PuMaQC: R-based pipeline for the search, import and QC/QA of public microarray data

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

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

[*] ***Corresponding author:*** *petr.nazarov@crp-sante.lu*

BACKGROUND

Data-driven studies such as inference of gene regulatory networks and translational cancer research normally require large amounts of transcriptomic data. One simple and cost free solution comes from importing microarray data from public repository databases such as NCBI Gene Expression Omnibus (GEO), integrating hundreds of thousand experiments. Despite the existence of the MIAME guidelines for standard microarray information, there is still a lack of information related to the quality of submitted data. Given that low quality samples can add noise and impair the statistical and biological significance of microarray analysis, quality control and quality assessment (QC/QA) becomes an important step when using public microarray data. Taking this into account we have developed R-based PuMaQC (**Pu**blic **M**icro**a**rray **Q**uality **C**ontrol) pipeline.

METHODS

PuMaQC is a robust, easy to use, all-in-one pipeline for public microarray data handling based on 3 sequential steps: i) search for raw Affymetrix data in GEO, ii) import and preprocessing of CEL files; and iii) QC/QA with identification and removal of low quality arrays.

The pipeline incorporates functions from *GEOmetadb*, *GEOquery*, *arrayQualityMetrics* R/Bioconductor packages and uses *Affymetrix Power Tools (APT)* for raw data extraction and normalization. We have included the possibility to filter out unwanted samples at step (i), and a Gpl-platform dictionary that allows broadening sample search to several related GEO platforms (Gpl).

RESULTS

To test PuMaQC we have applied it to 3 possible cases when searching for healthy human lung samples generated with Affymetrix HG-U133plus2 chips:

1. All lung-related samples existing for GPL570.
2. Similar to case 1, but filtering out cancer- and embryo-related samples.
3. Similar to case 2 but broadening search to all Gpl related to HG-U133plus2 chips.

The search for human lung related samples returned a total of 1370 found GEO samples (Gsm) (Case 1). By filtering out cancer related samples we were able to exclude 1313 unfitting samples, leaving a total of 57 arrays (Case 2), hence avoiding and exhaustive manual curation of query results. The incorporation of a Gpl-platform translation dictionary (Case 3) doubled the number of found arrays (105 in case 3).

CONCLUSIONS

PuMaQC pipeline allows for performing effective search, import and QC of public Affymetrix microarray data, with identification and removal of outliers. PuMaQc is a simple-to-use, relatively fast, but powerful tool, which makes it attractable to both bioinformaticians and biologists. See http://sablab.net/PuMaQC for the details.