Investigation of allelic polymorphisms in the negative selection gene *AIRE* on the risk of rheumatoid arthritis using clinical samples

Ph.D. thesis

Bálint Bérczi

Clinical Medical Sciences Doctoral School
Molecular epidemiology of tumours
Doctoral School Leader: Lajos Bogár, M.D., PhD, DSc
Program Leader: István Kiss, M.D., PhD, DSc
Tutor: Dr. Zoltán Gyöngyi, PhD
University of Pécs, Medical School
Department of Public Health Medicine

Pécs, 2024.

List of abbreviations

aCCP	anti-cyclic citrullinated protein antibody
AIRE	autoimmune regulator
AIRE	the AIRE coding DNA sequence
APECED	autoimmune polyendocrinopathy-candidiasis ectodermal dystrophy
CARD	caspase-activation and recruitment domain
CRP	C-reactive protein
DAS28	disease activity score 28
ESR	erythrocyte sedimentation rate
HSR	homogeneously staining region
mTEC	medullar thymoepitelial cells
NLS	nuclear localisation signal
RA	rheumatoid arthritis
Rf	rheumatoid factor
SAND	Sp100 AIRE NucP41/75 and DEAF domain
SNP	single nucleotide polymorphism
TRA	tissue-restricted antigen

Investigation of allelic polymorphisms in the negative selection gene AIRE on the risk of rheumatoid arthritis using clinical samples

1. INTRODUCTION

Rheumatoid arthritis (RA) is a chronic autoimmune disease characterized by synovial inflammation, leading to symmetrical polyarticular arthritis and functional impairment. In developed countries, the prevalence of RA in the adult population ranges from 0.5% to 1%, with an annual incidence of 5 to 50 per 100,000 individuals. While the exact cause of the disease remains unclear, numerous studies suggest that autoimmune T cells evading the adaptive immune system play a crucial role in initiating the process.

Numerous environmental and genetic factors play a role in determining disease susceptibility. Recent inter-ethnic genome-wide association studies involving 69,825 cases and 330,798 matched controls have successfully identified all currently detectable genes involved in the development of the disease. These genes contain single nucleotide polymorphisms (SNPs) that may contribute to the disease's onset. Notably, one of these genes, the autoimmune regulator (*AIRE*) sequence, located on chromosome 21 at band 21q22.3, is about 12.5 kilobases long and encodes AIRE protein of 545 amino acids and 58 kDa in size, featuring 14 exonic sequences. It plays a crucial role in controlling autoimmunity.

The AIRE protein functions as a transcription factor responsible for regulating the negative selection of immature T cells (thymocytes). A critical aspect of its function is the promotion of promiscuous gene expression in medullary cells of the thymus, where it collaborates with DNA-binding proteins to enable the generation of tissue-restricted antigens (TRA) that correspond to the body's entire tissue protein pool. These antigens then move to the MHC receptors of the major histocompatibility complex on the surface of the thymic epithelial cells, effectively serving as a reflection for the developing T cells.

When an immature T-cell receptor recognises body antigens on the surface of medullary thymoepithelial cells (mTEC), it establishes its autoimmune nature and is clonally deleted by apoptosis. Due to its multidomain structure, AIRE belongs to a group of proteins capable of binding to chromatin and regulating gene transcription. AIRE comprises five functionally important structural domains from the N-terminus to the C-terminus: Homogeneously staining region (HSR), also known as caspase-activation and recruitment domain (CARD), nuclear localisation signal (NLS) domain, Sp100 AIRE NucP41/75 and DEAF domain (SAND), and two PHD domains (PHD1, PHD2).

The SAND domain is a crucial component of the AIRE protein, essential for its proper function in the medulla. First described in 1998, the SAND domain is the second structural sequence of AIRE, following the PHD domain. Positioned in the middle of the amino acid chain encoding AIRE, the structure of the SAND domain consists of a globular α/β element with a conserved KDWK motif in the α-helix. This element allows the AIRE protein to bind to DNA phosphate groups, which is the primary chromatin-dependent transcriptional regulatory function of the SAND domain. It plays a key role in binding AIRE to the DNA of medullary mTEC cells. Upon binding, the SAND domain initiates the expression of clustered TRA genes, contributing to the medullary negative selection for maturing T cells. This domain is encoded by exons 5, 6, and 7 in the coding region of AIRE.

Loss-of-function mutations in the DNA sequence that encodes the AIRE protein or the SAND region lead to a loss of selection. This allows autoreactive T cells to mature and escape to the periphery, causing a relatively rare autoimmune disease known as autoimmune polyendocrine syndrome type I (APECED, APS-1). Multiple reports consistently suggest that single nucleotide polymorphisms (SNPs) in the coding gene sequence might impact AIRE's transcription and, consequently, the protein's function, potentially increasing susceptibility to certain diseases.

The initial study to associate allelic polymorphisms with central tolerance was conducted by *Lovewell et al.* The study focused on examining a 591 bp segment upstream of the *AIRE* promoter region and the potential impact of allelic polymorphisms on AIRE transcription. Human lymphocyte DNA was analyzed using Denaturing High-Performance Liquid Chromatography (dHPLC). The study identified two allelic polymorphisms in these regions, namely *AIRE-655R* (rs117557896) and *AIRE-230Y* (rs751032), which were then ligated into a pCR2.1 TA vector and transfected into TOP10 cells. Each promoter variant was cloned into the pGL3-basic vector upstream of the luciferase reporter gene to generate an *AIRE* reporter construct. The transcription assay revealed that the *AIRE-655G AIRE-230T* allelic polymorphism genotype exhibited significantly reduced levels of promoter activity, almost at an undetectable level, while the *AIRE-655G AIRE-230C* genotype displayed the highest levels. The study's intriguing finding demonstrates that the genotype of a specific locus polymorphism can have a substantial impact on transcription, potentially resulting in barely detectable expression levels. This suggests that genotype-specific negative selection for central tolerance during the neonatal period may vary among individuals and that specific allelic polymorphisms could fundamentally determine its effectiveness. Furthermore, a genotype with low AIRE expression could potentially increase the incidence of autoimmune diseases from the postnatal period.

The first molecular epidemiological study on AIRE allele polymorphisms was performed in 2002 by *Tazi-Ahnini et al.* in alopecia areata autoimmune disease. During the course of the disease, T-cell populations filter into the follicles and induce baldness by autoantibody and cytokine production. In a severe form of this - alopecia universalis - the frequency of the *G961C* allele was found to be significantly higher in 202 patients compared to 175 Caucasian controls. Systemic sclerosis (SSc), also known as scleroderma, is a complex autoimmune disease characterised by fibrosis, vascular abnormalities and immune activation with T-cell infiltration and B-cell activation resulting in autoantibody and cytokine production. In 2007, Ferrara et al. found that the odds of *G11107A* allelic intron polymorphism were significantly higher in the patient population in a study of 41 SSc patients and 100 controls. The association was even stronger in the SSc population in the autoimmune thyroiditis subgroup. The association with autoimmune skin disease vitiligo was discovered in 2008 by *Tazi-Ahnini et al.* In a study of 86 patients and 363 controls, the *AIRE 7215C* allele was present at a significantly higher odds ratio (odds ratio, OR=3.12; 95%CI 1.87-5.46; $p=1.36\times10^{-5}$) in the vitiligo patient population. Inefficient central tolerance due to AIRE allelic polymorphism may affect the body's ability to recognise and destroy tumour cells and their antigens. *Conteduca et al.* wondered whether specific allelic polymorphisms might be associated with melanoma. In their 2010 study, as part of an inverse association, rs1055311 (C8385T), rs1800520 (C8723G) and rs1800522 (T16366C) were significantly more frequent in healthy individuals than in melanoma patients, regardless of sex, age and stage of melanoma, suggesting a possible protective effect of the allele. Autoimmune-based severe muscle weakness, myasthenia gravis, was associated in 2017 by *Zhang et al.* in a Chinese case-control study with significantly higher odds of having the G allele of rs3761389 or a genotype homozygous for this allele in patients compared to controls (OR=1.68; 95%CI 1.14-2.48; p=0.027). An inverse association was found with Addison's disease, an autoimmune destruction of the adrenal cortex, in 2018. The alleles rs2075875 C-, rs2075876 A-, rs62220374 G-, and rs9983695 C- were present at significantly higher odds in the control population; these alleles were found to be protective for the disease. In systemic lupus erythematosus (SLE), the immune system mistakenly attacks tissues of multiple organs throughout the body, particularly the skin, kidneys, brain and joints. In 2021, this rare autoimmune disease was associated with the rs2075876 A allele in the intron five region of rs2075876, which controls the negative selection of thymocytes and determines AIRE protein expression and was found to be protective in the case group (OR=0.16, 95%CI=0.09-0.28, p<0.001). Finally, the most recent molecular epidemiological result on *AIRE* allelic polymorphisms was published in 2023 and found to be associated with thrombocytopenic purpura, a haemophilia caused by autoimmune processes resulting in autoreactive platelet destruction. In this disorder, the A allele and homozygous genotype of rs2075876 located in the 5 region of the locus intron of the rs2075876 locus are also associated with risk (OR=4,299; 95%CI 1,650-11,202; p=0.003). As we can see, no association studies have been performed for these autoimmune diseases by mass. However, according to the literature, the association of one autoimmune disease with allelic polymorphisms in the AIRE coding region has been extensively studied, and this is rheumatoid arthritis.

This disease is a common autoimmune manifestation associated with chronic arthritis. The resulting symmetric polyarticular arthritis combined with extraarticular complications leads to functional impairment. Polymorphisms in the AIRE promoter region were first investigated in a genome-wide association study in 2011 by *Terao et al.* In this study, they found that among 5415 RA and 6489 Asian controls, 277 420 SNPs were examined and that the A and G alleles of rs2075876 and rs760426 in the intron region and homozygous genotypes of these alleles were significantly more likely to be present in the patient population (OR=1.18; 95%CI 1.11-1.24; $p=3.6\times10^{-9}$; OR=1.16; 95%CI 1.10-1.22; $p=4.4\times10^{-8}$). They further published results showing that despite the fact that rs2075876 occupies a splicing region, expression is significantly lower in the presence of the AA genotype ($p=6.8\times10^{-5}$) compared to the GG genotype, raising the possibility that exposure to the disease starts in neonatal life through inadequate negative selection. Similar results have been obtained in Spanish, Japanese, Chinese and Egyptian studies, the vast majority of which are case-control studies. In conclusion, the allelic polymorphism genotypes themselves may contribute to a large extent to the development of AIRE expression, as seen in *Lovewell et al.*, which can result in low expression, which may be a fundamental determinant of negative selection in the neonatal period in the case of immunological tolerance. In the presence of low AIRE expression, autoreactive $CD4^+$ effector cells are more likely to be released to the periphery and, in the presence of autoantigen, may initiate the disease initiation phase in the neonatal period.

Among these diseases, rheumatoid arthritis is the only one that has been investigated by several molecular epidemiological observations, so it is optimal for us to further analyse positive or negative associations with allelic polymorphisms and, based on the literature, the SAND region is considered to be a particularly important domain in the process of negative selection, so we focus our investigation on the DNA segments encoding this domain.

2. AIMS OF THE STUDY

2.1. Meta-analysis of *AIRE* allelic polymorphisms in autoimmune diseases

The first molecular epidemiological observation of AIRE allelic polymorphisms was performed in 2002, and since then, several case-control analyses have been performed in various autoimmune diseases. Before starting our research, we performed a systematic literature review, the aim of which was to clarify which autoimmune disease had the largest number of molecular epidemiological case-control studies and, within the disease, to find new findings and associations between candidate allelic polymorphisms and autoimmune diseases by a summary analysis of individual case-control studies.

2.2. *In silico* study to determine *AIRE* allelic polymorphisms

In order to have a database of allelic polymorphisms of the loci that we want to deal with in the future, we aim to create a comprehensive database of all allelic polymorphisms with allele frequencies of 1% or more in the 5'-3' direction.

2.3. Pilot study to investigate *AIRE* rs878081 and rs1003854 allelic polymorphisms in rheumatoid arthritis patient samples

As this was initially an unknown area for us, we needed to see if the allele frequencies of the selected AIRE domain coding region loci and our test design could give us valuable results. In order to clarify whether the allele-specific assays chosen to discriminate allelic polymorphisms and the protocol developed for genotyping are appropriate, we aimed to implement a pilot study to answer these questions. To this end, we selected two loci of the SAND region encoding the DNA-binding domain, rs878081 and rs1003854, and aimed to investigate the association of allelic polymorphisms and genotypes at these loci with the risk of rheumatoid arthritis in a case-control study using six different genetic models. We targeted this region because we believe the SAND domain encoded by this exon sequence is the key component of AIRE protein binding to mTEC DNA and, thus, negative selection. Within this, the allelic polymorphism at the rs878081 locus is a synonymous variant that can significantly affect the amount of protein produced and, thus, the efficiency of negative selection. For our study, we aimed to create a comprehensive database that includes all the artefacts from the RA and control populations of the pilot study and the possible subsequent study, recording all parameters that can be extracted from the artefacts, including demographic data, lifestyle data, laboratory test results and past infections. A further aim was to compare statistically, where possible, the mean values of clinical parameters for each genotype subgroup and, where significant, to investigate correlations or associations between clinical parameters and genotypes.

2.4. Analysis of *AIRE* exon 5 and 6 and intron 5 and 7 regions on the risk of rheumatoid arthritis

Depending on the results of our pilot study, we aim to continue the study with a larger number of cases and controls and more allelic polymorphisms to clarify whether allelic polymorphisms in the DNA-binding region are associated with the disease. To this end, using the in silico database and the results of the pilot study, we designed a panel including all potential loci and their allelic polymorphisms that define the SAND domain coding region (exons 5 and 6) and the splicing intronic region between them (introns 5 and 7). Our aim is to investigate the allelic polymorphisms in this panel and the genotypes of the alleles on the risk of rheumatoid arthritis using six different genetic models.

2.5. Examination of the association of *AIRE* exon 5 and 6 and intron 5 and 7 regions with disease-determining clinical parameters

Further, we wondered whether the genotypes of each allelic polymorphism are associated with clinical parameters and disease activity in RA. We aimed to investigate the genotypes defined by alleles of the loci in the panel, using five different genetic models to investigate the following clinical parameters: erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), rheumatoid factor (Rf), anti-cyclic citrullinated peptide antibody (aCCP), disease activity score 28 (DAS28).

2.6. *In silico* analysis of the effects of allelic polymorphisms on regulatory binding motifs

To conclude the study, we were interested in how the alleles of each locus might affect the DNA binding sites to which regulatory units, transcription factors (TFs) may bind, so we aimed to include all allelic polymorphisms in the panel for which a statistically significant result was obtained in the allele discrimination test with RA or, if not, for which the genotype of the allele was statistically significantly associated with clinical parameters. We aimed to analyse the effect on binding motifs and their associated regulatory unit in the HaploReg v4.2 in silico database.

3. MATERIALS AND METHODS

3.1. Meta-analysis of *AIRE* allele polymorphisms in autoimmune diseases

Literature was searched in databases using keyword searches. Publications were selected according to inclusion and exclusion criteria. Statistical analysis was performed using five different genetic models, Hardy-Weinberg equilibrium, heterogeneity, sensitivity analysis and publication bias.

3.2. *In silico* study to determine AIRE allele polymorphisms

A variant is included in our database if the frequency of the rare allele is between 0.01 (1%) and 0.5 (50%) or if it is included in a reference annotation database. The variant was excluded if the frequency of the rare allele fell below 0.01. Two genome-wide annotation systems, UCSC Genome Browser (GRCh38 assembly) and Ensembl automated human annotation system, were used.

3.3. Pilot study to investigate *AIRE* rs878081 and rs1003854 allelic polymorphisms in rheumatoid arthritis patient samples

Our study is interventional research, so we applied for ethical approval to the Deputy State Secretariat of the National Medical Officer for Health Administration (hereinafter: the National Medical Officer) in 2018 for a four-year interval, with the involvement of the Harkányi Thermal Rehabilitation Centre Nonprofit Nonprofit Ltd. To this end, the National Medical Officer has asked the Scientific Council and Research Ethics Committee for Health (ETT TUKEB) to formulate an expert opinion. The ETT TUKEB found the application for authorisation of the research to be professionally and ethically sound and gave its consent to the authorisation of the research by the national medical officer. On this basis, the National Medical Officer approved our application on 24 April 2018 for four years with the registration number 11871-7/2018/EÜIG.

All persons included in the study were outpatients and inpatients of the Harkány Thermal Rehabilitation Centre. Enrolment for the study was voluntary and was accompanied by a patient information leaflet and informed consent form. Blood samples were collected in 4mL citrate tubes per person. Whole blood was stored at -75˚C until analysis. A volunteer was included in the patient group if (1) he/she was over 18 years of age, (2) signed the informed consent form, and (3) had a diagnosis of rheumatoid arthritis according to the diagnostic and classification criteria of the American College of Rheumatology. Patients with a history of concomitant autoimmune conditions such as Sjogren's syndrome, psoriasis, SSc, and SLE were excluded from further analysis in order to reduce confounding effects of autoimmune diseases other than RA on RA risk.

As we knew from the beginning of the study that we would need to design an appropriate control population carefully, we wondered how to limit the incidence of possible false negatives for RA outcomes. Our experience of the disease course is that the average age of onset and hospitalisation is around 65 years. However, from a review of the literature and patients' medical reports, we have seen that in a minority of cases, the first hospitalisation may occur at the age of 70 years. Therefore, we decided to raise the age inclusion criterion for controls to 75 years or older because, without this, we might have included subjects in our control group with RA in the preclinical stage and with pathomechanisms of the disease in the incipient or central stage without the presence of symptoms. By the age of 75 years, patients are more likely to present with symptoms and become part of the care system at some point. Thus, we reduced confounding effects in our study. Thus, our inclusion criteria for control subjects were (1) age 75 years or older, (2) signed a consent form, (3) free of RA according to the diagnostic and classification criteria of the American College of Rheumatology, (4) free of other autoimmune diseases according to their medical records (e.g. SLE, SSc, psoriasis, Sjögren's syndrome), (5) CRP $\leq$ 20 mg/dl and Rf $\leq$ 25 IU/ml (153). Finally, we collected blood samples from 100 RA and 100 control subjects in our pilot study.

RA and control blood samples were primarily used to obtain complete blood count, cholesterol, blood glucose, ESR, CRP, sodium, potassium, urea, creatinine, glutamate oxaloacetate transaminase (GOT) levels, glutamate pyruvate transaminase (GPT), gamma-glutamyl transferase (gamma-GT), alkaline phosphatase, lactate dehydrogenase, creatine kinase, protein albumin, magnesium, phosphorus and serum bilirubin were determined. To determine the seropositivity and seronegativity of RA, Rf (IgG) and aCCP autoantibody levels were determined in both RA and control groups, the latter mainly to exclude autoimmune pseudo-negatives.

Following extraction and purification of genomic DNA, multiplex qPCR was performed on genotype rs878081 and rs1003854 allelic polymorphisms in the SAND region in both RA patients and controls using TaqMan technology. To ensure accurate allelic discrimination, we used Minor Groove Binder (MGB) TaqMan® Assays (Thermo Fisher Scientific, Waltham, USA), where MGB increases the melting temperature at the 3′ end of the oligonucleotide probe, thus stabilizing the probe-target hybrids.

Statistical analyses were performed using SPSS software version 28.0 (IBM Corp. Released 2021, IBM SPSS Statistics for Windows, Version 28.0; IBM Corp, Armonk, NY, USA). Expected genotype frequencies were calculated using the Hardy-Weinberg equilibrium (HWE), and the difference with observed frequencies was statistically compared using a non-parametric χ^2 test to see how well the observed genotype frequencies fit the expected frequencies according to the HWE. Continuous variables representing clinical parameters were plotted as mean ± standard deviation (SD). Shapiro-Wilk test was used to determine normality, and Levene's test was used to analyse the agreement of variances. Independent samples t-test and analysis of variance were used to compare groups for continuous variables. To compare the medians of non-normally distributed variables, Mann-Whitney U and Kruskal-Wallis H tests with two-sided significance were used. The correlation was tested using bivariate correlation, Pearson correlation coefficient and two-sided significance. Binary logistic regression was used to calculate ORs using 95% CI intervals. Statistical significance was determined when p-values were less than 0.05. For several reasons, we chose the conventional p-value of 0.05. First, it is a standardized threshold in scientific research. In addition, our study is basic biological research aimed at detecting new associations between hypothetical exposure and outcome, so we wanted to avoid correcting or modifying this standardized threshold. In this case, the choice of 0.05 is reasonable because it strikes an acceptable balance between first- and second-order errors. This means that the value strikes a reasonable balance between being too permissive, which would cause an increase in the risk of false positives, and being too strict, where the risk of false negatives would have increased. Although the significance is set at 0.05, higher significance levels (<0.01, <0.001) are also found in the tables.

3.4. Analysis of *AIRE* exon 5 and 6 and intron 5 and 7 regions on the risk of rheumatoid arthritis

In order to continue our study, we needed to collect additional samples and time, which required an extension of the ethics approval for the pilot study. In doing so, we expanded the source of our sample collection in cooperation with the existing Harkány Thermal Rehabilitation Centre and the St. Andrew Hospital for Rheumatology of Hévíz and applied for an extension of the ethical license for another four years. The ETT TUKEB, which was asked to provide an expert opinion, found our submitted amendment application to be professionally and ethically appropriate and agreed to approve the amendment to the research plan, based on which the National Center for Public Health approved our amendment application for another four years on November 8, 2022, by decision No. 57142-5/2022/EÜIG.

From then on, the sample collection was carried out in parallel at two institutions. The Harkány Thermal Rehabilitation Centre collected blood samples from RA and control subjects, while St. Andrew Hospital for Rheumatology of Hévíz collected blood samples exclusively from RA patients. In the present case, the study application was voluntary and accompanied by a consent form and a patient information leaflet. Blood samples were collected in 4mL citrate tubes per person and stored at -75˚C in our institute until analysis. Subjects who volunteered for the study were included in the RA group if they (1) were over 18 years of age, (2) signed a consent form, and (3) were diagnosed with RA according to the diagnostic and classification criteria of the American College of Rheumatology. RA patients with a history of concomitant autoimmune diseases such as Sjogren's syndrome, psoriasis, SSc, SLE were excluded from further analysis in order to exclude the possible association of autoimmune diseases other than RA with selected allelic polymorphisms and to reduce the confounding effect of autoimmune diseases other than RA on the risk of RA. In the present study, we used the age criterion of the control group used in the pilot study to avoid false RA negative cases, and thus, the age inclusion criterion for controls was raised to 75 years or older. Without this, we might have included subjects in our control group who were in the preclinical, asymptomatic stage of RA. Accordingly, our inclusion criteria for volunteer controls were (1) age 75 years or older, (2) signed a consent form, (3) free of RA according to the diagnostic and classification criteria of the American College of Rheumatology, (4) free of other autoimmune diseases according to their medical records (e.g. SLE, SSc, psoriasis, Sjögren's syndrome), (5) CRP ≤ 20 mg/dl and Rf ≤ 25 IU/ml (153). By the end of the total sample collection in December 2023, a total of 592 subjects had been collected, of which 270 were RA patients, and 322 were controls.

Blood samples from 592 subjects were primarily used to determine the following clinical parameters: complete blood count, cholesterol, blood glucose, ESR, CRP, sodium, potassium, urea, creatinine, GOT, GPT, gamma-GT, alkaline phosphatase, lactate dehydrogenase, creatine kinase, protein albumin, magnesium, phosphorus and serum bilirubin. Serum Rf (IgG) and aCCP were determined in order to confirm our RA group and to exclude pseudo-negatives from the control.

We used 20 ng gDNA per reaction for our multiplex qPCR assay. In this more extensive second study, we also used TaqMan technology to genotype allelic polymorphisms. By using Minor Groove Binder (MGB) TaqMan® Assays (Thermo Fisher Scientific, Waltham, USA), we were able to generate more stable probe-target hybrids for allelic discrimination. In our pilot study, we successfully used the products for genotyping, so we have chosen the same assay set-up for the present study. TaqMan® assays include a primer pair, a VIC dye-labelled oligonucleotide assay (530 nm yellow channel) that detects allele 1 sequence, and a FAM dye-labelled assay (470 nm green channel) that detects allele 2 sequence. A genotype is homozygous if only the VIC or the FAM channel shows a fluorescent signal. The genotype is heterozygous if a signal is present in both channels.

Our statistical analyses, performed with the trusted SPSS software version 28.0 (IBM Corp. Released 2021, IBM SPSS Statistics for Windows, Version 28.0; IBM Corp, Armonk, NY, USA), were a key part of our research. We tested the expected and observed genotype frequencies based on HWE using a non-parametric χ^2 test to determine if there was a significant difference between the two. Linkage was assessed using Lewontin's "D" method, and the haplotype block capturing the LD structure was generated using Haploview software (version 4.2). Clinical continuous variables were plotted as mean ± SD. Shapiro-Wilk test was used to determine normality, and Levene's test was used to analyse the variance agreement. Independent samples t-test and analysis of variance were used to compare groups for continuous variables. Mann-Whitney U and Kruskal-Wallis H test with two-tailed significance were used to compare medians of non-normally distributed variables. Correlation was tested using bivariate correlation, Pearson correlation coefficient and two-tailed significance. Binary logistic regression was used to calculate OR values, generating 95% CI and p-value. Statistical significance was defined as a p-value less than 0.05.

3.5. *In silico* analysis of the effects of *AIRE* allelic polymorphisms on regulatory binding motifs

The effect of each allele polymorphism on the binding affinity of transcription factors was analysed *in silico* using the HaploReg v4.2 database. The database was used to calculate potential transcription factor binding affinity values for the rare and common alleles.

4. RESULTS

4.1. Meta-analysis of *AIRE* allelic polymorphisms in autoimmune diseases

After removing duplicates, a search of four literature databases identified 11 relevant publications, five of which dealt with rheumatoid arthritis and the remaining publications dealt with one disease. Thus, we focused our meta-analysis on this study. The five selected case-control analyses identified 11 SNPs, of which rs2075876 and rs760426 allele polymorphisms were examined in the majority of case-control analyses, with 7145 cases, 8579 controls and 6696 cases, 8164 controls, respectively. The 50 weighted odds ratios of the five genetic models selected to examine these SNPs showed statistically significant evidence that the A allele of rs2075876 and the G allele of rs760426 are risk factors for disease development in Asians, except for 9.

4.2. *In silico* study to determine *AIRE* allelic polymorphisms

We set our search to the AIRE coding sequence in UCSC Genome Browser and Ensembl's automatic annotation system (HGNC Symbol; Acc: HGNC:360). We identified 23 023 unconstrained variants, including point mutations, indels and SNPs. Based on the global frequency of occurrence of the rare allele, 68 SNPs were detected, which were examined individually in a European population, and finally, 44 SNPs with rare allele frequencies between 0.01 (1%) and 0.5 (50%) were selected for inclusion in our database or in a reference annotation database. We then selected allelic polymorphisms in the DNA segment encoding the SAND region for further analysis based on their role in negative selection.

4.3. Pilot study to investigate *AIRE* rs878081 and rs1003854 allelic polymorphisms in rheumatoid arthritis patient samples

For our pilot study, we collected 100 cases and 100 control samples from the Harkány Thermal Rehabilitation Centre, where patients showed significantly higher rheumatoid arthritis marker values.

In the framework of our allelic model, we searched for the risk allele at the SAND region coding loci rs878081 and rs1003854, i.e., we first examined which allele frequency was higher in the patient population, and then calculated odds ratios using binary logistic regression to find out which allele was associated with disease risk. We then developed our dominant, recessive, codominant heterozygous and codominant homozygous and over-dominant population genetic models for the disease risk allele for both the rs878081 and rs1003854 loci. All control groups were Hardy-Weinberg balanced.

Based on the recessive genetic model, we found a statistically significant association with RA, where the frequency of TT homozygotes was significantly higher in RA patients than in controls, supporting the risk of RA (OR=5.44, 95% CI=1.16-25.52, p=0.032). The genotypic frequency of TT homozygotes was also higher in RA patients using the codominant homozygote genetic model than in controls; however, the association was not significant (OR=4.68, 95% CI=0.98-22.27, p=0.052), but was in the same direction as the recessive model (152). No further significant association was found in the allelic, dominant, codominant heterozygote or over-dominant genetic models.

Allelic polymorphisms of rs1003854 were examined in allelic, dominant, recessive, codominant heterozygous, codominant homozygous and over-dominant genetic models. Genotype frequencies were statistically significantly higher in patients with RA in TT homozygotes than in controls, supporting the risk of RA (OR=4.84, 95% CI=1.02-23.02, p=0.047). No further significant association was found in allelic, dominant, codominant heterozygous or over-dominant genetic models.

In RA patients, the CC genotype subgroup had significantly higher ESR levels compared to TT homozygotes in the rs1003854 codominant homozygote model (p=0.048). The CC genotype subgroup had significantly higher CRP levels compared to TT homozygotes in the codominant homozygote model and compared to the TT+TC subgroup in the recessive model (p=0.043 and p=0.029). Furthermore, patients with CC homozygous RA showed a significant correlation with CRP in the codominant homozygous and recessive models (p=0.006 and p=0.002 and r=0.350 and r=0.338, respectively).

4.4. Examination of allelic polymorphisms in *AIRE* exon 5 and 6 and intron 5 and 7 regions on the risk of rheumatoid arthritis

In addition to the two loci tested in the pilot, we selected three loci that contribute strongly to the expression of the SAND domain. Thus, the loci selected for this study in the 5'-3' direction are rs878081 in exon 5, rs1003853 and rs2075876 in intron 5, rs1055311 in exon 6 and rs1003854 in intron 7.

In this study, two groups were formed, with Group 1 representing the total population of 592 with 270 cases and 322 controls, to which we selected the three additional *AIRE* loci rs1003853, rs2075876 and rs1055311. Group 2 is a subgroup of the total population with 170 RA and 222 controls, for which we selected *AIRE* rs878081 and rs1003854 squirrels.
In the present case, we first used the allelic genetic model to find which alleles had a higher prevalence in the patient population, and then calculated OR values using binary logistic regression to find out the direction of disease risk. We then developed our dominant, recessive, codominant heterozygous and codominant homozygous and over-dominant population genetic models for the allele that conferred disease risk. This methodology was performed for all five loci, rs878081, rs1003853, rs2075876, rs1055311 and rs1003854, for the total study population of 592 individuals in groups 1 and 2, using discriminant test results for 1184 alleles.

In our study, a significant association was determined by testing 1184 alleles of the total 592-strong Group 1 population. The frequency of the C allele of the rs1003853 splice variant located in the intron 5 region of *AIRE* was significantly higher in RA patients than in controls, and binary logistic regression results showed a statistically significant association between the C allele of *AIRE* rs1003853 and RA disease (OR=1.33, 95% CI 1.01-1.74, p=0.037). In the population of 592 patients, the frequency of rs1003853 CC homozygous genotype was also statistically significantly higher in RA patients than in controls, resulting in a statistically significant association between CC homozygous genotype and RA in the recessive genetic model (OR=1.618, 95% CI 1.16-2.25, p=0.004).

Among the five loci, the C allele of rs878081 encoding the SAND domain of *AIRE* SAND exon 5 in group 2 showed a significant association with RA risk according to the allelic model (OR=1.48, 95% CI 1.05-2.09, p=0.023). Here, based on the recessive genetic model, the genotype frequencies of CC homozygous RA patients were significantly higher compared to the sum of genotype frequencies of CT heterozygotes and TT homozygotes. Based on a recessive genetic model, the association between CC genotype and RA was statistically significant (OR=1.64, 95% CI 1.09-2.48, p=0.01).

In group 2, we next identified the rs1003854 locus of *AIRE* intron 7, where the odds of the T allele occurring were significantly higher among RA alleles compared to the C allele (OR=1.52, 95% CI 1.08-2.12, p=0.014). Based on the recessive genetic model, the odds of TT homozygous genotype were significantly higher in RA patients compared to controls (OR=1.72, 95% CI 1.15-1.44, p=0.008). Among RA patients, statistically significant differences were found in the rs878081 recessive model of group 2, where CC homozygotes had significantly higher mean ESR levels (p=0.027) compared to the TT+CT genotype subpopulation. Using Pearson's bivariate correlation, we found a statistically significant correlation between ESR and CC homozygote genotype (p=0.023, r=0.190), and using binary logistic regression, we also found a statistically significant correlation (OR=1.01, 95% CI 1.002-1.032, p=0.026) between the two variables. The genotype group of rs1055311 locus codominant heterozygous model CT and dominant model CC+CT with 592 subjects in 270 RA patients in group 1 showed significantly higher aCCP levels (p=0.028, r=0.250 and p=0.044, r=0.150, respectively).

4.5. Allele-specific affinity of regulatory binding motifs for transcription factors

The allelic polymorphisms rs878081 on exon 5 and rs1055311 on exon 6 in the SAND domain are synonymous polymorphisms that can cause no structural loss of function in the protein, but allele-specific functional differences. In order to clarify the possible transcriptional consequences, we performed allele-specific *in silico* analysis using the HaploReg v4.2 database to see if each allele polymorphism occupies a position in the DNA binding region motifs of certain TFs and if so, whether each allele affects the binding affinity of the TF (Regulatory motifs altered).

We investigated the four loci whose allelic polymorphisms were associated with RA risk or clinical parameters such as ESR or aCCP within RA groups. A site-weighted matrix prediction model for rs878081 revealed that the locus itself is located in the sequence of NF-κB and Yin-Yang 1 (YY1) binding motifs.

The LOD value calculated for the NF-κB NF-kappaB_disc2 DNA binding motif is +11.1 higher for the T-allele and +2 for NF-kappaB_known4, indicating that NF-κB binds less strongly to DNA in the presence of the reference C-allele. At the same locus, the LOD value of affinity for the alternative T-allele binding motif YY1_disc1 is +0.9 higher, also indicating that this TF binds more weakly in the presence of the reference C-allele.

Further, the site-weighted matrix prediction model of rs1003853 revealed that the locus position itself is located in the sequence of the Krüppel-like zinc finger protein AP-2rep (AP-2rep) and the binding motifs of the transcription factors Rad21. The LOD value calculated for the AP-2rep DNA binding motif is +7.1 and +0.5 higher for the Rad21_disc5 binding motif for the T allele, i.e. in the presence of reference allele C, these two TFs bind more weakly.
According to the site-weighted matrix prediction model, the locus position for rs1055311 is located in the myeloid zinc finger 1 (MZF1) binding motif MZF1::1-4_1, the transcription factor NF-κB NF-kappaB_disc2 and the p300 transcriptional coactivator p300_known1 binding motif. Predictive LOD values are consistently +1.6 and +11.1 for MZF1::1-4_1 and NF-kappaB_disc2, respectively, indicating stronger affinity for TF binding in the presence of the alternative T allele and weaker affinity in the presence of the reference C allele. For p300_known1, the LOD value is reversed in favour of the reference C allele with a value of +1.3, which in the reverse situation, unlike the previous one, the alternative T allele binds the p300 transcriptional co-activator more weakly.

Based on the prediction model, the rs1003854 locus is located in the GATA_disc3 binding motif sequence of the zinc finger protein family GATA, which can bind to the (T/A)GATA(A/G) consensus DNA sequence. The predictive LOD value is +1 higher for the alternative C allele, i.e. the reference T allele binds this transcription factor more weakly. The results of this analysis are demonstrated in Table 1.

Locus	Ref	Alt	Position Weight Matrix ID	LOD value Ref	LOD value Alt
rs878081	C	T	NF-kappaB_disc2	1,5	12,6
			NF-kappaB_known4	12.5	14.5
			YY1_disc1	8,6	9,5
rs1003853	C	T	AP-2rep	3,0	10,1
			Rad21_disc5	10,2	10,7
rs1055311	C	T	MZF1::1-4_1	10,3	11,9
			NF-kappaB_disc2	1.7	12.8
			p300_known1	11,1	9,8
rs1003854	T	C	GATA_disc3	12.2	13.2

Ref: reference allele; Alt: alternative allele; LOD: log odds ratio

Table 1. Based on the prediction model of the HaploReg4.2 database, the position of *AIRE* allele polymorphisms in the DNA binding region motifs of transcription factors and their allele-specific impact on TF binding affinity.

5. DISCUSSION AND CONCLUSION

One of our research's main aims was to better understand the association of the AIRE transcription factor with possible autoimmune diseases. As AIRE functions to filter and delete autoreactive thymocytes, its expression is locally undetectable in young adulthood following the fatty atrophy of the thymus, which begins at puberty, supporting the hypothesis that AIRE transcription factor function is associated with infancy. Furthermore, the age-related natural loss of thymus function results in the immune system operating with an existing T-cell repertoire that is no longer largely replenished by adulthood, with naïve T-cell production significantly reduced. Thus, if the existence of a genetic-based disease risk is shown to be associated with the transcription factor AIRE, it would also be associated with infancy, which would shed new light on the pathomechanism of the disease. AIRE is responsible for tissue antigen production underlying negative selection. Suppose it is not produced or is not produced in sufficient amounts. In that case, this step of central tolerance may be defective or absent, increasing the clonal survival of autoreactive naive T cells. Above, I have described the work of *Lovewell et al.* in which two genotypes with *AIRE-655G AIRE-230T* allelic polymorphisms at loci rs117557896, rs751032 had little promoter activity and no negative selection. Suppose the presence of two allelic polymorphisms can result in such a drastic expression difference. In that case, it is conceivable that a set of allelic polymorphisms in the *AIRE* promoter region and their haplotypes could generate an individualized *AIRE* expression profile. For the population as a whole, this will provide sufficient expression in the majority of cases, however, there may be a combination of haplotype combinations affecting *AIRE* transcription that do not provide sufficient AIRE protein for tissue antigen expression, negative selection may fail, leading to clonal expansion of autoreactive naive T cells. Many autoimmune diseases are caused by autoreactive T-cell populations, so if AIRE is a key player in immunological tolerance, it is conceivable that it could be a common denominator, or causal factor, of many autoimmune diseases through allelic polymorphism, and hence AIRE could be the starting point of many autoimmune diseases. To investigate this, infant *AIRE* expression profiling, haplotype analysis on thymic tissue samples, and lifetime cohort follow-up of infants in a large-element count study against controls would be required. In addition, a study of the entire actor pool of central tolerance could provide a more detailed picture. We have not attempted to perform studies of this degree of complexity. Each step of our series of tests evolved from the others. With our meta-analysis, we aimed at the very beginning of the study to determine which autoimmune disease is most closely associated with *AIRE* allelic polymorphisms according to the current state of the art. In our publication, having determined that it is rheumatoid arthritis, we clarified which allelic polymorphisms give a significant association with which race. The largest population we could include was Asian, in which we determined for RA the locus A allele of rs2075876 and the G allele of rs760426, and genotypes carrying these alleles, whether homozygous or heterozygous, were significantly associated with the disease in almost all genetic models. Our meta-analysis therefore identified which allele polymorphisms are at risk for RA in Asian populations based on the state of the literature up to 2018. After allelic polymorphisms were shown to be associated with the disease, however, no significant association was obtained for European populations so we were curious to see if any allelic polymorphisms existed that might be at risk among Caucasian individuals. This led us to plan the next steps, whereby we detected 44 SNPs in the *AIRE* region with *in silico* analysis that could potentially contribute to the development of autoimmune disease. Of these 44, two were selected to study the SAND domain encoding the DNA-binding region: loci rs878081 and rs1003854. The primary role of our pilot study was to search for correlations and to optimize the assay methods. Based on our results, rs878081 and rs1003854 loci showed significant association with disease, carrying the disease risk allele was associated with risk of blood sedimentation and CRP elevation. Our publication is the first to associate genotypes from the rs878081 and rs1003854 loci with RA, and also the first to find an association between genotypes and clinical parameters (ESR and CRP) in the patient population. RA patients who carried two risk alleles at the rs1003854 locus had significantly higher CRP in both recessive (p=0.029) and codominant (p=0.043) genetic models, which genetic constitution also showed a positive correlation with CRP in both models (p=0.002 and p=0.006). This may be explained by the study of *Bergström et al.* During the central phase of the disease, an inflammatory environment develops in the synovial capsule, during which fibroblast like synoviocyte (FLS) cells produce chemokines (CXCL10, CCL2, CCl5, CCL8), with the presence of VCAM-1 cells promoting the extravasation of circulating autoreactive T cells. Experimentally, it was found that when AIRE expression was silenced by small non-coding siRNA in FLS populations, cells exhibited 26x higher chemokine production compared to unstimulated cells where chemokine production was undetectable, i.e. the AIRE-silenced FLS cell population had higher inflammatory activity. Allelic polymorphisms as we have seen can contribute greatly to AIRE expression, i.e. it is conceivable that the genotype homozygous for the risk allele at the rs1003854 locus of *AIRE* greatly reduced expression, which decades later, through FLS cells, created a chemokine-rich environment for autoreactive T cells. This is supported by the strong association of chemokines (CXCL10, CCL8 and CCL5) with CRP levels.

Following our pilot study, we extended our study to 3 additional loci and 592 subjects. To investigate the association between genetic exposure and disease, we examined allelic polymorphisms at five loci, of which rs878081 in exon 5 of the SAND domain, rs1003853 in intron splice site 5 and rs1003854 in intron splice site 7 showed a significant positive association with disease risk when the genotype was homozygous for the risk allele (p=0. 017, p=0.004, p=0.008). And for two loci (rs878081, rs1055311), there was an association between genetic subgroups of RA patients and ESR and aCCP clinical parameters. Based on our results, allele and genotype frequencies of the rs1055311 locus do not show a significant association with disease, but correlate with aCCP levels. This may occur, among other reasons, because the allelic polymorphism acts epistatically on an intermediate variable that determines autoantibody production, i.e. an intermediate phenotype, which is thus not directly associated with disease risk. It may influence the survival of autoreactive T cells that intrinsically promote the production of aCCP antibodies, characterising an intermediate disease phenotype.

Based on all these disease risk results, we therefore believe that in the presence of intronic allelic polymorphisms of both exon synonymous and splice region intronic allelic polymorphisms and their genotype, AIRE expression is sufficiently low to be a candidate for this disease risk, to allow expression of tissue antigens, allowing the release of autoreactive thymocytes from control, which, transformed into naive T cells, survive clonally in secondary lymphoid tissues and remain hidden until the extravasation of the central stage. From there on, they trigger the pathomechanism in several directions, on the one hand producing cytokines (IFN-γ, IL-17) which subsequently attract macrophages and maintain a circulus vitious, and on the other hand autoreactive CD4+ T cells causing expansion of the FLS cell population resulting in the formation of scar connective tissue. In individuals homozygous for risk alleles, it is conceivable that low AIRE expression due to genetic risk creates a more active inflammatory environment than average, by producing a broad spectrum of chemokines in the absence of AIRE in FLS cells, resulting in the production of additional cytokines, autoantibodies, and accumulated immune complexes.

As a final step in our study, we analysed *in silico* the molecular consequences of our risk alleles to further explore causal relationships. In doing so, we found that 4 of the 5 loci that showed a statistically significant association with either disease risk or clinical parameters are located in the DNA binding motif sequence of transcription factors and *in silico* data suggest that they may affect their binding affinity. The results were dichotomized according to whether transcription factor binding is stronger or weaker in the presence of the reference (Ref) allele to the DNA binding motif sequence in which the allele polymorphism is located. It is not always possible to determine the transcription fate, with one or two exceptions, they are highly context-dependent. We have found that in the presence of the Ref allele (C) of exon 5 rs878081 and exon 6 rs1055311, NF-κB binds more weakly. Studies by *Haljasorg et al.* suggest that strong affinity binding of this transcription factor to mTEC DNA is required for negative selection AIRE expression in thymus (35). For both loci, the allele at risk was the Ref C allele, whose homozygous genotype was associated with disease risk for rs878081 and with aCCP for rs1055311. Based on our case-control and *in silico* results, we hypothesize that in the presence of this allele, NF-κB was unable to bind with sufficient affinity to mTEC DNA, resulting in low AIRE expression and consequent tissue antigen expression, which may have led to partial or complete failure of negative selection. Further, rs878081 was associated with a weaker binding of YY1, a zinc finger multifunctional transcription factor, to the Ref allele, which is an activator or repressor depending on the context. An example of the latter, when bound to the target promoter region in the presence of histone deacetylase (HDAC), causes transcriptional repression, chromatin condensation. Orders of magnitude weaker binding to the Ref (C) allele of rs1003853 Ref (C) of intron 5 is associated with the transcription factor AP-2, which binds to promoter and enhancer regions with its DNA-binding domain, its strong affinity activating transcription. Also in the presence of the Ref allele of this locus, Rad21, whose strong affinity binding is required for DNA repair, is part of a cohecin complex that promotes chromosome segregation during meiosis and mitosis. Its weak affinity binding may influence these functions (178). Our results suggest that the Ref (C) allele of rs1055311 locus has weaker binding of the transcription factor MZF1, which has a role in, among other things, myeloid DC cell development. The GATA family of transcription factors binds more weakly to the TF binding motif of DNA in intron 7 of rs1003854 Ref (T). GATA is essential for the maturation of thymocytes, affects β-selection and determines the commitment of cell lines to central tolerance (180). As this allelic polymorphism is intronic, it is conceivable that it may be able to affect AIRE expression through a process of alternative splicing (181). 1 of the TFs bound more strongly to the Ref allele, this is p300 to the C allele of rs1055311. Histone acetyl transferase (HAT) is an enzyme that transports an acetyl group onto the lysine side chain of histone molecules, relaxing chromatin as a consequence of electrical charges and making DNA available to RNA polymerase II (RNA pol II), i.e. activating transcription. As a result of our*in silico* study, 4 out of 5 loci had a number of TFs with weak affinity for the Ref allele-containing DNA binding motif, with a strong affinity for transcription factors that essentially determine gene expression, thus predicting TRA-driven negative selection for *AIRE* allele polymorphism-dependent expression.

In addition to our results, our studies have several limitations. The main limitation that could negatively affect the strength of the association is the limited number of elements. This was primarily due to the drop in patient volume due to the COVID-19 pandemic. Secondarily, the net patient turnover of the two hospitals that were able to enroll so many RA patients and controls in the study between 2019 and 2023. This was the maximum that the two hospitals were able to provide during a pandemic-loaded 4-year sample collection period, for which we cannot but be grateful to the physicians, doctors and nurses of the Harkány Thermal Rehabilitation Centre and St. Andrew Hospital for Rheumatology of Hévíz, without whom this series of studies would not have been possible. Increasing the number of participants in our study would have significantly increased both the number of correlations and the statistical significance. An additional limitation is that we included a domain coding region in the study, the results of which cannot provide comprehensive insights into the function of *AIRE* and the measured expression consequences of certain allelic polymorphisms *on AIRE*. We envision that genome-wide analysis and results of all allelic polymorphisms and AIRE partners in all domain coding regions would not provide valuable results without examining them in isolation and in isolation from each other, which we plan to pursue in the future. A further limitation is that we were not able to investigate the mRNA expression of *AIRE*, which would have greatly contributed to a better understanding of allele- or genotype-dependent transcriptional consequences. Our future plans include a more comprehensive understanding of the relationship between RA and central tolerance and the role of AIRE partners in further genetic association studies of the remaining 11 intronic and 12 exonic coding sequences.

6. SUMMARY OF NOVEL FINDINGS

In total, our study consisted of 6 parts. First, we performed a meta-analysis of allelic polymorphisms at the rs2075876 and rs760426 loci in Asian populations. Here, for the first time in 2018, we determined that the risk A allele of rs2075876 and the risk G allele of rs760426 and their genotypes and genotype groups were statistically significantly associated with RA in Asian populations only. Our research team published this association for the first time.

Next, based on MAF allele frequencies, we identified 44 SNPs that could potentially be tested in European Caucasian populations and selected two of them, rs878081 and rs1003854, to perform our pilot study, which concluded with the first determination of whether the risk alleles and genotypes of rs878081 and rs1003854 were statistically associated with disease risk and clinical parameters of ESR and CRP, respectively. As the first study in a European population, our research team conducted a pilot study to analyse the association between

allelic polymorphisms and RA and clinical parameters using six different genetic models (allelic, dominant, recessive, codominant hetero- and homozygous).
Based on the results obtained so far, we extended the number of elements of our next study to 529 individuals and the population of loci to 3 additional positions and found that the risk C- and T-alleles of rs878081 and rs1003853 and their genotypes, respectively, of the five selected loci were statistically significantly associated with RA disease risk. Among clinical parameters, ESR showed a statistically significant association and aCCP showed a correlation with the risk C-allele homozygous genotype for rs878081 and rs1055311.
As a final step in our series of studies, we analysed *in silico* the transcriptional consequences of each allele polymorphism, by first determining the behaviour of 7 different transcription factors in the presence of risk allele polymorphisms. Of these, it is consistent that in the presence of risk SNPs, 6 out of 7 TFs bind more weakly to the DNA binding motif, of which 2 out of 4 SNPs showed weaker binding of NF-κB, which is responsible for *AIRE* transcription that ensures TRA expression during neonatal negative selection. Our results suggest that one possible consequence of the presence of these risk alleles is a weaker binding of NF-κB, resulting in insufficient production of AIRE or TRA during neonatal negative selection, which predicts the possible clonal survival of autoreactive thymocytes, thus providing the initial immunological basis for the development of autoimmune diseases, including RA.

7. REFERENCES

Bérczi B, Gerencsér G, Farkas N, Hegyi P, Veres G, Bajor J, Czopf L, Alizadeh H, Rakonczay Z, Vigh É, Erőss B, Szemes K, Gyöngyi Z. Association between AIRE gene polymorphism and rheumatoid arthritis: a systematic review and meta-analysis of case-control studies. Sci Rep. 2017 Oct 26;7(1):14096. doi: 10.1038/s41598-017-14375-z. PMID: 29074995; PMCID: PMC5658331 **(D1, IF: 4,6)**

Berczi B, Nusser N, Peter I, Nemeth B, Gyongyi Z. Association Between AIRE Polymorphisms rs870881(C>T), rs1003854(T>C) and Rheumatoid Arthritis Risk: A Hungarian Case-control Study. In Vivo. 2024 Mar-Apr;38(2):774-784. doi: 10.21873/invivo.13501. PMID: 38418155; PMCID: PMC10905445. **(Q2, IF: 2,4)**

Bérczi B, Nusser N, Péter I, Németh B, Kulisch Á, Kiss Z, Gyöngyi Z. Genetic Polymorphisms in Exon 5 and Intron 5 and 7 of AIRE Are Associated with Rheumatoid Arthritis Risk in a Hungarian Population. Biology. 2024; 13(6):439. https://doi.org/10.3390/biology13060439 **(Q1, IF: 4,2)**

8. ACKNOWLEDGEMENTS

I would like to thank Prof. Dr. István Kiss, Head of the Department of Public Health Medicine for allowing me to work in the department he leads.
I would also like to thank Dr. Zoltán Gyöngyi, Associate Professor, my tutor, for his professional and human support during my work.
I am grateful to Dr. Iván Péter, the Director of the Harkány Thermal Rehabilitation Centre, for allowing me to do my work in the hospital he runs.
I would like to thank Dr. Balázs Németh, Assistant Professor, for his help in the initial steps of the research.
I am grateful to Dr. Nóra Nusser, Medical Director of the Harkány Thermal Rehabilitation Centre, for the immense professional help, organisation, kindness, trust and humanity she has

shown me during my work, without which this study in its present form would have been unthinkable.

I owe my thanks to Zsuzsanna Feledi, head of the laboratory of the Harkány Thermal Rehabilitation Centre, for bringing together the team working on the samples, and to Zita Tóthné Kamanczi, laboratory technician, for systematically and tirelessly collecting and cataloguing our samples. I am also grateful to Natália Pappné Netik, Anita Óbert and Adrienn Szűcs for their help in collecting samples in Harkány.

I am also grateful to Dr. Zsuzsanna Holléné Dr. Mándó, the Director of the St. Andrew Hospital for Rheumatology of Hévíz, for allowing me to work in collaboration with the hospital she runs.

I am grateful to Dr. Ágota Kulisch, Chief Physician of the St. Andrew Hospital for Rheumatology of Hévíz, for her great help in making the collaboration and sample collection possible. Without her, it would also have been inconceivable that our study would have been carried out.

I would also like to thank Dr. Jánosné Fülöp, the head nurse of the St. Andrew Hospital for Rheumatology of Hévíz, for her immeasurable help during the sample collection in Hévíz.

I would also like to thank all the staff of the Department of Public Health Medicine for their help during my work.

Last but not least, I would like to thank my wife for her support and encouragement, which contributed greatly to the completion of this thesis. I would also like to thank my entire family for their support and encouragement, which has contributed greatly to the completion of this thesis.

For the support of this research I would like to thank the PTE ÁOK, Faculty Research Fund (ÁOK-KA) grant KA-304516.