Functional and association analysis of polymorphisms in the *WFS1* gene

Doctoral Theses

Zsuzsanna Molnár

Semmelweis University
Molecular Medicine Doctoral School

Supervisor: Dr. Zsolt Rónai, MD, Ph.D

Official reviewers: Beáta Törőcsik MD, Ph.D
Tibor Füle, Ph.D

Head of the Final Examination Committee: Dr. Ilona Kovalszky Ds.C
Members of the Final Examination Committee: Ágnes Semsei, Ph.D
Zoltán Gáspári, Ph.D

Budapest
2019

INTRODUCTION

Discovery of the 3 billion-basepair long DNA-sequence of the human genome was a principal milestone in human genetics and genomics, this information has been available in databases on the Internet for more than 15 years. This result, however, did not mean the end of genetic researches, but instead it has provided numerous novel possibilities and trends in the analysis of genetic variations and molecular biological procedures.

One of the most thoroughly investigated fields is the analysis of the heritable components of complex phenotypes, which is of essential theoretical and practical importance. Identification of the genetic components responsible for the development of diseases contributes to prevention, prediction of prognosis, moreover it can help in understanding the molecular patomechanism. It can suggest molecular targets for novel medicaments, and thus can result in the development of targeted and individualized therapeutic protocols. Although several large-scale collaborative projects analyzing thousands of patients have aimed the identification of the genetic components of various complex phenotypes, none of these researches have managed to shed light on the complete list of any complex disorders so far.

The *WFS1* gene codes for the wolframin, which has a molecular weight of 100 kDa and consists of 890 amino acids. Wolframin is localized in the membrane of the endoplasmic reticulum, the protein possesses 9 transmembrane, hydrophobic, tetramer regions, its hydrophilic N terminal region faces to the cytosol, whereas the C terminal is in the lumen of the ER. The protein is abundantly expressed in practically all tissue types, highest expression levels can be observed in the heart, brain and beta-cells of the pancreas.

Loss of function mutations are responsible for the development of Wolfram-syndrome (WFS; OMIM 222300). The disease is also known as DIDMOAD-syndrome, it is a rare, neurodegenerative

illness including numerous disorders, such as **d**iabetes **i**nsipidus, **d**iabetes **m**ellitus **o**ptic **a**trophy and **d**eafness often combined by various mental deficiencies. A *WFS1* knock-out animal model was generated to model the disease, in these mice decreased glucose-induced insulin secretion can be observed, and lower number of beta-cells can be detected by immunohistochemistry. Isolated islet cells of these animals show apoptosis when treated with glucose or ER-stress inducers. These data suggest that wolframin might play an important role in maintaining normal number of beta-cells, as well as in the exocytosis of insulin. Besides these, wolframin was shown to bind calmodulin, and together with several other proteins it influences numerous Ca^{2+}-dependent signal transduction pathways and plays an important role in the regulation of Ca^{2+}-homeostasis.

All these data proposed the assumption that genetic polymorphisms of the gene might be the risk factors of psychiatric diseases and diabetes mellitus. Association between the rs4689388 SNP and type 2 diabetes mellitus was first shown by a genome-wide association study (GWAS) in 2009. Metaanalysis of several association studies proved that the rs1046320 and the rs10010131 SNPs can be considered as the genetic risk factors of type 2 diabetes. The former polymorphism is localized in the 3’ UTR, whereas the latter one in intron 4 of the gene, respectively, and linkage disequilibrium is high between the two loci. Techniques analyzing SNPs and DNA-sequences have been dramatically improved in the recent years, however the time- and labor-intensive analysis of the biological effect of the polymorphisms cannot keep step. This is the reason, why no data about the functional role of these SNPs have been available so far, this question is however of importance, because the two loci are localized in rather different gene regions. This is not a unique situation: biological function of numerous hits of the genome wide association studies have not yet been discovered so far.

OBJECTIVES

The major goal of the project was the functional and association analyses of some selected polymorphisms localized in the regulatory regions of the *WFS1* gene. Most important tasks of the research were as follows:

- selection of polymorphisms of interest by *in silico* approaches,
- elaboration of reliable genotyping technique for the genotype analysis of the rs148797429 6 bp insertion / deletion polymorphism,
- elaboration of a method that is capable of the simultaneous genotype determination of the SNPs localized in the regulatory region of the *WFS1* gene in a multiplex format,
- analysis of linkage disequilibreium (LD) between the investigated SNPs and determination of the haplotype blocks formed by these loci,
- performance of genetic association studies to investigate, whether allelic variants of the analyzed polymorphisms show significantly different frequencies among patients with type 1 or type 2 diabetes mellitus compared to a control group of healthy individuals,
- optimization and application of *in vitro* luciferase system for the analysis of biological function of the SNPs localized in the promoter and the 3' UTR of the gene (regulation of transcription or miRNA binding, respectively).

MATERIALS AND METHODS

Patient Groups, DNA-sampling and DNA-purification

DNA samples of 900 patients with diabetes mellitus as well as that of 892 healthy individuals were analyzed in the genetic association study. The project was approved by the Local Ethical Committee (ETT-TUKEB, Scientific and Research Ethics Committee of the Medical Research Council, 4514-0/2010-1018EKU). Participants were informed in detail about the project in written form and signed informed consent prior to providing DNA sample.

DNA sampling was carried out by non-invasive method: buccal cells were collected from the inner surface of the cheek or the gum using cotton swabs. This method provided sufficient amount of DNA for the genotype analysis. First steps of the DNA purification were the lysis of the cells, denaturing and digestion of the proteins by SDS and proteinase K, respectively. It was followed by salting out the proteins using NaCl, then DNA was precipitated by isopropanol and ethanol. Finally, DNA precipitations were resuspended in TE buffer solution.

In silico Methods

Several on-line tools and databases were employed to prepare the genetic and molecular biological experiments. Data about gene sequence and polymorphisms of the *WFS1* gene were downloaded from genebanks of NCBI and Ensembl. SNPs putatively altering miRNA binding were identified by the Patrocles, the PolymiRTS and the miRWalk 2.0 databases. Sequence of the micro-RNAs of interest were obtained form miRBase. Primers for the PCR-based analyses were designed by the Primer Blast tool of NCBI. Recognition sites of restriction endonucleases were searched using NEBcutter v2.0.

Methods for Genotype Determination

Genotype determination of 7 SNPs of the *WFS1* gene was carried out by real-time PCR-based technique. The rs4273545 (T/G) SNP was analyzed by allele-specific amplification. This method employs two outer and two allele-specific primers (annealing to the polymorphic locus with the 3' end) in two separate reaction mixtures. PCR products were visualized by traditional agarose gel-electrophoresis. The rs4689388 SNP was genotyped by PCR–RFLP. The principle of this technique is the application of a retriction endonuclease, which can cut the PCR product in the presence of one allele, whereas the recognition site is eliminated by the other variant. PCR primers were designed to include a control, non-polymorphic recognition site of the enzyme serving as control of the digestion. Genotype analysis of the rs4689388, rs4273545, rs1064320, rs1046322 and rs9457 SNPs was carried out by primer extension. First the adjacent regions of the SNPs were amplified, genotypes were then determined applying the four labeled terminator acyclo-nucleotides. Analyses were performed in multiplex format using extension primers of different lengths, products were visualized by capillary gel-electrophoresis.

The first step of the investigation of the rs148797429 insertion / deletion was the amplification of the adjacent gene region by PCR. Products were subjected to downstream analyses employing traditional or – for higher accuracy – capillary gel-electrophoresis as well as melting curve analysis.

In vitro Functional Analysis of Selected Polymorphisms

The pGL3B and the pMIR-REPORT luciferase vectors were used for the analysis of the promoter and the 3' UTR of the *WFS1* gene, respectively. Constructs with different lengths were generated for the investigation of the promoter region. The entire 3' UTR of the *WFS1* gene was subcloned into pMIR-REPORT vector for the analysis of the regulatory effect of micro-RNAs. Restriction digestion followed by gel-electrophoresis was used to

verify the proper incorporation of the insert into the vector, furthermore each generated construct was confirmed by Sanger sequencing as well. Site-directed mutagenesis was used to prepare the constructs with the different allelic variants.

HEK293T cell line was used for the functional analyses of the polymorphisms. The different pGL3B constructs were cotransfected with a construct coding for beta-galactosidase for the analysis of the promoter. The pMIR-REPORT constructs were applied together with the miR185 precursor and a construct coding for beta-galactosidase for the investigation of the SNPs in the 3' UTR. Each measurement was carried out in triplicates.

Cells were harvested 24 hours after transfection and extracted by three consecutive freeze–thaw cycles. Enzyme activity assays were carried out with a Varioskan Flash instrument, capable of measuring both luminescent (luciferase) signals as well as light absorption (photometry for beta galactosidase).

RNA-purification, Micro-RNA Assays

The TRI reagent was used to harvest cells for micro-RNA assays. This solution prevented RNA from degradation, RNA was isolated using chloroform, isopropanol and ethanol for extraction and purification, respectively. Expression levels of micro-RNAs were assessed by real time PCR. As micro-RNAs are rather short, a special step was required during cDNA synthesis, which attached an adapter sequence to the products serving for the annealing site of a universal primer during real-time PCR.

Statistical Analysis

The Hardy–Weinberg equilibrium was tested to confirm reliability of sampling and genotyping protocols. The case–control studies were evaluated by χ^2-test, the Bonferroni method was employed for correction for multiple testing. Luciferas enzyme activity data were normalized by beta-galactosidase in each case, results were tested by one-way ANOVA.

RESULTS

Selection of candidate polymorphisms

Our aim was the recruitment of polymorphisms possessing molecular biological role, thus putatively influencing the amount of the generated wolframin protein. Consequently, SNPs were selected that were suggested to alter the binding efficiency of transcription factors in the promoter region, or were localized in the binding site of micro-RNAs (miR-SNPs) in the 3' UTR.

Genotyping techniques

Independent methods were elaborated for the reliable genotype analysis of the rs148797429 6-bp-long insertion / deletion polymorphism. Each technique was initiated by the PCR-based amplification of the adjacent gene region. The obtained products were analyzed by (1) traditional horizontal gel electrophoresis, (2) melting curve analysis and (3) denaturing multicapillary gel electrophoresis.

A novel, primer extension-based technique employing multicapillary gel electrophoresis was optimized for the simultaneous genotype determination of the 5 SNPs in the regulatory regions of the gene. First, the adjacent regions were amplified by PCR. It was followed by the single base extension reaction. Unlabeled primers were applied that annealed to the very adjacent nucleotide of the SNP of interest. This primer was elongated by one single chain terminator, acyclo nucleotide. These nucleotides were labeled with different fluorescent dyes, thus the genotype could be determined based on the color of the obtained product. Multiplex analysis of the 5 SNPs could be carried out using extension primers with different lengths: products corresponding to the appropriate SNPs could been unambiguously separated from each other.

Linkage Disequilibrium and Halotype Analysis

Assessment of linkage disequilibrium between SNPs is of importance because of several reasons. Association studies often reveal polymorphisms that do not possess relevant biological function, they are just in linkage disequilibrium with the locus with molecular biological, functional importance. When identification of "functional" SNPs (possessing molecular effect) is also aimed, determination of linkage disequilibrium is of high significance.

Linkage disequilibrium between the investigated polymorphisms in the *WFS1* gene was determined by calculating the *D*' (Lewontin's standardized linkage disequilibrium coefficient) as well as the R^2 (correlation coefficient) values. The 3 polymorphisms localized in the promoter region formed a haplotype block. This high level of linkage disequilibrium was confirmed by the observation, that the total frequency of two out of the theoretically possible eight haplotypes was higher than 90% in the investigated population (A–insertion–T: 62.9%, G–insertion–G: 28.5%). Besides these variants, only two further haplotypes could be detected with notably lower frequencies (G–deletion–G: 7%, A–deletion–G: 1.6%)

Association Analyses

Association between diabetes mellitus and polymorphisms of the *WFS1* gene was analyzed in case–control settings: we compared the allele frequencies of the selected polymorphisms measured in the patient and healthy control group, respectively. An outstandingly high value of statistical significance as well as odds ratio could be observed for the rs1046322 SNP when investigating patients with type 1 diabetes mellitus, this suggests the significant role of this locus in the background of the disease. Although it is

not exceptional but notable that – except the rs1046322 locus – the major allele proved to be the risk factor in case of all the other SNPs showing an association with diabetes. Please note, that the association proved to be significant in case of these polymorphisms even using the Bonferroni correction for multiple testing, however the odds ratio values were relatively low. This can be explained by the complex etiology of diabetes mellitus, and demonstrates that each genetic component has only a minor role in the development of the disease.

A haplotype analysis was also carried out for the polymorphisms in the 3' UTR (rs1046320, rs1046322 and rs9457). Results confirmed the significance of the role of rs1046322 SNP in the background of type 1 diabetes mellitus, whereas case–control study suggested, that the C allele of the rs9457 locus might be the genetic component of type 2 form of the disease.

Functional Studies of the Polymorphisms Localized if the Regulatory Regions

The rs148797429 and the rs4273545 polymorphisms are localized in the promoter of the *WFS1* gene, their putative effect on gene expression was analyzed using cell culture system.

Two DNA-constructs with different lengths were generated to investigate the biological effect of the polymorphisms. The shorter construct included the "minimal promoter", it contained the rs4273545 SNP only. Analyses with this system did not show any difference in the effect on the regulation of gene expression between the promoter regions containing "G" or "T" allele at the SNP site. The longer construct contained both the rs148797429 insertion / deletion, as well as the rs4273545 SNP. Although the rs148797429 variant did not influence the activity of transcription, the adjacent sequence seems to contribute to the regulation

of the activity of gene expression. It was namely observed that the “T” and the “G” variants of the rs4273545 SNP lead to different transcriptional activity in this system. In the presence of the “T” allele the relative luciferase activity was approximately 2.5 times higher than in case of the “G” form. This difference could be observed regardless of the rs148797429 locus, i.e. in case of both the insertion and the deletion variant at this site.

Putative effect of the rs9457 SNP on micro-RNA binding was analyzed using *in vitro* cell culture system as well. The SNP is localized in the “seed” region of miR-185 according to the PolymiRTS database. The entire 3’ UTR of the *WFS1* gene was subcloned into the pMIR-Report vector to assess the regulatory role of the polymorphism on protein expression. Site directed mutagenesis was used to generate the construct with the other allelic variant as well as the “seed” mutation. Furthermore, a control construct was also prepared, which contained an insert of similar length, however lacking the recognition site of miR-185. Lowest relative luciferase activity was detected in case of the C variant of the rs9457 SNP, this value was only 35% of that of the control construct. The presence of the G allele resulted in as high as 1.7 times increase in enzyme activity, and this value was practically identical with the activity of the construct with the “seed mutation”. Based on the sequence alignment, the C allele results in 6 complementary bases in the “seed” region of miR-185, thus the G variant leads to 5 base pairs only. It is also notable that the SNP is in the middle of the binding site, which probably also contributes to its prominent effect.

CONCLUSIONS

Rapid development of technology and computer engineering significantly contributes to the advancement of natural sciences. This effect could already be observed during the Human Genome Project. A couple of years after the initiation of the work results were disappointing: approximately 100,000 basepairs could be sequenced during a year, and based on this number the discovery of the entire genome seemed to be hopeless. Because of this, relatively slow progress theoretical innovations were also introduced, although the rapid development and success of the project would have been impossible without the application of the fluorescently labeled dideoxy nucleotides and the multicapillary gel electrophoresis instruments, which opened new perspectives in Sanger-sequencing. The advancements have been recently undiminished: the novel and sophisticated technological applications offer a wide range of innovative tools in practically all fields of theoretical and clinical sciences from research to diagnosis and therapy.

Applying these innovative methods dysfunctions in the background of several diseases can be determined at molecular level. This provides significant advancements in diagnostics and therapy. Whereas healing often used to be palliation only, today molecular dysfunctions in the background of diseases can be improved in case of more and more disorders. As an example, dysfunction of wolframin was suggested to play a role in the development of diabetes mellitus. This protein is localized in the membrane of endoplasmic reticulum and contributes to Ca^{2+}-homeostasis. Data suggest that alteration of gene expression regulated by micro-RNAs might be one of the molecular components of the disease. It is also notable that these kind of molecular procedures are only small pieces of the whole complex system. Although results of similar researches will help to complete the puzzle: discovery of the molecular patomechanism might contribute to diagnostics, prevention as well as targeted and individualized therapy.

BIBLIOGRAPHY OF THE CANDIDATE'S PUBLICATIONS

Publications related to the theses

1. Elek Z, Denes R, Prokop S, Somogyi A, Yowanto H, Luo J, Souquet M, Guttman A, Ronai Z. (2016) Multicapillary gel electrophoresis based analysis of genetic variants in the WFS1 gene. Electrophoresis, 37:2313-2321. (IF: 2.744)
2. Elek Z, Nemeth N, Nagy G, Nemeth H, Somogyi A, Hosszufalusi N, Sasvari-Szekely M, Ronai Z. (2015) Micro-RNA Binding Site Polymorphisms in the WFS1 Gene Are Risk Factors of Diabetes Mellitus. PloS one, 10:e0139519. (IF: 3.057)

Publications unrelated of the theses

1. Z, Elek Z, Nanasi T, Szekely A, Nemoda Z, Sasvari-Szekely M, Ronai Z. (2015) Polymorphism in the serotonin receptor 2a (HTR2A) gene as possible predisposal factor for aggressive traits. PloS one, 10:e0117792. (IF: 3.057)

2. Kis A, Bence M, Lakatos G, Pergel E, Turcsan B, Pluijmakers J, Vas J, Elek Z, Bruder I, Foldi L, Sasvari-Szekely M, Miklosi A, Ronai Z, Kubinyi E. (2014) Oxytocin receptor gene polymorphisms are associated with human directed social behavior in dogs (Canis familiaris). PloS one, 9:e83993. (IF: 3.234)

3. Kovacs-Nagy R, Elek Z, Szekely A, Nanasi T, Sasvari-Szekely M, Ronai Z. (2013) Association of aggression with a novel microRNA binding site polymorphism in the wolframin gene. American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics, 162B:404-412. (IF: 3.271)

4. Ronai Z, Kovacs-Nagy R, Szantai E, Elek Z, Sasvari-Szekely M, Faludi G, Benkovits J, Rethelyi JM, Szekely A. (2014) Glycogen synthase kinase 3 beta gene structural variants as possible risk factors of bipolar depression. American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics, 165B:217-222. (IF: 3.416)

5. Kotyuk E, Biro V, Bircher J, Elek Z, Sasvari M, Szekely A. (2017) ABCA1 polymorphism, a genetic risk factor of harm avoidance. Journal Of Individual Differences, 38:189-195. (IF: 1.283)

6. Spiro Z, Arslan MA, Somogyvari M, Nguyen MT, Smolders A, Dancso B, Nemeth N, Elek Z, Braeckman BP, Csermely P, Soti C. (2012) RNA interference links oxidative stress to the inhibition of heat stress adaptation. Antioxidants & redox signaling, 17:890-901. (IF: 7.189)

7. Szantai E, Elek Z, Guttman A, Sasvari-Szekely M. (2009) Candidate gene copy number analysis by PCR and multicapillary electrophoresis. Electrophoresis, 30:1098-1101. (IF: 3.077)

8. Wan M, Hejjas K, Ronai Z, Elek Z, Sasvari-Szekely M, Champagne FA, Miklosi A, Kubinyi E. (2013) DRD4 and TH gene polymorphisms are associated with activity, impulsivity and inattention in Siberian Husky dogs. Animal genetics, 44:717-727. (IF: 2.210)

ACKNOWLEDGEMENT

Hereby I would like to thank

- József Mandl Ds.C. and Gábor Bánhegyi Ds.C. for providing me the opportunity to carry out my research in the Department of Medical Chemistry, Molecular Biology and Pathobiochemistry.
- Mária Sasvári Ds.C., the head of the laboratory for supporting my research with her ideas as well as for always helping in both practical and theoretical work.
- my supervisor, Zsolt Rónai for his patience, professional knowledge and guidance, enthusiasm, help and humor,
- Anikó Somogyi Ds.C., the head of the clinical part of the studies in the 2nd Department of Internal Medicine,
- Eszter Szántai for her patience and help in theoretical and practical work,
- as well as all members of our lab for cooperation, kindness and cheerful atmosphere.