EXCLI Journal 2015;14:1256-1258 – ISSN 1611-2156
Received: December 15, 2015, accepted: December 18, 2015, published: December 21, 2015

Guest editorial:

HIGHLIGHT REPORT: ERRONEOUS SAMPLE ANNOTATION IN A HIGH FRACTION OF PUBLICLY AVAILABLE GENOME-WIDE EXPRESSION DATASETS

Marianna Grinberg

Department of Statistics, TU Dortmund University, 44139 Dortmund, Germany,
E-mail: grinberg@statistik.tu-dortmund.de

http://dx.doi.org/10.17179/excli2015-760

Recently, Lohr et al. have published a method that identifies sample annotation errors in gene expression data (Lohr et al., 2015). Surprisingly, 40 % of 45 analyzed publicly available datasets including 4913 patients were affected by erroneous sample annotation. The authors conclude that sample annotation errors may be a more widespread phenomenon as previously expected (Lohr et al., 2015). The authors used two strategies for identifying sample mix-up. First, a classifier was established that differentiates between samples from female and male patients. This classifier is based on the x-chromosomal gene XIST and the y-chromosomal genes RPS4Y1 and DDX3Y (Lohr et al., 2015). In datasets with similar numbers of male and females, approximately half of sample mix-ups will result in sex mislabeling. A further possible error is sample duplication, where the same sample is analyzed twice and the duplicate is erroneously labeled with another patient (Lohr et al., 2015). To identify such duplications, a correlation-based strategy was used. A strength of the techniques presented by Lohr et al. is that they include normalization steps which make it possible to apply the same algorithm on samples of all datasets. The algorithm then differentiates between 'correctly classified' and 'misclassified' samples. In the analyzed 45 publicly available cohorts 18 contained at least one misclassification. The authors also show that deleting the erroneous samples can strongly influence the number of statistically significant prognostic genes.

Currently, genome-wide data are frequently used in cancer research (Stock et al., 2015; Sicking et al., 2014; Cadenas et al., 2014; Mattson et al., 2015). Intensively studied fields are breast- and ovarian cancer (Siggelkow et al., 2012; Godoy et al., 2014; Stewart et al., 2012; Schmidt et al., 2012). It can be expected that cohorts with only samples from either females or males have a lower risk of sex mislabeling. Therefore, it was surprising that an example of mislabeled patients was also identified in breast cancer patients. For example the well-known TRANSBIG cohort contains one female node-negative breast cancer patient who in reality is a man (Lohr et al., 2015).

Besides its intensive use in cancer research (Micke et al., 2014; Schmidt et al., 2008; Botling et al., 2013) genome-wide expression data are also frequently used in toxicology (Campos et al., 2014; Stöber et al., 2014; Marchan, 2014a, b; Bolt, 2013; Song et

al., 2013; Godoy et al., 2013; Bolt et al., 2010). The goal of these studies is to obtain first evidence of the mechanism of action of chemicals (Glahn et al., 2008; Shimada et al., 2010; Dika Nguea et al., 2008; Hendrickx et al., 2014; Shinde et al., 2015; Yao and Costa, 2014; Gagné et al., 2013; Fang et al., 2014; Kim et al., 2012) or to establish classifiers of co-regulated gene clusters (Grinberg et al., 2014; Shao et al., 2014; Krug et al., 2013; Godoy et al., 2015; Rempel et al., 2015; Doktorova et al., 2012). In these datasets the correlation-based classifier for sample duplication may be helpful. In conclusion, the easy to use classifiers published by Lohr and colleagues (2015) should be routinely included into the analysis of gene array but also RNA seq data to reduce the number of erroneous sample annotations.

REFERENCES

Bolt HM, Marchan R, Hengstler JG. Gene array screening for identification of drugs with low levels of adverse side effects. Arch Toxicol. 2010;84:253-4.

Bolt HM. Transcriptomics in developmental toxicity testing. EXCLI J. 2013;12:1027-9.

Botling J, Edlund K, Lohr M, Hellwig B, Holmberg L, Lambe M, et al. Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. Clin Cancer Res. 2013;19:194-204.

Cadenas C, van de Sandt L, Edlund K, Lohr M, Hellwig B, Marchan R, et al. Loss of circadian clock gene expression is associated with tumor progression in breast cancer. Cell Cycle. 2014;13:3282-91.

Campos G, Schmidt-Heck W, Ghallab A, Rochlitz K, Pütter L, Medinas DB, et al. The transcription factor CHOP, a central component of the transcriptional regulatory network induced upon CCl4 intoxication in mouse liver, is not a critical mediator of hepatotoxicity. Arch Toxicol. 2014;88:1267-80.

Dika Nguea H, de Reydellet A, Lehuédé P, De Meringo A, Le Faou A, Marcocci L, et al. Gene expression profile in monocyte during in vitro mineral fiber degradation. Arch Toxicol. 2008;82:355-62.

Doktorova TY, Ellinger-Ziegelbauer H, Vinken M, Vanhaecke T, van Delft J, Kleinjans J, et al. Comparison of hepatocarcinogen-induced gene expression profiles in conventional primary rat hepatocytes with in vivo rat liver. Arch Toxicol. 2012;86:1399-411.

Fang JL, Han T, Wu Q, Beland FA, Chang CW, Guo L, et al. Differential gene expression in human hepatocyte cell lines exposed to the antiretroviral agent zidovudine. Arch Toxicol. 2014;88:609-23.

Gagné F, André C, Turcotte P, Gagnon C, Sherry J, Talbot A. A comparative toxicogenomic investigation of oil sand water and processed water in rainbow trout hepatocytes. Arch Environ Contam Toxicol. 2013;65: 309-23.

Glahn F, Schmidt-Heck W, Zellmer S, Guthke R, Wiese J, Golka K, et al. Cadmium, cobalt and lead cause stress response, cell cycle deregulation and increased steroid as well as xenobiotic metabolism in primary normal human bronchial epithelial cells which is coordinated by at least nine transcription factors. Arch Toxicol. 2008;82:513-24.

Godoy P, Hewitt NJ, Albrecht U, Andersen ME, Ansari N, Bhattacharya S, et al. Recent advances in 2D and 3D in vitro systems using primary hepatocytes, alternative hepatocyte sources and non-parenchymal liver cells and their use in investigating mechanisms of hepatotoxicity, cell signaling and ADME. Arch Toxicol. 2013;87:1315-530.

Godoy P, Cadenas C, Hellwig B, Marchan R, Stewart J, Reif R, et al. Interferon-inducible guanylate binding protein (GBP2) is associated with better prognosis in breast cancer and indicates an efficient T cell response. Breast Cancer. 2014;21:491-9.

Godoy P, Schmidt-Heck W, Natarajan K, Lucendo-Villarin B, Szkolnicka D, Asplund A, et al. Gene networks and transcription factor motifs defining the differentiation of stem cells into hepatocyte-like cells. J Hepatol. 2015;63:934-42.

Grinberg M, Stöber RM, Edlund K, Rempel E, Godoy P, Reif R, et al. Toxicogenomics directory of chemically exposed human hepatocytes. Arch Toxicol. 2014;88:2261-87.

Hendrickx DM, Boyles RR, Kleinjans JC, Dearry A. Workshop report: Identifying opportunities for global integration of toxicogenomics databases, 26-27 June 2013, Research Triangle Park, NC, USA. Arch Toxicol. 2014;88:2323-32.

Kim JS, Song KS, Lee JK, Choi YC, Bang IS, Kang CS, et al. Toxicogenomic comparison of multi-wall carbon nanotubes (MWCNTs) and asbestos. Arch Toxicol. 2012;86:553-62.

Krug AK, Kolde R, Gaspar JA, Rempel E, Balmer NV, Meganathan K, et al. Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. Arch Toxicol. 2013;87:123-43.

Lohr M, Hellwig B, Edlund K, Mattsson JS, Botling J, Schmidt M, et al. Identification of sample annotation errors in gene expression datasets. Arch Toxicol. 2015;89:2265-72.

Marchan R. Highlight report: Validation of prognostic genes in lung cancer. EXCLI J. 2014a;13:457-60.

Marchan R. Cancer research: from prognostic genes to therapeutic targets. EXCLI J. 2014b;13:1278-80.

Mattsson JS, Bergman B, Grinberg M, Edlund K, Marincevic M, Jirström K, et al. Prognostic impact of COX-2 in non-small cell lung cancer: a comprehensive compartment-specific evaluation of tumor and stromal cell expression. Cancer Lett. 2015;356:837-45.

Micke P, Mattsson JS, Edlund K, Lohr M, Jirström K, Berglund A, et al. Aberrantly activated claudin 6 and 18.2 as potential therapy targets in non-small-cell lung cancer. Int J Cancer. 2014;135:2206-14.

Rempel E, Hoelting L, Waldmann T, Balmer NV, Schildknecht S, Grinberg M, et al. A transcriptome-based classifier to identify developmental toxicants by stem cell testing: design, validation and optimization for histone deacetylase inhibitors. Arch Toxicol. 2015;89:1599-618.

Schmidt M, Hellwig B, Hammad S, Othman A, Lohr M, Chen Z, et al. A comprehensive analysis of human gene expression profiles identifies stromal immunoglobulin κ C as a compatible prognostic marker in human solid tumors. Clin Cancer Res. 2012;18:2695-703.

Shao J, Berger LF, Hendriksen PJ, Peijnenburg AA, van Loveren H, Volger OL. Transcriptome-based functional classifiers for direct immunotoxicity. Arch Toxicol. 2014;88:673-89.

Shimada M, Kameo S, Sugawara N, Yaginuma-Sakurai K, Kurokawa N, Mizukami-Murata S, et al. Gene expression profiles in the brain of the neonate mouse perinatally exposed to methylmercury and/or polychlorinated biphenyls. Arch Toxicol. 2010;84:271-86.

Shinde V, Stöber R, Nemade H, Sotiriadou I, Hescheler J, Hengstler J, et al. Transcriptomics of hepatocytes treated with toxicants for investigating molecular mechanisms underlying hepatotoxicity. Methods Mol Biol. 2015;1250:225-40.

Sicking I, Rommens K, Battista MJ, Böhm D, Gebhard S, Lebrecht A, et al. Prognostic influence of cyclooxygenase-2 protein and mRNA expression in node-negative breast cancer patients. BMC Cancer. 2014;14:952.

Siggelkow W, Boehm D, Gebhard S, Battista M, Sicking I, Lebrecht A, et al. Expression of aurora kinase A is associated with metastasis-free survival in node-negative breast cancer patients. BMC Cancer. 2012;12:562.

Song M, Song MK, Choi HS, Ryu JC. Monitoring of deiodinase deficiency based on transcriptomic responses in SH-SY5Y cells. Arch Toxicol. 2013;87:1103-13.

Stewart JD, Marchan R, Lesjak MS, Lambert J, Hergenroeder R, Ellis JK, et al. Choline-releasing glycerophosphodiesterase EDI3 drives tumor cell migration and metastasis. Proc Natl Acad Sci U S A. 2012;109:8155-60.

Stock AM, Klee F, Edlund K, Grinberg M, Hammad S, Marchan R, et al. Gelsolin is associated with longer metastasis-free survival and reduced cell migration in estrogen receptor-positive breast cancer. Anticancer Res. 2015;35:5277-85.

Stöber R. Transcriptome based differentiation of harmless, teratogenetic and cytotoxic concentration ranges of valproic acid. EXCLI J. 2014;13:1281-2.

Yao Y, Costa M. Toxicogenomic effect of nickel and beyond. Arch Toxicol. 2014;88:1645-50.