Animal Science Papers and Reports vol. 37 (2019) no. 3, 323-327
Institute of Genetics and Animal Breeding, Jastrzębiec, Poland

Annotation of Illumina microarray probes: similarities and differences between various bioinformatic tools*

Short Report

Adrian M. Stankiewicz, Aneta Jaszczyk, Grzegorz R. Juszczak**

Institute of Genetics and Animal Breeding, Polish Academy of Sciences,
Postępu 36A, 05-552 Jastrzebiec, Poland

(Accepted March 29, 2019)

Analysis of transcriptomes is crucial for understanding animal physiology. However, comparison and interpretation of transcriptomic data is hindered by inconsistent gene nomenclature that has evolved over time leading to existence of numerous alternative names used for the same gene. A solution is the re-annotation of retrieved data but there is no commonly agreed way to perform such standardization. Therefore, we compared the results of re-annotation performed on a sample of data using various tools. The re-annotation was facilitated by application of a custom-made script that is provided in this report. Complete overlap between gene symbols originally derived from the GEO database and gene symbols obtained during re-annotation of probe id's irrespective of the applied method was found only in case of 50% of microarray probes. The inconsistent re-annotation resulted mainly from inclusion of synonyms that are not currently used as official gene symbols, missing information in the re-annotated datasets and even inclusion of transcripts that were officially withdrawn because they are not currently considered as genuine genes. These data show that conclusions drawn from microarray studies depend heavily on the applied method of probe annotation.

KEY WORDS: annotation / microarrays / probe

*Supported by NCN Grant 2017/27/B/NZ2/02796 and STAT/GRZJUS/2019/01 intramural grant.
**Corresponding author: g.juszczak@ighz.pl

Transcriptomic technologies provide researchers with huge amount of data that are important for understanding animal physiology including problems such as muscle growth, lactation, reproduction efficiency, and response to diseases [Parreira and De Sousa 2018]. These data can help, therefore, to develop new strategies to improve animal health, welfare, and production [Parreira and De Sousa 2018]. One of the most popular transcriptomic tools used to study gene expression are microarrays that measure the abundances of a defined set of transcripts via their hybridisation to an array of complementary probes [Lowe *et al*. 2017]. Microarray technology allows researcher to study entire transcriptoms but interpretation of these data and assessment of replicability between studies constitute a huge challenge [Juszczak and Stankiewicz 2018]. One of the problems is inconsistent gene nomenclature that has evolved over time leading to existence of numerous alternative names used for the same genes [Juszczak and Stankiewicz 2018]. A solution is the re-annotation of retrieved data but there is no commonly agreed way to perform such standardization. Therefore, we compared the results of re-annotation performed on a sample of data using various tools.

Material and methods

We have compared annotation of 100 microarray probes derived from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/, Accession number GSE100086) using the GEO2R online tool. Gene symbols derived from the GEO database were compared with genes symbols obtained by subsequent re-annotation of probe id's obtained from GEO. To re-annotate the data we used Gemma, a website dedicated for meta-analysis of genomic data, which re-annotates microarray probes at the sequence level [Zoubarev *et al*. 2012] (https://gemma.msl.ubc.ca/arrays/showArrayDesign.html?id=395) and data provided by the manufacturer of microarrays (Illumina, USA: http://emea.support.illumina.com/array/array_kits/mousewg-6_v2_expression_beadchip_kit/downloads.html?langsel=/de/). Text files containing probe annotation were downloaded and data on probes of interest were extracted from them. Because of the striking differences between obtained data we performed an additional annotation with data from most current release of Ensembl genome database using the biomaRt package [Durinck *et al*. 2009] for the R programming language. The script used to perform probe re-annotation can be accessed here: https://github.com/AdrianS85/varia/blob/master/Small_biomaRt_extraction.R. Finally, the re-annotation with all mentioned methods was repeated to exclude differences in time of data retrieval. In case of discrepancies we have manually checked the gene symbols in NCBI GENE database (https://www.ncbi.nlm.nih.gov/gene/) [Maglott *et al*. 2011] to discriminate between currently used official symbols and synonyms. In case of conflicting information (two different official symbols connected to the same probe) we have manually checked probes in NCBI PROBE database (https://www.ncbi.nlm.nih.gov/probe/) and Ensembl (https://www.ensembl.org/index.html) [Aken *et al*. 2016].

Results and discussion

Complete overlap between gene symbols originally derived from the GEO database and gene symbols obtained during re-annotation of probe id's irrespective of the applied method was found in case of 50% of microarray probes (Fig. 1). The least congruence was found between the GEO data and genes re-annotated using manufacturer data (63%). The manufacturer data contained considerable number of synonyms instead of official symbols (19%), transcripts that were not found in the NCBI Gene (10%) and discontinued items (7%) that were withdrawn by NCBI Gene for example because the model on which they were based were not predicted in a later annotation. Much higher congruence was found in case of data re-annotated with other tools. The same gene symbols were found in 73% and 80% of probes when the GEO data were compared with probes re-annotated either with Gemma or with most current Ensembl data using the biomaRt package. Frequently, the differences resulted from amount of data retrieved from different sources. It means that one source linked probe with a gene while another source provided no gene specific for the probe or both compared sources provided the same gene but one of them listed additional transcripts detected the probe. We found also that 4% of probes are linked to completely different genes depending on annotation method and this differences cannot be explained by presence of synonyms. This problem was found in case of the following probes: ILMN_1213278, ILMN_1214703, ILMN_1212967

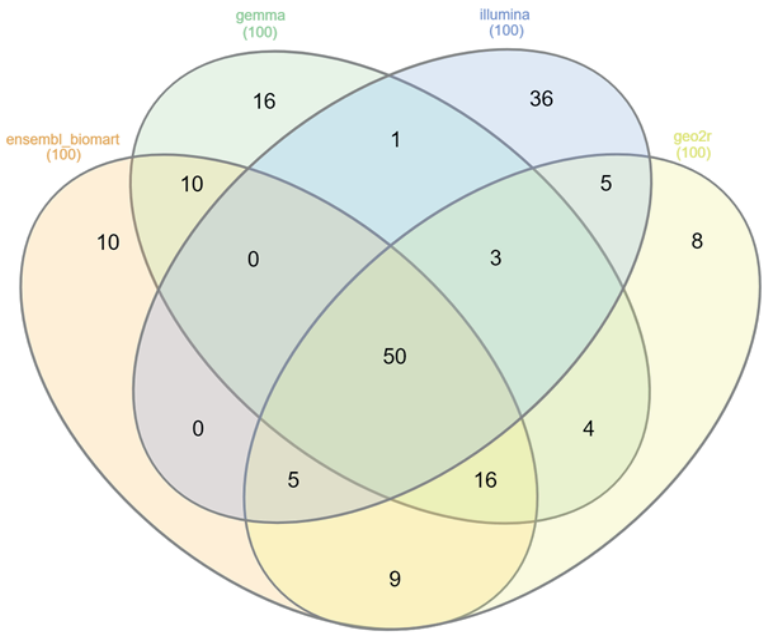

Fig. 1. Venn diagram showing differences and similarities in probe annotation.

and ILMN_1217074. The annotation made with biomaRt (ILMN_1212967) or both biomaRt and Gemma (ILMN_1214703; ILMN_1213278) were consistent with data provided by the European genomic database Ensembl, while other sources (GEO and/or manufacturer of microarrays) were consistent with data provide by American database NCBI Probe. In case of the ILMN_1217074 probe none of the major databases (NCBI and Ensembl) provided matching gene although this probe was linked with two different genes by Gemma, manufacturer of microarrays and the GEO-retrieved data. This shows complex pattern of similarities and differences between different sources of transcriptomic data. Finally, complete overlap between Gemma and biomaRt was found in case of 76% of probes. Most often (18 probes) differences resulted from less information retrieved by the Gemma (no gene symbol provided or one gene linked to the probe while biomaRt provided more than one genes). Less frequently (6 probes) Gemma provided more information than biomaRt.

These data show that conclusions drawn from microarray studies depend heavily on the method of annotation. Usage of databases that are not updated frequently enough can result in inclusion of synonyms that are not currently used as official gene symbols and even in inclusion of transcripts that have been withdrawn from databases because they were not considered as genuine genes. Probes that cannot be unequivocally attributed to defined genes, making the result of annotation dependent on the source of genomic information, introduce additional confusion while concluding is based on different annotation methods' outputs. The discrepancies between major databases (NCBI and Ensembl) can be explained by the fact that identification of gene coding sequences is a multistep process requiring many assumption leading to differences in annotation of probes [Aken *et al*. 2016].

Previously, it has been proposed that the best solution is rejection of all inconsistent annotations [Allen e*t al*. 2012]. However, our study showed that a large number of inconsistent annotations results from inclusion of alternative names that are applied for the same gene. Rejection of these data leads to a loss of considerable number of significantly regulated genes from datasets. This, in turn, negatively affects subsequent pathway analysis, which identifies groups of genes that are functionally related and appear in datasets significantly more frequently than expected from random chance. Therefore, we advise to annotate transcriptomic data with the most recently updated tool that is available at the time of the analysis.

Authors declare no conflict of interest.

REFERENCES

1. AKEN B.L., AYLING S., BARRELL D., CLARKE L., CURWEN V., FAIRLEY S., FERNANDEZ BANET J., BILLIS K., GARCIA GIRON C., HOURLIER T., HOWE K., KAHARI A., KOKOCINSKI F., MARTIN F.J., MURPHY D.N., NAG R., RUFFIER M., SCHUSTER M., TANG Y.A., VOGEL J.H., WHITE S., ZADISSA A., FLICEK P. and SEARLE S.M., 2016 – The Ensembl gene annotation system. ***Database*** 2016.

2. ALLEN J.D., WANG S., CHEN M., GIRARD L., MINNA J.D., XIE Y., XIAO G., 2012 – Probe mapping across multiple microarray platforms. ***Briefings in Bioinformatics*** 13, 547-554.
3. DURINCK S., SPELLMAN P.T., BIRNEY E. and HUBER W., 2009 – Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. ***Nature Protocols*** 4, 8, 1184-1191.
4. JUSZCZAK G.R. and STANKIEWICZ A.M., 2018 – Glucocorticoids, genes and brain function. ***Progress in Neuro-psychopharmacology and Biological Psychiatry*** 82, 136-168.
5. LOWE R., SHIRLEY N., BLEACKLEY M., DOLAN S. and SHAFEE T., 2017 – Transcriptomics technologies. ***PLoS Computational Biology*** 13, 5, e1005457.
6. MAGLOTT D., OSTELL J., PRUITT K.D. and TATUSOVA T., 2011 – Entrez Gene: gene-centered information at NCBI. ***Nucleic Acids Research*** 39, D52-57.
7. PARREIRA J.R. and DE SOUSA A.S. 2018 – Studying the Animal Transcriptome: State of the Art and Challenges in the Context of Animal and Veterinary Sciences. Proteomics in Domestic Animals: from Farm to Systems Biology. A. de Almeida, D. Eckersall and I. Miller, Springer**,** 421-446.
8. ZOUBAREV A., HAMER K.M., KESHAV K.D., MCCARTHY E.L., SANTOS J.R., VAN ROSSUM T., MCDONALD C., HALL A., WAN X., LIM R., GILLIS J. and PAVLIDIS P., 2012 – Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. ***Bioinformatics*** 28, 17, 2272-2273.