genes

MDPI

Article

Inherited and De Novo Variation in Lithuanian Genomes: Introduction to the Analysis of the Generational Shift

Alina Urnikyte *, Laura Pranckeniene *, Ingrida Domarkiene, Svetlana Dauengauer-Kirliene, Alma Molyte, Ausra Matuleviciene, Ingrida Pilypiene and Vaidutis Kučinskas

Department of Human and Medical Genetics, Biomedical Science Institute, Faculty of Medicine, Vilnius University, Santariskiu Street 2, LT-08661 Vilnius, Lithuania; ingrida.domarkiene@mf.vu.lt (I.D.); svetlana.dauengauer-kirliene@mf.vu.lt (S.D.-K.); alma.molyte@mf.vu.lt (A.M.); ausra.matuleviciene@mf.vu.lt (A.M.); ingrida.pilypiene@mf.vu.lt (I.P.); vaidutis.kucinskas@mf.vu.lt (V.K.)
* Correspondence: alina.urnikyte@mf.vu.lt (A.U.); laura.pranckeniene@mf.vu.lt (L.P.)

Abstract: Most genetic variants are rare and specific to the population, highlighting the importance of characterizing local population genetic diversity. Many countries have initiated population-based whole-genome sequencing (WGS) studies. Genomic variation within Lithuanian families are not available in the public databases. Here, we describe initial findings of a high-coverage (an average of 36.27×) whole genome sequencing for 25 trios of the Lithuanian population. Each genome on average carried approximately 4,701,473 (±28,255) variants, where 80.6% (3,787,626) were single nucleotide polymorphisms (SNPs), and the rest 19.4% were indels. An average of 12.45% was novel according to dbSNP (build 150). The WGS structural variation (SV) analysis identified on average 9133 (±85.10) SVs, of which 95.85% were novel. De novo single nucleotide variation (SNV) analysis identified 4417 variants, where 1.1% de novo SNVs were exonic, 43.9% intronic, 51.9% intergenic, and the rest 3.13% in UTR or downstream sequence. Three potential pathogenic de novo variants in the *ZSWIM8*, *CDC42EP1*, and *RELA* genes were identified. Our findings provide useful information on local human population genomic variation, especially for de novo variants, and will be a valuable resource for further genetic studies, and medical implications.

Keywords: whole genome sequencing; SNV; de novo variation; newborns; trios

Citation: Urnikyte, A.; Pranckeniene, L.; Domarkiene, I.; Dauengauer-Kirliene, S.; Molyte, A.; Matuleviciene, A.; Pilypiene, I.; Kučinskas, V. Inherited and De Novo Variation in Lithuanian Genomes: Introduction to the Analysis of the Generational Shift. *Genes* **2022**, *13*, 569. https://doi.org/10.3390/genes13040569

Academic Editors: Aristotelis Chatziioannou and Yudong Zhang

Received: 27 January 2022
Accepted: 22 March 2022
Published: 23 March 2022

1. Introduction

In recent years, the accessibility of whole-genome sequencing (WGS) together with new computational and statistical methods has allowed us to not only to analyze genomic patterns of variation, but also infer de novo mutations and selection events, and unravel their potential phenotypic consequences under different environmental conditions. To identify which region of the human genome might be evolving, we first need to identify mutation prevalence changes across different generations and test their effect on fitness.

Lithuania is a country in the Baltic region of Europe with a population of 2.8 million. Previous research has revealed the Lithuanian population's partial isolation and genetic distinctiveness within the European context [1]. Structure analysis performed by Urnikyte et al., 2021 [2] identified the close genetic proximity of Lithuanians to Latvians, Estonians and Belarusians, with moderate impact of Finno-Ugrians on Lithuanians. Many articles about various inherited or single de novo mutations identified in the Lithuanian families related to specific pathologies can be found in the NCBI database. However, there is no summarizing data that could cover whole-genome sequencing data for the Lithuanian families nor for inherited, nor for the novel genome variants for at least one generation. Summarized de novo germline variants were studied only in one whole exome general Lithuanian population scale study. According to previous (Pranckeniene et al., 2018) research, the rate of de novo variants (DNV) was identified as significantly higher than in other population studies—2.4×10^{-8} and 2.74×10^{-8} for single nucleotides, and for

de novo indels it was 1.77 × 10^{-8} per position per generation [3]. The higher DNV rate was elucidated by the only exome analysis model whereas exomes exhibit significantly higher (by 30%) mutation rates than whole genomes regarding the base pair composition of the whole genome is different from that of exomes. Researches showed that DNAse 1 hypersensitivity, context of CpG islands, GERPP++ conservation values, and expression level explained 68–93% of the DNV rate. There also four possible pathogenic DNVs were found in the genes encoding proteins that are essential for chromatin modeling, regulation of the cytoskeleton, modulation of cell growth and vitality, function of cytoplasmic signaling pathways, and initiation of neuronal response. These variants were not deleterious enough to reduce mean fitness, therefore individuals with these novel variants are healthy.

Many countries have initiated population-based WGS studies [4–6]. However, there are still many geographical regions, for example, Lithuania, lacking genomic information in public-databases. Previous population studies have been performed using exome sequencing [3,7] or genome-wide genotyping [1,8,9] data. To widen our knowledge of patterns of genomic variation in the Lithuanian population, we aimed to analyze high-coverage (an average of 36.27×, see Table S1) genome sequencing data of 25 trios. This study represents the first WGS data analysis of the genome variation within Lithuanian families providing more comprehensive characterization of local human population genomic variation, especially for de novo, and structural variants.

2. Materials and Methods

2.1. Study Population

In this study, Lithuanian families who lived in Lithuania for at least three generations according to the pedigree were included. A total of 35 trios were collected to the study group: 35 newborns and 70 parents (105 individuals in total). Two standardized questionnaires about participants' behavior and medical data were filled by investigators. Only healthy Lithuanian nationality adult participants were included in the cohort. The criteria for the inclusion of newborns were: born to Lithuanian parents, a healthy term-born neonate with a population mean birth weight. Twenty-five trios meeting the established inclusion criteria were selected for the whole-genome sequencing. After digitizing the collected questionnaires and performing the analysis of descriptive statistics, we found that maternal and paternal age was 30.4 (±3.62) and 33.96 (±4.88) years old on average, respectively. All the newborns were screened for whole blood count and leucogram, C reactive protein, pH from the umbilical cord, blood group, Rh factor as well as for inherited abnormalities using head, heart, abdominal and renal ultrasound. Based on the total laboratory and instrumental data, only those who had no pathologies were considered healthy. All 25 newborns enrolled in the study (17 (62%) boys and 8 (32%) girls) were born naturally at a median gestational age of 39.70 (±0.93) weeks in 96% from the primiparous pregnancy. The weight, height, and head circumference of all neonates were in line with the Lithuanian population average regardless of gender. The mean neonatal weight was 3580.8 g (±326.48 g), height was 54.04 cm (±1.65 cm), and head circumference was 35.64 cm ± 0.86 cm.

Demographic and health information of the newborn's parents, data on the pregnancy and childbirth history, and data from a newborn clinical examination and laboratory and instrumental studies were collected. DNA samples from parental venous blood and neonatal umbilical cord blood were collected for whole-genome sequencing. DNA was extracted from whole blood (3 mL) using QIAGEN GENTRA® Puregene® Blood Kit (Qiagen GmbH) according to the manufacturer's protocol. DNA concentration and quality were assessed using NanoDropR ND-1000 spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE, USA).

2.2. DNA Sequencing

Whole-genome sequencing (WGS) was performed for 25 trios of Lithuanian origin at coverage of 26.88–61.38× (an average of 36.27×), Table S1. Overall, the sample size was

75 individuals: 25 newborns, 25 mothers, and 25 fathers. WGS was performed at the CeGaT company (Tubingen, Germany). 100 ng DNA was paired-end sequenced in 2 × 150 bp mode on state-of-the-art Illumina NovaSeq™ 6000 Sequencing System using TruSeq® Nano DNA Library Prep Kit (Illumina Inc., San Diego, CA, USA).

Demultiplexing of the sequencing reads was performed with Illumina bcl2fastq (2.20). Adapters were trimmed with Skewer (version 0.2.2) [10]. Quality trimming of the reads has not been performed. Analysis of sequencing data was performed using the Illumina DRAGEN platform (version 3.6.4). The DRAGEN DNA Pipeline uses the current industry standard, BWA-MEM and GATK-HC software. Reads were mapped to the reference genome hg19 (present on the Illumina DRAGEN platform v.3.6.4) and duplicates were marked. Calling of small variants, regions of homozygosity, and structural variants was performed with default parameters. SNVs found at higher frequencies than 1% in the population were qualified as SNPs. The quality of the FASTQ files was analyzed with FastQC (version 0.11.5-cegat) [11]. Sequencing quality control Q30 values were above 88.59% (Figures S1–S3).

2.3. Structural Variation Detection

Calling of structural variants (SV) such as translocations, inversions, large and medium-sized indels was performed by Illumina DRAGEN platform v.3.6.4 using the same methods as Manta [12]. Two individuals were removed from further analysis as the outliers according to the number of SVs. We have calculated mean, median, mode for deletions, duplications, and insertions in parent and newborn groups. The comparison between groups was performed using R v.4.0.2. [13], with a significance level of 0.05. To determine known and new SVs, SV annotation was performed on *.vcf* format files using AnnotSV software [13] with default parameters. After annotation, we analyzed only SVs with the FILTER status: PASS. Novel SVs were determined according to the annotation from gnomAD [14], ClinVar [15], ClinGen [16], DGV (dgv, nsv or esv) [17], DDD [18], 1000 genomes [19], Ira M. Hall's lab [20], and Children's Mercy Research Institute data [21].

2.4. De Novo Mutation Detection

To detect de novo variants, the whole genomes in *.bam* format of each family that consisted of father, mother, and child were combined using Samtools [22]. Initial identification of de novo variants was performed using the merged trio's *.bam* file and the open-access VarScan v.2.4.4 software [23]. A potential de novo variant is identified if the child has a genomic variant when neither parent has it in the same genome position. The results are provided in a generated *.vcf* format data file. SnpSift v.4 software [24] was used for the initial rejection of false-positive results derived from VarScan [23]. The following conservative filtering criteria were applied: (1) a genotype quality of the individual $\geq$ 50; and (2) the number of reads at each site > 30.

Furthermore, to discard the remaining variants that were somatic (only present in a fraction of the sequenced blood cells) with low allele balance or sequencing artefacts, de novo variants were filtered by setting a threshold for the observed fraction of the reads in individuals with the alternative allele (the allele balance) for the trios [0.3; 0.7]. In cases where de novo variants were detected significantly more frequently than in all other trios, biological paternity verification using WGS data of specific regions was performed.

2.5. Variant Analysis

Prior to the analysis we carried out principal component analysis (PCA) to both sample groups: parents and newborns. PCA was performed with SmartPCA from EIGENSOFT 7.2.1 [25] using independent SNPs obtained with the indep-pairwise option of PLINK v.1.07 [26] with parameters: window size of 50 SNPs, a step size of 5, and a r^2 threshold of 0.5. Genetic relationship was computed with VCFtools (0.1.16) [27] option relatedness. Samples detected as outliers in PCA plot (Figures S4 and S5) as well as individuals with relatedness coefficient higher than expected for unrelated individuals were removed from

further analysis (Tables S5 and S6). To perform the analysis of the variants allele frequency distribution, all identified SNPs were annotated using gnomAD genome database (v.2.1.1) [14] in ANNOVAR [23] software. SNP frequencies determined by the gnomAD genome data were compared with the frequencies of the same SNPs in the Lithuanian population. Data visualization was performed by R software Rcmdr package (version 2.7-2) [28]. The structure analysis was carried out on the merged 1000 Genomes Project Phase3 [19] dataset with the SmartPCA program from EIGENSOFT 7.2.1 [25].

2.6. Variant Annotation

Single nucleotide variant annotation was performed using ANNOVAR v.20210123 [23] using hg19, cytoBand [29], RefSeqGene [30], avsnp150, dbnsfp30a [31], dbnsfp31a_interpro [31], dbnsfp33a [31], exac03 [32], kaviar_20150923 [33] SIFT [34], PolyPhen [32], LRT, MutationTaster [35], MutationAssessor [36], FATHMM [37], PROVEAN [38], CADD [39], GERP++ [40], PhyloP [41], SiPhy [42], and COSMIC [41] algorithm annotation tools. Databases providing information on the gene identifying the de novo variant and the dbSNP database [43] were also added, thus assigning the *rs* code to each genome variant, evaluating the genomic SNVs association with pathogenicity in ClinVar [15], evaluating the 1000 Genome Project Phase 3 dataset [19], and ExAC database [44] genome variant frequencies in different populations. For the de novo variants annotation Gene4Denovo201907 database [45] was used additionally.

3. Results

3.1. Sample Collection

We found that 25 women included in the statistical analysis were 30.4 (±3.62) years of age on average, first-time delivering with higher education, did not smoke, and did not use drugs. Most of them (88%) did not complain about their health condition, a third (36%) did not drink alcohol at all. Overall, 64% of respondents reported consuming alcohol before pregnancy. Additionally, 88% of women took dietary supplements during pregnancy, and 96% of them used folic acid. Prenatal screening for chromosomal abnormalities was performed in 29% of cases. The risk of infection was minimal for most of the pregnant women because the amounts of amniotic fluid were within the normal range in 92%, 80% of group B streptococcal tests were negative, nobody had asymptomatic bacteriuria, and a maximum latency period did not exceed 18 hours. The mean age of the 25 fathers was 33.96 (±4.88) years. Overall, 84% of them had higher education. Most of them (88%) reported consuming alcohol in moderation and 16% smoking. The most detected ABO blood groups were B (44%) and O (32%), and the Rh factor was positive in 72% of cases. An increase in C-reactive protein on the first day after birth was observed in 20.83% of cases, therefore, seven infants underwent recurrent inflammatory markers: CRB and leukogram counts.

3.2. Genomic Variation Characterization

A total of 25 trios of Lithuanian origin were sequenced at coverage of 26.88–61.38× (an average of 36.27×). On average, 94.72% of the reads were mapped to the reference genome hg19. Statistics of mapped reads are shown in Table S1. Sequencing quality control Q30 values were above 88.59% (Figures S1–S3). As presented in Table 1, after variant calling, an average of 4,704,096 SNVs per genome were discovered across autosomes in the parent group and 4,696,226 in the newborn cohort, with a transition/transversion ratio of 2.03, and the heterozygote/homozygote (het/hom) ratio of 1.6 (Tables S2–S4). In total, an average of 446,107 insertions and 433,072 deletions in parents, and 430,017 insertions and 442,233 deletions in newborns were identified (Tables S1–S3). The estimated het/hom ratios were 1.67−1.87 for insertions and 1.74−1.97 for deletions in both groups.

Table 1. The average number of autosomal single nucleotide genetic variants per genome identified in the Lithuanian cohort.

	Parents (n = 50)	Newborns (n = 25)
Raw reads in M	78,447	83,755
Bases (Gb)	11,832	12,633
Coverage depth	35.46	37.88
SNVs	4,704,096	4,696,226
SNPs	3,791,674	3,783,578
Insertions (Hom)	155,784	153,755
Insertions (Het)	277,288	276,262
Deletions (Hom)	152,914	150,766
Deletions (Het)	293,193	291,467
Indels (Het)	22,332	22,377

On average, we identified 3,791,674 SNPs per genome in parents and 3,783,584 SNPs in newborns (Table 1). Of all newborn chromosomes 88.8% SNPs were present in dbSNP (build 150) and 11.2% were novel, and in the parent group 13.7% were novel according to dbSNP (build 150) [43].

We detected on average 120,300 (2.5%) SNPs and indels in X and Y chromosomes. Based on variant distribution across different genomic regions 0.9% SNPs and <49 bp indels are located in exonic regions, 41.5% in intronic, 54.1% in intergenic, and 3.6% in downstream, and upstream sequences of the genome. The distribution of SNVs in genome sections is different: 0.24% is in the exonic regions, 24.5% in the intronic, 30.9% in the intergenic, and 44.4% in the downstream, upstream sequences of the genome (Figures S6 and S7). Each variant was annotated using ANNOVAR software [23] (see Section 2).

The analysis of the variant allele frequency collected in the gnomAD [14] database, compared to the distribution of the same genome variants in Lithuanian genomes, showed that the Lithuanian population is distinguished by alleles whose frequency in the gnomAd database is identified as rare or unique. This is best reflected by comparing allele frequencies between the Lithuanian and African American, Finish, uncertain and Ashkenazi Jewish ancestry genomes (Figure S8). An exclusive comparison of results of the Lithuanian genome variants with Ashkenazi Jewish (AJN) ancestry genomes indicates that about half of the common variants with frequency 0.05–0.5 present in AJN genomes, in Lithuanian's genomes becomes rare allelic variants with minor allele frequency (MAF) < 0.05. Assessing the distribution of MAF for the Lithuanian population data set of 1,508,407 SNPs identified 4.3% SNVs are rare with MAF $\leq$ 0.01, and 9.5% are low frequency (MAF 0.01–0.05), and the rest are common (MAF > 0.05) (Figure 1).

To infer population structure we combined Lithuanian and 1000 Genomes Project Phase3 [19] dataset, generating a pooled dataset of 242,188 autosomal SNPs in a total of 2553 individuals. We analyzed 28 populations from main geographical regions: Africa including the Yoruba in Ibadan, Nigeria (YRI), Luhya in Webuye, Kenya (LWK), Gambian in Western Divisions in the Gambia (GWD), Mende in Sierra Leone (MSL), and Esan in Nigeria (ESN) populations; Europe including Utah residents with ancestry from northern and western Europe (CEU), Toscani in Italy (TSI), Finnish in Finland (FIN), British in England and Scotland (GBR), Lithuanians (LT), Iberian population from Spain (IBS); East Asia including Han Chinese in Bejing, China (CHB), Japanese in Tokyo, Japan (JPT), Southern Han Chinese, China (CHS), Chinese Dai in Xishuangbanna, China (CDX), Kinh in Ho Chi Minh City, Vietnam (KHV), Denver Chinese in Denver, Colorado (CHD); South Asia including Gujarati Indians in Houston, Texas (GIH), Punjabi from Lahore, Pakistan (PJL), Bengali from Bangladesh (BEB), Sri Lankan Tamil from the UK (STU) and Indian Telugu from the UK (ITU); America including African American in Southwest US (ASW),

African Caribbean in Barbados (ACB), Mexican-American in Los Angeles, California (MXL), Puerto Rican in Puerto Rico (PUR), Colombian in Medellin, Colombia (CLM), and Peruvian in Lima, Peru (PEL).

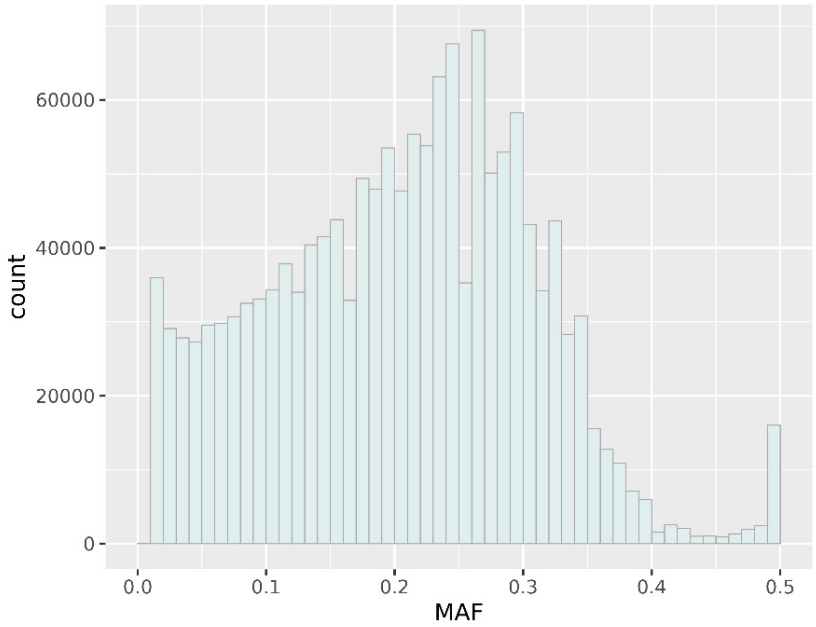

Figure 1. The minor allele frequency distribution for 1,508,407 SNPs in the Lithuanian population samples.

The first two PCs explained 57.32% and 25.34% of the variance, respectively. The results show all populations clustered according their continental origin (Figure 2).

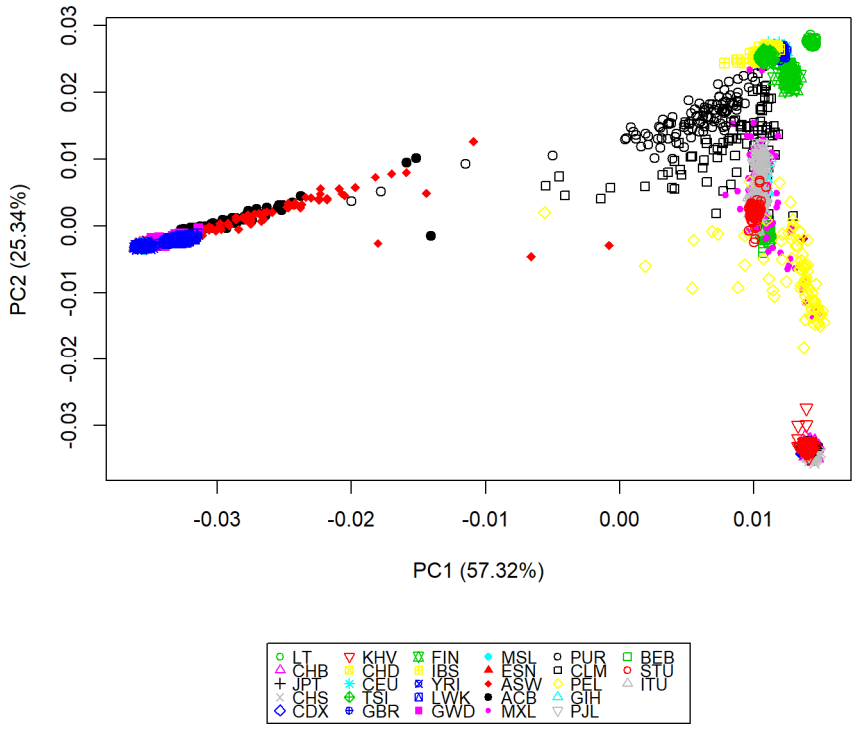

Figure 2. Principal component analysis of the first two PCs of individuals from Lithuania and 27 populations from the 1000 Genomes Project Phase3 dataset. Abbreviations as indicated in the text.

3.3. Structural Variation

Summary statistics for structural variation (>49 bp) in parent and newborn groups are presented in Table 2. There was no statistically significant difference among parents and newborns in the analyzed structural variant groups. The average number of SVs was 9133 (not including newborns, as they inherit most of their variation from parents). On average there were 4159 (92% novel) deletions, 349 (93% novel) duplications, and 4621 (99% novel) insertions. Deletions and insertions are more abundant if compared with the number of duplications.

Table 2. Summary statistics for deletions, duplications, and insertions in parent and newborn groups.

	Parents (*n* = 49)	**Newborns (*n* = 24)**	
Statistics	**Deletions**		***p*-Value**
Mean	4159.39	4191.33	0.5611
SE	33.63	47.08	
Median	4126.0	4172.5	
Mode	3963	#N/A	
SD	23,543	23,064	
	Duplications		
Mean	34,884	35,408	0.2747
SE	3.94	4.76	
Median	346	356	
Mode	365	355	
SD	27.61	23.31	
	Insertions		
Mean	4620.49	4661.54	0.5033
SE	49.98	63.05	
Median	4612.0	4697.5	
Mode	#N/A	#N/A	
SD	34,983	30,889	

SE—standard error, SD—standard deviation.

3.4. De Novo Mutation Discovery

We further performed de novo variant (DNV) analysis. An exceptionally high number of de novo variants were identified for two trios (no. 2 and 4): 21,127 and 44,641, respectively. Their data were evaluated as an exclusion, thus, data for these trios were excluded from further analysis. In the final set of 23 trios, on average 158 single nucleotide variants and 34 indels (<49 bp) were identified. All participants had at least few de novo variants. Two de novo single nucleotides and two de novo indels were placed in chromosomes X and Y.

Analysis of 4417 DNVs identified by VarScan software showed that on average 1.1% de novo SNVs were exonic (37), 43.9% intronic (69), 51.9% intergenic (83), and the rest 3.13% in UTR or downstream sequence. Analysis of de novo indels revealed 1.07% exonic, 42.7% intronic, 48.4% intergenic, and 2.4% variants in UTR or downstream sequence (Figure 3).

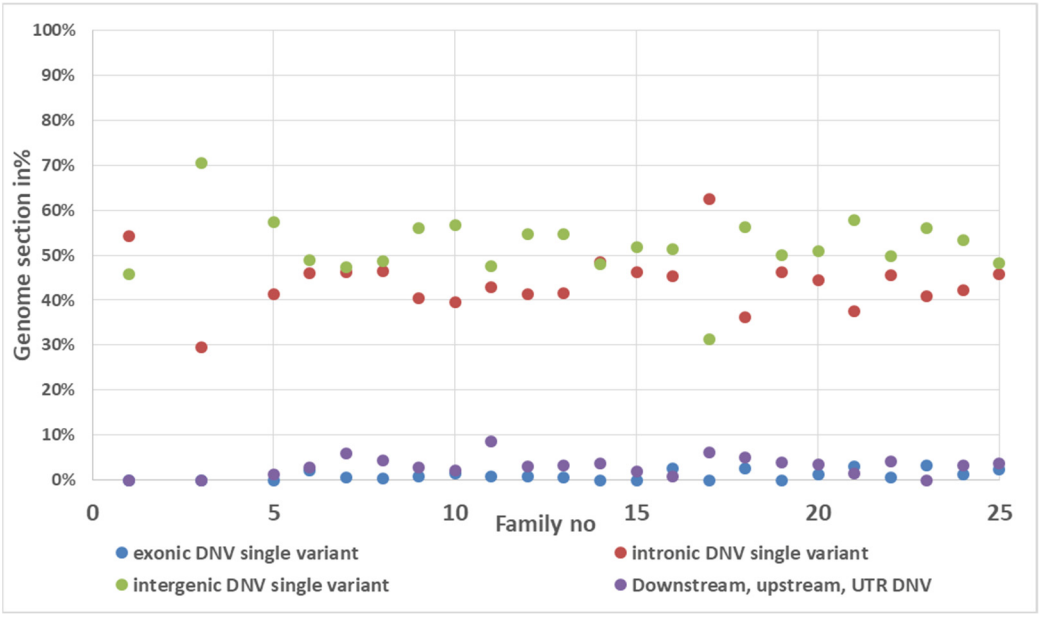

Figure 3. The distribution of de novo indels in genome regions according genome sequence function.

To assess whether there is a potentially pathogenic variant among the identified DNVs, predicted categorical scores for the damage induced by DNVs were analyzed. The following 10 values were considered: polyphen HDIV and HVAR, LRT, PROVEAN, CADD, FATHMM, MutationTester, MutationAssessor, SIFT, Fathmm-MKL coding, and GERP++. According to evaluated pathogenicity scores, three DNVs scored five or more estimates as pathogenic or potentially pathogenic therefore were identified them as possibly pathogenic: *ZSWIM8* (NM_001242487:c.3814G>C; p.G1272R), *CDC42EP1* (NM_152243:c.763C>A; p.P255T; rs77417880), and *RELA* (NM_001243985:c.1265A>G; p.N422S, rs746519095). All three DNVs are in a homozygous state.

4. Discussion

To study human genetics for many purposes, researchers intended to create a fully mapped sequence of the human genome and initiated the Human Genome Project (HGP) in 1990 [46]. Since then, the human reference genome has provided the foundation for genetic discovery and research, but recently, multiple authors of papers, such as Kaye A.M. et al., 2021 [47], Ballouz S. et al., 2019 [48], and Yang X. et al., 2019 [49], have highlighted the deficiencies of the linear reference, leading to a growing consensus that a richer reference structure is needed [47–49]. Continued improvements in the era of widespread whole-genome sequencing must improve the ability to predict how an individual's inherited genome contributes to aging, complex disease, and even some monogenic diseases [50]. Furthermore, de novo mutations have increasingly been proposed to affect disease onset and progression [45]. As we move towards population-specific sequencing studies, the reference genome is no longer sufficiently static, with personalized reference genomes providing more accurate analysis results [51].

Here, in this article, we summarized the high-coverage WGS data obtained for this study that is analyzed and reported for the Lithuanian population for the first time. The data is of high quality with 36X coverage on average across the entire read length of 2 × 150 bp, with coverage distribution almost identical across all samples. WGS produced up to 757.973 M effective reads per sample (approximately 121 GB data per single sample run) (see Figure S2). A duplication rate of 5.21% in our WGS data is twice lower than in previously reported 10.12% Glanzmann et al. data [52], which indicates high levels of coverage for a target sequence, whereas high duplication rate indicates an enrichment bias.

Primary analysis of the all variant allele frequency showed that the Lithuanian population distinguishes and is unique by alleles whose frequency in the gnomAd database is

identified as rare or unique. The comparison results of the Lithuanian genome variants with Ashkenazi Jewish (AJN) ancestry genomes revealed that about half of the common variants with frequency 0.05–0.5 present in AJN genomes, in Lithuanian's genomes become rare allelic variants with MAF < 0.05 (Figure S8). PCA analysis identified that Lithuanians position within the European context.

Novel SNVs and indel variants in 75 individuals were defined in the group of newborns and parents as absent from dbSNP build 150, respectively, 11.2% and 13.7%. Our study detected 4417 novel variants (SNs and indels) for 23 individuals, demonstrating the genetic diversity present in Lithuanian individuals. This finding underscores the value of sequencing Lithuanian individuals, as it allows the comprehensive cataloging and characterization of variants, which will in the future aid the clinical interpretation of genetic results [52].

In addition, during the study, we analyzed de novo variants. Aware of the fact that all individuals of the general population have only potentially neutral variants because survey subjects assessed themselves as "healthy" (although they may become ill in the future), we identified three potential pathogenic de novo variants in the *ZSWIM8*, *CDC42EP1*, and *RELA* genes for three different families. According to protein function prediction, *ZSWIM8* codes Zinc Finger SWIM-Type containing protein. Although we know that no pathological phenotype is expressed in the examined individual, disease such as Acromelic Frontonasal Dysostosis is associated with *ZSWIM8*. An important paralog of this gene is *ZSWIM6*. The frequency of identified NM_001242487: c.3814G>C rare allele was not previously reported. *CDC42EP1* codes for serum protein MSE55, which is a non-kinase CRIB (Cdc42/Rac interactive-binding) domain-containing molecule of unknown function. These findings indicate that MSE55 is a Cdc42 effector protein that mediates actin cytoskeleton reorganization at the plasma membrane. The third DNV was in the *RELA* gene, which encodes NF-kappa-B, a ubiquitous transcription factor involved in several biological processes. It is held in the cytoplasm in an inactive state by specific inhibitors. Upon degradation of the inhibitor, NF-kappa-B moves to the nucleus and activates transcription of specific genes. NF-kappa-B is composed of NFKB1 or NFKB2 bound to either REL, RELA, or RELB. The most abundant form of NF-kappa-B is NFKB1 in a complex with the product of this gene, RELA. Four transcript variants encoding different isoforms have been found for this gene. Despite the pathogenicity prediction the persistence of the variant in the genome and the absence of effect phenotypes are due to many factors, therefore, it is necessary to examine the variant in appropriate conditions and of course to confirm by Sanger sequencing. Those include environment, age of parents, genomic context, epigenetics, and other factors because all of them influence the value of mean relative fitness that increases monotonically, whereas the strength of selection decreases [53].

In summary, WGS follows the current trends in the convergence of fundamental genomic research and clinical implications of the presence or absence of certain genes. Now, WGS is becoming one of the most widely used applications and is providing tremendous quantities of genome sequences relative to the past through public and private human genome sequencing projects throughout the world. The ability to compare and contrast genomic findings across a unified database containing information from diverse populations is crucial to advancing our understanding of the human genome [47].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes13040569/s1, Figure S1: Sequence lengths of trimmed FASTQ reads (average of all samples) (DNA); Figure S2: Sequence quality of trimmed FASTQ reads (average of all samples) (DNA); Figure S3: GC content of trimmed FASTQ reads (average of all samples) (DNA); Figure S4. Principal component analysis (PCA) of 50 individuals (parents) from Lithuania included in the study. Figure S5. Principal component analysis (PCA) of 25 newborns from Lithuania included in the study. Figure S6. Summary of the identified variants. Distribution of SNPs and indels. Figure S7. Summary of the identified variants. Distribution of structural variants. Figure S8. Allele frequency distribution comparison between the Lithuanian and African American, Latino ancestry, Finish, uncertain, Ashkenazi Jewish, and east Asian ancestry genomes. Table S1:

Statistics of mapped reads (DNA); Tables S2–S4: Number of identified variants (passing QC filter); Table S5: Relatedness analysis performed for trios; Table S6: Detected outliers in our study.

Author Contributions: Methodology, formal analysis and writing—original draft preparation, A.U., L.P., I.D., S.D.-K., A.M. (Alma Molyte), A.M. (Ausra Matuleviciene) and I.P.; supervision, V.K.; project administration, A.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Research Council of Lithuania (LMTLT), grant number S-MIP-20-34.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee, No. 2020/6-1243-724, 22-06-2022.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The WGS variation data have been deposited on https://figshare.com/articles/dataset/Inherited_and_de_novo_variation_in_Lithuanian_genomes_introduction_to_the_analysis_of_the_generational_shift/19354817 (accessed on 14 March 2022).

Acknowledgments: We would like to acknowledge Vilnius University Hospital Santaros Klinikos, Center of Obstetrics and Gynecology, Vilnius, Lithuania, for parental venous blood and neonatal umbilical cord blood collection, newborn screening for whole blood count and leucogram, C reactive protein, pH from the umbilical cord, blood group, Rh factor as well as for inherited abnormalities using head, heart, abdominal and renal ultrasound. We thank all the study participants, who volunteered in this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Urnikyte, A.; Flores-Bello, A.; Mondal, M.; Molyte, A.; Comas, D.; Calafell, F.; Bosch, E.; Kučinskas, V. Patterns of Genetic Structure and Adaptive Positive Selection in the Lithuanian Population from High-Density SNP Data. *Sci. Rep.* **2019**, *9*, 9163. [CrossRef] [PubMed]
2. Urnikyte, A.; Molyte, A.; Kučinskas, V. Genome-Wide Landscape of North-Eastern European Populations: A View from Lithuania. *Genes* **2021**, *12*, 1730. [CrossRef] [PubMed]
3. Pranckėnienė, L.; Jakaitienė, A.; Ambrozaitytė, L.; Kavaliauskienė, I.; Kučinskas, V. Insights Into. *Front. Genet.* **2018**, *9*, 315. [CrossRef] [PubMed]
4. Huang, J.; Howie, B.; McCarthy, S.; Memari, Y.; Walter, K.; Min, J.L.; Danecek, P.; Malerba, G.; Trabetti, E.; Zheng, H.F.; et al. Improved Imputation of Low-Frequency and Rare Variants Using the UK10K Haplotype Reference Panel. *Nat. Commun.* **2015**, *6*, 8111. [CrossRef] [PubMed]
5. Hindorff, L.A.; Bonham, V.L.; Brody, L.C.; Ginoza, M.E.C.; Hutter, C.M.; Manolio, T.A.; Green, E.D. Prioritizing Diversity in Human Genomics Research. *Nat. Rev. Genet.* **2018**, *19*, 175–185. [CrossRef] [PubMed]
6. Han, E.; Sinsheimer, J.S.; Novembre, J. Characterizing Bias in Population Genetic Inferences from Low-Coverage Sequencing Data. *Mol. Biol. Evol.* **2014**, *31*, 723–735. [CrossRef]
7. Pranckėnienė, L.; Kučinskas, V. The Relative Fitness of the de Novo Variants in General Lithuanian Population vs. in Individuals with Intellectual Disability. *Eur. J. Hum. Genet.* **2021**, *30*, 332–338. [CrossRef]
8. Urnikyte, A.; Domarkiene, I.; Stoma, S.; Ambrozaityte, L.; Uktveryte, I.; Meskiene, R.; Kasiulevičius, V.; Burokiene, N.; Kučinskas, V. CNV Analysis in the Lithuanian Population. *BMC Genet.* **2016**, *17*, 64. [CrossRef]
9. Urnikytė, A.; Molytė, A.; Kučinskas, V. Recent Effective Population Size Estimated from Segments of Identity by Descent in the Lithuanian Population. *Anthropol. Sci.* **2017**, *125*, 53–58. [CrossRef]
10. Jiang, H.; Lei, R.; Ding, S.W.; Zhu, S. Skewer: A Fast and Accurate Adapter Trimmer for next-Generation Sequencing Paired-End Reads. *BMC Bioinform.* **2014**, *15*, 182. [CrossRef]
11. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010. Available online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 1 March 2022).
12. Chen, X.; Schulz-Trieglaff, O.; Shaw, R.; Barnes, B.; Schlesinger, F.; Källberg, M.; Cox, A.J.; Kruglyak, S.; Saunders, C.T. Manta: Rapid Detection of Structural Variants and Indels for Germline and Cancer Sequencing Applications. *Bioinformatics* **2016**, *32*, 1220–1222. [CrossRef] [PubMed]
13. Geoffroy, V.; Herenger, Y.; Kress, A.; Stoetzel, C.; Piton, A.; Dollfus, H.; Muller, J. AnnotSV: An Integrated Tool for Structural Variations Annotation. *Bioinformatics* **2018**, *34*, 3572–3574. [CrossRef] [PubMed]
14. Karczewski, K.J.; Francioli, L.C.; Tiao, G.; Cummings, B.B.; Alföldi, J.; Wang, Q.; Collins, R.L.; Laricchia, K.M.; Ganna, A.; Birnbaum, D.P.; et al. The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans. *Nature* **2020**, *581*, 434–443. [CrossRef] [PubMed]

15. Landrum, M.J.; Lee, J.M.; Riley, G.R.; Jang, W.; Rubinstein, W.S.; Church, D.M.; Maglott, D.R. ClinVar: Public Archive of Relationships among Sequence Variation and Human Phenotype. *Nucleic Acids Res.* **2014**, *42*, D980–D985. [CrossRef] [PubMed]
16. Rehm, H.L.; Berg, J.S.; Brooks, L.D.; Bustamante, C.D.; Evans, J.P.; Landrum, M.J.; Ledbetter, D.H.; Maglott, D.R.; Martin, C.L.; Nussbaum, R.L.; et al. ClinGen-the Clinical Genome Resource. *N. Engl. J. Med.* **2015**, *372*, 2235–2242. [CrossRef]
17. MacDonald, J.R.; Ziman, R.; Yuen, R.K.C.; Feuk, L.; Scherer, S.W. The Database of Genomic Variants: A Curated Collection of Structural Variation in the Human Genome. *Nucleic Acids Res.* **2013**, *42*, D986–D992. [CrossRef]
18. Firth, H.V.; Richards, S.M.; Bevan, A.P.; Clayton, S.; Corpas, M.; Rajan, D.; Van Vooren, S.; Moreau, Y.; Pettett, R.M.; Carter, N.P. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **2009**, *84*, 524–533. [CrossRef]
19. Sudmant, P.H.; Rausch, T.; Gardner, E.J.; Handsaker, R.E.; Abyzov, A.; Huddleston, J.; Zhang, Y.; Ye, K.; Jun, G.; Fritz, M.H.; et al. An Integrated Map of Structural Variation in 2,504 Human Genomes. *Nature* **2015**, *526*, 75–81. [CrossRef]
20. Abel, H.J.; Larson, D.E.; Regier, A.A.; Chiang, C.; Das, I.; Kanchi, K.L.; Layer, R.M.; Neale, B.M.; Salerno, W.J.; Reeves, C.; et al. Mapping and Characterization of Structural Variation in 17,795 Human Genomes. *Nature* **2020**, *583*, 83–89. [CrossRef]
21. Children's Mercy Research Institute Data. Available online: https://grch38.warehouse.cmh.edu/ (accessed on 1 March 2022).
22. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. Subgroup, 1000 Genome Project Data Processing the Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
23. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data. *Nucleic Acids Res.* **2010**, *38*, e164. [CrossRef] [PubMed]
24. Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W1118; Iso-2; Iso-3. *Fly Austin* **2012**, *6*, 80–92. [CrossRef] [PubMed]
25. Patterson, N.; Price, A.L.; Reich, D. Population Structure and Eigenanalysis. *PLoS Genet.* **2006**, *2*, e190. [CrossRef] [PubMed]
26. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.R.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.W.; Daly, M.J.; et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [CrossRef]
27. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; et al. The Variant Call Format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [CrossRef]
28. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2014.
29. Robinson, J.T.; Thorvaldsdóttir, H.; Winckler, W.; Guttman, M.; Lander, E.S.; Getz, G.; Mesirov, J.P. Integrative Genomics Viewer. *Nat. Biotechnol.* **2011**, *29*, 24–26. [CrossRef]
30. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation. *Nucleic Acids Res.* **2016**, *44*, D733–D745. [CrossRef]
31. Liu, X.; Wu, C.; Li, C.; Boerwinkle, E. DbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum. Mutat.* **2016**, *37*, 235–241. [CrossRef]
32. Adzhubei, I.A.; Schmidt, S.; Peshkin, L.; Ramensky, V.E.; Gerasimova, A.; Bork, P.; Kondrashov, A.S.; Sunyaev, S.R. A Method and Server for Predicting Damaging Missense Mutations. *Nat. Methods* **2010**, *7*, 248–249. [CrossRef]
33. Glusman, G.; Caballero, J.; Mauldin, D.E.; Hood, L.; Roach, J.C. Kaviar: An Accessible System for Testing SNV Novelty. *Bioinformatics* **2011**, *27*, 3216–3217. [CrossRef]
34. Vaser, R.; Adusumalli, S.; Leng, S.N.; Sikic, M.; Ng, P.C. SIFT Missense Predictions for Genomes. *Nat. Protoc.* **2016**, *11*, 1–9. [CrossRef] [PubMed]
35. Schwarz, J.M.; Cooper, D.N.; Schuelke, M.; Seelow, D. MutationTaster2: Mutation Prediction for the Deep-Sequencing Age. *Nat. Methods* **2014**, *11*, 361–362. [CrossRef] [PubMed]
36. Reva, B.; Antipin, Y.; Sander, C. Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics. *Nucleic Acids Res.* **2011**, *39*, e118. [CrossRef] [PubMed]
37. Shihab, H.A.; Gough, J.; Mort, M.; Cooper, D.N.; Day, I.N.; Gaunt, T.R. Ranking Non-Synonymous Single Nucleotide Polymorphisms Based on Disease Concepts. *Hum. Genom.* **2014**, *8*, 11. [CrossRef] [PubMed]
38. Choi, Y.; Sims, G.E.; Murphy, S.; Miller, J.R.; Chan, A.P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE* **2012**, *7*, e46688. [CrossRef]
39. Kircher, M.; Witten, D.M.; Jain, P.; O'Roak, B.J.; Cooper, G.M.; Shendure, J. A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants. *Nat. Genet.* **2014**, *46*, 310–315. [CrossRef]
40. Davydov, E.V.; Goode, D.L.; Sirota, M.; Cooper, G.M.; Sidow, A.; Batzoglou, S. Identifying a High Fraction of the Human Genome to Be under Selective Constraint Using GERP++. *PLoS Comput. Biol.* **2010**, *6*, e1001025. [CrossRef]
41. Pollard, K.S.; Hubisz, M.J.; Rosenbloom, K.R.; Sepal, A. Detection of Nonneutral Substitution Rates on Mammalian Phylogenies. *Genome Res.* **2010**, *20*, 110–121. [CrossRef]
42. Garber, M.; Guttman, M.; Clamp, M.; Zody, M.C.; Friedman, N.; Xie, X. Identifying Novel Constrained Elements by Exploiting Biased Substitution Patterns. *Bioinformatics* **2009**, *25*, i54–i62. [CrossRef]
43. Sherry, S.T.; Ward, M.; Sirotkin, K. DbSNP-Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. *Genome Res.* **1999**, *9*, 677–679. [CrossRef]

44. Karczewski, K.J.; Weisburd, B.; Thomas, B.; Solomonson, M.; Ruderfer, D.M.; Kavanagh, D.; Hamamsy, T.; Lek, M.; Samocha, K.E.; Cummings, B.B.; et al. The ExAC Browser: Displaying Reference Data Information from over 60 000 Exomes. *Nucleic Acids Res.* **2017**, *45*, D840–D845. [CrossRef]
45. Zhao, G.; Li, K.; Li, B.; Wang, Z.; Fang, Z.; Wang, X.; Zhang, Y.; Luo, T.; Zhou, Q.; Wang, L.; et al. Gene4Denovo: An Integrated Database and Analytic Platform for de Novo Mutations in Humans. *Nucleic Acids Res.* **2019**, *48*, D913–D926. [CrossRef] [PubMed]
46. Watson, J.D. The Human Genome Project: Past, Present, and Future. *Science* **1990**, *248*, 44–49. [CrossRef] [PubMed]
47. Kaye, A.M.; Wasserman, W.W. The Genome Atlas: Navigating a New Era of Reference Genomes. *Trends Genet.* **2021**, *37*, 807–818. [CrossRef] [PubMed]
48. Ballouz, S.; Dobin, A.; Gillis, J.A. Is It Time to Change the Reference Genome? *Genome Biol.* **2019**, *20*, 159. [CrossRef]
49. Yang, X.; Lee, W.P.; Ye, K.; Lee, C. One Reference Genome Is Not Enough. *Genome Biol.* **2019**, *20*, 104. [CrossRef]
50. Zahn, L.M. The Human Genome. *Science* **2021**, *373*, 1458–1459. [CrossRef]
51. Grytten, I.; Rand, K.D.; Nederbragt, A.J.; Sandve, G.K. Assessing Graph-Based Read Mappers against a Baseline Approach Highlights Strengths and Weaknesses of Current Methods. *BMC Genom.* **2020**, *21*, 282. [CrossRef]
52. Glanzmann, B.; Jooste, T.; Ghoor, S.; Gordon, R.; Mia, R.; Mao, J.; Li, H.; Charls, P.; Douman, C.; Kotze, M.J.; et al. Human Whole Genome Sequencing in South Africa. *Sci. Rep.* **2021**, *11*, 606. [CrossRef]
53. Peck, J.R.; Waxman, D. What Is Adaptation and How Should It Be Measured? *J. Theor. Biol.* **2018**, *447*, 190–198. [CrossRef]