UDC 577.218 https://doi.org/10.15407/biotech14.01.38

CONSECUTIVE INTEGRATION OF AVAILABLE MICROARRAY DATA FOR ANALYSIS OF DIFFERENTIAL GENE EXPRESSION IN HUMAN PLACENTA

O. LYKHENKO, A. FROLOVA, M. OBOLENSKA

Institute of Molecular Biology and Genetics
of the National Academy of the Sciences of Ukraine, Kyiv

E-mail: lykhenko.olexandr@gmail.com

Received 18.12.2020
Revised 18.12.2020
Accepted 27.01.2021

The purpose of the study was to provide the pipeline for processing of publicly available unprocessed data on gene expression via integration and differential gene expression analysis.

Data collection from the open gene expression databases, normalization and integration into a single expression matrix in accordance with metadata and determination of differentially expressed genes were fulfilled. To demonstrate all stages of data processing and integrative analysis, there were used the data from gene expression in the human placenta from the first and second trimesters of normal pregnancy.

The source code for the integrative analysis was written in the R programming language and publicly available as a repository on GitHub. Four clusters of functionally enriched differentially expressed genes were identified for the human placenta in the interval between the first and second trimester of pregnancy.

Immune processes, developmental processes, vasculogenesis and angiogenesis, signaling, and the processes associated with zinc ions varied in the considered interval between the first and second trimester of placental development. The proposed sequence of actions for integrative analysis could be applied to any data obtained by microarray technology.

Key words: microarray, transcript, integrative analysis, Bayesian empirical method, meta-analysis, differentially expressed genes, placenta.

The last two decades in biology are marked by the emergence and rapid advancement of high throughput gene expression studies, allowing to explore the systemic patterns over the entire gene expression profiles as opposed to focusing on a small subset of particular genes. The most frequently asked question is what is the difference in gene expression between the physiological and pathological cases, between the two time points of the developing system or between the medical therapy and placebo, etc. One of the answers to for this question is the list of differentially expressed genes which the researcher analyses and interprets.

As high-throughput technology usage is often quite expensive compared to qPCR, researchers usually tend to resort to a small number of samples. In order to increase the statistical power, it often makes sense to merge or integrate unprocessed data from several similar studies to get a larger sample size. The method of merging unprocessed data is called the integrative analysis, which has advantages [1–3] over meta-analysis that is a merging of differentially expressed genes lists from different studies. The crucial point in gene expression integrative analysis is a cross experiment normalization of the expression data. Gene expression tables from two datasets cannot be simply concatenated to form an integrated dataset. Gene expression values in both datasets correspond to the relative fluorescence intensity values of microarray chips. The expression values are systematically differ in two different datasets due to technical reasons. Similar like the photos of the same object taken in different lighting environments using different photo cameras. This systematic technical difference, called batch effect, can be removed by adjusting location and scale of gene expression distribution (mean and variance if the last is normal) in samples from different batches.

A more advanced technique would be to separate samples from each batch into groups according to their metadata and adjust location and scale in these groups separately. For example, let's assume we have two studies that are batch A and batch B. Batch A and B have norm (AN, BN) and disease (AD, BD) respectively. The idea is to adjust gene expression distributions in AN with BN and in AD with BD instead of just in A with B. This principle is used in the ComBat algorithm (or Empirical Bayes method) [4, 5] which we use for the batch effect removal.

Despite rapid emergence of next-generation sequencing, microarray based studies remain dominant by the amount of available datasets. We made a search over the ArrayExpress database which returned 3602 homo sapiens RNA sequencing studies (210 146 samples) versus 18 829 microarray studies (871 721 samples), 4 times more than sequencing. In cases when the subject of research is not popular, namely, non-cancer studies, this difference can be even larger. For instance, the same Array Express query for placenta studies only returned 39 RNA sequencing studies (10 409 samples) and 148 microarray studies (48 214 samples), that is 5 times as much.

Affymetrix microarray chip consists of about 500,000 cells (up to ~6,500,000) each filled with multiple copies of a unique short 25 bp long oligonucleotide sequence, called probes. A set of probes matching the same gene is called a probeset. Microarray chips are intended to be manufactured in a way that each probe uniquely maps onto a single gene. However, as our knowledge about the genome extends, some probes turn out to be matching several genes and it makes sense to reanalyze existing unprocessed data using newer gene annotation. Brainarray (available at http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/) project provides such up-to-date gene annotation for Affymetrix chips.

A technical variation between samples is inherent to microarray experiments. It is addressed by normalization, log-transforming and background correcting.

The aim of this article is to exemplify the sequence of specific procedures or the pipeline from the unprocessed data of the fluorescence intensity to the list of differentially expressed genes. To this end, we use two groups of samples, the gene expression in the human placenta from the first and the second trimester of gestation. We intend to show how to recognize the difference in gene expression between two time points of placental development and represent it in the list of differentially expressed genes.

The pipeline

Here we provide a step-by-step walkthrough of our pipeline. We chose four datasets containing samples of the healthy human placenta at I and II trimesters of pregnancy. The R packages used for the pipeline can be downloaded from the Bioconductor website at https://bioconductor.org/packages/release/bioc/html/ArrayExpress.html. The entire pipeline code used in this study along with the supplementary files are available at GitHub: https://github.com/Sashkow/r-affymetrix-integration-pipeline

Table 1. **General characteristics of the samples**

GSE Accession number	Platform	First trimester	Second trimester	Number of genes
GSE122214 [20]	Affymetrix HG-U133_Plus_2	4	–	20261
GSE22490 [21]	Affymetrix HG-U133_Plus_2	5	1	20261
GSE37901 [22]	Affymetrix HG-U133_Plus_2		4	20342
GSE9984 [23]	Affymetrix HG-U133_Plus_2	4	4	20342
Total number		13	9	–

Data extraction. We downloaded the corresponding expression data from ArrayExpress by accession numbers with ArrayExpress R package (Table1). Only the samples from the first and second trimester healthy placentas were filtered in from these datasets according to metadata in our specialized database [6].

Expression data reprocessing

We used affy [7] and R package to access and preprocess expression data from Affymetrix initial. CEL files that store the results of the intensity calculations on the pixel values. A single representative intensity value is stored per cell (feature) of the image. The Affy package uses robust multi-array average (RMA) algorithm [8] based on quantile normalization algorithm [9]. RMA maps probes to genes, then the initial intensity values are background corrected, log2 transformed and quantile normalized.

Mapping probes to genes. When mapping probes to genes, one must decide which probe or a combination of probes to choose as an indicator of each particular gene expression. A comparative analysis of such methods is provided in [10]. It concludes that for Affymetrix chips one-to-one probeset to gene correspondence is best established in a form of custom chip definition Files (CDF) provided in Brainarray database [11]. Reannotation leads to improvement in microarrays accuracy.

Cross-experiment normalization. Non-biological experimental variation has to be excluded before integration of different datasets. We used the empirical Bayes method, a procedure for statistical inference in which the prior distribution is estimated from the data contrast to standard Bayesian methods, for which the prior distribution is fixed before any data are observed. Particular implementation of this method (ComBat function in limma R package) incorporates systematic batch biases common across genes by making adjustments, assuming that phenomena resulting in batch effects often affect many genes in similar ways (i.e. increased expression). Specifically, it estimates parameters that represent the batch effects by pooling information across genes in each batch (usually a dataset from an independent project) to shrink the batch effect parameter estimates toward the overall mean of the batch effect estimates across genes [12]. The data are then transformed to remove the effects of the different batch effect parameters across experiments. Shrinkage is a general technique of moving the observed data toward the mean. For example, two extreme mean values can be combined to make one more centralized mean value; repeating this for all means in a population sample will result in a revised mean that has “shrunk” towards the true population mean.

Since fetal sex affects the genes expression in sex and autosomal chromosomes [13], we needed to include fetal sex as a confounding biological variable into a model matrix during batch effect removal. To this end, we used massiR R package [14] and applied it to the data on gene expression associated with Y-chromosome.

*Batch effect removal validation.*We used principal component analysis [15] (PCA, prcomp function in R) and t-distributed stochastic neighbor embedding [16] (TSNE, Rtsne package in R) (to check the results of batch effect removal (Fig. 1, *A–E*). Fig. 1, *A* represents the distribution of all the data from four datasets before batch removal, indicating that the largest source of variation is the dataset’s id, which is captured by the first and second PCA components (dim1, dim2). This is no longer the case after the batch effect removal as shown on Fig. 1, *B*. Oppositely, biological variation of the same data caused by trimester number, which was invisible before batch effect removal (Fig. 1, *C*), was preserved and became apparent after batch effect removal (Fig. 1, *D*). Biological variation by fetus sex also became visible and captured by components 3 and 4 (dim3, dim4) (Fig. 1, *E*).

Figures *A* and *B* represent the distribution of data along with the first and second components of PCA (dim1, dim2) before and after the data integration and the batch effect removal from four data sets, correspondingly. Figures *C* and *D* represent the distribution of data from the first and the second trimesters of gestation before (Fig. 1, *C*) and after the batch effect removal (Fig. 1, *D*). Fig. 1, *E* represents the separation of the data according to the sex of the fetus after the batch effect removal. The influence of the batch-effect caused by different datasets is present before integration and the batch effect removal (Fig. 1, *A*) and is no longer seen afterwards, while the biological variability (Fig. 1, *D*, *E*) is preserved and becomes apparent after the batch effect removal (*D*, *E*). Note that fetus sex (*E*) is shown for principal components 3 and 4.

Differential gene expression and gene ontology enrichment analysis. For

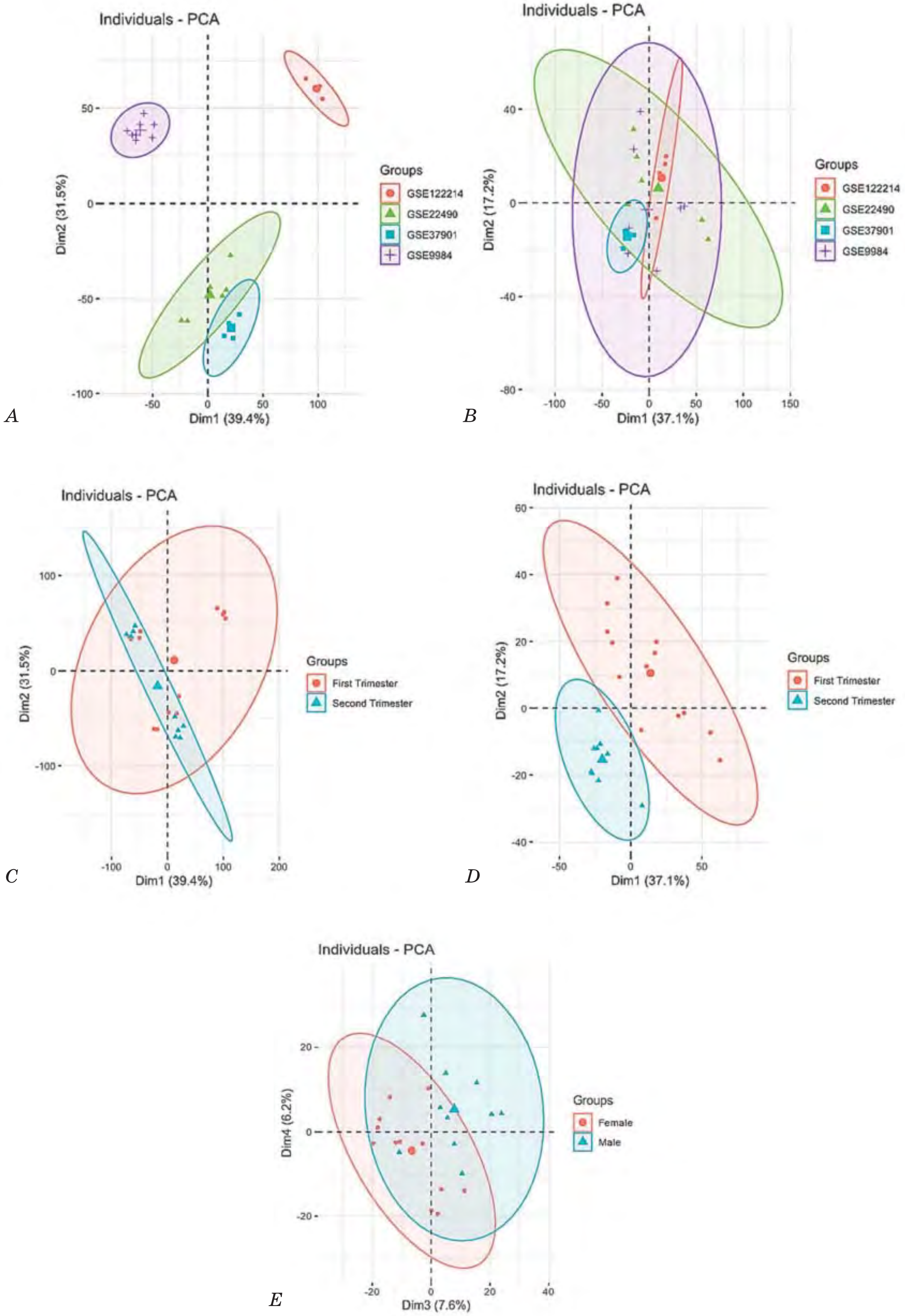

Fig. 1. **The results of PCA analysis before and after data integration and the batch effect removal**

the integrated datasets, we found 3912 differentially expressed genes between the second and the first trimesters of gestation and 327 genes with a value |logfc>1|. We used generalized linear models implemented in the limma R package for differential expression analysis. A linear model was fit to each gene with moderated t-statistics computed. P-value was adjusted to account for multiple gene comparisons with Benjamini & Hochberg method (FDR) [17].

Gene Ontology enrichment analysis. In order to provide biological interpretation for the newly found differentially expressed genes we clustered closely related genes relying on the interaction data from the String database [18]. After that, we run Gene Ontology enrichment analysis on each cluster. We downloaded an interaction graph for the 327 differentially expressed. The graph's vertices are proteins coded by differentially expressed genes, and the edges are interactions (either direct protein-protein interactions or indirect relations such as belonging to the same cellular pathway). String interaction data is manually curated in a way that each interaction is scored with a confidence (a float value ranging from 0 to 1) according to the strength of the evidence provided for a particular interaction. Choosing a lower confidence for a graph building gives more interactions to work with, which leads to less isolated genes, while a larger confidence gives better clustering modularity, a clustering quality metric, which measures a fraction of graph edges that connect vertices from the same cluster. Choosing too low confidence will lead to the formation of few or even one big cluster with few and very general enriched categories. Oppositely, choosing too high confidence will lead to lots of disconnected genes and too specific enrichment with low p-value. Therefore we optimized the confidence to get a graph that gives the highest total cluster coverage, which we define as a fraction of genes in clusters that belong to at least one enriched biological process. In our case the confidence value turned out to be 0.1.

After mapping the differentially expressed genes to the String identifiers and excluding isolated vertices, 268 genes were left in the graph. We then applied a fastgreedy clustering to the graph data and got 8 clusters, 4 of which having coverage > 0. Table 2 contains cluster statistics for these clusters. Cluster enrichment coverage column indicates a fraction of genes which are a part of at least one enriched process. The graph itself is on Fig. 2.

Cluster names are given by the category name in the cluster with the lowest p-value. Two clusters, the immune response process and the vessel and organ development, contain the majority of differentially expressed genes. During pregnancy, the mother's immune system has to tolerate the persistence of paternal alloantigens without affecting the anti-infectious immune response. Consequently, several mechanisms aimed at preventing allograft rejection, occur during a pregnancy. In fact, the early stages of pregnancy are characterized by the correct balance between inflammation and immune tolerance, in which proinflammatory cytokines contribute to both the remodeling of tissues and to neo-angiogenesis, thus, favoring the correct embryo

Table 2. **Enriched clusters of differentially expressed genes between first and second trimesters of physiological gestation**

Cluster names	**Amount of genes**			**Log Fc Min**	**Log Fc Max**	**Log Fc Mean**	**Cluster enrichment coverage**
	Total	Up*	Down**				
1. Immune system process	114	81	33	−3.2	2.97	0.48	0.87
2. Vessel and organ development	122	62	60	−3.2	2.15	−0.06	0.88
3. Zinc ion response	10	3	7	−4.4	1.55	−1.21	1
4. Cell surface signalling pathways	7	6	1	−2.7	2.25	1.1	1
Total	253	152	101	–	–	–	–

Note: Up* — up-regulated genes, Down** — down-regulated genes.

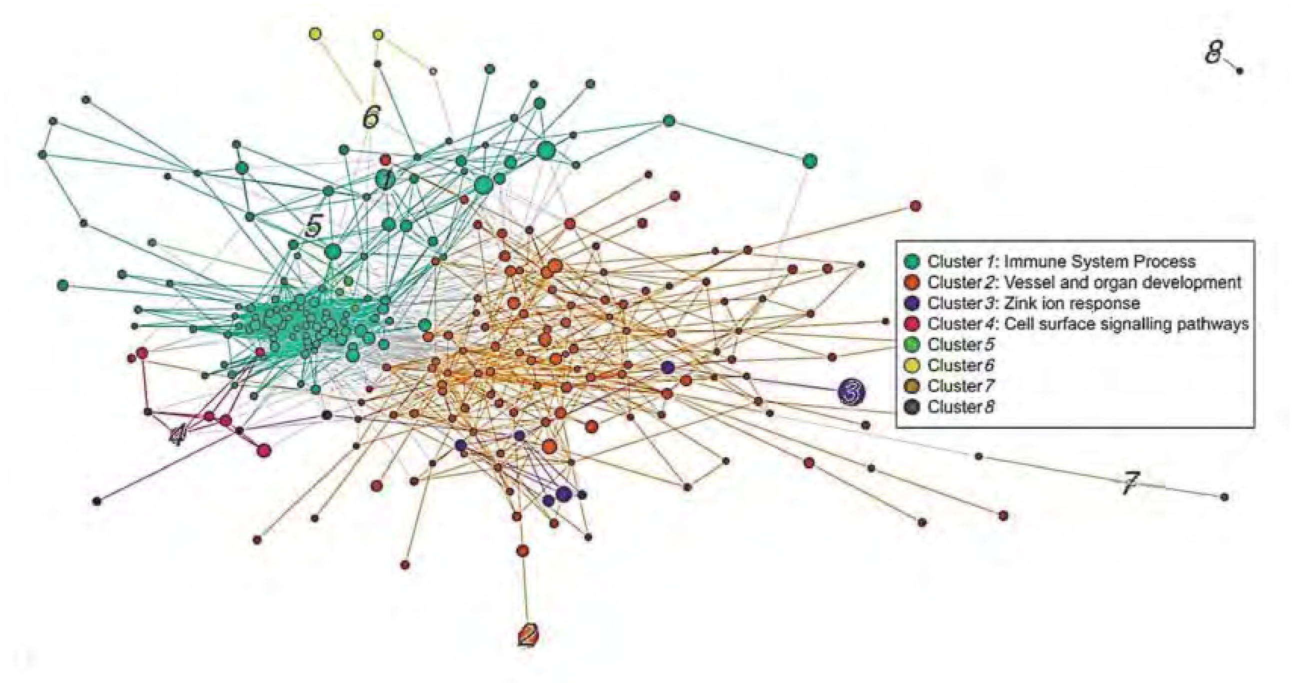
6
8
5
1
3
4
7
2
● Cluster 1: Immune System Process
● Cluster 2: Vessel and organ development
● Cluster 3: Zink ion response
● Cluster 4: Cell surface signalling pathways
● Cluster 5
● Cluster 6
● Cluster 7
● Cluster 8

Fig. 2. **Gene interaction graph for differentially expressed genes in human placenta (I vs II trimester) based on String data**

implantation. In addition to the creation of a microenvironment able to support the immunological privilege, the trophoblast supports the spiral artery remodeling and angiogenesis. An increase in the number of regulatory T (Treg) cells prevents excessive inflammation and avoids fetal immune-mediated rejection [19].

Cell surface signalling pathways are indispensable for the morphogenetic processes. We now know that each type of a cell has a different set of proteins in its surfaces, and that some of these differences are responsible for forming the structure of the tissues and organs during development. Different cell types have different types and different amounts of cell adhesion molecules, which define the intercellular binding and the mutual cellular localization in the tissue [20].

Zn ion response

Zinc is as common as iron in biology. Involvement of zinc is common in biology but it is at very different analytical concentrations. It is usually thought to be a trace element required only for catalysis but it's much more fundamental. Zinc ions are involved in regulating intracellular signaling pathways in innate and adaptive immune cells and play the role of gatekeepers of the immune function [21]. A property peculiar to zinc is the absence of redox chemistry. So, many zinc enzymes are used specifically in situations where the presence of redox reactions, which are typical for the developing placenta, would lead to damaging radicals and/or preferential reactions with oxygen or hydrogen peroxide [22].

Biological interpretation of the differentially expressed genes, obtained through integration of four openly available microarray datasets, reveals the biological processes that occur in human placenta at the early periods of gestation, namely, immune inflammation and immune tolerance, angiogenesis, organ development and morphogenesis, regulation of growth. The described pipeline is applicable to a wide range of biological cases investigated by the microarray-technology.

The study was conducted in the frame of the budget project #2.2.4.18 and supported by the National Academy of Sciences of Ukraine. The authors declare that they have no conflicts of interest.

REFERENCES

1. *Taminau J., Lazar C., Meganck S., Nowé A.* Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis. *ISRN Bioinform*. 2014,V. 2014, P. 345106. https://doi.org/10.1155/2014/345106
2. *Uitert M., Moerland P. D., Enquobahrie D. A., Laivuori H., Joris A. M. van der Post, Ris-Stalpers C., Afink G. B.* Meta-analysis of placental transcriptome data identifies a novel molecular pathway related to preeclampsia. *PLoS One*. 2015, V. 10, P. e0132468.
3. *Cosmin L., Meganck S., Taminau J., Steenhoff D., Coletta A., Molter C., Weiss-Solís D. Y., Duque R., Bersini H., Nowé A.* Batch effect removal methods for microarray gene expression data integration: a survey. *Briefings Bioinf*. 2013, 14 (4), 469–490. https://doi.org/10.1093/bib/bbs037
4. *Turnbull A. K., Kitchen R. R., Larionov A. A., Renshaw L., Dixon J., Sims A. H.* Direct integration of intensity-level data from Affymetrix and Illumina microarrays improves statistical power for robust reanalysis. *BMC Medical Genomics*. 2012, 5 (1), 35.
5. *Tseng G. C., Ghosh D., Feingold E.* Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research*. 2012, 40 (9), 3785–99.
6. *Lykhenko O., Frolova A., Obolenska M.* Creation of gene expression database on preeclampsia-affected human placenta. *Biopolym. Cell*. 2017, 33 (6), 442–452. https://doi.org/10.7124/bc.000967
7. *Gautier L., Cope L., Bolstad B. M., Irizarry R. A.* Affy — analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. 2004, 20 (3), 307–315
8. *Irizarry R. A., Bolstad B. M., Collin F., Cope L. M., Hobbs B., Speed T. P.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*. 2002, 31 (4), e15.
9. *Zhao Yaxing, Limsoon Wong, Wilson Wen Bin Goh.* How to do quantile normalization correctly for gene expression data analyses. *Sci. Rep*. 2020, 10 (15534), 1–11. https://doi.org/10.1038/s41598-020-72664-6
10. *Frolova A. O, Bondarenko V. S., Obolenska M. Yu.* Cross-platform integration of experimental microarrays and its effect on the value of gene expression in the analysis of human breast cancer samples. *Medychna informatyka ta inzheneriia*. 2016, No 2, P. 5–14.
11. *Sandberg R., Larsson O.* Improved precision and accuracy for microarrays using updated probe set definitions. *BMC Bioinf*. 2007, 8 (1), 48.
12. *Johnson W. Evan, Cheng Li, Rabinovic A.* Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007, 8 (1), 118–127. https://doi.org/10.1093/biostatistics/kxj037
13. *Clifton Vicki.* Sex-based functional features of human placentae. *Zdorove zhenshchyni*. 2011, No 4, P. 24–29. (In Ukrainian).
14. *Buckberry Sam, Stephen J. Bent, Tina Bianco-Miotto, Claire T. Roberts, Author Notes. massiR*: a method for predicting the sex of samples in gene expression microarray datasets. *Bioinformatics*. 2014, 30 (14), 2084–2085. https://doi.org/10.1093/bioinformatics/btu161
15. *Venables W. N., Ripley B. D.* Modern Applied Statistics with *S. Springer Springer-Verlag. New York*. 2002, 562 p.
16. *Van Der Maaten, Hinton L. J. P., Hinton G. E.* Visualizing High-Dimensional Data Using t-SNE. *J. Machine Learning Res*. 2008, V. 9, P. 2579–2605.
17. *Smyth G. K.* Bioinformatics and Computational Biology Solutions Using R and Bioconductor. limma: Linear Models for Microarray Data. *Springer*. 2005, P. 397–420. https://doi.org/10.1007/0-387-29362-0_23
18. *Szklarczyk D., Gable A. L., Lyon D., Junge A., Wyder S., Huerta-Cepas J., Simonovic M., Doncheva N. T., Morris J. H., Bork P.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019, 47 (Database), D607–D613. https://doi.org/10.1093/nar/gky1131
19. *Albonici L., Benvenuto M., Focaccetti C., Cifaldi L., Miele M.T., Limana F., Manzari V., Bei R.* PlGF Immunological Impact during Pregnancy. *Int. J. Mol. Sci*. 2020, 21 (22), 8714. https://doi.org/10.3390/ijms21228714
20. *Scott G. F.* Morphogenesis and Cell Adhesion. *Development Biology, 6th edition. Sunderland (MA): Sinauer Associates*. 2000, P. 325–335.
21. *Wessels I., Maywald M., Rink L.* Zinc as a Gatekeeper of Immune Function. *Nutrients*. 2017, 9 (12), 1286. https://doi.org/10.3390/nu9121286
22. *Williams R. J.* Zinc: what is its role in biology? *Endeavour*. 1984, 8 (2), 65–70. https://doi.org/10.1016/0160-9327(84)9004=0-1

ПОЕТАПНА ІНТЕГРАЦІЯ НАЯВНИХ У ВІДКРИТОМУ ДОСТУПІ ДАНИХ МІКРОЧИПІВ ДЛЯ АНАЛІЗУ ДИФЕРЕНЦІЙНОЇ ЕКСПРЕСІЇ ГЕНІВ У ПЛАЦЕНТІ ЛЮДИНИ

О. Лихенко
А. Фролова
М. Оболенська

Інститут молекулярної біології та генетики
НАН України, Київ

E-mail: lykhenko.olexandr@gmail.com

Метою роботи було навести послідовні етапи оброблення наявних у відкритому доступі даних із використанням біочипів для проведення їх інтеграції та аналізу диференційної експресії генів.

З відкритих баз даних було зібрано дані з генної експресії в плаценті з першого і другого триместрів вагітності людини. Дані нормалізовали, інтегрували їх в єдину матрицю експресії згідно з метаданими і визначили диференційно експерсовані гени.

Початковий код послідовності дій для проведення інтегративного аналізу написаний мовою програмування R і є у відкритому доступі у вигляді репозиторію на GitHub. З використанням інтегративного аналізу виявлено чотири кластери функціонально збагачених диференційно експресованих генів у плаценті людини в інтервалі між першим і другим триместрами вагітності.

Встановлено, що імунні процеси, процеси розвитку, васкулогенез і ангіогенез, сигналінг, а також ті, що пов'язані з іонами цинку, змінюються між першим і другим триместром розвитку плаценти. Запропоновану послідовність дій для проведення інтегративного аналізу можна застосовувати до будь-яких даних, отриманих за допомогою мікрочипів.

Ключові слова: мікромасив, транскриптом, інтегративний аналіз, емпіричний метод Баєса, метааналіз, диференційно експресовані гени, плацента.

ПОЭТАПНАЯ ИНТЕГРАЦИЯ ИМЕЮЩИХСЯ В ОТКРЫТОМ ДОСТУПЕ ДАННЫХ МИКРОЧИПОВ ДЛЯ АНАЛИЗА ДИФФЕРЕНЦИАЛЬНОЙ ЭКСПРЕССИИ ГЕНОВ В ПЛАЦЕНТЕ ЧЕЛОВЕКА

О. Лихенко
А. Фролова
М. Оболенская

Институт молекулярной биологии и генетики
НАН Украины, Киев

E-mail: lykhenko.olexandr@gmail.com

Целью работы было изложить последовательные этапы обработки имеющихся в открытом доступе данных с использованием микрочипов для проведения интеграции и анализа дифференциальной экспрессии генов.

Из открытых баз данных были собраны данные генной экспрессии в плаценте женщин с первого и второго триместров беременности. Данные нормализовали и интегрировали в единую матрицу экспрессии в соответствии с метаданными и определили дифферениально экспрессированные гены.

Исходный код последовательности действий для проведения интегративного анализа написан на языке программирования R и находится в открытом доступе в виде репозитория на GitHub. При использовании интегративного анализа выявлены четыре кластера функционально обогащенных дифференциально экспрессированных генов в плаценте человека между первым и вторым триместрами беременности.

Иммунные процессы, процессы развития, васкулогенез и ангиогенез, сигналинг и процессы, связанные с ионами цинка изменяются между первым и вторым триместром развития плаценты. Предложенная последовательность действий для проведения интегративного анализа может быть применена к любым данным, полученным с помощью микрочипов.

Ключевые слова: микромассив, транскриптом, интегративный анализ, эмпирический метод Байеса, метаанализ, дифференциально экспрессированные гены, плацента.