R-based Pipeline for the Search, Import and QC/QA of Public Microarray Data

Joana Corte-Real, Petr V. Nazarov, Arnaud Muller, Tony Kaoma, Laurent Vallar

http://sablab.net/PuMaQC

Microarray Center, CRP-Santé, 84, Val Fleuri, L-1526, Luxembourg

Introduction

Importing microarray data from public repository databases such as NCBI Gene Expression Omnibus (GEO), is a simple and cost free solution for **data-driven studies** such as inference of gene regulatory networks and translational cancer research requiring **large amounts of transcriptomic information.** Handling public data can be time consuming and not always straightforward. Also it is important to assure that data is not of bad quality since it may compromise statistical and biological significance in downstream analysis[1]. Taking this into account we created PuMaQC tool.

Here we aimed at creating **an easy to use all-in-one pipeline** incorporating the different stages important for handling public Affymetrix microarray data including quality control, that allows saving time and is accessible to both biologists and bioinformaticians.

Tool

We propose a **free and simple-to-use tool PuMaQC** to work with public microarray data (Affymetrix) from the earlier stage of searching for samples in public databases up to quality control and quality assessment (QC/QA).

The pipeline is divided into several steps described below.

PuMaQC Pipeline

INI-file: text description file*

```
[Project]
  Title = Lung project
  Description = Look for healthy lung microarrays
[Search]
  GEOmetadb = ../GEOmetadb.sqlite
  Platform = HG-U133_Plus_2
  Organism = Homo sapiens
  Include = lung, bronchus, respiratory
  Exclude = cancer, carcinoma, tumor, embryo
  Results = selectedGSM.txt
[Import]
  DownloadTo = download
  CDF = N:/Data/_CDF/HG-U133_Plus_2.cdf
  APT = C:/APT-1.14.2/Bin
  Results = importedGSM.txt
[QC]
  Alpha = 0.05
  arrayQualityMetrics = TRUE
  Results = dataGoodQuality.txt
```

*) Format is similar to the one of standard Windows INI files. "[...]" define sections, variables are followed by "=". Paths can be either relative to INI file or absolute. Comments start from ";" or "#" character.

Database Query
- Search GEO database by *include* and *exclude* keywords.
- Perform custom multiple search

Manual Curation
- **Optional** manual curation of the selected list of GSMs

Data Download
- Download and unpack files
- Assign CEL file formats

Preprocessing
- Convert CEL formats
- Summarize probesets and normalize data (by RMA[2])

Quality Control
- Compute quality metrics
- Quality assesment
- Detect and remove outliers

Data in text format, ready for downstream analysis

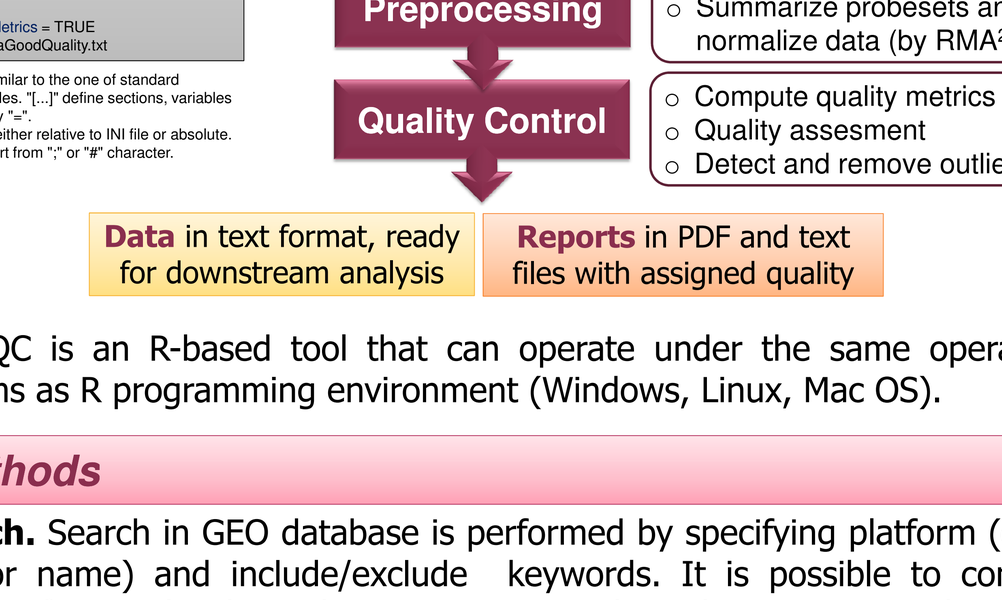
Reports in PDF and text files with assigned quality

PuMaQC is an R-based tool that can operate under the same operating systems as R programming environment (Windows, Linux, Mac OS).

Methods

Search. Search in GEO database is performed by specifying platform (either GPL or name) and include/exclude keywords. It is possible to combine results of several independent queries using logical operators: and, or, not, e.g `Expression = ( 1 and 2 ) or ( 1 and 3 )`, where 1,2,3 are names of the queries, for which *Include1*, *Include2* and *Include3* parameters are specified in the INI-file (*Exclude1-3* are optional).

QC/QA. Four quality metrics are currently used for QC/QA:



Alternatively QC/QA can be performed by a standard R package ***arrayQualityMetrics*** [1] (set corresponding parameter to TRUE)

Getting Started

PuMaQC can be run under Windows, Linux or MacOS. Prerequisites:

- R/Bioconductor
- Affymetrix Power Tools and proper CDF (see www.affymetrix.com)
- User-defined a configuration file (an INI-file)

PuMaQC can be run simply by typing 2 lines in R console:

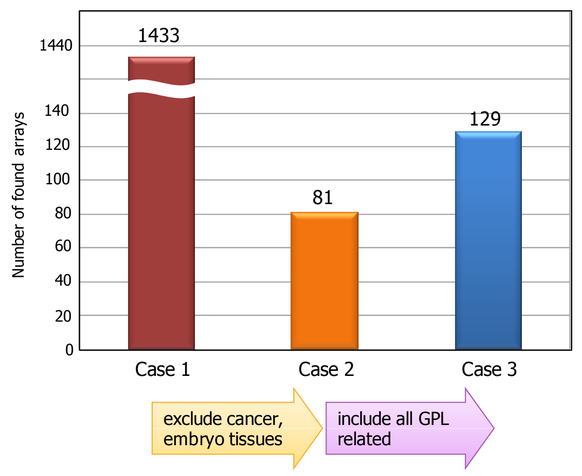
```
source("http://sablab.net/PuMaQC/PuMaQC.r")
PuMaQC()
```

The current version of PuMaQC, along with manual, INI-flie template and all R-scripts are freely available at http://sablab.net/PuMaQC

Results

To test our tool, we applied the pipeline to look for ***human lung*** samples. 3 possible search cases were considered:

	GEO Platform	Cancer studies
Case 1	GPL570	Included
Case 2	GPL570	Filtered out
Case 3	All related	Filtered out

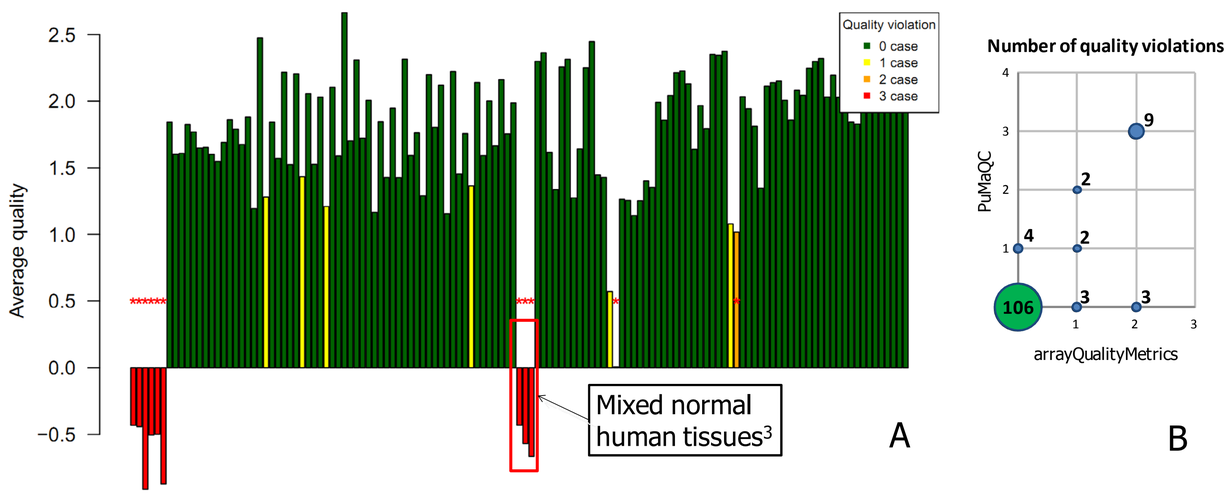


Case 1. Here we queried the database for human lung samples, annotated as standard GPL570.

Case 2. To show power of "exclude" keywords we reduced number of samples excluding cancer, cell lines and embryonic related samples.

Case 3. In GEO many platforms (GPL) can correspond to the same array. By specifying Affymetrix array name (e.g. HG-U133_Plus_2) we expand our search to other related GPLs and increase the odds of finding samples.

Resulting quality of the arrays is summarized below:



Array quality summarization for Case 3. (A) Bars represent the average quality score for each array, based on computed quality metrics. Bar colors represents the number of metrics that do not meet the defined quality threshold (i.e. Quality violation).
(B) Comparison of PuMaQC and arrayQualityMetrics quality assessment. Quality problems are detected by both methods for 13 arrays. 4 arrays are found problematic uniquely by PuMaQC and 6 – uniquely by arrayQualityMetrics.

Conclusions

➢ PuMaQC search method allows highly customizable querying GEO database for microarray samples in simple and effective way. *Exclude* keywords allow focusing on relevant samples while filtering out undesired studies or samples.

➢ Automatic data download and processing significantly speeds-up data analysis.

➢ QC/QA methods used are sensitive to both technical and biological variability and give similar results to *arrayQualityMetrics*, however our calculations are less memory and computationally intensive.

References

1. Kauffmann, A. and W. Huber. (2010) Genomics 95(3): 138-42
2. Gentleman R., et al. Bioinformatics and computational biology solutions using R and Bioconductor. Springer 2005.
3. GSM304263, GSM304264, GSM304265 from GEO series GSE12034