THE APPROACH TO PATIENT CLUSTERING BASED ON THE MICROCHIP DATA CONFINED TO DISTINCT LOCI USING THE COMBINATIONS OF VARIANTS

Iulmetova LN, Kulemin NA, Sharova EI ✉

Lopukhin Federal Research and Clinical Center of Physical-Chemical Medicine of the Federal Medical Biological Agency, Moscow, Russia

Fuchs' endothelial corneal dystrophy is a socially significant hereditary disease. More than a half of cases in the European population are caused by the increased number of trinucleotude repeats in the *TCF4* gene. The study was aimed to develop and test the approach of dividing patients into groups based on the chip-based genotyping and genome-wide association study (GWAS) results. The analysis was conducted using FECD Genetics Multi-center Study and AREDs project datasets containing the data of 1721 clinical cases and 2408 control patients. When analyzing the GWAS results, the patients and the control group were divided into two groups by means of hierarchical clustering suggesting that patients with the increased number of repeats in the *TCF4* gene are carriers of specific combinations of genomic variants (haplotypes). It was shown that individual variants cannot be used for the molecular genetic stratification of patients with the increased number of repeats in *TCF4* due to inconsistent results obtained for the variants. Furthermore, the haplotype-based approach outperformed the SNPs in terms of odds ratio. The paper proposes a method that enables further search for the biologically relevant combinations of genomic variants.

Keywords: genome wide association study, Fuchs endothelial corneal dystrophy, trinucleotide repeat expansion, patient stratification, locus

Funding: the study was supported by the President of the Russian Federation (grant for young postdocs MK-2951.2022.1.4)

Acknowledgements: the authors thank dbGaP for providing access to the phs000421.v1.p1 and phs000001.v3.p1 datasets. The dataset with the dbGaP registration number phs000421.v1.p1 was obtained from the genetic study of Fuchs' endothelial corneal dystrophy (FECD) available from https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000421.v1.p1. The authors recognize the grants that have been used to support registration of cases and controls and would be used in this GWAS: R01EY016514 (DUEC, PI: Gordon Klintworth), R01EY016482 (CWRU, PI: Sudha Iyengar) and 1X01HG006619-01 (PI: Sudha Iyengar, Natalie Afshari). The authors express their gratitude to the FECD study participants and the FECD research team for their valuable contribution to this study. The dataset with the dbGaP registration number phs000001.v3.p1 was obtained from the Age-Related Eye Disease Study (AREDS) database available from https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000001.v3.p1. AREDS was funded by the National Eye Institute (N01-EY-0-2127). The authors thank the AREDS participants and research team for their valuable contribution to this study. The authors express their gratitude to L.O. Skorodumova, research fellow at the Lopukhin Federal Research and Clinical Center of Physical-Chemical Medicine, for valuable suggestions, comments, and support.

Author contribution: Sharova EI — concept and selection of data; Sharova EI, Iulmetova LN — planning and selection of methods; Kulemin NA — project funding and management; Iulmetova LN — design and computation; Sharova EI, Iulmetova LN, Kulemin NA — discussion, manuscript writing and editing.

Compliance with ethical standards: the study was performed according to the principles of the Declaration of Helsinki using the data of the phs000421.v1.p1 and phs000001.v3.p1 projects, the access to which was approved and provided by dbGaP in accordance with the policy of approval and access to specific datasets.

✉ **Correspondence should be addressed:** Elena I. Sharova
Malaya Pirogovskaya, 1, str. 3, Moscow, 119435, Russia; sharova78@gmail.com

Received: 12.12.2022 **Accepted:** 20.01.2023 **Published online:** 12.02.2023

DOI: 10.24075/brsmu.2023.001

ПОДХОД К КЛАСТЕРИЗАЦИИ ПАЦИЕНТОВ ПО МИКРОЧИПОВЫМ ДАННЫМ ВНУТРИ ОТДЕЛЬНЫХ ЛОКУСОВ С ИСПОЛЬЗОВАНИЕМ КОМБИНАЦИЙ ВАРИАНТОВ

Л. Н. Юльметова, Н. А. Кулемин, Е. И. Шарова ✉

Федеральный научно-клинический центр физико-химической медицины имени Ю. М. Лопухина Федерального медико-биологического агентства, Москва, Россия

Дистрофия роговицы Фукса является социально значимым наследственным заболеванием. Более половины случаев в европейской популяции вызваны увеличением числа тринуклеотидных повторов в гене *TCF4*. Целью исследования было разработать и проверить подход разделения пациентов на группы на основе результатов чип-генотипирования и полногеномного ассоциативного исследования (GWAS). В качестве исходных данных использовали датасеты FECD Genetics Multi-center Study и проекта AREDs в количестве 1721 клинических случаев и 2408 контрольных пациентов. При анализе результатов GWAS было проведено разделение пациентов и группы контроля на две группы методом иерархической кластеризации с учетом предположения, что пациенты с увеличенным числом повторов в гене *TCF4* имеют определенные сочетания геномных вариантов (гаплотипов). Показано, что одиночные варианты не могут быть использованы для молекулярно-генетической классификации пациентов с увеличенным числом повторов в гене *TCF4* из-за рассогласованности результатов по вариантам. При этом гаплотипный подход превзошел анализируемые варианты по параметру отношения шансов, перекрывая 95%-й доверительный интервал выборок двух экспериментальных исследований. Предложенный метод позволяет продолжать поиск биологически обоснованных сочетаний геномных вариантов.

Ключевые слова: полногеномное ассоциативное исследование, эндотелиальная дистрофия роговицы, экспансия тринуклеотидных повторов, классификация пациентов, локус

Финансирование: работа выполнена в рамках гранта Президента РФ для молодых ученых-кандидатов наук МК-2951.2022.1.4.

Благодарности: авторы благодарят dbGaP за предоставление доступа к наборам данных phs000421.v1.p1 и phs000001.v3.p1. Набор данных с регистрационным номером dbGaP phs000421.v1.p1 получен из исследования генетики эндотелиальной дистрофии роговицы Фукса (FECD) https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study. cgi?study_id=phs000421.v1.p1. Авторы признают гранты, финансировавшие регистрацию случаев и контролей, которые будут использоваться в этом GWAS: R01EY016514 (DUEC, PI: Gordon Klintworth), R01EY016482 (CWRU, PI: Sudha Iyengar) и 1X01HG006619-01 (PI: Sudha Iyengar, Натали Афшари). Авторы благодарят участников FECD и исследовательскую группу FECD за их ценный вклад в это исследование. Набор данных с регистрационным номером dbGaP phs000001.v3.p1 получен из базы данных исследования возрастных заболеваний глаз (AREDS) по адресу https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study. cgi?study_id=phs000001.v3.p1. Финансовая поддержка AREDS была предоставлена Национальным глазным институтом (N01-EY-0-2127). Авторы благодарят участников AREDS и исследовательскую группу AREDS за их ценный вклад в это исследование. Авторы благодарят научного сотрудника ФГБУ ФНКЦ ФХМ имени Ю. М. Лопухина Л. О. Скородумову за ценные предложения, замечания и поддержку.

Вклад авторов: Е. И. Шарова — идея и подбор данных; Е. И. Шарова, Л. Н. Юльметова — планирование и выбор методов; Н. А. Кулемин — финансирование и администрирование проекта; Л. Н. Юльметова — разработка и вычисления; Е. И. Шарова, Л. Н. Юльметова, Н. А. Кулемин — обсуждение результатов, написание и редактирование статьи.

Соблюдение этических стандартов: исследование проведено с соблюдением принципов Хельсинкской декларации, с использованием наборов данных проектов phs000421.v1.p1 и phs000001.v3.p1., доступ к которым одобрен и получен через dbGaP в соответствии с их политикой одобрения и доступа к конкретным сетам данных.

✉ **Для корреспонденции:** Елена Ивановна Шарова
ул. Малая Пироговская, д. 1с3, г. Москва, 119435, Россия; sharova78@gmail.com

Статья получена: 12.12.2022 **Статья принята к печати:** 20.01.2023 **Опубликована онлайн:** 12.02.2023

DOI: 10.24075/vrgmu.2023.001

Finding a biological basis for the inheritability of phenotypes is one of the main tasks of modern medical genetics. Generally, approaches aimed at the detection of pathogenic genomic variants can be divided into two categories: biological and mathematical. Biological methods include the approaches that explain phenotypes based on the studied biochemical processes. When it is impossible to directly trace the biochemical pathway underlying phenotype formation, but the disease shows a familial tendency, various statistical approaches are applied: genome-wide association studies (GWAS) [1], polygenic risk score (PRS) [2], haplotype identification approaches [3], and other methods. Basic GWAS methodology performs single nucleotide polymorphism (SNP) association testing to identify SNP loci exceeding a genome-wide significance *p*-value threshold. Thus, the GWAS results for any disorder representing a combination of rare inherited mutations could be inaccurate, since the number of rare polymorphisms don't meet the significance criteria. The PRS approach might be considered as an extension of GWAS, however, it also evaluates the effect of each SNP independently. For some disorders the genetic basis can't be explained by biological or popular statistical methods. The inheritance of such phenotypes is based on the haplotype architectures. We define a haplotype as a linear combination of a certain number (up to several hundred) of the linked variable variants that together forma small number (less than 100, 10–20 on average) of allele variants. The approach involving identification of specific haplotype variants is actively used in pharmacogenetics for analysis of P450 cytochromes. For example, there are more than 120 haplotype variants for *CYP2D6* resulting from more than 500 polymorphisms [4]. However, this approach is extremely rarely discussed with reference to the majority of loci of polygenic diseases.

GWAS is commonly applied to the nervous system disorders, polygenic developmental disorders and neurodegenerative diseases, such as amyotrophic lateral sclerosis, Parkinson's disease, schizophrenia, autism spectrum disorders. GWAS method allows to identify genome regions, the alterations of which are overrepresented in affected individuals relative to the general (control) population. GWAS also handles the structural variations that can't be detected directly by the chip SNPs but are in linkage disequilibrium with those ones.In particular, the amyotrophic lateral sclerosis GWAS detects the *C9orf72* gene locus comprising the G4C2 expanded six-nucleotide repeat (GGGGCC) [5], however, the repeat variants are not detected directly with the chip. The Huntington's disease GWAS reveals the chromosome 15 *HTT* gene locus comprising trinucleotide repeats [6], but there are no probes matching the repeat region in the chip.

Fuchs Endothelial Corneal Dystrophy (FECD) is a hereditary eye disease characterized by a decrease in the number of corneal endothelial cells that maintain the corneal stroma water balance. FECD is a polygenic disease that is of considerable interest for genetic research [7]. There are two FECD forms: early onset and late onset FECD. These forms have different genetic bases. Early onset FECD is diagnosed at the age below 50 and is a very rare disorder associated with the *COL8A2* gene pathogenic variants [8]. The late onset FECD manifests at the age greater than 50 and it is the most common form of FECD. It was shown that late onset FECD is associated with the intronic CTG18.1 trinucleotide repeat expansion in *TCF4* [9]. According to our data and the data provided by foreign authors, the CTG18.1 intronic trinucleotide repeat expansion in *TCF4* is the most common FECD-associated variant among Caucasoid populations. The expansion of at least one allele of CTG18.1 trinucleotide repeat was detected in approximately two thirds of the FECD patients in European descent cohorts. Later Afshari et al. [10] made an attempt to find other variants associated with FECD in a bigger cohort also using GWAS. They confirmed the association of the *TCF4* locus and identified three new loci in the genes KANK4, LAMC1 genes and near the ATP1B1 gene, however, their independence from trinucleotide repeat expansion was not tested [10]. The role of mutations in *ZEB1* [11], *SLC4A11* [12], *AGBL1* [13], and *LOXHD1* [14] in the development of FECD is also discussed. The question, whether FECD is a set of phenocopies or a polygenic disease, still remains open. The reported asymptomatic carriers of the repeat expansion [9, 15] and the disputable nature of the clear monogenic link of FECD to some other genes suggest that late onset FECD is a set of polygenic phenocopies. This makes it similar to other late onset repeat expansion diseases.

Thus it leads to the question if it's possible to split the patients into groups within the loci using GWAS results and what accuracy can be achieved. And is it possible to stratify late onset FECD patients by expansion/no expansion based on the microchip-based data? Are the haplotype stratification results and simple patient grouping based on the minor allele of SNPs comparable for these purposes? The study was aimed to develop and test the approach of dividing patients into groups based on the chip-based genotyping and genome-wide association study (GWAS) results.

METHODS

The analysis was carried on dbGaP datasets corresponding to two studies: the FECD Genetics Multi-center Study [16] and the Age-Related Eye Disease Study (AREDS, Refractive Error Substudy) [17–18]. All samples were genotyped on Illumina HumanOmni2.5-4v1 arrays. Clinical manifestations of the disease were classified using a modified Krachmer grading scale based on the slit lamp biomicroscopy data [19].

Both sample-level and variant-level quality control (QC) was performed. The genotyping data were preprocessed using the PLINK 1.9 software [20], GRAF 2.4 [21–22], and code written in R version 4.1.0.

First genotypes with GenCall (GC) scores below 0.3 were removed. Subsequent QC selected markers met the following criteria: missing genotype rate < 10%, minor allele frequency > 1%, number of Mendel errors, a Hardy-Weinberg Equilibrium *p*-value $> 1 \times 10^{-10}$ for control samples and p-value $> 1 \times 10^{-15}$ for FECD patients. Duplicate markers, i.e. markers with different IDs but identical genetic positions and allele coding, were detected and analyzed separately. Both markers of each pair of duplicates were excluded from consideration. One marker with the lowest missing genotype rate was excluded from each pair of duplicates showing 10 differences or more. A total of 1,580,746 SNP markers were included in the analysis after applying all the filters.

The following inclusion criteria were defined for the group of FECD patients: age 47 or older; keratoplasty in at least one eye or grade 2 or above disease (according to the modified Krachmer grading scale) in at least one eye.

Inclusion criteria for the control group: age 60 or older; normal cornea with no epithelial, endothelial, or stromal abnormalities except corneal injuries.

Exclusion criteria: samples with Mendel errors, samples with mismatch between annotated and genetic sex (determined based on the X chromosome heterozygosity rate and Y chromosome genotype counts); samples with genotype missingness above 5%; relatives up to the second degree of relationship (according to GRAF-rel).

Table 1. Characteristics of the study paticipants

	Patients with FECD	Control samples	
Sample	According to Afshari et al, 2017	According to Afshari et al, 2017	According to ARED
Number of participants	1287	2373	
		562	1811
Males	408	989	
		245	744
Females	879	1384	
		317	1067
Median age	71	72	68
		69	

The population structure was estimated using GRAF-pop in order to obtain a genetically homogenous sample. The samples identified as outliers in the genetic distance coordinates were filtered out. The patients were divided into groups according to the potential carrier state of repeat expansion in three stages:

Stage 1: selection of significant variants;

Stage 2: clustering the study participants based on the haplotypes/combinations of the selected variants, calculation of the repeat expansion rate;

Stage 3: evaluation of the concordance between the obtained repeat expansion rate and the percentage of the repeat expansion carriage according to the experimental data reported in previous studies. The repeat allele was considered as expanded if the number of the repeats was ≥ 40 and as unexpanded if the number of the repeats was < 40.

In the first stage, variants were tested for association with FECD using logistic regression with sex and the first six principal components as covariates. *p*-values were adjusted for multiple testing using the Benjamini–Hochberg method. The chromosome 18 (carrying the locus with the repeats) variants were first filtered by *p*-value $< 1 \times 10^{-15}$. For comparison with the haplotype-based approach, three SNPs showing the lowest *p*-values in the resulting set of variants were considered as the potential markers of the increased number of repeats. Additionally marker pruning based on LD ($r^2 > 0.6$) was performed. The genotype matrix was encoded according to the dominant inheritance model.

In the second stage, we used the assumption that the patients with the repeat expansion in the *TCF4* gene carried the certain combinations of SNPs. We expected that the FECD samples would cluster within the *TCF4* locus based on the haplotypes and the combinations of individual variants. However, individuals with phenocopy due to expansion would fall into common clusters based on the similarity of the combinations of minor variants. Asymptomatic control repeat carriers from the control sample (2–10%) and a fraction of the control sample carrying minor haplotypes with no repeats would fall into the same clusters. Furthermore, the combinations of major variants and haplotypes showing predominance of major alleles would form clusters mostly of the control sample representatives. However, these clusters would also include some FECD patients with phenocopy and some patients with the expansion no longer linked to minor haplotypes (in 7% of FECD patients, linked haplotypes and repeats sometimes break apart, which has been earlier demonstrated for the rs613872 variant [23]). That is why the percentage of FECD patients and subjects with no FECD can be used as a surrogate marker of the carriage of the repeats in specific clusters.

Agglomerative hierarchical clustering implemented in the hclust function of the stats R package was used for clustering. The algorithm arranges the data into a tree representation by merging the pairs of clusters with the minimum distance into a new cluster. The algorithm takes the matrix of pairwise distances between the points (samples) as input; initially each point represents a distinct cluster. Since the haplotypes are not identical, we expected that there would be more than two clusters, while the optimal number of clusters was defined by the Silhouette metrics.

For each cluster we calculated the percentage of patients and controls. Clusters with a patient predominance we considered associated with FECD. For the selected three SNPs, carriage of the minor allele was considered as a marker of the repeat expansion carriage.

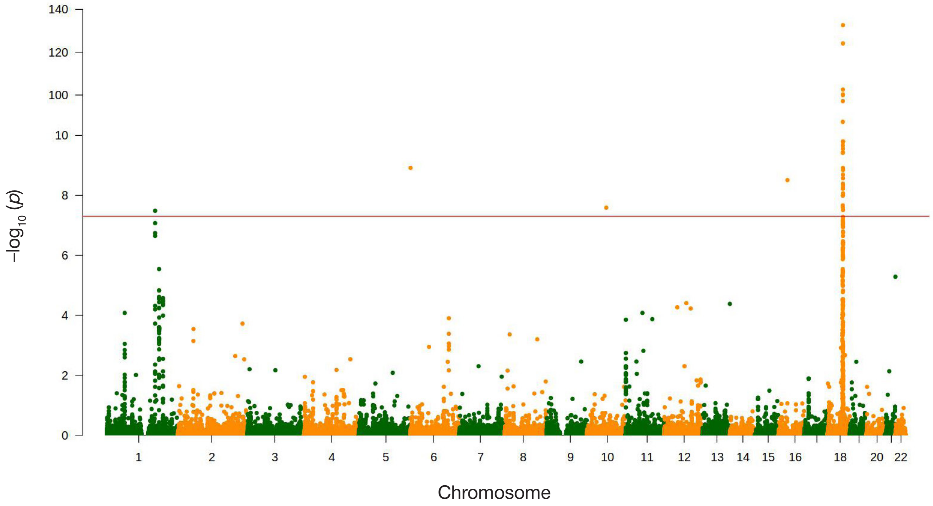

Fig. 1. Genome-wide association study results presented through a Manhattan plot. Points represent the assessed variants. Their positions on the *x* axis are determined by positions of the variants in the genome, while positions on the *y* axis represent the degree of the disease association ($-\log_{10}$ of the *p*-value)

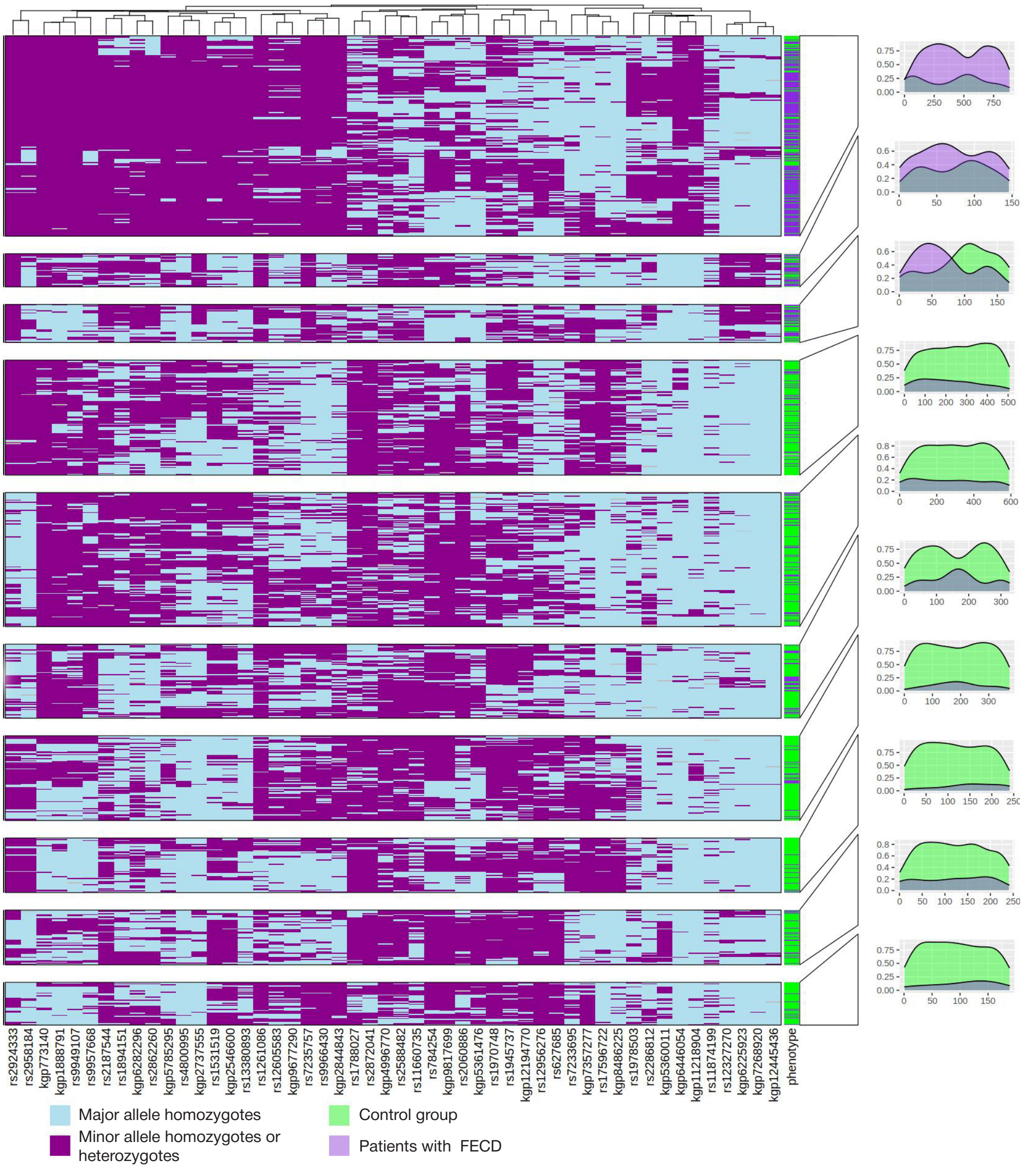

Fig. 2. Heatmap visualization of clustering output. Columns represent variants patients, and rows represent patients various genomic variants. Annotations on the right show the distribution of the FECD samples and the control samples within each cluster

To evaluate the resulting partition, we calculated the odds ratio from the estimates of the expansion status in each group.

The last step was to compare the results obtained by the proposed approach with the experimental data reported in the previous studies. We selected studies based on the following conditions:

1. The number of repeats in *TCF4* was determined by means of fragment analysis or triplet repeat primed PCR.
2. The study participants were individuals of European ancestry.
3. The sample size was at least 50 people for each comparison group.

RESULTS

After the quality control procedure, the discovery dataset consists of 3,660 samples of European ancestry (Table 1) and 1,580,746 SNP markers.

Since GWAS was performed on the same cohort of samples that were studied in the Afshari et al [10], its results (Fig. 1) are comparable to those described in the article. The genomic inflation factor was 1.05, which indicated slight population stratification.

For further analysis, only the locus of chromosome 18 was considered. Filtering by *p*-values resulted in 134 SNPs,

Table 2. Distribution of probable repeat expansion carriers across the comparison groups

Marker of increased repeat count	Patients with FECD			Control sample		
	Predicted repeat expansion	Predicted no repeat expansion	Predicted repeat expansion carriers, %	Predicted repeat expansion	Predicted no repeat expansion	Predicted repeat expansion carriers, %
Group of clusters	764	523	59.4	264	2109	11.1
rs784257	1046	237	81.5	765	1602	32.3
rs72932578	698	583	54.5	286	2082	12.1
rs618869	852	431	66.4	588	1780	24.8

Table 3. Results of experimental studies of the percentage of people with the repeat expansion in *TCF4* among patients with FECD and control samples of European ancestry

	Country	FECD		Total	
		Total	Repeat expansion	Total	Repeat expansion
Skorodumova et al, 2018 [15]	Russia	100	72	100	5
Viberg et al., 2022 [24]	Sweden	85	76	102	4
Foja et al., 2017 [25]	Germany	61	47	113	13
Kuot et al., 2017 [26]	Australia	189	107	183	9

three of which with the lowest *p*-values were rs784257, rs72932578, and rs618869 (and according to gnomAD v3.1.2, the frequencies of the C, T, and C minor alleles in the European population are 0.17932, 0.05649, and 0.13451, respectively). These variants were further tested in terms of dividing patients into groups.

The haplotype block size was 50 variants left after pruning. After clustering the samples were divided into 10 subgroups (Fig. 2).

Clustering has shown that clusters with a predominance of control group participants are homogeneous in terms of their representation. However, three clusters with the potentially increased repeat counts (in which FECD patients prevail) are heterogeneous in terms of haplotypes. This is reflected by the uneven distribution of patients with various phenotypes within each cluster. This may be due to both asymptomatic carriers of the increased repeat number in this locus and the resolution of the population-variable chip SNPs that is not enough for accurate division of samples based on the repeats of varying length.

Our analysis has shown that that the proportion of people from clusters with a presumptive carriage of expanded repeats in the group of samples with FECD is significantly higher than in the control group (Table 2). Furthermore, the calculated rate of probable repeat expansion carriers varies significantly depending on the selected method (prediction of expansion based on the haplotypes/combinations of variants or based on the genotypes of certain variants with low *p*-values).

To verify the results obtained we have selected the studies involving experimental determination of the repeat expansion. The number of repeats is routinely defined by conventional fragment analysis or triplet repeat primed PCR with subsequent fragment analysis. A total of five papers with appropriate samples have been found (Table 3).

To compare the predicted and reported frequencies of the expansion carriers we have merged the samples from the papers. Comparative analysis has shown that markers reproduce the frequency of the expansion carriers in the comparison groups to a different extent (Fig. 3).

None of the applied approaches represent the repeat frequency in the group of patients and the control group accurately enough compared to the results of direct typing reported in the papers (Table 4). However, the haplotype-based approach outperformed the SNPs in terms of odds ratio by covering the 95% confidence intervals of the samples used in two studies.

It is interesting to note that the individual variants we have considered produce extremely discordant results (Fig. 4), i.e. quite different people are carriers of minor allele in these variants, which makes the applied metrics volatile. rs784257 differs most from the haplotype-based approach in terms of the allele carrier state, it is also the most significant variant according to the GWAS results. At the same time, it shows the maximum discrepancy in the proportions of potential expansion carriers in the control group and no better correspondence with the FECD group. This allele is most likely to show weaker linkage to the repeat carrier state than the other two alleles.

DISCUSSION

Molecular genetic stratification of patients with polygenic diseases is a useful tool for studying the disease genetics. Furthermore, there could be patients with the groups of causal variants linked to various haplotypes within one phenotype.

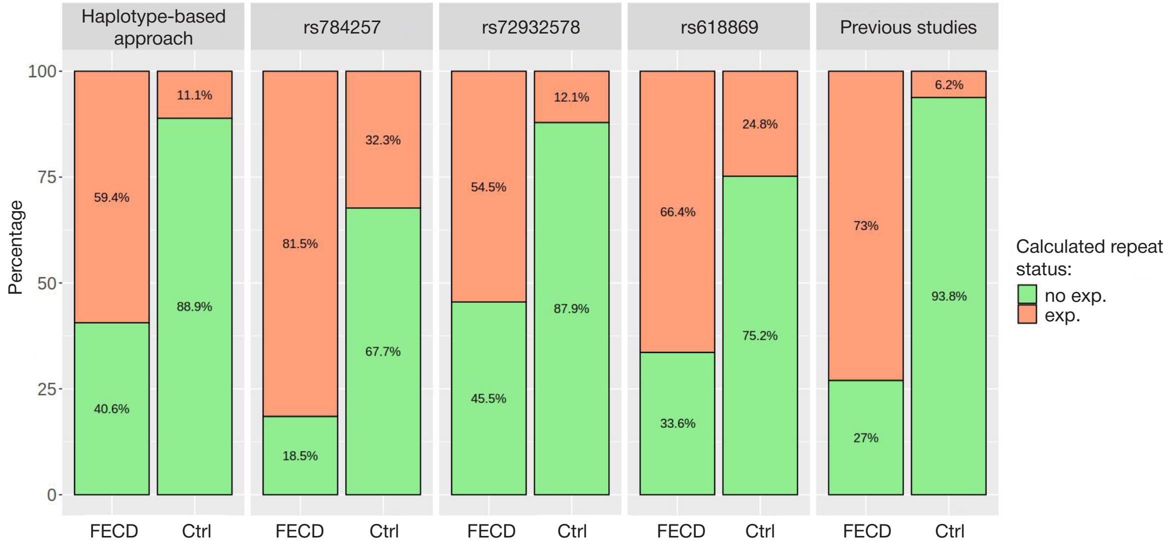

Fig. 3. Percentage of people with the repeat expansion and no repeat expansion in the *TCF4* gene intron based on the data of our study (haplotype-based approach, carriers of minor alleles of the rs784257, rs72932578, rs618869 variants) and other studies. FECD — individuals with Fuchs' endothelial corneal dystrophy, Ctrl — control group

Table 4. Odds ratio of finding the expansion in the group of patients with FECD compared to the control group

	Odds ratio	95% confidence interval
Haplotype-based approach (h-clust)	11.67	9.85–13.83
rs784257	9.24	7.83–10.90
rs72932578	8.72	7.39–10.28
rs618869	5.98	5.16–6.94
Skorodumova et al, 2018 [15]	48.86	17.98–132.76
Viberg et al., 2022 [24]	206.89	61.37–697.50
Foja et al., 2017 [25]	25.82	11.25–59.26
Kuot et al., 2017 [26]	25.23	12.17–52.31

Despite the fact that the gene is definitely associated with the disease, *p*-values of the variants would be higher due to the large number of the groups of linked variants, i.e. the variants that are significant for every group do not surpass the generally accepted significance threshold (*p*-value $< 5 \times 10^{-8}$) due to the features of the disease genetic basis. Moreover, many loci, the significance of which is close to the generally accepted threshold, are characterized by the marked sparseness of the significant variants, under which only a few variants are strongly associated with the disease. Thus, it is impossible to choose between genomic variants as the population outliers (the significance of which results from random population frequency shift) or assigning these variants to the potentially significant group of variants. That is why the genetic data structuring methods that involve assessing interactions between both variants within haplotype blocks and haplotype blocks are a promising tool for the disease genetic basis clarification.

GWAS makes it possible to obtain more information about the disease genetic basis than exclusion of variants based on p-values and loci formation with reference to the nearest gene that represents the transition from the "variant" level to the "gene" level. However, questions remain about unequal contribution of various loci to the genetic basis of the disease in specific groups of people with the same phenotype. This is due, among other things, to the lack of advanced approaches to formation of the combinations of variants, i.e. to working at the intermediate level between the "variant" level and "gene" level. Since the variants show incomplete linkage, it would be reasonable to consider the sets of haplotypes/combinations of variants that define the differentiated disease risk instead of the specific risk haplotypes or protective haplotypes. This means that the variant with the highest population attributable risk (combination of allele frequency and relative risk) is likely to be the most significant one in the locus.

Assessment of the groups of haplotypes linked to the causal variants is still a challenging task, however, it more and more often outperforms GWAS, even despite the lack of the high throughput standard approaches. The GWAS performed in 2005 showed that the *CFH* gene was associated with age-related macular degeneration [27]. Later it was reported that this association was not confined to individual variants and was also observed in the groups of patients with structural alterations, such as partial deletions of the *CFHR1-5* genes [27]. Furthermore, it was found that most of the variation attributed to individual variants was in fact the marker of haplotypes showing large-scale structural alterations in this region. And these are haplotype variants of the locus structure, including those with different population abundance, that show much stronger correlation with the risk of retinal degeneration than the majority of individual variants in this locus [28].

In this study we have implemented sample clustering by variants of the region containing the expansion based on the data on the association of individual variants with the repeats [14, 23, 29], particularly, allele G of the rs613872 variant, and haplotype blocks [29]. After clustering the samples of the group of patients with FECD turned out to be distributed unevenly across the clusters, which was indirect evidence of clustering by haplotypes linked to the expansion. All clusters except one (cluster 3) had a clear status of the repeats. Uncertainty in defining the status was due to parity between patients and controls in the cluster. In the future we have to decide what to do with such clusters: re-cluster people in these clusters in the case-by-case manner or leave them with uncertain status. It is also necessary to select another clustering metrics. This requires additional data that include both sample genotyping results and information about the repeat length. Regardless of these limitations, the results obtained using the haplotype-based approach were better than the results shown by

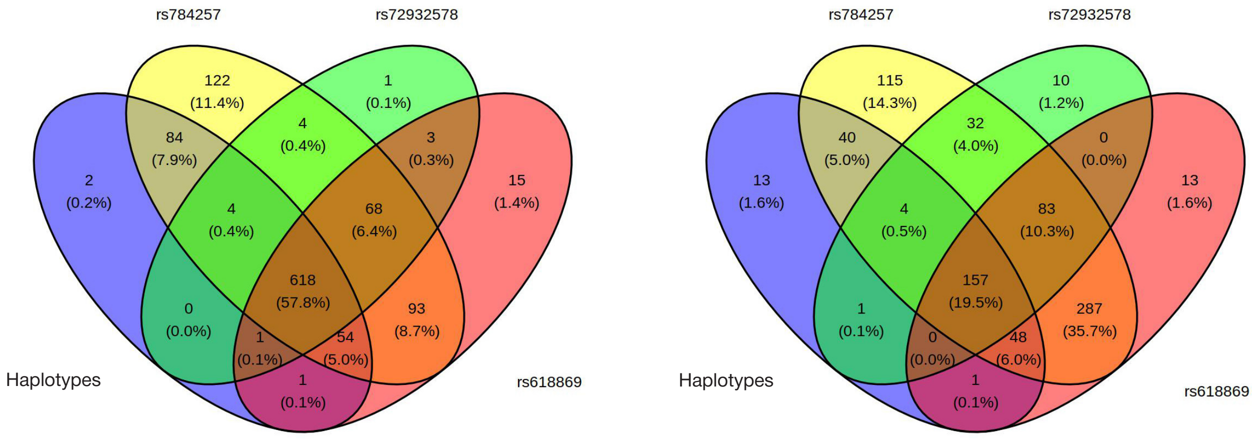

Fig. 4. Consistency of the repeat status determination results obtained by different approaches: based on the genotypes of the rs784257, rs72932578, rs618869 variants and haplotype-based. Left — for the repeat expansion carriers in the FECD group, right — for the repeat expansion carriers in the control group

individual variants. However, these results turned out to be not precise enough to consider our method optimal.

This work accomplished two goals.

1. Initial testing of an approach that allows stratification of patients and control groups at the intermediate level (not the level of a single variant and not the level of the gene closest to the locus) without first understanding the haplotype structure of the locus. The proportion of patients with FECD and control samples in clusters is used as a measure, allowing this approach to be used for diseases in which the approximate proportion of individuals with a phenotype closely related to or due to changes in a given locus is not known in advance.

2. Obtaining a subsample of patients with FECD and no expansion carrier state for precision re-analysis of GWAS in order to refine genetic structure in this particular category of patients.

In the future, patient clustering will make it possible not only to allocate groups within the phenotype showing a strong contribution from distinct genetic variants, including structural variants, but also to propose the basis and approaches to predicting the patients' responses to various types of therapy.

CONCLUSIONS

The study has shown the possibility of using the haplotype-based approach for genetic stratification of patients based on the cause of the genetic disorder, namely the presence of the repeat expansion. The findings have made it possible to draw the following conclusions: 1) the haplotype-based approach is better suited for detection of the association of loci with certain groups of patients than individual variants; 2) for a more accurate picture we should reconsider the approach to defining the haplotype composition and modeling the data matrix for clustering. In particular, it is planned to analyze some methods of computing the genetic similarities (genetic distances (genetic distances) among samples and apply more specific methods for initial selection of variants; 3) the results obtained show that clustering splits the patients with FECD and the control group based on the groups of haplotypes/combinations of variants associated with the repeat expansion. Further testing of the approach requires additional evidence base that demands the use of more validation data.

References

1. Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021; 1 (1): 59. Available from: https://doi.org/10.1038/s43586-021-00056-9.
2. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome medicine. 2020; 12 (1): 1–1. Available from: https://doi.org/10.1186/s13073-020-00742-5.
3. Stram DO. Multi-SNP haplotype analysis methods for association analysis. Statistical Human Genetics: Methods and Protocols. 2017: 485–504. Available from: https://doi.org/10.1007/978-1-4939-7274-6_24.
4. Twesigomwe D, Wright GE, Drögemöller BI, da Rocha J, Lombard Z, Hazelhurst S. A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: a focus on CYP2D6 genotyping. NPJ genomic medicine. 2020; 5 (1): 30. Available from: https://doi.org/10.1038/s41525-020-0135-2.
5. Van Rheenen W, Van Der Spek RA, Bakker MK, Van Vugt JJ, Hop PJ, Zwamborn RA, et al. Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. Nature genetics. 2021; 53 (12): 1636–48.
6. Lee JM, Wheeler VC, Chao MJ, Vonsattel JP, Pinto RM, Lucente D, et al. Identification of genetic factors that modify clinical onset of Huntington's disease. Cell. 2015; 162 (3): 516–26.
7. Fautsch MP, Wieben ED, Baratz KH, Bhattacharyya N, Sadan AN, Hafford-Tear NJ. et al. TCF4-mediated Fuchs endothelial corneal dystrophy: Insights into a common trinucleotide repeat-associated disease. Progress in retinal and eye research. 2021; 81: 100883.
8. Biswas S, Munier FL, Yardley J, Hart-Holden N, Perveen R, Cousin P. et al. Missense mutations in COL8A2, the gene encoding the α2 chain of type VIII collagen, cause two forms of corneal endothelial dystrophy. Human molecular genetics. 2001; 10 (21): 2415–23.
9. Wieben ED, Aleff RA, Tosakulwong N, Butz ML, Highsmith WE, Edwards AO, et al. A common trinucleotide repeat expansion within the transcription factor 4 (TCF4, E2-2) gene predicts Fuchs corneal dystrophy. PLoS One. 2012; 7 (11): e49083.
10. Afshari NA, Igo Jr RP, Morris NJ, Stambolian D, Sharma S, Pulagam VL, et al. Genome-wide association study identifies three novel loci in Fuchs endothelial corneal dystrophy. Nature communications. 2017; 8 (1): 14898.
11. Chung DW, Frausto RF, Ann LB, Jang MS, Aldave AJ. Functional impact of ZEB1 mutations associated with posterior polymorphous and Fuchs' endothelial corneal dystrophies. Investigative Ophthalmology & Visual Science. 2014; 55 (10): 6159–66.
12. Chaurasia S, Ramappa M, Annapurna M, Kannabiran C. Coexistence of congenital hereditary endothelial dystrophy and Fuchs endothelial corneal dystrophy associated with SLC4A11 mutations in affected families. Cornea. 2020; 39 (3): 354–7.
13. Riazuddin SA, Vasanth S, Katsanis N, Gottsch JD. Mutations in AGBL1 cause dominant late-onset Fuchs corneal dystrophy and alter protein-protein interaction with TCF4. The American Journal of Human Genetics. 2013; 93 (4): 758–64.
14. Riazuddin SA, Parker DS, McGlumphy EJ, Oh EC, Iliff BW, Schmedt T, et al. Mutations in LOXHD1, a recessive-deafness locus, cause dominant late-onset Fuchs corneal dystrophy. The American Journal of Human Genetics. 2012; 90 (3): 533–9.
15. Skorodumova LO, Belodedova AV, Antonova OP, Sharova EI, Akopian TA, Selezneva OV, et al. CTG18. 1 expansion is the best classifier of late-onset Fuchs' corneal dystrophy among 10 biomarkers in a cohort from the European part of Russia. Investigative Ophthalmology & Visual Science. 2018; 59 (11): 4748–54.
16. Louttit MD, Kopplin LJ, Igo Jr RP, Fondran JR, Tagliaferri A, Bardenstein D, et al. A multi-center study to map genes for Fuchs' endothelial corneal dystrophy: baseline characteristics and heritability. Cornea. 2012; 31 (1): 26.
17. Age-Related Eye Disease Study Research Group. The age-related eye disease study (AREDS): design implications AREDS report no. 1. Controlled clinical trials. 1999; 20 (6): 573.
18. Stambolian D, Wojciechowski R, Oexle K, Pirastu M, Li X, Raffel LJ, et al. Meta-analysis of genome-wide association studies in five cohorts reveals common variants in RBFOX1, a regulator of tissue-specific splicing, associated with refractive error. Human molecular genetics. 2013; 22 (13): 2754–64.
19. Krachmer JH, Purcell JJ Jr, Young CW, Bucher KD. Corneal endothelial dystrophy. A study of 64 families. Arch Ophthalmol. 1978; 96 (11): 2036–9. DOI: 10.1001/archopht.1978.03910060424004.
20. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4: 7. DOI: 10.1186/s13742-015-0047-8.
21. Jin Y, Schäffer AA, Sherry ST, Feolo M. Quickly identifying identical and closely related subjects in large databases using genotype data. PLoS One. 2017; 12 (6): e0179106. DOI: 10.1371/journal.pone.0179106.

22. Jin Y, Schaffer AA, Feolo M, Holmes JB, Kattman BL. GRAF-pop: A Fast Distance-Based Method To Infer Subject Ancestry from Multiple Genotype Datasets Without Principal Components Analysis. G3 (Bethesda). 2019; 9 (8): 2447–61. DOI: 10.1534/g3.118.200925.
23. Okumura N, Hayashi R, Nakano M, Tashiro K, Yoshii K, Aleff R. et al. Association of rs613872 and Trinucleotide Repeat Expansion in the TCF4 Gene of German Patients With Fuchs Endothelial Corneal Dystrophy. Cornea. 2019; 38 (7): 799–805. DOI: 10.1097/ICO.0000000000001952.
24. Viberg A, Westin IM, Golovleva I, Byström B. TCF4 trinucleotide repeat expansion in Swedish cases with Fuchs' endothelial corneal dystrophy. Acta Ophthalmol. 2022; 100 (5): 541–8. DOI: 10.1111/aos.15032. Epub 2021 Oct 13.
25. Foja S, Luther M, Hoffmann K, Rupprecht A, Gruenauer-Kloevekorn C. CTG18.1 repeat expansion may reduce TCF4 gene expression in corneal endothelial cells of German patients with Fuchs' dystrophy. Graefes Arch Clin Exp Ophthalmol. 2017; 255 (8): 1621–31. DOI: 10.1007/s00417-017-3697-7. Epub 2017 Jun 12.
26. Kuot A, Hewitt AW, Snibson GR, Souzeau E, Mills R, Craig JE, et al. TGC repeat expansion in the TCF4 gene increases the risk of Fuchs' endothelial corneal dystrophy in Australian cases. PLoS One. 2017; 12 (8): e0183719. DOI: 10.1371/journal.pone.0183719.
27. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005; 308 (5720): 385–9. DOI: 10.1126/science.1109557. Epub 2005 Mar 10.
28. Spencer KL, Hauser MA, Olson LM, Schmidt S, Scott WK, Gallins P, et al. Deletion of CFHR3 and CFHR1 genes in age-related macular degeneration. Hum Mol Genet. 2008; 17 (7): 971–7. DOI: 10.1093/hmg/ddm369. Epub 2007 Dec 15.
29. Baratz KH, Tosakulwong N, Ryu E, Brown WL, Branham K, Chen W, et al. E2-2 protein and Fuchs's corneal dystrophy. N Engl J Med. 2010; 363 (11): 1016–24. DOI: 10.1056/NEJMoa1007064. Epub 2010 Aug 25.

Литература

1. Uffelmann E, Huang QQ, Munung NS, De Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021; 1 (1): 59. Available from: https://doi.org/10.1038/s43586-021-00056-9.
2. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome medicine. 2020; 12 (1): 1–1. Available from: https://doi.org/10.1186/s13073-020-00742-5.
3. Stram DO. Multi-SNP haplotype analysis methods for association analysis. Statistical Human Genetics: Methods and Protocols. 2017: 485–504. Available from: https://doi.org/10.1007/978-1-4939-7274-6_24.
4. Twesigomwe D, Wright GE, Drögemöller BI, da Rocha J, Lombard Z, Hazelhurst S. A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: a focus on CYP2D6 genotyping. NPJ genomic medicine. 2020; 5 (1): 30. Available from: https://doi.org/10.1038/s41525-020-0135-2.
5. Van Rheenen W, Van Der Spek RA, Bakker MK, Van Vugt JJ, Hop PJ, Zwamborn RA, et al. Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. Nature genetics. 2021; 53 (12): 1636–48.
6. Lee JM, Wheeler VC, Chao MJ, Vonsattel JP, Pinto RM, Lucente D, et al. Identification of genetic factors that modify clinical onset of Huntington's disease. Cell. 2015; 162 (3): 516–26.
7. Fautsch MP, Wieben ED, Baratz KH, Bhattacharyya N, Sadan AN, Hafford-Tear NJ. et al. TCF4-mediated Fuchs endothelial corneal dystrophy: Insights into a common trinucleotide repeat-associated disease. Progress in retinal and eye research. 2021; 81: 100883.
8. Biswas S, Munier FL, Yardley J, Hart-Holden N, Perveen R, Cousin P. et al. Missense mutations in COL8A2, the gene encoding the α2 chain of type VIII collagen, cause two forms of corneal endothelial dystrophy. Human molecular genetics. 2001; 10 (21): 2415–23.
9. Wieben ED, Aleff RA, Tosakulwong N, Butz ML, Highsmith WE, Edwards AO, et al. A common trinucleotide repeat expansion within the transcription factor 4 (TCF4, E2-2) gene predicts Fuchs corneal dystrophy. PLoS One. 2012; 7 (11): e49083.
10. Afshari NA, Igo Jr RP, Morris NJ, Stambolian D, Sharma S, Pulagam VL, et al. Genome-wide association study identifies three novel loci in Fuchs endothelial corneal dystrophy. Nature communications. 2017; 8 (1): 14898.
11. Chung DW, Frausto RF, Ann LB, Jang MS, Aldave AJ. Functional impact of ZEB1 mutations associated with posterior polymorphous and Fuchs' endothelial corneal dystrophies. Investigative Ophthalmology & Visual Science. 2014; 55 (10): 6159–66.
12. Chaurasia S, Ramappa M, Annapurna M, Kannabiran C. Coexistence of congenital hereditary endothelial dystrophy and Fuchs endothelial corneal dystrophy associated with SLC4A11 mutations in affected families. Cornea. 2020; 39 (3): 354–7.
13. Riazuddin SA, Vasanth S, Katsanis N, Gottsch JD. Mutations in AGBL1 cause dominant late-onset Fuchs corneal dystrophy and alter protein-protein interaction with TCF4. The American Journal of Human Genetics. 2013; 93 (4): 758–64.
14. Riazuddin SA, Parker DS, McGlumphy EJ, Oh EC, Iliff BW, Schmedt T, et al. Mutations in LOXHD1, a recessive-deafness locus, cause dominant late-onset Fuchs corneal dystrophy. The American Journal of Human Genetics. 2012; 90 (3): 533–9.
15. Skorodumova LO, Belodedova AV, Antonova OP, Sharova EI, Akopian TA, Selezneva OV, et al. CTG18. 1 expansion is the best classifier of late-onset Fuchs' corneal dystrophy among 10 biomarkers in a cohort from the European part of Russia. Investigative Ophthalmology & Visual Science. 2018; 59 (11): 4748–54.
16. Louttit MD, Kopplin LJ, Igo Jr RP, Fondran JR, Tagliaferri A, Bardenstein D, et al. A multi-center study to map genes for Fuchs' endothelial corneal dystrophy: baseline characteristics and heritability. Cornea. 2012; 31 (1): 26.
17. Age-Related Eye Disease Study Research Group. The age-related eye disease study (AREDS): design implications AREDS report no. 1. Controlled clinical trials. 1999; 20 (6): 573.
18. Stambolian D, Wojciechowski R, Oexle K, Pirastu M, Li X, Raffel LJ, et al. Meta-analysis of genome-wide association studies in five cohorts reveals common variants in RBFOX1, a regulator of tissue-specific splicing, associated with refractive error. Human molecular genetics. 2013; 22 (13): 2754–64.
19. Krachmer JH, Purcell JJ Jr, Young CW, Bucher KD. Corneal endothelial dystrophy. A study of 64 families. Arch Ophthalmol. 1978; 96 (11): 2036–9. DOI: 10.1001/archopht.1978.03910060424004.
20. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4: 7. DOI: 10.1186/s13742-015-0047-8.
21. Jin Y, Schäffer AA, Sherry ST, Feolo M. Quickly identifying identical and closely related subjects in large databases using genotype data. PLoS One. 2017; 12 (6): e0179106. DOI: 10.1371/journal.pone.0179106.
22. Jin Y, Schaffer AA, Feolo M, Holmes JB, Kattman BL. GRAF-pop: A Fast Distance-Based Method To Infer Subject Ancestry from Multiple Genotype Datasets Without Principal Components Analysis. G3 (Bethesda). 2019; 9 (8): 2447–61. DOI: 10.1534/g3.118.200925.
23. Okumura N, Hayashi R, Nakano M, Tashiro K, Yoshii K, Aleff R. et al. Association of rs613872 and Trinucleotide Repeat Expansion in the TCF4 Gene of German Patients With Fuchs Endothelial Corneal Dystrophy. Cornea. 2019; 38 (7): 799–805. DOI: 10.1097/ICO.0000000000001952.
24. Viberg A, Westin IM, Golovleva I, Byström B. TCF4 trinucleotide repeat expansion in Swedish cases with Fuchs' endothelial corneal dystrophy. Acta Ophthalmol. 2022; 100 (5): 541–8. DOI: 10.1111/aos.15032. Epub 2021 Oct 13.
25. Foja S, Luther M, Hoffmann K, Rupprecht A, Gruenauer-

Kloevekorn C. CTG18.1 repeat expansion may reduce TCF4 gene expression in corneal endothelial cells of German patients with Fuchs' dystrophy. Graefes Arch Clin Exp Ophthalmol. 2017; 255 (8): 1621–31. DOI: 10.1007/s00417-017-3697-7. Epub 2017 Jun 12.

26. Kuot A, Hewitt AW, Snibson GR, Souzeau E, Mills R, Craig JE, et al. TGC repeat expansion in the TCF4 gene increases the risk of Fuchs' endothelial corneal dystrophy in Australian cases. PLoS One. 2017; 12 (8): e0183719. DOI: 10.1371/journal.pone.0183719.

27. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005; 308 (5720): 385–9. DOI: 10.1126/science.1109557. Epub 2005 Mar 10.

28. Spencer KL, Hauser MA, Olson LM, Schmidt S, Scott WK, Gallins P, et al. Deletion of CFHR3 and CFHR1 genes in age-related macular degeneration. Hum Mol Genet. 2008; 17 (7): 971–7. DOI: 10.1093/hmg/ddm369. Epub 2007 Dec 15.

29. Baratz KH, Tosakulwong N, Ryu E, Brown WL, Branham K, Chen W, et al. E2-2 protein and Fuchs's corneal dystrophy. N Engl J Med. 2010; 363 (11): 1016–24. DOI: 10.1056/NEJMoa1007064. Epub 2010 Aug 25.