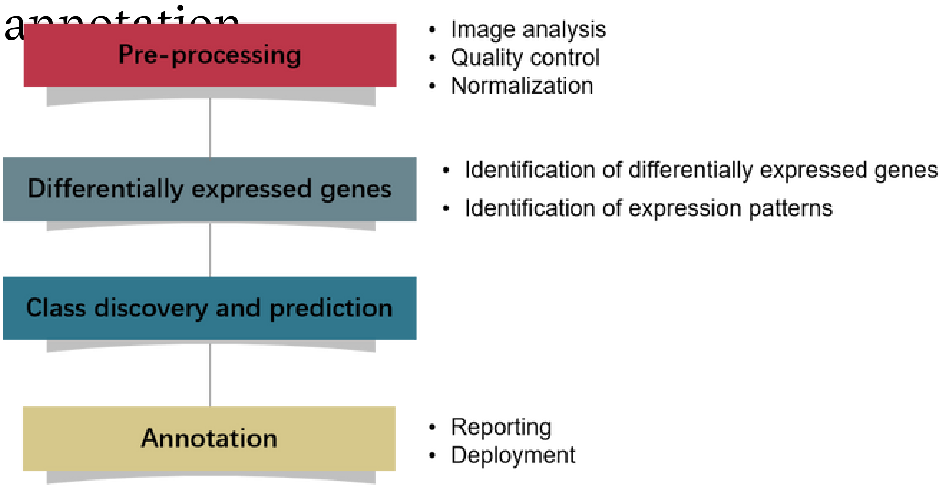
DNA microarray data analysis

Introduction:

What is a microarray?

Microarray is a technology which enables researchers to analyse and address issues that were thought to be non-traceable by facilitating the simultaneous measurement of the expression levels of thousands of genes. The microarray is a glass slide where the DNA molecules are fixed at specific locations in order and it's called spots or probes. The DNA (or RNA) in a spot may either be a complete copy of genomic DNA or a short stretch of oligonucleotides that correspond to a gene. Microarrays can be used in many types of experiments, including genotyping, epigenetics, translation profiling,gene expression profiling and they may be used in treatment and are the best candidates for specific treatments. The main goal of microarray data analysis involves the generation of raw expression data and determination of their biological significance. The typical process can be divided into the following step: pre-processing, differentially expressed genes analysis, class discovery and prediction, as well as annotation


Pre-processing
Image analysis
Quality control
Normalization
Differentially expressed genes
Identification of differentially expressed genes
Identification of expression patterns
Class discovery and prediction
Annotation
Reporting
Deployment

Pre-processing:

The “pre-processing” steps of analysis of array data are somewhat platform specific, and it includes image analysis, quality assessment, and normalization.

Image analysis

Different scanning settings can result in different images that may affect the experimental results. Generated pictures are then quantified using packages such as Imagene or GenePix. The intensities are generally measured as either mean or median pixel value in the given region.

Quality assessment

The quality assessment starts with visual examination of the images and plots of the raw data. Experienced researchers are able to tell which arrays in the set have inferior quality or whether some regions are unusual due to scratches, etc. Spatial plots can also help identify regions with unusual signals.

Normalization

An important part of data preprocessing is normalization, which adjusts the individual intensities so as to make comparisons both within an array and between arrays in the experiment. Examples of differences needed to be adjusted are unequal RNA quantities, differences in labeling, and systematic biases.

Microarray Data Analysis

Microarray data sets are commonly very large, and analytical precision is influenced by a number of variables. So it is extremely useful to reduce the dataset to those genes that are best distinguished between the two cases or classes (e.g. normal vs. diseased). Such analyses produce a list of genes whose expression is considered to change and known as differentially expressed genes. Identification of differential gene expression is the first task of an in depth microarray analysis.

There are two common methods for in depth microarray data analysis, i.e. **clustering** and **classification**

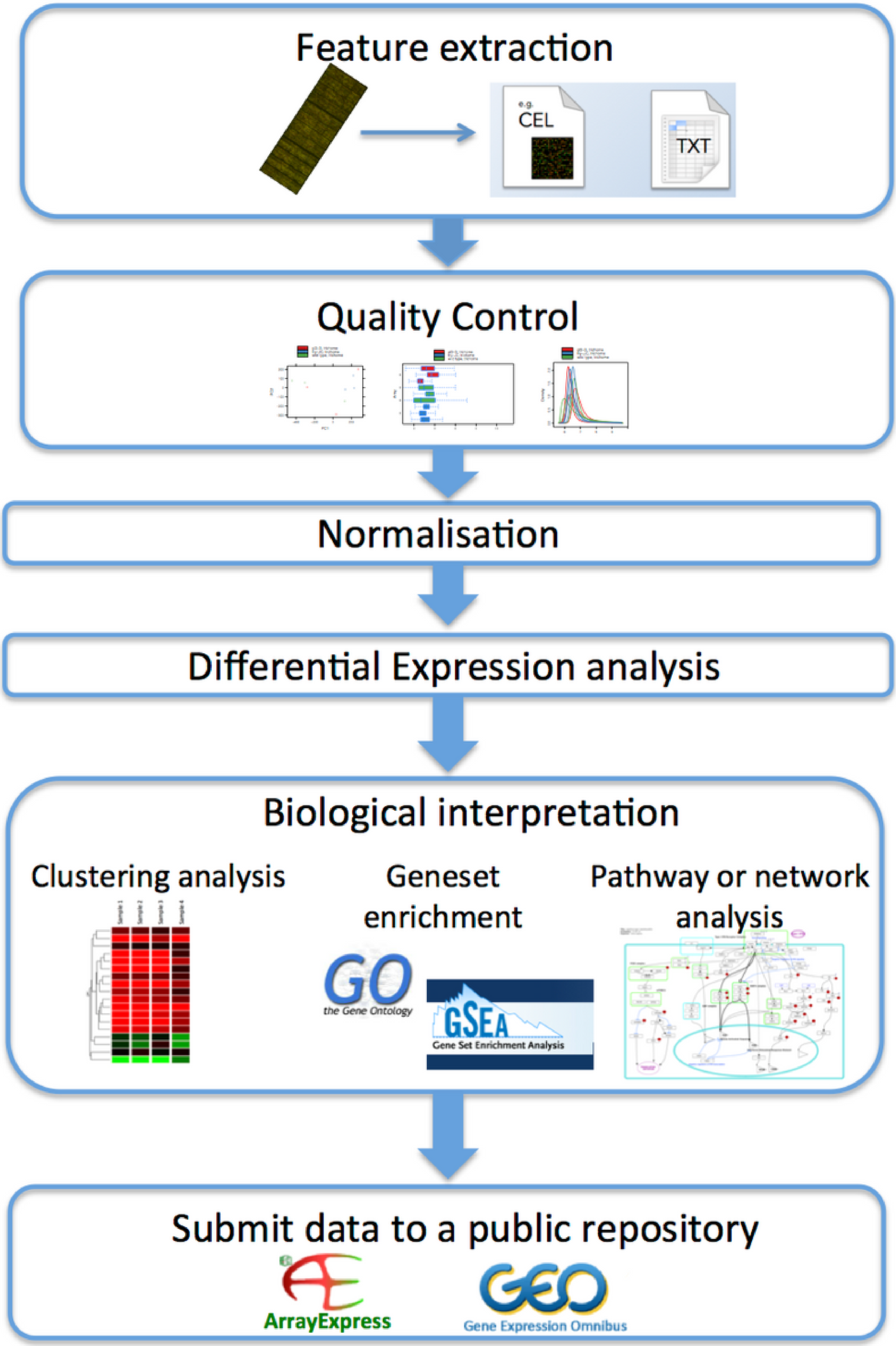
Cluster Analysis

Clustering is the most popular method currently used in the first step of gene expression data matrix analysis. It is used for finding co-regulated and functionally related groups. Clustering is particularly interesting in cases when we have complete sets of an organism's genes. There are three common types of clustering methods (i.e.) hierarchical clustering, k-means clustering and self-organizing maps.

Hierarchical clustering: is a commonly used unsupervised technique that builds clusters of genes with similar patterns of expression .

K-means clustering: is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships.

A self-organizing map (SOM): is a neural network based non-hierarchal clustering approach. (SOMs) work in a manner similar to K-means clustering.


Feature extraction
e.g.
CEL
TXT
Quality Control
Normalisation
Differential Expression analysis
Biological interpretation
Clustering analysis
Sample 1
Sample 2
Sample 3
Sample 4
Geneset enrichment
GO
the Gene Ontology
GSEA
Gene Set Enrichment Analysis
Pathway or network analysis
Submit data to a public repository
ArrayExpress
GEO
Gene Expression Omnibus

Classification

Classification is also known as class prediction, discriminatory analysis, or supervised learning. Given a set of pre-classified examples, a classifier will find a rule that will allow them to assign new samples. For a classification task, one must have sufficient sample numbers to allow an algorithm to be trained, known as a training test, and then to have it tested on an independent set of samples known as a test set. Using normalized gene expression data as input vectors, classification rules can be built. There are a wide range of algorithms that can be used for classification, including k Nearest Neighbors (kNN), Artificial Neural Networks, weighted voting and support vector machines (SVM). The promising application of classification is in clinical diagnostics to find disease types and sub-types. The general data mining and machine learning application tools that are used for classification tasks are illustrated in the.

Classification, clustering and identification of differential genes can be considered as basic microarray data analysis tasks with gene expression profiles alone. However, Gene expression profiles can be linked to other external resources to make new discoveries and knowledge.

Conclusion

DNA microarrays are a revolutionary technology, and microarray experiments generate far more data than other technologies. However, innovative statistical techniques and computer software are critical to the successful analysis of microarray data. This overview shows current bioinformatics tools and promising applications for analyzing microarray experimental data.