

ToxCastTM Workflow: Big Data Processing and Analysis of High-Throughput Screening Assays for Environmental Chemicals

PG Kothiya¹, DM Reif², RS Judson¹, DL Filer¹, AM Richard¹, MT Martin¹

1 National Center for Computational Toxicology, Office of Research and Development, US EPA, RTP, NC

2 Department of Genetics, North Carolina State University, Raleigh, NC.

US EPA's ToxCast program is generating data in high-throughput screening (HTS) and high-content screening (HCS) assays for thousands of environmental chemicals, for use in developing predictive toxicity models. Currently the ToxCast screening program includes over 1800 unique chemicals and over 700 bioassays. Data collection from multiple assay vendors and collaborators in diverse formats, data standardization, and normalization, coupled with the needs for data transparency and consistency have posed major technical and scientific challenges to the program. We developed the ToxCast data workflow to allow for consistent cross-technology analysis and data reporting. The ToxCast workflow consists of 8 levels: "Data Processing/Normalization" (levels 1-4) and "Curve-fitting/Hit-calling" (levels 5-8). The first half of the pipeline includes data preparation, mapping, batch effect correction, normalization and transformation to standard file-format. The curve-fitting and hit-calling process consists of cytotoxicity-point and outlier detection/masking, concentration activity estimates using dose-response modeling, and identifying & filtering confounded activity calls based on other assays. To increase curve-fitting and hit-calling accuracy, the ToxCast workflow incorporates a data-scan feature in the early stage of processing to identify noise and response variation (across wells, plates, and assay results). Moreover, the ToxCast workflow standardizes the analysis of highly-heterogeneous chemical assay data sets by accepting heterogeneous data formats, allowing for rapid processing with a convenient interface enabling easier access and interpretation of data at all levels for repeatable and transparent analyses. The finalized results have been uploaded into the ToxCast Dashboard for data integration and analysis, as well as to serve as the primary portal for publication and data release. *This work does not necessarily reflect Agency policy.*