Data curation for GENEVESTIGATOR®

Last updated: November 2015

The Problem

Public repositories like Gene Expression Omnibus (GEO) and ArrayExpress collect and make publicly available a large number of gene expression studies. Typically, experiments are submitted to these repositories by the authors themselves or by a lab technician. The submission process is loosely controlled and leads to content with highly variable quality. In particular, the following problems frequently occur:

- Imprecise descriptions, generally without use of controlled vocabularies
- Missing patient characteristics and clinical parameters
- Unordered/ungrouped samples
- Falsely labeled samples
- Redundant samples within and between studies
- Low statistical quality of array or sequencing data

Curation by NEBION

To create a high quality compendium of public experiments, NEBION biocurators quality control, normalize and manually annotate these studies after searching and assembling all the essential information needed to understand them. Our biocurators typically have a PhD and several years of lab/clinical experience.

NEBION has an established SOP for curation, including the following levels:

Quality control:

- Using various public and proprietary statistical analysis tools

Normalization:

- Global normalization across all experiments

Annotation:

- Experimental design, title, authors, and links to original repositories
- Annotate samples using controlled vocabularies (ontologies) searching for detailed information about genetic background, clinical parameters, patient characteristics, etc.
- Group samples into biological replicates
- Define group comparisons (e.g. diseased vs. healthy; treatment vs. control)
- Detect duplicates and merge multiple GEO/ArrayExpress experiments containing the same controls or duplicated samples
- Create and update ontologies using a versioning system

Verification:

- Each experiment is curated by a first curator and **peer-reviewed** independently by a second one
- Gender annotations are verified using markers
- Experiments and sample consistency are verified by visualizing the experiment and using reference genes

Criteria for choosing public experiments to curate

NEBION chooses to curate public experiments based on three criteria:

1. Customers' wishes. Companies having licensed GENEVESTIGATOR® can request the curation of public experiments for chosen diseases or therapeutic areas.

2. Priority areas defined by NEBION. Although we have curated data from over 500 diseases and 28 therapeutic areas, we have particularly enriched datasets for the following areas:
    a. Respiratory diseases (e.g. IPF and COPD)
    b. Cardiovascular diseases (e.g. ACS, atherosclerosis)
    c. Auto-immune diseases (e.g. rheumatology, diabetes)
    d. Neurology
    e. Oncology and immunology

3. Large reference datasets of general interest, e.g. Connectivity Map, TCGA, CCLE, ExpO, etc.

Examples illustrating the added value of the curation

Examples for the correction of annotation errors:

Example 1: MM-00431 / GSE25640: Expression data from wild type or FIZZ2 knockout murine lungs. The overall description of the experiment in GEO states that bleomycin treatment was performed for 21 days. By contrast, the corresponding published article states it was performed for 7 days.
=> NEBION curators clarified the situation with the authors, who confirmed that the information in GEO is wrong: the correct duration is 7 days.

Example 2: HS-01135 / GSE12385: Gene expression changes in Peripheral Blood Mononuclear cells (PBMC) induced by physical activity. In the part of matrix file from GEO, the same clinical values were annotated for samples before and after 24 weeks of physical exercise.
=> NEBION curators corrected this information based on the description provided in the corresponding published article.

Examples for the annotation enrichment by NEBION:

Experiment HS-01045
GSE52724 / GSE35713: Molecular signatures differentiate immune states in Type 1 Diabetes families / Transcriptional Signatures as a Disease-Specific and Predictive Inflammatory Biomarker for Type 1 Diabetes. Almost no information is available in GEO
=> 14 sample variables were annotated by NEBION curators based on two corresponding publications.

Experiment HS-01018
GSE41177: Region-specific gene expression profiles in left atria of patients with valvular atrial fibrillation. Almost no information is available in GEO. =>18 sample variables were annotated by NEBION curators.

Experiment HS-01147
GSE32512: Hyperglycemia and a Common Variant of GCKR Are Associated with the Levels of Eight Amino Acids in 9,371 Finnish Men
Almost no information was available in GEO.
=> 14 sample variables were annotated by NEBION curators based on two corresponding publications.