QLUCORE

Analysis in an instant

Plots in focus: PCA and Heatmap uses, configurations and settings for OMICs experiments

Yana Stackpole, PhD

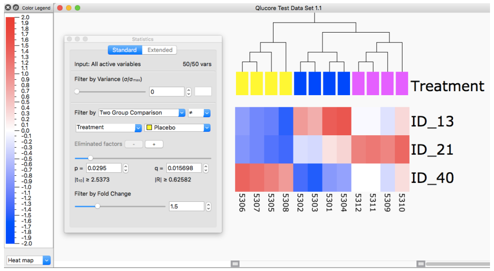
Application Support

Visualize and Explore

- QC (outliers, mislabeled samples)
- Make observations - identify structures, patterns
- Generate new hypotheses
- Browse the genome

Analysis

- t-test, ANOVA, Regressions, R scripts. Open API to R, Batch exec
- Variant calling
- Easy generation and export of reports, status, and plots
- Save your session, share



Input: All active variables 50/50 vars

Filter by Variance (σ/σ$_{max}$) 0

Filter by Two Group Comparison ≠

Treatment Placebo

Eliminated factors - +

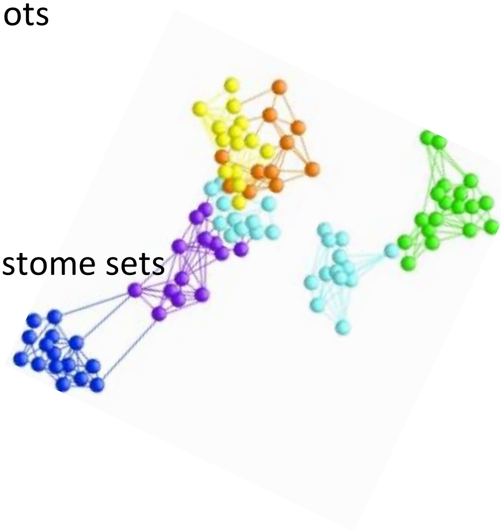
p = 0.0295 q = 0.015698

|t$_{10}$| ≥ 2.5373 |R| ≥ 0.62582

Filter by Fold Change 1.5

Biological Insight

- Gene Set Enrichment using MSigDB, reactome, or custome sets
- Explore annotations
- GO Browser



Classify and Predict

- Build classifiers
- kNN, SVM, RT
- Predict sample class, outcome, etc.

QLUCORE

Enjoy fantastic computing speed on a laptop to boost your discovery and scientific creativity

Benchmark examples (static). Compared to R	Times faster
ANOVA (22k var. + 130 samples)	2800
t-test (two-groups, selected from 22k var. + 130 samples)	1000
Kruskal-Wallis (22k var. + 130 samples)	900
Mann-Whitney U-test (two groups, 30k var. + 5k samples)	480
ANOVA (30k var. + 5000 samples)	180
PCA calculations (30k var. + 150 samples)	77
UMAP (22k var. + 130 samples)	13

The speed enables a more flexible workflow – generating better results faster.

Details at: https://qlucore.com/calculation-benchmarks

"This tool might literally save you years of your life"

Prof. Ulrich Steidl at Albert Einstein College of Medicine

PCA

What we see – samples OR variables in 2D or 3D separately. Biplots – synchronized sample and variable plot (or in one plot). Distance between samples reflects the degree of similarity.

Dimensionality reduction – YES, we plot artificial dimensions=components
Process of creating PC helps structures to be revealed by reducing noise.

Transformation - Threshold and Log2

Normalization

- **mean 0** - covariance based PCA
- **mean 0 and variance 1** - correlation based PCA (default). More informative to assess relationship between random variables
- **none**

Missing value vs Zero – keep in mind, big difference for how algorithms treat them. If unsure – check with your core. PCA cannot be drawn with holes (missing values), hence – reconstruction.

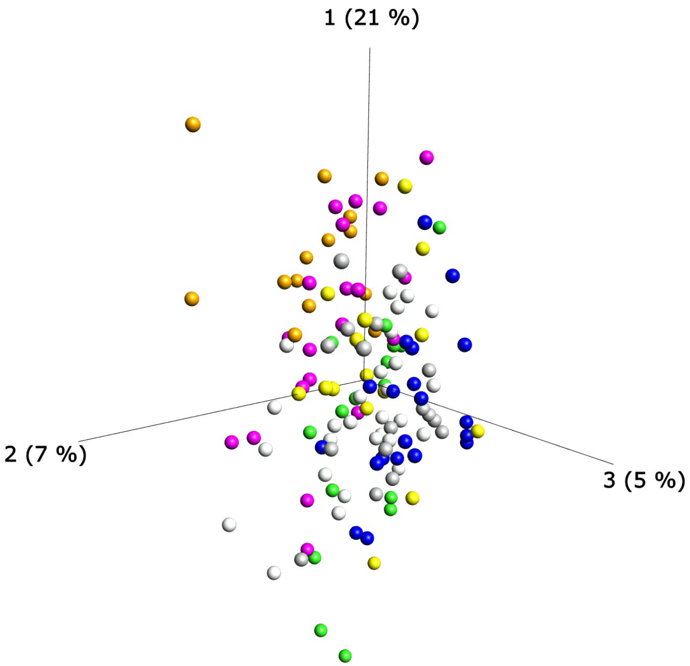
PCA - Visualization

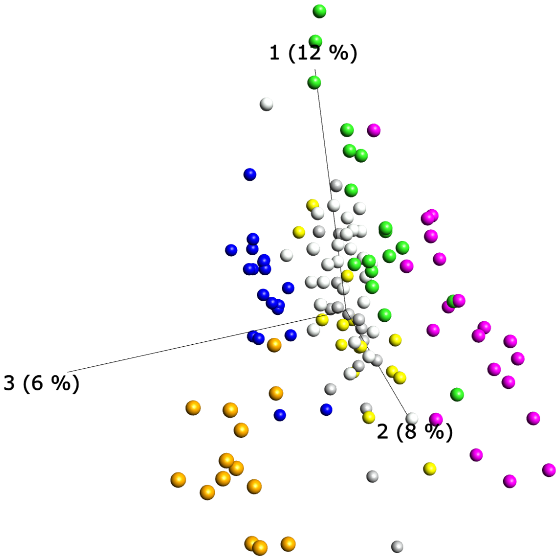
What we control:

- Sample and Variable input
- Normalization (z score), thresholding, Log
- PC chosen to be plotted
- Color of samples/variables
- Shapes of samples
- Labels
- PCA is heavily affected by outlier samples - exclude

Manipulation or Not?

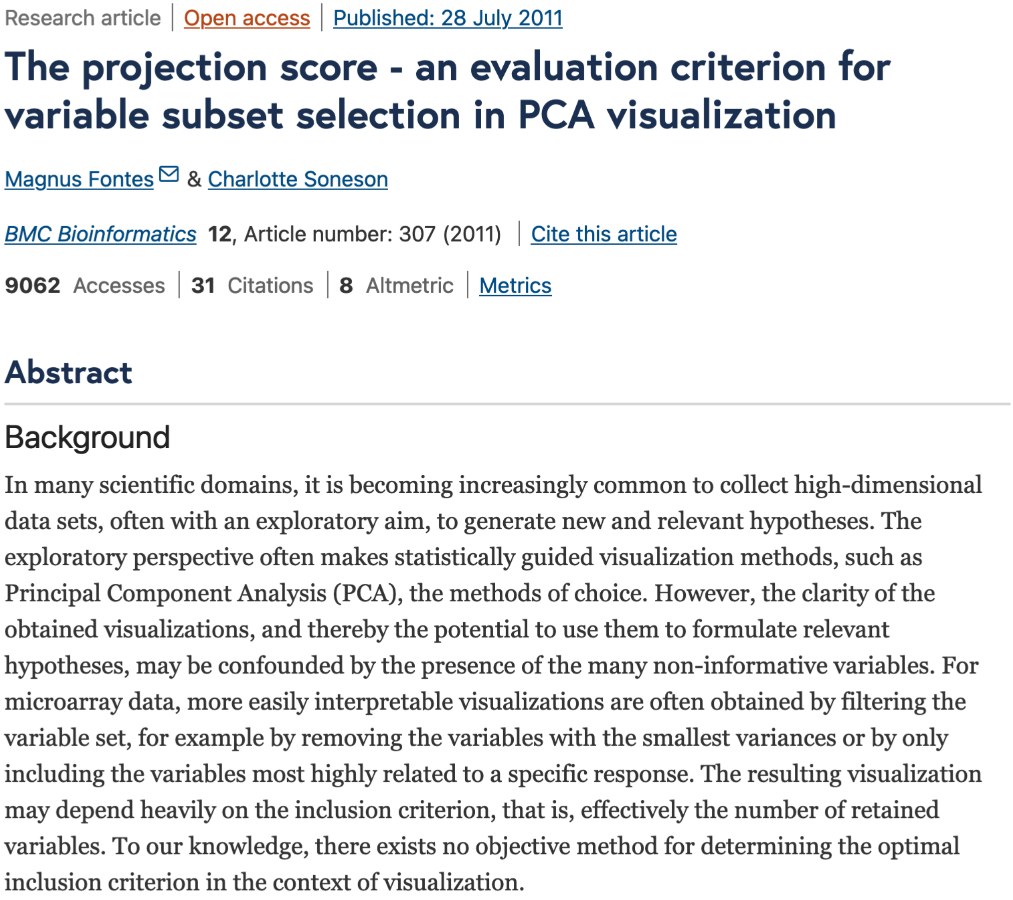
- Unjustified sample or sample group exclusion to enhance pattern
- Picking variable input based on the pattern visualization only
- Discovery when we explore patterns from multiple variable subsets, uses QC like stability of a pattern, like information (=variance captured) loss, etc





QLUCORE

Qlucore offers a tool you can use as a guidance for variance filtering when using PCA – Projection Score



Research article | Open access | Published: 28 July 2011

The projection score - an evaluation criterion for variable subset selection in PCA visualization

Magnus Fontes & Charlotte Soneson

BMC Bioinformatics **12**, Article number: 307 (2011) | Cite this article

9062 Accesses | **31** Citations | **8** Altmetric | Metrics

Abstract

Background

In many scientific domains, it is becoming increasingly common to collect high-dimensional data sets, often with an exploratory aim, to generate new and relevant hypotheses. The exploratory perspective often makes statistically guided visualization methods, such as Principal Component Analysis (PCA), the methods of choice. However, the clarity of the obtained visualizations, and thereby the potential to use them to formulate relevant hypotheses, may be confounded by the presence of the many non-informative variables. For microarray data, more easily interpretable visualizations are often obtained by filtering the variable set, for example by removing the variables with the smallest variances or by only including the variables most highly related to a specific response. The resulting visualization may depend heavily on the inclusion criterion, that is, effectively the number of retained variables. To our knowledge, there exists no objective method for determining the optimal inclusion criterion in the context of visualization.

Heatmaps

What we see – samples AND variables in 2D, in one plot. Measurements are represented by colors according to the normalization choice (artificial scale).

Dimensionality reduction – None

Data set can contain unnecessary complexity which may hide structures.

Transformation - Threshold and Log2

Normalization

- **mean 0, mean 0 and variance 1,** or **none.**
- use **None** for raw values on a heatmap

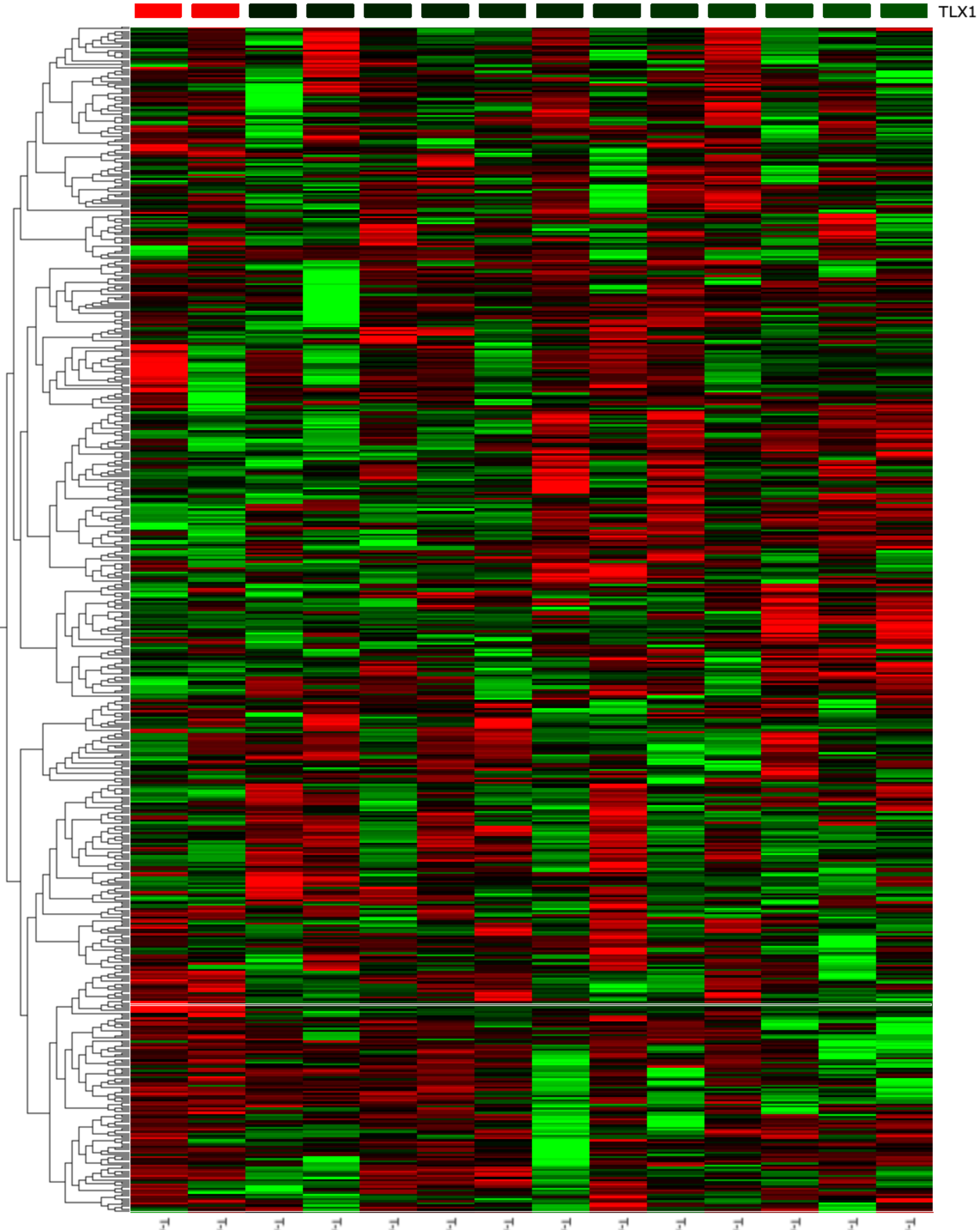
Missing value vs Zero – keep in mind, big difference for how algorithms treat them. If unsure – check with your core.

Heatmaps - Visualization

What we control:

- Sample and Variable input
- Normalization, transformation
- Color scale
- Labels
- With z score, If you look at the color legend, you will see that it goes from -2.0 to 2.0 which means the scale goes from two standard deviations less than the mean to two standard deviations greater than the mean.
- Ordering of Samples and variables
- Hierarchical Clustering – reflects a degree of similarity. User can flip branches for better pattern representation (save your session to preserve exact config.)

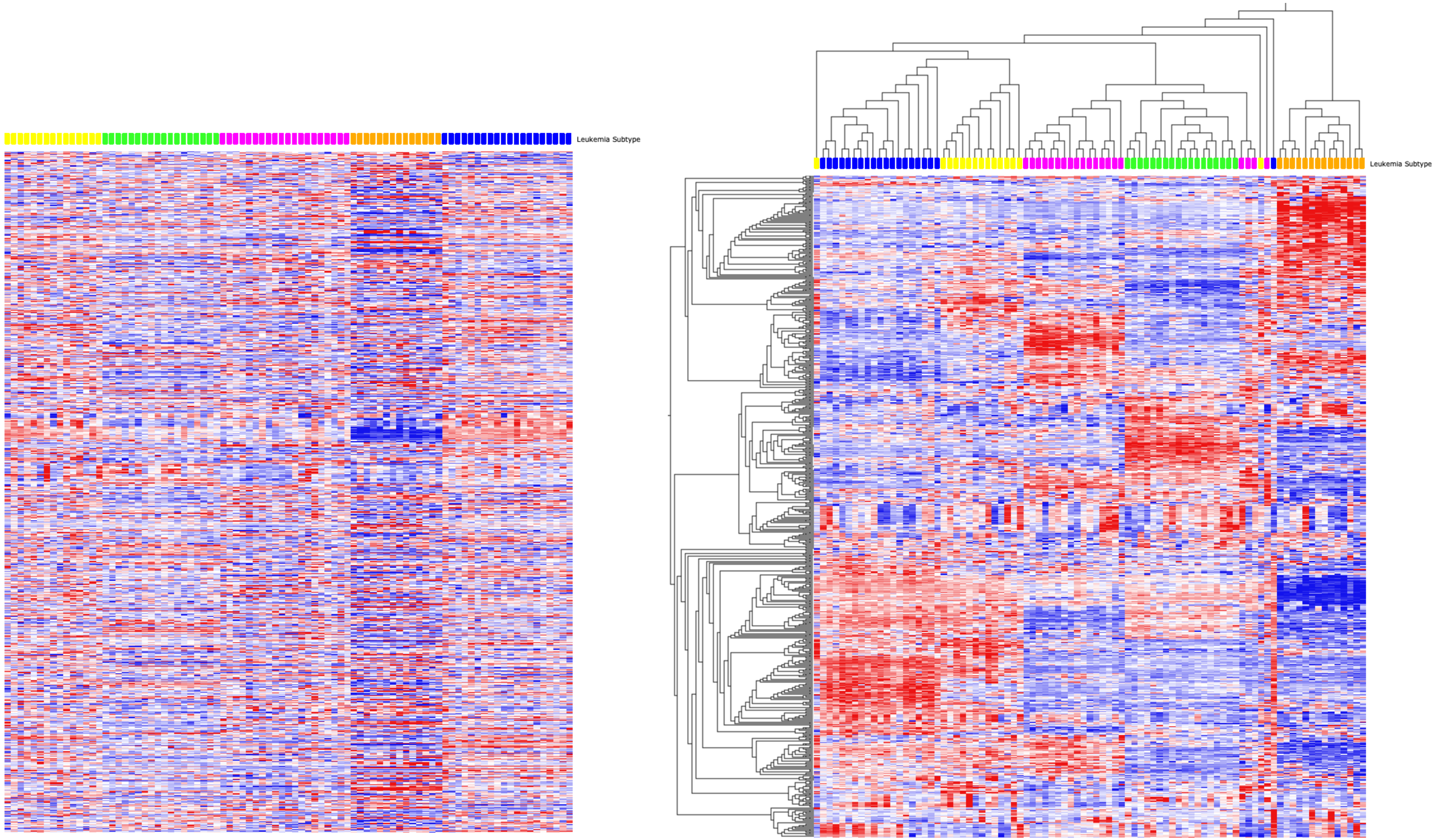
Sample stratification -- marker expression



As with other leukaemias, understanding the heterogeneity within T-ALL holds clinical significance as T-ALL subtypes may influence risk stratification and treatment decisions. For instance, **TLX1** and TLX3 **lesions indicate good prognosis**, TAL1/2 lesions indicate intermediate prognosis, and the rare SPI rearrangement subtype indicates very poor prognosis5.

Manipulation or Not?

Hierarchical clustering for a heatmap is unsupervised method – no user input into the algorithm (Kmeans is supervised, class-based machine learning is supervised)



The same heatmap, ordered in different ways. Hierarchical clustering reveals data patterns

PCA vs Heatmap

We hear that Qlucore heatmaps are the most good looking out there ☺
Widely used in publications, including many high impact journals

PCA (sample + variable)	Heatmap with samples and variables clustered hierarchically
Unsupervised – fit for Exploratory Discovery	Unsupervised - fit for Exploratory Discovery
Dimensionality reduction method. Pre-processing creates artificial dimensions (PCs) working with variance. Weakest signal, lowest variance (=noise) are removed.	Does not filter data - no pre-processing
Does not partition data - random cloud if no noticeable signal. May reveal structures linked to high variance, but not aiming on that specifically	Partitions data creating artificial objects - clusters by degrees of similarity/dissimilarity (will cluster even when no strong signal present)
The most dominant patterns (captured by the first PCs), are those separating different subgroups of the samples from each other. In this case, the results from PCA and hierarchical clustering support similar interpretations. PCA patterns are cleaner, at a cost of some lost info (=variance).	
Works with Variance Visualization	Works with Degrees of Similarity Both Visualization and Partitioning

Next steps

1. Access your Qlucore site access – contact BTEP
2. Book a session with us – get help importing your data and setting up your analysis. Here is a link https://calendly.com/yana-stackpole/30min

3. Explore your data in a cool GUI in a way that makes sense to you!

Global support team Support@Qlucore.com
Local support Yana.Stackpole@Qlucore.com

System requirements
Base module

FOR WINDOWS

- Windows 7, Windows 8 or Windows 10
- 512 MB of RAM memory
- A graphical card with support of at least Open GL 2.1
- 5 GB of free hard disk space
- Qlucore Omics Explorer is available in both a 32-bit and 64-bit versions. The program takes full advantage of processors with multiple cores and computers with multiple processors.

FOR MAC

- Max OS X 10.15 or 10.14
- 512 MB of RAM memory
- A graphical card with support of at least Open GL 2.1
- 5 GB of free hard disk space