Generate hypotheses assisted by protection score and enable new discoveries

INTRODUCTION

In this example we will show and discuss how the combination of visualization techniques and projection score can be used both to avoid very clear pit-falls associated with data set analysis and how projection score can be used to improve the outcome of the analysis and enable new discoveries.

This example is for users who are familiar with Qlucore Omics Explorer (QOE)

RESULTS

1. By starting the analysis through exploration we will show how you within minutes detect potential unwanted factors (batch effects or not controlled factors).

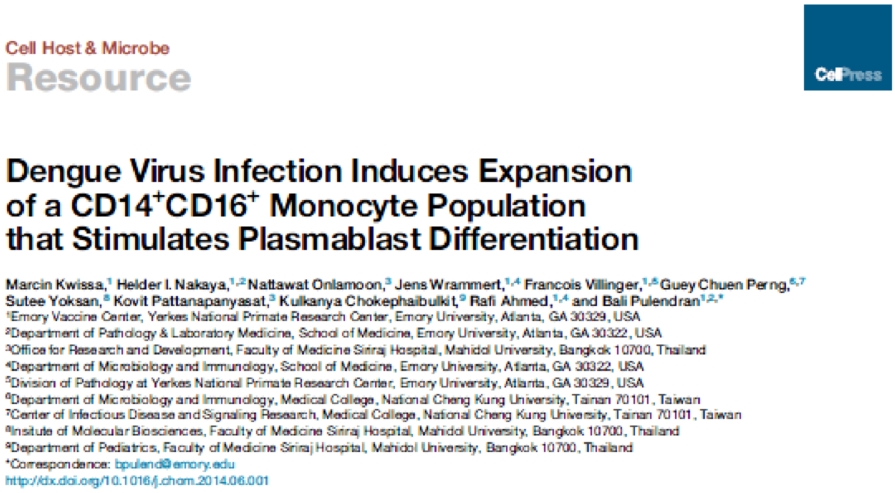
2. We will then show how to identify the key subgroups in data by using the unique projection score metric.

3. Finally we will show how to avoid missing unknown subgroups in the data by starting with data exploration and hypothesis generation instead of directly applying a statistical test that matches the main hypothesis.

DATA SET

The example is based on a data set downloaded from Gene Expression Omnibus - GDS5093. The data set is from a study related to Dengue Fever. This specific data set is not vital for the example; the approach can be used on any data set.

Note: The outcome of the analysis below shall not be seen as a review or comment to the original analysis, it is an example on how to perform data analysis.



Cell Host & Microbe

Resource

CellPress

Dengue Virus Infection Induces Expansion of a CD14$^+$CD16$^+$ Monocyte Population that Stimulates Plasmablast Differentiation

Marcin Kwissa,[1] Helder I. Nakaya,[1,2] Nattawat Onlamoon,[3] Jens Wrammert,[1,4] Francois Villinger,[1,5] Guey Chuen Perng,[6,7] Sutee Yoksan,[8] Kovit Pattanapanyasat,[3] Kulkanya Chokephaibulkit,[9] Rafi Ahmed,[1,4] and Bali Pulendran[1,2,*]
[1]Emory Vaccine Center, Yerkes National Primate Research Center, Emory University, Atlanta, GA 30329, USA
[2]Department of Pathology & Laboratory Medicine, School of Medicine, Emory University, Atlanta, GA 30322, USA
[3]Office for Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand
[4]Department of Microbiology and Immunology, School of Medicine, Emory University, Atlanta, GA 30322, USA
[5]Division of Pathology at Yerkes National Primate Research Center, Emory University, Atlanta, GA 30329, USA
[6]Department of Microbiology and Immunology, Medical College, National Cheng Kung University, Tainan 70101, Taiwan
[7]Center of Infectious Disease and Signaling Research, Medical College, National Cheng Kung University, Tainan 70101, Taiwan
[8]Insitute of Molecular Biosciences, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand
[9]Department of Pediatrics, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand
*Correspondence: bpulend@emory.edu

http://dx.doi.org/10.1016/j.chom.2014.06.001

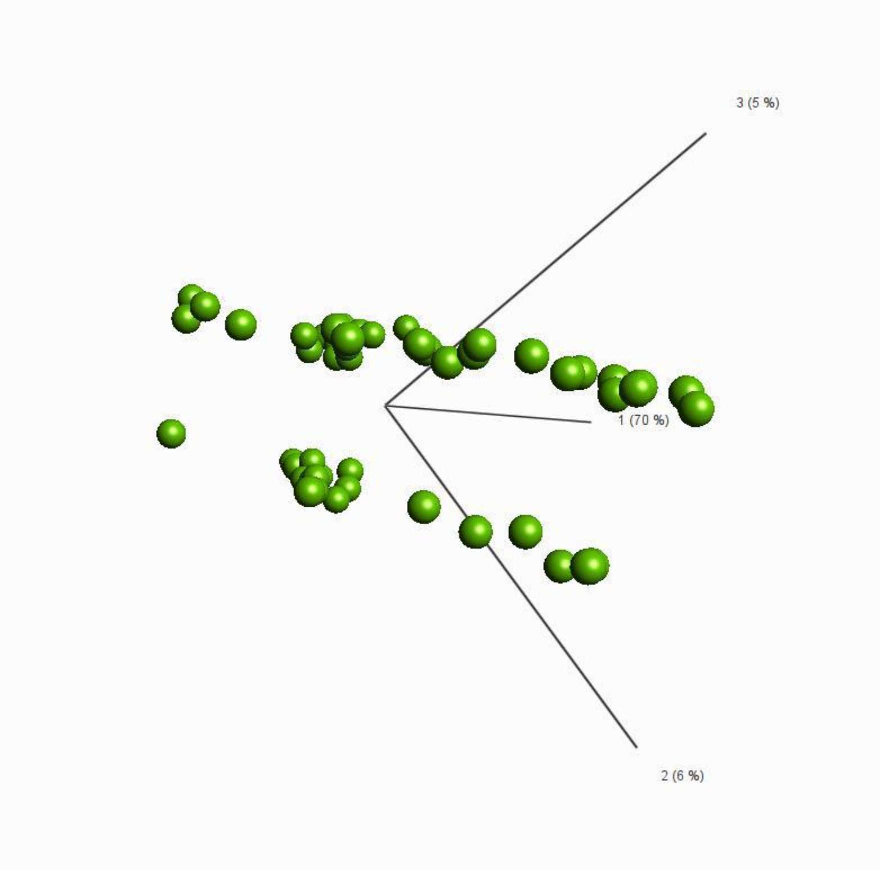
Load the data set from GEO by selecting the direct download option of Data Sets in the File menu in QOE.

ANALYSIS

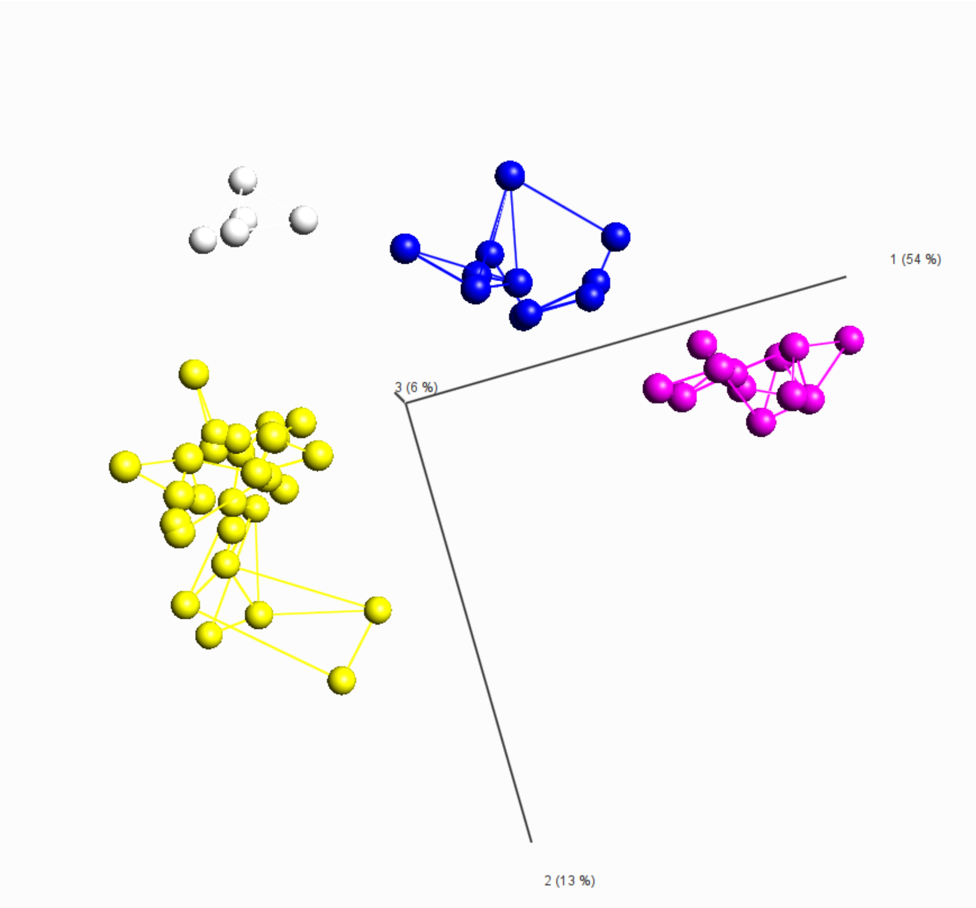
Start the analysis by following the steps below

ANALYSIS STEPS

- Filter on variance and maximize projection score. Max is 0,59 and there will be around 383 variables left. (Statistics dialog)
- Note the two groups. See image 1 below. Color according to the included annotations and observe that the groups are not explained by any of the annotations.
- Create a new sample annotation and classify the two groups
- Remove the variance filtering and perform a “two group test” to identify the variables that best explains the two groups. Filter until there are 40 variables left.
- Look at variable list and the chromosome annotation. It is clear that the variables are all associated with either X or Y chromosome and the conclusion is that the two observed groups are related to gender. (Variable tab)
- Since gender is a factor that, at this stage at least, is not relevant for the next steps use the inbuilt functionality to remove eliminate this factor. (Statistics dialog and Eliminated Factor)

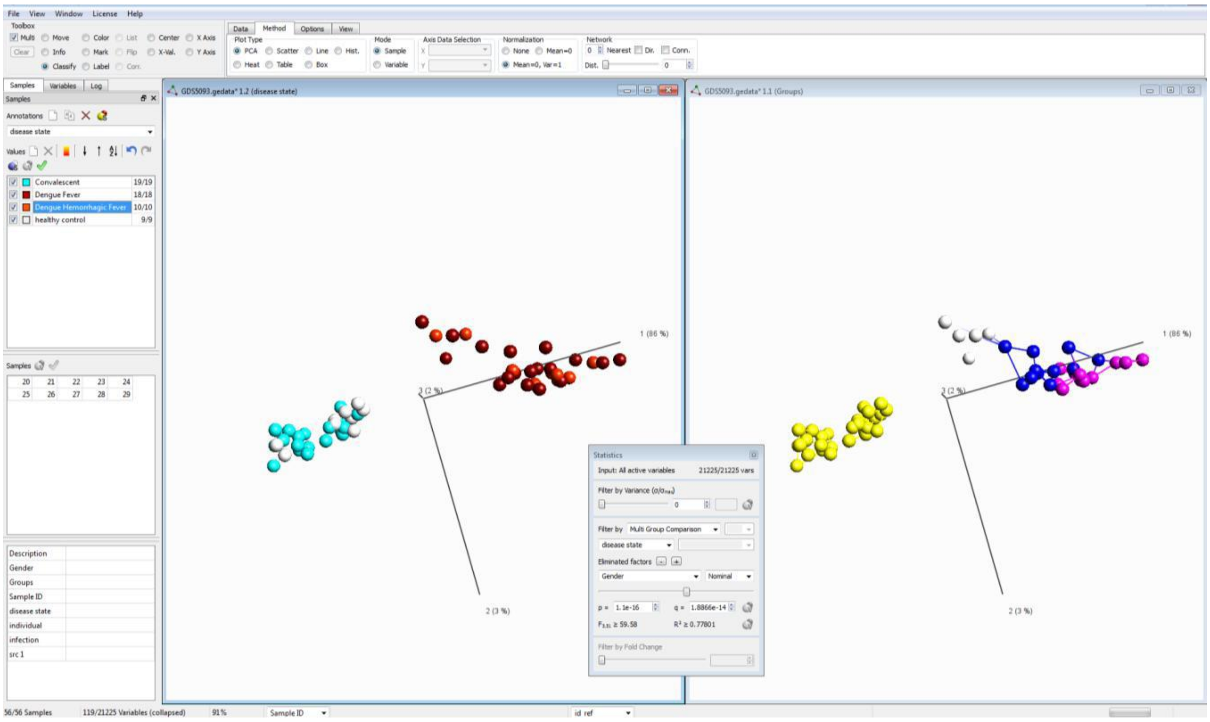


- Change to gene symbol and collapse. This is not required but shows how easy it is to work with different annotations as the base for the data. (Data tab)
- Use the variance filter and maximize the projection score again.
- Add the links to the two nearest neighbors. (Method tab)
- There are clearly 4 groups that can be identified in the PCA plot. See image 2.
- Classify the groups and color the sample PCA plot according to this annotation.



- Open a second synchronized sample PCA plot.
- Color the samples according to the annotation “disease state”.
- Change the settings in the statistics dialog so that there is no variance filtering and instead perform a “multi group comparison” in disease state. **This represents the type of test that many would do as the first approach.** Note in Image 3 how you find statistically relevant results with 119 variables with a qvalue less than 1,8 e-14. In the PCA plot you can see good separation between Dengue fever and control and convalescent.
- **IMPORTANT:** You are however not observing the interesting split of the Dengue group into three subgroups that we identified earlier and you can see in the right part of Image 3. If the analysis would have started with the last steps above (which is a common approach) the potential new discovery that there are multiple subgroups in the Dengue fever samples would have been missed.

This is a very typical example on how you can use QOE to identify unknown subgroups and also generate new hypothesis.



GDS5093.gedata* 1.2 (disease state)

GDS5093.gedata* 1.1 (Groups)

Plot Type: ● PCA ○ Scatter ○ Line ○ Hist. ○ Heat ○ Table ○ Box

Mode: ● Sample ○ Variable

Axis Data Selection: X Y

Normalization: ○ None ○ Mean=0 ● Mean=0, Var=1

Network: 0 Nearest □ Dir. □ Conn. Dist. 0

disease state

☑	Convalescent	19/19
☑	Dengue Fever	18/18
☑	Dengue Hemorrhagic Fever	10/10
☑	healthy control	9/9

Input: All active variables 21225/21225 vars

Filter by Variance (σ/σ_{max}) 0

Filter by Multi Group Comparison

disease state

Eliminated factors

Gender Nominal

p = 1.1e-16 q = 1.8866e-14

$F_{3,51} \geq 59.58$ $R^2 \geq 0.77801$

56/56 Samples 119/21225 Variables (collapsed) 91% Sample ID id ref

DISCLAIMER

The contents of this document are subject to revision without notice due to continuous progress in methodology, design, and manufacturing. Qlucore shall have no liability for any error or damages of any kind resulting from the use of this document. Qlucore Omics Explorer is only intended for research purposes.