What should be considered when using high-throughput screening data from CTD^2

The note is meant for consumers who do not frequently work with large-scale, high-content data generated through high-throughtput (HT) methods; specifically it highlights the factors to consider when interpreting these primary data. While many investigators are developing analytic methods to deconvolve some of the confounding variables, this is still a work in progress and error-proof solutions are not available. We therefore urge users of CTD^2 data to **"caveat emptor"**.

CTD^2 Data Portal is an open-access resource that hosts diverse datasets. The data in the Portal have been assessed for quality based on technology-relevant controls, but have not been independently validated. The Network members release their data through the Portal to accelerate identification of targets and facilitate discoveries related to understanding cancer biology.

HT approaches are exceedingly valuable for our improved understanding of cancer etiology and are the foundation for the discovery of many novel targets, perturbagens, or marker signatures. However, every HT experiment or technology has inherent imprecision and data variations related to the protocol, implementation of the screen, analytical methods, and instrumentation used. These errors and variation effect reproducibility and can give rise to conclusions that are false positive or false negative. When interpreting the HT data and/or using the information in follow-up experiments, the user needs to determine whether the initial design is suitable for their downstream goals, understand the bases of the errors and apply appropriate filters.

The CTD^2 Network has defined and implemented multi-tier definitions to validate HT results with additional experimental evidence and thereby increase the likelihood that the results are biologically relevant and can therefore move a study closer to the realm of clinical application.