, the smaller panel on the top left of the main plot shows the histogram of the correlation coefficients and also serves as the correlation color key. The heatmaps are clustered; positive correlations are indicated by the color red, whereas negative correlations are indicated by the color blue. The X and Y axes are labeled with the gene symbols.

Use Case 1: Using a Pre-Filtered Dataset

A heatmap generated by BioCPR is depicted in Figure 1, and the correlation matrix generated by BioCPR to create the heatmap is depicted in Figure 2. The heatmap shows a representation of the direction and strength of association between the expression of individual genes.

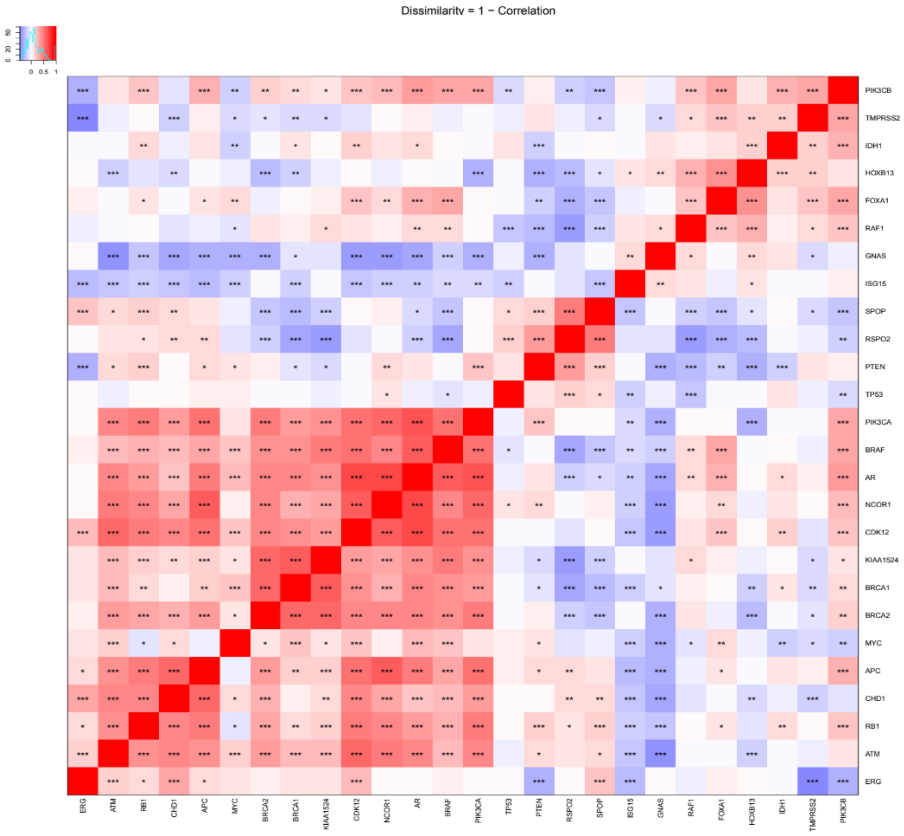

Figure 1. Correlation heatmap of the subset of the TCGA dataset. The smaller panel on the top left of the plot shows the color key and a histogram of correlation coefficients. Positive correlations are indicated in shades of red, whereas negative correlations are indicated in shades of blue. Significant correlations, depending on the *p*-value, are indicated by asterisks.

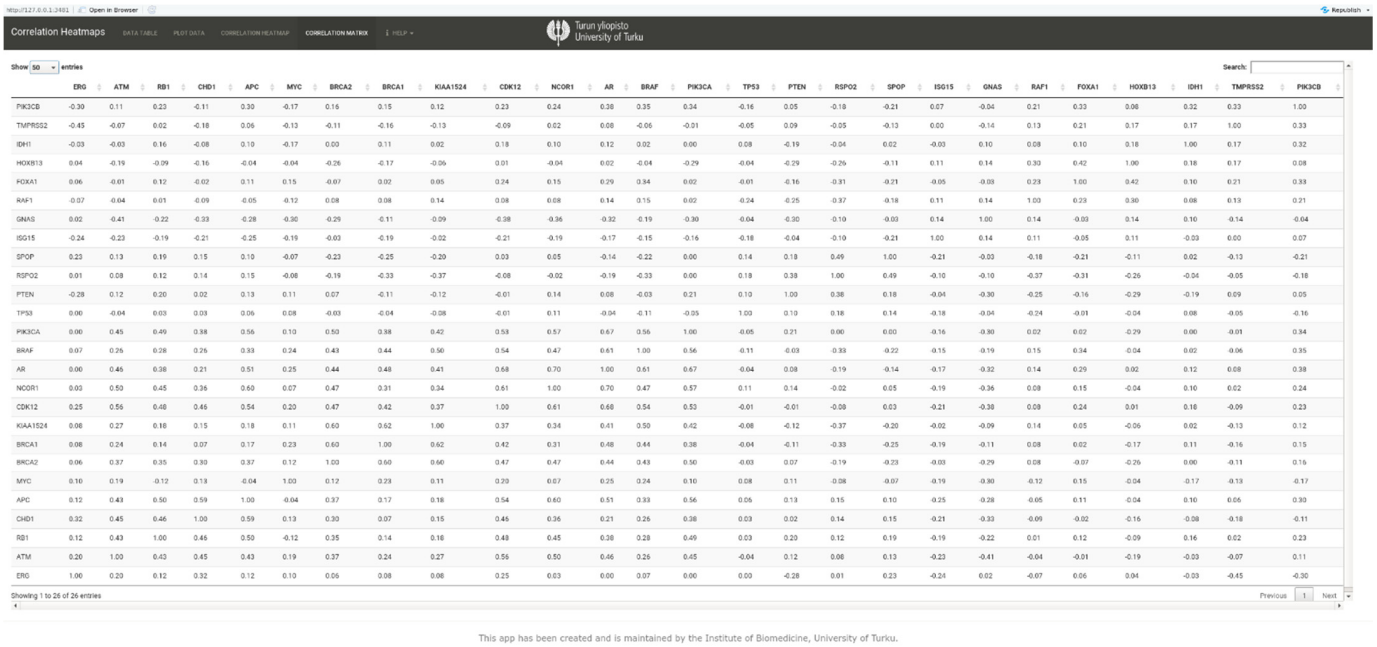

Figure 2. Correlation matrix of the subset of the TCGA dataset.

A glance at the plot reveals which genes are positively correlated and which are negatively correlated. Strong positive correlations can be seen, for example, with gene pairs BRCA1-BRCA2, APC-NICOR1, NICOR1-CDK12, CDK12-AR, ATM-CDK12, KIAA1524-BRCA1, and KIAA1524-BRCA2, whereas strong negative correlations can be seen, for example, in gene pairs HOXB13-BRCA2, HOXB13-BRCA1, HOXB13-PTEN, BRCA1-RSPO2, ATM-GNAS, NICOR1-GNAS, and CDK12-GNAS.

Use Case 2: Using an Unfiltered Dataset

For the second use case, we randomly sampled 5000 genes from the TCGA dataset and used the BioCPR tool to select the most significant genes according to their variance. This was carried out to enable us to have a closer look at genes and understand how they cluster with each other and how far from the gene of interest the correlation is significant.

From the 5000 genes selected from the TCGA dataset, the 250 genes with highest significant variance were selected to be plotted using the tool. It can be observed from Figure 3 that the gene cluster at the bottom left has a higher correlation significance and could contain novel gene interactions that have not been studied before. The plot was then filtered down by cutting the clustered area into smaller dendrograms, as shown in Figure 4, to focus on a smaller number of genes. The plot can be zoomed in on and studied, and the genes can be taken for further studies or filtered down using more stringent criteria before proceeding further.

The correlation represented in the above plots are supported by their strong *p*-values indicated in the form of significance asterisks. The plots give an idea of which genes could be co-expressed and which could not. Using this as base, the analysis could be built further to include:

1. Co-expression studies to create a gene co-expression network to determine, for example, their functional similarity, shared regulatory inputs, and functional pathways.
2. A combination with genotype information from DNA-sequencing data to perform eQTL studies to identify the relationship between genetic variants and gene expression, and which variants influence the expression of genes of interest.
3. Gene set enrichment analysis using gene pairs to see which genes are over-represented and possibly associated with disease phenotypes.

These observations not only open up avenues for further investigation into the expression of genes but also stand to demonstrate the utility of BioCPR. Although the application was developed for calculating and visualizing gene co-expression, it can be further applied to explore correlation in non-biological datasets to visualize patterns or to check the reliability of regression analysis.

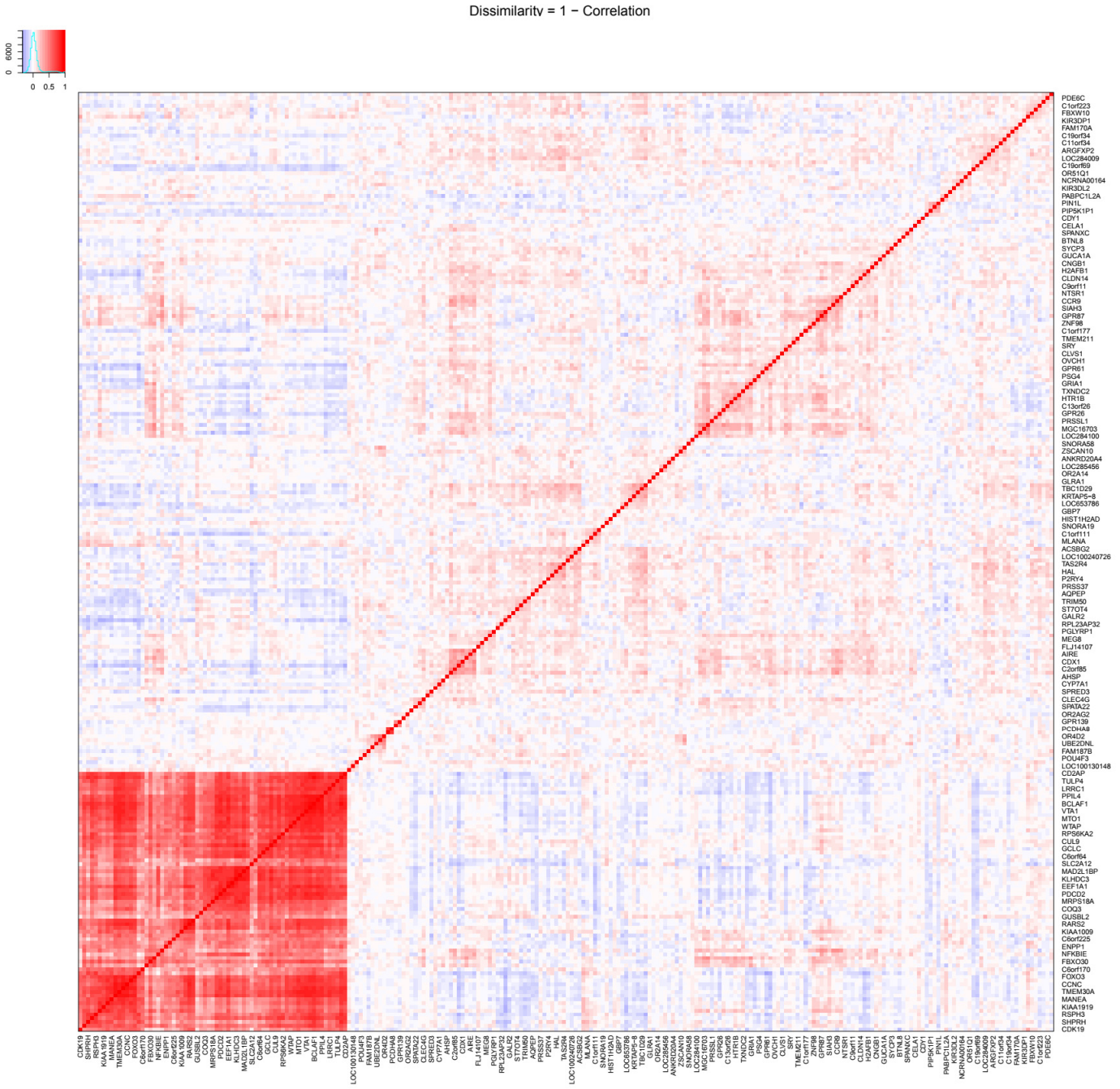

Figure 3. Correlation heatmap of the subset of the TCGA dataset that has been filtered using the BioCPR tool. See Figure 1 for a description of the color scheme.

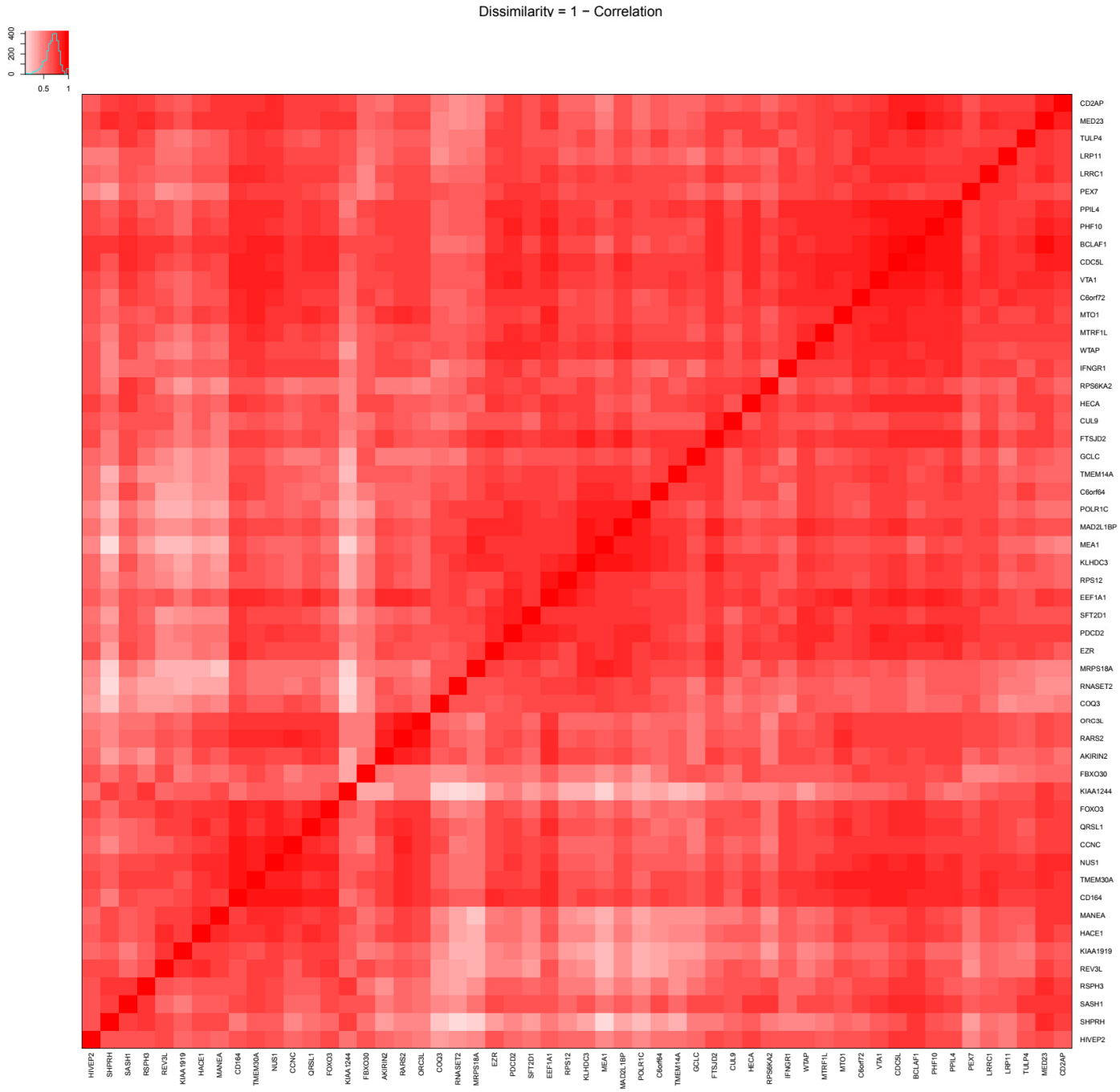

Figure 4. Dendrogram that has been filtered from the correlation heatmap in Figure 3. See Figure 1 for a description of the color scheme.

3.1. Availability and Accessibility

A live version of BioCPR tool is hosted in the Shinyapps server and can be accessed via the shinyapps.io platform (https://biocpr.shinyapps.io/corplot/, accessed on 29 August 2021). The source code is available through the Github repository (https://github.com/vfey/biocpr, accessed on 29 August 2021), and the source code can be run on a personal computer or hosted on a private server. Further instructions for installation and usage of the tool is provided in the attached Supplementary Materials.

3.2. Input File Size

The maximum permitted file size is currently set at 40 MB. This can be changed as per user requirements when running the tool from its source code. In practice, cramming in too much information can defeat the purpose of visualizing the data and could eclipse meaningful correlations.

3.3. Privacy and Security of Data

Genetic data are usually subjected to data protection regulations of the respective state (in Europe, the General Data Protection Regulation (GDPR)), and transferring/uploading the data to an external server can violate policy. BioCPR offers a significant advantage in this regard, as the tool can be run on a personal computer without having to transfer/upload the data elsewhere.

3.4. Future Developments

BioCPR is under continuous development, and we have the following updates planned for future versions:

1. Improving visualization options by providing choices for different color palettes, including those to help people with color blindness.
2. Making the correlation matrix downloadable as a text file as it is not currently downloadable.
3. Adopting plotly to improve the interactivity of the plot.

Questions, feedback or expression of possible contributions can be addressed to the above contact details.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/data6090097/s1, Figure S1: Screenshot of the download page on the official R website (https://cran.r-project.org/, accessed on 29 August 2021), Figure S2: Screenshot of the download page on the official RStudio website (https://rstudio.com/products/rstudio/download/, accessed on 29 August 2021), Figure S3: Screenshot of the RStudio startup page, Figure S4: Screenshot of the ui.R file in the RStudio interface with the "Run App" button highlighted, Figure S5: Screenshot of the startup page of BioCPR tool, Figure S6: Screenshot of the startup page of BioCPR tool after loading the dataset, Figure S7: Screenshot of the "PLOT DATA" tab where we select the genes for plotting, Figure S8: Screenshot of the selection criteria under the plot data tab, Figure S9: Screenshot of the correlation heatmaps generated from data selected in "PLOT DATA" tab, Figure S10: Screenshot of the options under "CORRELATION HEATMAP" tab, Figure S11: Screenshot of the heatmap with significance stars added, Figure S12: Screenshot of the advanced options for editing heatmap, Figure S13: Screenshot of the "CORRELATION MATRIX" tab.

Author Contributions: Conceptualization, V.F.; Funding acquisition, J.S.; Methodology, V.F. and H.S.; Project administration, J.S.; Resources, D.J., S.H. and C.S.; Software, V.F., D.J. and H.S.; Supervision, J.S.; Validation, V.F., D.J. and S.H.; Visualization, D.J.; Writing–original draft, D.J.; Writing–review and editing, D.J., V.F., S.H., C.S. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: The work was financially supported by Cancer Foundation Finland sr. (grant to JS).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were used in this study. This data can be found here: https://www.cbioportal.org/study/summary?id=prad_tcga_pub accessed on 29 August 2021.

Acknowledgments: The authors gratefully thank Amanda Tursi and Leigh Ann Lindholm for their efforts in proofreading the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BioCPR	Biological correlation plots in R
HGNC	HUGO Gene Nomenclature Committee
CRAN	The Comprehensive R Archive Network
GUI	Graphical User Interface
TCGA	The Cancer Genome Atlas
WGCNA	Weighted Correlation Network Analysis

References

1. Boas, F. Determination of the Coefficient of Correlation. *Science* **1909**, *29*, 823–824. [CrossRef] [PubMed]
2. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **2015**, *163*, 1011–1125. [CrossRef] [PubMed]
3. Ethan, C.; Gao, J.; Gogrusoz, U.; Gross, B.E.; Sumer, S.O.; Bülent Arman Aksoy, A.J.; Byrne, C.J.; Heuer, M.L.; Larsson, E.; Antipin, Y.; et al. The CBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.* **2012**, *2*, 401–404. [CrossRef]
4. Winston, C.; Cheng, J.; Allaire, J.J.; Sievert, C.; Schloerke, B.; Xie, Y.; Allen, J.; McPherson, J.; Dipert, A.; Borges, B. Shiny: Web Application Framework for R. *CRAN* **2021**. Available online: https://cran.r-project.org/package=shiny (accessed on 15 August 2021).
5. Avijit, H.; Gogtay, N. Biostatistics Series Module 6: Correlation and Linear Regression. *Indian J. Dermatol.* **2016**, *61*, 593–601. [CrossRef]
6. Alboukadel, K. Ggcorrplot: Visualization of a Correlation Matrix Using 'Ggplot2'. *CRAN* **2019**. Available online: https://cran.r-project.org/package=ggcorrplot (accessed on 15 August 2021).
7. George, K.M. A New Measure of Rank Correlation. *Biometrika* **1938**, *30*, 81–93. [CrossRef]
8. Taiyun, K.; Chen, I.R.; Lin, Y.; Wang, A.Y.; Yang, J.Y.H.; Yang, P. Impact of Similarity Metrics on Single-Cell RNA-Seq Data Clustering. *Brief. Bioinform.* **2019**, *20*, 2316–2326. [CrossRef]
9. Ming, L.H.; Yang, D.; Liu, Z.F.; Hu, S.Z.; Yan, S.H.; He, X.W. Density Distribution of Gene Expression Profiles and Evaluation of Using Maximal Information Coefficient to Identify Differentially Expressed Genes. *PLoS ONE* **2019**, *14*, e0219551. [CrossRef]
10. Jeremy, M.; Banyard, P. *Understanding and Using Statistics in Psychology: A Practical Introduction*; SAGE Publications Ltd.: London, UK, 2007. Available online: https://sk.sagepub.com/books/understanding-and-using-statistics-in-psychology (accessed on 15 August 2021).
11. Mavuto, M. Statistics Corner: A Guide to Appropriate Use of Correlation Coefficient in Medical Research. *Malawi Med. J.* **2012**, *24*, 69–71. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/ (accessed on 15 August 2021).
12. Karl, P. Determination of the Coefficient of Correlation. *Science* **1909**, *30*, 23–25. [CrossRef]
13. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020. Available online: https://www.r-project.org/ (accessed on 15 August 2021).
14. Barret, S.; Crowley, J.; Cook, D.; Wickham, H.; Briatte, F.; Marbach, M.; Thoen, E.; Elberg, A.; Larmarange, J. GGally: Extension to 'Ggplot2'. *CRAN* **2021**. Available online: https://cran.r-project.org/package=GGally (accessed on 15 August 2021).
15. Charles, S. The Proof and Measurement of Association between Two Things. *Am. J. Psychol.* **1904**, *15*, 101. [CrossRef]
16. Statistics Solutions. Correlation (Pearson, Kendall, Spearman). Statisticssolutions.Com. Available online: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/ (accessed on 15 August 2021).
17. Taiyun, W.; Simko, V. R Package 'Corrplot': Visualization of a Correlation Matrix. *CRAN* **2021**. Available online: https://github.com/taiyun/corrplot (accessed on 15 August 2021).
18. Clark, W. The Spearman Correlation Formula. *Science* **1905**, *22*, 309–311. [CrossRef]