Integrating Expression Data from Different Microarray Platforms in Search of Biomarkers of Radiosensitivity.

Anna Papiez[1], Paul Finnon[2], Christophe Badie[2],
Simon Bouffler[2], and Joanna Polanska[1]

[1] Silesian University of Technology, Institute of Automatic Control,
Akademicka 2A, 44-100 Gliwice, Poland
{anna.papiez,joanna.polanska}@polsl.pl
[2] Public Health England, Centre for Radiation, Chemical and Environmental Hazards
Chilton, Didcot, Oxfordshire, OX11 ORQ, United Kingdom
{paul.finnon,christophe.badie,simon.bouffler}@phe.org.uk

Abstract. The goal of this study was to propose a method for meta-analysis of expression sets of single and dual channel intensity microarray data. This involved solving the issue of computational and biological consistency of the expression measures. We tested this approach on two sets from microarray data acquired in experiments performed to search for genetic biomarkers of radiosensitivity. The expression sets were unified taking into account the technical aspects of the design of the experiment and commonly used algorithms for the removal of batch effects. The resulting genes were subject to annotation analysis for enrichment in ontologies and signaling pathways.

Keywords: microarray, cDNA, oligonucleotide, data integration

1 Introduction

Breast cancer is the most common occurring in women in terms of new cases observed yearly and the second type of tumor that causes death in the female population [1]. One of the means of treatment in this case is radiotherapy that relies on the application of ionizing radiation to tumor cells. This type of remedy, depending on the patient, may result in late adverse effects that diminish the quality of life after therapy. This case, known as radiosensitivity, is studied with the goal of creating diagnostic tests that would allow to personalize the treatment in terms of dose and time intervals between fractions. One of the approaches aims at finding a genetic signature of radiosensitivity.

Microarray technology is a method of high-throughput analysis of gene expression that enables simultaneous examination of thousands of features in search of markers involved in a studied biological condition. The Gene Expression Omnibus (GEO) [2] is a constantly growing public repository that holds the possibility of merging data sets in order to forge a comprehensive image of examined cases. However, meta-analysis of gene chips often brings forward issues implying the need of data processing for biological as well as numerical consistency.

In previous studies, various methodologies have been considered for combining biochip data sets across platforms. As simple approaches such as standardization and mean-centering had their limitations, more complex concepts started to emerge. Parmigiani *et al.* [3] introduced the Probability of Expression method, which transforms expression data to signed probabilities. Benito *et al.* proposed Distance Weighted Discrimination, relying on Support Vector Machines [4]. Breitling *et al.* present the Rank Product computation scheme [5], Johnson *et al.* created the Empirical Bayes methods [6] and Shabalin *et al.* developed cross-platform normalization (XPN) based on iterative k-means clustering [7]. These algorithms have been evaluated in numerous studies on merging multiple microarray data sets [8, 9], yet it seems that the question of integration of platforms of different nature has not been attended to.

We developed an approach that addresses the particularly intricate issue of combining data sets from two types of microarrays: oligonucleotide and cDNA. In this case the problem lies in the different nature of the signal resulting from the platforms being respectively one channel and two channel data. The data sets were produced in the course of two studies which were designed for assessing differential expression in genes related with radiosensitivity.

2 Materials and Methods

2.1 Data Sets

For the purpose of this study data sets from two independent experiments were used. One provided blood samples from 60 breast cancer patients, of which 30 were classified as radiosensitive (RS) and 30 as radioresistant (RR). The blood was divided into two portions - one sample per patient left as control, the other was irradiated with a 2 Gy high dose of X-rays. After 24h, lymphocytes were filtered and RNA extracted for amplification and labeling in the microarray experiment. In this case the HuGene 1.0 ST Affymetrix oligonucleotide chips were used, which provide raw intensity CEL files.

As for the second experiment, samples were gathered from 59 patients: 31 radiosensitive and 28 radioresistant [10]. The irradiated samples were subject to 4 Gy of X-rays. Again, after 24h, lymphocytes were filtered and RNA extracted. This procedure was carried out on custom cDNA Breakthrough 20K arrays. The experiment was designed in a dye-swap manner, such that each sample was labeled with the cy3 and cy5 dye and hybridized to the chip against a reference sample from a pooled set of 30 breast cancer cell lines. This data was obtained as a set of GPR files produced by the GenePix 5.1 scanning software.

The scheme in Figure 1 presents the typical course of a meta-analysis of expression values. The diagram in Figure 2 illustrates the work flow for our strategy.

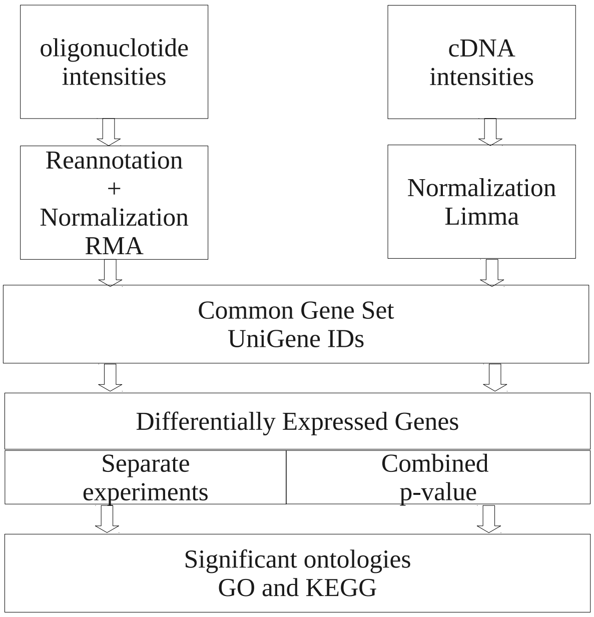
oligonuclotide intensities
cDNA intensities
Reannotation + Normalization RMA
Normalization Limma
Common Gene Set UniGene IDs
Differentially Expressed Genes
Separate experiments
Combined p-value
Significant ontologies GO and KEGG

Fig. 1. Work flow for a standard microarray comparative analysis.

2.2 Preprocessing

The oligonucleotide single channel data was normalized using the Robust Multichip Average (RMA) method [11] which consists of background correction of perfect-match (PM) intensities, quantile normalization and summarization using the median polish algorithm. Probes were reannotated with a custom chip description file [*hugene10st_Hs_ENTREZG version 1.36.0 May 10, 2013*] from the Brainarray database [12].

For the sake of comparison of data from two different microarray platforms, we adopted an approach where the main concern was to obtain data of the same character, regarding the numerical as well as biological aspect. Thus, we extracted the intensity data for separate channels and included intensities for the patients' samples (red or green channel) for further investigation, omitting the information on breast cancer cell lines. This was motivated by retaining consistency in terms of the biological representation of the signal, as there is no such reference available in the oligonucleotide array experiment, and additionally, the unfeasibility of juxtaposing expression values in oligonucleotide chip data with ratios of expression from cDNA arrays. Standard normalization resulting in ratios of the intensities was also performed in the separate experiment normalization scheme for comparative purposes.

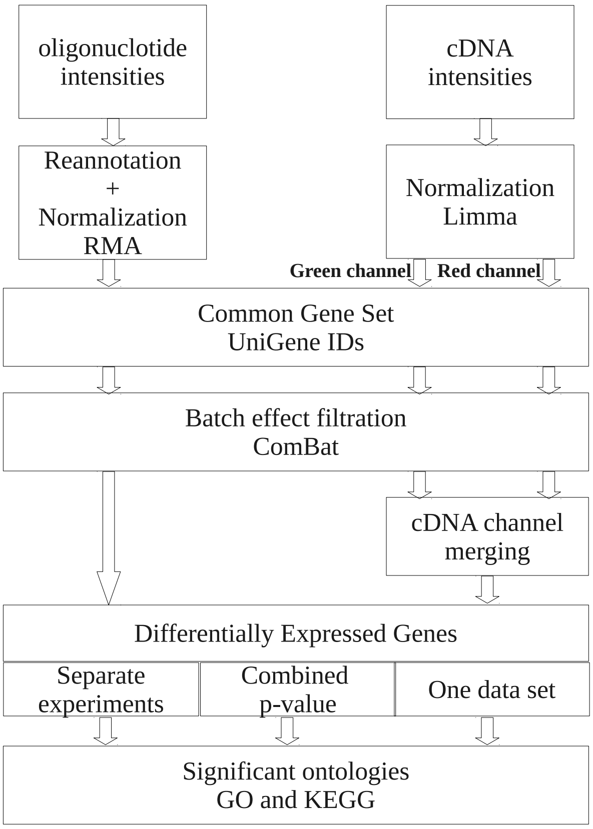
oligonuclotide intensities
cDNA intensities
Reannotation + Normalization RMA
Normalization Limma
Green channel
Red channel
Common Gene Set UniGene IDs
Batch effect filtration ComBat
cDNA channel merging
Differentially Expressed Genes
Separate experiments
Combined p-value
One data set
Significant ontologies GO and KEGG

Fig. 2. Scheme of the proposed microarray data integration procedure.

The cDNA microarrays were preprocessed with the Bioconductor Limma package [13]. The values were background adjusted using the *normexp* algorithm. For reasons described in the previous paragraph, within array normalization was left out and between array normalization was performed with the *quantile* method. This resulted in an expression set of patients' samples in two replicates, one for each color channel.

2.3 Data Integration

The first step of data set integration was to extract a set of genes common for both platforms. This was done on the basis of UniGene identifiers. For the common gene sets, the expression values were transferred to a unified scale with the use of Empirical Bayes Methods implemented in the ComBat software for removal of batch effects provided in the R SVA package. Since the red and green channel data have been filtered for batch effects, their expression was merged as for technical replicates. For determining a signature of radiosensitivity, both irradiated and control samples were tested.

2.4 Statistical Analysis

The genes in individual groups of irradiated and control samples were tested for statistically significant differential expression with either of the three tests: Student's t-test, Welch's t-test or U-Mann-Whitney test, depending on prior determination of normality (Lilliefors test) and variance homogeneity (F-test). This analysis was carried out for both types schemes: the typical separate normalization of data from two experiments and the proposed here unification of data using batch effect filtration.

Apart from considering results for the two data sets and creating a list of common differentially expressed genes, an approach reported by Liu *et al.* [14] was implemented to procure combined p-values from the two experiments. Moreover, as the data sets after applying batch effect removal can be considered numerically compatible, the samples for this method have been merged into one set and tested for differential expression.

The genes classified as differentially expressed were investigated for annotations to ontologies in the GO [15] and KEGG [16] databases.

3 Results

3.1 Data Integration

In the course of this study, two expression data sets were combined to gain numerical consistency using empirical Bayes methods for filtering batch effects. The results of this stage of integration are presented in Figure 3.

Moreover, the data was merged so as to attain coherence in the biological sense. For this purpose, we retained raw intensity signals in the cDNA experiment for one channel - red or green referring to the patient sample and, after batch effect removal, this information was averaged over the two dye-swap replicates. Furthermore, the intersection of probes from the two biochips based on UniGene identifiers was incorporated in subsequent research. The quantity of genes retained at this point is illustrated in Figure 4.

3.2 Identification of Differentially Expressed Genes

Statistical testing was performed on the genes for samples normalized separately within the two experiments as well as samples processed with batch effect filtration. These tests were carried out for the determination of differentially expressed genes for RR vs RS patients in experiments considered independently, with the

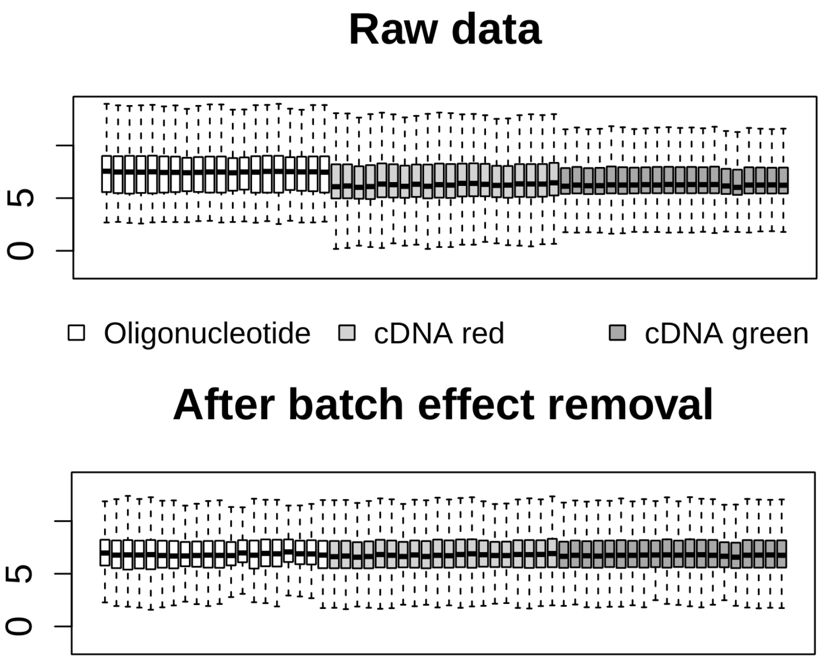

Fig. 3. Exemplary boxplots from each of the normalized groups of signals before and after batch effect filtration.

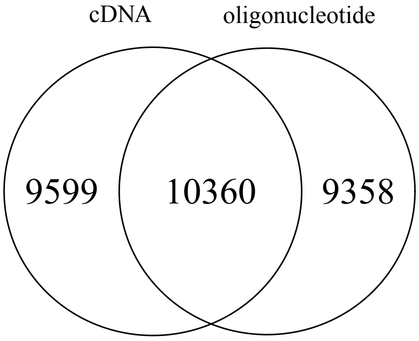

Fig. 4. Venn diagram illustrating the proportion of genes common for both microarray platforms.

combined p-values approach described in Section 2.4 and with data from the two experiments integrated into one sample. The number of genes for control lymphocytes is reported in Table 1, for irradiated - in Table 2.

	(A) separate normalization	(B) batch effect adjustment	(A∩B) Intersection
oligonucleotide	577	577	577
cDNA	922	1093	380
Common	44	53	12
Combined p-value	115	111	29
One data set	–	3146	–

Table 1. Number of differentially expressed genes at the significance level of 5% for control samples.

	(A) separate normalization	(B) batch effect adjustment	(A∩B) Intersection
oligonucleotide	633	633	633
cDNA	669	1159	289
Common	38	51	12
Combined p-value	71	100	18
One data set	–	3526	–

Table 2. Number of differentially expressed genes at the significance level of 5% for irradiated samples.

It can be seen that when integrating the data from both experiments, in most cases the algorithm involving batch effect removal results in a larger amount of differentially expressed genes. Additionally, the possibility of merging the expression sets into one produced notably considerable numbers of genes classified as differentially expressed. This may have been expected when handling meta-analysis, as increasing the number of samples enhances the power of statistical testing, yet such discrepancy in relation to the single-study concept requires further investigation. Another factor that clearly affected the outcome is the single-channel approach applied to the cDNA array data for improved comparability of the results.

The differentially expressed genes common for the two studies, resulting from combined p-values and merging the data into one set, were tested for statistically significant ontologies in the GO and KEGG databases, using the hypergeometric test. Ontologies and pathways were assumed to be statistically significant at the confidence level of 95%. In general the differentially expressed genes reported in the batch effect filtration approach were linked to a wider range of ontologies and pathways, in particular those related to radiation induced processes that have not been revealed when using data normalized separately.

Among others, a strong group of processes and functions pointed to the MAPK signaling pathway which has been reported to play a key role in the molecular background of radiosensitivity [17]. Additionally, annotations to the radiation-related p53 regulation [18] and mTor [19] pathways occurred. Other annotations such as cellular response to stress, apoptosis and regulation of cell death may confirm the link between the identified genes and radiosensitivity.

4 Conclusions

Cross-studies of high-throughput genomic data constitute a valuable solution to the problem of overdimensionality, though they hold a challenge in terms of transformation of the expression values to achieve computational and biological consistency. This issue becomes more complex when the design of the compared platforms is of different nature. We established a procedure for integrated study of data from oligonucleotide and cDNA microarrays, that enables merging of the expression sets equivalent in the numerical form and within the analyzed biological condition. This method has its limitations, ie. the loss of information about features unique for either of the platforms. However, due to the raise of statistical power after enabling data merging our investigation resulted in an increase of information about differentially expressed genes and the additional features have been shown to be annotated to radiosensitivity linked processes. The validation of the presented methods on supplementary analogous data and examination of the applicability for profile identification will be performed in the future.

References

1. Siegel, R., Naishadham, D., Jemal, A.: Cancer statistics, 2012. CA Cancer J Clin **62**(1) (2012) 10–29
2. Edgar, R., Domrachev, M., Lash, A.E.: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30**(1) (2002) 207–210
3. Parmigiani, G., Garrett, E.S., Anbazhagan, R., Gabrielson, E.: A statistical framework for expression-based molecular classification in cancer. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**(4) (2002) 717–736

4. Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D.and Perou, C.M., Marron, J.S.: Adjustment of systematic microarray data biases. Bioinformatics **20**(1) (2004) 105–114
5. Breitling, R., Armengaud, P., Amtmann, A., Herzyk, P.: Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS letters **573**(1) (2004) 83–92
6. Johnson, W.E., Li, C., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics **8**(1) (2007) 118–127
7. Shabalin, A.A., Tjelmeland, H., Fan, C., Perou, C.M., Nobel, A.B.: Merging two gene-expression studies via cross-platform normalization. Bioinformatics **24**(9) (2008) 1154–1160
8. Sîrbu, A., Ruskin, H.J., Crane, M.: Cross-platform microarray data normalisation for regulatory network inference. PloS one **5**(11) (2010) e13822
9. Liu, Z., Xie, M., Yao, Z., Niu, Y., Bu, Y., Gao, C.: Three meta-analyses define a set of commonly overexpressed genes from microarray datasets on astrocytomas. Molecular neurobiology **47**(1) (2013) 325–336
10. Finnon, P., Kabacik, S., MacKay, A., Raffy, C., A'Hern, R., Owen, R., Badie, C., Yarnold, J., Bouffler, S.: Correlation of in vitro lymphocyte radiosensitivity and gene expression with late normal tissue reactions following curative radiotherapy for breast cancer. Radiother Oncol **105**(3) (Dec 2012) 329–336
11. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4**(2) (Apr 2003) 249–264
12. Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., Watson, S.J., Meng, F.: Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. **33**(20) (2005) e175
13. Smyth, G.K.: Limma: linear models for microarray data. In Gentleman, R.and Carey, V., Dudoit, S., Irizarry, R., Huber, W., eds.: Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, New York (2005) 397–420
14. Liu, Z., Niu, Y., Li, C., Yang, Y., Gao, C.: Integrating multiple microarray datasets on oral squamous cell carcinoma to reveal dysregulated networks. Head Neck **34**(12) (Dec 2012) 1789–1797
15. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25**(1) (May 2000) 25–29
16. Kanehisa, M., Goto, S.: KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **28**(1) (Jan 2000) 27–30
17. Chung, E.J., Brown, A.P., Asano, H., Mandler, M., Burgan, W.E., Carter, D., Camphausen, K., Citrin, D.: In vitro and in vivo radiosensitization with AZD6244 (ARRY-142886), an inhibitor of mitogen-activated protein kinase/extracellular signal-regulated kinase 1/2 kinase. Clin. Cancer Res. **15**(9) (May 2009) 3050–3057
18. Mirzayans, R., Andrais, B., Scott, A., Wang, Y.W., Murray, D.: Ionizing Radiation-Induced Responses in Human Cells with Differing TP53 Status. Int J Mol Sci **14**(11) (2013) 22409–22435
19. Steelman, L.S., Navolanic, P., Chappell, W.H., Abrams, S.L., Wong, E.W., Martelli, A.M., Cocco, L., Stivala, F., Libra, M., Nicoletti, F., Drobot,

L.B., Franklin, R.A., McCubrey, J.A.: Involvement of Akt and mTOR in chemotherapeutic- and hormonal-based drug resistance and response to radiation in breast cancer cells. Cell Cycle **10**(17) (Sep 2011) 3003–3015

Acknowledgment

This work was supported by the National Science Center grant Harmonia 4 DEC-2013/08/M/ST6/00924.