Arch Toxicol (2015) 89:2459–2461
DOI 10.1007/s00204-015-1644-0

EDITORIAL

Highlight report: quality control for genome-wide expression data: how to identify sample mix-up

Marianna Grinberg[1]

Published online: 27 November 2015

In this issue of the Archives of Toxicology, Miriam Lohr and colleagues from TU Dortmund University in Germany contributed a method how to identify sample mix-up in gene expression datasets (Lohr et al. 2015). Today, large sets of genome-wide expression data obtained either by gene array or by RNAseq are publicly available. This offers the possibility to reuse previously analyzed data and to combine several datasets in order to improve statistical power (Lohr et al. 2015). However, a risk is that samples in public databases have been misannotated. This may lead to errors which are particularly critical, when the same error compromises several studies which all rely on the same publicly available expression data. Therefore, Lohr et al. (2015) established a biostatistical method that allows the retrospective identification of misannotated samples. The authors show that two types of error occur surprisingly often in public databases: (1) A sample is analyzed twice, and the duplicate is erroneously labeled with the identification number of another patient. (2) Two samples are mixed up. When this mix-up occurs between samples from male and female individuals, it can clearly be identified by a set of sex-specific genes. The authors applied the mix-up identification procedures to 45 publicly available datasets, including 4913 individuals. Erroneous sample annotation was identified in a surprisingly high fraction of 40 % of the analyzed datasets (Lohr et al. 2015). The authors also demonstrate that the removal of the identified erroneous samples may critically influence the results of statistical analysis of individual studies.

✉ Marianna Grinberg
grinberg@statistik.tu-dortmund.de

[1] Fakultät Statistik, Technische Universität Dortmund, Vogelpothsweg 87, 44221 Dortmund, Germany

Publicly available datasets have been particularly often used in clinical cancer research to identify prognostic genes (Schmidt et al. 2008, 2012; Cadenas et al. 2010, 2014; Stewart et al. 2012; Godoy et al. 2014; Stock et al. 2015). However, genome-wide expression data are also increasingly used in toxicological research (Song et al. 2013; Faust et al. 2013; Shao et al. 2014; Shinde et al. 2015; Reif 2014; Stöber 2014; Marchan 2014; Hammad and Ahmed 2014; Hengstler 2011; Stewart 2010). Cutting-edge topics are the establishment of classifiers and signatures in developmental toxicity (Krug et al. 2013; Rempel et al. 2015; Balmer et al. 2014; Zimmer et al. 2014; Waldmann et al. 2014) and hepatotoxicity (Campos et al. 2014; Yafune et al. 2013; Doktorova et al. 2012; Zellmer et al. 2010; Godoy et al. 2015). For these purposes, large public databases are available (Grinberg et al. 2014). Using the novel methods for the identification of sample annotation errors described by Lohr et al. (2015) will improve the reliability of genome-wide biostatistical analyses in future.

References

Lohr M et al (2015) Identification of sample annotation errors in gene expression datasets. Arch Toxicol (this issue, epub ahead of print)

Balmer NV, Klima S, Rempel E, Ivanova VN, Kolde R, Weng MK, Meganathan K, Henry M, Sachinidis A, Berthold MR, Hengstler JG, Rahnenführer J, Waldmann T, Leist M (2014) From transient transcriptome responses to disturbed neurodevelopment: role of histone acetylation and methylation as epigenetic switch between reversible and irreversible drug effects. Arch Toxicol 88(7):1451–1468. doi:10.1007/s00204-014-1279-6

Cadenas C, Franckenstein D, Schmidt M, Gehrmann M, Hermes M, Geppert B, Schormann W, Maccoux LJ, Schug M, Schumann A, Wilhelm C, Freis E, Ickstadt K, Rahnenführer J, Baumbach JI, Sickmann A, Hengstler JG (2010) Role of thioredoxin reductase

1 and thioredoxin interacting protein in prognosis of breast cancer. Breast Cancer Res 12(3):R44. doi:10.1186/bcr2599

Cadenas C, van de Sandt L, Edlund K, Lohr M, Hellwig B, Marchan R, Schmidt M, Rahnenführer J, Oster H, Hengstler JG (2014) Loss of circadian clock gene expression is associated with tumor progression in breast cancer. Cell Cycle 13(20):3282–3291. doi:10.4161/15384101.2014.954454

Campos G, Schmidt-Heck W, Ghallab A, Rochlitz K, Pütter L, Medinas DB, Hetz C, Widera A, Cadenas C, Begher-Tibbe B, Reif R, Günther G, Sachinidis A, Hengstler JG, Godoy P (2014) 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 88(6):1267–1280. doi:10.1007/s00204-014-1240-8

Doktorova TY, Ellinger-Ziegelbauer H, Vinken M, Vanhaecke T, van Delft J, Kleinjans J, Ahr HJ, Rogiers V (2012) Comparison of genotoxicant-modified transcriptomic responses in conventional and epigenetically stabilized primary rat hepatocytes with in vivo rat liver data. Arch Toxicol 86(11):1703–1715. doi:10.1007/s00204-012-0946-8

Faust D, Vondráček J, Krčmář P, Smerdová L, Procházková J, Hrubá E, Hulinková P, Kaina B, Dietrich C, Machala M (2013) AhR-mediated changes in global gene expression in rat liver progenitor cells. Arch Toxicol 87(4):681–698. doi:10.1007/s00204-012-0979-z

Godoy P, Cadenas C, Hellwig B, Marchan R, Stewart J, Reif R, Lohr M, Gehrmann M, Rahnenführer J, Schmidt M, Hengstler JG (2014) Interferon-inducible guanylate binding protein (GBP2) is associated with better prognosis in breast cancer and indicates an efficient T cell response. Breast Cancer 21(4):491–499. doi:10.1007/s12282-012-0404-8

Godoy P, Schmidt-Heck W, Natarajan K, Lucendo-Villarin B, Szkolnicka D, Asplund A, Björquist P, Widera A, Stöber R, Campos G, Hammad S, Sachinidis A, Chaudhari U, Damm G, Weiss TS, Nüssler A, Synnergren J, Edlund K, Küppers-Munther B, Hay DC, Hengstler JG (2015) Gene networks and transcription factor motifs defining the differentiation of stem cells into hepatocyte-like cells. J Hepatol 63(4):934–942. doi:10.1016/j.jhep.2015.05.013

Grinberg M, Stöber RM, Edlund K, Rempel E, Godoy P, Reif R, Widera A, Madjar K, Schmidt-Heck W, Marchan R, Sachinidis A, Spitkovsky D, Hescheler J, Carmo H, Arbo MD, van de Water B, Wink S, Vinken M, Rogiers V, Escher S, Hardy B, Mitic D, Myatt G, Waldmann T, Mardinoglu A, Damm G, Seehofer D, Nüssler A, Weiss TS, Oberemm A, Lampen A, Schaap MM, Luijten M, van Steeg H, Thasler WE, Kleinjans JC, Stierum RH, Leist M, Rahnenführer J, Hengstler JG (2014) Toxicogenomics directory of chemically exposed human hepatocytes. Arch Toxicol 88(12):2261–2287. doi:10.1007/s00204-014-1400-x

Hammad S, Ahmed H (2014) Biomarker: the universe of chemically induced gene expression alterations in human hepatocyte. EXCLI J 13:1275–1277

Hengstler JG (2011) Cutting-edge topics in toxicology. EXCLI J 10:117–119

Krug AK, Kolde R, Gaspar JA, Rempel E, Balmer NV, Meganathan K, Vojnits K, Baquié M, Waldmann T, Ensenat-Waser R, Jagtap S, Evans RM, Julien S, Peterson H, Zagoura D, Kadereit S, Gerhard D, Sotiriadou I, Heke M, Natarajan K, Henry M, Winkler J, Marchan R, Stoppini L, Bosgra S, Westerhout J, Verwei M, Vilo J, Kortenkamp A, Hescheler J, Hothorn L, Bremer S, van Thriel C, Krause KH, Hengstler JG, Rahnenführer J, Leist M, Sachinidis A (2013) Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. Arch Toxicol 87(1):123–143. doi:10.1007/s00204-012-0967-3

Marchan R (2014) Cancer research: from prognostic genes to therapeutic targets. EXCLI J 13:1278–1280

Reif R (2014) Concepts of predictive toxicology. EXCLI J 2014(13):1292–1294

Rempel E, Hoelting L, Waldmann T, Balmer NV, Schildknecht S, Grinberg M, Das Gaspar JA, Shinde V, Stöber R, Marchan R, van Thriel C, Liebing J, Meisig J, Blüthgen N, Sachinidis A, Rahnenführer J, Hengstler JG, Leist M (2015) A transcriptome-based classifier to identify developmental toxicants by stem cell testing: design, validation and optimization for histone deacetylase inhibitors. Arch Toxicol 89(9):1599–1618. doi:10.1007/s00204-015-1573-y

Schmidt M, Böhm D, von Törne C, Steiner E, Puhl A, Pilch H, Lehr HA, Hengstler JG, Kölbl H, Gehrmann M (2008) The humoral immune system has a key prognostic impact in node-negative breast cancer. Cancer Res 68(13):5405–5413. doi:10.1158/0008-5472

Schmidt M, Hellwig B, Hammad S, Othman A, Lohr M, Chen Z, Boehm D, Gebhard S, Petry I, Lebrecht A, Cadenas C, Marchan R, Stewart JD, Solbach C, Holmberg L, Edlund K, Kultima HG, Rody A, Berglund A, Lambe M, Isaksson A, Botling J, Karn T, Müller V, Gerhold-Ay A, Cotarelo C, Sebastian M, Kronenwett R, Bojar H, Lehr HA, Sahin U, Koelbl H, Gehrmann M, Micke P, Rahnenführer J, Hengstler JG (2012) 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 18(9):2695–2703. doi:10.1158/1078-0432.CCR-11-2210

Shao J, Berger LF, Hendriksen PJ, Peijnenburg AA, van Loveren H, Volger OL (2014) Transcriptome-based functional classifiers for direct immunotoxicity. Arch Toxicol 88(3):673–689. doi:10.1007/s00204-013-1179-1

Shinde V, Klima S, Sureshkumar PS, Meganathan K, Jagtap S, Rempel E, Rahnenführer J, Hengstler JG, Waldmann T, Hescheler J, Leist M, Sachinidis A (2015) Human pluripotent stem cell based developmental toxicity assays for chemical safety screening and systems biology data generation. J Vis Exp 17(100):e52333. doi:10.3791/52333

Song M, Song MK, Choi HS, Ryu JC (2013) Monitoring of deiodinase deficiency based on transcriptomic responses in SH-SY5Y cells. Arch Toxicol 87(6):1103–1113. doi:10.1007/s00204-013-1018-4

Stewart JD (2010) In vitro test systems in toxicology. EXCLI J 9:156–158

Stewart JD, Marchan R, Lesjak MS, Lambert J, Hergenroeder R, Ellis JK, Lau CH, Keun HC, Schmitz G, Schiller J, Eibisch M, Hedberg C, Waldmann H, Lausch E, Tanner B, Sehouli J, Sagemueller J, Staude H, Steiner E, Hengstler JG (2012) Choline-releasing glycerophosphodiesterase EDI3 drives tumor cell migration and metastasis. Proc Natl Acad Sci USA 109(21):8155–8160. doi:10.1073/pnas.1117654109

Stöber R (2014) Transcriptome based differentiation of harmless, teratogenetic and cytotoxic concentration ranges of valproic acid. EXCLI J 13:1281–1282

Stock AM, Klee F, Edlund K, Grinberg M, Hammad S, Marchan R, Cadenas C, Niggemann B, Zänker KS, Rahnenführer J, Schmidt M, Hengstler JG, Entschladen F (2015) Gelsolin is associated with longer metastasis-free survival and reduced cell migration in estrogen receptor-positive breast cancer. Anticancer Res 35(10):5277–5285

Waldmann T, Rempel E, Balmer NV, König A, Kolde R, Gaspar JA, Henry M, Hescheler J, Sachinidis A, Rahnenführer J, Hengstler JG, Leist M (2014) Design principles of concentration-dependent transcriptome deviations in drug-exposed differentiating stem cells. Chem Res Toxicol 27(3):408–420. doi:10.1021/tx400402j

Yafune A, Taniai E, Morita R, Nakane F, Suzuki K, Mitsumori K, Shibutani M (2013) Expression patterns of cell cycle proteins in the livers of rats treated with hepatocarcinogens for 28 days. Arch Toxicol 87(6):1141–1153. doi:10.1007/s00204-013-1011-y

Zellmer S, Schmidt-Heck W, Godoy P, Weng H, Meyer C, Lehmann T, Sparna T, Schormann W, Hammad S, Kreutz C, Timmer J, von Weizsäcker F, Thürmann PA, Merfort I, Guthke R, Dooley S, Hengstler JG, Gebhardt R (2010) Transcription factors ETF, E2F, and SP-1 are involved in cytokine-independent proliferation of murine hepatocytes. Hepatology 52(6):2127–2136. doi:10.1002/hep.23930

Zimmer B, Pallocca G, Dreser N, Foerster S, Waldmann T, Westerhout J, Julien S, Krause KH, van Thriel C, Hengstler JG, Sachinidis A, Bosgra S, Leist M (2014) Profiling of drugs and environmental chemicals for functional impairment of neural crest migration in a novel stem cell-based test battery. Arch Toxicol 88(5):1109–1126. doi:10.1007/s00204-014-1231-9