PCAGO: An interactive web service to analyze RNA-Seq data with principal component analysis

Ruman Gerst[1,2] *and Martin Hölzer*[3,4,*]

[1] *Applied Systems Biology, Leibniz Institute for Natural Product Research and Infection Biology, HKI, Jena, Germany*
[2] *Faculty of Biological Sciences, Friedrich Schiller University Jena, Jena, Germany*
[3] *RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University Jena, Leutragraben 1, Jena, Germany*
[4] *European Virus Bioinformatics Center, Friedrich Schiller University Jena, Leutragraben 1, Jena, Germany*

[*]To whom correspondence should be addressed.

ABSTRACT

The initial characterization and clustering of biological samples is a critical step in the analysis of each transcriptomic study. In most studies, principal component analysis (PCA) is the clustering algorithm of choice to predict the relatedness of samples or cells based solely on differential gene expression. Besides the pure quality assessment of the data, a PCA can also give first insights into the biological background of an experiment and help researchers to interpret the data and to design the following computational steps accordingly. However, to avoid misleading clusterings and interpretations, an appropriate selection of the underlying gene sets to build the PCA and the choice of the most fitting principal components for the visualization are crucial parts. Here, we present PCAGO, an easy-to-use and interactive web service to analyze gene quantification data derived from RNA sequencing (RNA-Seq) experiments with PCA. The tool includes features such as read-count normalization, filtering of read counts by gene annotation, and various visualization options. Additionally, PCAGO helps to select appropriate parameters such as the number of genes and principal components to create meaningful visualizations.

Availability and implementation: The web service is implemented in R and freely available at pcago.bioinf.uni-jena.de.

Contact: martin.hoelzer@uni-jena.de

Introduction

In the last decade, the sequencing of whole transcriptomes (RNA-Seq) emerged as a powerful technique to understand versatile molecular mechanisms and to tackle various biological questions. The initial characterization and clustering of biological samples based on their different gene expression levels are crucial steps in the early access of RNA-Seq experiments.

Principal component analysis (PCA)[1] is a technique used to emphasize variation and to carve out strong patterns in a high-dimensional data set, such as a matrix of gene expression values[2]. The output of a PCA transformation can be used to make data easy to explore and visualize. PCAs are therefore widely used approaches in bioinformatics[3] and should be an important part of almost all RNA-Seq analysis pipelines today[4].

Although PCAs are simple to compute and frequently used in transcriptomic, metabolomic, and medical studies, there are several challenges and disadvantages that can especially occur when analyzing experimental data[5]. For example, the visualization and interpretation of the data can be heavily influenced by the final selection of the principal components, the number of most-variant genes and prior normal-

ization of gene expression[2,5]. Therefore, a careful evaluation of the PCA and its components is crucial to derive meaningfully interpretations from the transformation.

As a dimension reduction approach, a PCA searches for a small set of 'metagenes', represented by linear combinations of all genes[3]. Such linear combinations of the original measurements (one read count per gene and sample) are called principal components (PCs). In other words, a PCA merges the quantified expression information from correlated genes into one metagene or PC. The PCs are orthogonal to each other and may have a much lower dimensionality than the original data. Ultimately, the PCs can be plotted against each other to show the grouping of the measured expression data. For a detailed description and formal definition of a PCA with regard to gene expression data we refer to[3]. As a best practice, a PCA should first be used for quality control, e.g. to check whether biological replicates have a similar expression profile and tend to cluster together[4].

By definition, the first principal component (PC1) explains the greatest possible variance in the data (e.g. for an infection experiment, 40 % of the variance between control and infected samples is explained by the genes that define PC1). PC2 explains the second greatest portion of variance, and so on. Traditionally, the first few PCs are used for visualization since they capture most of the variation from the original data set.

However, less important PCs such as PC3 or PC4 might capture strong patterns comparable to PC1 and PC2 and help to distinguish between gene expression patterns more clearly[2]. Different gene sets that can be used as the backbone of the PCA, as well as a varying number of PCs, can influence the effectiveness in capturing the cluster structure[2,3]. Accordingly, there is a great need to investigate the effectiveness of the PCA transformation before drawing conclusions on the gene expression data[2].

Here, we present PCAGO (pcago.bioinf.uni-jena.de), an interactive and easy-to-use web service to analyze and visualize quantification data derived from RNA-Seq with PCA. The web service's functionality can be easily tested by loading sample RNA-Seq read counts obtained from Klassert et al.[6] and Riege et al.[7] (execute *Start analysis* $\longrightarrow$ *Load all sample data* online).

In contrast to other PCA tools[8,9], PCAGO gives the user the opportunity to comprehensively inspect and vary key features of the transformation, such as the PCs, and to easily compare different outcomes. Furthermore, PCAs can be built on annotation-filtered gene subsets, that might better reflect the biological question of interest in the final visualization. A minimum set of genes that produces the same clustering as all genes can be calculated and an animated PCA shows how PCs and the clustering change by incrementally including less variant genes in the transformation. Finally, all computational steps are well documented to allow full reproducibility of the performed PCA.

PCAGO workflow and features

PCAGO requires a table of raw or already normalized read count data as produced by any standard RNA-Seq pipeline[4] as input (Fig. 1A). Based on the raw read counts, PCAGO can perform the following steps: normalization (DESeq2-based, TPM); sample and gene set annotation; Ensembl and gene ontology (GO) integration from online databases; filtering of gene subsets by gene annotation and variance cutoff; visualization of dendrograms, 2D/3D-PCAs, scree plots, and PC loadings; PCA plot animation; export of all results in various formats; and the creation of an HTML-based documentation for full reproducibility. While most of these steps are optional, at least one read count table and corresponding sample annotation must be provided to perform the PCA (Fig. 1A). All output of PCAGO can be downloaded in various formats (pdf, svg, png, tiff, mp4) and is directly visualized in the web app (Fig. 1B).

Based on microarray-derived expression data, De Haan et al.[10] showed the advantages of performing PCAs only on certain gene subsets. One of the major used methods to categorize functional gene information is the Gene Ontology (GO) database[11]. With PCAGO, different gene subsets filtered by features such as chromosome, Ensembl[12] biotype (protein coding, miRNA, lncRNA, ...) or GO terms can be defined and directly used for the PC calculations. An annotation can be applied by the user or

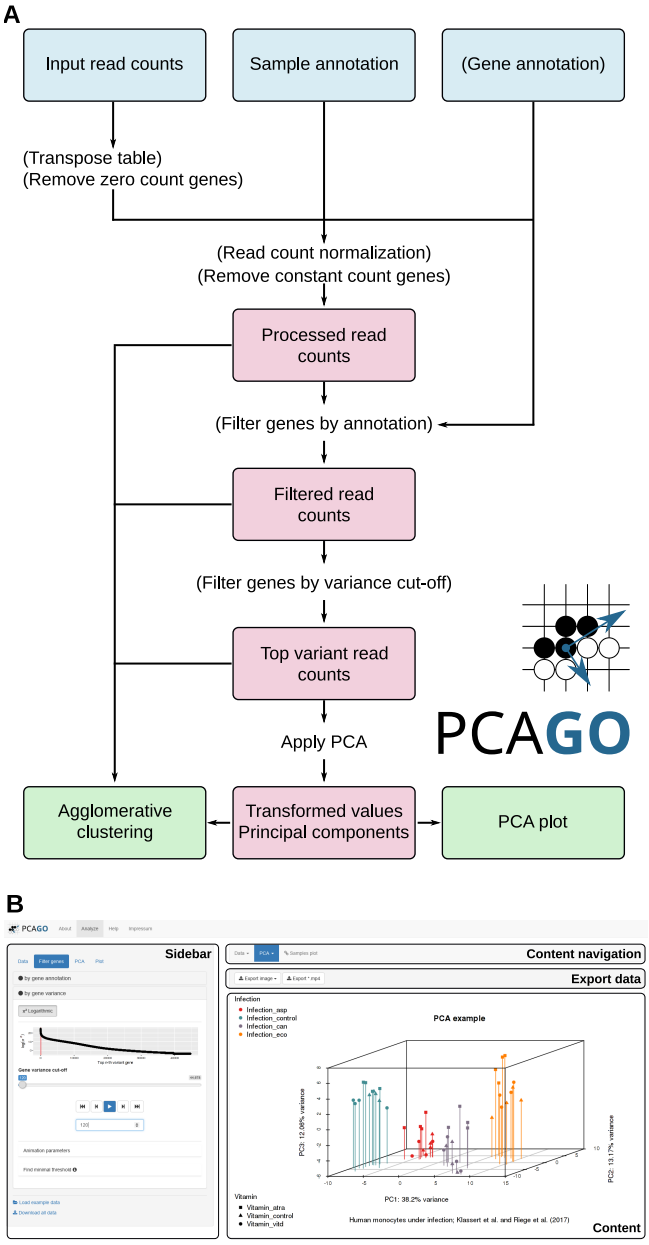

Figure 1: (A) Workflow of the PCAGO web service. Optional steps are written in brackets. **(B)** Interface of the web service. Example data[6,7] was loaded and a 3D-PCA plot drawn.

directly extracted from online databases such as Ensembl BioMart[13].

The core of PCAGO is the visualization of interactive 2D/3D PCA and hierarchical clustering plots as well as scree and loading plots of the PCs (see online manual). An animation feature is included to analyze how adding more and more variant genes into the calculation changes the outcome. We included highly flexible customization and coloring options directly within the web application to generate publication-ready vector graphics. All generated data, as well as detailed reports about the processing steps, can be viewed and exported in various formats. As a special feature, PCAGO can export an MP4 movie of the animated PCA plot.

Availability and implementation

PCAGO is implemented in R[14] and distributed on the powerful R Shiny framework. The web service, including a comprehensive online documentation with how-to and help sections, as well as an easy-to-load example data set, is freely available at pcago.bioinf.uni-jena.de. The source code, as well as a list of all used software packages, are available on GitHub (github.com/rumangerst/pcago). PCAGO can be locally installed and executed based on a functional R (v3.4.3 or higher) and MariaDB installation. Furthermore, a stand-alone desktop version developed with Electron) is available at github.com/rumangerst/pcago-electron.

Funding MH has been supported by the Deutsche Forschungsgemeinschaft (DFG) through the Priority Program SPP-1596 MA5082/7-1 and DFG CRC 1076 “AquaDiva”, subproject A06. RG is funded by the International Leibniz Research School for Microbial and Biomolecular Interactions (ILRS).

Conflict of interest statement None declared.

Author Contributions MH developed the research idea. RG did the methods development and implementation of the tool. RG implemented all functionalities of the web service. MH wrote the main manuscript. MH and RG proofread and finalized the manuscript for publication.

References

[1] J. E. Jackson. *A user's guide to principal components*, volume 587. John Wiley & Sons, 2005.

[2] K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

[3] S. Ma and Y. Dai. Principal component analysis based methods in bioinformatics studies. *Briefings Bioinf*, 12(6):714–722, 2011.

[4] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biol*, 17:13, January 2016. ISSN 1474-760X.

[5] T. Konishi. Principal component analysis for designed experiments. *BMC Bioinf*, 16(18):S7, 2015.

[6] T. E. Klassert, J. Bräuer, M. Hölzer, M. Stock, K. Riege, C. Zubiría-Barrera, M. M. Müller, S. Rummler, C. Skerka, M. Marz, et al. Differential effects of vitamins A and D on the transcriptional landscape of human monocytes during infection. *Sci Rep*, 7, 2017.

[7] K. Riege, M. Hölzer, T. E. Klassert, E. Barth, J. Bräuer, M. Collatz, F. Hufsky, N. Mostajo, M. Stock, B. Vogel, H. Slevogt, and M. Marz. Massive effect on lncRNAs in human monocytes during fungal and bacterial infections and in response to vitamins A and D. *Sci Rep*, 7: 40598, 2017.

[8] A. A. Sharov, D. B. Dudekula, and M. S. Ko. A web-based tool for principal component and significance analysis of microarray data. *Bioinformatics*, 21(10):2548–2549, 2005.

[9] T. Metsalu and J. Vilo. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Res*, 43:W566–W570, July 2015. ISSN 1362-4962.

[10] J. R. De Haan, E. Piek, R. C. van Schaik, J. de Vlieg, S. Bauerschmidt, L. M. Buydens, and R. Wehrens. Integrating gene expression and GO classification for PCA by preclustering. *BMC Bioinf*, 11(1):158, 2010.

[11] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, G. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25 (1):25–29, 2000.

[12] D. R. Zerbino, P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Girón, et al. Ensembl 2018. *Nucleic Acids Res*, 46(D1):D754–D761, 2017.

[13] R. J. Kinsella, A. Kähäri, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, P. Kersey, and P. Flicek. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)*, 2011:bar030, 2011. ISSN 1758-0463.

[14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.