VILNIAUS UNIVERSITETAS

Laura
PRANCKĖNIENĖ

De novo mutacijų intensyvumo įvertinimas bendrosios Lietuvos populiacijos ir asmenų su intelektine negalia egzomuose

DAKTARO DISERTACIJOS SANTRAUKA

Medicinos ir sveikatos mokslai,
Medicina M 001

VILNIUS 2019

Disertacija rengta 2015 – 2019 metais Vilniaus universitete
Mokslinius tyrimus rėmė Lietuvos mokslo taryba

Mokslinis vadovas:
Akad. prof. habil. dr. Vaidutis Kučinskas (Vilniaus universitetas, medicinos ir sveikatos mokslai, medicina – M 001)

Gynimo taryba:

Pirmininkė – **prof. habil. dr. Zita Aušrelė Kučinskienė** (Vilniaus universitetas, medicinos ir sveikatos mokslai, medicina, M 001).
Nariai:
prof. dr. Loreta Cimbalistienė (Vilniaus universitetas, medicinos ir sveikatos mokslai, medicina, M 001);
prof. habil. dr. Limas Kupčinskas (Lietuvos sveikatos mokslų universitetas, medicinos ir sveikatos mokslai, medicina, M 001);
prof. dr. Česlovas Venclovas (Vilniaus universitetas, gamtos mokslai, biologija, N 010);
prof. dr. Alexandra Zhernakova (Groningeno universitetas, Olandija, medicinos ir sveikatos mokslai, medicina, M 001).

Disertacija ginama viešame Gynimo tarybos posėdyje 2019 m. gruodžio mėn. 4 d. 14 val. Vilniaus universiteto ligoninės Santaros klinikų Raudonojoje auditorijoje, Santariškių g. 2, Vilnius, Lietuva.

Disertaciją galima peržiūrėti Vilniaus Universiteto bibliotekoje ir VU interneto svetainėje adresu: *https://www.vu.lt/naujienos/ivykiu-kalendorius*

VILNIUS UNIVERSITY

Laura
PRANCKĖNIENĖ

An Evaluation of the Intensity of *De Novo* Mutations in the Exomes of the General Population of Lithuania and of Individuals with Intellectual Disabilities

SUMMARY OF DOCTORAL DISSERTATION

Medicine and Health Sciences,
Medicine M 001

VILNIUS 2019

This dissertation was written between 2015 and 2019 at Vilnius University. The research was supported by the Research Council of Lithuania.

Academic supervisor:
Acad. Prof. Habil. Dr. Vaidutis Kučinskas (Vilnius University, Medical and Health Sciences, Medicine, M 001)

This doctoral dissertation will be defended in a public meeting of the Dissertation Defence Panel:
Chairman – **Prof. Habil. Dr. Zita Aušrelė Kučinskienė** (Vilnius Univesity, Medical and Health Sciences, Medicine, M 001).
Members:
Prof. Dr. Loreta Cimbalistienė (Vilnius Univesity, Medical and Health Sciences, Medicine, M 001);
Prof. Habil. Dr. Limas Kupčinskas (Lithuanian University of Health Sciences, Medical and Health Sciences, Medicine, M 001).
Prof. Dr. Česlovas Venclovas (Vilnius University, Natural Sciences, Biology, N 010);
Assoc. Prof. Dr. Alexandra Zhernakova (University Medical Centre Groningen, the Netherlands, Medical and Health Sciences, Medicine, M 001).

The dissertation shall be defended at a public meeting of the Dissertation Defence Panel at 14:00 PM on December 4, 2019 in the Red Auditorium of the Vilnius University Hospital Santaros Clinics, 2 Santariškių St., Vilnius, Lithuania.

The text of this dissertation can be accessed through the library of Vilnius University as well as on the website of Vilnius University: www.vu.lt/lt/naujienos/ivykiu-kalendorius.

SUMMARY

INTRODUCTION

The analysis of *de novo* mutations, which is generalized on the Lithuanian population scale and not based on individual case descriptions, is important because the research strategy of trios (the father, mother, and at least one child) allows us to determine the mutation rate of different genome variants as well as the mutation rate of individual genomic regions in the studied Lithuanian and heterogeneous genetic disease model – in the groups of individuals with intellectual disability (ID) and their siblings. The analysis of whole-exome sequencing data in trios allows for a direct identification of different *de novo* mutations unaffected by natural selection that may be associated with both common and rare diseases or adaptive regions of the genome [1–7]. A functional and contextual analysis of all the identified genomic variants make it possible to identify the possible mechanisms of mutation formation and mechanisms of intellectual disabilty pathogenesis in the group of individuals with intellectual disability.

Aim of the Study

To analyze the distribution, mechanisms of formation, and effects of exome *de novo* mutations in the Lithuanian population and in groups of individuals with intellectual disability.

Tasks of the Research

1. To select the optimal algorithm for detecting *de novo* mutations.

2. To identify the *de novo* mutation spectrum specific to the Lithuanian population, a group of individuals with intellectual

disability, and a group of individuals with intellectual disability sibs; to compare it with the composition of the spectrum of Utah from Northern and Western Europe (CEU), Ibadan Yoruba (YRI), Icelandic, Scottish, and South Dakota populations; to identify the potential mechanisms of *de novo* mutation formation.

3. To calculate the intensity rate of *de novo* mutations in the genome and compare it with those identified in Icelandic, Danish, Dutch, Canadian, and American general populations as well as the Autism Spectrum Disease group.

4. To determine the influence of parental age, genomic and epigenetic factors on the intensity and formation of *de novo* mutation emergence.

5. To determine *in silico* the influence of the function of *de novo* mutations on the phenotype of the Lithuanian population and individuals with intellectual disabilities; to identify the potential mechanisms of the pathogenesis of intellectual disability;

6. To evaluate whether *de novo* mutations affect the relative fitness of amino acids, and, if affected, what patterns of relative fitness describe the Lithuanian population.

Statements to be Defended

1. The Lithuanian population is characterized by a distinctive spectrum, intensity, and mechanisms of *de novo* mutation patterns.

2. The *de novo* mutations that occur in the Lithuanian population have an overall relative fitness that does not adversely affect an individual's phenotype and health, and may therefore persist in the genome for many generations.

3. The exomes of individuals with intellectual disabilities have significantly more harmful *de novo* mutations than in the other groups studied, and thus may be the cause of intellectual disability.

MATERIALS AND METHODS

1.1 Samples

Data from two research studies (LITGEN and UNIGENE) were used to fulfill the tasks and purpose of the research.

The data of 147 individuals (49 trios, the first group) participating in the LITGEN project were used to complete the study. This study is part of the LITGEN project, which was approved by the Vilnius Regional Research Ethics Committee 235 No. 158200-05-329-79 on May 3, 2011.

The individuals involved in the project UNIGENE were divided into two groups (the second and the third group). The second group consisted of 37 trios participating in the UNIGENE project, where the proband has an intellectual disability or autism spectrum disease (111 individuals in total). In this paper, intellectual disability was chosen as a model for genetic disease to investigate *de novo* mutations. It is a highly heterogeneous genetic disease the causes and consequences of which are different in each analyzed case, while the number of individuals with ID is constantly increasing [8]. This study consisted of subjects meeting the main selection criterion, that of intellectual disabilities, when the intelligence quotient on the Wecksler scale is <70, or when a developmental delay was observed (in children under 6 years of age).

The third group consisted of 17 trios, the proband being the healthy sibs of a person with intellectual disability (51 individuals in total). During the study, only the individuals whose whole exome had been sequenced were analyzed. This study is part of the UNIGENE project, which was approved by the Vilnius Regional Research Ethics Committee 235 No. 158200-12-563-164 on November 29, 2012.

In accordance with the Declaration of Helsinki, forms of written informed consent were received from all of the study participants. Genomic DNA was extracted from venous blood using either the phenol-chloroform extraction method or the automated DNA extraction platform TECAN Freedom EVO® (Tecan Schweiz AG, Switzerland) based on the paramagnetic particle method.

1.2 Whole Exome Sequencing and Data Analysis

The proband and his healthy parents were assessed using whole exome sequencing (WES). The exomes of the LITGEN project were sequenced using the *SOLiD 5500* sequencing system (75 bp reads). Sequencing data were processed and prepared using *Lifescope* software. Exomes were mapped according to the human reference genome build 19. An average read depth of sequencing was 38.5. BAM-formatted files of the mother, father, and child generated by *Lifescope* were combined using *Samtools* software for each trio.

To analyze the UNIGENE project data set, the exomes were captured using the Agilent SureSelect Human All Exon V5 enrichment kit and multiplex sequenced (6-plex) using the *Illumina HiSeq 2500* platform to reach an about 100-fold coverage on average and were mapped, using the *Burrows-Wheeler Aligner* algorithm, according to the human reference genome build 38. BAM-formatted files of the mother, father, and child, generated by the GATK, were combined using *Samtools* software for each trio.

1.3 Identification and Verification of *De Novo* Mutations

An initial identification of *de novo* mutations was performed using the open-access *VarScan* v.2.4.2 software [9] in accordance with protocol and an alternative commercial *VarSeq*™ software. A potential *de novo* mutation was identified if the child has a genomic variant when neither parent has it in the same genome position. The

results are provided in a generated *.vcf* format data file. *SnpSift v.4* software was used for the initial rejection of false positive results derived from *VarScan* software. The following conservative filtering criteria were applied: 1) a genotype quality of the individual ≥ 50; 2) the number of reads at each site > 20. The same filtering criteria were applied to the *dn*M list obtained by *VarSeq*™ software in the *Trio Workflow* part.

Furthermore, in order to discard the remaining variants that were somatic (only present in a fraction of the sequenced blood cells) with low allele balance or sequencing artefacts, *dn*Ms were filtered by setting a threshold for the observed fraction of the reads in individuals with the alternative allele (the allele balance) for the trios [0.3;0.7]. In addition, all possible identified and filtered *de novo* single nucleotide variants were manually reviewed using *Integrative Genomics Viewer* [10]. In cases where *dn*M was detected significantly more frequently than in all other trios, biological paternity verification using the *AmpFLSTR*™ *Identifiler*™ kit (Applied Biosystems™, USA) was performed according to the manufacturer's recommendations.

Due to the large number of identified *dn*Ms, for the validation of variants by *Sanger* sequencing, 50 *de novo* single nucleotide variants were randomly selected for the first group, and 50 – for the second and third groups. Sanger sequencing was performed using an ABI PRISM 3130xl Genetic Analyzer.

Verificated *dn*M annotation was performed using ANNOVAR (v.20170601 hg19 and v.20170912 hg38 reference genome) software [11]. The annotation tool package consisted of *SIFT*, *PolyPhen*, *LRT*, *MutationTaster*, *MutationAssessor*, *FATHMM*, *PROVEAN*, *CADD*, *GERP++*, *PhyloP*, *SiPhy*, and *COSMIC* algorithm annotation tools. Databases providing information on the gene identifying the *dn*M and the *dbSNP* database were also connected, thus assigning the *rs* code to each genome variant, evaluating the genomic variant pathogenicity in *ClinVar*, and evaluating the *1000 genome* project

and ExAC database genome variant frequencies in different populations.

1.4 Calculating the *De Novo* Mutation Rate

The probability that a calling position was a *dn*M in the trio was calculated independently for each trio. The *de novo* rate per position per generation (PPPG) was calculated as follows:

$$De\ novo\ rate_{PPPG} = \frac{\sum_{i=1}^{f} n_i}{2\sum_{i=1}^{f}\sum_{j=1}^{N} P_j^i(de\ novo)}$$

where *f* is the number of trios and *N* is the number of callable sites that may potentially be identified as *de novo* sites for each trio separately, regardless of the sequencing depth. This number varies depending on the trio. n_i is the number of identified *dn*Ms for trio *i*. The probability $P_j^i(de\ novo)$ (*de novo* single nucleotide) for the called single nucleotide site *j* and trio *i* to be mutated was calculated as follows:

$$P_j^i(dnM) = P_j^i\left(C_{Hetero}\middle|M_{HomR,}T_{HomR}\right) + P_j^i\left(C_{Hetero}\middle|M_{HomA,}T_{HomA}\right)$$

The probability $P_j^i(de\ novo\ indel)$ for the called indel site *j* and trio *i* to be mutated was calculated as:

$$P_j^i(de\ novo\ indel) = P_j^i\left(C_{HomR}\middle|M_{HomA,}F_{HomA}\right)$$

where *C*, *M*, and *F* stand for offspring, mother, and father, respectively, and Hetero, HomR, and HomA denote heterozygous, homozygous for reference, and homozygous for alternative allele, respectively. The probability $P_j^i(de\ novo)$ was calculated with respect to the sequencing coverage. Confidence intervals for rate

estimates were calculated as for binomial proportions. For the estimation of the *dn*M rate and for further calculations, R package (version 3.4.3) was used [12].

1.5 Analysis of Parental Age, Genomic and Epigenetic Factors That May Influence *dn*M Intensity

In order to test the hypothesis that variations in the *dn*M rate across different regions of the genome could be explained by intrinsic characteristics of the genomic region itself and parent age, a linear regression analysis was performed, for which the "secondary" annotation of each *dn*M was carried out using data from the ENCODE [13], LITGEN, and UNIGENE projects. First, according to a previous study [14], in order to collect records regarding the genomic landscape of the identified *dn*Ms, lymphoblastoid cell lines (LCL and GM12878) [15] were chosen. Data were collected for: (1) expression rates (eQTL) [13, 16, 17] in different tissues. According to the expression of regions with *de novo* mutations, these were divided into positions with specific and nonspecific expression; (2) measurements of DNase1 hypersensitivity sites (DHS) [16, 17]. The DHS status was assigned to 0 if outside the DHS peak and 1 if within; (3) measurements of context of CpG islands [16, 17]. If *dn*M was within the CpG islands, a status of position was assigned 1, and if outside – 0; (4) three histone marks (H3K27ac, H3K4me1, and H3K4me3 [16]) from the ENCODE project. If *dn*M was in a position marked with histone, it was assigned with 1, and if not – 0; 5) GERPP++ conservation values were collected using the ANNOVAR annotation tool. According to conservation values, positions with *de novo* mutations were assigned to conservative (GERP++ score >12) and nonconservative positions (GERP++ score < 12) [11]. Based on questionnaire records from the LITGEN project, data on parental age were collected. After the collection of parameters, a number of positions with each parameter was calculated for each trio. A

correlation analysis, followed by linear regression modeling of the *dn*M rate and parameters, was then performed.

1.6 Identification and Evaluation of the Composition and Patterns of *De novo* Mutation Spectrum in the Lithuanian Population, Individuals with ID, and Individuals with ID Sibs Groups

To analyze the mutational spectrum and templates, reference mutational templates from the https://cancer.sanger.ac.uk/cosmic/signatures/SBS/ database were used. During the study, each triplet, having a unique *dn*M in the middle position and nucleotides at the 3' and 5' ends of the reference genome sequence, were respectively assigned to one of the 96 possible triplets. Using the Spearman's correlation coefficient, a correlation analysis was performed between the identified triplet frequencies in the Lithuanian population, the group of individuals with ID, the group of individuals with ID sibs, and the mutation templates from the COSMIC database. The identified templates with *dn*Ms were further compared to the catalog of 6570 *de novo* mutation-based templates compiled by Rahbari et al. for the Ibadan Yoruba (Nigeria) (YRI) population, the population of Utah with Northern and Western European Ancestry (CEU), and Icelandic, Scottish, and South Dakota (USA) populations [29].

1.7 Evaluation of the Relative Fitness of Sequences with *de novo* Mutations

To analyze whether *dn*M affects the relative fitness of individual amino acid types, and, if so, what relative fitness patterns characterize the Lithuanian population, an evaluation of the relative fitness of amino acid sequences with *dn*M was conducted.

Calculations of the selection coefficients or the otherwise relative fitness effect distribution of protein coding amino acid sequences in which the *dn*M had occurred were performed using the *TdG12 site-wise mutation-selection calculation model* (swMutSel0) [18]. This model was chosen because of the highest specificity of the models developed to evaluate the mutability process, where the evolution of each amino acid under study is characterized by an individual fitness value for the amino acid position and a mutation model at the nucleotide level. In this model, it was assumed that the sequences of the genes were formed independently of each other and the effect of natural selection is very small, therefore insignificant. The evolution process has been evaluated over a long period of time (millions of years), and the time from the formation of new sequences with *dn*Ms to their capture is instantaneous. According to the sequences being analyzed, κ (transition and transversion bias), π (base nucleotide composition), and μ (branch scaling) were used. Variants with the relative fitness value <0 were considered as a damaging variant, while those of >0 as a beneficial variant. Using the parameters described above for TdG12, the relative fitness of the amino acid sequences before and after the mutations was evaluated. To analyze whether *dn*M affects the relative fitness of individual amino acid types, and, if so, what relative fitness patterns characterize the Lithuanian population, an evaluation of the relative fitness of amino acid sequences with *dn*M was conducted.

1.8 *In silico* Identification of the Pathogenicity Mechanisms of *de novo* Mutations

Possible pathogenicity mechanisms of pathogenic genome variants with identified *dn*Ms were *in silico* investigated using the following bioinformatic tools and databases:

- *SWISS-MODEL* sofware [19];

- *STRING* sofware [20];
- *OMIM* (*Online Mendelian Inheritance in Man*);
- *GeneCards*®;
- *Ensembl*;
- *ClinVar*;
- *Human Splicing Finder*;
- *denovo-db*;
- *IDGenetics* (http://www.ccgenomics.cn/IDGenetics/);
- *SFARI gene* (https://gene.sfari.org/database/human-gene/);
- *AutDB* (http://autism.mindspec.org/autdb/Welcome.do).

In one case, when a probably pathogenic variant of the *ARID1B* gene was identified, all proband RNA was isolated from venous blood to determine its pathogenicity mechanism using *Tempus™ Blood RNA Tube* and *Tempus™ Spin RNA Isolation* kits according to the manufacturer's recommendations. Copy DNA from RNA was synthesized using the High-Capacity RNA-to-cDNA kit according to the manufacturer's recommendations. A PCR reaction of the cDNA cutting sequence was performed using specific oligonucleotide primers designed with the Primer Blast (NCBI) tool. PCR products were scanned using the standard Sanger sequencing method. Mutation Taster (http://www.mutationtaster.org/) and Human Splicing Finder (http://www.umd.be/HSF3/) databases were used to evaluate the pathogenicity of the change in the splicing site. Possible effects of gene variants with *dn*M were evaluated using the ExPASy Bioinformatics (https://www.expasy.org/) and Pfam 32.0 (https://pfam.xfam.org/) databases.

1.9 Data Analysis and Statistical Methods

The Shapiro-Wilk test was used to test the distribution of samples in the normal distribution. Pearson or Spearman coefficients were

used to evaluate correlation. Linear regression analysis, calculations of Mann-Whitney *U* test, Shapiro-Wilk Test, and the Fisher exact criterion were performed using the Open Access *R* package, version 3.4.3, and *R studio*, version 1.2.1335 [12]. A *p-value* of less than 0.05 was considered statistically significant in the analysis. All other calculations were performed using *Microsoft Office Professional Plus 2013* Excel.

RESULTS AND DISCUSSION

Almost all common variants of the genome are already known. Thousands of genetic variants and/or genes are associated with relevant human traits and diseases. Meanwhile, the identification of rare genetic variants is still progressing rapidly. The need for this research is based on the fact that the diversity of rare variants directly depends on the regions of origin of the individuals, so the more of these regions, the more rare variants are found. For this reason, the analysis of rare variants, most of which are unique *de novo* variants, requires different populations, which requires considerable effort. This work involved a detailed analysis of the distribution and rate of *de novo* mutations in the whole human exome regions of the Lithuanian population (hereafter referred to as the first group), the trio when the proband has an intellectual disability (hereafter referred to as the second group), and the trio when the proband is a sibling of a person with an intellectual disability (hereafter referred to as the third group) (275 individuals studied in total).

2.1 Selection of the Algorithm for Detecting *de novo* Mutations

When investigating the first group, it was aimed at selecting the optimal *dn*M detection software for further work. It should be noted that *VarScan* software was selected for *dn*M detection, because the

exome sequencing of the Lithuanian population group was performed using the *5500 SOLiD*™ Sequencer genetic analyzer, and the primary sequencing files in *.XSQ* format were immediately converted to *.BAM* format files using *LifeScope*™ Genomic Analysis Software 2.5.1. In these files, there is either no information about the software used or it is formatted differently than in the *.BAM* files obtained by sequencing exomes with the now widely used Illumina genetic analyzer and processed by the GATK software according to best practice guidelines. Regarding the missing or improperly formatted information about sequenced fragments using LifeScope™ software, the *.BAM* files have resulted from similarly invalid *.vcf* files to the widely used, sensitive, and specific *dn*M identification algorithm, *PhaseByTransmission* [21]. Meanwhile, *VarScan* software is appropriate for analyzing the *.BAM* files obtained by sequencing with the use of both the *5500 SOLiD*™ Sequencer and *Illumina HiSeq 2500* genetic analyzer regardless of the software used to generate the *.BAM* files. Based on the data of Warden CD et al. (2014) [22], this algorithm provides the results that are >97% consistent with the high quality data obtained using *GATK UnifiedGenotyper* and *HaplotypeCaller* softwares. *VarSeq*™ software has been chosen as an analogue of the open-access *VarScan* software, as it is commercial, more visually appealing, and user-friendly for Windows users, used for both scientific and clinical research. *VarScan* runs on Linux.

Thus, according to the study data, a total of 95 *dn*M was set for 34 trios in the first group using *VarScan* software, while using *VarSeq*™ – 84 *dn*M for 31 trios. Although the two algorithms used are designed to identify *dn*M in the proband exome when those variants are not present in the parent exomes, only 5.37% of all *dn*Ms detected by both softwares have matched. After the *dn*M filtration stage, it was determined that the *VarScan* algorithm has a higher sensitivity (5.42%) than the *VarSeq*™ algorithm (1.76%), while the efficiency of mutation detection, as verified by the *Sanger*

sequencing method, varied by only 4%. Thus, it can be concluded that none of the tools could correctly detect true *dn*M due to its extremely high specificity (> 99%) and low sensitivity. For this reason, it is important to emphasize the need for additional rigorous data filtering methods when detecting *dn*M. When applied, the number of true *dn*Ms identified ranged from 1752 to 95 (by *VarScan* data) and from 4756 to 84 (by *VarSeq*™ data). Since it is known that softwares with higher (≥100) horizontal sequence overlaps should also have a higher sensitivity, the *VarScan* algorithm was chosen due to its higher sensitivity and open access for the further identification of *dn*M in other groups of subjects. When applied in the second and third groups, 78% of the mutations were confirmed and verified by *dn*M *Sanger* sequencing.

2.2 Rate of *de novo* Mutations in the Lithuanian Population, in Groups of Individuals with Intellectual Disability, and Individuals with Intellectual Disability Sibs

Germ cell mutations represent a fundamental force of evolution that shapes phenotypic diversity; therefore, the accurate detection of the *de novo* mutation rate is extremely necessary and can be applied in a wide variety of ways: interpretations of disease-causing mutations [23–25], dates from natural selection studies [25, 26], and demographic events based on genetic analysis [27–30] to human mutagenesis studies [31]. In addition, the *de novo* mutation rate is a key and central biological parameter that helps us understand evolutionary phenomena. One of the most important is the "molecular clock": the continuous process by which genomic changes accumulate during evolution, and the neutrality theory that explains the causes of the effect of neutral genetic changes, such as changes without effect on protein fit, on the genome.

Many different methods have been developed to determine the rate of *dn*M based on population genetics models or fossil dating

data, which require many additional implicit estimates to be realized. Some methods require demographic population models (based on the time to first common ancestor of the subjects), rate of recombination of a genome divided into short segments, or the approximate duration of one generation in years to estimate the annual *dn*M rates. Using this additional data, the rate of *dn*M is determined from the autozygous regions of the genome that originate from the closest common ancestor that can be reliably identified. This technique is advantageous in that the potential contribution of somatic mutations is minimal when estimating the rate of *dn*M, but only the *dn*M average of both genders is determined, and for several generations at once.

Another approach is to determine the average rate of *dn*M at different periods of evolution from fossil DNA dating data. In this way, the divergence time of the analyzed object, such as the surviving human DNA, and the outer object – for example, chimpanzee DNA – are compared with the fossil genomes, thus estimating the rate of *dn*M according to their differences [32, 33].

In this thesis, the rate of *dn*M was calculated and evaluated using one of the most accurate methods – based on a directly determined *dn*M throughout whole exome sequencing. Based on the obtained data, the estimated *de novo* single nucleotide mutation rate in the Lithuanian population is 2.74×10^{-8}, and the *de novo* indels – 1.77×10^{-8} per position per generation (PPPG). In the second group, the rate of a single *de novo* nucleotide is 3.05×10^{-8} PPPG, and in the third group – 7.19×10^{-8} PPPG. The incidence of *de novo* indel mutations in the genome of the second group is 1.74×10^{-7} PPPG, and in the third group – 1.08×10^{-6} PPPG. In all three groups, the estimated rate of *dn*M is significantly higher (*p-value* = 4.49×10^{-9} for the first group, *p-value* = 5.15×10^{-11} for the second group, and *p-value* = 4.8×10^{-6} for the third group, respectively) than that found in previously published sustainable studies of the Icelandic, Danish, Dutch, Canadian, and American general populations and the group

of autists [32, 33], where *de novo* single-nucleotide mutations have a rate between 0.96×10^{-8} and 1.28×10^{-8} PPPG, and *de novo* indels – 1.1×10^{-9} PPPG. The higher rate of *dn*M found in this study can be based on the fact that the study was performed using only exome sequencing data. It is worth noting that exomes exhibit a significantly higher (about 30%) mutation rate than the whole genome. This is explained by the fact that the composition of heterocyclic bases throughout the genome differs from that of the exome: in the exome, the regions with repeats of guanine and cytosine dinucleotides account for about 50% of the total size of the exome, whereas for the whole genome, they account for about 40% [34]. Cytosine and guanine dinucleotide-rich regions methylated in human genomes are known to be highly mutable due to a spontaneous deamination of cytosine bases [34]. Based on comparative genomics studies, higher mutation rates in CpG dinucleotide-rich regions are thought to have occurred approximately during the increase of taxonomic diversity in mammals (also known as mammalian radiation) [35]. This hypothesis was raised based on the knowledge that, as species diverged, CpG dinucleotide-rich coding regions had a higher mutation rate than the noncoding DNA regions [36]. During evolution, as the mutations progressed, the coding regions turned into noncoding regions, so the context effect of the CpG dinucleotide genome, like the average mutation rate, declined over time until it reached the mutation rate detected in the noncoding DNA surrounding the coding sequences. However, although there was sufficient time in the genomic regions with a different CpG dinucleotide context to create neutral genomic regions and thus achieve equilibrium, purifying natural selection in functional genomic regions retained the share of hypermutable CpG dinucleotides [35–38]. This study identified a higher rate of *dn*M than the genomic studies of other researchers, presumably because of the specific context of the CpG exome sequences and whether natural selection affects the exome. Based on linear regression models, it was

found that ~68–94% of the total rate of *dn*M in the genome of the Lithuanian population is explained by DNAse1 hypersensitivity, CpG context, and regional conservatism (based on GERPP++ conservatism estimates) regional expression levels (**TABLE 1**).

The rate of *dn*M in the group of individuals with ID was also moderately (13%, *p-value* = 0,02) explained by the expression levels of the exome regions (Table 1). Based on these data, it was concluded that the *dn*M exome is formed in promoter-like, transcriptionally active regions of the genome, regardless of DNA sequence conservatism, although the rate of *dn*M was higher in genes with nonspecific expression. This coincides with the location of gene expression, transcription factor binding, and histone modifications in many species of organisms and their tissue map data [13, 17, 39]. They have revealed that the expression of identical or similar tissue genes in different species is conservative [40], but that the enhancer and promoter sequences are characterized by rapid changes in DNA structure [41–43]. For example, a recent study of hepatocyte promoters and enhancers in 20 mammalian species revealed that 25% of enhancers and 10% of promoters were unique to all species, even when sequences were particularly conservative [41, 44]. Similar results were obtained in human, macaque, and mouse limb tissues [45].

Contrary to other research work, this study found that fathers' ages did not correlate with the *dn*M rate in the Lithuanian population (fathers $R_S = 0.2$, *p-value* = 0.09, mothers $R_S = 0.08$, *p-value* = 0.98) (**FIGURE 1**).

TABLE 1. Influence of the number of DNase1 highly sensitive regions, regions abundant with CpG dinucleotides, number of conservative regions, and number of genomic regions with different expression on the intensity of the *dnM* rate in Lithuanian populations and the group of individuals with intellectual disability.

Response variable	**Independent variable**	***β*±SE***	***p-value***	***R²***
Lithuanian population				
***dn*M rate**	DNase1 hypersensitivity sites	$2{,}83\times10^{-6}\pm1{,}96\times10^{-7}$	$2{,}31\times10^{-130}$	0,89
	Conservative positions	$5{,}28\times10^{-6}\pm5{,}67\times10^{-7}$	$1{,}49\times10^{-6}$	0,88
	Nonconservative positions	$2{,}56\times10^{-6}\pm2{,}67\times10^{-7}$	$2{,}45\times10^{-9}$	0,8
	Context of CpG islands	$2{,}76\times10^{-6}\pm4{,}94\times10^{-7}$	$8{,}93\times10^{-6}$	0,68
	Positions with nonspecific expression	$2{,}05\times10^{-6}\pm1{,}144\times10^{-7}$	$8{,}34\times10^{-140}$	0,94
The group of individuals with intellectual disability				
***dn*M rate**	Positions with nonspecific expression	$8{,}56\times10^{-7}\pm3{,}52\times10^{-7}$	0,02	0,13

β – regression coefficient showing how much the rate of *dn*M increases as the independent variable increases by one (one position or region, respectively), **SE –** standard error, **R^2 –** coefficient of determination.

This can be explained by the fact that the age of the fathers of the trios analyzed in the study is too similar, and that the power of age and *dn*M rate analysis is therefore not sufficient. Also, the study analyzed only a small (~1.5%) part of the entire genome – in the exome, therefore, on average, 1.9 (*VarScan*) and 1.7 (*VarSeq™) de novo* single-nucleotide mutations were detected in each of the analyzed exomes, while approximately 40 to 82 *de novo* single-nucleotide substitutions were identified throughout the genome [34, 35, 46–49]. However, in the Lithuanian population, the number of *de novo* indels (on average 0.0625 *de novo* of indel per person) was similar to that found in Danish and Dutch studies [14, 49]. The *de novo* SNP was higher in the group of individuals with intellectual disability – an average of 4.2 *de novo* single nucleotide and 1.3 *de novo* indel, ranging in size from one to 30 nucleotides. Meanwhile, in the group of individuals with intellectual disability sibs, on average, each proband studied had 2.4 *de novo* SNP and 0.92 *de novo* indels between three and 31 nucleotides in size. In the third group, only one *dn*M was detected in the coding sequence (8%) among the *de novo* indels, whereas ten (24%) were found in the second group, of which as many as four *dn*M are of the frame shift type. Hence, the higher number of *dn*M detected in the second group suggested that the *dn*M may have been the cause of intellectual disability in this group.

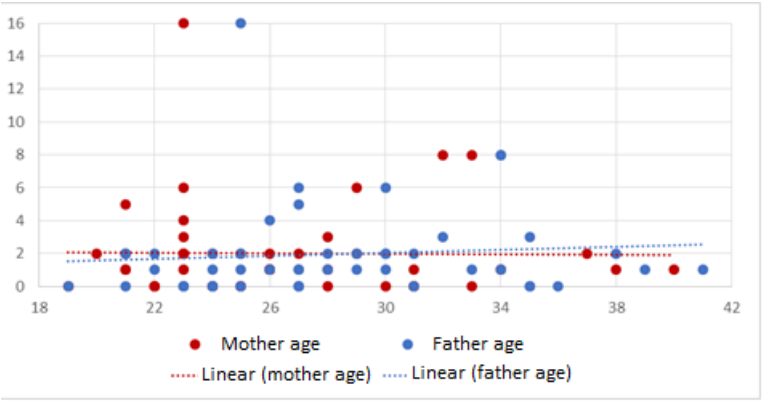

FIGURE 1. Graph of correlation between the *dn*M number and fathers' ages.

Lithuanian populations and groups of individuals with intellectual disability sibs were similar in functional regions where *dn*Ms are detected. In these, the majority (54–64%) consisted of *dn*M present in introns and only 21–24% were in exons. Meanwhile, in the group of individuals with intellectual disability, *dn*M was 48% in introns and 34% in exons. Analogous group similarities and differences can be seen in the analysis of groups by *dn*M types: an abundance of mutations leading to altered protein composition due to altered amino acids was evident in the group of individuals with intellectual disability: two stop codon *dn*M, five frameshift *dn*M, 13 synonymous, 28 nonsynonymous *dn*M, and five *de novo* deletions of three to 31 nucleotides in size, when the basis for *dn*M in the Lithuanian population and in the groups of individuals with intellectual disability sibs were 13 synonymous and 18 nonsynonymous *dn*M.

Based on the analysis of the *de novo* mutation spectrum composition, there is no difference in the composition of the *dn*M spectrum between the groups of the Lithuanian population, individuals with ID, and those with ID sibs (*p-value* = 0.575, $\chi 2$ criterion). Furthermore, the composition of the *dn*M spectrum is

close to populations of Utah Yoruba, originating from Northern and Western Europe (CEU), Nigeria (YRI), Iceland, Scotland, and South Dakota (USA) [29]. For the most part of all groups, C> T (33% in the first group, 35% in the second group, and 59% in the third group) and T> C (28% in the first group, 27% in the second group, and 18.5% in the third group) substitutions were identified.

However, an analysis of the correlation of the triplets with a *dn*M frequency in the Lithuanian population with the 49 exome templates of the COSMIC database revealed that all groups analyzed in this work differ in their composition from the *de novo* template catalog based on the research conducted by Rahbari et al. [50] and are unique in the following way: the Lithuanian population is characterized in almost equal proportions (42–49%) by six mutation templates: 26, 10a, 15, 41, and 7a. The 44, sixth, fifth, 20, and seventh b mutation templates describe the group of individuals with ID, and the fifth, 33, and 23 mutation templates describe the group of siblings for individuals with ID (**FIGURE 2**). Meanwhile, in the CEU, YRI, Icelandic, Scots, and South Dakota populations, the fifth template is assigned for ~ 75% of all *dn*Ms and the first for ~ 25%.

The *dn*M templates identified in the Lithuanian population are thought to result from errors in the DNA strand mismatch repair mechanism (template 26, 15), ε polymerase exonuclease domain mutations (template 10a), and ultraviolet-induced cyclobutane pyrimidine dimers (template 7a). The etiology of the 41 mutation template is unknown. Mutation patterns identified in individuals with the ID group are thought to result from errors in the repair mechanism of DNA strand mismatches (templates 44, 6, and 20), also mutations in the *POLD1* gene (Template 20), and UV radiation (Template 7b). The etiology of the fifth mutation template is unknown, but the expression of this template was found to be characterized by bladder cancer cells with *ERCC2* mutations. In the group of siblings for individuals with ID, the etiology of any of the identified mutation templates is unknown.

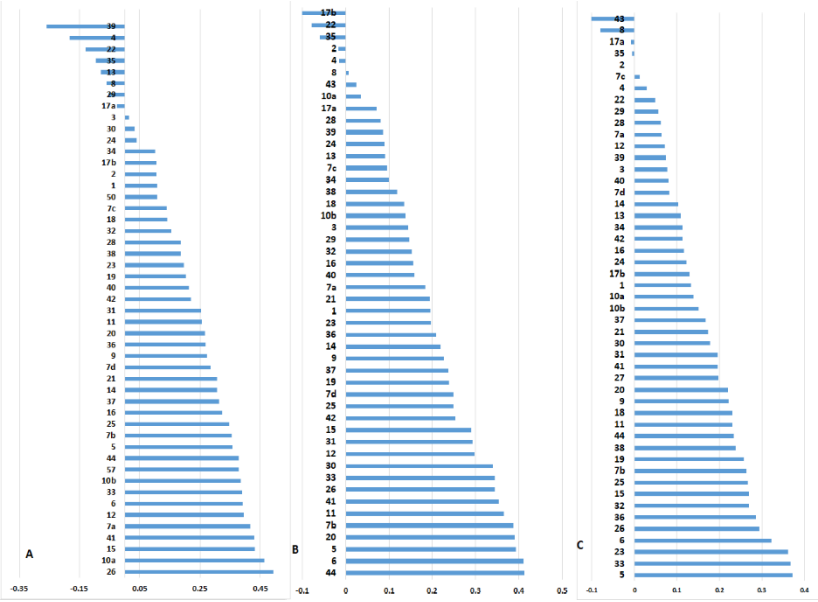

FIGURE 2. Correlation of the frequency of triplets with *dn*M detected in (A) the Lithuanian population, (B) the group of individuals with ID, and (C) the group of siblings for individuals with ID with 49 templates of exomes in the COSMIC database. The horizontal axis represents the Spearman correlation coefficient, the vertical indicates the number of the mutation template.

2.3 Discussion of the Results of the Analysis of the Relative Suitability of Protein Sequences in which *de novo* Mutations are Detected

In population genetics studies, the rate distribution of individuals with different types of *de novo* mutations and their associated selection coefficients is an essential subject of population genetics research [51, 52]. The mutation rate and relative fitness affect a wide range of evolutionary and biological phenomena, such as the maintenance of genetic variation, alteration of quantitative traits,

recombination, sex evolution, evolution of aging, persistence of gene copies, and so on [53]. When a *de novo* mutation occurs in the genome, it can have one of three effects on the fit of an individual's phenotype (hereafter $w = s + 1$, where w is the relative fit and s is the selection factor): mutation can be harmful ($s < 0$), resulting in the reduced survival or fertility of individuals, neutral ($s \approx 0$), where the mutation has such a small effect on the fitness of individuals that its rate in the population is directly dependent on gene drift, or beneficial ($s > 0$), when individuals with such a mutation have a longer survival time or greater reproductive success than others. The final fate of *dn*M, whether captured or removed from the population genome, depends on the effect of natural selection on the strength and gene drift on the effective population size; therefore, fitness is described by a fitness coefficient S equal to the product of the effective population size N_e and the natural selection s ($S=2N_e \times s$). In this work, the suitability of each amino acid with *dn*M to factor S prior to natural selection was analyzed, as *dn*M was evaluated instantaneously. Using the *swMutSel0* model, it was found that in the Lithuanian population, neutral ($-2 < S < 0$) mutations accounted for only ~ 9%, while harmful ($S < -2$) did up to 91%, of which the highly harmful ($S < -10$) were ~ 73% (**FIGURE 3**). In the group of individuals with intellectual disability, neutral mutations accounted for about 27.5%, harmful ~72.5%, of which highly harmful accounted for ~60%, and in the group of individuals with intellectual disability sibs, neutral mutations were similar to those in the first group – 12.5 %, and harmful – 87.5 %. The part of neutral and pathogenic mutations in the group of individuals with intellectual disability was similar to that found by Eyre-Walker A and Fay JC [52, 54] – 20% and 80%, respectively. Adaptive mutations have not been detected in any of the groups, although recent data indicate that the rate of adaptive mutations in genome $\hat{\alpha}$ is as high as 13.5% [55].

Paradoxically, although this study is based on a study conducted in 2016 by Lelieveld SH et al. [56], it was hypothesized that, in

Lithuania, the proportion of harmful mutations in the group of individuals with intellectual disability should be significantly higher than in the other studied groups; based on Fisher's exact criterion, there was no statistically significant difference in the rate distribution of amino acids with different relative fitness between all three groups (*p-value* = 0,63). Moreover, during the study, it was found that in Lithuania and in the group of individuals with intellectual disability sibs, the harmful mutations accounted for a ~18% higher *dn*M share than in the group of individuals with intellectual disability. Assuming that this result was potentially influenced by the fact that in the first and third groups of all of the detected *dn*M, relative fitness was assessed only for the respective 11% and 17% of the detected *dn*M located in the coding regions of the genome; the part of the harmful *dn*M for each group was re-evaluated by including in the analysis those *dn*Ms that could not be included in the relative fitness analysis due to their positions in non-coding genomic sequences. Considering that additionally included *dn*M are potentially neutral because they do not significantly influence the health of the subjects (in the questionnaire, survey subjects assessed themselves as "healthy" although they may become ill in the future), the above hypothesis was confirmed, in which case the part of harmful *dn*M in the group of individuals with intellectual disability reaches ~20%, and in the first and third groups – ~10% and ~13%, respectively.

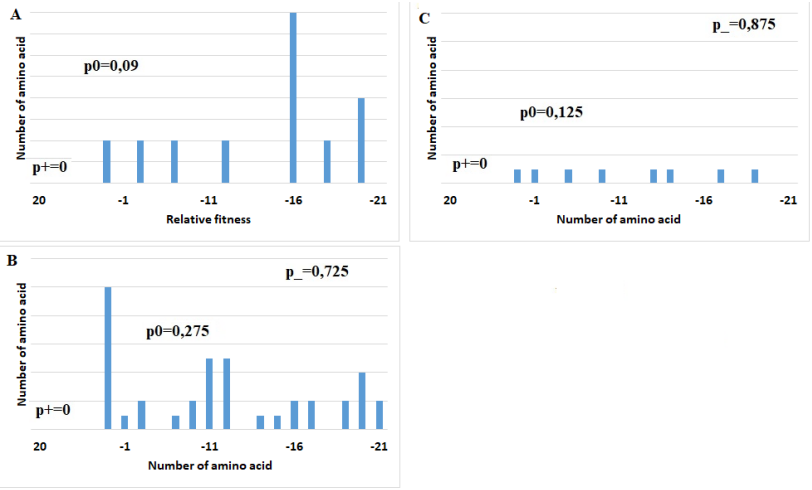

FIGURE 3. Distribution of relative fitness coefficient *S* in the genes for which the determined *dn*M. p_- corresponds to the harmful ($S < -10$) *dn*M part, $p+$ beneficial ($S > 0$) *dn*M part, and p_0 – neutral ($-2 < S < 0$) *dn*M part. Data of the analysis of (A) the Lithuanian population, (B) the group of individuals with intellectual disability, and (C) the group of individuals with intellectual disability sibs.

At the same time, the data of the relative fitness analysis reveal that before the effects of natural selection on the number of *dn*M (and at the same time intensity) in the genome gives a direct indication of the possible effects *de novo* mutations on the genome and on the individual's phenotype. Since a significantly higher number of *dn*M increases the likelihood that unique mutations (especially in the coding regions of the genome) may have a harmful effect on an individual's phenotype, and thus may be the cause of intellectual disability [57]. Meanwhile, when exposed to natural selection, the effects of harmful mutations and the rate of *dn*M in the genome that individuals can tolerate depend only on whether it affects the absolute or relative fitness of the person's phenotype with intellectual disability [58].

The theory of neutrality is evidenced by the "mirror reflection effect" observed in the first group, where most of the amino acids with high negative relative fitness ($S < -16$) turned into neutral (0) or low (up to -5) relative amino acids after *dn*M (**FIGURE 4**). All non-polar amino acids had less negative relative fitness values after *dn*M, whereas the relative fitness of polar amino acids changed slightly after *dn*M. The number of amino acids by type remained unchanged, and *dn*M had no significant effect on the structure and function of the amino acids and the proteins they formed. Meanwhile, in the second group, in the group of individuals with intellectual disability, *dn*M significantly changed the amino acid composition – non-polar amino acids increased by 1.9 times, while charge-loaded amino acids had, on the contrary, decreased by 1.7 times. Also, the relative fitness of only the three polar amino acids increased from -3, -10, and -21 to 0 and -16, the relative fitness of the resulting non-polar amino acids increased for six of the 11 (54.5%) amino acids, and for the charged amino acids, the relative fitness increased for four (33%) and decreased for five (42%) of the 12 amino acids with charge (**FIGURE 5)**. Thus, when summarized, both in terms of the number of amino acids and changes in their relative fitness, it can be predicted that the structure of the proteins encoded by the studied amino acids should undergo significant changes following *dn*M, leading to possible changes in protein function associated with phenotypic traits.

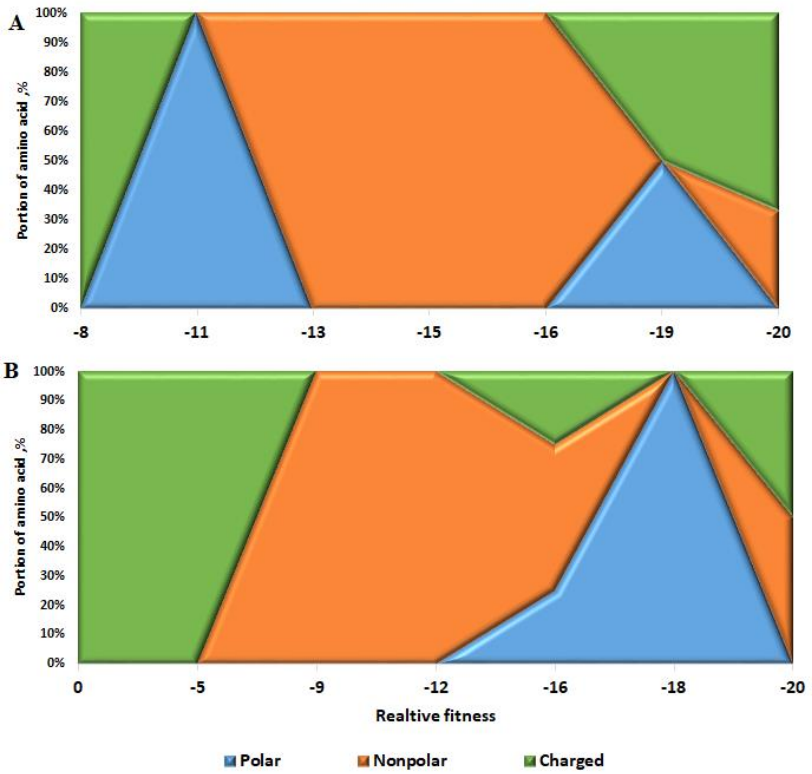

FIGURE 4. Summarized estimates of relative fitness in the group of general population of Lithuania. **A** estimates of the relative fitness of amino acids before mutations, **B** – following *de novo* mutations.

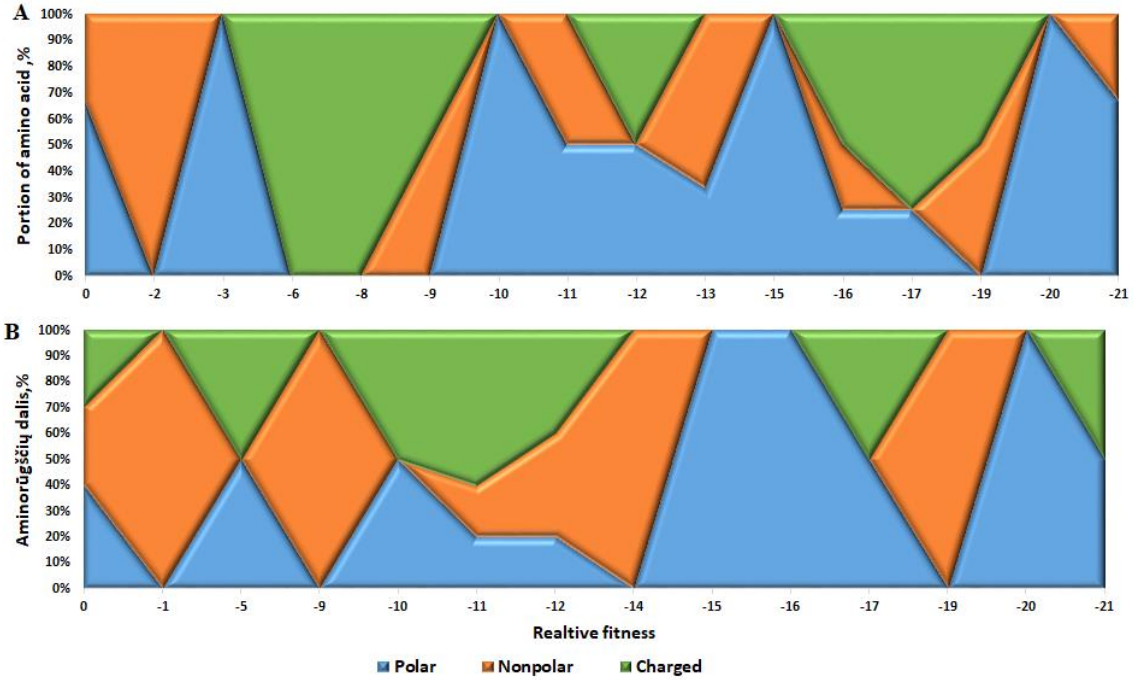

FIGURE 5. Summarized estimates of relative fitness in the group of individuals with intellectual disability. **A** figure shows the estimates of the relative fitness of amino acids before mutations, **B** – following *de novo* mutations.

The relative fitness data in the third group replicates the first group, because the relative fitness effect values of all types (except for one amino acid with charge) of amino acids decreased with *dn*M, while their number by type did not change. This indicates that all types of amino acids have become more neutral following *dn*M than before, suggesting that the analyzed *dn*M has no significant effect on changes in the structure or function of the amino acids and thus the protein.

2.4 A Discussion of the Functional *in silico* Analysis of *de novo* Mutation Results

The results of an extended functional *in silico* analysis indicated that in the Lithuanian population group, among all the identified *dn*M, 4 *dn*M detected by *VarScan* software and 48 by *VarSeq*™ software were potentially pathogenic. The large difference in the number of pathogenic *dn*M can be explained by the fact that, depending on the algorithm used to detect *dn*M, the number of detected *dn*M in the coding sequences differed significantly. For example, the *dn*M detected with the *VarScan* algorithm was only 21.05%, and 95.24% with the *VarSeq*™ algorithm. Based on the knowledge of protein structure and function algorithms for assessing the effects of mutations in the coding part of the genome are abundant, there is little knowledge of regulatory regions of the genome, and thus of algorithms to evaluate their structural changes, so mutations in coding sequences are more often assessed as pathogenic. Meanwhile, regulatory regions of the genome that affect protein structure and function indirectly make up a large part of the non-coding genome, accounting for ~11% of the total genome size per protein. Their scattering in the genome ensures minimal impact of pathogenic genomic variants [15]. The pathogenic *dn*M detected in the first group were in genes the encoding proteins of which were important for DNA repair, chromatin remodeling, ribosome biogenesis, immune response, lipid biosynthesis, post-translational

labeling of proteins in the endoplasmic network, signaling pathways in the cytoplasm, transport of molecules across membranes, regulation of cell cytoskeleton, cell growth and survival, and the initiation of neuronal responses that influence odor perception. However, despite the fact that these *dn*Ms were found to be pathogenic, all individuals considered themselves "healthy" during the LITGEN survey. Thus, it demonstrates that, despite the possible pathogenicity of *dn*M, such changes are apparently tolerated by the genomes containing these *dn*Ms, so that the phenotypic effects of these *dn*Ms do not occur. This demonstrates that the pathogenic *dn*Ms detected in this study are not harmful enough to diminish the overall fitness of the encoded proteins and therefore have the ability to persist for many generations and avoid the effects of natural selection. According to Szamecz's study, the more frequently *dn*M occurs in conservative regions of the genome, the stronger the effect of natural selection is through genetic compensatory mechanisms on genetic changes [59]. The harmful effects of change can be mitigated in four ways. Some genes may tolerate truncated protein variants because their functional effects may be masked by (1) incomplete gene expression, (2) compensatory protein variants, or (3) functional significance due to low protein truncation [60]. Other cases of gene changes associated with non-synonymous *dn*M can be compensated (4) by the accumulation of the required number of useful mutations in the genome [59].

Functional *dn*M analysis in the group of individuals with ID (second group) allowed 13.5% of the individuals to identify the genetic cause of ID and to explain the etiopathogenesis of the disease. Five pathogenic *dn*Ms in the *ARID1B*, *PACS1*, *TCF4*, *CHD7*, and *MECP2* genes have been identified and confirmed in the group of individuals with ID, possibly pathogenic mutations, three of which have so far not been described in the literature. For a pathogenic *de novo* mutation in the *CREBBP* gene that is not described in all

evaluation estimates, the patient's phenotype will be reassessed to determine whether the mutation detected results in a phenotype specific to the *Rubinstein-Taybi* syndrome. According to the OMIM database, ~800 genes are now known to cause intellectual disability, accounting for ~19% of the known genetic disease genes, so in many cases, as in this study, the causes of intellectual disability remain unclear.

All of the detected *dn*M were in the heterozygous condition. DNA and RNA analysis revealed that NC_000006.12 (NM_020732.3): c.5025+2T>C *dn*M in the *ARID1B* gene induces the autosomal dominantly inherited Coffin-Siris syndrome (MIM #135900). Proband cDNR sequencing by the *Sanger* sequencing method revealed that *dnM* disrupts the cleavage center at the 5' end of intron 19, causing exon 19 to be omitted. Missing one exon during splicing is a frequent result of pathogenic pre-mRNA, resulting in open DNA reading frame shifts, the formation of premature termination codons, or shorter protein synthesis [61]. Due to the presence of both wild-type and mutant transcripts in the cells, the mutant transcripts are not completely disintegrated, and instead the truncated protein NP_065783.3:p.(Thr1633Valfs*11) is synthesized. Truncated ARID1B results in the loss of the BAF250 domain, which is part of the ATP-dependent SWI/SNF-like chromatin modification complex that regulates gene expression. BAF250 may act similarly to the E3 ubiquitin ligase targeting H2B histones [62]. In addition, previous studies of mouse embryonic cells with an inactive BAF250b complex have shown that BAF250 is particularly important for early embryogenesis, as modified cell regeneration is impaired and differentiation is increased [63–65]. ARID1B-associated BAF complexes are involved in neuronal differentiation and mammalian brain development. In particular, their expression is important for the development of pyramidal neurons during the formation of complex dendritic architecture [62, 63]. The haploinsufficiency of ARID1B is also known to cause a decrease in GABAergic interneurons in the

cerebral cortex. In this way, excitatory and neurotransmission-inhibiting processes are affected during brain development. Furthermore, due to the interaction of the BAF complex with histones, ARID1B haploinsufficiency may also function through epigenetic regulation.

NM_018026.4:c.607C>T (Arg203Trp) in the *dn*M PACS1 gene induces autosomal dominantly inherited intellectual disability with a disease-specific phenotype (MIM #615009). In 2015, Gadzicki D et al. [66]. identified this mutation and explained the etiopathogenesis of intellectual disability [67]. PACS1 is a membrane regulator of the Gold's apparatus that directs proteins in the required direction. During embryogenesis, high levels of expression occur in the brain and are also regulated in the postnatal period [68]. To investigate this change, it was hypothesized that c.607C> T *dn*M near the proximal CK2-binding site motif alters the binding properties of PACS1 [66]. Studies on zebrafish and antisense DNA sequences have confirmed this hypothesis. These genomic variants were found to have a dominant negative effect on the survival of zebrafish neurons with SOX10 as a result of the formation of facial dysmorphia [69].

De novo duplication of 18 nucleotide was detected in the *CHD7* gene (c.6341_6358dup, p.Asp2119_Pro2120ins6); NM_017780; NP_060250; MIM# 214800). The mutation was found to be in the evolutionarily conservative region, and protein modeling revealed that the conformation of the DNA sequence containing *dn*M changed from a linear strand to a short α helix. CHD7 is known to encode the chromodomain DNA-binding helicase, which is important in early embryogenesis and controls gene expression through chromatin modification during the cell cycle [70]; therefore, even in three out of four cases, mutations in individuals with the phenotypic features of CHARGE (MIM #214800) syndrome are specifically identified by the *CHD7* gene. The subject was also confirmed to have CHARGE syndrome, and the mutation was not only *de novo*, but also unique,

so far not described. CHARGE syndrome is characterized by a combination of symptoms of coloboma, heart defect, growth or developmental malformation, genital and ear abnormalities, deafness, and chuan atresia (occlusal nasal passages).

NM_001243234:c.1259G>A:p.420R>Q (rs121909121) *dn*M was detected in the *TCF4* gene leading to autosomal dominantly inherited *Pitt-Hopkins* syndrome (MIM #610954). The *TCF4* gene encodes a helix-loop-helix structure transcription factor (bHLH), also known as a member of the E protein family. It is evolutionarily conservative and is particularly important for neuronal determination and differentiation because of its ability to bind DNA as homodimers or heterodimers [71, 72]. When the resulting mutation does not allow TCF4 to interact properly with ASCL1, the tissue-specific HLH protein, the development of individual brain structures becomes impaired in the PHOX-RET pathway [73]. In addition, mouse studies have shown that embryo development is directly dependent on the number of copies of the TCF4 protein and, when their number is less than 3 – the mouse embryos did not survive [74]. The expression of this protein is high in the human brain, lungs, heart, and other muscles. Individuals with *Pitt-Hopkins* syndrome have intellectual disabilities, a wide mouth, distinctive facial features, and intermittent hyperventilation followed by apnea.

*Dn*M NM_001316337:c.194C>T :p.65T>M (rs28934906) in the *MECP2* gene eads to *Rett* syndrome (MIM #312750). MECP2 is a multifunctional protein that can alter gene expression and metabolism in a number of ways: it can activate the AKT/mTOR signaling pathway, is involved in alternative gene cutting, and is involved in microRNA and long non-coding RNA expression, as well as in epigenetic chromatin modification [75–79]. Its main function is to recognize and bind specifically to methylated and A/T-based cytosine residues (5MeCyt) in the DNA strand. Mutations in the *MECP2* gene result in the loss of specificity for binding to

5MeCyt and are at the same time the cause of the *Rett* syndrome. The *Rett* syndrome defines the phenotype as the expression of genes that alter neurobiological activity, network formation, and function through epigenetic regulation of MECP2. Individuals with *Rett* syndrome (predominantly females as it is a predominantly hereditary disease of the X chromosome) have impaired development between the ages of 6 and 18 months, a regression of acquired skills, loss of speech, stereotypical (usually hand) movements, microcephaly, seizures, and, of course, intellectual disability. The *Rett* syndrome is classified as a neurological and autistic spectrum disease [75, 77, 78]. Differences in dose compensation of the *MECP2* gene on the X chromosome may result in different phenotypic expression of the syndrome.

In all five cases of the confirmed pathogenic *dn*M, the phenotypes of the subjects studied were very similar to the phenotypes of published individuals with identical intellectual disability syndromes.

In conclusion, the analysis of the distribution of *dn*M, spectrum and their genetic and epigenetic context in this work provided knowledge and insights into the genetic diversity of the Lithuanian genome and its possible causes. These data and results enhance researchers' capabilities and at the same time facilitate the separation of true pathogenic *dn*M from a tolerable *dn*M background and help identify potential causal *dn*M for intellectual disability.

CONCLUSIONS

1. For the detection of *de novo* single nucleotide and small indels in the genome, when it is sequenced by different technologies, the *VarScan* open access algorithm is more appropriate because it has a higher degree of sensitivity, is suitable for working with target sequences, exome, or the whole genome, analyzes data from various

genetic analyzers (Illumina, *SOLID*, *Life/PGM*), is based on a heuristic / statistical method for identifying genomic variants that meet the desired estimates of horizontal region coverage, base quality, variant allele rate, and statistical significance.

2. a) The Lithuanian population, a group of individuals with an intellectual disability, and group of siblings of a person with an intellectual disability did not differ from each other in the *de novo* mutations spectrum (*p-value* = 0.575) and are close to the population of Utah from Northern and Western Europe (CEU), Ibadan Yoruba (Nigeria), Icelandic, Scottish, and South Dakota (USA) *dn*M spectra of: C> T (33% in the first, 35% in the second, 59% in the third group) and T> C (28% in the first, 27% in the second and 18.5% in the third group).

b) Based on the mechanisms of the *dn*M spectrum formation, all studied groups are unique:

- The *dn*M detected in the Lithuanian population is thought to be due to errors in the repair mechanism of DNA strand mismatches, mutations in the exonuclease domain of ε polymerase, and cyclobutane pyrimidine dimers formed by UV radiation.
- The *dn*Ms detected in the goup of individuals with an intellectual disability are believed to be due to errors in the repair mechanism of DNA strand mismatches, mutations in the *POLD1* gene, and UV radiation.
- The etiology of any identified *dn*M in the group group of sibs of a person with an intellectual disability is unknown.

3. The estimated *de novo* mutation rate in all the studied groups is significantly higher (*p-value* = 4.49×10^{-9} for the first group, *p-value* = 5.15×10^{-6} for the second group, and *p-value*

= 4.8×10^{-6} for the third group) than in the Icelandic, Danish, Dutch, Canadian, and American general populations and the group of autists, and this was determined by the genetic structure of the analyzed genome, the exome.

- The rate of *de novo* single nucleotide and *de novo* indels in the Lithuanian population is 2.74×10^{-8} and 1.77×10^{-8} per position per generation, respectively; for the group of individuals with intellectual disability – 3.05×10^{-8} and 1.74×10^{-7} per position per generation, respectively, and for the individuals with intellectual disability sibs – 7.19×10^{-8} and 1.08×10^{-6} per position per generation, respectively.

4. *De novo* mutations in the exome are generated in promoter-like, transcriptionally active regions of the genome, regardless of the conservatism of the DNA sequence.

- The intensity of *de novo* mutations (~68–94%) in the exomes of the Lithuanian population is influenced by regions particularly sensitive to DNAse1, CpG context, regional expression levels.

- In the group of individuals with intellectual disabilities, the intensity of *de novo* mutations was slightly (13%) influenced by the expression levels of exome regions.

5. There are significantly more pathogenic mutations (*p-value* = 0.02) in the exomes of individuals with intellectual disabilities than in the Lithuanian population and the groups of individuals with intellectual disabilities sibs; therefore, unique *de novo* mutations (especially in the coding regions of the genome) are likely to have harmful effects on a person's phenotype and to cause intellectual disability.

- Five pathogenic *de novo* mutations in the *ARID1B*, *PACS1*, *TCF4*, *CHD7*, and *MECP2* genes were detected in 13.5% of

subjects with intellectual disabilities, three of which (in the *ARID1B*, *CHD7*, and *PACS1* genes) have not yet been described in the literature previously, causing *Coffin-Siris*, CHARGE, and autosomal dominantly inherited intellectual disability, respectively.

6. The Lithuanian population is characterized by a "mirror reflection effect," where high negative relative fitness ($S < -16$) amino acids are converted to neutral (0) or low (up to -5) relative fitness amino acids after *dn*M, thus confirming the theory of neutrality.

REFERENCES

1. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014.

2. Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, et al. The Genome of the Netherlands: Design, and project goals. Eur J Hum Genet. 2014.

3. Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, et al. The 1000 Genomes Pproject: Data management and community access. Nature Methods. 2012.

4. Ségurel L, Wyman MJ, Przeworski M. Determinants of Mutation Rate Variation in the Human Germline. Annu Rev Genomics Hum Genet. 2014.

5. Korona DA, Lecompte KG, Pursell ZF. The high fidelity and unique error signature of human DNA polymerase ε. Nucleic Acids Res. 2011.

6. Schmitt MW, Matsumoto Y, Loeb LA. High fidelity and lesion bypass capability of human DNA polymerase δ. Biochimie. 2009.

7. Kunkel TA, Erie DA. Eukaryotic Mismatch Repair in Relation to DNA Replication. Annu Rev Genet. 2015.

8. Oliver C, Woodcock K, Adams D. The importance of aetiology of intellectual disability. In: Learning disability: a life cycle approach to valuing people. 2010.

9. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012.

10. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform. 2013.

11. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010.

12. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2016.

13. Bernstein B, Birney E, Dunham I, Green E, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012.

14. Francioli LC. Inherited and de novo variation in human genomes. 2015.

15. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci. 2014.

16. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, et al. Genetic effects on gene expression across human tissues. Nature. 2017.

17. The GTEx Consortium, Ardlie KG, Deluca DS, Segre A V., Sullivan TJ, Young TR, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-). 2015.

18. Tamuri AU, dos Reis M, Goldstein RA. Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. Genetics. 2012.

19. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: Homology modelling of protein structures and complexes. Nucleic Acids Res. 2018.

20. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: Quality-

controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017.

21. Depristo MA, Banks E, Poplin R, Garimella K V., Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011.

22. Warden CD, Adamson AW, Neuhausen SL, Wu X. Detailed comparison of two popular variant calling packages for exome and targeted exon studies. PeerJ. 2014.

23. Zuckerkandl E, Pauling L. Molecular Disease, Evolution and genic heterogeneity. In: Horizons in biochemistry. 1962.

24. Kumar S. Molecular clocks: Four decades of evolution. Nature Reviews Genetics. 2005.

25. Bromham L, Penny D. The modern molecular clock. Nature Reviews Genetics. 2003.

26. Takahata N, Kimura M. The Neutral Theory of Molecular Evolution. Princ Med Biol. 1994.

27. JUKES TH, CANTOR CR. Evolution of Protein Molecules. In: Mammalian Protein Metabolism. 2013.

28. Birky CW, Walsh JB. Effects of linkage on rates of molecular evolution. Proc Natl Acad Sci. 1988.

29. Kumar S, Subramanian S. Mutation rates in mammalian genomes. Proc Natl Acad Sci. 2002.

30. Kondrashov FA, Kondrashov AS. Measurements of spontaneous rates of mutations in the recent past and the near future. Philosophical Transactions of the Royal Society B: Biological Sciences. 2010.

31. Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. Genetics. 2000.

32. Ciochon RL, White JL. Walter C. Hartwig (ed.). 2002. The Primate Fossil Record. Cambridge University Press, Cambridge, 503 p., cloth, ISBN 0-521-66315-6. J Paleontol. 2004.

33. Steiper ME, Young NM. Timing primate evolution: Lessons from the discordance between molecular and paleontological estimates. Evol Anthropol. 2008.

34. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, et al. Patterns and rates of exonic de novo mutations in autism

spectrum disorders. Nature. 2012.

35. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, et al. Genome-wide patterns and properties of de novo mutations in humans. Nat Genet. 2015.

36. Subramanian S, Kumar S. Neutral substitutions occur at a faster rate in exons than in noncoding DNA in primate genomes. Genome Res. 2003.

37. Schmidt S, Gerasimova A, Kondrashov FA, Adzuhbei IA, Kondrashov AS, Sunyaev S. Hypermutable non-synonymous sites are under stronger negative selection. PLoS Genet. 2008.

38. Branciamore S, Chen Z-X, Riggs AD, Rodin SN. CpG island clusters and pro-epigenetic selection for CpGs in protein-coding exons of HOX and other transcription factors. Proc Natl Acad Sci. 2010.

39. Sakabe NJ, Savic D, Nobrega MA. Transcriptional enhancers in development and disease. Genome Biology. 2012.

40. Weirauch MT, Hughes TR. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends in Genetics. 2010.

41. Villar D, Flicek P, Odom DT. Evolution of transcription factor binding in metazoans-mechanisms and functional implications. Nature Reviews Genetics. 2014.

42. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014.

43. Necsulea A, Kaessmann H. Evolutionary dynamics of coding and non-coding transcriptomes. Nature Reviews Genetics. 2014.

44. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, et al. Enhancer evolution across 20 mammalian species. Cell. 2015.

45. Cotney J, Leng J, Yin J, Reilly SK, Demare LE, Emera D, et al. XThe evolution of lineage-specific regulatory activities in the human embryonic limb. Cell. 2013.

46. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father-s age to disease risk. Nature. 2012.

47. Crow JF. The origins, patterns and implications of human

spontaneous mutation. Nature Reviews Genetics. 2000.

48. Wong WSW, Solomon BD, Bodian DL, Kothiyal P, Eley G, Huddleston KC, et al. New observations on maternal age effect on germline de novo mutations. Nat Commun. 2016.

49. Besenbacher S, Liu S, Izarzugaza JMG, Grove J, Belling K, Bork-Jensen J, et al. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. Nat Commun. 2015.

50. Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, et al. Timing, rates and spectra of human germline mutation. Nat Genet. 2016.

51. Bustamante CD, Wakeley J, Sawyer S, Hartl DL. Directional selection and the site-frequency spectrum. Genetics. 2001.

52. Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D. Quantifying the slightly deleterious mutation model of molecular evolution. Mol Biol Evol. 2002.

53. Katju V, Bergthorsson U. Old trade, new tricks: Insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with high-throughput genomic approaches. Genome Biol Evol. 2019.

54. Fay JC, Wyckoff GJ, Wu CI. Positive and negative selection on the human genome. Genetics. 2001.

55. Uricchio LH, Petrov DA, Enard D. Exploiting selection at linked sites to infer the rate and strength of adaptation. Nat Ecol Evol. 2019.

56. Lelieveld SH, Reijnders MRF, Pfundt R, Yntema HG, Kamsteeg E-J, de Vries P, et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nat Neurosci. 2016.

57. Cormack RM, Hartl DL, Clark AG. Principles of Population Genetics. Biometrics. 2006.

58. Keightley PD. Rates and fitness consequences of new mutations in humans. Genetics. 2012.

59. Szamecz B, Boross G, Kalapis D, Kovács K, Fekete G, Farkas Z, et al. The Genomic Landscape of Compensatory Evolution. PLoS Biol. 2014.

60. Bartha I, Rausell A, McLaren PJ, Mohammadi P, Tardaguila

M, Chaturvedi N, et al. The Characteristics of Heterozygous Protein Truncating Variants in the Human Genome. PLoS Comput Biol. 2015.

61. Anna A, Monika G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. Journal of Applied Genetics. 2018.

62. Li XS, Trojer P, Matsumura T, Treisman JE, Tanese N. Mammalian SWI/SNF-A Subunit BAF250/ARID1 Is an E3 Ubiquitin Ligase That Targets Histone H2B. Mol Cell Biol. 2010.

63. Yan Z, Wang Z, Sharova L, Sharov AA, Ling C, Piao Y, et al. BAF250B-Associated SWI/SNF Chromatin-Remodeling Complex Is Required to Maintain Undifferentiated Mouse Embryonic Stem Cells. Stem Cells. 2008.

64. Boyer LA, Tong IL, Cole MF, Johnstone SE, Levine SS, Zucker JP, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell. 2005.

65. Flores-Alcantar A, Gonzalez-Sandoval A, Escalante-Alcalde D, Lomelí H. Dynamics of expression of ARID1A and ARID1B subunits in mouse embryos and in cells during the cell cycle. Cell Tissue Res. 2011.

66. Gadzicki D, Döcker B, Schubach M, Menzel M, Schmorl B SF et al. Expanding the phenotype of a recurrent de novo variant in PACS1 causing intellectual disability. Clin Genet. 2015;88:300–2.

67. Wan L, Molloy SS, Thomas L, Liu G, Xiang Y, Rybak SL, et al. PACS-1 defines a novel gene family of cytosolic sorting proteins required for trans-Golgi network localization. Cell. 1998.

68. Youker RT, Shinde U, Day R, Thomas G. At the crossroads of homoeostasis and disease: roles of the PACS proteins in membrane traffic and apoptosis. Biochem J. 2009.

69. Schuurs-Hoeijmakers JHM, Oh EC, Vissers LELM, Swinkels MEM, Gilissen C, Willemsen MA, et al. Recurrent de novo mutations in PACS1 cause defective cranial-neural-crest migration and define a recognizable intellectual-disability syndrome. Am J Hum Genet. 2012.

70. Sanlaville D, Etchevers HC, Gonzales M, Martinovic J, Clément-Ziza M, Delezoide AL, et al. Phenotypic spectrum of CHARGE syndrome in fetuses with CHD7 truncating mutations

correlates with expression during human development. J Med Genet. 2006.

71. Kageyama R, Nakanishi S. Helix-loop-helix factors in growth and differentiation of the vertebrate nervous system. Curr Opin Genet Dev. 1997.

72. Soosaar A, Chiaramello A, Zuber MX, Neuman T. Expression of basic-helix-loop-helix transcription factor ME2 during brain development and in the regions of neuronal plasticity in the adult brain. Mol Brain Res. 1994.

73. Brockschmidt A, Todt U, Ryu S, Hoischen A, Landwehr C, Birnbaum S, et al. Severe mental retardation with breathing abnormalities (Pitt - Hopkins syndrome) is caused by haploinsufficiency of the neuronal bHLH transcription factor TCF4. Hum Mol Genet. 2007.

74. Zhuang Y, Cheng P, Weintraub H. B-lymphocyte development is regulated by the combined dosage of three basic helix-loop-helix genes, E2A, E2-2, and HEB. Mol Cell Biol. 1996.

75. Schwartzman JS, Velloso R de L, D'Antino MEF, Santos S. The eye-tracking of social stimuli in patients with Rett syndrome and autism spectrum disorders: a pilot study. Arq Neuropsiquiatr. 2017.

76. Ricciardi S, Boggio EM, Grosso S, Lonetti G, Forlani G, Stefanelli G, et al. Reduced AKT/mTOR signaling and protein synthesis dysregulation in a Rett syndrome animal model. Hum Mol Genet. 2011.

77. Gonzales ML, Adams S, Dunaway KW, LaSalle JM. Phosphorylation of Distinct Sites in MeCP2 Modifies Cofactor Associations and the Dynamics of Transcriptional Regulation. Mol Cell Biol. 2012.

78. Long SW, Ooi JYY, Yau PM, Jones PL. A brain-derived MeCP2 complex supports a role for MeCP2 in RNA processing. Biosci Rep. 2010.

79. Young JI, Hong EP, Castle JC, Crespo-Barreto J, Bowman AB, Rose MF, et al. Regulation of RNA splicing by the methylation-dependent transcriptional repressor methyl-CpG binding protein 2. Proc Natl Acad Sci U S A. 2005.

CURRICULUM VITAE

Laura Pranckėnienė was born on December 25, 1990 in Vilnius, Lithuania. In 2009, she graduated from the Jonas Basanavičius school. In 2013, she graduated with a bachelor's degree in biochemistry from Vilnius University, Faculty of Chemistry; In 2015, she graduated with a master's degree in medical genetics from Vilnius University, Institute of Biomedical Science, Department of Human and Medical Genetics, and began her PhD in medicine under the supervision of Prof. Habil Dr. Vaidutis Kučinskas. She extended her knowledge and skills in the field of bioinformatic analysis of genome sequencing during her traineeship on January 21–31, 2017 at the University Medical Center Groningen (UMCG) in the Netherlands. During 2015–2017 she worked as a medical geneticicist at the Center of Medical Genetics, Vilnius University Hospital Santaros Clinics. Since 2017, Laura is a junior research associate in the Dept. of Human and Medical Genetics, Institute of Biomedical Sciences, Faculty of Medicine, Vilnius University.

LIST OF PUBLICATIONS

Published articles:

1. **Pranckėnienė** L, Kučinskas V. Trumpų *de novo* iškritų ir intarpų intesnyvumo įvertinimas skirtinguose žmogaus genomo regionuose. *Laboratorinė medicina*, 2016; 18, No. 3 (71), p. 118–122.

2. **Pranckėnienė L**, Jakaitienė A, Ambrozaitytė L, Kavaliauskienė I, Kučinskas V. Insights Into *de novo* Mutation Variation in Lithuanian Exome. *Frontiers in Genetics*. 2018;9:315. doi:10.3389/fgene.2018.00315. Impact factor – 4,151.

3. **Pranckėnienė L**, Preikšaitienė E, Gueneau L, Reymond A, Kučinskas V. *De novo* duplication in the CHD7 gene associated with severe CHARGE syndrome. *Genomics Insights*. 2019. Factor H – 5.

4. **Pranckėnienė L**, Bumbulienė Ž, Dasevičius D, Utkus A, Kučinskas V, Preikšaitienė E. Novel Androgen Receptor gene variant containing a premature termination codon in a patient with androgen insensitivity syndrome. *J Pediatr Adolesc Gynecol*. 2019 Aug 8. pii: S1083-3188(19)30257-8. doi: 10.1016/j.jpag.2019.08.001.

Poster presentations:

1. **Pranckėnienė L**, Kučinskas V. Trumpų *de novo* iškritų ir intarpų intensyvumo įvertinimas skirtinguose žmogaus genomo regionuose. *Laboratorinė medicina*, 2016; 18, No. 3 (71), p. 118–122.

2. **Pranckėnienė L**, Jakaitienė A, Ambrozaitytė L, Kavaliauskienė I, Kučinskas V. Insights Into *de novo* Mutation Variation in Lithuanian Exome. *Frontiers in Genetics*. 2018;9:315. doi:10.3389/fgene.2018.00315. Impact factor – 4,151.

3. **Pranckėnienė L**, Preikšaitienė E, Gueneau L, Reymond A, Kučinskas V. *De novo* duplication in the CHD7 gene associated with severe CHARGE syndrome. *Genomics Insights*. 2019. Factor H – 5.

Oral presentations:

1. Pranckėnienė L, Kučinskas V. An evaluation of *de novo* mutation process intensity in the different regions of human genome. International Conference "Evolutionary Medicine: Pre-Existing Mechanisms and Patterns of Current Health Issues," 14–17 06 2016, Vilnius, Lithuania.

2. **Pranckėnienė L**, Jakaitienė A, Ambrozaitytė L, Kučinskas V. *De novo* mutacijų ir jų intensyvumo analizė lietuvių triosų grupėje. 07 12 2017, LSA Conference of Young Scientists "Bioateitis." Diploma for Best Oral presentation, 3rd place. Vilnius, Lithuania.

3. **Pranckėnienė L.** Distribution and rates of exonic *de novo* mutations in patients with intellectual disability. 10–12 05 2018, Baltic Congress in Laboratory Medicine (BALM), Vilnius, Lithuania.

4. **Pranckėnienė L**, Jakaitienė A, Kučinskas V. *De novo* mutacijų ir jų intensyvumo analizė lietuvių triosų grupėje. 14 12 2018, LSA Conference of Young Scientists “Bioateitis.” Diploma for Best Oral presentation, 3rd place. Vilnius, Lithuania.

Traineeship:

21–31 01 2017. The University Medical Center Groningen (UMCG), the Netherlands.

Vilniaus universiteto leidykla
Saulėtekio al. 9, LT-10222 Vilnius
El. p. info@leidykla.vu.lt,
www.leidykla.vu.lt
Tiražas 25 egz.