Meta-analysis of microarray raw data using a virtual integrated platform

Bisognin A[1], Mazza E[2], Ferrari F[1], Forcato M[2], Bicciato S[2], Bortoluzzi S[1]

Motivation

Publicly available datasets of microarray gene expression profiles represent an unprecedented opportunity to extract genomic relevant information and to validate biological hypotheses without the need of novel experiments. However, their exploitation is still limited by cross-platform comparison and chip annotation issues. To facilitate the meta-analysis of Affymetrix data contained in Gene Expression Omnibus (GEO), we recently developed A-MADMAN, a web application that allows the retrieval, annotation, organization and analysis of public available gene expression datasets (Bisognin et al., 2009). In its original version, A-MADMAN addressed the meta-analysis of multiple experiments adopting custom chip definition files (custom CDFs) and a meta-normalization strategy based on a quantile distribution transformation. Although effective, we improved A-MADMAN meta-normalization approach and developed a novel procedure to effectively cope with platformspecific signals and distortions. Based on the construction of a virtual platform, this approach integrates probe intensities from different generation of Affymetrix arrays before the quantification of the signal, thus allowing the application of standard signal generation and normalization procedures (e.g., RMA, GCRMA).

Methods

Raw expression data (i.e., CEL files) obtained from at least two different platforms are integrated using an approach inspired by the generation of custom CDFs. In custom CDFs, probes matching the same transcript, but belonging to different probes sets, are aggregated into putative custom-probe sets, each one including only those probes with a unique and exclusive correspondence with a single transcript. Similarly, probes matching the same transcript but located at different coordinates on different type of arrays may be merged in custom-probe sets and arranged in a virtual platform grid. As for any other microarray geometry, this virtual grid may be used as a reference to create i) the virtual-CDF file, containing the probes, shared among the platforms of interest, and their coordinates on the virtual platform, and ii) the virtual-CEL files containing the intensity data of the original CEL files properly re-mapped on the virtual grid. Once defined the virtual platform through the creation of its custom-CDF and transformed the CEL files into virtual-

[1] Department of Biology, University of Padova, Padova, Italy [2] Department of Biomedical Sciences, Center for Genomic Research, University of Modena and Reggio Emilia, Modena, Italy

CELs, raw data, originally obtained from different platform, are homogeneous in terms of platform and can be preprocessed and normalized adopting standard approaches, as RMA or GCRMA.

Results

The data integration procedure was tested on benchmark datasets with different combinations of pre-processing steps to minimize platform related biases as well as false positive and false negative results in detecting differential expression. A dataset, including HG-U133A and HG-U133Plus2.0 arrays hybridized with Stratagene Universal Human RNA, was used as baseline case to assess the false positives due to mixedplatforms effect. In addition, a dataset of different tumor subgroups, hybridized to both HG-U133A and HG-U133Plus2.0, was analyzed to evaluate the sensitivity of the integration procedure. The virtual platform was finally applied to generate a gene expression data matrix from 30 GEO series comprising 411 samples of monocytes, macrophages and dendritic cells collected from a broad range of experimental conditions and hybridized on different Affymetrix arrays. The results of this large meta-dataset confirmed that samples integrated and normalized with the virtual platform show minimal platform-related biases and are correctly classified according to cellular type and physiological state.

Contact e-mail

silvio.bicciato@unimore.it

Supplementary information

References

[1] Bisognin, A, Coppe, A, Ferrari, F, Risso, D, Romualdi, C, Bicciato, S, Bortoluzzi, S. A-MADMAN: annotation-based microarray data meta-analysis tool. BMC Bioinformatics. 2009 Jun 29; 10:201.