DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2013, Article ID bat010, doi:10.1093/database/bat010

Original article

Uncovering hidden duplicated content in public transcriptomics data

Marta Rosikiewicz[1,2,†], Aurélie Comte[1,2,†], Anne Niknejad[1,2], Marc Robinson-Rechavi[1,2] and Frederic B. Bastian[1,2,*]

[1]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland and [2]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

***Corresponding author:** Tel: +41 21 692 4221; Fax: +41 21 692 4165; Email: frederic.bastian@unil.ch

†These authors contributed equally to this work.

Submitted 30 November 2012; Revised 16 January 2013; Accepted 19 February 2013

Citation details: Rosikiewicz,M., Comte,A., Niknejad, A. *et al.* Uncovering hidden duplicated content in public transcriptomics data. *Database* (2013) Vol. 2013: article ID bat010; doi:10.1093/database/bat010

As part of the development of the database Bgee (a dataBase for Gene Expression Evolution), we annotate and analyse expression data from different types and different sources, notably Affymetrix data from GEO and ArrayExpress, and RNA-Seq data from SRA. During our quality control procedure, we have identified duplicated content in GEO and ArrayExpress, affecting ~14% of our data: fully or partially duplicated experiments from independent data submissions, Affymetrix chips reused in several experiments, or reused within an experiment. We present here the procedure that we have established to filter such duplicates from Affymetrix data, and our procedure to identify future potential duplicates in RNA-Seq data.

Database URL: http://bgee.unil.ch/

Introduction

Bgee (a dataBase for Gene Expression Evolution) (1) is a resource to perform high-throughput and automated comparisons of gene expression patterns between species. To this end, it integrates expression data of different types (ESTs, *in situ* hybridization, Affymetrix and RNA-Seq data, as of Bgee release 12), from different species (fruit fly, human, mouse, Xenopus and zebrafish, as of Bgee release 12). Each data type is analysed using dedicated statistical tests, to generate present/absent expression calls, and differential expression calls (overexpression and underexpression). These data are then integrated into a homology framework (relations of homology between organs and between genes), facilitating the comparison of gene expression patterns, between and within species, as well as the study of their evolution.

To date, we have manually curated 15 988 Affymetrix chips from 1285 experiments, retrieved from Gene Expression Omnibus (GEO) (2) and ArrayExpress (3). The curation process includes the following: (i) controlling for samples 'normality' (e.g. no treatments, no diseases, wild type genotypes); (ii) mapping the chips to anatomical and developmental ontologies, to determine 'where and when' genes are expressed; and (iii) performing quality controls to remove low-quality and incompatible chips (detailed in the documentation section of our website; http://bgee.unil.ch/?page=documentation#AffyQC).

Following improvements in our quality control procedures, we noticed that many chips shared identical values of quality parameters, which was highly unlikely. After further examination, we found that these chips were truly identical. Indeed, it seems to be common practice among the community to reuse data, especially those used as controls, and to resubmit them to public repositories into new experiments. Although this might not be an issue when analysing individual experiments, it is a problem when performing meta-analyses. Moreover, as this practice is

Database, Vol. 2013, Article ID bat010, doi:10.1093/database/bat010

not documented, it is misleading for the construction of secondary databases such as Bgee based on the primary transcriptome data. We therefore developed a pipeline to filter for such duplicates.

Bgee release 12 also integrates some RNA-Seq data from Sequence Read Archive (SRA) (4). In future releases, this is expected to be the fastest growing data type. Although we have not yet detected problems of duplicated runs, we have also built a pipeline to allow their detection. This is even more important for RNA-Seq data, as we cannot always store the extremely large run files locally, but we need to re-download and re-process the data at each new release. We therefore need not only to check for run file identity but also to track any change between releases in the run files that constitute a sample.

Materials and methods

Affymetrix data

We have established a procedure to detect duplicated Affymetrix data. The procedure is different depending on whether the raw CEL files or only the processed MAS5 files (5) are available.

When raw CEL files are available:

(1) Computation of the present/absent MAS5 calls, and retrieval of the percentage of probesets flagged as 'present' for each chip.
(2) Computation of the 'Inter Quartile Range of average rank' (arIQR), a new quality score, which we have developed. Briefly, it is obtained by ranking all the probes intensities from a given array, then computing the average rank for each probeset; arIQR is the interquartile range of the probesets average ranks for a given array (M. Rosikiewicz and M. Robinson-Rechavi, unpublished). For the purpose of this article, the important feature is that it is very unlikely that two CEL files have the same arIQR.
(3) Generation for each chip of an expression based unique identifier, by concatenating the percentage of present probesets and the arIQR score.
(4) Computation of the SHA512 checksum [a cryptographic hash function, which will provide a unique identifier for each unique data file (6)] of the CEL files.
(5) Retrieval of the scan dates (second accuracy) from the CEL files.

If the checksums or the expression based unique identifiers are equal between two chips, they are likely identical and are manually checked by curators. The unique identifier approach might seem redundant to the SHA512 checksum, but it allowed us to detect duplicates for which the files were different, because of differences in the CEL file format used (i.e. CEL format version 3 and 4 generated with MAS5 and Gene Chip Operating Software, respectively). Although scan date equality is a useful red flag, it is not sufficient to consider two chips identical and is therefore mostly used as a control (two chips can be scanned together at the same time).

When only processed MAS5 files are available:

(1) Generation of an expression based unique identifier for each MAS5 file, by concatenating the number of 'present' calls, 'absent' calls, 'marginal' calls and 'undefined' calls.
(2) Computation of the SHA512 checksum of the original MAS5 files.
(3) Standardization of the original files to make them comparable. Depending on which software was used [e.g. MAS5, R (7)], the files generated from a same sample can be slightly different: change of the header line, change in the order of the probesets, changes in the values representing the calls. This limits the use of a checksum for file comparison. We thus modify all MAS5 files: (i) only the columns of the probeset identifiers, the signal intensities and the MAS5 calls are kept; (ii) all lines with non-standard formats are removed (this notably removes the header line); (iii) lines are ordered based on the probeset identifiers; (iv) all calls are standardized ('p' and 'present' in different cases transform to 'present'; 'a' and 'absent' in different cases transform to 'absent'; 'm' and 'marginal' in different cases transform to 'marginal'; 'u', 'undefined' and 'rp', in different cases, transform to 'undefined').
(4) Computation of the SHA512 checksum of the standardized files.

If any of the checksums or the expression based unique identifiers are equal between two chips, they are likely identical and are manually checked by curators.

The information generated for all CEL files and MAS5 files used in Bgee release 12 are available as supplementary material (see Supplementary Table S1).

RNA-Seq data

RNA-seq samples are often composed of several runs. As the run files are re-downloaded and re-processed for each new release of Bgee, it is important, not only to identify duplicated run files but also to track changes in the run composition of samples, between releases. We thus established a procedure to uniquely identify run files:

(1) Computation of the SHA512 checksum of the run files.
(2) Retrieval of the size (in bytes) of the run files.

If the checksums are equal between two runs, they are considered identical. If the sizes are equal, an inspection of the data is required. This procedure is preliminary, and other

parameters, such as expression-based unique identifiers, might be added in the future.

To track potential changes in sample composition, at each release of Bgee, we store the identifiers of the runs associated with each sample, and for each run identifier, the checksum and the size of the related file. We then compare this information between releases: (i) addition or deletion of a run file associated with a sample and (ii) change in the checksum or the size of a run file.

The information generated for the run files used in Bgee is available as supplementary material (see Supplementary Table S2).

Results

Identification of duplicated data

Using the procedure described earlier in the text, we have identified in our data set (Bgee release 12) 1065 groups of two to four identical Affymetrix chips, reused in one to four experiments (see Table 1, Supplementary Tables S3 and S4). All files were manually confirmed to be duplicates. This represents 2173 chips from 111 experiments (13.6% of our annotated chips and 8.6% of our annotated experiments, respectively). These experiments included from 1 to 340 chips used more than once (see Table 2).

We have identified several typical scenarios explaining data duplication:

Error in data submission: this is for instance the case with samples GSM8982 and GSM8979 from experiment GSE591 (8). They are supposed to be two replicates of the expression profiles of 3-week-old female wild type FVB mice while they are actually identical. As the experiment includes other wild type samples with true replicates, this is likely an error. Of note, a simple checksum comparison of the two files in MAS5 format would have been misleading as the headers of the files are different (they include the identifier of the sample). The expression-based identifier and the checksum of the filtered MAS5 file allowed us to identify the correspondence between these samples.

Reuse of samples in several experiments: this is for instance the case in the five experiments, used as the core data of five papers, GSE9692 (9), GSE26378 (10), GSE8121 (11), GSE13904 (12), and GSE26440 (13). It appears that over the 101 samples that we have annotated from these experiments, 72 were reused several times: 15 were duplicated in four experiments (60 annotated samples, GSE9692, GSE8121, GSE13904, GSE26440); three were duplicated in two experiments (six annotated samples, GSE13904 and GSE26440), leading these experiments to have a total of 18 samples in common; yet, three others in two experiments (six annotated samples, GSE26378 and GSE26440). The reason is that these experiments reused blood samples from healthy individuals to be compared with blood samples from septic shock patients.

Complete duplication of an experiment: all of the 120 samples of the experiment GSE9676 (14) were reused in the experiment GSE10760 (15). The former experiment studied expression profiles of healthy human skeletal muscles, whereas the latter compared expression profiles of muscle samples affected by facioscapulohumeral dystrophy with control samples that were being reused.

Table 1. Distribution of groups of identical Affymetrix chips

Number of chip groups	Number of chips per group	Number of experiments per group
4	2	1
13	3	3
15	4	4
1033	2	2

Improvement of the Bgee data set

When files have been manually confirmed to be redundant, only one representative of the group is kept. When a chip is duplicated between different experiments, the one that is part of the experiment with the largest number of remaining chips is kept, as normalization procedures are more efficient with a larger number of chips. This led us to remove 1119 chips from our data set (7% of our annotated chips, see Supplementary Table S4).

In some cases, the information provided about duplicated samples was inconsistent or provided with different granularity, depending on the focus of the study for which they were used. For instance, although the samples GSM322066 and GSM336955 from experiment GSE12826

Table 2. Distribution of pairs of experiments sharing identical chips

Number of shared identical chips	1	2	3	4	5	6–10	11–20	21–50	51–340
Number of experiment pairs	2	5	14	8	3	14	13	3	4

Note that an experiment can be part of several pairs, depending on the number of experiments it shares chips with, and that the four experiments using duplicated chips within themselves (GSE591, GSE9750, GSE6196 and GSE6490) are not considered.

Database, Vol. 2013, Article ID bat010, doi:10.1093/database/bat010

(16) and GSE13348 (17) are identical, the first one is supposed to have been obtained from a whole zebrafish embryo and the second one from the brain of a zebrafish embryo. We have tracked annotation inconsistencies between identical chips and have corrected them in Bgee.

Discussion

The problem of redundancy has long been identified in sequence databases [e.g. (18)] but has not been addressed as far as we know at the level of functional genomics data. Detecting redundancy in these data can be more complicated than for DNA or protein sequences, but it is important to allow unbiased meta-analyses and comparisons of results between conditions or species. For example, duplicated samples might provide a false sense of confidence in a result, which is in fact only supported by one experimental data point.

We have implemented different measurements to identify redundant content, through several iterations, as we noticed that some of them failed to detect duplicated Affymetrix chips. For instance, using the SHA512 checksums failed most of the time to identify duplicated MAS5 files, as the chip identifier from the source database seems to be automatically added to the header of these files. Such duplicates were identified owing to the expression-based unique identifier approach and to the SHA512 checksums from the standardized MAS5 files. These two measurements have been so far always congruent but are nevertheless both useful. We cannot rule out that two files from different samples could contain the same number of expression calls (used to build the expression-based unique identifier of MAS5 files), or that our filtering step of the original files could fail to standardize some of them.

Similarly, a third of the CEL files found to be duplicated were not identified using their SHA512 checksums, but using their expression-based unique identifiers (all the other CEL files were identified by both these measurements). This is likely caused by the generation of CEL files, from a same scan, using different software. The scan date is in such cases a useful control, which so far always confirmed the identity revealed by the expression-based unique identifier. But this scan date is not sufficient to consider two files as duplicated: in one case, it was identical between two different CEL files, from a same experiment.

Of note, our pipeline is currently only suitable for the comparison of MAS5 files on the one hand, and of CEL files on the other hand, and is not applicable to other file types. The normalization of a same chip using different procedures [e.g. gcRMA, RMA (19,20)] leads to slightly different results, which are neither reversible nor comparable. If two files were generated from a same chip, but using different normalization procedures, they would not be identified as duplicates. This is one of the reasons why raw CEL files are always included preferentially in Bgee, as the normalization step, irreversibly hiding information, is not yet performed.

Our pipeline is also only capable of detecting different files generated from a same assay. It is neither capable of detecting, for instance, technical replicates, where a same sample is assayed on different chips, nor is it aimed at detecting them.

Using this approach, we were able to identify duplicated content that would have otherwise remained unnoticed. Indeed, only in a few cases did the authors take care to link their duplicated samples when submitting their data. For instance, the samples GSM2334, GSM2335, GSM2336, from the experiments GSE760 (21) and GSE75 (22), have the same identifiers and are clearly part of the two experiments. But this represents only 55 groups of duplicated chips out of 1065 (5%).

In some other cases, the authors mentioned the data duplication in the full text description of the experiment, as for the experiment GSE10760 (see example earlier in the text). Yet this information is not available to automatic tools; in the majority of cases, the authors did not update the first experiment submitted to add links to the later one (i.e. GSE9676 has no link to GSE10760).

Most often, it is nowhere explicitly stated that some samples are redundant: the identifier of the samples are different; the names of the data files are different. By a careful examination of the experiments descriptions, it might be possible to realize that the data have been submitted by the same authors, and that one experiment studied healthy samples, whereas the other one used healthy samples as a control. But without a control procedure, such as the one we have established, this cannot be distinguished from the case of a laboratory submitting several different experiments related to its research subject.

Conclusion

We have set up procedures to identify duplicated content in Affymetrix and RNA-Seq data and to track changes in run composition of samples used in RNA-Seq analyses. We have discovered that a large number of Affymetrix chips were actually redundant in the GEO and ArrayExpress repositories. As far as we know, this issue has not previously been reported, and the present study provides insight into precautions that should be taken when using public data.

As our data set is focused on 'normal' samples (e.g. no drug treatments, no gene knock-outs), which are more likely to be reused as control samples in different experiments, our result that 13.6% of Affymetrix data content is duplicated might represent an upper boundary; but it remains clear that a large proportion of these public data are redundant, and that precautions should be taken when performing meta-analyses on them.

We have not identified such problems with RNA-Seq data so far, but our data set is much smaller, as this data type was added into Bgee recently (Bgee release 12), and includes for now only 33 samples from one experiment (23), composed of 39 runs. Similar issues are likely to appear for RNA-Seq in the future.

By removing duplicated content from our database and by correcting annotations inconsistencies between identical samples, Bgee provides now what we believe to be a unique data set of duplicate-free, high quality Affymetrix data. This will enable us to prevent such problems from appearing in our growing RNA-Seq data set.

All Bgee data are freely available from our website (http://bgee.unil.ch/).

Supplementary Data

Supplementary data are available at *Database* Online.

Funding

This work was supported by the Swiss Institute of Bioinformatics, by the Swiss National Science Foundation [grant number 31003A 133011/1], and by Etat de Vaud.

Conflict of interest. None declared.

References

1. Bastian,F., Parmentier,G., Roux,J. *et al*. (2008) Bgee: integrating and comparing heterogeneous transcriptome data among species. In: Bairoch,A., Cohen-Boulakia,S. and Froidevaux,C (eds), *Data Integration in the Life Sciences*, Vol. 5109. Springer, Berlin/Heidelberg, pp. 124–131.
2. Barrett,T., Troup,D.B., Wilhite,S.E. *et al*. (2011) NCBI GEO: archive for functional genomics data sets, 10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
3. Parkinson,H., Sarkans,U., Kolesnikov,N. *et al*. (2011) ArrayExpress update – an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **39**, D1002–D1004.
4. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
5. Liu,W.-m., Mei,R., Di,X. *et al*. (2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, **18**, 1593–1599.
6. Federal Information Processing Standards (2002) Secure Hash Standard. *Publication* 180–2.
7. R Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
8. Tang,Z.H. and Feldman,A.M (2003) *Heart Failure in TNF-Alpha Transgenic Mice*, NCBI GEO, GSE591.
9. Cvijanovich,N., Shanley,T.P., Lin,R. *et al*. (2008) Validating the genomic signature of pediatric septic shock. *Physiol. Genomics*, **34**, 127–134.
10. Wynn,J.L., Cvijanovich,N.Z., Allen,G.L. *et al*. (2011) The influence of developmental age on the early transcriptomic response of children with septic shock. *Mol. Med.*, **17**, 1146–1156.
11. Shanley,T.P., Cvijanovich,N., Lin,R. *et al*. (2007) Genome-level longitudinal expression of signaling pathways and gene networks in pediatric septic shock. *Mol. Med.*, **13**, 495–508.
12. Wong,H.R., Cvijanovich,N., Allen,G.L. *et al*. (2009) Genomic expression profiling across the pediatric systemic inflammatory response syndrome, sepsis, and septic shock spectrum. *Crit. Care Med.*, **37**, 1558–1566.
13. Wong,H.R., Cvijanovich,N., Lin,R. *et al*. (2009) Identification of pediatric septic shock subclasses based on genome-wide expression profiling. *BMC Med.*, **7**, 34.
14. Welle,S., Tawil,R. and Thornton,C.A. (2008) Sex-related differences in gene expression in human skeletal muscle. *PLoS One*, **3**, e1385.
15. Osborne,R.J., Welle,S., Venance,S.L. *et al*. (2007) Expression profile of FSHD supports a link between retinal vasculopathy and muscular dystrophy. *Neurology*, **68**, 569–577.
16. Krishnan,K., Salomonis,N. and Guo,S. (2008) Identification of Spt5 target genes in zebrafish development reveals its dual activity in vivo. *PLoS One*, **3**, e3621.
17. Guo,S. (2008) *Foggy Mutant: Brain (Guo-1R01NS042626-01A2)*. NCBI GEO, GSE13348.
18. Suzek,B.E., Huang,H., McGarvey,P. *et al*. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.
19. Wu,Z. and Irizarry,R.A. (2004) Preprocessing of oligonucleotide array data. *Nat. Biotech.*, **22**, 656–658.
20. Irizarry,R.A., Hobbs,B., Collin,F. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
21. Schinke,M., Shioi,T., Riggi,L. *et al*. (2003) *Overexpression of dn-p21ras as a Model System for Severe Dilated Cardiomyopathy*, NCBI GEO, GSE760.
22. Schinke,M., Riggi,L., Chen,I. and Izumo,S (2002) *FVB Benchmark Set for Cardiac Development, Maturation, and Aging*, NCBI GEO, GSE75.
23. Brawand,D., Soumillon,M., Necsulea,A. *et al*. (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.