VILNIUS UNIVERSITY

TAUTVYDAS RANČELIS

# DIVERSITY ANALYSIS OF PATHOGENIC GENOMIC VARIANTS AND GENES THAT CAUSE THE AUTOSOMAL RECESSIVE DISEASES USING WHOLE EXOME SEQUENCING

Summary of doctoral dissertation

Biomedical science, medicine (06 B)

Vilnius, 2016

This doctoral dissertation research was prepared at the Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University in 2012-2016.

**Scientific supervisor** – **prof. habil. dr. Vaidutis Kučinskas** (Vilnius University, biomedical sciences, biology – 01 B).

**Scientific consultant** – **prof. dr. Loreta Cimbalistienė** (Vilnius University, biomedical sciences, medicine – 06 B).

The defence of the dissertation will be held at Vilnius University Scientific Council on Medicine:

**Chairman:**

**prof. habil. dr. Kęstutis Sasnauskas** (Vilnius University, biomedical sciences, biology – 01 B).

**Members:**

1) **dr. Renata Posmyk** (Medical University of Bialystok, Poland, biomedical sciences, medicine – 06 B);

2) **prof. habil. dr. Limas Kupčinskas** (Lithuanian University of Health Sciences, biomedical sciences, medicine – 06 B);

3) **prof. (HP) Arvydas Kaminskas** (Vilnius University, biomedical sciences, medicine – 06 B);

4) **prof. dr. Janina Tutkuvienė** (Vilnius University, biomedical sciences, medicine 06 B).

The doctoral dissertation will be defended at the open session of the Council of Medical Sciences at 14:00 on 24 November, 2016 in the Red audience of Vilnius University Hospital Santariškių Clinics.

Address: Santariškių str. 2, LT-08661, Vilnius, Lithuania.

The dissertation is available at the library of Vilnius University and on the website: *www.vu.lt/lt/naujienos/ivykiu-kalendorius*

The summary of the doctoral dissertation was distributed on 24 October, 2016.

VILNIAUS UNIVERSITETAS

TAUTVYDAS RANČELIS

# PATOGENINIŲ GENOMO VARIANTŲ IR JŲ GENŲ, LEMIANČIŲ AUTOSOMINES RECESYVIĄSIAS LIGAS, ĮVAIROVĖS ANALIZĖ, PANAUDOJANT VISO EGZOMO SEKOSKAITĄ

Daktaro disertacija

Biomedicinos mokslai, medicina (06 B)

Vilnius, 2016

Disertacija rengta 2012 – 2016 metais Vilniaus universiteto Medicinos fakulteto Žmogaus ir medicininės genetikos katedroje.

**Mokslinis vadovas – prof. habil. dr. Vaidutis Kučinskas** (Vilniaus universitetas, biomedicinos mokslai, biologija - 01 B).

**Mokslinis konsultantas – prof. dr. Loreta Cimbalistienė** (Vilniaus universitetas, biomedicinos mokslai, medicina – 06 B).

Disertacija bus ginama Vilniaus universiteto Medicinos mokslo krypties taryboje:

**Pirmininkas – prof. habil. dr. Kęstutis Sasnauskas** (Vilniaus universitetas, biomedicinos mokslai, biologija - 01 B).

**Nariai:**

**dr. Renata Posmyk** (Balstogės medicinos universitetas, Lenkija, biomedicinos mokslai, medicina - 06 B);

**prof. habil. dr. Limas Kupčinskas** (Lietuvos sveikatos mokslų universitetas, biomedicinos mokslai, medicina - 06 B);

**prof. dr. Arvydas Kaminskas** (Vilniaus universitetas, biomedicinos mokslai, medicina - 06 B);

**prof. dr. Janina Tutkuvienė** (Vilniaus universitetas, biomedicinos mokslai, medicina – 06 B).

Disertacija bus ginama viešame Medicinos mokslo krypties tarybos pasėdyje 2016 m. lapkričio 24 d. 14 val. Santariškių klinikų Auditorijų korpuso Raudonojoje auditorijoje.

Adresas - Santariškių g. 2, Vilnius, Lietuva.

Su disertacija galima susipažinti Vilniaus universiteto bibliotekoje ir VU interneto svetainėje adresu: *www.vu.lt/lt/naujienos/ivykiu-kalendorius*

Disertacijos santrauka išsiuntinėta 2016 m. spalio 24 d.

## INTRODUCTION

Analysis of whole exome (all protein coding DNA sequences) is valuable diagnostic approach for discovering the genetic basis of many diseases. After 2009, when Sarah Ng and co-workers successfully applied Whole Exome Sequencing (WES) to causal study of Miller syndrome, WES is widely used for finding causative genome variants of a long list diseases. Despite whole exome comprising approximately only 1 % of the genome, it is involved in determining of about 85 % of known Mendelian diseases (Majewski et al., 2011; Gilissen et al., 2012).

Stargart disease, pigmented retinitis, Usher, Perrault, Joubert syndromes – are only a few diseases for which genomic causes were identified by WES (Pierce et al., 2010; Becker et al., 2011; Zuchner et al., 2011; Cabral et al., 2012; Eisenberger et al., 2012; Strom et al., 2012; Srour et al., 2015). .

The dissertation "Diversity analysis of pathogenic genomic variants and genes that cause the autosomal recessive diseases using whole exome sequencing" is related to several important human genetic areas:

1) next generation sequencing (NGS) and its use in research of human genetics and practice of medicine;

2) a population-based research for the identifying pathogenic variant's frequency in healthy individuals;

3) the use of genomic data of healthy individuals to improve diagnostics of genetic diseases;

4) the use of bioinformatical methods in modern human genetics.

In the research of human genetic diseases NGS is used for finding genomic cause of particular disease in patients. However, the research of pathogenic variants could be carried out not only in sick individuals, but also in healthy ones, because every healthy individual is considered to be a carrier of some pathogenic genomic variants (Xue et al., 2012; MacArthur et al., 2012; Lazarin et al., 2013). The study of diseases in healthy individuals is enabled by the published studies of other researchers, where particular genomic causes of diseases were identified.

In this dissertation NGS and data from ClinVar database, which gather known pathogenic genome variants, were applied to find pathogenic variants in exome of healthy individuals from Lithuanian population. Pathogenic variants frequencies in Lithuanians were compared with the data from other populations.

While NGS has tremendous benefit, it requires a proper analysis pipeline. Due to this reason, this thesis gives a particular attention to solving bioinformatical problems.

**Aim of the study**

Aim of the study is to identify and assess pathogenic genome variants in Lithuanian population using the whole exome sequencing.

**Tasks of the research**

1. To select and perform NGS analysis pipeline, which would be the most appropriate for the analysis of the primary *SOLiD* sequencing data.

2. To find in Lithuanian population exome data found genomic variants that are known to be pathogenic.

3. To use the information stored in *ClinVar* database.

4. To determine frequency differences of pathogenic variants in Lithuanian population and in other populations.

5. To perform comparison of intra-populational distribution of pathogenic variants in Lithuania's ethnolinguistic groups.

**The statements to be defended**

1. Frequencies of widespread pathogenic variants in Lithuanian population differ in comparison to frequencies of pathogenic variants in other populations.

2. Frequencies of pathogenic variants differ in the intra-populational level as well – there are some pathogenic variants observed only in Aukštaičiai, or only in Žemaičiai.

3. In the studied Lithuanian population group, the majority of the identified pathogenic variants were associated with metabolic disorders.

4. *ClinVar* database contains a high number of genomic variants that are considered as pathogenic, but actually are not pathogenic.

**Scientific novelty of the study**

The study of pathogenic variants in healthy individuals of Lithuanian population, using next generation sequencing, is the first of this kind in Lithuania.

**Practical and scientific significance**

The relevance of this thesis on Lithuanian scale is due to Lithuanian exomes' data usage. It is the first study of such kind, providing new knowledge about the occurrence of pathogenic variants among healthy Lithuanians.

While progress of this study in the world is growing rapidly, this thesis is relevant also on global scale, as on the ground of research data, it evaluates widely used *ClinVar* database.

The study of the genomic data of Lithuanian people reveals the presence of pathogenic variants, with statistically significantly higher frequency in Lithuanian population, in comparison with other populations, acquired from the 1000 genome project and ExAC database. This knowledge enables us to use such data for diagnostic purposes – allowing the creation of special pathogenic variants panel for Lithuanian population, or even promoting incorporation of more frequent hereditary diseases into the newborn screening of Lithuanian population.

**The presentation and approbation of the results**

The results were published in two local Lithuanian scientific journals and in two international publications, one of them is in the ISI Web of Science Journal List and the other is in the ISI Web of Science Journal List of journals with impact factor. The results of the research were presented at two international and two Lithuanian conferences.

**The structure and volume of the dissertation**

The dissertation contains the following chapters: List of acronyms, Introduction, List of publications and Conference abstracts, Review of the literature, Material and methods, Results, Discussion, Conclusions, References (116), Acknowledgements, and Supplement. The volume of dissertation is 115 pages. The dissertation is illustrated with 10 figures and 16 tables. The dissertation is written in Lithuanian with a summary in English.

## MATERIAL AND METHODS

To achieve the objectives of the dissertation, the data from the project "Genetic diversity of the population of Lithuania and changes of its genetic structure related with evolution and common diseases" (acronym: LITGEN) was used. The LITGEN project was performed at the Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University. The project manager was prof. habil. dr. Vaidutis Kučinskas.

During the LITGEN project venous blood and questionnaire data were collected from 1 000 individuals (279 trios). All of the researched individuals were Lithuanians, with at least three generations born in Lithuania. The obtained data was divided into six ethnolinguistic groups of Lithuania: south Aukštaitija, east Aukštaitija, west Aukštaitija, south Žemaitija, north Žemaitija, west Žemaitija (Uktverytė et al., 2013; Uktverytė , 2014).

The samples were collected by these researchers: dr. Neringa Burokienė, Raimonda Meškienė, Aidas Pranculis, dr. Laima Ambrozaitytė, dr. Ingrida Domarkienė and dr. Ingrida Kavaliauskienė.

The sequenced exome data of the LITGEN project was researched in this thesis. From total 144 sequenced exoms (in trios), the data of 96 individuals was used (excluding data from probands).

### Sequencing

The 5500 SOLiD™ System (Applied Biosystems; Thermo Fisher Scientific, Inc., USA) was used to sequence the samples. Sequencing was carried out according to the manufacturer's protocols (Thermo Fisher Scientific, Inc., USA) using the Agilent SureSelect$^{XT}$ Target Enrichment System or the Life Technologies TargetSeq™ Exome Enrichment System. Using the 5500 SOLiD™ System, 75-bp short-read sequences were generated.

Optimized sequencing protocols were implemented and carried out by dr. Ingrida Kavaliauskienė, dr. Laima Ambrozaitytė, dr. Ingrida Domarkienė. NGS data verification by Sanger sequencing was carried out by Kristina Aleknavičienė.

**Bioinformatic analysis**

The SOLiD™ System uses a specific ligation-based sequencing strategy and color-space encoding.

For NGS analysis proprietary software for the SOLiD™ System LifeScopeTM 2.5.1 genomic analysis software was used, but to check if analysis pipeline was applied appropriately and to get more results, the alternative NGS analysis pipeline was used (Figure 1) (Pranckevičienė et al., 2015).

To check if during NGS analysis pipeline acquired results were correct and not random, the rate of transitions (Ti)/ transversions (Tv) was obtained. In human genome this rate is 2.1, and in its exome this rate is around 2.8. In the studied data Ti/Tv ratio was 2.2 - 2.8, which indicates that obtained data is not random.

After alignment step in NGS analysis pipeline, the coverage was verified as it is one of the most important factors in the following variant calling phase. It is recommended that 85 % of the targeted regions should be covered at least 20 × for population based studies. In our study 80 % of target regions were covered 20 ×.

Additionally, to ensure that results from NGS were correct, the Sanger sequencing for NGS called variants was applied.

For the overview of genomic variance in Lithuanian population, the Genome Analysis Toolkit's (GATK) CombineVariants software was used to pool the .vcf files of all 96 individuals (DePristo et al., 2011). Functional annotation of genomic variants was performed using Annovar v.2014nov12. This program annotated frequencies of genomic variants from the 1000G and ExAC databases and pathogenic variants from the ClinVar database to studied data (Wang et al., 2010).

The Integrative Genomics Viewer (IGV) was used for visualization of the data (Thorvaldsdóttir et al., 2013).

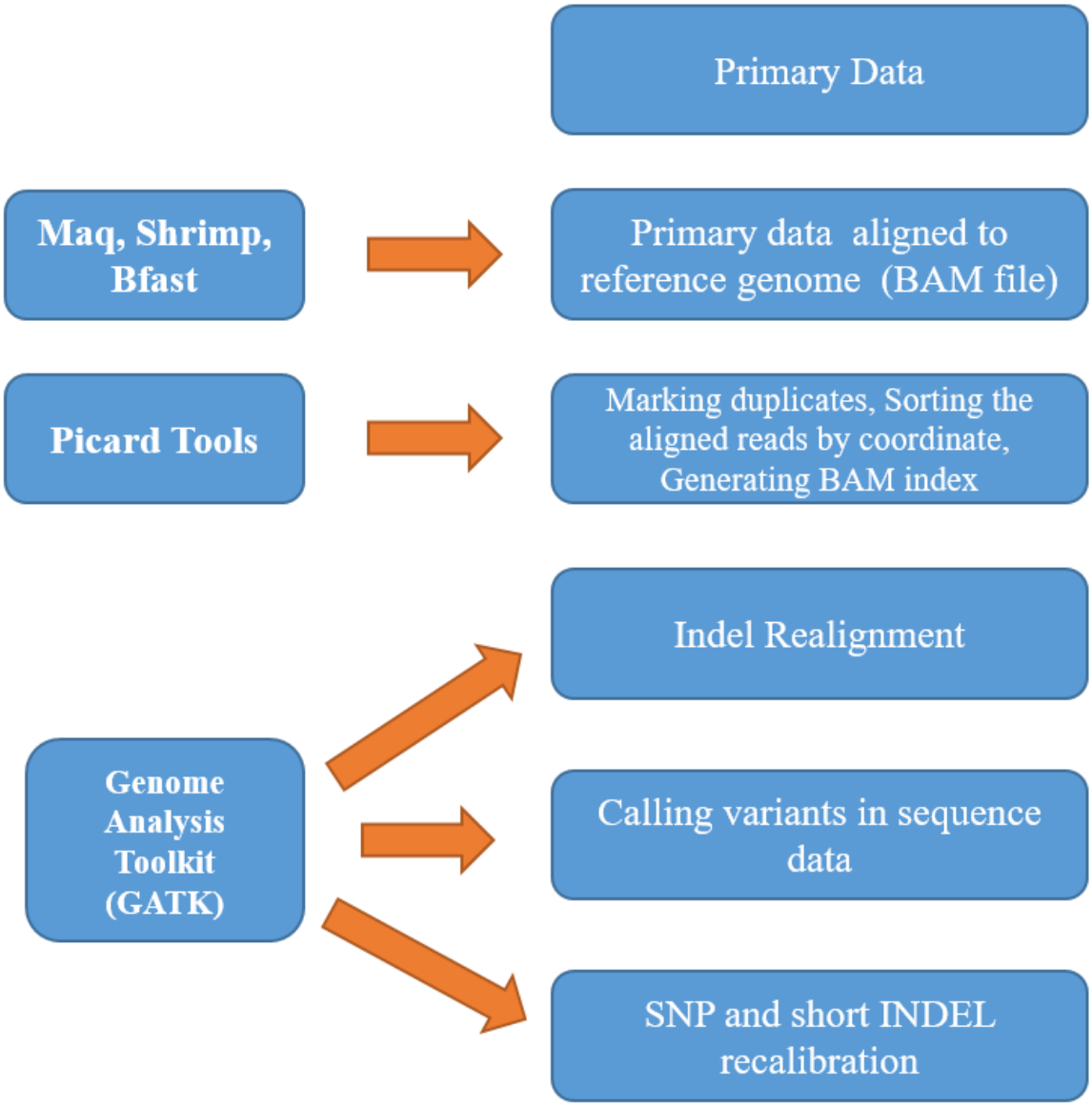


Fig. 1 Alternative to Lifescope program NGS pipeline.

**Statistics**

In the comparative frequency analysis of pathogenic variants in the Lithuanian population study group and other population data (1000G, ExAC), Fisher's exact test and the chi-squared test were used. Statistically significant differences were considered with $p < 0.05$.

## RESULTS

### Pathogenic variants in exomes data

Exome sequencing of each individual exome allowed to identify an average of 44,000 SNVs and 2300 short INDELs, that differed from the reference genome (hg19). Over 275 thousand substitutions were determined in all 96 individuals.

An average of 45 SNVs per individual exome were indicated as pathogenic by the *ClinVar* database. Three hundred and twenty-one unique SNVs and 30 unique short INDELs were classified as pathogenic in all 96 exomes (Table 1).

**Table 1.** Pathogenic variants from general Lithuanian population statistics according to their frequency.

| **Total amount of variants in 96 exomes** | **SNV** | **Short INDEL** |
|---|---|---|
| Frequency in 1000G and ExAC < 50% | 301 | 30 |
| Intronic variants | 13 | 2 |
| Splicing variants | 6 | – |
| Exonic variants | 282 | 28 |
| Nonsynonymous, frameshift variants | 268 | 25 |
| Synonymous, nonframeshift variants | 14 | 3 |
| Frequency in 1000G and ExAC ≤ 1% | 150 | 24 |
| Intronic variants | 2 | 0 |
| Splicing variants | 6 | – |
| Exonic variants | 141 | 24 |
| Nonsynonymous, frameshift variants | 138 | 23 |
| Synonymous, nonframeshift variants | 3 | 1 |

Frequency comparison with other population data from ExAC and 1000G projects was performed. Such comparison with global ExAC data (from 60,706 individuals) showed that 103 pathogenic variants in this study statistically significantly differed from global ExAC data.

To acquire more Lithuanian-specific pathogenic variants, comparison with 1000G European data was performed for all assigned pathogenic variants.

From unique 30 pathogenic INDELs, 25 appeared only in 1 or 2 individuals, which does not allow to calculate statistical significance and separate them for random appearance. The rest 5 INDELs demonstrated no statistically significant difference from 1000G European data.

The comparison of the SNVs identified in the present study with the 1000G European data showed 22 statistically significant varying pathogenic variants: 14 of them had statistically significantly higher frequencies and 8 variants had statistically significantly lower frequencies.

**Identification of pathogenic genomic variants causing recessive diseases among Lithuanians**

The majority of apparently healthy individuals, examined among the LITGEN project data, had pathogenic variants, inherited as autosomal recessive. With this inheritance, individuals, having heterozygous genotype, are healthy.

Among pathogenic variants, having autosomal recessive inheritance, 10 genomic variants had statistically significantly higher (Table 2) and 5 variants had statistically significantly lower (Table 3) frequencies than in other European populations.

Notably there is a pathogenic genomic variant that is associated with severe combined immunodeficiency disease, caused by lack of *DCLRE1C* gene product. This variant occurred even among 10 of 96 individuals, in all cases having heterozygous genotype.

During additional Sanger sequencing of individuals, by using NGS, a pathogenic variant, related to the severe combined immunodeficiency disease, was found. It proved that in all individuals this variant was determined correctly. This pathogenic variant is important because the severe combined immunodeficiency disease in USA is already included into the newborn screening program. As its frequency in Lithuanian population is higher in comparison with global and European population, this variant could be perspective candidate for the newborn screening in Lithuania.

**Table 2.** dbSNP database codes of pathogenic variants which have autosomal recessive inheritance and statistically significantly higher frequencies in the self-reported healthy Lithuanian individuals in comparison with data from the 1000G and ExAC projects.

| Genomic variant dbSNP 142 | Gene | Disease | 1000G EUR, % | LT, % | 1000G EUR chi | 1000G EUR Fisher |
|---|---|---|---|---|---|---|
| **rs1800553** | *ABCA4* | Stargardt disease | 0.40 | 2.55 | $1.419 \times 10^{-3}$ | $8.006 \times 10^{-3}$ |
| **rs115532916** | *ACAD9* | Deficiency of Acyl-CoA dehydrogenase | 2.19 | 7.14 | $2.6454 \times 10^{-4}$ | $9.75 \times 10^{-4}$ |
| **rs142181517** | *PHYKPL* | Phosphohydroxy-lysinuria | 0.40 | 3.57 | $2.0602 \times 10^{-5}$ | $5.2093 \times 10^{-4}$ |
| **rs41297018** | *DCLRE1C* | Severe combined immunodeficiency disease | 2.19 | 5.10 | 0.0241 | 0.04919 |
| **rs41298135** | *MYO7A* | Usher syndrome | 0.40 | 2.04 | $9.8591 \times 10^{-3}$ | 0.0283 |
| **rs45495503** | *EPB42* | Spherocytosis | 0.20 | 2.04 | $8.2718 \times 10^{-4}$ | $7.9094 \times 10^{-3}$ |
| **rs113298164** | *LIPC* | Hepatic lipase deficiency | 0.50 | 2.04 | 0.0224 | 0.0447 |
| **rs34526199** | *AMPD1* | Muscle AMP deaminase deficiency | 3.68 | 7.14 | 0.0380 | 0.053 |
| **rs2229738** | *CPT1A* | Carnitine palmitoyltransferase I deficiency | 8.65 | 16.33 | 0.044 | 0.0519 |
| **rs104895094** | *MEFV* | Familial Mediterranean Fever | 0.55 | 6.63 | $3.2363 \times 10^{-9}$ | $1.1564 \times 10^{-6}$ |

**Table 3.** dbSNP database codes of pathogenic variants which have autosomal recessive inheritance and statistically significantly lower frequencies in the self-reported healthy Lithuanian individuals in comparison with data from the 1000G and ExAC projects.

| Genomic variant dbSNP 142 | Gene | Disease | 1000G EUR, % | LT, % | 1000G EUR chi | 1000G EUR Fisher |
|---|---|---|---|---|---|---|
| rs1736557 | *FMO3* | Trimethylaminuria | 6.26 | 2.55 | 0.0264 | 0.0296 |
| rs11887534 | *ABCG8* | Gallbladder disease | 7.95 | 2.04 | 0.0006 | 0.0001 |
| rs45547231 | *RAPSN* | Congenital myasthenic syndrome | 15.01 | 5.10 | $1.7077 \times 10^{-6}$ | $1.2823 \times 10^{-7}$ |
| rs2269219 | *FECH* | Erythropoietic protoporphyria | 20.08 | 17.35 | $2.096 \times 10^{-3}$ | $1.773 \times 10^{-3}$ |
| rs12483377 | *COL18A1* | Knobloch syndrome | 8.35 | 4.59 | 0.01 | 0.009 |

The dominant pathogenic variants were also revealed among the LITGEN project data. They were divided into two groups: frequencies in Lithuanian population that are higher, as in 1000 G project (Table 4), or lower, as in1000 G project (Table 5).

**Table** 4. dbSNP database codes of pathogenic variants, which have autosomal dominant inheritance and statistically significantly higher frequencies in the self-reported healthy Lithuanian individuals in comparison with data from the 1000G and ExAC projects.

| Genomic variant dbSNP 142 | Gene | Disease | 1000G EUR, % | LT, % | 1000G EUR chi | 1000G EUR Fisher |
|---|---|---|---|---|---|---|
| rs76308115 | *PDE11A* | Primary pigmented nodular adrenocortical disease | 0.50 | 2.04 | 0.0224 | 0.0447 |
| rs121917710 | *GLI3* | Polydactyly | 0.70 | 3.57 | $6.4 \times 10^{-4}$ | $3.53 \times 10^{-3}$ |
| rs77724903 | *RET* | Familial medullary thyroid carcinoma | 0.10 | 2.04 | $1.1281 \times 10^{-4}$ | $3.0128 \times 10^{-3}$ |
| rs41549716 | *POLG* | Progressive external ophthalmoplegia | 0.70 | 3.06 | 0.0036 | 0.0113 |

An interesting pathogenic variant is rs76308115 C>T(c.169C>T (p.Arg57Ter)) in *PDE11A* gene, causing primary pigmented nodular adrenocortical disease, which has similar symptoms to Cushing syndrome (Horvath et al., 2006). It is rare in the global population, but has relatively high (2 %) frequency in Lithuanian population. This frequency is statistically higher in comparison with global or European populations. Interesting part is, that this variant occurred only among individuals from north Žemaitija.

**Table** 5. dbSNP database codes of pathogenic variants which have autosomal dominant inheritance and statistically significantly lower frequencies in the self-reported healthy Lithuanian individuals in comparison with data from the 1000G and ExAC projects.

| **Genomic variant *dbSNP* 142** | **Gene** | **Disease** | **1000G EUR, %** | **LT, %** | **1000G EUR chi** | **1000G EUR Fisher** |
|---|---|---|---|---|---|---|
| rs1805124 | *SCN5A* | Progressive familial heart block, type 1A | 21.67 | 13.78 | $1.4802 \times 10^{-5}$ | $7.5158 \times 10^{-6}$ |
| rs28757581 | *OR2J3* | Ability to smell cut grasses | 11.93 | 6.63 | $7.662 \times 10^{-3}$ | $7.765 \times 10^{-3}$ |
| rs3749977 | *OR2J3* | Ability to smell cut grasses | 22.86 | 17.86 | $9.5588 \times 10^{-5}$ | $6.6306 \times 10^{-5}$ |

## Pathogenic recessive genome variants in ethnolinguistic groups

The 321 unique pathogenic SNV are distributed almost equally among six ethnolinguistic groups, at an average 150 pathogenic genome variants in each group (Table 6).

Among ethnolinguistic groups there was no pathogenic variant that reliably occurred only in one of the groups. However, several variants that were found are specific only to Žemaičiai, or to Aukštaičiai.

**Table 6**. Number of pathogenic genomic variants in each ethnolinguistic group according to frequencies of different genomic variants obtained from 1000g, ExAC data.

| SA | EA | WA | SŽ | NŽ | WŽ | Total number of genomic variants |
|---|---|---|---|---|---|---|
| **Total number of identified pathogenic genomic variants** | | | | | | |
| **158** (49.22 %)* | 148 (46.11 %) | 158 (49.22 %) | 158 (49.22 %) | 168 (52.34 %) | 155 (48.29 %) | 321 |
| **When the genomic variant frequency in 1000G, ExAC data is less than 50%** | | | | | | |
| **140 (46.51 %)** | 131 (43.52 %) | 141 (46.84%) | 141 (49.22 %) | 150 (49.83 %) | 138 (45.85 %) | 301 |
| **When the genomic variant frequency in 1000G, ExAC data is less than 25 %** | | | | | | |
| **119 (42.5 %)** | 111 (39.64 %) | 122 (43.57%) | 121 (46.84 %) | 130 (46.43 %) | 118 (42.14 %) | 280 |
| **When the genomic variant frequency in 1000G, ExAC data is less than 1 %** | | | | | | |
| **33 (22.0 %)** | 28 (18.67%) | 38 (25.33 %) | 38 (25.33 %) | 47 (31.33 %) | 34 (22. 67%) | 150 |

*numbers in brackets indicate the relative frequency

As it was mentioned above, the dominant pathogenic variant rs76308115 C>T(c.169C>T (p.Arg57Ter)) in *PDE11A* gene was found only in 4 individuals from north Žemaitija.

Additionally to it, several recessive pathogenic genome variants also were specific for one ethnolinguistic group. Substitution rs142181517 A>T (c.1310A>T (p.Glu437Val)) was found only among Žemaičiai. That pathogenic variant causes phosphohydroxylysinuria. This disorder occures in 1 of 16000 individuals (Veiga-da-Cunha et al., 2013), while among 96 persons studied even 7 individuals were carriers of that disorder: 4 from west Žemaitija, 2 from north Žemaitija, 1 from south Žemaitija. Substitution rs1799807 A>G (c.293A>G (p.Asp98Gly)) in *BCHE* gene was also found only among Žemaičiai, changing the structure and the function of butyrylcholinesterase. This pathogenic variant is important pharmacologically, because the post-anaesthetic apnoea could develop for persons who have it.

Autosomal recessive variant rs1800562 G>A (c.845G>A (p.Cys282Tyr)) was revealed only among Aukštaičiai, causing hereditary hemochromatosis that increases the level of iron in the blood. This pathogenic genome variant is an exclusive feature of

Europeans (Pedersen and Milman, 2009). Among Lithuanians this pathogenic variant was found in all Aukštaičiai ethnolinguistic groups.

Unusually high frequency was determined for substitution rs104895094 *A>G (c.2084 A>G (p. Lys695Arg))* in *MEFV* gene among east and south Aukštaičiai. It caused familial Mediterranean fever, also known as Armenian disease, a hereditary inflammatory disorder (Bernot et al., 1998). Spreading of this substitution among Lithuanians in south regions of Lithuania may suggest about formerly migration from Mediterranean regions.

**Errors in ClinVar database**

According to Bell et al. (2011), about 10 % of disease-causing mutations depicted in widely used databases are misinterpreted, and such databases should be carefully scrutinized.

In the dataset of this study, as many as 90 genomic variants, that were indicated as pathogenic by ClinVar, had 50 % and higher frequencies in the 1000G and ExAC projects. Pathogenic variants with such high frequency in the global population are highly questionable, and therefore these variants were excluded from further analysis. Examples of genome variants that have high frequency in global population are shown in Table 7.

Of all genomic variants indicated as pathogenic, 233 had a frequency lower than 50 %, and 124 had a frequency lower or equal to 1 % in the 1000G and ExAC projects.

**Table 7.** Examples of pathogenic genomic variants from research data, which are identified as pathogenic, but have a very high frequency in general population.

| Gene | *dbSNP* | 1000G World data | 1000G Europe data | ExAC Europe | LITGEN |
|---|---|---|---|---|---|
| *DPYD* | rs1801265 | 73.98 % | 78.53 % | 76.52 % | 72.45 % |
| *DBT* | rs12021720 | 89.18 % | 92.15 % | 91.38 % | 92.86 % |
| *NOS3* | rs1799983 | 82.37 % | 65.61 % | 75.30 % | 70.41 % |
| *HPD* | rs1154510 | 87.64 % | 87.18 % | 85.02 % | 84.18 % |
| *BBS2* | rs4784677 | 99.64 % | 99.11 % | 99.38 % | 100 % |
| *PRODH* | rs450046 | 90.56 % | 92.84 % | 91.71 % | 63.78 % |

Another debatable point is that there were individuals who had homozygous genotypes for alleles identified as pathogenic by ClinVar, meaning that these individuals

may have disease symptoms. Since the data in this study is acquired from self-reported healthy Lithuanian individuals, possible explanations are that the pathogenic variant causes a very subtle alteration, that there was an error in this study's NGS data, that the phenotype was not determined in detail, or that the variant is incorrectly attributed as pathogenic. For some pathogenic variants, this homozygous state is seen both in our data and in the data of large-scale population studies.

Several of proposed dominant pathogenic variants are clearly annotation artefacts in database. Among the LITGEN project data such artefacts were found for three genomic variants – two in *MYO1A* gene: rs33962952 G>A and rs55679042 G>A, and one in *MYH14* gene, all attributed in ClinVar database as causing deafness, but detail studies showed that they are not pathogenic (Mikštienė et al., 2016). The same is for substitution rs119103280G>T (c.1150G>T (p.Gly384Cys)) in *MYH14* gene. It is also doubtful that there is pathogenicity of substitution rs140926412 C>T (c.796C>T (p.Arg266Cys)) in *GDF3* gene, in database proposed as causing rare Klippel-Feil syndrome. The disease has clearly expressed strong phenotype, and its pathogenicity is scarcely probable in healthy individuals. Only if there is a weak phenotype gene effect or low penetrance, interaction protein-protein takes place or gene redundancy is involved, individual may be healthy despite pathogenic variant in this genome.

## DISCUSSION

The results of the study represent statistically significant differences in frequencies of genomic variants between individuals from Lithuanian population and other populations.

When the data of the present study was compared with all ExAC project data, a statistically significant difference was observed in 103 of 321 pathogenic variants, but most of them correlated with the frequencies of European data. The comparison of Lithuanian data with European data of the 1000 Genome Project revealed 15 statistically significant pathogenic variants that differed from European population data.

The study showed that whole exome sequencing and analysis of the general population is an effective way to find pathogenic variants with statistically significant

differences in a particular population even if cohort studied is relatively small. This could be valuable information for genetic counselling and may facilitate focusing on the diagnosis of specific variants that are more frequent in a particular population.

Based on the present study data, *ClinVar* is currently the best freely available database of genomic variants of different clinical significance. However, a considerable amount of the variants classified as pathogenic in *ClinVar* have a high frequency in the 1000G and ExAC and may be considered as false pathogenic variants. A similar pattern we observed in LITGEN data.

Another matter of concern is, that there were individuals who had homozygous genotypes for alleles identified as pathogenic. This case should be investigated further. Cautious interpretation of *ClinVar* data for pathogenic variants should be undertaken by researchers and medical specialists.

## CONCLUSIONS

1. Comparison of NGS analysis pipeline programs showed that *Lifescope*$^{TM}$ program is best suited for analysis of *SOLiD* sequencing data.

2. Each person in average had 45 SNV genomic variants that were identified as pathogenic. In each person occurred from zero to three INDEL, identified as pathogenic. In all studied individuals 321 unique SNV and 30 unique INDEL were identified as pathogenic. In the majority, the greater part of pathogenic variants studied in Lithuanian population group was associated with metabolic disorders.

3. The database of *ClinVar* pathogenic variants contains genomic variants, named as pathogenic, that have extremely high frequencies both in Lithuanian and in the global samples' data.

4. The comparison of pathogenic variant frequencies in studied individuals with frequencies of pathogenic variants from the genomic data of Europeans allowed to identify 15 pathogenic genomic variants, that are inherited in autosomal recessive manner and statistically significantly differ from other European populations, 10 of them had a higher frequency and 5 had a lower frequency than the frequencies of Europeans.

There were determined 103 pathogenic variants, that differed statistically from the global frequencies of respective genomic variants (in comparison with 1000G and ExAC data).

5. The 321 pathogenic SNV were distributed about equally in ethnolinguistic groups, on average 150 pathogenic variants for each ethnolinguistic group. Two pathogenic variants rs142181517 A>T and rs1799807 A>G of studied data were found only in Žemaitija ethnolinguistic groups. One pathogenic variant rs1800562 G>A was found only in Aukštaitija ethnolinguistic groups.

6. The studies of pathogenic variants in healthy individuals could have significant practical benefits for identification and diagnosis of diseases. Among the identified pathogenic variants there was variant rs41297018:G>A in *DCLRE1* gene, causing severe combined immunodeficiency disease, which had statistically significantly higher frequency in comparison to the global and European data, and which is included into the newborn screening in other countries.

**ACKNOWLEDGEMENTS**

First of all, I would like to express my thank to my scientific supervisor **acad. prof. habil. dr. Vaidutis Kučinskas** for the opportunity to do this work, suggestions, patience, concern, for the opportunity to internship in Estonia and Switzerland. Also I would like to thank my scientific consultant **prof. (HP) dr. Loreta Cimbalistienė** for comments and consultations in clinical issues.

This thesis would not have appeared, if not of the contribution from many colleagues to whom I am very grateful. Especially I would like to thank to the entire of LITGEN project, to manager of project **acad. prof. habil. dr. Vaidutis Kučinskas** and to executors which data were used in this thesis. Many thanks dedicated **dr. Laima Ambrozaitytė**, **dr. Ingrida Domarkienė**, **dr. Ingrida Kavaliauskienė**, **dr. Neringa Burokienė, Raimonda Meškienė**, **Aidas Pranculis** for their work of collecting used samples.

Background of this thesis are Lithuanian exomes, which would not be possible to study without lab work - DNA extraction, sequencing. Therefore, I would like to thank to **Daiva Kazlauskaitė**, **Dalytė Pliaugo, dr. Ingrida Kavaliauskienė**, **dr. Laima Ambrozaitytė**, **dr**. **Ingrida Domarkienė** and **Kristina Aleknavičienė**.

An important part of the thesis is the next-generation sequencing analysis pipeline. My sincerely thank is to **dr. Erinija Prackevičienė** for the given instructions and knowledge, teamwork and for the ability to enthuse me into bioinformatics and programming. My thank is to **Aidas Pranculis** for our shared work with NGS pipeline. I would like to thank **Justas Arasimavičius** for the advices on statistical matters. I am also thankful to **Violeta Mikštienė** for our common work and for giving opportunity for deeper analysis of single disability among the Lithuanians.

For the assistance, valuable advices that was given during the writing of the thesis, I am very grateful to **acad. habil prof. dr. Zita Aušrelė Kučinskienė**, **prof. dr. (HP) Audrone Jakaitienė**, **dr. Aušra Morkūnienė**, **dr. Aušra Matulevičienė**, **dr. Alma Molytė**, **dr. Laima Ambrozaitytė** and **prof. (HP) dr. Algirdas Utku**s.

Finally, I want to thank to all colleagues from Department of Human and Medical Genetics and family for their support and assistance.

## REFERENCES

1. Becker, J., O. Semler, C. Gilissen, Y. Li, H. J. Bolz, C. Giunta, C. Bergmann, M. Rohrbach, F. Koerber, K. Zimmermann, P. de Vries, B. Wirth, E. Schoenau, B. Wollnik, J. A. Veltman, A. Hoischen, C. Netzer. 2011. 'Exome sequencing identifies truncating mutations in human SERPINF1 in autosomal-recessive osteogenesis imperfecta', Am J Hum Genet, 88: 362-71.

2. Bell, C. J., D. L. Dinwiddie, N. A. Miller, S. L. Hateley, E. E. Ganusova, J. Mudge, R. J. Langley, L. Zhang, C. C. Lee, F. D. Schilkey, V. Sheth, J. E. Woodward, H. E. Peckham, G. P. Schroth, R. W. Kim, S. F. Kingsmore. 2011. 'Carrier testing for severe childhood recessive diseases by next-generation sequencing', Sci Transl Med, 3: 65ra4.

3. Biesecker, L. G., and R. C. Green. 2014. 'Diagnostic clinical genome and exome sequencing', N Engl J Med, 370: 2418-25.

4. Bernot, A., C. da Silva, J. L. Petit, C. Cruaud, C. Caloustian, V. Castet, M. Ahmed-Arab, C. Dross, M. Dupont, D. Cattan, N. Smaoui, C. Dodé, C. Pêcheux, B.Nédelec, J.Medaxian, M.Rozenbaum, I.Rosner, M.Delpech, G.Grateau, J.Demaille, J. Weissenbach, I.Touitou. 1998. 'Non-founder mutations in the MEFV gene establish this gene as the cause of familial Mediterranean fever (FMF)', Hum Mol Genet.,7:1317-25.

5. Cabral, R. M., M. Kurban, M. Wajid, Y. Shimomura, L. Petukhova, A. M. Christiano. 2012. 'Whole-exome sequencing in a single proband reveals a mutation in the CHST8 gene in autosomal recessive peeling skin syndrome', Genomics, 99: 202-8.

6. Donaudy, F., A. Ferrara, L. Esposito, R. Hertzano, O. Ben-David, R.E. Bell, S. Melchionda, L. Zelante, K. B. Avraham, P. Gasparini. 2003. Multiple mutations of MYO1A, a cochlear-expressed gene, in sensorineural hearing loss, Am. J Hum Genet., 72:1571-7.

7. DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly. 2011.

'A framework for variation discovery and genotyping using next-generation DNA sequencing data', Nature Genet, 43: 491-8.

8. Eisenberger, T., R. Slim, A. Mansour, M. Nauck, G. Nurnberg, P. Nurnberg, C. Decker, C. Dafinger, I. Ebermann, C. Bergmann, H. J. Bolz. 2012. 'Targeted next-generation sequencing identifies a homozygous nonsense mutation in ABHD12, the gene underlying PHARC, in a family clinically diagnosed with Usher syndrome type 3', Orphanet J Rare Dis, 7: 59.

9. Eisenberger, T., N. Di Donato, S. M. Baig, C. Neuhaus, A. Beyer, E. Decker, D.Mürbe, C. Decker, C. Bergmann, H. J. Bolz. 2014. 'Targeted and genomewide NGS data disqualify mutations in MYO1A, the "DFNA48 gene", as a cause of deafness', Hum Mutat., 35:565-70.

10. Gilissen, C., A. Hoischen, H. G. Brunner, J. A. Veltman. 2012. 'Disease gene identification strategies for exome sequencing', Eur J Hum Genet, 20: 490-7.

11. Horvath, A, Boikos S, Giatzakis C, Robinson-White A, Groussin L, Griffin KJ, Stein E, Levine E, Delimpasi G, Hsiao HP, Keil M, Heyerdahl S, Matyakhina L, Libè R, Fratticci A, Kirschner LS, Cramer K, Gaillard RC, Bertagna X, Carney JA, Bertherat J, Bossis I, Stratakis CA. 2006. 'A genome-wide scan identifies mutations in the gene encoding phosphodiesterase 11A4 (PDE11A) in individuals with adrenocortical hyperplasia', Nauret Genet, 38(7):794-800.

12. Lazarin, G. A., I. S. Haque, S. Nazareth, K. Iori, A. S. Patterson, J. L. Jacobson, J. R. Marshall, W. K. Seltzer, P. Patrizio, E. A. Evans, B. S. Srinivasan. 2013. 'An empirical estimate of carrier frequencies for 400+ causal Mendelian variants: results from an ethnically diverse clinical sample of 23,453 individuals', Genet Med, 15: 178-86.

13. MacArthur, D. G., S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M. M. Suner, T. Hunt, I. H. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero,

Consortium Genomes Project, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, C. Tyler-Smith. 2012. 'A systematic survey of loss-of-function variants in human protein-coding genes', Science, 335: 823-8.

14. MacArthur, D. G., and C. Tyler-Smith. 2010. 'Loss-of-function variants in the genomes of healthy humans', Hum Mol Genet, 19: R125-30.

15. Majewski, J., J. Schwartzentruber, E. Lalonde, A. Montpetit, N. Jabado. 2011. 'What can exome sequencing do for you?', J Med Genet, 48: 580-9.

16. Mikštienė, V., A. Jakaitienė, J. Byčkova, E. Gradauskienė, E. Preikšaitienė, B. Burnytė, B. Tumienė, A. Matulevičienė, L. Ambrozaitytė, I. Uktverytė, I. Domarkienė, T. Rančelis, L. Cimbalistienė, E. Lesinskas, V. Kučinskas, A. Utkus. 2016.'The high frequency of GJB2 gene mutation c.313_326del14 suggests its possible origin in ancestors of Lithuanian population'. BMC Genetics, 19, 17(1): 45.

17. Ng, S. B., K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, M. J. Bamshad. 2010. 'Exome sequencing identifies the cause of a mendelian disorder', Nature Genet, 42: 30-5.

18. Pedersen, P., and N.Milman. 2009. 'Genetic screening for HFE hemochromatosis in 6,020 Danish men: penetrance of C282Y, H63D, and S65C variants', Ann Hematol., 88:775-84.

19. Pierce, S. B., T. Walsh, K. M. Chisholm, M. K. Lee, A. M. Thornton, A. Fiumara, J. M. Opitz, E. Levy-Lahad, R. E. Klevit, M. C. King. 2010. 'Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome', Am J Hum Genet, 87: 282-8.

20. Pranckeviciene, E., T. Rancelis, A. Pranculis, V. Kucinskas. 2015. 'Challenges in exome analysis by LifeScope and its alternative computational pipelines', BMC Res Notes, 8: 421.

21. Srour, M., F. F. Hamdan, D. McKnight, E. Davis, H. Mandel, J. Schwartzentruber, B. Martin, L. Patry, C. Nassif, A. Dionne-Laporte, L. H. Ospina, E. Lemyre, C. Massicotte,

R. Laframboise, B. Maranda, D. Labuda, J. C. Decarie, F. Rypens, D. Goldsher, C. Fallet-Bianco, J. F. Soucy, A. M. Laberge, C. Maftei, Consortium Care4Rare Canada, K. Boycott, B. Brais, R. M. Boucher, G. A. Rouleau, N. Katsanis, J. Majewski, O. Elpeleg, M. K. Kukolich, S. Shalev, J. L. Michaud. 2015. 'Joubert Syndrome in French Canadians and Identification of Mutations in CEP104', Am J Hum Genet, 97: 744-53.

22. Strom, S. P., Y. Q. Gao, A. Martinez, C. Ortube, Z. Chen, S. F. Nelson, S. Nusinowitz, D. B. Farber, M. B. Gorin. 2012. 'Molecular diagnosis of putative Stargardt Disease probands by exome sequencing', BMC Med Genet, 13: 67.

23. Thorvaldsdottir, H., J. T. Robinson, J. P. Mesirov. 2013. 'Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration', Brief Bioinform, 14: 178-92.

24. Uktverytė, I., R. Meškienė, L. Ambrozaitytė, I. Domarkienė, A. Pranculis, N. Burokienė, A. Coj, A. Mažeikienė, V. Kasiulevičius, Z. A. Kučinskienė, V. Kučinskas. 2013. LITGEN - revealing genetic structure of the population of Lithuania. Europ J Human Genet: European Human Genetics Conference 2013, Paris, France, June 8-11.

25. Uktverytė, I. 2014. „Lietuvos etnolingvistinių grupių genetinės struktūros analizė remiantis informatyviais genomo žymenimis”: Daktaro dis. biomedicinos mokslai:medicina (06 B).

26. Veiga-da-Cunha M, Verhoeven-Duif NM, de Koning TJ, Duran M, Dorland B, Van Schaftingen E. 2013. 'Mutations in the AGXT2L2 gene cause phosphohydroxylysinuria', J Inherit Metab Dis., 36:961-6

27. Wang, Q., Q.Lu, H. Zhao. 2015. 'A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing', Front Genet, 6:149.

28. Xue, Y., Y. Chen, Q. Ayub, N. Huang, E. V. Ball, M. Mort, A. D. Phillips, K. Shaw, P. D. Stenson, D. N. Cooper, C. Tyler-Smith, and Consortium Genomes Project. 2012. 'Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing', Am J Hum Genet, 91: 1022-32.

29. Ye, M., K. M. Berry-Wynne, M. Asai-Coakwell, P. Sundaresan, T. Footz, C. R. French, M. Abitbol, V. C. Fleisch, N. Corbett, W. T. Allison, G. Drummond, M. A. Walter, T. M. Underhill, A.J. Waskiewicz, O. J. Lehmann. 2010. 'Mutation of the bone morphogenetic protein GDF3 causes ocular and skeletal anomalies', Hum Mol Genet, 19:287-98.

## LIST OF PUBLICATIONS

1) **Tautvydas Rančelis**, Loreta Cimbalistienė, Vaidutis Kučinskas. Next-generation whole-exome sequencing contribution to identification of rare autosomal recessive diseases. Acta Medica Lituanica, 2013, Vol. 20, N. 1. p. 43-51.

2) **Tautvydas Rančelis**, Erinija Pranckevičienė, Vaidutis Kučinskas. Anotaciniai įrankiai ir kompiuterinės programos genomo/egzomo duomenų analizei. Laboratorinė medicina, 2013, T. 15, N. 4(60), p. 206-212.

3) Erinija Pranckevičienė, **Tautvydas Rančelis**, Aidas Pranculis, Vaidutis Kučinskas. Challenges in exome analysis by LifeScope and its alternative computational pipelines. BMC Research Notes, 2015, 8:1.

4) Violeta Mikštienė, Audronė Jakaitienė, Jekaterina Byčkova, Eglė Gradauskienė, Eglė Preikšaitienė, Birutė Burnytė, Birutė Tumienė, Aušra Matulevičienė, Laima Ambrozaitytė, Ingrida Uktverytė, Ingrida Domarkienė, **Tautvydas Rančelis**, Loreta Cimbalistienė, Eugenijus Lesinskas, Vaidutis Kučinskas, Algirdas Utkus. The high frequency of *GJB2* gene mutation c.313_326del14 suggests its possible origin in ancestors of Lithuanian population. BMC Genetics, 2016, 19;17(1):45. Epub 2016, doi: 10.1186/s12863-016-0354-9.

## CONFERENCE PRESENTATIONS

1. Local:
    1) **T. Rančelis** „Žinomų patogeninių variantų anotacija bendroje Lietuvos populiacijoje“. 2015.03.06 konferencija „Lietuvos populiacijos genetinė įvairovė ir sandaros kitimai, susiję su evoliucija ir dažniausiai paplitusiomis ligomis“.
    2) **T. Rančelis** „Žalingų genomo variantų nustatymas bei identifikavimas panaudojant viso egzomo sekoskaitą“. 2015.12.06 Jaunųjų mokslininkų konferencija „BIOATEITIS: gamtos ir gyvybės mokslų perspektyvos“.
2. International:
    1. Abstract and Poster: **T. Rančelis**, E. Pranckevičienė, A. Pranculis, V. Kučinskas, Comparison of SOLiD sequencing data analysis pipelines. In European Journal of Human Genetics. London, Nature Publishing Group. ISSN 1018-4813. Vol. 23, Supplement 1. 2015. p. 314-314.
    2. Abstract and Poster: V. Kučinskas, **T. Rančelis**, I. Domarkienė, E. Pranckevičienė, I. Uktverytė, L. Ambrozaitytė. Profile of pathogenic alleles in healthy Lithuanian population / 65th Annual Meeting of The American Society of Human Genetics, October 6-10, 2015, Baltimore MD, poster abstracts. Baltimore, The American Society of Human Genetics, 2015, p. 483.

## SANTRAUKA

Disertacija „Patogeninių genomo variantų ir jų genų, lemiančių autosomines recesyviąsias ligas, įvairovės analizė, panaudojant viso egzomo sekoskaitą“ susijusi su keletu šiuo metu žmogaus genetikai aktualių sričių:

1) naujos kartos sekoskaita (NKS) ir jos naudojimu žmogaus genetikos tyrimuose ir medicinos praktikoje,

2) populiacinio pobūdžio tyrimais, kuriais nustatomas paveldimas ligas lemiančių genomo variantų dažnis sveikuose asmenyse,

3) sveikų asmenų genetinių duomenų naudojimu genetinių ligų diagnostikai pagerinti,

4) bioinformacinių metodų taikymu šiuolaikinėje žmogaus genetikoje.

Tiriant žmogaus genetines ligas pasitelkus NKS, gali būti tiriami sergantys asmenys – nustatomi genomo variantai, lemiantys paveldimas ligas. Tačiau dėl genetinių ligų gali būti tiriami ne tik sergantys, bet ir sveiki asmenys, nes kiekvienas sveiku laikomas asmuo gali būti iki 400 patogeninių genomo variantų nešiotojas (Xue et al., 2012; MacArthur et al., 2012; Lazarin et al., 2013). Patogeninių genomo variantų tyrimą sveikuose asmenyse palengvina kitų tyrėjų paskelbti darbai, kuriuose yra nurodyti konkretūs patogeniniai genomo variantai.

Šioje disertacijoje, panaudojant NKS ir specializuotą duomenų bazę, kurioje saugoma iki šiol mokslininkų sukaupta informacija apie patogeninius genomo variantus, ieškota, ar sveiki lietuvių populiacijos asmenys turi patogeninių genomo variantų, atliktas jų dažnio palyginamas su kitų populiacijų duomenimis.

Nors didelės apimties sekoskaitos nauda neabejotina, sekoskaita reikalauja tinkamai atlikti analizės algoritmą, todėl disertacijoje bioinformacinėms problemoms spręsti yra skiriama itin daug dėmesio.

**Darbo naujumas, aktualumas ir reikšmė**

Tyrimas, ar sveiki asmenys turi patogeninių genomo variantų, panaudojant naujos kartos sekoskaitą, yra pirmas tokio pobūdžio Lietuvoje.

Disertacinis darbas yra aktualus Lietuvos mastu, nes lietuvių egzomų duomenys pirmą kartą naudojami tokio pobūdžio tyrimuose ir suteikiama naujų žinių apie patogeninių variantų paplitimą tarp sveikų lietuvių.

Nors šioje srityje pažanga pasaulyje vyksta itin sparčiai, atlikto darbo pobūdis yra aktualus ir pasauliniu mastu, nes, remiantis tyrimo duomenimis, įvertinama *ClinVar* duomenų bazė.

Darbe analizuojami asmenų iš bendros lietuvių populiacijos genomo duomenys atskleidžia, kurie patogeniniai genomo variantai statistiškai patikimai dažniau pasitaiko lietuvių populiacijoje, palyginti su kitų populiacijų duomenimis iš 1000 genomų ir ExAC duomenų bazių. Tai leidžia tokius tyrimo duomenis panaudoti diagnostikai – sudaryti specialią lietuvių populiacijai skirtą patogeninių variantų lentelę ir gali paskatinti lietuvių populiacijoje dažnesnę paveldimą ligą įtraukti į visuotinę naujagimių patikrą.

**Tyrimo tikslas** – nustatyti ir įvertinti lietuvių populiacijoje esančius patogeninius genomo variantus panaudojant viso egzomo sekoskaitą.

**Tyrimo tikslui įgyvendinti iškelti uždaviniai:**

1. Parinkti ir atlikti NKS analizės algoritmą, kuris būtų tinkamiausias analizuojant turimus pirminius *SOLiD* sekoskaitos duomenis.
2. Nustatyti lietuvių populiacijoje esančius genomo variantus, kurie yra žinomi kaip patogeniniai.
3. Kaip patogeninių variantų šaltinį naudoti *ClinVar* duomenų bazę.
4. Nustatyti patogeninių genomo variantų dažnių skirtumus, lyginant lietuvių populiaciją su kitomis populiacijomis.
5. Atlikti vidupopuliacinį patogeninių genomo variantų palyginimą pagal lietuvių etnolingvistines grupes.

**Ginamieji teiginiai**

1. Lietuvių populiacijoje paplitusių patogeninių genomo variantų dažnis skiriasi nuo patogeninių genomo variantų dažnio kitose populiacijose.
2. Patogeninių genomo variantų dažnių skirtumas yra ir vidupopuliaciniu lygmeniu – yra tik aukštaičiams ir žemaičiams būdingų patogeninių genomo variantų.
3. Tirtoje lietuvių populiacijoje didžioji dalis identifikuotų patogeninių genomo variantų yra susiję su metabolizmo sutrikimais.
4. *ClinVar* duomenų bazėje yra daug patogeniniais įvardytų genomo variantų, kurie nėra patogeniški.

## IŠVADOS

1. Atliktas NKS analizės algoritmo programų lyginimas parodė, kad *Lifescope*$^{TM}$ programa geriausiai tinka *SOLiD* sekoskaitos duomenims analizuoti.
2. Kiekvienas asmuo vidutiniškai turėjo 45 VNV genomo variantus, kurie įvardyti kaip patogeniniai. Patogeniniais įvardytų iškritų/intarpų pasitaikydavo nuo nulio iki trijų vienam asmeniui. Tarp 96 tirtų asmenų patogeniniais įvardytas 321 unikalus VNV ir 30 unikalių iškritų/ intarpų. Tirtoje lietuvių populiacijoje didžioji dalis identifikuotų patogeninių genomo variantų buvo susiję su metabolizmo sutrikimais.
3. *ClinVar* patogeninių genomo variantų duomenų bazėje aptikta patogeniniais įvardijamų genomo variantų, kurių dažnis itin didelis Lietuvos ir pasauliniuose duomenyse.
4. Lyginant tirtų asmenų patogeninių variantų dažnius su europiečių genomo variantų dažniais, identifikuota 15 patogeninių genomo variantų, kurie paveldimi autosominiu recesyviuoju būdu ir kurie statistiškai patikimai skyrėsi nuo kitų europiečių populiacijų – 10 iš jų turėjo didesnį dažnį, o 5 turėjo mažesnį dažnį nei europiečių dažniai. Nustatyti 103 patogeniniai variantai, kurie statistiškai patikimai skyrėsi nuo pasaulinių žmogaus genomo variantų dažnių (lyginant su 1000G, ExAC duomenimis).

5. Tarp etnolingvistinių grupių 321 patogeninis VNV pasiskirstė maždaug vienodai, vidutiniškai po 150 patogeninių genomo variantų kiekvienoje etnolingvistinėje grupėje. Identifikuoti du patogeniniai genomo variantai – rs142181517 A>T ir rs1799807 A>G, kurie buvo būdingi visoms Žemaitijos etnolingvistinėms grupėms ir nė vienai aukštaičių lingvistinei grupei. Išskirtinai aukštaičiams buvo būdingas vienas patogeninis genomo variantas rs1800562 G>A.
6. Patogeninių variantų, būdingų sveikiems asmenims, tyrimai gali turėti didelę praktinę naudą ligų diagnostikai ir prognozavimui. Tarp Lietuvos asmenims identifikuotų patogeninių variantų, turinčių statistiškai patikimai didesnį dažnį, palyginti su pasaulio arba Europos šių variantų dažniais, pasitaikė ir variantas rs41297018:G>A DCLRE1 gene, lemiantis sunkų kombinuotą imunodeficitą, kitų šalių praktikoje tiriamą visuotinės naujagimių patikros metu.

# CURRICULUM VITAE

## PERSONAL INFORMATION


| | |
|---|---|
| Name | Tautvydas Rančelis |
| Address | Department of Human and Medical Genetics,<br>Faculty of Medicine, Vilnius University<br>Santariškių street 2, LT-08661, Vilnius, Lithuania |
| Telephone | +37060176123 |
| E-mail | tautvydas.rancelis@mf.vu.lt |
| Citizenship | Republic of Lithuania |
| Date and place of birth | November 23,1987, Vilnius, Lithuania |


## WORK EXPERIENCE

From 27.06.2012 junior research associate at Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University

## EDUCATION AND TRAINING

2006 finished Vilnius Žemynos gymnasium

2006 – 2010 Vilnius University biology bachelor degree

2010 – 2012 Vilnius University genetics master degree

2012 – 2016 PhD student, Department of Human and Medical Genetics,Vilnius University

## SCIENTIFIC INTEREST

Human genetics, next generation sequencing, analysis of large scale data, annotation, rare disease.

# GYVENIMO APRAŠYMAS

## ASMENINĖ INFORMACIJA


| | |
|---|---|
| Vardas ir pavardė | Tautvydas Rančelis |
| Adresas | Žmogaus ir medicininės genetikos katedra |
| | Santariškių g. 2, LT-08661, Vilnius, Lietuva |
| Telefonas | +37060176123 |
| Elektroninis paštas | tautvydas.rancelis@mf.vu.lt |
| Pilietybė | Lietuvos Respublikos |
| Gimimo data ir vieta | 1987 m. lapkričio mėn. 23 d., Vilnius |


## DARBO PATIRTIS

Nuo 2012.06.27 Vilniaus universiteto Žmogaus ir medicininės genetikos jaunesnysis mokslo darbuotojas

## IŠSILAVINIMAS

2006 baigė Vilnius Žemynos gimnaziją

2006 – 2010 Biologijos bakalauras, Vilniaus universitetas

2010 – 2012 Genetikos magistras, Vilniaus universitetas

2012 – 2016 Vilniaus universiteto Žmogaus ir medicininės genetikos katedros doktorantas

## MOKSLINIAI INTERESAI

Žmogaus genetika, naujos kartos sekoskaita, didelės apimties duomenų analizė, anotacija, retos ligos.