MEDICALLY RELEVANT GENOME DIVERSITY IN UKRAINE

by

KHRYSTYNA SHCHUBELKA

A dissertation submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY IN BIOLOGICAL AND BIOMEDICAL SCIENCES

2023

Oakland University
Rochester, Michigan

Doctoral Advisory Committee:

Taras Oleksyk, Ph.D., Chair
Fabia Battistuzzi, Ph.D.
Thomas Raffel, Ph.D.

To my beloved country and family

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my mentor, Taras Oleksyk, who not only served as my academic advisor but also became a father figure during my time away from home. His constant support, guidance, trust, and mentorship have been invaluable to my personal and academic growth.

Also, I would like to thank my admired family who all went above and beyond to help me balance research work and care for my son.

I would also like to express my appreciation to Walter Wolfsberger, Stephanie Castro, and Kenneth Hilkert, my fellow graduate students. Together, we have worked on numerous projects, collaborating, and pushing each other to reach our true potential. Your dedication, friendship, and camaraderie have been instrumental in making my academic experience richer and more fulfilling.

Furthermore, I would like to thank all the faculty members at OU who have provided valuable guidance and expertise throughout my research endeavors. Your, insights, and willingness to share knowledge and resources have significantly contributed to the success of my projects. I am grateful for the opportunities to learn from each of you.

I would also like to express my heartfelt gratitude to the Armed Forces of Ukraine for their unbelievable efforts in protecting our families and collaborators in Ukraine. Your bravery and dedication inspire the world. I would also like to thank all the collaborators in Ukraine who have contributed to making many of our projects possible in these unprecedented times. Your hard work and stive to do quality science amidst the war have been a source of even more inspiration to me.

Finally, I want to extend my gratitude to the United States of America for providing me with the opportunity to pursue my studies with the best and learn from the best. I am forever grateful for the opportunities and experiences that this great nation has offered me.

Khrystyna Shchubelka

ABSTRACT

MEDICALLY RELEVANT
GENOME DIVERSITY IN UKRAINE
by

Khrystyna Shchubelka

Adviser: Taras Oleksyk, Ph.D.

Ukraine, the second-largest country in Europe, has a rich history characterized by migrations, epidemics, famines, wars, and occupations, all of which have contributed to the formation of the modern Ukrainian population. However, the genetic composition of Ukraine remains understudied. Previous population genetic studies have largely overlooked the unique genetic makeup of Ukraine due to lack of publicly available genomic data. Yet, recent global assessments of genome diversity have highlighted the presence of numerous endemic variants, underscoring the significance of investigating genetic variation in underrepresented regions like Eastern Europe and, specifically, Ukraine. Such research is vital in the context of worldwide clinical trials and the development of personalized medicine. Local populations, shaped by their distinct histories, harbor a wealth of unexplored genomic variation, which may impact drug response, influence the risk of common and rare diseases, and shape lifestyle adaptations. Comprehensively understanding the genetic makeup of specific populations facilitates more effective identification of genetic markers for disease gene mapping, including family linkage and genome-wide association studies. These investigations aim to uncover

susceptibility genes or loci for both Mendelian and complex diseases. Therefore, in this discovery-based study, our objective is to characterize the extensive genetic variation of medically relevant alleles in contemporary Ukraine across various levels of population structure: within major regions of Ukraine, and among the multiethnic population of Transcarpathia. Additionally, we will explore the relationship between lactose persistence genotypes and phenotypes, as well as the genetic factors underlying global developmental delay and intellectual disability in Ukraine.

TABLE OF CONTENTS

TABLE OF CONTENTS—Continued

TABLE OF CONTENTS—Continued

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GATK	Genome analysis toolkit
IRB	Institutional review board
GNOMAD	Genome Aggregation Database
LP	lactase persistence
GDD	Global Developmental Delay
ID	Intellectual Disability
WES	Whole Exome Sequencing
NGS	Next Generation Sequencing
AD	Autosomal Dominant
AR	Autosomal Recessive
XLD	X-Linked Dominant
XLR	X-Linked Recessive
CNVs	Copy Number Variants
NDD	Neurodevelopmental Disorder
HapMap	Famous worldwide populations surveys of genome-wide diversity
GWAS	Genome-wide association study
SNP	Single nucleotide polymorphism
UK10K project	United Kingdom 10,000 Genomes Project
FDA	Food and Drug Administration
OU	Oakland University
EDTA	Ethylenediaminetetraacetic acid
uL	microliter
DCEG	Division of Cancer Epidemiology and Genetics
BGI	Beijing Genomic Institute
LP	Lactase persistence
CI	Confidence interval
LNP	Lactose non-persistent
UV	Ultraviolet
HBT	Hydrogen breath test
CADD score	Combined Annotation Dependent Depletion score
SIFT score	Sorting Intolerant From Tolerant score
OMIM	Online Mendelian inheritance in Man database
TRC150	Ukrainians from Transcarpathia (current study)
UAU97	Ukrainians from Ukraine (current study)

CHAPTER ONE

MEDICALLY RELEVANT GENETIC DIVERSITY IN UKRAINE

1.1 Introduction

Genetic diversity in the modern Ukrainian population cannot be understood outside of the historic background of migrations, admixture, numerous wars and occupations, epidemics, and famines. Although, at present time ethnic Ukrainians constitute approximately three quarters of the total population of the modern Ukraine, the country is not ethnically uniform: a large minority of ethnic Russians compose around one-fifth of the total population mostly in the southeast, and smaller ethnic groups with quite sizeable minorities are in different parts of the country: Poles, Moldovans, Bulgarians, Belarusians, Hungarians, Romanians, Roma (Gypsies), Jews and others [1]. As people have moved and settled across this land, they have contributed unique genome variation to the existing population shaping its unique and distinct structure in Europe [2], but the full extent and the medical impact of it is currently unknown, partially because it hasn't been captured yet by next generation sequencing and reported on the genomic scale to date. Setting genomic data in the context of population history can help explain the patterns of variation observed today and help design effective local approaches in public health and clinical medicine[3]. Generally, the prehistoric and historic events that contributed to formation of genomic diversity in Ukrainian population today were mostly driven by its geographic location on the crossroads between Europe and Asia, at the edge of the great forests and the beginning of the great steppe[4].

Understanding population genetics in the context of past demographic events including founder effects, bottlenecks, migration, admixture and possibly episodes of

natural selection, can help explain the origin and distribution of rare and common medically related variants. For example, the mutation c.5266dupC in *BRCA1* gene was originally thought to occur mostly in Ashkenazi Jewish (AJ) population [5]. However, it has also been argued that it may have originated in the Scandinavian peninsula or in the north of what is today Russia, spread by Vikings to many countries of Europe, and only around 500 years ago was transferred into the AJ community in Poland where it rose in frequency due to the nonrandom mating and severe population bottlenecks that characterized Jewish population in the predominantly Christian Europe. Interestingly, this mutation is quite common in France, but uncommon in among the French-derived population of Quebec in Canada, due to the founder effect following the first establishment of the in the colonists' community [5]. In addition, studying genetic adaptations also showed that mutations causing same adaptive effect to arouse independently at the different loci of the genome in various human populations due to convergent evolution (for instance, lactose tolerance alleles in Europeans differ from those in African populations[6][7]; light skin color as an adaptation to effective vitamin D synthesis arose due to a different mutation in Europeans and East Asians[8]). Consequently, we see that both old and recent historic events can dramatically alter the spread of disease-causing variants in different populations[9].

While mutations are the ultimate source of all genetic variation, observed human genetic diversity is influenced by historic, demographic, and evolutionary forces, which leads to significant differences in their population frequencies among geographic locations [10]. Understanding the distribution of this heritable variation on the genome level can help us understand human history as well as the relative contribution of genomic versus the

environmental components in the distribution of population traits, including medically related heritable conditions, as well as susceptibility and resistance to cancer and infectious diseases [11].

Famous worldwide populations surveys of genome-wide diversity like “HapMap” and "1000 Genomes" project did not include any representative population from Eastern Europe [12]. This is due to a fact that systematic genome research has never been done in Eastern Europe. This apparent underrepresentation of certain ethnic populations creates an ascertainment bias in the clinical applications of the genome research - the GWAS results are frequently obtained by genotyping samples using SNPs arrays discovered in other populations [13]. To compensate for this bias, research groups in many countries have already launched their own local genome projects [14]. With the current era of genomics pushing down the cost of whole-genome sequencing (further WGS), a number of large-scale sequencing projects on hundreds of thousands of samples have recently been completed on whole genome sequencing, such as the UK10K project[15], and Genome Russia Project etc., which identified new important disease-causing genetic variation [15]. Thanks to a combination of the results from these studies, over 100 million SNPs are now cataloged in the dbSNP and GNOMAD (Genome Aggregation Database)[16] In the recent decades, great efforts have been made to accelerate the advent of precision medicine in local population[17]. Yet, given that there are many underrepresented populations, there are still many obstacles to overcome before the human genome variation is fully described at the local level.

In the process of variant discovery, which is the foundation of personalized medicine, genomic DNA is sequenced, and the resulting data are compared against a

reference sequence to catalogue differences[18]. The first challenge is to find the local medically related variants that have not been previously identified in the global surveys. During a population genomics project, large numbers of genome sequences from local populations are screened for both common and rare mutations[18]. These can be reported and used to design locus-specific genotyping essays that can augment the current commercial panels making them more useful for the local populations and eliminating the ascertainment bias[19].

There are two main practical uses of the variation captured from population specific genome sequences. First, the most obvious, and direct application of genomic screening for mutations is diagnostics of common and rare diseases, as well as cancer[20]. Diagnostic genotyping panels are developed based on whole genome sequences of local populations to include population-specific common and rare local mutations to save the cost and to make the testing available to the local patients [21]. On the other hand, the genome data can also be analyzed and used for the clinical research. In particular, basket trials [22] are a new and evolving form of clinical trial design and are predicated on the hypothesis that the presence of a molecular marker alone can predict response to a targeted therapy or disease outcome. The goal of a "basket study" is first to identify a mutation that causes a clinical response to a specific agent, and then to enroll all patients with that mutation into a clinical trial regardless of their disease type. Basket trials enroll a dozen subjects per disease type (e.g. cancer type), putting each type into its own separate category or "basket". Each disease type can then be then analyzed separately, as well as combined with other outcomes. A basket trial is particularly useful when the disease or the candidate mutation is rare, as is often the case with endemic diseases based on local mutations. These

are particularly difficult to plan for, due to the small number of target patients, that are recruited gradually over time. One of possible approaches to facilitate such a trial is to use a centralized bioinformatics resource: every time a patient with the rare type is diagnosed, his or her genetic profile is characterized, and the information is stored, so it can be recalled when a sufficient number of potential positive responders are identified population-wide.

Once relevant mutations are captured, a number of new research opportunities arise from the resulting data. For instance, in the clinical trials for cancer treatments, rather than consider the specific type of tissue corresponding to a particular cancer as the major determinant of the treatment, researchers may now consider mutations in different genes as a classifier [23]. More specifically, instead of crafting the same treatment for all lung cancer patients, treatments can be developed based on a specific mutation in a particular gene. In 2017 FDA approved the first immunotherapy drug "Keytruda®" that relied on genetic characteristics of the tumors rather than their anatomical location [24].

Because of the inequality of research funding, developed countries often can afford to have genome surveys in their populations to be studied first, creating a socioeconomic bias in the medical research, as those countries with known public genome databases tend to receive and enjoy the early benefits of advancing research, innovation and treatment options based on documented genetic variation[25]. Local catalogs of genome variation, along with the analytical tools for its interpretation and presentation to the researchers benefit the local populations by providing opportunities for medical research, collaboration, innovation and modern health treatment options by establishing a firm foundation for personalized medicine in the region[26]. Once the common and rare

diversity in a population in understood, researchers can start focusing on specific and endemic diseases that are confined in local areas within such populations[27].

Thus, ***my first objective*** is to annotate and to evaluate the type and frequencies of the known and new genomic variants in the population from different regions of Ukraine, with the focus on medically related alleles.

1.2 Methods

1.2.1 Sampling strategy and the informed consent

We have established a network of doctors and medical professionals at health care centers across Ukraine to coordinate collection of blood samples. Sampling centers were distributed across the country to facilitate capture of the genomic diversity across the entire population (**Figure 1.1**).

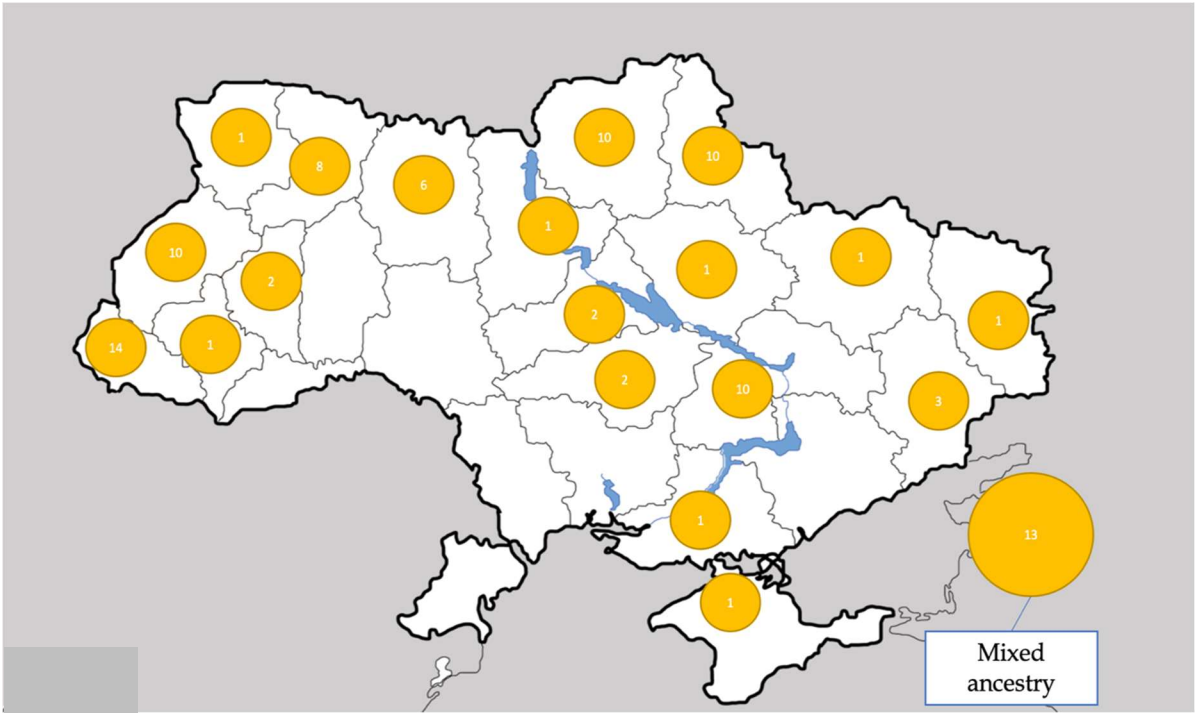

Figure 1.1. Location of sampling centers and numbers of samples in the initial effort to collect DNA to evaluate the frequency of the known medically related genomic variants from across the country of Ukraine. Mixed ancestry means the grandparents of the individual were born in at least two different regions of Ukraine.

According to the sampling protocol, non-hospitalized healthy volunteers were invited for interviews at outpatient departments of different clinics. Each volunteer received all necessary information regarding the study and the blood collection procedure and provided full consent to participate and have their genotypic and phenotypic data to be publicly available. The study participants also filled a questionnaire indicating their self-reported region of origin, place of birth of maternal and paternal grandparents (if could be recalled), sex and several phenotypical features, such as family history of disease. In the initial phase of the project, a total of 113 samples were collected from participating individuals (**Figure 1.1**).

The sampling protocol was approved by the Institutional Review Board (IRB) of Uzhhorod National University in Uzhhorod, Ukraine (Protocol #1 from 09/18/2018) as part of the "*Genome Diversity in Ukraine*" project. The sampling procedure and the informed consent that have already been approved and used during the initial collection effort. The hard copies of the consents and personal interviews data remain sealed and stored at the Biology Department of Uzhhorod National University in Ukraine. The project is considered to be non-human research at Oakland University, as OU employees handled only the depersonalized genomic data passed on by the Ukrainian collaborators.

1.2.2 Whole blood DNA extraction

After signing an informed consent, a whole blood sample was collected from a cubital vein of an individual using 5 ml EDTA tubes by a doctor or a certified nurse and assigned a code number. After finalizing the interview and taking the blood sample, all the vials were deidentified, and assigned an alphanumeric identifier and a barcode. All the

downstream sample processing and data analysis after this point has been done blindly, where neither the participants nor the researchers are able to identify the person who donated the sample.

The tubes containing fresh blood were shipped on dry ice overnight by a courier to a biomedical laboratory certified to handle blood samples in Uzhhorod, Ukraine (Astra Dia Inc.) where DNA extraction commenced immediately on arrival of a sample. Any remaining blood that was not used for extraction (~2 ml), and the DNA extracted from samples is stored frozen at the biobank of the Biology Department, Uzhhorod National University, Ukraine as a backup for the future use.

Immediately on arrival to the clinical laboratory, 200 uL of blood upon arrival to the clinical laboratory, DNA is purified from the sample using innuPREP DNA Blood Minikit (Analitik Gena, Germany). The initial DNA concentration was measured by photometric method (Implen C40 Nanophotometer; München, Germany), and quality visualized on a 2% agarose gel. The concentration of the successfully extracted DNA samples was normalized to 20-30 ng/µl for subsequent application.

In the initial project, the HMW DNA samples were aliquoted, assigned new codes and shipped to Division of Cancer Epidemiology and Genetics (DCEG), National Cancer Institute, National Institute of Health (Bethesda, MD, USA) for whole genome genotyping, to Beijing Genomic Institute facility (BGI-Shenzhen, China), or to Psomagen Inc. (Gaithersburg, MD, USA) for the whole genome sequencing (WGS).

1.2.3 Genotyping and sequencing

In the current study97 individuals were sequenced with DNBSEQ-G500 at the BGI-Shenzhen, CHINA and 87 individuals were cross validated at DCEG by genotyping using Illumina Global Screening Array (Illumina Inc., San Diego, CA) [2]. In addition, a single sample (EG600036) was also resequenced at ~60x coverage on the Illumina NovaSeq 600 S4 Psomagen Inc. (Gaithersburg, MD, USA). Upon the receipt at one of the sequencing or genotyping facilities, samples were once more checked for quality, concentration is once more detected by fluorometric method, and sample integrity and purity are evaluated using agarose gel electrophoresis (1%; 150 V; 40 min).

As much as 1 μg genomic DNA was aliquoted and fragmented by Covaris (Woburn, Massachusetts), and prepared for the DBNSEQ-G50 sequencing. The fragmented genomic DNA was selected by Agencourt, XP-Medium kit (Beckman Coulter, Life Sciences, IN, USA) to an average size of 200-400 bp. Fragments were end repaired and then 3' adenylated. Adaptors were ligated to the ends of these 3' adenylated fragments. PCR products were purified by the Agencourt AMPure XP-Medium kit. The double stranded PCR products were heat denatured and circularized by the splint oligo sequence. The single strand circle DNA (ssCir DNA) was formatted as the final library. The qualified libraries were sequenced by DNBSEQ-G50: ssCir DNA molecule formed a DNA nanoball (DNB) containing more than 300 copies through a rolling-cycle replication. The DNBs are loaded into the patterned nanoarray by using high density DNA nanochip technology. The raw, pair-end 100 bp reads are obtained by combinatorial Probe-Anchor Synthesis (cPAS, BGI-

Shenzhen, China). Finally, the raw reads were filtered removing adaptor sequences, contamination and low-quality reads.

The whole genome sequences were further aligned to the reference human genome, and for each whole genome sample, we estimated the number of passing bi-allelic SNP calls (i.e., loci with the non-reference genotypes relative to the most current major human genome assembly, GRCh38 [28]. The resulting SNP databases were also compared to the existing global resources of population variation such as gnomAD [29] and the 1,000 Genomes (1KG) [30]. Specifically, under our search criteria, the small variants (SNPs and small indels) were considered "novel" if they were absent from all the samples in the 2 global datasets. The comparison SNP calls on three independent platforms allow for cross validation of technologies to access our confidence in objectivity of sequencing methodology.

DBNSEQ-G500 sequencing was cross-validated on 87 samples with the Illumina Infinium Global Screening BeadChip Array-24 v1.0 (GSAMD-24v1-0) for 700,078 loci at the NCI's DCEG (Bethesda, MD), and a single sample (EG600036) was additionally sequenced by Illumina NovaSeq 600 S4. The high level of concordance between the three platforms has been reported. Approximately 97.7% of the SNPs identified in the BGISEQ-500 were verified by the Illumina NovaSeq 600 S4, and 95.8% of all the SNP genotypes called by the Illumina method were also detected by the BGISEQ-500. This allows us to be confident that the sequencing method we used resulted in robust dataset of sequence variants that can be verified across different platforms.

1.2.4 Variant calling and annotation

The sequencing data from 97 samples obtained with the BGISEQ-500 platform were analyzed using Sentieon tools. The analysis involved aligning the reads to the GRCh38 human reference genome with BWA-MEM (Version: 0.7.16a-r1181), preparing mapped reads for variant calling using GATK (v3.8-1-0-gf15c1c3ef by Broad), and performing SNP and indel discovery with GATK HaplotypeCaller[31]. The results were merged into a single pVCF using bcftools. For sample EG600036, joint calling was not performed, and concordance between Illumina and BGISeq variant callsets was calculated using TiTvtools and visualized with plotTiTvv[32].

The sequence variant files were annotated using ANNOVAR (RRID:SCR_012821)[33]and SNPEff (RRID:SCR_005191)[34] software with GRCh38 reference databases. ANNOVAR annotations utilized the RefSeq Gene, 1GP superpopulation, dbSNP150 (with allelic splitting and left-normalization) databases. Medically related and functional variants were annotated using ClinVar Version 20200316[35], InterVar gnomeAD Version 3.0, and dbnsfp Version 35c[36]. SNPEff utilized the default GRCh38 annotation database supplemented with ClinVar (RRID:SCR_006169)[35] and GWAS catalog[37] database annotations using the snpSift tool (RRID:SCR_015624)[38].

1.2.5 Population structure

For studying population structure, we used principal component analysis (PCA) analysis, where we incorporated WGS variants from our samples and merged them with samples from neighboring countries, including European populations from 1KG (CEU,

TSI, FIN, GBR, IBS)[39][30], French (FRA) and Russians (RUS) from HGDP[40], as well as high-coverage human genomes from EGDP[41] and the Simons Genome Diversity Project[42] (CRO, CZ, EST, GER, GRE, HUN, MOL, POL, RUS, UKR). Eigensoft (RRID:SCR_004965) was employed for the analysis.

To ensure a meaningful dataset, we filtered the resulting dataset by genotyping rate (1) and pruned variants in linkage disequilibrium by excluding highly correlated variants within a moving window (–indep-pairwise 50 10 0.5). This led to 677 samples with 208,945 variants. The eigenvectors were calculated using EIGENSOFT [48], and PC1 and PC2 were visualized using Python programming language with pandas, matplotlib, and seaborn libraries [59]. Two outlier samples (EG600056 and EG600052) were excluded from the visible range of the PCA plot as they clustered together far from any known European group.

1.3 Results

The study produced a set of data that includes more than 13 million distinct variations in Ukrainians (UAU). This dataset was analyzed to determine if the variants have a functional impact and if they are relevant to medically related phenotypes. The data can be found in Table 1.1 and database GigaDB[2]. Of the detected variants, 3.7%, or 478,000, have never been recorded in the Genome Aggregation Database (gnomAD[29]) (Table 1.1), and this number is similar to what was found in two populations in European Russia (3–4%[43]). A significant portion of the discovered variants, 12.6%, is also not registered in the 1000 Genomes Project (1KG) [30]. The majority of these described variants are rare or very rare (frequency <5%).

Table 1.1: Summary of variation in the 97 whole-genome sequences from Ukraine

	All samples			**Mean per sample**	
Sequencing results			**Novel % gnomAD (1000 Genomes)[a]**	**All**	**Novel %**
Total sequence reads	99.8 Bn			1.03 Bn	
Mean coverage	97 Samples at 30×			30×	
Variation					
	No. of total unique variants	Novel gnomAD count			
SNPs	13,010,979	477,564	3.7 (12.6)	3,488,083	0.1 (0.7)
Bi-allelic	12,667,283	470,667	3.7 (12.7)	3,340,557	0.3 (0.6)
Multi-allelic	343,696	6,897	2.0 (7.4)	146,340	0.8 (4.7)
Small indels[b]	2,727,604	76,484	2.8 (7.4)	917,731	0.3 (1.0)
Deletions	1,805,739	55,599	3.1 (9.0)	624,919	0.3 (2.4)
Insertions	14,459,87	30,453	2.1 (6.7)	571,461	0.2 (2.1)
Structural variants [c]					
Large deletions	16,078	10,914	67.9 (48.3)	3,524	52.6 (19.1)
Large duplications	1,845	1,356	73.5 (42.3)	562	89.4 (35.2)
Inversions	337	314	93.2 (47.8)	185	94.1 (48.6)
Mobile element insertions					
Alu	2,316	1,805	77.9 (38.1)	473	68.1 (18.0)
L1	451	289	64 (50.1)	79	60.8 (27.8)
SVA	100	75	75 (52.0)	20	70 (50)
NUMT	714			16	

a
Defined as percent not reported in gnomAD (1000 Genomes).
b
Small indels are insertions and deletions <50 bp called by GATK [44].
c
Large deletions and duplications are those called by Lumpy[45], which are >50 bp.

A particular interest in this study is the distribution of functional variation, due to their potential impact on phenotypes, especially to those with medical relevance. To ensure the reliability of our findings and eliminate any potential confounding effects, we thoroughly examined the ancestry composition of the studied populations. This analysis aimed to identify any population substructure that could lead to variations in allele frequencies.

The Ukrainian genomes analyzed in this study, along with ones from other studies, exhibit a distinct clustering pattern. They form a defined cluster positioned between the Northern European populations, including Russians and Estonians, on one side, and Western European populations like CEU, French, British, and Germans on the other side (refer to Fig. 1.2). Furthermore, there is a noteworthy overlap with other populations from Central and Eastern Europe, such as Czechs, Poles, Croats, Greeks, and Moldovans from the Balkans. Only one sample clustered with Italian samples indicating a possible migration event.

SNP databases were compared to the existing global resources of population variation containing variants that are already known as medically related, either in ClinVar[46], an NCBI database of reports of the relationships among human variations and phenotypes with supporting evidence), or in Genome-wide association studies (GWAS)[37]. We compared genetic mutations related to medical conditions and found 43,892 benign, 189 likely pathogenic, and 20 protective variants based on the ClinVar database. On average, each study participant had 19 pathogenic and 12 protective mutations (summarized in the Table 1.2). Even though some individuals had two copies of the disease-causing mutation, none of the participants were symptomatic for the associated

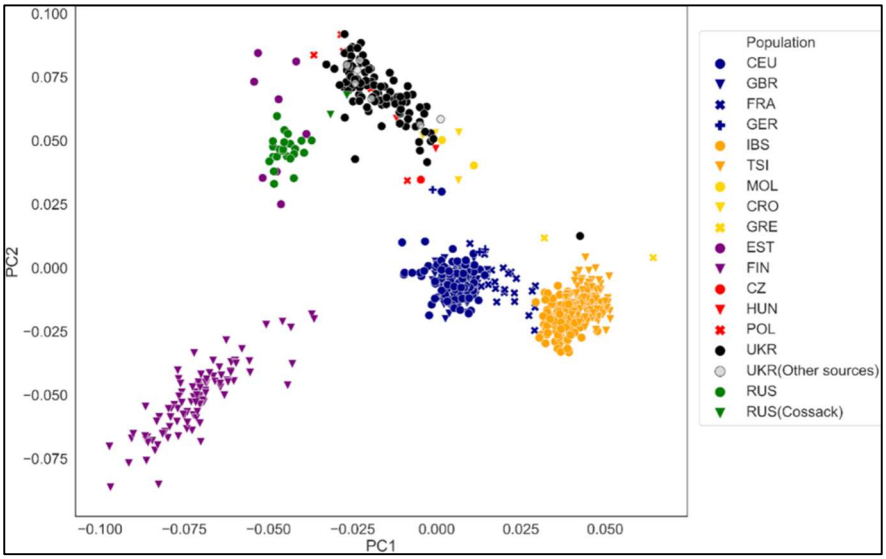

Figure 1.2: Principal component analysis (PCA) was conducted on a merged genetic dataset comprising European populations. The colors assigned in the analysis correspond to the population assignments derived from European samples in the 1,000 Genomes Project (specifically, Utah residents [CEU] with Northern and Western European ancestry, Toscani in Italy [TSI], Finnish in Finland [FIN], British in England and Scotland [GBR], and Iberian population in Spain [IBS])[39]. Additionally, French individuals (FRA) and Russians (RUS) from the Human Genome Diversity Project (HGDP)[40], as well as relevant high-coverage human genomes from the Estonian Biocentre Human Genome Diversity Panel (EGDP)[41] and the Simons Genome Diversity Project[42] , including Croatian (CRO), Czech (CZ), Estonian (EST), German (GER), Greek (GRE), Hungarian (HUN), Moldovan (MOL), Polish (POL), Russian Cossack (RUS), and Ukrainian (UKR) populations, were included. The PCA analysis was performed using Eigen

Table 1.2: ClinVar and GWAS documented variants found in Ukrainian population

Source of annotation	**Number of Unique substitutions**	**Total number of alleles**	**Average per sample**
GWAS catalogue	102, 551	6 479 953	66804
ClinVar: pathogenic or likely pathogenic	189	1830	19
ClinVar: benign or likely benign	43892	1842668	18997
ClinVar: protective or likely protective	20	1209	12

diseases. This could be due to various factors such as age, genetic diversity, and environmental factors.

Frequency of genetic diseases and traits can vary between populations due to evolutionary forces like genetic drift, mutation, migration, non-random mating, and natural selection. To highlight the distinctiveness of this population and its potential for future genomic research, we compiled a list of known disease-causing genetic variants that have different frequencies in Ukrainians compared to other European populations, as seen in the EUR sample from the 1000 Genomes Project.[47]).

Several examples of these variants are presented in **Table 1.3**. Among those are the variants involved in a number of diseases such as hyperglycinuria/iminoglycinuria (*rs35329108; SLC6A19*), warfarin (*rs7089580; CYP2C9*), bisphosphonates (*rs2297480; FDPS*) and cisplatin (*rs2228001; XPC*) response/toxicity (Level of evidence 3, according to https://www.pharmgkb.org), longevity and diabetes type II susceptibility (rs1805097; *IRS2*), risk of asthma and lung function (*rs4950928; CHI3L1*) (**Table 1.3**). The rest of these alleles are reported in the Supplementary Table 1.1.

1.4 Discussion

The wealth of genomic variation uncovered in Ukraine and, to the lesser extent, in Russia was not due to the particularly high variation within the Ukrainian population but to the absence of sampling variation from the countries around Ukraine. Since the Ukrainian population was surrounded by unexplored genome deserts, the study of fewer than 100 Ukrainian genomes uncovered 478,000 novel genomic SNPs that have never been previously registered in the Genome Aggregation Database (gnomAD). The effect

Table 1.3: Examples of the functional SNPs with highly differentiating functional markers reported in ClinVar [29], with high different frequencies in the Ukrainian population compared to the neighboring populations in other European countries (the combined sample from Western and Central Europe from 1000Genomes Project (EUR). Non-reference allele frequency (NAF) is reported compared to the reference allele in GRCh38. Differences are evaluated by the Fisher Exact Test (FET). Modified from Oleksyk et al., 2021.

SNP	Chr	Gene	REF/ alt [P]	Associated medical condition	Function	NAF UKR	NAF EUR	FET vs. EUR (p-value)
rs35329108	*5*	*SLC6A19*	g/A	Hyperglycinuria| Iminoglycinuria	Pathogenic, Intronic	0.32	0.23	*0.004*
rs2297480	*1*	*FDPS*	t/G	Bisphosphonates response	Drug response, intronic	0.13	0.26	*0.0000536*
rs7089580	*10*	*CYP2C9*	a/T	Warfarin response	Drug response, intronic	0.31	0.22	*0.006*
rs2228001	*3*	*XPC*	G/t	Cisplatin response-toxicity	Drug response, exonic, nonsynonymous	0.51	0.66	*0.04*
rs1805097	*13*	*IRS2*	c/T	Longevity, Diabetes type II susceptibility	Risk factor, exonic, nonsynonymous	0.39	0.31	*0.04*
rs4950928	*1*	*CHI3L1*	G/c	Risk of asthma and lung function	Risk factor, upstream	0.75	0.82	*0.03*

[P] The reference allele is set according to the reference allele in GrCH38.p13 .

observed is due to what is known as a “first-mover” or “pioneer advantage” in the business and marketing world[26].

This magnitude is striking, particularly when compared to the most genetically heterogeneous populations found in sub-Saharan Africa. By incorporating genetic information from 426 individuals across 50 distinct ethnolinguistic groups across the entire continent into existing databases, approximately 3 million genetic variants were unveiled [27]. While Africa is expected to have the highest genetic variation as the source population, Ukraine’s genome diversity is more recent in origin and is not expected to be higher than in other European countries where genetic diversity is comparatively low. Yet, since no studies looked at Ukraine and her neighbors, the first glance at genomic diversity in the region yielded a treasure trove of previously undiscovered variation. This straightforward juxtaposition emphasizes the criticality of obtaining comprehensive insights into the global panorama of genomic diversity and eliminating remaining genomic gaps worldwide.

Among the discovered variation, medically relevant alleles were of particular interest because they contribute to the difference in disease, treatment and prophylactic outcomes in Ukrainians compared to other populations elsewhere in Europe and worldwide. These endemic variants will become a valuable resource for designing future population and clinical studies, help address questions about ancestry and admixture, and will fill a missing place in the puzzle characterizing human population diversity in Eastern Europe. As expected, our study shared a lot more variants with the GWAS than with the ClinVar catalogue[35], [48] While GWAS has recently become the tool of choice to identify genetic variants associated with complex disease and other phenotypes of interest,

since the amount of genetic variance explained by these variants is low, they are generally not very useful for prediction pathogenic phenotypes[48].

Comprehending the full spectrum of geographic dimensions encompassing genome diversity in Europe holds paramount importance for the local biomedical community, as it enables the utilization of localized genomics data instead of relying on extrapolations from genome projects conducted in other nations[26]. The significance of obtaining sequence data from multiple populations cannot be understated, given their distinctive historical trends of genetic drift, natural selection, migration, genetic admixture, and socioeconomic structures[49]. Consequently, we propose that each country should establish its own national genome database to facilitate the formulation of regionally relevant and objective public health policies. There still remains a wealth of critical genetic variations waiting to be unearthed, and it is imperative that these findings are made accessible to provide an informative framework for future biomedical research endeavors[26].

CHAPTER TWO

GENETIC DIVERSITY IN UKRAINIAN TRANSCARPATHIA AND ITS IMPLICATION FOR THE DEVELOPMENT OF PERSONALIZED MEDICINE

2.1 Introduction

The number of reference populations that are whole genome sequenced are rapidly increasing because genetic disease studies benefit from knowing the genetic variation typical for the geographical area of interest. The recent (<250 years) explosive population growth of human populations in Europe resulted in an abundance of rare variants with relatively recent origin that tend to be geographically clustered to a higher degree than variants that are common in different parts of the world [50].

The major problem is that still much of the genetic variation remains unknown, especially in local and isolated populations. While large-scale projects (including Genome Diversity in Ukraine described in Chapter 1) we have already mapped much of the most common variants, the detailed map of human genetic diversity is far from complete, in particular with respect to rare and endemic variants specific to certain areas.

Population of Transcarpathia (Zakarpatska oblast) in Ukraine is an important sample for genetic studies due to its numerous rural communities, relative geographic isolation, and admixture due to its history and proximity to neighboring countries. It is a region in Ukraine that is situated in the southwestern part of the country and shares borders with Slovakia, Hungary, Romania, and Poland. This region has been relatively geographically isolated from the rest of modern Ukraine due to its location behind Carpathian Mountains and history of being a part of different states and empires. It was

isolated from the rest of Ukraine while it was ruled by the Austria-Hungary and was later incorporated into Czechoslovakia, only to be occupied by Hungary once more during WWII before being reunited with Ukraine in 1946 [51]. This tumultuous history resulted in it developing a diverse population and culture in this unique region, as change of allegiances has brought migrants from different parts of Europe.

Transcarpathia is generally defined as a region delimited by the right bank of the Tizsa river in the south and its watershed in the Carpathian Mountains in the north. The majority of its population are ethnic Ukrainians who have settled here in historic times by moving across the mountains from the north, and forming four different sub-ethnic groups, Dolynyany in the lowlands and Boykos, Hutsuls, and Lemkos, the three population groups of the Carpathian highlands[52]. Hungarian and Hungarian-speaking population is especially prevalent in the southern part of Transcarpathia, along the Tizsa river and attest to the centuries of Hungarian domination in the region[52], [53].

Historic records indicate that in 17-20th century Transcarpathia also had steady immigration of Germans, Austrians, Czechs as well as Hungarians. Moreover, during and immediately after World War II, the region was the site of significant population movements, with many people being resettled in the region from other parts of Europe and especially from the other parts of the USSR[54], [55]. As a result, significant minorities that exist here reflect the history of domination in the region. These include 12.7% Hungarians, 2.9% Russian, 2.6% Romanian (or Wallachians), and smaller but geographically defined areas with Slovakian, Roma (Gypsies) and German (Swabian) populations (see **Figure 2.1**).

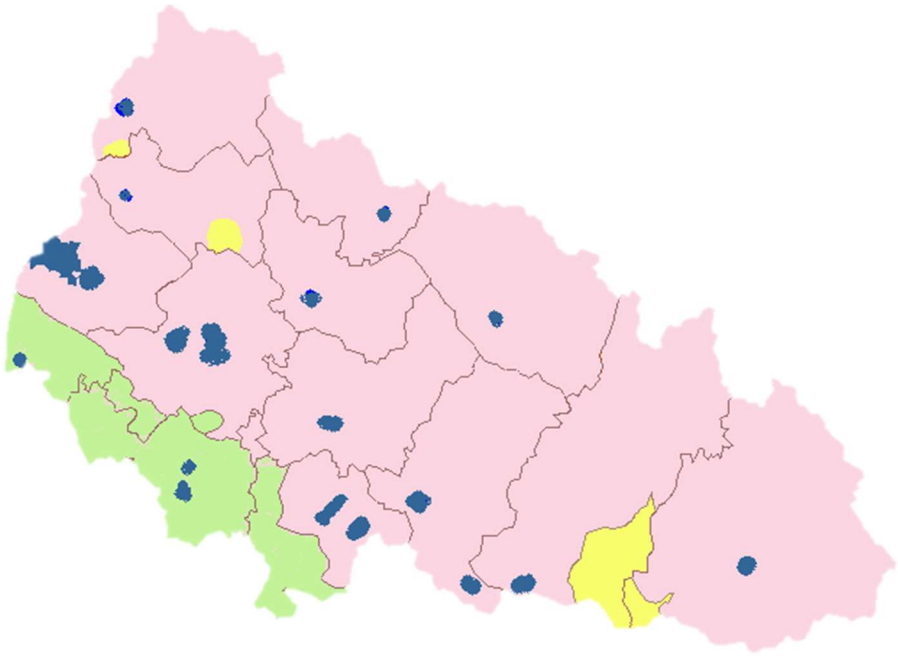

Figure 2.1: The self-identified ancestry map of Transcarpathia (circa 2001). Majority of the population is Ukrainian (pink) while there are defined areas populated by majority Hungarians (green) and Romanians (yellow). Large urban centers indicate mixed populations (blue). The Roma, Slovakians and Germans are not represented. Source "Regions of Ukraine / Zakarpattia region". 2001 Ukrainian Census. Retrieved 2021.

Socio-economic studies show that the region has been heavily reliant on agriculture and currently has a lower level of industrialization compared to other regions of Ukraine. Due to this still majority of population of Transcarpathia, officially defined as administrative region of Zakarpatska Oblast, resides in villages or small towns, many of which are relatively isolated culturally and geographically[44]. A total estimated population of 1.25 million individuals lives in an area of approximately 5,000 sq miles, slightly bigger than the size of the US state of Connecticut, with majority population inhabiting the piedmont region[56].

Such unique status of Zakarpatska oblast makes it extremely interesting to further zoom in in population genetic studies, especially having countrywide genomic data

available for comparison[2]. Unfortunately, very little is known about genome diversity of this region, with a handful of previous studies focusing on mitochondrial haplogroups[57], [58] that are quite uninformative, especially on the population scale. Genome wide exploration in this region is likely to have a "pioneer advantage" and has a high potential to uncover many previously undetected genetic variants[26]. Specifically, many rural communities in Transcarpathia are the potential harbors of rare genetic variation due to founder effect and limited interbreeding. Such communities worldwide have been proven to be excellent samples for genetic studies as exemplified by Amish, French Canadian, Mennonite communities among others, who being studied provided valuable information on such rare conditions as Gaucher's, Tay-Sachs disease, Leigh syndrome etc[59]–[61].

With the justification provided, in my second chapter I will focus on characterizing of the ethnically and genetically diverse population of Transcarpathia for further discovery-based research on existing genetic variation and identifying intriguing facts for subsequent hypothesis-driven research.

2.2 Methods

2.2.1 Sampling strategy

This study was part of the cross-border cooperation project between Romania and Ukraine funded by European Union (Grant Joint Operational Program Romania - Ukraine 2014-2020, funded by ENI CBC-2SOFT / 1.2 / 48), sampling protocol is approved by IRB of Uzhhorod National University. Our main sampling strategy was to represent genetic diversity in Ukrainians in all the districts (regions) in Transcarpathia (Zakarpatska Oblast) (**Figure 2.2**). Recently in 2020 the 13 old administrative regions of Transcarpathia been combined into 6 new regions for the 2020[62]. While the current population sizes are

difficult to estimate due to population movements, the old 13 regions were formed to represent historic population sizes, and used for administration and electoral purposes in the past when the current population of Transcarpathia was formed and may be better approximation to the historical population distribution than the 6 modern regions. In addition, several samples were collected from minority groups (Wallachians and Roma). These self-reported ethnicities that were not represented in any other genome surveys available to us for comparison[26].

During the study, 150 healthy volunteers have been recruited by the certified medical personnel in all 13 old administrative regions. All participants were chosen to fit the strict inclusion criterion of self-reported ancestry from Transcarpathian region at least in third generation. Similarly, to chapter one[2], each volunteer received all necessary information regarding the study and the blood collection procedure and provided full consent to participate and have their genotypic and phenotypic data to be publicly available. The study participants also filled out a questionnaire indicating their self-reported region of origin, place of birth of maternal and paternal grandparents (if recalled), sex and several phenotypical features, such as anamnesis of diseases. Three samples of blood were collected from every participant, placed in 2 ml EDTA tubes, and frozen immediately.

Out of a total of 150 samples were collected from participating individuals in the initial phase of the project, 143 were self-declared ethnic Ukrainians, and 130 of them had all 4 grandparents from a single or neighboring settlements in the same region, while 13 had grandparents from different places within the region (**Figure 2.2**).

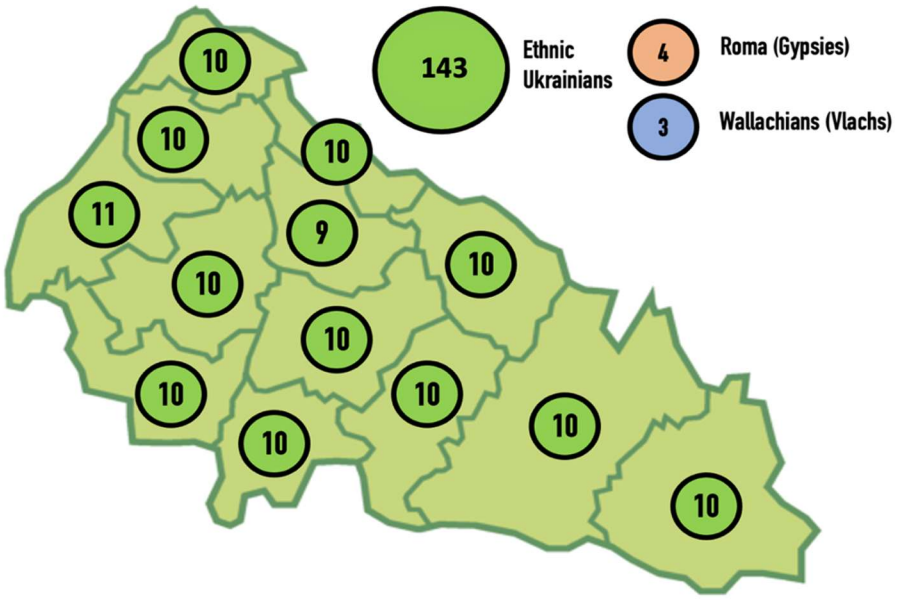

Figure 2.2: Structure of the population sample collection in Transcarpathia. 150 samples were collected from participating individuals, including 143 ethnic Ukrainians, 4 Roma, 3 Wallachians.

2.2.2 DNA extraction, Next Generation Sequencing Data and Analysis

DNA extraction was performed at the Molecular Genetics laboratory at Biology department of Uzhhorod National University. Monarch® DNA purification kit (New England biolabs, Inc., Rowley, MA, USA) is used to extract genomic DNA from the whole blood in the original frozen sample. As much as 1 μg genomic DNA was aliquoted and fragmented by Covaris (Woburn, Massachusetts), and prepared for the DBNSEQ-G50 sequencing at the BGI-Shenzhen facility CHINA. Protocol of library preparation, sequencing, bioinformatic analysis pipeline for ancestry composition and medically relevant alleles search is consistent with the methods section of Chapter 1. Exploratory

ancestry composition analysis was performed using DIYDodecad tool with Eurogenes K36 and Tolani K54 calculators[63]. These methods are commonly used in population genetics to infer ancestral components and assess population structure[64]. Specifically, Eurogenes is a tool used for population structure analysis and ancestry determination based on genome-wide SNP data. It employs a model-based clustering approach called ADMIXTURE[65], which assigns individuals to different ancestral populations or clusters based on genetic similarity. The parameter "K"[65] in Eurogenes represents the hypothesized number of populations or clusters in the analysis. In our case, "K=36" indicates that the analysis assumes the presence of 36 distinct ancestral populations. On the other hand, Tolani is another method used for population structure analysis and ancestry estimation. Similar to Eurogenes, it utilizes a model-based clustering algorithm to assign individuals to different ancestral populations. In the case of Tolani, "K=52" indicates that the analysis assumes the presence of 52 distinct ancestral populations. Due to the inherent biases present in these calculators, we have used the euro-centric Eurogenes K36 for the Ukrainian-identifying part of our sample[66].

2.3 Results

In order to investigate if there is any underlying substructure within the population, in order to facilitate further analysis of medically relevant alleles, we conducted an ancestry analysis of the population of Transcarpathia. Compared to our estimates of overall genome diversity in 97 Ukrainians from across Ukraine described in Chapter 1, Ukrainians in Transcarpathia exhibit a significantly higher proportion of Balkan and Romanian ancestry, nearly doubling the averages admixture fractions when compared to the reference group

(14% vs. 7%, and 11% vs. 5%, respectively). At the same time, the proportions of unique Ukrainian and Polish ancestry were found to be relatively similar between the two groups (21% vs. 24% and 16% vs. 17%, respectively). The detailed ancestry composition in Transcarpathia is shown in the Figure 2.3 compared to the average composition across Ukraine estimated by the same methods in Figure 2.4.

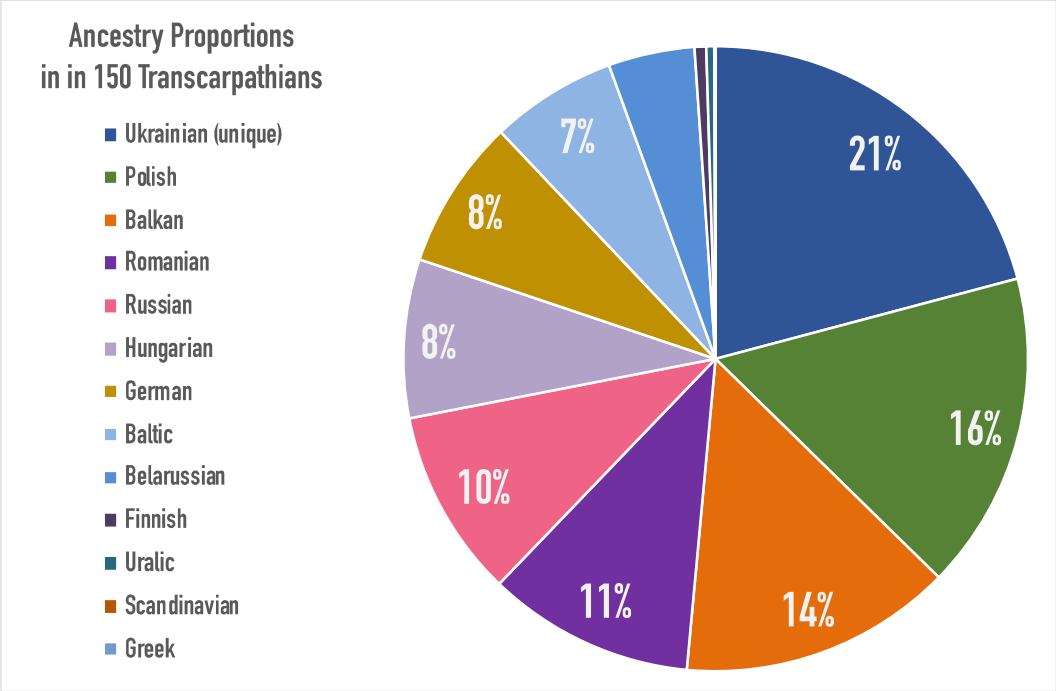

Figure 2.3: Ancestry composition of 150 Ukrainians in Transcarpathia (TRC 150) estimated by the Eurogenes method with K=36[63]

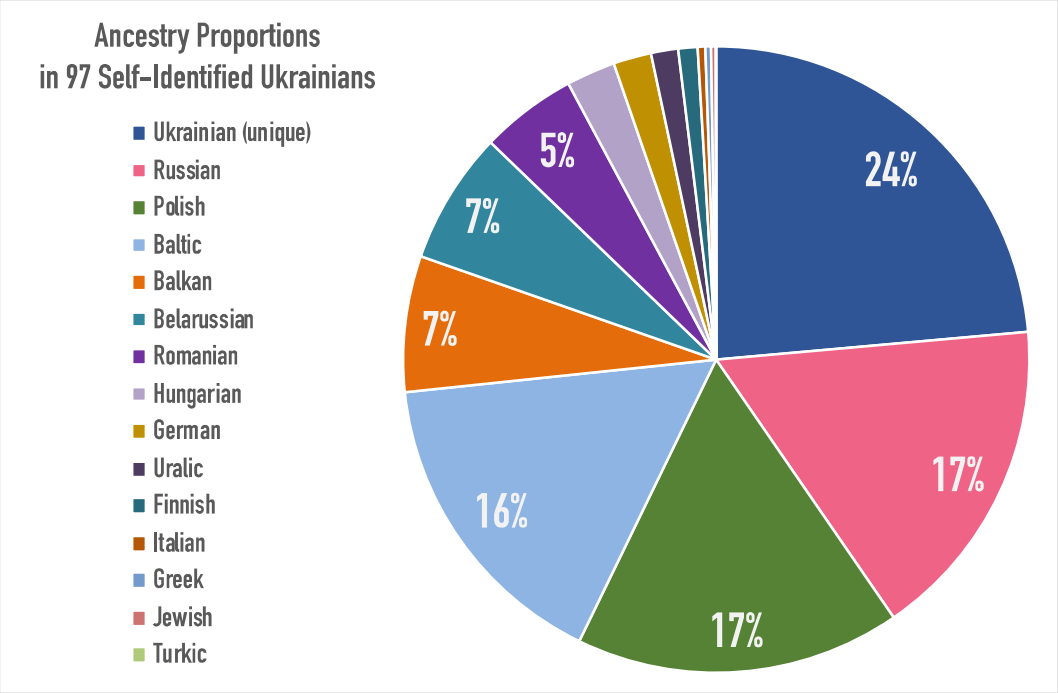

Figure 2.4: Ancestry composition of 97 Ukrainians in Ukraine (UAU97) from Genome Diversity in Ukraine study (described in Chapter 1) estimated by the Eurogenes method with K=36[63]

In this study, we compared the population frequency of alleles detected in 143 individuals from Transcarpathia, Ukraine, to those found in our previous study of 97 Ukrainian individuals (UAU97), European population (EUR), Romanians (RO) and Russians (RU). We did not include the 3 Roma and 4 Wallachian individuals into the comparisons with Ukrainians, as they clearly show different genetic ancestry. The origin of this admixture is mostly from the Indian subcontinent for Roma, and mixed South European for Wallachians. This difference is consistent with existence of distinct genetic subpopulations and structure and supports the isolated status of Roma populations within Transcarpathia (Figure 2.5). Because Roma has presence of the non-European ancestry,

Eurogenes K=36 method would not be able to identify all the ancestral components. Therefore Tolani (K=52) model was used for the estimation of ancestry proportions.

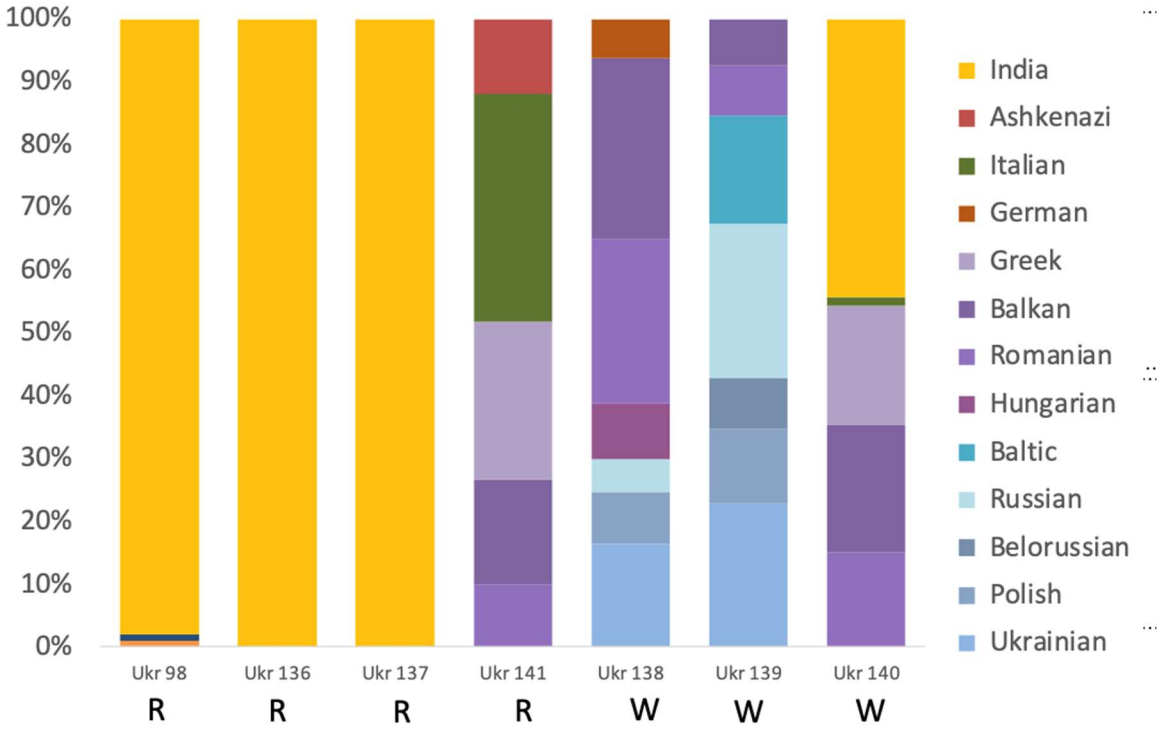

Figure 2.5: Ancestry composition of 4 Roma and 3 Wallachian individuals using by Tolani method with K=54[63].

We identified 5,496 variants that significantly differed in frequency between Transcarpathian sample (TRC150)[67] and surrounding populations (150 Romanians (ROR)[67], European (EUR)[30] and Ukrainians from Ukraine (UAU97)[2]. Of these alleles, 1,742 were also significantly different in frequency from previously studied Ukrainian population from other regions of Ukraine (UAU97), Fisher exact test $p<0,05$ (Supplementary Table 2.1)).

Among these medically relevant variants previously classified in ClinV[35] were evaluated on their impact on human health. By impact, 389 variants are predicted to have

HIGH effect, 305 MODERATE, and 518 LOW effect. The rest 4283 variants are MODIFIERS. The breakdown of the medically relevant variant classified by different criteria are shown in the Table 2.1.

Table 2.1: Number of variants with significantly different frequencies in TRC150 from surrounding populations classified by ClinVar effect, type of mutation and predicted impact (Fisher exact test p<0,05)

CLINVAR EFFECT	NUMBER	TYPE OF MUTATION	NUMBER	IMPACT	NUMBER
Benign or likely benign	4504	**3-prime-UTR**	392	**Low**	517
Conflicting interpretations	24	**5-prime-UTR**	137	**Moderate**	305
				High	389
Pathogenic or likely pathogenic	23	**Conservative in frame deletion**	16	**Modifiers**	4283
		Frameshift	180		
Drug response and association	6	**Downstream gene variant**	427		
		Intergenic region	2		
Risk alleles	6				

Table 2.1 – continued

CLINVAR EFFECT	NUMBER	TYPE OF MUTATION	NUMBER	IMPACT	NUMBER
Uncertain significance	61	**Intron variants**	2791		
		Missense variant	268		
Unknown effect	369				
		Splice acceptor	268		
		Synonymous variant	394		
		Upstream gene variant	547		

Next, we focused on identifying variants that show a robust association with certain conditions that exhibit significant differences in frequency between TRC150 and the UAU97 samples from the rest of Ukraine, as well as European samples where genome sequences are available. In our analysis, we aimed to find specific variants that contribute to the unique genetic landscape of Transcarpathia and potentially shed light on the underlying factors contributing to the disparities in disease prevalence. Several examples of highly functional variants are presented in the Table 2.2 and the rest are reported in Supplementary Table 2.1. One of such variants is *rs202003805* in *PRSS1* gene known as c.47C>T; (p. Ala16Va), strongly associated with chronic pancreatitis. We estimated its

frequency in Transcarpathia is 34% while in our previous Ukrainian sample and Europe it is completely absent. Another variant, *rs1799724* near the *TNF* gene, was identified as an independent risk factor for Alzheimer's disease and a modifier of risk factor in individuals with *ApoE4* alleles. This variant is seen in 23% of Ukrainians in Transcarpathia, resembling frequency in the neighboring Romanian population (RO150) while it is completely absent in the UAU97 and less prevalent in Western Europe.

Table 2.2: Medically relevant variants with significantly different frequency in the cohort from Transcarpathia comparatively to surrounding populations, Fisher exact test p<0,05

Gene and allele	**Effect on**	**TRC150**	**UAU97 [1]**	**RO 150**[67]	**EUR**[30]	**GnomAD**
TNF rs1799724	Heritable Alzheimer's	23%	0%	18%	5%	8%
PRSS1 rs202003805	Heritable Pancreatitis	34%	0%	33%	0%	0.009%
XPNPEP2 rs1799724	ACE-inhibitor dependent angioedema	23%	0	0	4.5%	7.8%
PRC1 rs11852999	Breast cancer course	12.3%	7%	7%	7%	10%
HTRA1 rs11200638	Age-related macular degeneration	30	0	22	17	22%

2.4 Discussion

In this part of the study, we sequenced 150 individuals with self-reported Ukrainian ancestry from Transcarpathian region in Ukraine (TRC150) and several ethnic minorities (Roma and Wallachian) from the same region that were not represented in earlier genomic surveys. We conducted comparisons of genetic frequencies with neighboring populations in order to identify candidate variants among the medically related loci that could be useful for future studies. We identified a substantial number of variants that significantly differed in frequency between the Transcarpathian sample and the surrounding populations. These variants classified based on their ClinVar effect, type of mutation, and predicted impact, provide a comprehensive overview of the medically relevant genetic landscape specific to Transcarpathia. The alleles we identified exhibited significant differences between populations of Ukrainians in Transcarpathia and other European populations in close geographical proximity: Ukrainians from Ukraine (UAU97), Romanian (Romanians from Maramures (RO150)[67] and European superpopulation from 1000 genomes project (EUR)[30](Fisher exact test $p<0,05$, Table 2.3).

Of particular interest are the variants showing robust associations with certain conditions that exhibit significant differences in frequency between Transcarpathia and the rest of Ukraine, as well as European samples. By pinpointing these specific variants, we can gain insights into the unique genetic factors contributing to the disparities in disease prevalence observed in Transcarpathia (Table 2.3). For example, we identified

rs202003805 in the *PRSS1* gene, strongly associated with chronic pancreatitis, which showed a frequency of 34% in Transcarpathia while being absent in the previous Ukrainian sample and Europe. It leads to a non-conservative alteration of an amino acid in the protein sequence. The change occurs in position 16 of the trypsinogen activation peptide, which is released during the activation process and is not present in active trypsin. While it is unlikely to affect trypsin function directly, it may impact the processing of trypsinogen by chymotrypsin C (*CTRC*). In-silico tools predict a benign effect on protein function in four out of five cases. The variant allele has been observed at a frequency of $9*10^{-5}$ in 244,118 chromosomes within the Non-Finnish European subpopulation from the GnomAD database exomes dataset[16]. Although the variant is located in a region of segmental duplication, no quality control issues were identified in the GnomAD[16] database exomes dataset. The frequency of the variant does not exceed the estimated expected maximum for a pathogenic *PRSS1* variant causing chronic pancreatitis. Published studies indicate that this variant is the third most common inherited mutation in the *PRSS1* gene[68] and has been reported in affected individuals with a reported penetrance of approximately 50%[69]. However, it should be noted that the presence of pseudogenes has not been definitively ruled out in these reports. Co-occurrences with other pathogenic variants linked to pancreatitis, such as *SPINK1* c.101A>G (p.Asn34Ser), *CFTR* TG11-5T, and *PRSS1* c.364C>T (p.Arg122Cys), have been reported[70], suggesting an increased risk for pancreatitis, particularly in individuals who are trans-heterozygous for this variant along with pathogenic variants in *SPINK*1 and *CFTR* genes, and in the presence of risk factors like smoking. Functional studies indicate that the A16V variant does not affect trypsinogen secretion and autoactivation in the absence of chymotrypsin C. However, in the presence

of chymotrypsin C, it increases the rate of autoactivation by enhancing N-terminal processing, albeit to a lesser extent than highly penetrant *PRSS1* pathogenic variants like R122H and V39A[71]. Different submitters in ClinVar provided varying clinical-significance assessments for this variant, with some classifying it as pathogenic, variant of uncertain significance (VUS), or likely benign.[46]

Considering the co-occurrences with known pancreatitis-causing variants, the possibility of the variant being derived from a pseudogene, and the milder biochemical phenotype observed in functional studies, the variant has been reclassified as likely pathogenic. However, caution is advised when interpreting NGS findings for the *PRSS1* A16V allele due to the high degree of sequence homology with other trypsinogen genes, particularly *PRSS3P2*, which can lead to erroneous identification of the mutation and misinterpretation of closely located nucleotide changes by NGS analysis alone[72]. We plan to further investigate and confirm the presence of this variant in our cohort by Sanger sequencing as suggested by Weiss at al[72].

Another variant, *rs1799724* near the *TNF* gene (Table 2.3), was identified as an independent risk factor for Alzheimer's disease and a modifier of risk factor in individuals with *ApoE4* alleles[73]. This variant was present in 23% of the Ukrainians from Transcarpathia, while it was completely absent in the previous UAU97 samples and less prevalent in Western Europe (EUR). On its own, the *rs1799724* (T) allele led to an odds ratio of 1.63 (CI: 1.13-2.34, p=0.009). In carriers of *ApoE4* alleles, the odds ratio changed from 2.92 (CI: 2.00-4.27) in the absence of *rs1799724* (T) to an OR of 6.65 (CI: 3.26-13.55, p=0.03) in its presence. However, that this study is from 2005 and it does not appear to have been independently replicated since then[73].

Our findings highlight the importance of investigating population-specific genetic variations in order to comprehend disease prevalence and susceptibility locally[74]. In the current analysis, we discovered a considerable number of variants that exhibited significant frequency differences between Transcarpathia and the neighboring populations. These variants were found to be associated with various diseases. In addition to the known medically related alleles, our study also identified 389 variants with high impact but unknown function as defined in ClinVar (Table 2.1). Uncovering their potential clinical relevance to confirm their effects, should be addressed in the future association studies.

This knowledge contributes to our understanding of the genetic makeup of Transcarpathia and emphasizes the necessity of taking into account regional genetic substructure when conducting medical studies and implementing interventions. Moving forward, we plan to conduct additional association studies to validate the impact of these identified variants in Transcarpathia. This will further enhance our understanding of their role in disease susceptibility and treatment response, ensuring more accurate recommendations for effective public health policies for the population of the region.

CHAPTER THREE

LACTASE PERSISTENCE AND ITS GENETIC DETERMINANTS IN UKRAINE

3.1 Introduction

Among genetic variation related to environment adaptations in humans, lactase persistence alleles are probably the most studied and well-known, and often are referred to as a textbook example of gene-culture coevolution[75]. The ability to digest lactose, a sugar found in milk, varies worldwide due to genetic differences and lactose tolerance is a trait most common in populations with a long history of dairy farming and milk consumption pointing to its relatively recent origin[76]. In this part of my project, is focused on the distribution of lactase persistence alleles in genotype-phenotype relationship of this phenomenon Ukrainians from Ukraine.

Lactase persistence (LP) is the physiologic trait which allows some humans to digest lactose in unprocessed milk in adulthood. Lactose (milk disaccharide) is broken down in the small intestine by the enzyme lactase (lactase-phlorizin hydrolase)[77]. In most mammals, levels of lactase, thus ability to digest fresh milk, rapidly decline after weaning, but humans are the exception to this trend. Lactase-non-persistent (LNP) phenotype is not equal to lactose intolerant, because not all LNP individuals experience intestinal symptoms after consuming raw milk, which is diagnostic of lactose intolerance[78].

The combined data from 89 countries encompassing 84% of the world's population (standardized for the size of the country) show that the global prevalence of lactose intolerance is around 68% (95% CI 64–72) [79], but varies greatly from region to region. Human populations from Northern Europe[80], certain Middle East and African populations[81] have extremely high prevalence of lactase persistence (80-96%)[82]. The

fraction of intolerance increases21-31% in the populations of central Europe, while most of the individuals in East Asia, as well as Native Americans[83][84] are almost completely lactose non-persistent (LNP).

Studies showed that lactase persistence is driven by specific genetic alleles, most of which are in the gene *MCM6* (minichromosome maintenance complex component 6), acting as an enhancer to *LCT* gene promoter, thus regulating the expression of lactase coded by *LCT* gene[76]. Two SNPs discovered in in *MCM6*, *rs4988235* (−13910:C>T) and *rs182549* (22018:G>A) linked to lactase persistence were reported in 2002 by Enattah et al.[85]. The first one, *rs4988235* was highly associated with LP phenotype in Finnish families acting showing a dominant mode of inheritance where heterozygous individuals expressed complete lactase persistence [85]. However, later studies suggested that the phenotype of lactase activity may be a codominant quantitative trait with trimodal distribution where heterozygous individuals for regulating loci express intermediate enzymatic activity[86]. Furthermore, evidence shows that *rs4988235* (T allele) drives the LP phenotype in most of the Eurasian populations[87][88]. However, because lactase persistence evolved independently in different parts of the world, there are other, different alleles in the *MCM6* gene linked to this trait found in pastoralist populations from the Sub-Saharan Africa and the Middle East: *rs41525747* (13907:C>G), *rs41380347* (13915:T>G), *ss820486563* (14009:T>G), and *rs145946881* (14010:G>C)[89].

The frequency of genetic variations that retain lactase expression in adulthood has increased since the Neolithic Revolution (about 10,000 years ago)[90], when human societies first started domesticating diverse plants and animals. Studies show[91] that Neolithic farmers at around 7,500 were not yet lactase persistent, but people from Medieval

Western Europe already had LP phenotype at modern day frequency[92]. With a selection coefficient of around 0.04–0.05, the strength of natural selection for LP has been predicted to be among the greatest in the human genome[93]. This estimate gives rise to the question why adult milk consumption has such a significant impact on human reproduction and/or survival. The nature of the exact selection forces responsible for the rise in LP frequency in humans is still unknown[94].

It is interesting if, and if so, why, the capacity to consume unprocessed milk as an adult resulted in such a difference in fitness between LP and LNP individuals. There are several hypotheses which attempt to explain the evolution of lactase persistence phenotypes[95]. The first, and the most accepted is the cultural-historical hypothesis, which states that frequency of LP parallels the ancestral milk-drinking habits of populations, with high frequencies found in pastoral or agropastoral populations that traditionally incorporated large amounts of milk into their diets. Dairy consumption does have a cost for LNP people, with diarrhea being the most plausible reason for a selection disadvantage. As milk is a rich source of proteins, lipids, minerals, and vitamins, perhaps all people benefited from drinking it, but only people with LP could take use of these nutrients without experiencing any symptoms.

To understand the mechanism of selection under the cultural-historical hypothesis, the symptoms of lactose intolerance have been studied by feeding fasting individuals 50 g of lactose, which is comparable to 1 L of milk, and most LNP individuals appear to handle modest doses of lactose (such as one glass of milk, or up to 15 g of lactose)[96]. Additionally, the frequency of diarrhea and the severity of intestinal symptoms depends on each person's unique capacity for fermentation, which varies greatly[97] and is largely

influenced by the make-up of their colonic microbiota (the collective name for all the microorganisms found in the colon)[98]. For instance, if colonic bacteria are effective in fermenting lactose, the volume of gas may rise but the osmotic shock (which causes diarrhea) would be reduced. Parallel to this, when methanogenic archaebacteria are common in the gut, carbon dioxide would mostly be converted into methane, resulting in constipation instead of diarrhea, as seen in 30% of people[99], [100]. Overall, it seems that LNP population would be able to drink small amounts of milk throughout the day without any symptoms, especially if it is consumed with other meals.

The second proposed mechanism of LP evolution is referred to as the calcium absorption hypothesis[101]. According to this model, milk is a significant source of calcium, as well as of vitamin D, and lactose, both of which help people better absorb calcium. This trait would be especially helpful for farmers from high latitudes, like those in Europe, who have poor vitamin D intake from their shift to cereals and little UV exposure (which stimulates the production of vitamin D). However, recent research[102] questions the notion that lactose aids in the absorption of calcium in humans and suggests that LNP persons may benefit from ingesting derived dairy products (like yoghurt and cheese) more than the LP individuals could from drinking raw unprocessed milk[103].

It is important to mention that some observations seemingly contradict both models of adaptation. For instance, the intake of milk does not necessarily correspond to the frequency of LPs in some populations: the traditional herders from Mongolia and Central Asia [88], [104]are two well-known examples of pastoral and agro-pastoral groups with low population frequency of LP alleles despite daily consuming milk, while Hadza and Yaaku hunter-gatherers from East Africa [105], [106]have high frequency despite never

consuming any milk at all.

To date, the information on prevalence and degree of lactose intolerance in Ukraine is scarce. The review by Storhaug et al. reported the prevalence of lactose intolerance among Ukrainians to be 61% based on literature reviews and genotyping of common SNPs in various small cohorts [79][107][108]. Nevertheless, there has been no comprehensive evaluation of lactose intolerance in Ukraine, and never using evaluation of the phenotype. Genetic testing for the alleles associated with lactose intolerance can help in the diagnosis but cannot replace the testing by the golden standards: hydrogen breath test (HBT) or lactose tolerance test. Genetic differences between regions make this practice very controversial – a genotype association with the phenotype in one population do not automatically translate to other, unrelated, and unstudied human populations.

Our main objective in this study was to extract the genetic data available to us in Chapters 1 and 2 related to lactase persistence in Ukraine using WGS and evaluate it using the actual phenotype test of LP in Ukrainians from Ukraine. We started assessing the frequency of common and rare SNPs related to LP, estimated the haplotype length surrounding the SNPs, and searched for new variants related to the LP phenotype. Then we evaluated the relationship between the genotype and phenotypes by comparing alleles, genotypes and haplotypes in each individual to the results of a lactose hydrogen breath test. An estimation of the local prevalence of LP on the specific genetic background in Ukraine is of high importance because of a number of new alleles that have been discovered in this European population as a result of “pioneer advantage” after publishing new genome data[26]

3.2 Methods

3.2.1 Participants and study design

The study protocol was approved by IRB of Uzhhorod National University (protocol 5/2, 20.07.2021) and Oakland University (IRB-FY2022-108). A total of 154 individuals participated in this study, and all provided written informed consent before participation.

Among 154 participants, 100 were recruited from the pool of participants in the earlier projects described in this thesis: "Genome diversity in Ukraine"[2] (Chapter 1) and the cross-border cooperation project "Partnership for genomic research in Ukraine and Romania"[109]. All 100 participants provided written informed consent before participation to use their whole-genome sequences (WGS) to be matched with their phenotypes. For these participants, we extracted genetic data from the WGS data to determine their lactase persistence (LP) genotypes of the know associated alleles. For the remaining 54 participants, the *MCM6* gene, which is known to be associated with LP re-sequenced. The genomic and data were interrogated for genetic variation to be used for the subsequent genotype-phenotype correlation assessment.

3.2.2 Sample handling for WGS

Sampling strategy, DNA extraction, sequencing, reads alignment, mapping and variant calling was performed by the pipeline described in the Methods section of Oleksyk et al[2] and in Chapter 1. Variants were annotated using ANNOVAR[33] and SNPeff[34] compared to the latest version of the reference sequence of the human genome (GRCh38).

3.2.3 Sample handling for Sanger Sequencing

Genomic DNA was extracted from blood samples using a Genomic DNA purification kit (NEB®, Ipswich, MA, USA). The DNA was then amplified using standard

PCR with specific primers designed to amplify the target regions of interest in *MCM6* gene (Supplementary Table 3.1). The PCR products were then purified using an ExoSap-IT® (Thermofisher Scientific®, Detroit Michigan) enzymatic purification protocol[110] following the manufacturer's instructions.

The purified PCR products were then sequenced using the Sanger sequencing method with the Big Dye Terminator 3.1 kit (Thermofisher Scientific®, Detroit Michigan) on the Seq Studio Genetic Analyzer (Thermofisher Scientific®, Detroit Michigan). The sequencing reaction was set up according to the manufacturer's protocol, using the specific forward or reverse primer for each sample. The sequencing products were then purified using a purification beads purification method and then analyzed on the Seq Studio Genetic Analyzer using the default settings.

The resulting sequence data were then analyzed using the Geneious Prime® version 2022.2.2). The sequences were aligned to a reference human MCM6 gene (GRCh38), and the sequence variants were identified and annotated by Geneious Prime® inbuilt tools for chromatogram analysis.

All sequencing data were visually inspected for quality control, and any poor-quality data (variants with low signal intensities, baseline noise, Phered scores<30) were excluded from the analysis. The Geneious prime® software (version 2022.2.2) was used to automatically call bases and generate consensus sequences, and the resulting sequences were manually checked for accuracy.

3.2.4 Hydrogen Breath Test

All participants underwent the hydrogen breath test (HBT, Gastrolyzer® (Bedfont Scientific Ltd, Maidstone, UK).) to determine their lactase persistence status. Participants fasted overnight before the test and were asked to avoid consuming any food or beverages other than water for at least two hours before the test. The HBT was performed as follows: participants ingested a standard dose of lactose (50g) dissolved in 250 ml of water, and their breath hydrogen levels were measured at the baseline, 1 h and 2 h after ingestion using a Gastro+™Gastrolyzer® (Bedfont Scientific Ltd, Maidstone, UK). Lactose malabsorption was defined as a rise in breath hydrogen concentration of at least 20 parts per million (ppm) above the baseline.

3.2.5 Lactose intolerance symptoms assessment

The following five signs of lactose intolerance were assessed using a questionnaire validated by Casellas et al. [15]: audible bowel sounds, meteorism, stomach pain, vomiting, and diarrhea. The severity of these symptoms was rated by each research participant two hours following the test on a scale from 0 to 10. On this scale, “0” stood for no symptoms, and “10” stood for the strongest manifestation of lactose intolerance. The total score of the symptoms (SS), calculated as the sum of the intensities for each of the five symptoms, was 50 points. The participants were interviewed 2 h and 6 h after the lactose ingestion.

3.2.6 Phenotyping

Participants were classified as lactose intolerant or lactose tolerant based on their HBT results and score of the symptoms. Participants with a rise in breath hydrogen concentration of at least 25 ppm and any symptom present above baseline were considered

lactose intolerant, while those with a rise of less than 25 ppm and not symptoms were considered lactose tolerant. We then evaluated the concordance between the LP genotypes and the lactose intolerance phenotypes to determine the accuracy of predicting lactose intolerance status using genotyping tests.

3.2.7 Statistical Analysis

We performed statistical analyses to determine the sensitivity, specificity, positive likelihood ratio, and negative likelihood ratio of the LP genotyping in predicting lactose intolerance. We used the following formulas[111], [112]:

Sensitivity: Sensitivity = True Positives / (True Positives + False Negatives)

Specificity: Specificity = True Negatives / (True Negatives + False Positives)

Positive likelihood ratio (LR+): LR+ = Sensitivity / (1 - Specificity)

Negative likelihood ratio (LR-): LR- = (1 - Sensitivity) / Specificity

Accuracy: Accuracy = (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)

We utilized ANOVA (Analysis of Variance) to compare the means of Hydrogen pressure test across three groups genotype groups.

The threshold for statistical significance (p-value) was set at 0.05. The MS Excel 2010 was used for descriptive and inferential statistics. Omni calculator was used for sensitivity, specificity) positive likelihood ratio, and negative likelihood ratio and accuracy calculation (Omni Calculator®sp. z o.o., Poland).

3.3 Results

3.3.1 LP alleles frequencies in Ukraine

Of the two well-known LP-associated variants, rs4988235 (13910:C>T) and rs182549 (22018:G>A), both were present in the Ukrainian population, but at an almost half of the frequency recorded for the northern Europeans (CEU) [12]. The derived allele *rs4954490* (13495:C>T) located just outside the enhancer region showed a weaker association with *rs4988235* (13910:C>T) in Ukrainians compared to CEU population (Table 3.1). We also investigated the presence of known LP-associated variants from non-European populations in the Ukraine. However, none of the other known LP alleles were detected in our study population. This suggests that *rs4988235* and *rs4954490* may be the only two common LP-associated variants present in the Ukrainian population. These findings are consistent with previous studies that have reported a lower prevalence of LP in Eastern and Central Europe compared to Western Europe.

Table 3.1: Genotypes and allele frequency of SNPs associated with lactase persistence in Ukraine

N	***SNP***	**Known effect in populations**	**Gene position**	**MAF UAU97 and TRC150 (combined)**	**MAF EUR**[12]
1	*rs4954490*	-	13495:C>T	0.56	0.79
2	*rs56348046*	-	13603:C>T	0	0

Table 3.1: - continued

N	*SNP*	**Known effect in populations**	**Gene position**	**MAF UAU97 and TRC150 (combined)**	**MAF EUR**[12]
3	*rs4954492*	-	13730:T>G	1	1
4	*rs527791977*	-	13779:G>C	0	0
5	*ss820486563*	-	13806:A>G	0	0
6	*rs41525747*	-	13907:C>G	0	0
7	*rs4988235*	LP Europeans	13910:C>T	0.385	0.73
8	*rs41380347*	LP Arabs	13915:T>G	0	0
9	*rs869051967*	-	14009:T>C	0	0
10	*rs145946881*	LP Sub-Saharan Africans	14010:G>C	0	0
11	*rs759157971*	-	14028:T>C	0	0
12	*rs182549*	LP Europeans	22018:G>A	0.394	0.55

3.3.2 Hydrogen breath test results

The total of 154 participants were recruited for HBT, with 60 males (39.1%) and 94 females (60.9%). The table 3.2 summarizes the mean age and BMI of the participants, as well as their self-reported ancestry.

Table 3.2: Demographics of the cohort undergone HBT

SEX	Number of participants (%)	Mean age±SD	BMI±SD	Self-reported ancestry (% of participants)
Male	60 (39.1%)	43.2±15.6	28.2±3.5	96% Ukrainian
				4% Roma
Female	94 (60.9%)	39.9±14.2	26.5±4.1	100% Ukrainian

Among tested individuals 34.4% (53/154) had a positive HBT result (the rise of hydrogen level above 25 mm from the baseline) which is consistent with lactose malabsorption. Furthermore, 79.2% (42/53) of tested positive had LNP genotype of *rs4988235* (C/C) and lactose intolerance symptoms present at 2 h after the lactose load.

The mean total score of the symptoms (SS) was significantly higher in HBT+ group comparatively to HBT- group (10.4 and 1.04 respectively, $p<0.001$). Interestingly, only 2 *rs4988235* (C/T) heterozygous cases had a positive HBT, but no symptoms. All T/T homozygotes were HBT negative with no symptoms suggesting being complete lactose absorbers and lactose tolerant.

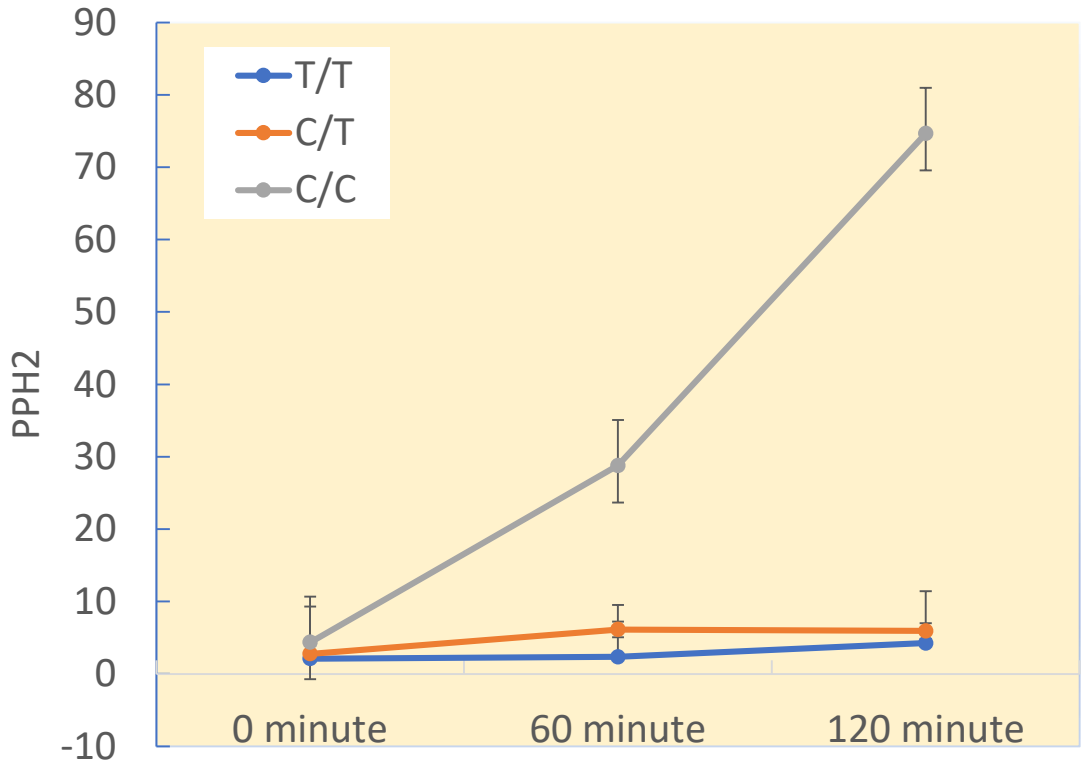

Figure 3.1: HBT results following the lactose load of 50 g in different *rs4988235* (C>T-13910) genotype groups. The y-axis represents ppmH2 and the x-axis represents time of the measurement

Table 3.3: Genotype-phenotype relation of *rs4988235* assessed by HBT and total score of symptoms.

rs4988235 **(C/T-13910) genotype**	**HBT Positive (%)**	**Mean total score of symptoms (SS)**	**Delayed symptoms (6 h, %)**
Homozygous dominant (TT)	**0**	**0**	**0**
Heterozygous (C/T)	**3.7**	**1.1**	**0**
Homozygous recessive (C/C)	**79.2**	**9.37**	**6**

3.3.3: Sensitivity and specificity of HBT test and *rs4988235* genotyping test

We also calculated sensitivity and the specificity[112] of the hydrogen breath test at 2h post ingestion to detect lactose malabsorption. (see table 3.4).

Table 3.4: Accuracy of human breath test (HBT) to detect lactose intolerance

		Lactose intolerant (symptomatic within 6 h)	
		Positive	**Negative**
HBT	**Positive**	50	3
	Negative	1	100
Sensitivity (%)		98.04%	
Specificity (%)		97.09%	
Positive likelihood ratio		33.66	
Negative likelihood ratio		0.02	
Accuracy		97.4%	

Next, we evaluated the widespread use of the RT-PCR method for detecting the *rs4988235* (C>T-13910) genotype in Ukraine, which is commonly employed as a diagnostic tool for predicting lactose intolerance. In addition to reporting the prevalence of lactase persistence and intolerance in the population, we also calculated the sensitivity and specificity of this method. These values are crucial in determining the accuracy and reliability of this diagnostic tool, as they provide insight into its ability to correctly identify individuals who are either lactose intolerant or lactase persistent. The results of these

calculations are presented in Table 3.5, which sheds light on the effectiveness of the RT-PCR method in predicting lactose intolerance in Ukraine.

Table 3.5: Accuracy of *rs4988235* (C>T-13910) genotype test to detect lactose intolerance.

		Lactose intolerant (symptomatic within 6 h)	
		Positive	**Negative**
rs4988235 **(C/T-13910) genotype**	**Positive (C/C)**	42	5
	Negative (C/T or T/T)	0	108
Sensitivity (%)		100%	
Specificity (%)		95.58%	
Positive likelihood ratio		22.6	
Negative likelihood ratio		0	
Accuracy		96.77%	

3.3.4 Rare alleles and their association with LP phenotype in Ukraine

There were several rare alleles in MCM6 gene associated with lactose persistence phenotypes in out cohort. We explored possible association of these alleles and LP. Rare allele *rs4988233* (C>T-14011) was reported to also upregulate LCT promoter activity by in vitro studies. It works in similar fashion to the well-studied neighboring variant

rs145946881 (C>T-14010), driving LP in sub-Saharan African populations. In our Ukrainian cohort this variant was present only in 2 heterozygous individuals (MAF 0.004). One of these individuals underwent HBT testing and did not confirm to be lactase persistent (positive HBT and symptoms of lactose intolerance (pH2 on 120 min: 56, total SS: 5). Hoever, this person was also a *rs4988235* (C/T-13910) C/C homozygote, which likely explains the expressed LNP phenotype.

Furthermore, we identified 5 individuals who were *rs4988235* (C/T-13910) C/C homozygotes, but expressed a complete LP phenotype (negative HBT, no symptoms). We assessed these subjects for their genotypes of well-known LP-associated alleles presented in the Table 3.1, but none of these alleles were identified. Thus, we continued our search for any other rare SNPs within the *MCM6* gene. We found that one out of the five individuals in question was also heterozygous for an extremely rare variant *rs1188475593* in the intron 13 of *MCM6* gene (T>C, global MAF 0.000019). This was also the only instance of this allele present in Ukrainian cohort (MAF 0.002). To explore further this genotype-phenotype relationship of *rs1188475593*, this individual was invited for the confirmation using HBT and lactose tolerance glucose test, which both confirmed complete lactase persistence phenotype. To further investigate the association between the *rs1188475593* genetic variant and lactose persistence (LP) phenotype, we conducted tests on a small sample size of seven blood relatives of the initial subject. Our aim was to determine if the same association observed in the initial subject was present in other members of the family, providing evidence for a hereditary link. Remarkably, we found a complete segregation between the *rs1188475593* variant and LP phenotype in 2 other relatives (Figure 3.2).

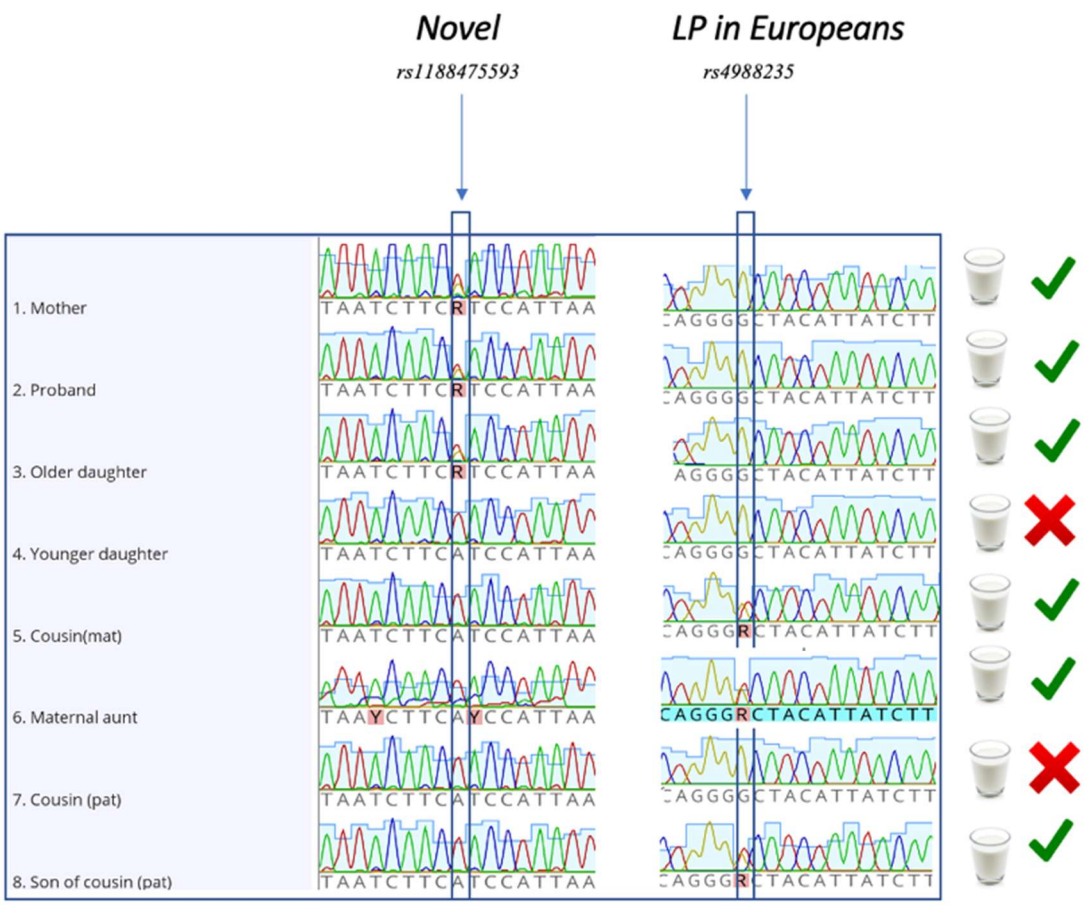

Figure 3.2. Association of *rs1188475593* and *rs4988235* in the family of LP rs4988235 C/C individual

3.4 Discussion

The current study employed both genetic and clinical testing methods to determine the prevalence of lactose persistence genotypes and phenotype in a population of Ukraine. Among the 12 SNPs evaluated in this study, only 2 (*rs4988235* (13910:C>T) and *rs182549* (22018:G>A) were variable in the Ukrainian population. Notably, *rs4988235* has a higher MAF in CEU compared to Ukrainian population, which is consistent with previous studies that have reported a higher prevalence of lactose tolerance in Western Europe compared to

Eastern and Central Europe[90]. None of the other known LP-associated variants were detected in the Ukrainian population, suggesting that *rs4988235* and *rs182549* may be the only two common LP-associated variants present in the Ukrainian population.

Only 34.4% of the participants expressed lactose malabsorption detected by HBT, which is significantly lower than reported by Storhaug et al. based on literature review. Expectedly, 94% of these individuals exhibited at least some symptoms of lactose intolerance of different severity according to the questionnaire. These results are consistent with the gradient prevalence of lactose intolerance in Europe which is reported to be at around 25% of the population[113].

We observed high correlation between LNP genotype C/C of *rs4988235* (C>T-13910) and HBT results supported the evidence of this SNP being a predominant driver of lactase persistence among Ukrainians likewise the rest of Europe. Nevertheless, five out of 47 *rs4988235* C/C homozygotes expressed absolute LP phenotype with complete absence of lactose intolerance symptoms, which suggests they might carry other LP-related alleles. Additionally, we have not observed the intermediate phenotype of *rs4988235* heterozygotes (C/T-13910), which indicates that this SNP determines the LP phenotype in the dominant mode of inheritance and the heterozygous phenotype must not be interpreted as a decreased lactase activity phenotype as it is instructed in the interpretations of some PCR-based tests systems currently used in Ukraine[114]

The results of our study demonstrate that both the hydrogen breath test (HBT) and *rs4988235* genotyping test can be used to detect lactose malabsorption and predict lactose intolerance. The HBT had a sensitivity of 98.04% and a specificity of 97.09%, while the *rs4988235* genotyping test had a sensitivity of 100% and a specificity of 95.58%. Our

results for these criteria lie on the high end of the previously reported numbers (sensitivity 76-94% and specificity 77-96% [115] The mean total symptom score reported in this study was 9.37, which is significantly higher than in heterozygotes (1.1) $p<0.05$. and T/T homozygotes (0) $p<0.05$. This observation supports the earlier inferences about the total score of 6.5 as an accurate cut-off to detect lactose intolerance using questionnaire method only[116].

Our findings suggest that both the HBT and *rs4988235* genotyping test are reliable and accurate methods for detecting lactose malabsorption and predicting lactose intolerance in population of Ukraine. Exceptional sensitivity of rs4988235 genotyping is attributable most probably to the low level of admixture from non-European countries[2] Nevertheless, its slightly lower specificity indicates that there might be other genetic or environmental drivers of LP in a small proportion of individuals in Ukraine[117][118].

Furthermore, the positive likelihood ratio for the HBT was 33.36, indicating that a positive HBT result is 33 times more likely to occur in lactose intolerant individuals than in those without lactose intolerance. Positive likelihood ratio for the *rs4988235* genotyping test was 22.6, indicating a lower chance that a positive result will occur in lactose intolerant individuals compared to HBT.

The identification of an extremely rare variant*, rs1188475593*, in the intron 13 of the *MCM6* gene possibly linked to lactase persistence (LP) phenotype as a rare driver is an important finding. This is the only variant which was unique in LP *rs4988235* C/C homozygote individual compared to the whole set of LNP *rs4988235* C/C homozygotes and located specifically in the intron 13, previously linked to lactase persistence[119]. It is

worth mentioning, that this subject had other unique variants up and downstream to the *LCT* and *MCM6* gene (Supplementary Table 1).

The testing of blood relatives of the individual with *rs1188475593* variant showed a 100% correlation between LP phenotype and *rs1188475593* polymorphism, indicating the possible heritability of LP phenotype trough this allele in this family. However, it is important to note that the small sample size of seven blood relatives limits the generalizability of these findings. Therefore, future studies should include a larger sample size to confirm the association between *rs1188475593* variant and LP phenotype. These findings also highlight the importance of considering rare variants in the analysis of the genotype-phenotype correlation. The absence of well-known LP-associated alleles in the five C/C homozygotes for *rs4988235* with complete LP phenotype suggests that there could be other rare variants that contribute to the expression of LP phenotype possibly through direct enhancing action or through DNA methylation [120].

3.5 Conclusions

This study found that lactose malabsorption and lactose intolerance are almost half as prevalent in the Ukrainian population than previously reported[79]. The hydrogen breath test and *rs4988235* genotyping test were both effective in predicting lactose intolerance, with the former having higher sensitivity, but slightly lower specificity. The identification of a rare variant, *rs1188475593*, possibly linked to lactase persistence phenotype in one individual highlights the importance of considering rare variants in genotype-phenotype correlation studies. However, the small sample size limits the generalizability of this finding, and future studies with larger sample sizes are needed to confirm the association.

CHAPTER FOUR

GENETICS OF GLOBAL DEVELOPMENTAL DELAY AND INTELLECTUAL DISABILITY IN CHILDREN AND ADOLESCENTS FROM UKRAINE

4.1 Introduction

Global developmental delay (GDD) and intellectual disability (ID) are terms used to describe individuals with significant delays in various developmental domains, including gross and fine motor skills, language and communication, and personal and social conduct [121]. While the designation GDD is reserved to children under the age of five, ID is applied for older children and adults. Both conditions are diagnosed when the standardized neurological tests fall two standard deviations below the age-appropriate mean [122]. While 40% of all GDD/ID cases are attributable to genetic disorders, other factors such as perinatal trauma, intrauterine infections, and toxic exposure, as well as postnatal events can also contribute to the developmental delay [123]. GDD/ID can also be accompanied by autism spectrum disorder (ASD)[124] as well as anatomical abnormalities of other organ systems[125].

Traditionally, chromosomal microarray (CMA) and fragile X syndrome testing have been the primary diagnostic approaches for GDD/ID[126] However, CMA can only detect chromosomal deletions or duplications in about 20% of genetic cases[127]. The advent of next-generation sequencing (NGS), specifically gene panel testing and whole exome sequencing (WES), has revolutionized the search for causative variants in neurodevelopmental disorders, increasing diagnostic success by an additional 25%[128]. NGS techniques, with improved testing methods and bioinformatic algorithms, can now

detect large copy number variations (CNVs) and chromosomal aberrations previously undetectable by NGS alone[129]. Even though CMA and WES are not interchangeable, with the current improvements in testing techniques and bioinformatic algorithms, NGS gene panels and WES can accurately find large CNVs and chromosomal aberrations previously doomed undetectable by NGS technique[130].

Syndromic genetic disorders are the leading cause of pediatric disability in Ukraine [131]. However, the diagnosis of developmental delay in Ukraine has been delayed due to limited newborn screening and only the recent adoption of NGS genetic testing by physicians[132] . Apart of isolated case reports, there has not been a comprehensive study on genetic causes of GDD/ID conducted in Ukraine, in particular, on diagnostic yield of NGS panels and WES techniques. This report will help pave the way to detecting locally significant candidate pathogenic variants for future variant resolution and familial studies in Ukraine and across Eastern Europe [131].

4.2 Methods

4.2.1 General study design

The cohort included a mixed set of individuals diagnosed with GDD/ID only, as well as GDD/ID patients with ASD and/or multiple congenital anomalies or other functional symptoms. The diagnostic protocol was performed according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5, APA 2013) [131] The patients were referred to a medical geneticist by either pediatrician or pediatric neurologist for consultation. All 416 children enrolled in the study underwent sequencing of whole exome sequencing (WES, Invitae Inc., San Francisco, CA) or custom broad neurodevelopmental

disorder (NDD) gene panel sequencing (Invitae Inc., San Francisco, CA). The medical geneticist obtained consent for testing and signed standardized requisition forms with optional clinical and demographic information. Family follow-up has not yet been performed with these patients' family members. The study of deidentified aggregated data was approved as "Not Human Subject Research" by the Institutional Review Board of Oakland University (Rochester, MI, Study #RB-FY2023-120).

4.2.2 NGS Neurodevelopmental disorders panel and whole exome sequencing

The neurodevelopmental disorders panel (NDD) included 1,813 genes (Supplementary Table 4.1, sequencing and report limitations specified in the Supplementary file 4.1). For the NDD panel, genomic DNA from the submitted samples was extracted and enriched for targeted regions using the hybridization-based protocol [133] For the WES sequencing, DNA libraries were prepared using the PCR-free method. The WES panel included a panel of more than 18,000 genes (Invitae Inc., San Francisco, CA).

All blood and saliva samples underwent double-step verification by visual identifiers (ID, Sex) and sex determined by sequencing, according to the company protocol (Invitae Inc., San Francisco, CA). All targeted regions were sequenced on Illumina NovaSeq 6000 platform (Illumina, San Diego, CA, USA). The average coverage for NDD panel testing was 50x, while for WES it was 35X across the entire exome.

Read mapping was performed to the reference GRCh37 human genome. To categorize the variants according to the laboratory, several pieces of evidence were

considered, such as variant frequency and type, clinical findings, experimental research, and indirect and computational approaches.

Before being reported, clinically important variation that failed to meet strict NGS quality parameters had its accuracy verified by alternative methods [133], Sherloc [134], a points-based framework based on the joint consensus recommendations from the American College of Medical Genetics and Genomics and the Association for Molecular Pathology [135], was used to analyze the variations discovered by the bioinformatics pipeline.

4.2.3 Genomic data analysis

First, a dataset of reported variants for each individual with only sex, age, and phenotypical description data was compiled in Ukraine and the data were transferred to Oakland University, MI (USA). Due to the heterogeneous nature of data provided in the reports, the alleles with missing explicit rs-code were cross-referenced by their respective allele and protein notation according to Sequence Variant Nomenclature specifications, using the ENSEMBL Variant Recoder[136], resulting in the genomic positions of each reported variant for GRCh38 genomic reference and their relevant rs-codes. After the validation, we performed a detailed search on reported pathogenic (P) or likely pathogenic (LP) variants using ClinVar[137] and Online Mendelian inheritance in Man (OMIM)[138] databases.

Variants of uncertain significance (VUS) or heterozygous variants related to the autosomal recessive condition were considered non-diagnostic. Using genome data from 97 individuals from the “Genome Diversity in Ukraine” database (Chapter 1) [139], [140],

as well as from the 150 whole genomes from the database of the cross-border cooperation project “Partnership for Genomic research in Ukraine and Romania” [141], we performed an additional annotation for the effect prediction among the VUS using CADD and SIFT scores and estimated their frequency in the general population of Ukraine. The CADD (Combined Annotation Dependent Depletion) Phred score[142] is a measure used to evaluate the potential pathogenicity of genetic variants. It quantifies the deleteriousness of a variant by integrating multiple annotations, such as conservation, functional impact, and disease association. A higher CADD Phred score corresponds to a higher predicted pathogenicity or potential functional impact of the variant. The SIFT (Sorting Intolerant from Tolerant) score[143] is a prediction tool used to assess the potential impact of amino acid substitutions caused by genetic variants. It determines whether a substitution is likely to affect protein function based on sequence conservation across species. The SIFT algorithm assigns a score to each variant, ranging from 0 to 1, where a score closer to 0 indicates a higher likelihood of being deleterious or damaging to protein function, while a score closer to 1 suggests the variant is likely to be tolerated or benign. The SIFT score helps prioritize variants in genetic analysis by identifying those with a higher potential for functional impact[143].

4.3 Results

4.3.1 Demographic and clinical characteristics of the cohort

The cohort included 416 exclusively pediatric patients under 18 years old (age ranged between 1-18 years), with 60.9% males. Both sexes, males and females, had a similar mean age of around 7 years old (Table 4.1). Genetic information either from the

NDD gene panel or from WES results as a first- or second-line test after inconclusive CMA was available for analysis in all individuals in this study. Diagnostic data for karyotypes, chromosomal microarrays, or FMR1 CGG-repeat expansion tests for the Fragile X syndrome, was not included in this study. Demographic and clinical information included in the analysis is summarized in Table 4.1.

Table 4.1: Demographic and clinical characteristics of the studied cohort

Characteristic	**Cohort (N=416)**
Sex	N (%)
Male	253 (60.9)
Female	163 (39.1)
Age (At The Time Of The Testing)	
Mean (SD) Males	7.41 (4.04)
Mean (SD) Females	7.37 (3.86)
Comorbidities	
ASD N (%)	230 (55.1)
Confirmed Congenital	69 (16.5)
Anomalies/Additional symptoms N (%)	
Epilepsy N (%)	121 (29.08)
Genetic Testing	
Neurodevelopmental Disorders Panel	379 (91.1)
N (%)	
WES	37 (8.8)

4.3.2 Yield of definitive molecular diagnosis

We identified a definitive molecular diagnosis of GDD/ID in 66 or 16.3% of all individuals (Fig. 1). In general, WES positively diagnosed 22 out of 37 ordered cases (59.4%), while the NGS testing panel yielded 44 definitive diagnoses among the 379 tested patients (12.1%). Non-diagnostic variants (VUS and carrier) were identified in 348 (83.4%) individuals (details in Fig. 4.1).

Most of the known diagnosed conditions followed the AD mode of inheritance (41, or 62.11%), four with AR, nine with XLD, and two with XLR modes (Table 4.2). Chromosome 15 was most affected by these types of variants. The most diagnosed same-gene condition was Rett syndrome: five cases were caused by single nucleotide variants (SNVs) or small indels in the *MECP2* gene. Among the other diagnoses, 12 different conditions were observed in two individuals each, all the rest diseases were isolated cases only (Table 2). Large copy number variations (CNVs) encompassing multiple genes were observed in 10 individuals (15.1% of diagnosed cases) (Table 4.3).

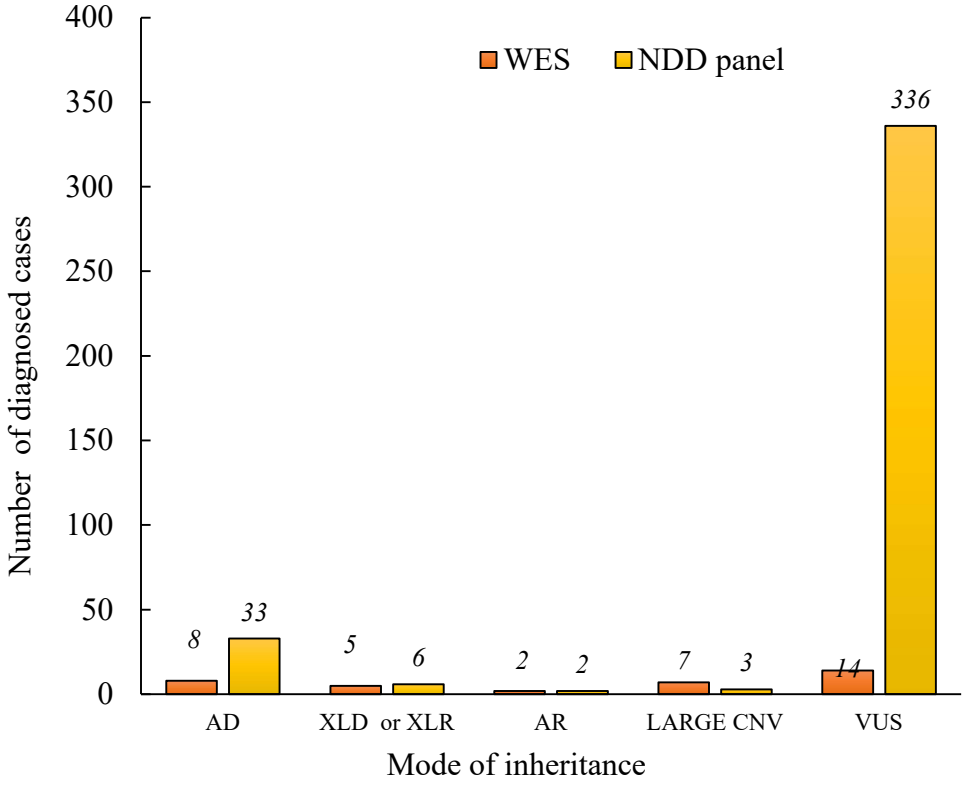

Figure 4.1: The absolute number diagnosed conditions by mode of inheritance: autosomal dominant (AD), autosomal recessive (AR), X-linked dominant (XLD), and X-linked recessive (XLR) (Table 4.2) diagnosed by either WES or NDD panel. Large CNVs are shown separately (Table 4.3).

Table 4.2: Summary of diagnosed conditions and causative variants in the cohort, classified by the mode of inheritance (see Fig .4.1)

N	*Patient ID*	*Gene*	Codon and amino acid change	Position on chromosome GRCh38.p13	Chromosome	Number of diagnosed cases	OMIM ID	Disease
		Autosomal dominant mode of inheritance (N=41)						
1	*190*	*ATP1A3*	c.2851G>A (p. Glu951Lys)	41967732	19q13.2	1	614820	Alternating hemiplegia of childhood 2
2	*55*	*C19OFR12*	c.204_214del (p. Gly69Argfs*10)	29702957-29702977	19q12	1	614298	Neurodegeneration with brain iron accumulation 4
3	*147*	*CHD1*	c.206C>G	98904946	5q15-q21.1	1	617682	Pilarowski-Bjornsson syndrome
4	139	CTBP1	c.991C>T(p.Arg331Trp)	1213028	4p16.3	1	617915	Hypotonia, ataxia, developmental delay, and tooth enamel defect syndrome
5	*278*	*DEPDC5*	c.1325-1G>A (Splice acceptor)	31810520	22q12.2-q12.3	1	604364	Epilepsy, familial focal, with variable foci 1
6	*14*	*DYRK1A*	c.1294G>T (stop gained)	37505364	21q22.13	1	614104	Intellectual developmental disorder, autosomal dominant 7

Table 4.2: - continued.

N	*Patient ID*	*Gene*	Codon and amino acid change	Position on chromosome GRCh38.p13	Chromosome	Number of diagnosed cases	OMIM ID	Disease
7	*101*	*FOXG1*	c.701C>T (p.Ser234Phe)	28767980	14q12	2	613454	Rett syndrome, congenital variant, FOXG1 syndrome
8	*392*		c.587A>C (p.Gln196Pro)	28767866				
9	*108*	*GABRB3*	c.288G>T (p.Arg96Ser)	26621487	15q12	2	617113	Developmental and epileptic encephalopathy 43
10	*23*		c.905A>G (p.Tyr302Cys)	26561107				
11	*314*	*GLRA1*	c.381dup (p.Phe128Leufs*11)	15185987	5q33.1	1	149400	Hyperekplexia 1
12	*161*	*GNB1*	c.239T>C (p.Ile80Thr)	1806503	1p36.33	1	616973	Intellectual developmental disorder, autosomal dominant 42
13	*151*	*KCNA2*	c.997T>C (p.Phe333Leu)	110603786	1p13.3	2	616366	Developmental and epileptic encephalopathy 32
14	*376*		c.1219C>T (p.Pro407Ser)	110603564				
15	*65*	*KCNQ2*	c.1639C>T (p.Arg547Trp)	63413574	20q13.33	1	613720	Developmental and epileptic encephalopathy 7
16	*237*	*KCNT1*	c.1309C>T (p.Leu437Phe)	135765732	9q34.3	2	614959	Developmental and epileptic encephalopathy 14

Table 4.2: - continued.

N	*Patient ID*	*Gene*	Codon and amino acid change	Position on chromosome GRCh38.p13	Chromosome	Number of diagnosed cases	OMIM ID	Disease
17	*323*		c.784C>T (p.Arg262Trp)	135758438				
18	*131*	*KDM1A*	c.2410dupA	23082329	1p36.12	1	616728	Cleft palate, psychomotor retardation, and distinctive facial features
19	*68*	*KMT2C*	c.8965_8970delinsAGTACCTT (p.Val2989Serfs*44)	118504857	7q36.1	1	617768	Kleefstra syndrome 2
20	*290*	*KMT2D*	c.14710C>T (p.Arg4904*)	49027256	12q13.13	1	147920	Kabuki syndrome 1
21	*134*	*MACF1*	c.7661A>G	39382151	1p34.3	1	618325	Lissencephaly 9 with complex brainstem malformation
22	*44*	*KMT2A*	c.2968_2969insAGAG (p.Cys990*)	118474126	11q23.3	2	605130	Wiedemann-Steiner syndrome
23	*72*		c.1038del (p.Val347fs)	118472196				
24	*172*	*NPRL3*	Exon 2-6 deletion	112622 – 138334	16p13.3	1	617118	Epilepsy, familial focal, with variable foci 3
25	*182*	*PAFAH1B1*	c.1159+2T>A (Splice donor)	2680322	17p13.3	2	607432	Lissencephaly 1

Table 4.2: - continued.

N	*Patient ID*	*Gene*	Codon and amino acid change	Position on chromosome GRCh38.p13	Chromosome	Number of diagnosed cases	OMIM ID	Disease
26	*284*		c.656G>A (p.Trp219*) (mosaic)	2674239				
27	*150*	*PTPN11*	c.922A>G (p.Asn308Asp)	112477719	12q24.13	1	163950	Noonan syndrome 1
28	*389*	*SCN1A*	c.4073G>T (p.Trp1358Leu)	166002683	2q24.3	2	607208	Dravet syndrome
29	*403*		c.4265A>G (p.Tyr1422Cys)	166002491				
30	*306*	*SCN2A*	c.2552C>A (p.Ser851*)	65342459	2q24.3	2	613721	Developmental and epileptic encephalopathy 11
31	*283*		Exon 17 deletion					
32	*311*	KCNC1	c.22G>T(p. .Glu8Ter)	17736024	11p15.1	1	616187	Epilepsy, progressive myoclonic 7
33	*130*	*STXBP1*	c.1606delC	127682459	9q34.11	2	612164	Developmental and epileptic encephalopathy 4
34	*224*		c.175G>A (p.Glu59Lys)	127658380				
35	*257*	*SYNGAP1*	c.1564del (p.Glu522Lysfs*5)	33438273	6p21.32	2	612621	Developmental disorder, autosomal Intellectual dominant 5
36	*126*		c.1534G>T (p.Glu512Ter)	33438777				
37	*142*	*TGFBR1*	c.844T>C (p.Tyr282His)	99142574	9q22.33	1	609192	Loeys-Dietz syndrome 1

Table 4.2: - continued.

N	*Patient ID*	*Gene*	Codon and amino acid change	Position on chromosome GRCh38.p13	Chromosome	Number of diagnosed cases	OMIM ID	Disease
38	*116*	*TLK2*	c.754C>T p.Gln252Ter	62560049	17q23.2	1	618050	Intellectual developmental disorder, autosomal dominant 57
39	*186*	*TREX1*	c.341G>A (p.Arg114His)	48466996	3p21.31	2	225750	Aicardi-Goutieres syndrome 1
40	*378*		c.341G>A (p.Arg114His)	48466996				
41	*145*	*TRRAP*	c.6653A>C	98967697	7q22.11	1	618454	Developmental delay with or without dysmorphic facies and autism
Autosomal recessive mode of inheritance (N=4)								
42	16	ARSA	c.465+1G>A (Splice donor)	50627165	22q13.33	1	250100	Metachromatic leukodystrophy
			c.542T>G (p.Ile181Ser)	50626976				
43	180	FBXL4	c.45T>G (p.Tyr15*) c.627_633del (p.Asn210Leufs*9)	98926944	6q16.1-16.2	1	615471	Mitochondrial DNA depletion syndrome 13 (encephalomyopathic type)

Table 4.2: - continued.

N	*Patient ID*	*Gene*	Codon and amino acid change	Position on chromosome GRCh38.p13	Chromosome	Number of diagnosed cases	OMIM ID	Disease
44	56	NPC1	c.2861C>T (p.Ser954Leu)	23539405	18q11.2	1	257220	Niemann-Pick disease, type C1
			c.1026G>A (p.Trp342*)	23556543				
45	332	PAH	c.1222C>T (p.Arg408Trp)	102840493	12q23.2	1	261600	Phenylketonuria
			c.473G>A (p.Arg158Gln)	102866632				
		X-linked dominant (N=9)						
46	*42*	*CDKL5*	c.372_385del (p.His124Glnfs*2)	18579937	Xp22.13	1	300672	Developmental and epileptic encephalopathy 2
47	*251*	*MECP2*	c.397C>T (p.Arg133Cys)	154031431	Xq28	5	312750	Rett syndrome
48	*83*		c.806del (p.Gly269Alafs*20)	154031025				
49	*414*		c.1084_1216del (p.Pro362Argfs*3)	154030743				
50	*106*		c.844del (p.Arg282fs)	154031020				
51	*95*		c.1115C>A (p.Ser372Ter)	154030749				
52	*157*	*SLC35A2*	c.845G>A	48905064	Xp11.23	1	300896	Congenital disorder of glycosylation, type IIm

Table 4.2: - continued.

N	*Patient ID*	*Gene*	Codon and amino acid change	Position on chromosome GRCh38.p13	Chromosome	Number of diagnosed cases	OMIM ID	Disease
53	*60*	*WDR45*	c.1013_1014del (p.Phe338Tyrfs*3)	49074874	Xp11.23	2	300894	Neurodegeneration with brain iron accumulation 5
54	*61*		c.64del (p.Cys22Alafs*16)	49077902				
		X-linked recessive (N=2)						
55	*400*	*ATP7A*	c.2938C>T (p.Arg980Ter)	78029271	Xq21.1	1	309400	Menkes disease
56	*200*	*ARX*	c.1058C>T (p.Pro353Leu)	25012937	Xp21.3	1	308350	Developmental and epileptic encephalopathy 1

Table 4.3: Diagnostic large CNV spanning multiple genes

Case N	Patient ID	Mutation	Genes duplicated/ deleted	# of allele copies	Condition	OMIM or Orphanet ID
1	276	2q37 deletion	AGXT; D2HGDH; KIF1A; NDUFA10	1	Chromosome 2q37 deletion syndrome	600430
2	352	2p16.3 deletion	*NRXN1*, Exons 2-3	1	2p16.3 deletion syndrome	614332
3	51	4.16 duplication	*CC2D2A; CPLX1; CTBP1; EVC; EVC2; FGFRL1; IDUA; KLB; LETM1; PIGG; QDPR; RBPJ; SEPSECS; WHSC1*	3	4.16 microduplication syndrome	ORPHA:96072

Table 4.3: - continued.

Case N	Patient ID	Mutation	Genes duplicated/ deleted	# of allele copies	Condition	OMIM or Orphanet ID
4	84	5p13 duplication	*SLC6A19; AMACR; HCN1; NADK2; NDUFS6; SLC6A3*	mosaic	Chromosome 5p13 duplication syndrome	613174
5	328	15q24 deletion	*CYP11A, SIN3A*	1	15q24 deletion syndrome, Witteveen-Kolk syndrome	613406
6	189	15q11.2 deletion	*UBE3A; GABRB3*	1	Angelman syndrome	615656
7	398	15q11.2 duplication	*UBE3A; GABRB3*	3	Chromosome 15q11-q13 duplication syndrome	608636
8	371	17p11.2 deletion	*ALDH3A2; TOP3A; ATPAF2*	1	17p11.2 deletion syndrome	182290
9	353	20P duplication	*ATRN; ITPA; NDUFAF5; PANK2; PDYN; PLCB1; PRNP; SNRPB; TBC1D20*	3	Trisomy 20p	ORPHA:261318
10	66	X28 duplication	*FLNA; NAA10; MECP2*	2	Intellectual developmental disorder, X-linked syndromic, Lubs type	300260

The prevailing group of diseases diagnosed were classified as developmental and epileptic encephalopathies (type 1, 2, 4, 7, 11, 14, 32 and 43) characterized by severe epilepsy and GDD/ID (14 cases). The specific variants reported for each definitive diagnosis are reported in the Table 4.2.

4.3.3 Annotation and analysis of variants of uncertain significance

A total of 3,317 heterozygous variants of uncertain significance (or VUS) were identified in our cohort of 417 patients. In this VUS dataset, a CADD-Phred score between 10 and 20 were associated with 245 variants (considered 10% most deleterious substitutions in the human genome), while for 723 variants it was above 20 (the top 1% most deleterious variants) [144]. A deleterious SIFT-prediction score [144] for least one alternative transcript was calculated for 527 variants (Figure 4.2; see details in Supplementary table 4.2).

Among these deleterious variants identified in the study, 165 were in the genes associated with AD conditions and 56 associated with X-linked dominant or recessive conditions according to OMIM (a total of 221). Here we reported these variants as potentially diagnostic and suggested a VUS resolution by proband's family VUS resolution testing in 138 undiagnosed cases (Supplementary table 4.2). However, 18 of these 221 variants had allele count 1 or 2 in 247 among the unaffected individuals from Ukraine (Supplementary Table 4.2). The rest of the variants (203) were absent in healthy individuals. Gene *PACS2* (associated with AD developmental and epileptic encephalopathy 66, OMIM 618067) was most frequently altered (seven individuals harboring rare highly deleterious heterozygous SNV) (Supplementary Table 4.2).

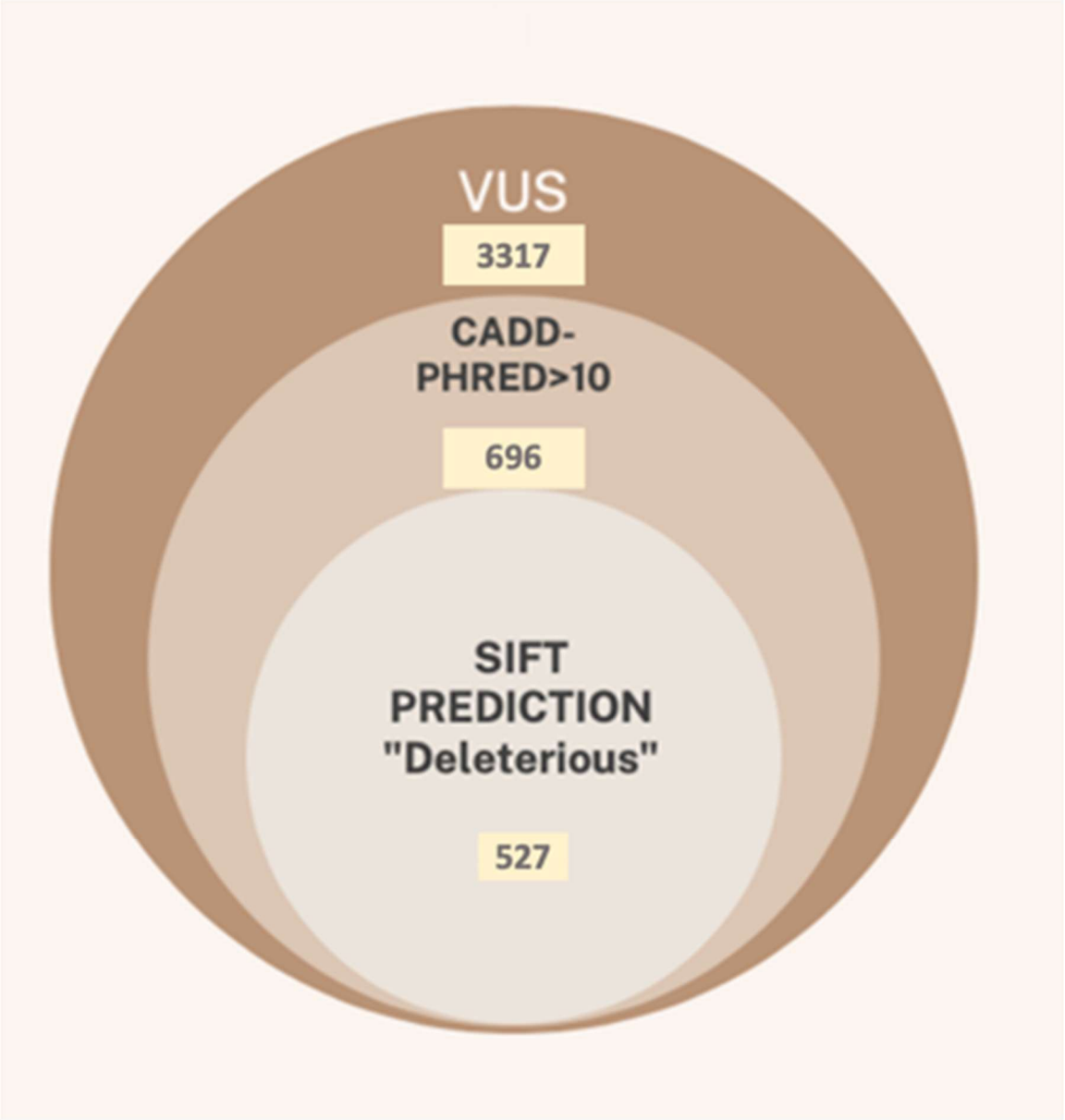

Figure 4.2: The Venn diagram showing an overlap in distributions of alleles with high CADD-Phred score [144] and “deleterious” SIFT-prediction score [144].

4.4 Discussion

Global developmental delay and intellectual disability (GDD/ID) is usually diagnosed for the patients with developmental delay before the exact genetic diagnosis is established[144]. GDD/ID is a complex set of symptoms with a wide range of genetic causes, including single nucleotide variants, large chromosomal indels, and copy number variants. However, in Ukraine a comprehensive study on the genetic causes of GDD/ID has not been conducted yet. This is mainly due to only the recent availability of NGS-based diagnostic tests. Also, the whole-genome data on the general genetic composition of the population has just been published recently [140], [145], [146], there was a need to use genome data available to evaluate the diagnostic yield of WES and NGS gene panel.

In this study, we report the largest to-date descriptive dataset of diagnosed genetic conditions which present with GDD/ID as a part of the clinical picture. Also, we report a combined diagnostic yield of the NDD gene panel of 1813 genes and WES at 16.3% on previously undiagnosed cases. Expectedly, individually WES had a much higher diagnostic yield compared to the NDD gene panel (59.4 % and 12.1% respectively). Similarly, the reported diagnostic yield of target exome sequencing in patients with ID ranges from 21% to 55.7%. Pekeles et al. [144]. used four distinct panels in a trial with a sample of 48 patients and achieved a 21% rate of definitive diagnoses. With a sample of 133 patients, Yamamoto et al [147] achieved a diagnostic rate of 29.3%. A similar rate of 34% was reported by Gieldon and colleagues in a study using four 813 gene panels in 106 patients[148]. It is possible that the significant difference in genetic yield between individuals who underwent exome sequencing and those who received targeted sequencing could be attributed to a chance. The limited sample size of 37 WES administered might have led to some variability

in the results. In such a small sample, the observed difference could have occurred randomly without any underlying phenotypic differences between the groups. Also, the NDD panels were prescribed in many cases due to their significantly lower cost. However, it is also important to consider other factors that could contribute to the observed discrepancy. Exome sequencing is a more comprehensive approach compared to targeted sequencing, as it examines a larger portion of the genome. This broader coverage increases the likelihood of identifying disease-causing genetic variants, leading to a higher genetic yield. Additionally, exome sequencing may capture variants in genes that are not initially suspected based on the clinical presentation but still contribute to the observed phenotype.

Our study showed that both the NGS gene panel and WES can be diagnostic of large CNVs associated with the clinical picture of known syndromes in the absence of CMA testing. In these 10 cases, both the NDD gene panel and WES reported multiple whole genes deleted or multiplied, which was indicative of the cytogenic location to determine large aberration from the set of genes mutated.

Most of the reported variants in our cohort were variants of uncertain significance (VUS). Both patients and medical geneticists face challenges as a result of the discovery of a VUS[149]. A VUS may ultimately be reclassified as pathogenic or benign, but this process often takes several years and may never be completed for rare VUS, particularly if the condition is uncommon to find enough cases or too expensive to test relatives[150]. The clinical relevance of a VUS has been increasingly determined by a phenotypically driven in-silico approach [151]. Furthermore, variant interpretation can be enhanced by quantitative analysis of consortium disease cohorts and population controls [152].

The absence of family members' genetic data was a major limiting factor to fully classifying or resolving the effect of the variants of uncertain significance in our study. Also, this prevented us from identifying de novo variants. Thus, we performed their annotation using CADD and SIFT predictive scores and found that as many as 527 variants were classified as deleterious by both scores (CADD-Phred>10 and SIFT prediction "Deleterious") with MAF<0.01 and should be resolved for disease causation by family testing or phenotype confirmation tests. Out of these, 221 variants were associated with AD or X-linked conditions making them potentially diagnostic. This number of variants resolved could potentially increase the diagnostic yield in 138 undiagnosed cases (by 33%). Interestingly, having WGS and phenotype data of 249 Ukrainians, we found that 18 of 221 potentially diagnostic variants are present at very low frequency in unaffected individuals (Supplementary Table 4.2), implying they might not be disease causative even with high prediction scores. Other 203 variants were absent in the sample of healthy individuals. Importantly, some of the VUS with high prediction scores and associated with AD or X-linked conditions were found in diagnosed individuals, potentially making their condition associated with multiple genetic aberrations.

4.5 Conclusions

This is the first comprehensive study on genetic causes of GDD/ID conducted in Ukraine, including diagnostic yield of NDD gene panel and WES techniques, comparing data from the cohort of pediatric patients and general population in the country. We report the largest to date descriptive dataset of diagnosed genetic conditions that present with GDD/ID as a part of the clinical picture. Our results support the important role of NGS gene panel and WES in the diagnostic approach to GDD/ID cases as a first line choice in

the scenario of scarce financial resources and logistic difficulty to perform multiple genetic tests in a family. A comprehensive approach to VUS resolution including computational effect prediction, comparative analysis of allele frequencies in population controls and phenotype assessment can be of extreme help to fully classify deleterious variants and narrow down the list for the further family follow-up testing.

APPENDIX A

IRB APROVALS

OAKLAND UNIVERSITY

Institutional Review Board

May 19, 2021

Protocol #: IRB-FY2021-346

Research Team:
Taras Oleksyk

Per federal regulations, the Oakland University IRB has determined that there is No Engagement in Research for Oakland University in the following study, "Genome diversity in Ukraine"

The IRB decision is based on the following:

The Oakland University PI, Dr. Oleksyk, will only be receiving coded biospecimens. A letter from Uzhhorod National University, assuring no codes will be released to any research collaborators has been provided in this submission.

Please note that since this project involves the use of biospecimens at Oakland University, the Biosafety Officer is copied on this IRB “Not Engaged” determination letter.

Please retain a copy of this correspondence for your records.

If you have any questions, please contact the IRB staff.

OAKLAND UNIVERSITY

Institutional Review Board

May 19, 2021

Protocol #: IRB-FY2021-347

Research Team:
Taras Oleksyk

Per federal regulations, the Oakland University IRB has determined that there is No Engagement in Research for Oakland University in the following study, "Partnership for Genomic Research in Ukraine and Romania"

The IRB decision is based on the following:

The Oakland University PI, Dr. Oleksyk, will only be receiving de-identified biospecimens that cannot be linked back to donors. A letter from the Uzhhorod National University confirming that only de-identified biospecimens will be provided to the OU PI has been provided in this submission.

Please note that since this project involves the use of biospecimens at Oakland University, the Biosafety Officer is copied on this IRB “Not Engaged” determination letter.

Please retain a copy of this correspondence for your records.

OAKLAND UNIVERSITY

Institutional Review Board

Date: October 18, 2021

Study #: IRB-FY2022-108

Study Tile: Evaluating prevalence and genotype-phenotype relationship of lactase persistence in Ukraine

Submission Type: Initial

IRB Decision: No Engagement in Research

Research Team:
Khrystyna Shchubelka
Taras Oleksyk

Per federal regulations, the Oakland University IRB has determined that there is No Engagement in Research for Oakland University in the above referenced study.

The IRB decision is based on the following:

- In this collaborative research project between Uzhhorod National University (Ukraine) and Oakland University, Oakland University researchers will not be conducting any human subjects research activities as they do not participate in the recruitment or consent processes, sample/data collection, or analysis of identifiable data.
- Researchers at the Uzhhorod National University (Ukraine) will be collecting hydrogen breath test (HBT) and responses to questionnaire under approval from the Uzhhorod National University (Ukraine) Ethics Committee Regarding the Observance of Moral and Legal Rules of Conducting Biomedical Research (a copy of the approval letter dated 7/20/2021 is included in the submission).
- Oakland University will only receive coded HBT test results associated with the participants from the Uzhhorod National University, and will not have access to any personal identifiers or the master list that deciphers the codes. A non-release confirmation letter from the dean of the biology department at the Uzhhorod National University (Ukraine) is included in this submission.

OAKLAND UNIVERSITY

Institutional Review Board

Date: December 5, 2022

Study #: IRB-FY2023-120

Study Tile: Genetic causes of global developmental delay in Ukraine

Submission Type: Initial

IRB Decision: No Human Subjects Research

Research Team:
Taras Oleksyk

The above referenced study has been has been determined to be No Human Subjects Research according to federal regulations.

The IRB decision is based on the following:

Research Using De-identified Human Biospecimens: The project is limited to the use of existing and/or prospectively collected human biospecimens provided ALL of the following criteria are met:
i. The biospecimens were/are not collected specifically for the currently proposed research through an intervention or interaction with living individuals; **and**
ii. The researcher can confirm that the use of the biospecimens is not in violation of the terms under which the biospecimens were originally collected; **and**
iii. The researcher will only receive biospecimens that are fully de-identified. De-identified means that the identity of the individual cannot be readily ascertained or does not include any of the 18 identifiers defined in the HIPAA Privacy Rule. De-identified biospecimens cannot be linked back to individuals directly or through coding systems; **and**
iv. Biospecimens are NOT being used to test the effectiveness of a medical device or as a control in an investigation of an investigational device and the results of the activity are to be submitted to the FDA or held for inspection by the FDA.

This is a retrospective analysis of 417 de-identified comprehensive patient records that the OU researcher, Dr. Taras Oleksyk, will be receiving from Dr. Liudmyla Turova, MD-PhD, Doctor of Genetic Medicine, Ukraine, who is the owner of the patient records.

The IRB submission includes a letter from Dr. Liudmyla Turova confirming the following:

- The OU researcher, Taras Oleksyk , will be provided with only de-identified patient records that were collected in Dr. Turova's practice for the purpose of conducting data analysis.
- Patients' identifying information is not known to the researchers at Oakland University and will not be released to them under any circumstances.
- In case of the need to re-contact the study participants, it will be done by the person not related to the study at Oakland University.

A fully executed data use agreement (DUA) indicating that indicate that Dr. Taras Oleksyk will only be provided with de-identified genomic information is also provided in this submission.

REFERENCES

[1] P. Eberhardt and J. Owsinski, *Ethnic groups and population changes in twentieth century Eastern Europe: history, data and analysis: history, data and analysis*. Routledge, 2015.

[2] T. K. Oleksyk *et al.*, "Genome diversity in Ukraine," *Gigascience*, vol. 10, no. 1, p. giaa159, 2021.

[3] M. A. Beaumont, "Recent developments in genetic data analysis: What can they tell us about human demographic history?," *Heredity*, vol. 92, no. 5. 2004. doi: 10.1038/sj.hdy.6800447.

[4] P. Kubicek, " The Gates of Europe: A History of Ukraine . By Serhii Plokhy. (New York, NY: Basic Books, 2015. Pp. xxiv, 395. $29.99.) ," *The Historian*, vol. 79, no. 4, 2017, doi: 10.1111/hisn.12732.

[5] N. Hamel *et al.*, "On the origin and diffusion of BRCA1 c. 5266dupC (5382insC) in European populations," *European Journal of Human Genetics*, vol. 19, no. 3, pp. 300–306, 2011.

[6] E. Check, "How Africa learned to love the cow." Nature Publishing Group, 2006.

[7] W. Lukito, S. G. Malik, I. S. Surono, and M. L. Wahlqvist, "From'lactose intolerance'to'lactose nutrition'," *Asia Pac J Clin Nutr*, vol. 24, no. Supplement, 2015.

[8] L. Deng and S. Xu, "Adaptation of human skin color in various populations," *Hereditas*, vol. 155, no. 1, pp. 1–12, 2018.

[9] M. L. Benton, A. Abraham, A. L. LaBella, P. Abbot, A. Rokas, and J. A. Capra, "The influence of evolutionary history on human health and disease," *Nature Reviews Genetics*, vol. 22, no. 5. 2021. doi: 10.1038/s41576-020-00305-9.

[10] S. A. Tishkoff and B. C. Verrelli, "Patterns of human genetic diversity: implications for human evolutionary history and disease," *Annu Rev Genomics Hum Genet*, vol. 4, no. 1, pp. 293–340, 2003.

[11] S. J. Chapman and A. V. S. Hill, "Human genetic susceptibility to infectious disease," *Nat Rev Genet*, vol. 13, no. 3, pp. 175–188, 2012.

[12] N. Siva, "1000 Genomes project." Nature Publishing Group, 2008.

[13] A. G. Clark, M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen, "Ascertainment bias in studies of human genome-wide polymorphism," *Genome Res*, vol. 15, no. 11, pp. 1496–1502, 2005.

[14] G. Pálsson and P. Rabinow, “Iceland: the case of a national human genome project,” *Anthropol Today*, vol. 15, no. 5, pp. 14–18, 1999.

[15] U. consortium and others, “The UK10K project identifies rare variants in health and disease,” *Nature*, vol. 526, no. 7571, pp. 82–90, 2015.

[16] K. Karczewski and L. Francioli, “The genome aggregation database (gnomAD),” *MacArthur Lab*, 2017.

[17] T. R. Magalhães *et al.*, “HGDP and HapMap analysis by Ancestry Mapper reveals local and global population relationships,” *PLoS One*, vol. 7, no. 11, p. e49438, 2012.

[18] J. Dopazo *et al.*, “267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation,” *Mol Biol Evol*, vol. 33, no. 5, 2016, doi: 10.1093/molbev/msw005.

[19] M. Via, C. Gignoux, and E. G. Burchard, “The 1000 Genomes Project: new opportunities for research and social challenges,” *Genome Med*, vol. 2, no. 1, pp. 1–3, 2010.

[20] S. Marwaha, J. W. Knowles, and E. A. Ashley, “A guide for the diagnosis of rare and undiagnosed disease: beyond the exome,” *Genome Medicine*, vol. 14, no. 1. 2022. doi: 10.1186/s13073-022-01026-w.

[21] I. Schrijver, E. Oitmaa, A. Metspalu, and P. Gardner, “Genotyping microarray for the detection of more than 200 CFTR mutations in ethnically diverse populations,” *The Journal of Molecular Diagnostics*, vol. 7, no. 3, pp. 375–387, 2005.

[22] A. J. Redig and P. A. Jänne, “Basket trials and the evolution of clinical trial design in an era of genomic medicine,” *J Clin Oncol*, vol. 33, no. 9, pp. 975–977, 2015.

[23] M. L. Kuijjer, J. N. Paulson, P. Salzman, W. Ding, and J. Quackenbush, “Cancer subtype identification using somatic mutation data,” *Br J Cancer*, vol. 118, no. 11, pp. 1492–1501, 2018.

[24] L. A. Raedler, “Keytruda (pembrolizumab): first PD-1 inhibitor approved for previously treated unresectable or metastatic melanoma,” *Am Health Drug Benefits*, vol. 8, no. Spec Feature, p. 96, 2015.

[25] S. Petrovski and D. B. Goldstein, “Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine,” *Genome Biol*, vol. 17, no. 1, 2016, doi: 10.1186/s13059-016-1016-y.

[26] T. K. Oleksyk, W. W. Wolfsberger, K. Schubelka, S. Mangul, and S. J. O’Brien, “The Pioneer Advantage: Filling the blank spots on the map of genome diversity in Europe,” *Gigascience*, vol. 11, 2022, doi: 10.1093/gigascience/giac081.

[27] A. R. Bentley, S. Callier, and C. N. Rotimi, “Diversity and inclusion in genomic research: why the uneven progress?,” *J Community Genet*, vol. 8, no. 4, 2017, doi: 10.1007/s12687-017-0316-6.

[28] V. A. Schneider *et al.*, “Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly,” *Genome Res*, vol. 27, no. 5, pp. 849–864, 2017.

[29] K. J. Karczewski *et al.*, “The mutational constraint spectrum quantified from variation in 141,456 humans,” *Nature*, vol. 581, no. 7809, pp. 434–443, 2020.

[30] A. Auton *et al.*, “A global reference for human genetic variation. A global reference for human genetic variation,” *Nature*, vol. 526, pp. 68–74, 2015.

[31] A. McKenna *et al.*, “The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data,” *Genome Res*, vol. 20, no. 9, 2010, doi: 10.1101/gr.107524.110.

[32] A. Mayakonda, D. C. Lin, Y. Assenov, C. Plass, and H. P. Koeffler, “Maftools: Efficient and comprehensive analysis of somatic variants in cancer,” *Genome Res*, vol. 28, no. 11, 2018, doi: 10.1101/gr.239244.118.

[33] K. Wang, M. Li, and H. Hakonarson, “ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data,” *Nucleic Acids Res*, vol. 38, no. 16, 2010, doi: 10.1093/nar/gkq603.

[34] P. Cingolani *et al.*, “A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff,” *Fly (Austin)*, vol. 6, no. 2, 2012, doi: 10.4161/fly.19695.

[35] M. J. Landrum *et al.*, “ClinVar: public archive of interpretations of clinically relevant variants,” *Nucleic Acids Res*, vol. 44, no. D1, pp. D862--D868, 2016.

[36] X. Zheng-Bradley, I. Streeter, S. Fairley, D. Richardson, L. Clarke, and P. Flicek, “Alignment of 1000 Genomes Project reads to reference assembly GRCh38,” *GigaScience*, vol. 6, no. 7. 2017. doi: 10.1093/gigascience/gix038.

[37] A. Buniello *et al.*, “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019,” *Nucleic Acids Res*, vol. 47, no. D1, pp. D1005--D1012, 2019.

[38] P. Cingolani *et al.*, “Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift,” *Front Genet*, vol. 3, no. MAR, 2012, doi: 10.3389/fgene.2012.00035.

[39] D. M. Altshuler *et al.*, “An integrated map of genetic variation from 1,092 human genomes,” *Nature*, vol. 491, no. 7422, 2012, doi: 10.1038/nature11632.

[40] A. Bergström *et al.*, “Insights into human genetic variation and population history from 929 diverse genomes,” *Science (1979)*, vol. 367, no. 6484, 2020, doi: 10.1126/science.aay5012.

[41] L. Pagani *et al.*, “Genomic analyses inform on migration events during the peopling of Eurasia,” *Nature*, vol. 538, no. 7624, 2016, doi: 10.1038/nature19792.

[42] S. Mallick *et al.*, “The Simons Genome Diversity Project: 300 genomes from 142 diverse populations,” *Nature*, vol. 538, no. 7624, 2016, doi: 10.1038/nature18964.

[43] D. v Zhernakova *et al.*, “Genome-wide sequence analyses of ethnic populations across Russia,” *Genomics*, vol. 112, no. 1, pp. 442–458, 2020.

[44] G. A. van der Auwera *et al.*, “From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline,” *Curr Protoc Bioinformatics*, no. SUPL.43, 2013, doi: 10.1002/0471250953.bi1110s43.

[45] R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall, “LUMPY: A probabilistic framework for structural variant discovery,” *Genome Biol*, vol. 15, no. 6, 2014, doi: 10.1186/gb-2014-15-6-r84.

[46] M. J. Landrum *et al.*, “ClinVar: improving access to variant interpretations and supporting evidence,” *Nucleic Acids Res*, vol. 46, no. D1, pp. D1062--D1067, 2018.

[47] 1000 Genomes Project Consortium and others, “An integrated map of genetic variation from 1,092 human genomes,” *Nature*, vol. 491, no. 7422, p. 56, 2012.

[48] D. Welter *et al.*, “The NHGRI GWAS Catalog, a curated resource of SNP-trait associations,” *Nucleic Acids Res*, vol. 42, no. D1, pp. D1001--D1006, 2014.

[49] K. L. Korunes and A. Goldberg, “Human genetic admixture,” *PLoS Genet*, vol. 17, no. 3, 2021, doi: 10.1371/journal.pgen.1009374.

[50] S. Gravel *et al.*, “Demographic history and rare allele sharing among human populations,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 29, pp. 11983–11988, 2011.

[51] J. Batt, “Transcarpathia: Peripheral Region at the’centre of Europe’,” *Regional & Federal Studies*, vol. 12, no. 2, pp. 155–177, 2002.

[52] K. Dávid and K. Áron, “Realignment? Transcarpathians in Hungary, Hungarians in Transcarpathia in the light of data up to 2017,” *Teruleti Statisztika*, vol. 60, no. 3, 2020, doi: 10.15196/TS600302.

[53] A. Eisfeld, “Germans of Ukraine in the Interwar Years (1918-1941),” *Problems of World History*, no. 10, 2020, doi: 10.46869/2707-6776-2020-10-7.

[54] I. Csernicskó and A. Beregszászi, “Different states, same practices: visual construction of language policy on banknotes in the territory of present-day Transcarpathia,” *Lang Policy*, vol. 18, no. 2, 2019, doi: 10.1007/s10993-018-9485-3.

[55] N. I. Dnistrianska, M. I. Senkiv, H. Ya. Ilnytska-Hykavchu, M. I. Haba, and O. P. Makar, “Tourism potential of Ukraine in the context of geography and current state of German cultural heritage,” *Journal of Geology, Geography and Geoecology*, vol. 30, no. 4, 2021, doi: 10.15421/112158.

[56] V. P. Miklovda, M. I. Pitiulych, V. v. Hoblyk, M. M. Pityulych, and N. I. Keretsman, “Socio-economic transformations of small rural settlements of Zakarpattia Amidst Decentralisation,” *Int J Criminol Sociol*, vol. 9, 2020, doi: 10.6000/1929-4409.2020.09.375.

[57] A. G. Nikitin, I. T. Kochkin, C. M. June, C. M. Willis, I. Mcbain, and M. Y. Videiko, “Mitochondrial DNA Sequence variation in the boyko, hutsul, and lemko populations of the carpathian highlands,” *Hum Biol*, vol. 81, no. 1, 2009, doi: 10.3378/027.081.0104.

[58] O. M. Utevska, M. I. Chukhraeva, A. T. Agdzhoyan, L. A. Atramentova, E. V. Balanovska, and O. P. Balanovsky, “Populations of Transcarpathia and Bukovina on the genetic landscape of surrounding regions,” *Visnyk of Dnipropetrovsk University. Biology, medicine*, vol. 6, no. 2, 2015, doi: 10.15421/021524.

[59] N. C. Orton, A. M. Innes, A. E. Chudley, and N. T. Bech-Hansen, “Unique disease heritage of the Dutch-German mennonite population,” *American Journal of Medical Genetics, Part A*, vol. 146, no. 8. 2008. doi: 10.1002/ajmg.a.32061.

[60] C. A. Francomano, V. A. McKusick, and L. G. Biesecker, “Medical genetic studies in the Amish: Historical perspective,” *American Journal of Medical Genetics - Seminars in Medical Genetics*, vol. 121 C, no. 1. 2003. doi: 10.1002/ajmg.c.20001.

[61] A. M. Laberge *et al.*, “Population history and its impact on medical genetics in Quebec,” *Clinical Genetics*, vol. 68, no. 4. 2005. doi: 10.1111/j.1399-0004.2005.00497.x.

[62] P. Roudik, “Ukraine: New Territorial Division of the Country Initiated,” *https://www.loc.gov/item/global-legal-monitor/2020-07-21/ukraine-new-territorial-division-of-the-country-initiated/*, Jul. 21, 2020.

[63] D. Kling, C. Phillips, D. Kennett, and A. Tillmar, “Investigative genetic genealogy: Current methods, knowledge and practice,” *Forensic Science International: Genetics*, vol. 52. 2021. doi: 10.1016/j.fsigen.2021.102474.

[64] K. Yuan, Y. Zhou, X. Ni, Y. Wang, C. Liu, and S. Xu, “Models, methods and tools for ancestry inference and admixture analysis,” *Quantitative Biology*, vol. 5, no. 3. 2017. doi: 10.1007/s40484-017-0117-2.

[65] P. G. Meirmans and P. W. Hedrick, “Assessing population structure: FST and related measures,” *Molecular Ecology Resources*, vol. 11, no. 1. 2011. doi: 10.1111/j.1755-0998.2010.02927.x.

[66] B. Cai *et al.*, “Matching phenotypes to whole genomes: Lessons learned from four iterations of the personal genome project community challenges,” *Hum Mutat*, vol. 38, no. 9, 2017, doi: 10.1002/humu.23265.

[67] Uzhhorod National University, “Partnership for genomic research in Ukraine and Romania,” *https://genomes.uzhnu.edu.ua/dataanalysis-results-and-availability/*, Jul. 13, 2023.

[68] R. A. Moran, N. Quesada-Vazquez, A. Sinha, E. De-Madaria, and V. K. Singh, “High Penetrance of the PRSS1 A16V Mutation in a Kindred with SPINK1 N34S and CFTR TG11-5T Co-mutations,” *Pancreas*, vol. 45, no. 1. 2016. doi: 10.1097/MPA.0000000000000445.

[69] C. J. Grocock *et al.*, “The variable phenotype of the p.A16V mutation of cationic trypsinogen (PRSS1) in pancreatitis families,” *Gut*, vol. 59, no. 3, 2010, doi: 10.1136/gut.2009.186817.

[70] J. Rosendahl *et al.*, “CFTR, SPINK1, CTRC and PRSS1 variants in chronic pancreatitis: Is the role of mutated CFTR overestimated?,” *Gut*, vol. 62, no. 4, 2013, doi: 10.1136/gutjnl-2011-300645.

[71] Z. Jancsó *et al.*, “Novel pathogenic PRSS1 variant p.Glu190Lys in a case of chronic pancreatitis,” *Front Genet*, vol. 10, no. FEB, 2019, doi: 10.3389/fgene.2019.00046.

[72] F. U. Weiss, F. Laemmerhirt, and M. M. Lerch, “Next generation sequencing pitfalls in diagnosing trypsinogen (PRSS1) mutations in chronic pancreatitis,” *Gut*, vol. 70, no. 8. 2021. doi: 10.1136/gutjnl-2020-322864.

[73] S. M. Laws *et al.*, “TNF polymorphisms in Alzheimer disease and functional implications on CSF beta-amyloid levels,” *Hum Mutat*, vol. 26, no. 1, 2005, doi: 10.1002/humu.20180.

[74] G. Sirugo, S. M. Williams, and S. A. Tishkoff, “The Missing Diversity in Human Genetic Studies,” *Cell*, vol. 177, no. 1. 2019. doi: 10.1016/j.cell.2019.02.048.

[75] T. M. Waring and Z. T. Wood, “Long-term gene-culture coevolution and the human evolutionary transition,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 288, no. 1952. 2021. doi: 10.1098/rspb.2021.0538.

[76] A. K. Campbell, J. P. Waud, and S. B. Matthews, “The molecular basis of lactose intolerance.,” *Science progress*, vol. 88, no. Pt 3. 2005. doi: 10.3184/003685005783238408.

[77] R. A. Forsgård, “Lactose digestion in humans: Intestinal lactase appears to be constitutive whereas the colonic microbiome is adaptable,” *American Journal of Clinical Nutrition*, vol. 110, no. 2. 2019. doi: 10.1093/ajcn/nqz104.

[78] W. Lukito, S. G. Malik, I. S. Surono, and M. L. Wahlqvist, “From ‘lactose intolerance’ to ‘lactose nutrition,’” *Asia Pacific Journal of Clinical Nutrition*, vol. 24. 2015. doi: 10.6133/apjcn.2015.24.s1.01.

[79] C. L. Storhaug, S. K. Fosse, and L. T. Fadnes, “Country, regional, and global estimates for lactose malabsorption in adults: a systematic review and meta-analysis,” *Lancet Gastroenterol Hepatol*, vol. 2, no. 10, pp. 738–746, 2017.

[80] C. J. E. Ingram, C. A. Mulcare, Y. Itan, M. G. Thomas, and D. M. Swallow, “Lactose digestion and the evolutionary genetics of lactase persistence,” *Hum Genet*, vol. 124, no. 6, pp. 579–591, 2009.

[81] M. C. Campbell and A. Ranciaro, “Human adaptation, demography and cattle domestication: an overview of the complexity of lactase persistence in Africa,” *Hum Mol Genet*, vol. 30, no. R1, pp. R98--R109, 2021.

[82] H. E. Busk, B. Dahlerup, T. Lytzen, B. Juul-Jorgensen, and E. Gudmand-Hoyer, “Incidence of lactose malabsorption among Danish students,” *Ugeskr Laeger*, vol. 137, no. 36, pp. 2062–2064, 1975.

[83] D. P. Bose and J. D. Welsh, “Lactose malabsorption in Oklahoma Indians,” *Am J Clin Nutr*, vol. 26, no. 12, pp. 1320–1322, 1973.

[84] J. D. Johnson *et al.*, “Lactose malabsorption among adult Indians of the Great Basin and American Southwest,” *Am J Clin Nutr*, vol. 31, no. 3, pp. 381–387, 1978.

[85] N. S. Enattah, T. Sahi, E. Savilahti, J. D. Terwilliger, L. Peltonen, and I. Järvelä, “Identification of a variant associated with adult-type hypolactasia,” *Nat Genet*, vol. 30, no. 2, pp. 233–237, 2002.

[86] Y. Wang *et al.*, “The lactase persistence/non-persistence polymorphism is controlled by a cis-acting element,” *Hum Mol Genet*, vol. 4, no. 4, pp. 657–662, 1995.

[87] N. S. Enattah *et al.*, “Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans,” *The american journal of human genetics*, vol. 81, no. 3, pp. 615–625, 2007.

[88] E. Heyer *et al.*, “Lactase persistence in central Asia: phenotype, genotype, and evolution,” *Hum Biol*, vol. 83, no. 3, pp. 379–392, 2011.

[89] C. J. E. Ingram *et al.*, “Multiple rare variants as a cause of a common phenotype: several different lactase persistence associated alleles in a single ethnic group,” *J Mol Evol*, vol. 69, no. 6, pp. 579–588, 2009.

[90] L. Ségurel and C. Bon, "On the Evolution of Lactase Persistence in Humans," *Annu Rev Genomics Hum Genet*, vol. 18, 2017, doi: 10.1146/annurev-genom-091416-035340.

[91] T. S. Plantinga *et al.*, "Low prevalence of lactase persistence in Neolithic South-West Europe," *European journal of human genetics*, vol. 20, no. 7, pp. 778–782, 2012.

[92] A. Krüttli, A. Bouwman, G. Akgül, P. della Casa, F. Rühli, and C. Warinner, "Ancient DNA analysis reveals high frequency of european lactase persistence allele (T-13910) in medieval Central Europe," *PLoS One*, vol. 9, no. 1, 2014, doi: 10.1371/journal.pone.0086251.

[93] T. Bersaglieri *et al.*, "Genetic signatures of strong recent positive selection at the lactase gene," *Am J Hum Genet*, vol. 74, no. 6, 2004, doi: 10.1086/421051.

[94] N. S. Enattah *et al.*, "Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture," *The American Journal of Human Genetics*, vol. 82, no. 1, pp. 57–72, 2008.

[95] J. C. K. Wells, E. Pomeroy, and J. T. Stock, "Evolution of Lactase Persistence: Turbo-Charging Adaptation in Growth Under the Selective Pressure of Maternal Mortality?," *Front Physiol*, vol. 12, 2021, doi: 10.3389/fphys.2021.696516.

[96] A. Szilagyi, "Digestion, absorption, metabolism, and physiological effects of lactose," in *Lactose: Evolutionary Role, Health Effects, and Applications*, 2019. doi: 10.1016/B978-0-12-811720-0.00002-7.

[97] T. He, K. Venema, M. G. Priebe, G. W. Welling, R. J. M. Brummer, and R. J. Vonk, "The role of colonic metabolism in lactose intolerance," *European Journal of Clinical Investigation*, vol. 38, no. 8. 2008. doi: 10.1111/j.1365-2362.2008.01966.x.

[98] M. F. Brandao Gois *et al.*, "Role of the gut microbiome in mediating lactose intolerance symptoms," *Gut*, vol. 71, no. 1. 2022. doi: 10.1136/gutjnl-2020-323911.

[99] M. C. E. Lomer, G. C. Parkes, and J. D. Sanderson, "Review article: Lactose intolerance in clinical practice - Myths and realities," *Alimentary Pharmacology and Therapeutics*, vol. 27, no. 2. 2008. doi: 10.1111/j.1365-2036.2007.03557.x.

[100] L. T. Cavalli-Sforza and A. Strata, "Double-blind study on the tolerance of four types of milk in lactose melabsorbers and absorbers," *Hum Nutr Clin Nutr*, vol. 41, no. 1, 1987.

[101] G. Flatz and H. W. Rotthauwe, "LACTOSE NUTRITION AND NATURAL SELECTION," *The Lancet*, vol. 302, no. 7820, 1973, doi: 10.1016/S0140-6736(73)93267-4.

[102] H. S. Kwak, W. J. Lee, and M. R. Lee, "Revisiting lactose as an enhancer of calcium absorption," *International Dairy Journal*, vol. 22, no. 2. 2012. doi: 10.1016/j.idairyj.2011.09.002.

[103] O. Ó. Sverrisdóttir *et al.*, “Direct estimates of natural selection in iberia indicate calcium absorption was not the only driver of lactase persistence in Europe,” *Mol Biol Evol*, vol. 31, no. 4, 2014, doi: 10.1093/molbev/msu049.

[104] “Yongfa, Wang, et al. ‘Prevalence of primary adult lactose malabsorption in three populations of northern China.’ Human genetics 67 (1984)- 103-106..bib”.

[105] R. R. Townley, K. T. Khaw, and H. Shwachman, “Quantitative assay of disaccharidase activities of small intestinal mucosal biopsy specimens in infancy and childhood.,” *Pediatrics*, vol. 36, no. 6, 1965, doi: 10.1542/peds.36.6.911.

[106] A. Ranciaro *et al.*, “Genetic origins of lactase persistence and the spread of pastoralism in africa,” *Am J Hum Genet*, vol. 94, no. 4, 2014, doi: 10.1016/j.ajhg.2014.02.009.

[107] J. Leichter, “Lactose tolerance in a Slavic population,” *Am J Dig Dis*, vol. 17, no. 1, pp. 73–76, 1972.

[108] M. Jasielska and U. Grzybowska-Chlebowczyk, “Lactose malabsorption and lactose intolerance in children with inflammatory bowel diseases,” *Gastroenterol Res Pract*, vol. 2019, 2019.

[109] “Partnership for genomic research in Ukraine and Romania.” https://genomes.uzhnu.edu.ua

[110] J. R. Bell, “A simple way to treat PCR products prior to sequencing using ExoSAP-IT®,” *Biotechniques*, vol. 44, no. 6, 2008, doi: 10.2144/000112890.

[111] A. Baratloo, S. Safari, M. Elfil, and A. Negida, “Evidence Based Emergency Medicine Part 3: Positive and Negative Likelihood Ratios of Diagnostic Tests.,” *Emerg (Tehran)*, vol. 3, no. 4, 2015.

[112] A. Swift, R. Heale, and A. Twycross, “What are sensitivity and specificity?,” *Evid Based Nurs*, vol. 23, no. 1, 2020, doi: 10.1136/ebnurs-2019-103225.

[113] S. Bhatnagar and R. Aggarwal, “Lactose intolerance,” *Bmj*, vol. 334, no. 7608. British Medical Journal Publishing Group, pp. 1331–1332, 2007.

[114] ООО «ДНК-Технология, “ ГЕНЕТИКА МЕТАБОЛИЗМА ЛАКТОЗЫ,” *http://dna-technology.ru/sites/default/files/laktoza_040-1.pdf*, Jul. 14, 2023.

[115] L. Robles and R. Priefer, “Lactose Intolerance: What Your Breath Can Tell You,” *Diagnostics*, vol. 10, no. 6, p. 412, 2020.

[116] F. Casellas, E. Varela, A. Aparici, M. Casaus, and P. Rodr\’\iguez, “Development, validation, and applicability of a symptoms questionnaire for lactose malabsorption screening,” *Dig Dis Sci*, vol. 54, no. 5, pp. 1059–1065, 2009.

[117] M. A. Azcarate-Peril *et al.*, “Impact of short-chain galactooligosaccharides on the gut microbiome of lactose-intolerant individuals,” *Proc Natl Acad Sci U S A*, vol. 114, no. 3, 2017, doi: 10.1073/pnas.1606722113.

[118] A. I. Lokki *et al.*, “Lactase persistence genotypes and malaria susceptibility in Fulani of Mali,” *Malar J*, vol. 10, 2011, doi: 10.1186/1475-2875-10-9.

[119] I. Järvelä, S. Torniainen, and K. L. Kolho, “Molecular genetics of human lactase deficiencies,” *Annals of Medicine*, vol. 41, no. 8. 2009. doi: 10.1080/07853890903121033.

[120] M. N. Leseva *et al.*, “Differences in DNA Methylation and Functional Expression in Lactase Persistent and Non-persistent Individuals,” *Sci Rep*, vol. 8, no. 1, 2018, doi: 10.1038/s41598-018-23957-4.

[121] M. I. Shevell *et al.*, “Practice parameter: Evaluation of the child with global developmental delay [RETIRED]: Report of the Quality Standards Subcommittee of the American Academy of Neurology and The Practice Committee of the Child Neurology Society,” *Neurology*, vol. 60, no. 3, pp. 367–380, 2003.

[122] M. Srour and M. Shevell, “Genetics and the investigation of developmental delay/intellectual disability,” *Arch Dis Child*, vol. 99, no. 4, pp. 386–389, 2014.

[123] D. Miclea, L. Peca, Z. Cuzmici, and I. V. Pop, “Genetic testing in patients with global developmental delay/intellectual disabilities. A review,” *Clujul Medical*, vol. 88, no. 3, p. 288, 2015.

[124] L. Shan, J. Y. Feng, T. T. Wang, Z. Da Xu, and F. Y. Jia, “Prevalence and Developmental Profiles of Autism Spectrum Disorders in Children With Global Developmental Delay,” *Front Psychiatry*, vol. 12, 2022, doi: 10.3389/fpsyt.2021.794238.

[125] S. A. Bélanger and J. Caron, “Evaluation of the child with global developmental delay and intellectual disability,” *Paediatrics and Child Health (Canada)*, vol. 23, no. 6, 2018, doi: 10.1093/pch/pxy093.

[126] M. Manning, L. Hudgins, P. Practice, G. Committee, and others, “Array-based technology and recommendations for utilization in medical genetics practice for detection of chromosomal abnormalities,” *Genetics in Medicine*, vol. 12, no. 11, pp. 742–745, 2010.

[127] D. Mordaunt, M. Gabbett, M. Waugh, K. O’Brien, and H. Heussler, “Uptake and diagnostic yield of chromosomal microarray in an Australian child development clinic,” *Children*, vol. 1, no. 1, 2014, doi: 10.3390/children1010021.

[128] G. Al-Kasbi *et al.*, “The diagnostic yield, candidate genes, and pitfalls for a genetic study of intellectual disability in 118 middle eastern families,” *Sci Rep*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-22036-z.

[129] J. M. Moreno-Cabrera *et al.*, "Evaluation of CNV detection tools for NGS panenetic diagnostics," *European Journal of Human Genetics*, vol. 28, no. 12, 2020, doi: 10.1038/s41431-020-0675-z.

[130] S. Srivastava *et al.*, "Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders," *Genetics in Medicine*, vol. 21, no. 11, pp. 2413–2421, 2019.

[131] "Annual reports.," *Annual reports.Center of Medical statistics of Ministry of Health Care.*, 2021. http://medstat.gov.ua/ukr/reports.html

[132] T. Vrijenhoek *et al.*, "Whole-exome sequencing in intellectual disability; cost before and after a diagnosis," *European Journal of Human Genetics*, vol. 26, no. 11, pp. 1566–1571, 2018.

[133] S. E. Lincoln *et al.*, "A rigorous interlaboratory examination of the need to confirm next-generation sequencing--detected variants with an orthogonal method in clinical genetic testing," *The Journal of Molecular Diagnostics*, vol. 21, no. 2, pp. 318–329, 2019.

[134] K. Nykamp *et al.*, "Sherloc: a comprehensive refinement of the ACMG--AMP variant classification criteria," *Genetics in Medicine*, vol. 19, no. 10, pp. 1105–1117, 2017.

[135] S. Richards *et al.*, "Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology," *Genetics in medicine*, vol. 17, no. 5, pp. 405–423, 2015.

[136] F. Cunningham *et al.*, "Ensembl 2022," *Nucleic Acids Res*, vol. 50, no. D1, pp. D988--D995, 2022.

[137] M. J. Landrum *et al.*, "ClinVar: public archive of interpretations of clinically relevant variants," *Nucleic Acids Res*, vol. 44, no. D1, pp. D862--D868, 2016.

[138] A. Hamosh, A. F. Scott, J. Amberger, D. Valle, and V. A. McKusick, "Online Mendelian inheritance in man (OMIM)," *Hum Mutat*, vol. 15, no. 1, pp. 57–61, 2000.

[139] T. K. Oleksyk *et al.*, "Genome diversity in Ukraine," *Gigascience*, vol. 10, no. 1, p. giaa159, 2021.

[140] T. K. Oleksyk and W. W. Wolfsberger, "Dataset 'Genome Diversity in Ukraine,'" Dec. 11, 2020. http://gigadb.org/dataset/100835# (accessed Sep. 26, 2021).

[141] "Partnership for genomic research in Ukraine and Romania." https://genomes.uzhnu.edu.ua

[142] P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, and M. Kircher, "CADD: predicting the deleteriousness of variants throughout the human genome," *Nucleic Acids Res*, vol. 47, no. D1, pp. D886--D894, 2019.

[143] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nat Protoc*, vol. 4, no. 7, pp. 1073–1081, 2009.

[144] H. Pekeles, A. Accogli, N. Boudrahem-Addour, L. Russell, F. Parente, and M. Srour, "Diagnostic yield of intellectual disability gene panels," *Pediatr Neurol*, vol. 92, pp. 32–36, 2019.

[145] T. K. Oleksyk *et al.*, "Genome diversity in Ukraine," *Gigascience*, vol. 10, no. 1, 2021, doi: 10.1093/gigascience/giaa159.

[146] T. K. Oleksyk, W. W. Wolfsberger, K. Schubelka, S. Mangul, and S. J. O'Brien, "The Pioneer Advantage: Filling the blank spots on the map of genome diversity in Europe," *Gigascience*, vol. 11, Sep. 2022, doi: 10.1093/gigascience/giac081.

[147] T. Yamamoto *et al.*, "Genomic backgrounds of Japanese patients with undiagnosed neurodevelopmental disorders," *Brain Dev*, vol. 41, no. 9, pp. 776–782, 2019.

[148] L. Gieldon *et al.*, "Diagnostic value of partial exome sequencing in developmental disorders," *PLoS One*, vol. 13, no. 8, p. e0201041, 2018.

[149] A. Peyser, K. Onel, and A. Hershlag, "CONFLICT RESOLUTION: DISCORDANT REPORTING OF CANCER GENE VARIANTS: VUS VS PATHOGENIC," *Fertil Steril*, vol. 118, no. 4, p. e356, 2022.

[150] L. Mohammadi *et al.*, "A simple method for co-segregation analysis to evaluate the pathogenicity of unclassified variants; BRCA1 and BRCA2 as an example," *BMC Cancer*, vol. 9, no. 1, pp. 1–11, 2009.

[151] N. M. Marinakis *et al.*, "Phenotype-driven variant filtration strategy in exome sequencing toward a high diagnostic yield and identification of 85 novel variants in 400 patients with rare Mendelian disorders," *Am J Med Genet A*, vol. 185, no. 8, pp. 2561–2571, 2021.

[152] R. Walsh *et al.*, "Enhancing rare variant interpretation in inherited arrhythmias through quantitative analysis of consortium disease cohorts and population controls," *Genetics in medicine*, vol. 23, no. 1, pp. 47–58, 2021.