Archives of Toxicology (2018) 92:2701–2702
https://doi.org/10.1007/s00204-018-2245-5

GUEST EDITORIAL

On the necessity of careful interpretation of omics data

Albert Braeuning[1] · **Falko Frenzel**[1] · **Alfonso Lampen**[1]

Received: 18 May 2018 / Accepted: 19 June 2018 / Published online: 25 June 2018

"Omics" technologies allow for unbiased analyses of molecular effects of toxicants. Bioinformatic tools assist with the extraction of information about underlying biological processes, e.g., deregulated signaling pathways. The quality of such in silico predictions is difficult to judge for toxicologists without bioinformatic background. "Ingenuity Pathway Analysis" (IPA; http://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/) is a well-established literature-based software that predicts pathway activation, transcription factor involvement, and functional interrelations. Complex cellular crosstalk functionally links virtually all molecules, and omics analyses generate reams of data points. While certainly providing advances, this also increases the risk of accidental findings.

We thus asked whether random nonsense omics data would yield "significant" biological predictions. As a model exercise, we performed comparative IPA analyses using true transcriptomics data and biologically meaningless nonsense data patterns, created by randomly re-assigning the fold-change values for transcript regulation of real data to the transcripts (see Supplement for details). IPA predicted statistically significant alterations in the so-called "canonical pathways" and "tox functions" to identify affected metabolic or signaling pathways using the software's default values. Absolute value and algebraic sign of IPA's so-called z score indicate confidence and up- or down-regulation of specific functions. "Upstream regulators" of transcriptional patterns and functionally related networks, with the number of contributing molecules and confidence scores as quality parameters, were also predicted.

"Canonical pathways", "tox functions", and "upstream regulators" results are summarized in Fig. 1 (see Supplement for details). Cumulative z scores (as overall measure of prediction quality) are plotted vs. the amount of underlying information (numbers of identified pathways, functions, or regulators). Numerous processes or molecules were, as expected, identically associated with the original data set and its clean counterpart without non-annotated probe sets. Resampled data sets also yielded numerous "tox functions" and "canonical pathways" predictions, with similar total numbers of identified pathways or functions as the original data set. However, cumulative z scores were mostly higher for biologically meaningful than for nonsense data. More pronounced differences appeared for "upstream regulators": here, predictions from the original data set reflected considerably more complex biological processes, as evident by higher numbers of effected entities and large differences in cumulative z scores (Fig. 1a). Visualization of consistency scores of regulator predictions and the number of target molecules generally assigned to the respective regulator showed that high-consistency predictions were more frequent with the original data set (Fig. 1b). Similarly, the original data set results displayed a tendency for higher network scores of functionally related molecules (Fig. 1c).

In essence, the quality of most predictions was superior with biologically meaningful data, especially for upstream regulator predictions. Nonetheless, many biological relationships were also deduced from nonsense data. Other software providing similar functionality may yield comparable results.

Data mining tools like IPA help generating biological hypotheses from comprehensive datasets, but were not designed for providing definite conclusions on biological processes. Some predictions might always be accidental, and the goal is to distinguish these from causative biological relationships. Managing limitations of omics data are essential for their future integration into toxicological risk assessment. To this end, advanced statistical methods for multiple testing corrections might improve the situation, and verification of predicted biological processes is important

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00204-018-2245-5) contains supplementary material, which is available to authorized users.

✉ Albert Braeuning
Albert.Braeuning@bfr.bund.de

[1] Department Food Safety, German Federal Institute for Risk Assessment, Max-Dohrn-Str. 8-10, 10589 Berlin, Germany

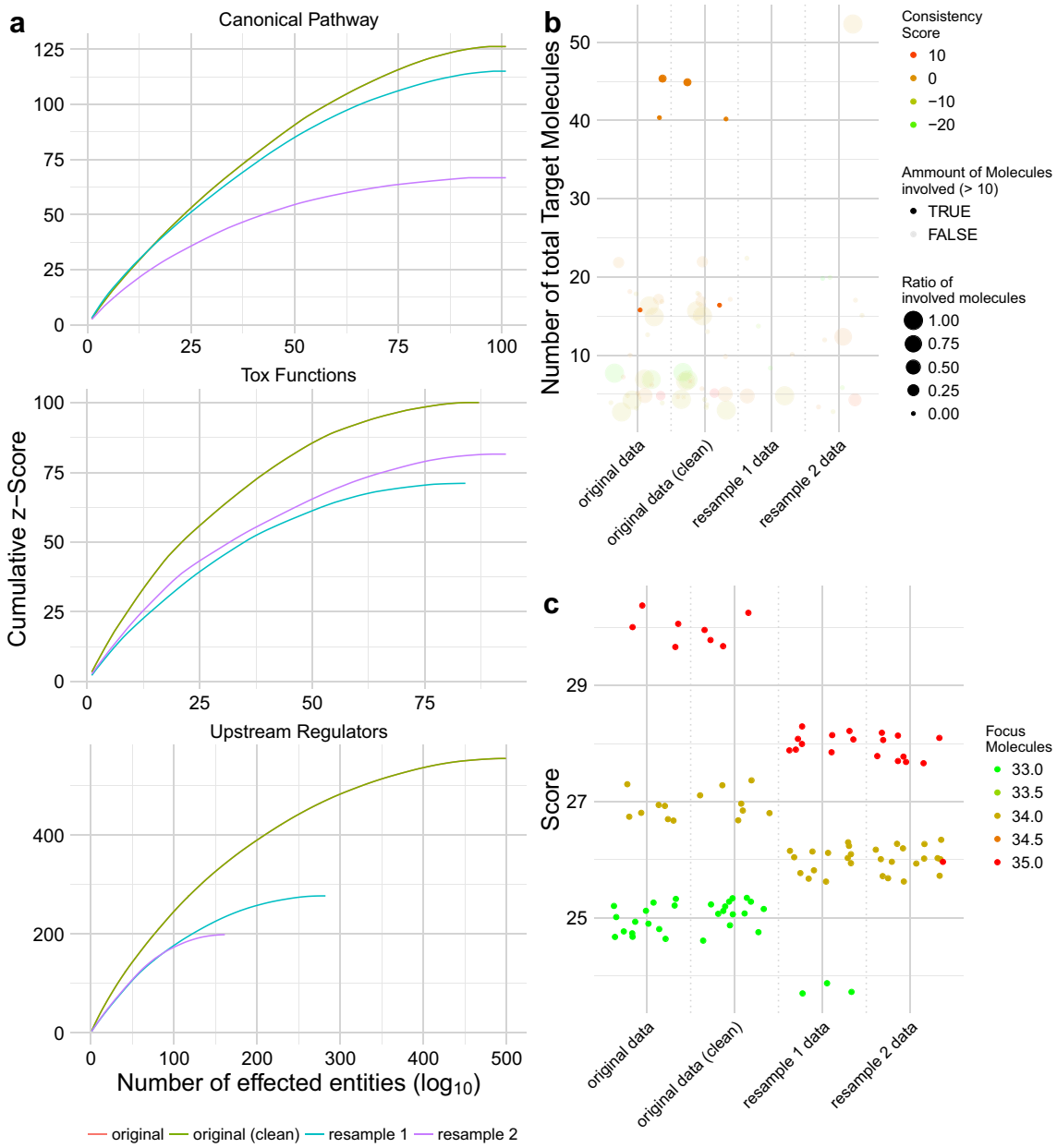

Fig. 1 Comparative IPA analysis. **a** Cumulative z scores of all predictions for the respective data set and analysis method are plotted vs. the number of affected entities ("canonical pathways", "tox functions", and "upstream regulators"). **b** Consistency scores of regulator predictions and the number of target molecules generally assigned to the respective regulator are plotted. Symbol size and opacity indicate the ratio of involved molecules (i.e., fraction of total molecules in the pathway and number of molecules actually deregulated in the analysis) and the general size of the group of deregulated molecules, respectively. **c** Network complexity is shown by confidence score and network size (color). Please note that original and clean (i.e., without the non-annotated probe sets) data sets yielded identical results. (Color figure online)

to increase confidence in bioinformatic data mining. The results of our model exercise here will increase awareness of this issue.

Acknowledgements The authors thank Dr. J. Rasinger (Bergen, Norway) for critical discussion and C. Knebel (Berlin, Germany) for sharing transcriptomic data.