Optimized Bioinformatic Strategy for the Analysis of Clinical Proteomic Data of the Endometrium in Chronic Endometritis

M.R. Gainullin, MD, PhD, Senior Researcher, Department of Molecular Cell Technology,
Institute of Fundamental Medicine[1]; Researcher[2, 3];
A.B. Yazykova, PhD, Assistant, Department of Biochemistry named after G.Ya. Gorodisskaya[1];
T.M. Motovilova, MD, PhD, Associate Professor, Department of Obstetrics and Gynecology[1];
H.M. Klemente Apumaita, MD, DSc, Professor, Department of Obstetrics and Gynecology No.1[4];
T.G. Khodosova, Obstetrician-Gynecologist[5];
Y.A. Gagaeva, Student[1];
E.S. Kolomina, Student[1];
M.M. Kovaleva, Student[1];
A.A. Militskaya, Student[1];
A.N. Shcherina, Student[1];
E.L. Boyko, MD, DSc, Senior Researcher, Department of Obstetrics and Gynecology[6];
V.G. Zgoda, DSc, Head of the Department of Proteomic Research and Mass Spectrometry[7];
G.O. Grechkanev, MD, DSc, Professor, Department of Obstetrics and Gynecology[1]

[1]Privolzhsky Research Medical University, 10/1 Minin and Pozharsky Square, Nizhny Novgorod,
603005, Russia;
[2]Norwegian PSC Research Center, Department of Transplantation Medicine, Division of Surgery,
Inflammatory Diseases and Transplantation, Oslo University Hospital Rikshospitalet, P.O. Box 4950,
Nydalen, Oslo, 0424, Norway;
[3]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, P.O. Box 1171, Blindern, Oslo,
0318, Norway;
[4]I.M. Sechenov First Moscow State Medical University (Sechenov University), 8/2 Trubetskaya St.,
Moscow, 119991, Russia;
[5]Regional Perinatal Center, 145 Kashtanovaya alleya, Kaliningrad, 236023, Russia;
[6]Ivanovo Research Institute of Motherhood and Childhood named after V.N. Gorodkov, 20 Pobeda St.,
Ivanovo, 153045, Russia;
[7]V.N. Orekhovich Research Institute of Biomedical Chemistry, Bldg 8, 10 Pogodinskaya St., Moscow,
119121, Russia

The aim of the study is to analyze the entire set of proteins (proteome) expressed in the endometrial tissue and to identify protein markers specific for carcinogenesis.

Materials and Methods. Tissue samples were obtained using endometrial pipelle biopsy in women with chronic endometritis. After homogenization the samples were subjected to protein electrophoresis in polyacrylamide gel in the presence of sodium dodecyl sulfate according to the Lamley method. The proteins separated according to their molecular weights were digested by modified trypsin using the standard method. Obtained tryptic peptides were analyzed and identified by high-performance liquid chromatography coupled with tandem mass spectrometry.

The Human Protein Atlas and Tissue-Specific Gene Expression and Regulation databases were used to analyze the tissue-specific protein expression.

Functional protein annotation and gene set enrichment analysis were performed using the Database for Annotation, Visualization and Integrated Discovery bioinformatics resource.

Results. In the obtained endometrial tissue samples, 103 proteins were identified by tandem mass spectrometry. Analysis of tissue specificity showed that 83 proteins were expressed in the tissues of the female reproductive system. Functional annotation followed by clustering revealed that 51 proteins (49.5% of the identified ones) were encoded by the genes differentially expressed in cell cultures of the female reproductive organs. Along with that, 4 groups of proteins were expressed both in tumors (serous ovarian adenocarcinoma, immortalized ovarian cystadenoma, ovarian carcinoma) and in the immortalized normal ovarian surface epithelium.

Conclusion. Endometrial tissue proteins were identified using a clinical proteomic analysis. The bioinformatic approach allowed us to annotate the functional clusters of the identified proteins based on their potential involvement in carcinogenesis. The obtained data can serve

Corresponding author: Anna B. Yazykova, e-mail: poet509@yandex.ru

as the starting point for further in-depth studies of the endometrium using the proteomic approach, as well as other OMICS technologies. Subsequent application of bioinformatic tools will allow revealing of molecular mechanisms of relationship between inflammation and endometrium tissue malignant transformation.

Key words: chronic endometritis; tandem mass spectrometry; functional clustering; tissue-specific expression.

Introduction

Impaired female fertility, recurrent miscarriages, and unsuccessful attempts of *in vitro* fertilization/embryo transfer often result in chronic inflammatory diseases of the uterus. (Among them is chronic endometritis — a frequent disorder in women of reproductive age). The disease develops "silently" without specific symptoms, and so the classical diagnostic methods are not applicable [1–4]. Hysteroscopy and morphological examination of the uterus is the gold standard of the current diagnostic procedures; however, the ability of hysteroscopy to diagnose initial signs of endometritis is limited by the resolution of the optical instrument. Consequently, hysteroscopy based on macroscopic evaluation can detect chronic endometritis in only 35–40% of cases. Therefore, clinicians need alternative diagnostic methods, safe and informative, for examining such a vulnerable tissue as endometrium [5, 6].

It has been proved by now that chronic endometritis is involved in the development of endometrial cancer by methylation of the tumor suppressor genes and by modifying the local immune and systemic inflammatory response, although the importance of this involvement is not precisely determined [7]. The molecular and pathogenetic mechanisms of this relationship are barely elucidated in the literature; rather the malignancy is attributed to the formation of focal or diffuse epithelium proliferation and the occurrence of hyperplasia, which is characterized by a high recurrence rate and potential for malignancy [8]. The above factors necessitate more research into the endometrial pre-cancer condition, specifically, into molecular biomarkers of endometritis, as well as into markers of neo-angiogenesis and proliferation. On that base, a screening system to identify early signs of malignancy in the endometrium can be developed.

It is known that inflammation is often associated with subsequent cancer by contributing to the development and progression of malignant tumors. In recent years, there is growing evidence of the significance of the local immune response and systemic inflammation in the progression of tumors and the survival of cancer patients [9, 10]. Therefore, the elucidation of inflammation and carcinogenesis molecular markers is highly important for early diagnosis of cancer and for new modalities of targeted therapy involving the patient's immune system.

The proteomic approach based on protein identification by tandem mass spectrometry coupled with high-performance liquid chromatography (LC–MS/ MS) meets most of the criteria for "screening" analysis of clinical material. The LC–MS/MS is used for large-scale qualitative identification of proteins in biological samples. It is characterized by the multiplexity within hundreds of individual proteins in a single sample, the sensitivity in range of pico- and femtomoles of single protein, and a dynamic range of 4–6th orders of magnitude, which is close to the protein concentrations in human tissues. Currently, the proteomic analysis of the endometrium is used in clinical medicine not only in endometriosis [11], but also in endometrial cancer [12], infertility [13] and in women with pregnancy-induced changes [14].

The OMICS technologies differ from other methods of molecular biology by the quantitative and qualitative characteristics of the acquired data. As a rule, these are global (as in the case of genomic and transcriptomic techniques) or local (typical for proteomics) sets of biomacromolecules (DNA, mRNA or proteins, respectively). To describe such a "molecular phenotype", genomic and post-genomic technologies are coupled with computer-assisted data processing.

The basic concept of the OMICS technologies implies that: a) essential biological information is contained in a set of identified genes or their products (mRNA, proteins); b) each element of a living system functions in conjunction with a specific set of other elements. Accordingly, the purpose of most OMICS data interpretation methods is to annotate (i.e., assign a functional characteristic to) individual biomacromolecules and then reconstruct the significant interactions between them. The computational tools of genomic and post-genomic technologies utilize the variety of available biological information obtained by experimental methods and deposited in systematized public databases. In most cases, the interpretation of OMICS data is predictive and based on specially developed methods of statistical analysis.

The most common computational approach in this area is the so-called gene set enrichment analysis (GSEA) [15]. In the GSEA, the principle of cluster analysis is applied to sets of biomacromolecules, which makes it possible to adapt this method to specific OMICS technologies.

It has to be emphasized though that today, the stage of generalization and data analysis is a "bottleneck" that significantly limits the scientific and practical value of transcriptomic and proteomic tests. Therefore, the optimization of systemic analysis and interpretation of OMISC data is an urgent task.

The present study is a pilot project aimed at

searching for protein markers of malignancy in women with clinically confirmed chronic endometritis; we used mass spectrometry for protein identification followed by bioinformatic analysis. For this purpose, we used a number of methods to assess the tissue specificity of the identified proteins; we also applied an optimized cluster analysis to the gene set enrichment method.

Materials and Methods

The study involved women of reproductive age with the histologically confirmed diagnosis of chronic endometritis. Informed consent was obtained from all participants. The work was carried out in accordance with the Helsinki Declaration (2013) and approved by the Ethics Committee of the Privolzhsky Research Medical University.

All patients underwent endometrial pipelle biopsy; the tissue samples were then placed in a buffer solution. After homogenization, the samples were subjected to protein electrophoresis in polyacrylamide gel in the presence of sodium dodecyl sulfate according to Lamley method. The proteins separated by their molecular weights were treated with modified trypsin (Promega, USA). The resulted tryptic peptides were analyzed and identified by high-performance liquid chromatography coupled with tandem mass spectrometry (Orbitrap Velos Pro apparatus; Thermo Scientific, USA). For protein identification, the UniProt database (release 2014) was used.

Tissue-specific expression was analyzed using the Human Protein Atlas [16] and Tissue-Specific Gene Expression and Regulation (TiGER) databases [17].

Functional annotation of the proteins and their analysis by GSEA were performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID v. 6.8) bioinformatics resource [18, 19].

Results

The endometrial tissue homogenates were first subjected to electrophoretic protein separation by molecular weights followed by trypsinolysis. The analysis of the tryptic fragments by using LC–MS/MS and the subsequent mass spectrometry allowed us to identify 103 different proteins.

In accordance with the objectives of this study, the presence of these proteins in various tissues was manually evaluated. According to the information from the Human Protein Atlas and TiGER, 83 of these proteins are present in tissues of the female reproductive system. Blood plasma proteins were also identified. This was an expected result because the endometrial biopsy material inevitably contained blood components.

It needs, however. to be emphasized that the manual analysis of tissue specificity with the help of databases cannot be considered the optimal method for interpreting clinical proteomic results. The impossibility of statistical evaluation of the results is a major disadvantage of the described approach. Considering the fact that the overwhelming majority of proteins are expressed at various concentrations in diverse tissues, practical significance of this approach is regrettably small.

At the next stage of this study, we used DAVID (v. 6.8) bioinformatic resource for functionally annotating the proteins identified in the endometrial tissues. The list of proteins was analyzed by the GSEA method using the standard clustering parameters. Unfortunately, the obtained results were not informative, because the proteins were combined into groups with low functional specificity; that was the blood micro-particles (GO: 0072562), the intercellular space proteins (GO: 0005615), and the signal peptides. We consider this result to be due to the high universality of the DAVID resource. It is known that the GSEA method is most commonly used for differentially expressed genes (i.e., transcriptome analysis), which is reflected in the set of parameters proposed "by default". However, the transcriptomic and proteomic data are fundamentally different from each other, both quantitatively and qualitatively.

Therefore, when choosing the optimal cluster parameters for protein detection with mass spectrometry from the CGAP_SAGE_QUARTILE (The cancer genome anatomy project / Serial analysis of gene expression) (see the Table), the main grouping factor was the association between the protein expression and the pathogenesis of female reproductive system cancers.

It was found that 51 proteins (49.5% of the original list) were encoded by genes differentially expressed in cell cultures of the female reproductive organs. In addition, four groups of proteins typical for tumor cells (immortal ovarian cystadenoma, ovarian carcinoma) as well as for the immortalized normal ovarian surface epithelium were detected.

These results demonstrate the significance of these proteins in the development of physiological and pathological processes in the cells of female reproductive system. Being a pilot project, this study provides a serious basis for further research in this area, for instance, an assessment of the clinical relevance of the identified proteins.

Conclusion

This pilot proteomic study resulted in identification of proteins expressed in the human endometrium. Using current bioinformatic methods, we were able to annotate these proteins into functional clusters according to their involvement in carcinogenesis.

The obtained data can serve as the starting point for further in-depth studies of the endometrium using the proteomic approach, as well as other OMICS technologies. The application of bioinformatics is expected to further elucidate the molecular basis of the inflammatory process in the endometrium and

Functional clusters of proteins identified in the endometrial tissue

Cluster characteristics	Identifier SAGE	The number of proteins (genes) in the cluster	P-value / Benjamini	Cluster enrichment factor	The protein-encoding gene*
Serous ovarian adenocarcinoma	SAGE_Ovary_ adenocarcinoma_ B_OVT-7	26	$1.4 \cdot 10^{-6}$/ $3.5 \cdot 10^{-5}$	5.82	*ATP5B; CNDP2; RAB13; ALDOA; ANXA11; ANXA2; CLIC1; C3; C4A; GSTP1; HMGB1P1; IGHG1; IGHG2; IGHG3; IGHG4; IGHM; IGKC; IGLC2; PRDX6; PYGB; P4HB; SERPINF1; TXNDC5; TALDO1; TGM2; UBA1*
Ovarian cystadenoma, immortalized	SAGE_Ovary_ cystadenoma_ CL_ML10-10	30	$2.9 \cdot 10^{-6}$/ $6.3 \cdot 10^{-5}$	4.55	*ATP5B; GNAI2; IQGAP1; RAB13; SET; ALDOA; ANXA11; ANXA2; ANXA5; CLIC1; CFL1; FLNA; GSTP1; GAPDH; HNRNPD; HMGB1P1; HYOU1; LDHA; LDHB; MYH9; PYGB; PFN1; P4HB; PDIA4; PDIA6; SERPING1; TXNDC5; TPI1; TPP1; UBA1*
Ovarian carcinoma	SAGE_Ovary_ carcinoma_CL_ A2780	28	$5.0 \cdot 10^{-4}$/ $2.5 \cdot 10^{-3}$		*ATP5B; CNDP2; RAB13; SET; ALDOA; ANXA2; CLIC1; CFL1; ENO1; GSTP1; HBB; HNRNPD; HMGB1P1; HYOU1; LDHA; LDHB; NEFH; PPIB; PRDX6; PFN1; P4HB; PDIA4; PPA1; SERPINF1; SERPING1; TXNDC5; TPI1; UBA1*
Normal ovarian surface epithelium, immortalized	SAGE_Ovary_ normal_CL_IO_ SE29EC-11	21	$1.3 \cdot 10^{-2}$/ $3.4 \cdot 10^{-2}$		*ATP5B; IQGAP1; RAB13; SET; ALDOA; ANXA2; CP; FLNA; GPI; GSTP1; GAPDH; HMGB1P1; LDHA; MYH9; PGK1; P4HB; SERPING1; TKT; TPI1; TPP1; UBA1*

* Gene abbreviations in accordance with the common protein names (for 51 proteins presented in the Table):
ALDOA — aldolase, fructose-bisphosphate A; *ANXA11* — annexin A11; *ANXA2* — annexin A2; *ANXA5* — annexin A5; *ATP5B* — ATP synthase (H^+-transporting, mitochondrial F1 complex, beta polypeptide); *CP* — ceruloplasmin; *CLIC1* — intracellular chloride channel; *CNDP2* — CNDP-dipeptidase 2, M20 metallopeptidases family; *CFL1* — cofilin 1; *C3* — component of the C3 complement system; *C4A* — component of the C4A complement system (Rodgers blood group); *ENO1* — enolase 1; *FLNA* — filamin A; *GNAI2* — alpha-i2 subunit of the G protein; *GPI* — glucose 6-phosphate isomerase; *GSTP1* — glutathione-S-transferase P1; *GAPDH* — glyceraldehyde-3-phosphate dehydrogenase; *HBB* — hemoglobin, beta subunit; *HNRNPD* — heterogeneous nuclear ribonucleoprotein D; *HMGB1P1* — amphoterin; *HYOU1* — hypoxia 1 activated protein (hypoxia up-regulated 1); *IGHG2* — constant region of the heavy chain of gamma globulin 2 (G2m marker); *IGHG3* — constant region of the heavy chain of gamma globulin 3 (G3m marker); *IGHG4* — constant region of the heavy chain of gamma globulin 4 (G4m marker); *IGHM* — constant region of the heavy chain of gamma globulin μ; *IGHG1* — constant region of the heavy chain of gamma globulin 1 (G1m marker); *IGKC* — immunoglobulin light kappa constant region; *IGLC2* — constant region of immunoglobulin 2 light kappa chain; *IQGAP1* — protein of the guanosine triphosphatase activators family containing IQ motifs 1; *LDHA* — lactate dehydrogenase A; *LDHB* — lactate dehydrogenase B; *MYH9* — myosin heavy chain; *NEFH* — heavy polypeptide neurofilament; *PPIB* — peptidyl prolyl isomerase B; *PRDX6* — peroxiredoxin 6; *PGK1* — phosphoglycerate kinase 1; *PYGB* — glycogen phosphorylase; brain fraction; *PFN1* — profilin 1; *P4HB* — prolyl-4-hydroxylase, beta subunit; *PDIA4* — protein disulfide isomerase, family A, number 4; *PPA1* — pyrophosphatase (inorganic) 1; *RAB13* — RAB13 protein of the oncogenic RAS protein family; *SERPINF1* — serpin protein 1 of the F protein family; *SET* — nuclear proto-oncogene SET; *TXNDC5* — thioredoxin-containing domain 5; *TALDO1* — transaldolase 1; *TGM2* — transglutaminase 2; *TKT* — transketolase; *TPI1* — triose phosphate isomerase 1; *TPP1* — tripeptidyl peptidase 1; *UBA1* — activating enzyme 1 of the ubiquitin-like modifier.

the occurrence of hyper- and neoplasia of the uterine mucosa.

References

1. Cicinelli E., Matteo M., Tinelli R., Lepera A., Alfonso R., Indraccolo U., Marrocchella S., Greco P., Resta L. Prevalence of chronic endometritis in repeated unexplained implantation failure and the IVF success rate after antibiotic therapy. *Hum Reprod* 2015; 30(2): 323–330, https://doi.org/10.1093/humrep/deu292.

2. Kitaya K., Matsubayashi H., Yamaguchi K., Nishiyama R., Takaya Y., Ishikawa T., Yasuo T., Yamada H. Chronic endometritis: potential cause of infertility and obstetric and neonatal complications. *Am J Reprod Immunol* 2016; 75(1): 13–22, https://doi.org/10.1111/aji.12438.

3. Kasius J.C., Fatemi H.M., Bourgain C., Sie-Go D.M., Eijkemans R.J., Fauser B.C., Devroey P., Broekmans F.J. The impact of chronic endometritis on reproductive outcome. *Fertil Steril* 2011; 96(6): 1451–1456, https://doi.org/10.1016/j.fertnstert.2011.09.039.

4. Tortorella C., Piazzolla G., Matteo M., Pinto V., Tinelli R., Sabbà C., Fanelli M., Cicinelli E. Interleukin-6, interleukin-1β, and tumor necrosis factor in menstrual effluents as biomarkers of chronic endometritis. *Fertil Steril* 2014; 101(1): 242–247, https://doi.org/10.1016/j.fertnstert.2013.09.041.

5. Viana G.A., Cela V., Ruggiero M., Pluchino N., Genazzani A.R., Tantini C. Endometritis in infertile couples: the role of hysteroscopy and bacterial endotoxin. *JBRA Assist Reprod* 2015; 19(1): 21–23, https://doi.org/10.5935/1518-0557.20150006.

6. Arlas T.R., Wolf C.A., Petrucci B.P., Estanislau J.F., Gregory R.M., Jobim M.I., Mattos R.C. Proteomics of endometrial fluid after dexamethasone treatment in mares susceptible to endometritis. *Theriogenology* 2015; 84: 617–623, https://doi.org/10.1016/j.theriogenology.2015.04.019.

7. Diakos C.I., Charles K.A., McMillan D.C., Clarke S.J. Cancer-related inflammation and treatment effectiveness. *Lancet Oncol* 2014; 15(11): 493–503, https://doi.org/10.1016/s1470-2045(14)70263-3.

8. Lax S.F. Pathology of endometrial carcinoma. *Adv Exp Med Biol* 2017; 943: 75–96, https://doi.org/10.1007/978-3-319-43139-0_3.

9. Nakamura K., Smyth M.J. Targeting cancer-related inflammation in the era of immunotherapy. *Immunol Cell Biol* 2017; 95(4): 325–332, https://doi.org/10.1038/icb.2016.126.

10. Zhang X., Meng X., Chen Y., Leng S.X., Zhang H. The biology of aging and cancer: frailty, inflammation, and immunity. *Cancer J* 2017; 23(4): 201–205, https://doi.org/10.1097/00130404-201707000-00002.

11. Adamyan L.V., Starodubtseva N., Borisova A., Stepanian A.A., Chagovets V., Salimova D., Wang Z., Kononikhin A., Popov I., Bugrova A., Chingin K., Kozachenko A., Chen H., Frankevich V. Direct mass spectrometry differentiation of ectopic and eutopic endometrium in patients with endometriosis. *J Minim Invasive Gynecol* 2017; 25(3): 426–433, https://doi.org/10.1016/j.jmig.2017.08.658.

12. Martinez-Garcia E., Lesur A., Devis L., Cabrera S., Matias-Guiu X., Hirschfeld M., Asberger J., van Oostrum J., Casares de Cal M.L.Á., Gómez-Tato A., Reventos J., Domon B., Colas E., Gil-Moreno A. Targeted proteomics identifies proteomic signatures in liquid biopsies of the endometrium to diagnose endometrial cancer and assist in the prediction of the optimal surgical treatment. *Clin Cancer Res* 2017; 23(21): 6458–6467, https://doi.org/10.1158/1078-0432.ccr-17-0474.

13. Kosteria I., Anagnostopoulos A.K., Kanaka-Gantenbein C., Chrousos G.P., Tsangaris G.T. The use of proteomics in assisted reproduction. *In Vivo* 2017; 31(3): 267–283, https://doi.org/10.21873/invivo.11056.

14. Moza Jalali B., Likszo P., Skarzynski D.J. Proteomic and network analysis of pregnancy-induced changes in the porcine endometrium on day 12 of gestation. *Mol Reprod Dev* 2016; 83: 827–841, https://doi.org/10.1002/mrd.22733.

15. Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S., Mesirov J.P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005; 102(43): 15545–15550, https://doi.org/10.1073/pnas.0506580102.

16. *The Human Protein Atlas.* URL: https://www.proteinatlas.org/.

17. *Tissue-Specific Gene Expression and Regulation (TiGER).* URL: http://bioinfo.wilmer.jhu.edu/tiger/.

18. *DAVID 6.8.* URL: https://david.ncifcrf.gov/.

19. Huang da W., Sherman B.T., Lempicki R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; 4(1): 44–57, https://doi.org/10.1038/nprot.2008.211.