A Warehouse of Quantified and Annotated Affymetrix Microarrays

Allen Day
Department of Human Genetics,
Geffen School of Medicine,
University of California, Los Angeles

May 1, 2006

We present a data warehouse consisting of more than 31,000 quantified and ontologically annotated Affymetrix microarrays. This warehouse continously imports primary gene expression and genotype data in the form of CEL files from eight major microarray repositories. CEL files are processed using several accepted transformation algorithms (e.g. RMA, GC-RMA), then manually and computationally annotated using a series of orthogonal, standard biomedical ontologies and finally, exposed through a DAS 2.0 web service. This service contains a powerful ontology-centric query API for selection subsets of the data based on sample annotation, taxonomy, and hybridization platform.

Affymetrix microarrays are inherently more reusable than alternate platforms, attributable to the technical consistency and independence from confounding competitive hybridization effects of multi-channel platforms. Yet, the reusability of published Affymetrix data has not yet realized its full potential. The main blockade in the creation of this warehouse, and thus microarray reuse, is the difficulty of locating primary data. Our survey of eight large public repositories reveals extensive fragmentation: less than 4% of primary data is available from any combination of two repositories. Data reuse is further hampered by the ineffective encoding of clinical and experimental metadata associated with the biological material hybridized to each array. These metadata are weakly typed, ambiguous, and often incomplete, and thus not amenable to structured search.

Methodology is presented for the efficient import, storage, retrieval, structured annotation and search of more than ten billion microaray data points. Novel biological findings include insights into the relationships between biological samples of differing sex, pathological state, and tissue/cell lineage. These findings were only possible through the use of our database, which is the largest repository of consistently annotated and quantified microarrays available.